Critical Reviews in Biochemistry and Molecular Biology, 37(3):121–147 (2002)

Utility of Low-Copy Nuclear Gene Sequences in Plant Phylogenetics

Tao Sang

Department of Plant Biology, Michigan State University, East Lansing, MI 48824 (sang@msu.edu; Ph. 517-355-4689; Fax 517-353-1926)

ABSTRACT: Low-copy nuclear genes in plants are a rich source of phylogenetic information. They hold a great potential to improve the robustness of phylogenetic reconstruction at all taxonomic levels, especially where universal markers such as cpDNA and nrDNA are unable to generate strong phylogenetic hypotheses. Low-copy nuclear genes, however, remain underused in plant phylogenetic studies due to practical and theoretical complications in unraveling the evolutionary dynamics of nuclear gene families. The lack of the universal markers or universal PCR primers of low-copy nuclear genes has also hampered their phylogenetic utility. It has recently become clear that low-copy nuclear genes are particularly helpful in resolving close interspecific relationships and in reconstructing allopolyploidization in plants. Gene markers that are widely, if not universally, useful have begun to emerge. Although utilizing low-copy nuclear genes usually requires extra lab work such as designing PCR primers, PCR-cloning, and/or Southern blotting, rapid accumulation of gene sequences in the databases and advances in cloning techniques have continued to make such studies more feasible. With the growing number of theoretical studies devoted to the gene tree and species tree problem, a solid foundation for reconstructing complex plant phylogenies based on multiple gene trees began to build. It is also realized increasingly that fast evolving introns of the low-copy nuclear genes will provide much needed phylogenetic information around the species boundary and allow us to address fundamental questions concerning processes of plant speciation. Phylogenetic and molecular evolutionary analyses of developmentally important genes will add a new dimension to systematic and evolutionary studies of plant diversity.

I. INTRODUCTION

Despite rapid advances of molecular phylogenetics, the goal of accurate reconstruction of the tree of life remains challenging. A major difficulty stems from the incongruence between gene phylogenies and the underlying organismal phylogeny. One way to identify this potential problem and appropriate solutions is to compare phylogenies of multiple unlinked genes (Hillis, 1995; Wendel and Doyle, 1998). Plant phylogenetic studies, however, still rely on a few universal molecular markers, primarily sequences of the chloroplast and nuclear ribosomal DNA. Although these markers have yielded enormous insights into the plant phylogeny, access to additional gene markers becomes increasingly important to the improvement of resolution and accuracy of plant phylogenetic reconstruction.

Slow rates of sequence divergence of chloroplast DNA have limited its phyloge-

1040-9238/02/$.50

netic utilities to high taxonomic levels. Although some rapidly evolving genes, such as *matK* and *ndhF*, as well as some noncoding regions of cpDNA have been used at low taxonomic levels, resolutions were often unsatisfactory. While phylogenetic utility of nuclear ribosomal DNA spans a wider taxonomic range, relationships among distantly related genera or closely related species remain poorly covered by nrDNA (Soltis and Soltis, 1998). For distantly related genera, the small and large subunits of nrDNA are usually too conserved to provide a sufficient resolution, whereas sequences of the internal transcribed spacers (ITS) have diverged too much to be aligned. The nrDNA ITS region has contributed tremendously to the understanding of relationships of congeneric species and closely related genera. Yet, ITS sequence variation appears to be inadequate for the study of closely related species or intraspecific relationships (Baldwin et al., 1995).

Another challenge to plant biologists is the reconstruction of hybrid speciation. The inheritance pathway of the chloroplast genome is predominantly uniparental in plants, and usually maternal in angiosperms. Consequently, cpDNA, as a single linkage group, traces the genealogy of one parent and thus is unable to provide direct evidence for hybrid speciation. Although nrDNA is biparentally inherited, it is not a reliable marker for reconstructing hybrid speciation because distinct nrDNA sequences derived from both parents can be homogenized by concerted evolution of the nrDNA gene family (Wendel et al., 1995).

The growing examples in the recent literature have demonstrated that low-copy nuclear genes have a great potential to compensate cpDNA and nrDNA for the improvement of resolution and robustness of plant phylogenetic reconstruction (e.g., Doyle and Doyle, 1999). In particular, low-copy nuclear genes provide a suitable marker for reconstruction of allopolyploidization (Small et al., 1998; Sang and Zhang, 1999; Doyle et al., 2000). Furthermore, the large number of genes in the nuclear genome serves as a virtually unlimited source of phylogenetic markers that can offer numerous independent estimates of the organismal phylogeny.

However, the phylogenetic utility of low-copy nuclear genes has been confounded by the complex evolutionary dynamics of nuclear gene families (Clegg et al., 1997). For example, gene duplication and deletion can potentially lead to the reconstruction of gene duplication events (paralogy) rather than speciation events (orthology). The extra effort required to disentangle orthology and paralogy apparently has discouraged the common application of low-copy nuclear genes in plant phylogenetics.

This article first reviews the recent plant phylogenetic studies using sequences of low-copy nuclear genes and discusses advantages and limitations concerning phylogenetic utility of low-copy nuclear genes at various taxonomic levels. It then focuses on the illumination of theoretical and practical problems associated with phylogenetic utility of low-copy nuclear genes. Important questions such as when and how many nuclear genes are needed for an accurate phylogenetic reconstruction of certain plant groups are addressed. Finally, it touches on the implications of low-copy nuclear gene phylogenies that are beyond the immediate goal of reconstructing organismal phylogenies.

II. PHYLOGENETIC UTILITY OF LOW-COPY NUCLEAR GENES

Most protein coding nuclear genes of plants have exons and introns. Exons, which

are usually under strong purifying selection that eliminates deleterious mutations, have relatively slow rates of nucleotide substitution. Unlike chloroplast genes, introns of low-copy nuclear genes have rather high rates of nucleotide substitution (Gaut, 1998; Li, 1998). A nuclear gene with regions diverging at variable rates can potentially provide phylogenetic markers at various taxonomic levels.

A. Interfamilial and Higher Levels

Phylogenetic reconstruction of major plant lineages has been difficult due to a high level of homoplasy and the problem of long branch attraction. Sequences of nuclear ribosomal DNA, chloroplast genes, and mitochondrial genes have proven useful at high taxonomic levels because they evolve at relatively slow rates (Qiu et al., 1999). Although much progress has been made recently toward a robust reconstruction of early diversification of vascular plants and angiosperms, controversy remains among studies using different genes or gene combinations (Pryer et al., 2001; Graham and Olmstead, 2000; Nickrent et al., 2000).

Few low-copy nuclear genes have been used to estimate relationships at such high taxonomic levels. Kolukisaoglu et al. (1995) investigated early evolution of vascular plants based on sequences of the phytochrome gene. The gene phylogeny inferred from substitutions at the first and second codon positions suggested that *Selaginella* and *Equisetum* were earlier vascular plant lineages than *Psilotum*. This result agrees in part with the evidence of chloroplast genome rearrangement, which suggested that *Selaginella* diverged prior to *Psilotum* (Raubeson and Jansen, 1992). Duplication and deletion of the phytochrome genes during the diversification of seed plants, however, prevented an assessment of relationships between gymnosperms and angiosperms. *Ephedra* was grouped with the A or C type of phytochrome gene of angiosperms, whereas conifers were grouped with the B type phytochrome gene of angiosperms.

The phytochrome genes, however, served as a useful marker for understanding the early diversification of angiosperms. Because the phytochrome genes were duplicated shortly before the diversification of angiosperms, they provided an opportunity for rooting the angiosperm phylogeny between the paralogous genes *PhyA* and *PhyC* (Mathews and Donoghue, 1999). The rooting strategy facilitated the identification of *Amborella* as the sister group of the remaining angiosperms, which is consistent with recent analyses of sequences of cpDNA, nrDNA, and mtDNA (Qiu et al., 1999; Soltis et al., 1999).

Denton et al. (1998) tested phylogenetic utility of the gene encoding RNA polymerase II (RPB2) by sampling nine angiosperm families, *Ginkgo*, and *Marchantia*. The phylogeny, inferred from amino acid sequences of the RPB2 gene, supported the monophyly of angiosperms and eudicots. The phylogenetic utility of this gene is being examined further with additional taxonomic sampling covering the major lineages of seed plants (B. D. Hall, personal communication).

Clearly, we are still at an early stage of testing the phylogenetic utility of low-copy nuclear genes at high taxonomic levels. The large number of low-copy nuclear genes would potentially allow selection of genes with extremely conserved rates of evolution, and consequently a robust reconstruction of deep-branch relationships of plants. Slow-evolving nuclear genes will also provide phylogenetic markers for the reconstruction of the tree of life, where certain

lineages lack chloroplast or mitochondrial genomes.

B. Intergeneric Level

In comparison with the interfamilial and interspecific levels, nuclear ribosomal DNA has been less frequently utilized for phylogenetic studies at the intergeneric level. This is probably because sequences of the small and large subunits of nrDNA diverge too slowly to yield sufficient phylogenetic information among closely related genera, whereas the internal transcribed spacers diverge too rapidly to be aligned unambiguously among distantly related genera. Low-copy nuclear genes have provided useful phylogenetic markers at the intergeneric level.

Galloway et al. (1998) tested the phylogenetic utility of the arginine decarboxylase gene (*Adc*) in Brassicaceae. Two loci of the *Adc* gene, *Adc1* and *Adc2*, were identified. Exon sequences of the genes were analyzed for 10 genera representing the major lineages of the Brassicaceae. Relationships among these genera are well resolved and supported and are congruent between the *Adc1* and *Adc2* phylogenies. The *Adc* genes have markedly higher sequence divergence at the synonymous sites (0.99% for Adc1 and 1.02% for Adc2) than the cpDNA *ndhF* gene (0.19%). While the *Adc* phylogenies are topologically identical to the *ndhF* phylogeny, the *Adc* genes provide stronger bootstrap support for almost all branches in comparison with the *ndhF* phylogeny.

Intergeneric relationships of the Poaceae were examined based on the phylogeny of the granule-bound starch synthase gene (*waxy*), a single-copy gene in grasses (Mason-Gamer et al., 1998). The phylogeny inferred from the exon sequences was well resolved but with relatively low bootstrap support. The intergeneric relationships of the *waxy* phylogeny are largely congruent with those inferred from morphological characters and cpDNA sequences. In Rosaceae, the *waxy* gene has been duplicated and exists at least two copies in a diploid genome (Evans et al., 2000). The *waxy* phylogeny of Rosaceae is largely congruent with the chloroplast gene phylogenies.

A portion of exon 1 of the phytochrome B gene (*PhyB*), another single copy gene in grasses, was used to reconstruct intergeneric relationships within Poaceae (Mathews and Sharrock, 1996; Mathews et al., 2000). The *PhyB* phylogeny had significantly improved resolution and support compared to the previous phylogenetic hypotheses of the family based on the chloroplast genes *ndhF* and *rbcL*. This may be attributed to faster sequence divergence of the *PhyB* genes compared with the chloroplast genes. Simmons et al. (2001) also used the exon region of the *PhyB* gene to investigate relationships within Celastraceae and found that the *PhyB* sequences provided particularly strong support for the relationships among closely related genera.

Wang et al. (2000) studied intergeneric relationships of Pinaceae using a low-copy nuclear gene encoding 4-coumarate : coenzyme A ligase (4CL) in the lignin biosynthetic pathway. All 4CL sequences from species of the same genus formed a strongly supported monophyletic group, whereas sequences cloned from an individual species did not always form a monophyletic group. This suggests that the 4CL gene has a short turnover time of duplication/deletion such that paralogous loci are maintained between species but not between genera. Therefore, the 4CL gene served as a useful phylogenetic marker for studying intergeneric relationships of Pinaceae. The phylogeny generated from 4CL gene sequences is congruent with those from the chloroplast *matK* gene and mitochondrial

nad5 gene. Of the three genes, the nuclear 4CL gene evolved most rapidly. The average sequence divergence of exon regions of the 4CL gene is approximately twice as great as that of the *matK* gene and five times that of the *nad5* gene.

C. Interspecific Level

At the interspecific level, sequences of chloroplast DNA, including noncoding regions, usually diverge too slowly to resolve close relationships. The nrDNA ITS region has been the only widely used sequence data at the interspecific level in plant phylogenetic studies (Baldwin et al., 1995). However, ITS sequence variation is not always sufficient to resolve closely related species. Thus, nuclear genes with rapidly evolving introns, are needed for a full resolution of interspecific phylogenies. Furthermore, because interspecific relationships of plants are often complicated by hybridization or introgression, multiple unlinked nuclear markers are often needed to ensure an accurate phylogenetic reconstruction.

Sequences of low-copy nuclear genes have contributed to a better understanding of interspecific relationships of various plant groups. Alcohol dehydrogenase genes (*Adh*) were used to study relationships within the genera *Paeonia* and *Gossypium*. The *Adh* phylogenies (*Adh1* and *Adh2*) of *Paeonia* are better resolved than those of ITS or the cpDNA *matK* gene and intergenic spacers (Sang et al., 1997a, 1997b). For the 11 diploid *Paeonia* species compared, sequence divergence of the *Adh* introns (4.68% for *Adh1* and 4.54% for *Adh2*) was approximately five times higher that of the *matK* gene (0.85%) and one and half times higher than the ITS region (3.11%) (Sang et al., 1997b).

Small et al. (1998) compared the phylogenetic utility of cpDNA noncoding regions and an *Adh* locus, *AdhC*, in *Gossypium*. They sampled more than 7 kb from seven noncoding regions of the chloroplast genome. Only one region, the *trnT-trnL* spacer, contained phylogenetically informative characters, accounting for only 0.05% of the total amount of nucleotides sampled from the chloroplast genome. Consequently, relationships among the five closely related tetraploid species were poorly resolved and weakly supported in the cpDNA phylogeny. A 1.6-kb region of the *AdhC* gene, containing approximately an equal amount of exon and intron sequences, was sampled from the same species. Approximately about 0.76% of the *Adh* sequences was phylogenetically informative. The resulting gene phylogeny was well resolved and strongly supported. The average sequence divergence of the *AdhC* gene was also higher than that of the nrDNA ITS region (~0.5%) between these species.

Another *Adh* locus, *AdhA*, was sequenced for investigating relationships of 13 D-genome diploid species of *Gossypium* (Small and Wendel, 2000b). Southern blotting suggested that the gene is present as a single copy in the majority of the species, but has been duplicated in four species of subsection *Erioxylum*. The gene phylogeny is well resolved and largely consistent with the monophyly of each subsection recognized in the current taxonomy. Within subsection *Erioxylum*, however, interspecific relationships are unresolved because sequences from each of the four species failed to form a monophyletic group. A combination of factors, including gene flow, unrecognized paralogy, and lineage sorting, may be responsible for the result.

The gene encoding glycerol-3-phosphate acyltransferase (*Gpat*) is a single-copy gene in plant species belonging to several eudicot families. Testing the phylogenetic utility of *Gpat* in *Paeonia* suggested that the gene is

likely present as a single copy in the majority of diploid *Paeonia* species (Tank and Sang, 2001). Phylogenetic analyses of a portion of the *Gpat* gene, including a large intron of more than 2 kb, yielded a well-resolved and strongly supported phylogeny. It resolved interspecific relationships that were previously unresolved on the ITS, *matK*, or *Adh* gene phylogenies. The better resolution of the *Gpat* gene phylogeny was apparently attributed to a sufficient amount of phylogenetic information yielded from the large intron that has diverged more rapidly than introns of the *Adh1* gene in *Paeonia*.

Emshwiller and Doyle (1999) explored the phylogenetic utility of the chloroplast-expressed glutamine synthetase gene (*ncpGS*) in the genus *Oxalis*. An approximately 670-bp region that contains about two-thirds intron sequence was sequenced from eight *Oxalis* species. The sequence divergence of the entire gene region (6.57%) is slightly lower than that of the ITS region (7.75%), whereas the sequence divergence of the introns (8.63%) is higher than that of the ITS region. Although the *ncpGS* gene region and ITS had an equivalent amount of phylogenetically informative characters, the former yielded more autapomorphies that may be informative when taxonomic sampling is increased.

A portion of the intron of a MADS-box gene, *pistillata*, was sequenced for the reconstruction of relationships within the genus *Sphaerocardamum* of Brassicaceae (Bailey and Doyle, 1999). The intron has higher sequence divergence (0.15 to 3.7%) than the nrDNA ITS region (0 to 2.5%) and cpDNA *trnL* intron (0 to 2.4%). The interspecific relationships of the genus were fully resolved by the *pistillata* intron sequences.

The entire coding region of the *PgiC* genes was used to infer phylogenetic relationships in the genus *Clarkia* (Onagraceae) (Gottlieb and Ford, 1996). The average sequence divergences among the *Clarkia* species, representing six sections of the genus are 3% for exons of *PgiC1*, 2.5% for exons of *PgiC2*, 7.7% for introns of *PgiC1*, and 7.2% for introns of *PgiC2*. Intersectional relationships on each gene tree were completely resolved and strongly supported.

A region of the *Vicilin* gene that contains approximately 0.7 kb exon and 0.4 kb intron sequences provided reasonably good resolution and support for relationships within and between two closely related genera, *Theobroma* and *Herriania*, of Sterculiaceae (Whitlock and Baum, 1999). The comparison of sequence divergences between the two genera indicated that the *Vicilin* gene evolved 5 to 10 times more rapidly than the chloroplast *ndhF* gene. A 450-bp region of the histone H3-D gene that contains two-thirds intron sequences provided sufficient resolution of relationships among the major groups of diploid species in *Glycine* subg. *Glycine* (Doyle et al., 1996, 1999).

D. Intraspecific Level

A robust phylogenetic reconstruction around the species boundary plays an increasingly important role in addressing fundamental evolutionary questions concerning speciation and adaptation. Unlike animals, plant organellar genes diverge too slowly to resolve population-level relationships. Low-copy nuclear genes with rapidly evolving introns provide an appealing source of DNA sequence data for phylogenetic reconstruction at the intraspecific level (Schaal and Olsen, 2000). In particular, low-copy nuclear gene phylogenies have already shed light on the evolution of a few crop plants.

Olsen and Schaal (1999) studied the domestication of cassava (*Manihot esculenta* subsp. *esculenta*) using sequences of a

single-copy nuclear gene *G3pdh*. They found that the level of phylogenetically informative variation in the *G3pdh* region far exceeded levels typically observed in the organellar genomes of plants at the intraspecific level. Comparison of *G3pdh* haplotypes of the crop and wild species indicated that *Manihot esculenta* subsp. *flabellifolia* alone could account for the genetic variation observed in cassava. The gene genealogy further suggested that cassava was domesticated from populations of this subspecies along the southern border of the Amazon basin and rejected the previous hypothesis that cassava was a compilospecies derived from one or more complexes of interbreeding wild species in the Neotropics.

Two single-copy nuclear genes, *Adh1* and *Glb1*, were sequenced to investigate maize (*Zea mays* spp. *mays*) domestication (Eyre-Walker et al., 1998; Hilton and Gaut, 1999). Phylogenetic reconstruction of both genes revealed that alleles of maize intermixed with those of *Zea mays* ssp. *parviglumis*, but clearly separated from clades containing alleles from *Zea luxurians*. The results suggest that the alleles of maize have not coalesced since its domestication from populations of *Zea mays* ssp. *parviglumis* 7500 years ago. On the other hand, alleles of both genes have reached coalescence between *Z. luxurians* and *Z. mays*, which diverged from each other 0.6 to 0.7 million years ago.

The phylogeny of a single-copy nuclear gene, *cl*, that regulates anthocyanin biosynthesis, presented a different picture of relationships among *Zea* species (Hanson et al., 1996). Maize shared haplotypes of the gene with *Z. mays* ssp. *parviglumis*, *Z. mays* ssp. *mexicana*, *Z. luxurians*, and *Z. diploperennis*. None of the *Zea* species or subspecies formed a monophyletic group on the *cl* gene phylogeny. Furthermore, the *cl* phylogeny suggested that maize is genetically more similar to *Z. mays* ssp. *mexicana* than its presumed wild progenitor *Z. mays* ssp. *parviglumis*. Apparently, the *cl* locus has experienced more extensive lineage sorting than either *Glb1* or *Adh1* locus. Introgression of this regulatory gene among the *Zea* species and subspecies may have also contributed to the complex phylogenetic relationships.

E. Duplicate Gene Rooting

In addition to the independent phylogenetic markers, low-copy nuclear genes offer a unique opportunity to solve difficult phylogenetic problems. Gene duplication events can serve as points where phylogenetic trees are rooted when appropriate outgroups are unavailable. If gene duplication occurred prior to the diversification of the ingroup taxa, the gene tree can be rooted at the gene duplication event, that is, between the phylogenies of the two paralogous genes. A good example of this application is the reconstruction of the tree of life by rooting between the paralogous gene pairs that were duplicated prior to the diversification of all major lineages of life (Brown and Doolittle, 1995; Lawson et al., 1996; Iwabe et al., 1989).

Duplicated-gene rooting can also be applied to groups that have uncertain or isolated systematic positions. The monotypic family Paeoniaceae has been placed in different subclasses of angiosperms based on morphological and molecular data (e.g., Soltis et al., 2000) and seems to be distantly related to its extant relatives. Consequently, it is difficult to find molecular markers that have diverged fast enough to resolve relationships within the genus *Paeonia* and are alignable between the ingroup and a distantly related outgroup. By rooting the phylogeny of *Paeonia* between a paralogous gene pair, *Adh1* and *Adh2* (based on exon sequence only), Sang et al. (1997b) were

able to identify that the earliest divergence within the genus was between the shrubby section *Moutan* and the herbaceous sections *Paeonia* and *Oneapia*. Subsequent analysis of sequences of both exons and introns of each *Adh* gene using section *Moutan* as a functional outgroup resolved the interspecific relationships of *Paeonia*.

Another recent application of this rooting strategy helped clarify the early diversification of angiosperms. Angiospermae may have undergone rapid radiation during the early diversification, and its closest relatives may be extinct. Thus, it has been difficult to resolve relationships of basal angiosperm lineages using distantly related outgroups, such as gymnosperms. By rooting the angiosperm phylogenies between *PhyA* and *PhyC*, a pair of paralogs that duplicated shortly before the diversification of angiosperms, the basal lineages of angiosperms were resolved (Mathews and Donoghue, 1999).

F. Reconstruction of Polyploidization

The accurate reconstruction of hybrid speciation has long been difficult in phylogenetics (Funk, 1985; McDade, 1995). Biparentally inherited nuclear markers are needed for the reconstruction of allopolyploidization. The most popular nuclear marker, nrDNA, however, is unreliable for this purpose because the hybrid genome could quickly homogenize one of the homoeologs through concerted evolution (Wendel et al., 1995). Low-copy nuclear genes, which are less susceptible to concerted evolution, can potentially serve as a very useful marker for reconstructing allopolyploidization (Small et al., 1998; Sang and Zhang, 1999).

Utility of low-copy nuclear genes in combination with the cloning procedure, one can also get around the analytical difficulty of direct reconstruction of reticulate evolution. By cloning the homoeologous loci derived from both parents, an allotetraploid genome is dissected into two units, with each homoeolog tracing to its own parental lineage. Phylogenetic analysis of the cloned homoeologs together with the genes of the putative parents converts the reconstruction of reticulate evolution into the reconstruction of the diverged histories of the parental lineages (e.g., Sang and Zhang, 1999).

Phylogenies of low-copy nuclear genes have proven effective in determining the origin of allotetraploids in several groups of flowering plants. The homoeologous loci of 16 nuclear genes were recovered from the allotetraploid species *Gossypium hirsutum*, with each pair of homoeologous loci forming sister groups with the corresponding diploid progenitors (Cronn et al., 1999). Moreover, the homoeologs have evolved independently of each other and at the same rate as those of their diploid progenitors, suggesting a genic stasis following polyploidization. For certain allotetraploid species of *Gossypium* and *Paeonia* that have fixed nrDNA ITS sequences of one parent, their allotetraploid origins were reconstructed successfully by *Adh* phylogenies (Small et al., 1998; Sang and Zhang, 1999).

Phylogenies of *Adh1* and *Adh2* genes provided evidence for the hybrid origin of tetraploid *P. officinalis* between allotetraploid peonies (Ferguson and Sang, 2001). Three distinct types of *Adh* sequences were identified from both accessions of *P. officinalis*. Two types were most closely related to the two homoeologous *Adh* loci of the allotetraploid *P. arietina* group, and the remaining type came from one of the two *Adh*

homoeologs of the allotetraploid species *P. peregrina*. The other *Adh* homoeolog of *P. peregrina* was apparently lost from the hybrid genome, possibly through backcrossing with the *P. arietina* group. This is the first documentation of homoploid hybrid speciation between allotetraploid species in nature. The study suggests that low-copy nuclear gene sequences can help unravel phylogenies of polyploid complexes.

In *Oryza*, *Adh* gene phylogenies supported the previous hypotheses of the allotetraploid nature of species with BBCC, CCDD, and JJHH genomes. They further suggested that the diploid EE genome species is the closest extant relative of DD-genome progenitor of the CCDD species (Ge et al., 1999). *Adh* and cpDNA

matK phylogenies provided evidence for the multiple origins of *Oryza* species with the BBCC genome. Based on the *Adh* gene phylogenies, the new genome type HHKK was assigned to *Oryza schlechteri* and *Poterisia corctata*.

The *PgiC* gene sequences were utilized to infer the allotetraploid origin of *Clarkia gracilis*. It was found that all homoeologous loci donated from the diploid parents were expressed in the allotetraploid species (Ford and Gottlieb, 1999). Origins of different morphological forms of an allotetraploid species, *Glycine tabacina*, were reconstructed through phylogenetic analysis of sequences of a single-copy nuclear gene histone H3-D (Doyle et al., 2000). The allotetraploid origin of the Hawaiian silversword alliance was revealed by phylogenies of two floral homeotic genes, *ASAP3/TM6* and *ASAP1* (Barrier et al., 1999). For each gene, two homoeologous loci identified from the Hawaiian silversword alliance were most closely related to those of the putative diploid parents in the genus *Raillardiopsis*.

III. THEORETICAL CONCERNS

The above examples have demonstrated the advantages of utilizing low-copy nuclear genes in plant phylogenetic studies. Low-copy nuclear genes remain underused in comparison to cpDNA and nrDNA due primarily to complex evolutionary dynamics of nuclear gene families. The following discussion attempts to elucidate, from a theoretical point of view, potential problems associated with the phylogenetic utility of low-copy nuclear genes.

A. Paralogy

A commonly recognized complication is the history of gene duplication and deletion. If there has been gene duplication prior to speciation, and it is followed by random deletions of the gene copies from the descendent lineages, the gene phylogeny may trace the gene duplication event, or paralogy, rather than the speciation. Theoretically, paralogy problems should be more frequent at higher taxonomic levels because more cycles of gene duplication/deletion could have occurred during a longer period of time. Obviously, this undermines the phylogenetic utility of low-copy nuclear genes at high taxonomic levels. A gene that yields paralogous relationships at high taxonomic levels, however, may serve as a useful marker for phylogenetic reconstruction at the lower levels. This is illustrated in the following hypothetical example.

A low-copy nuclear gene has undergone duplication and deletion during the diversification of a group of taxa, A through F (Figure 1a). Each terminal taxon maintains two copies of the gene. If all gene copies are sequenced and a gene tree is correctly in-

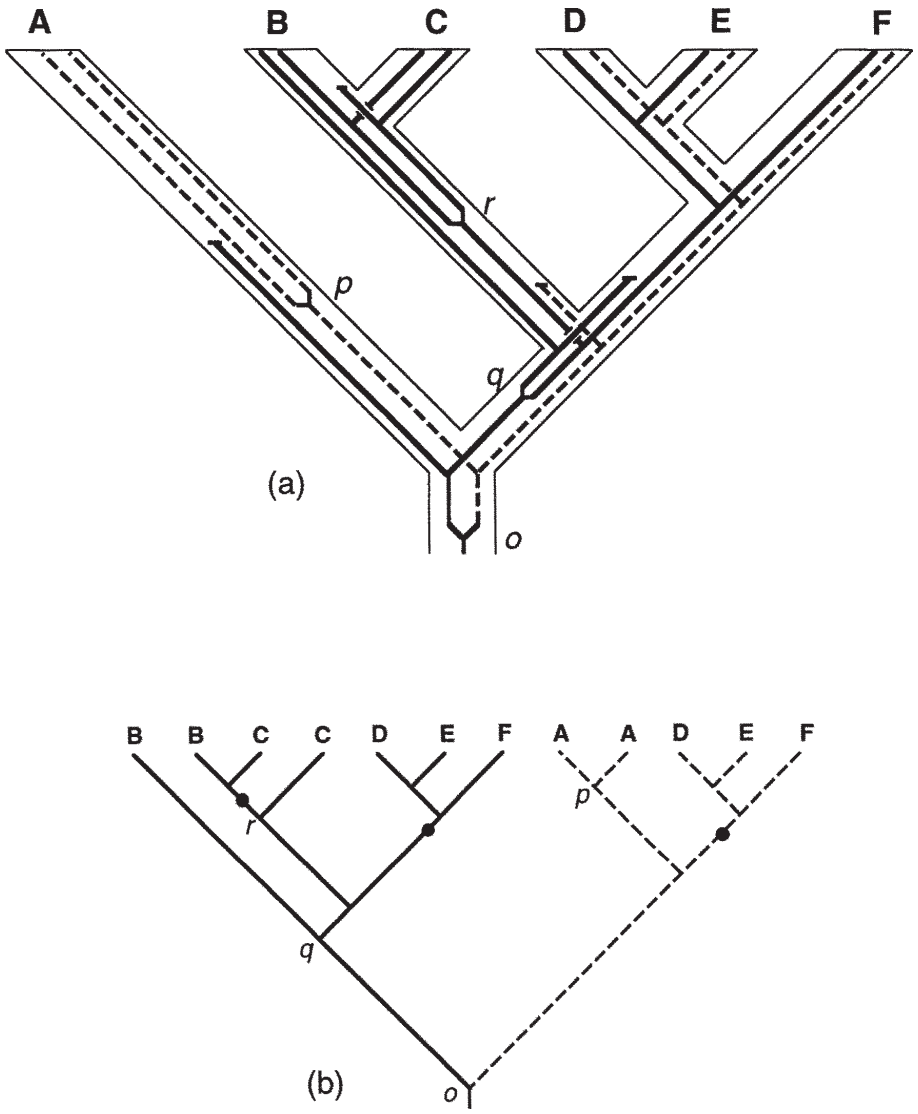

FIGURE 1. Species tree and gene tree of taxa A, B, C, D, E, and F. (a) Species tree (outlined by thin solid lines) and a contained gene phylogeny (illustrated by thick lines). (b) Gene tree. Letters, o, p, q, and r indicate gene duplication events. Solid and dashed lines represent two gene copies duplicated in the common ancestor of these taxa. Black dots on the gene tree indicate the clades that represent orthologous relationships.

ferred (Figure 1b), the gene tree contains both paralogous and orthologous relationships. Orthologous relationships are restricted to the close relationships, or lower taxonomic levels. The gene phylogeny fails to reflect the true phylogenetic relationships of deeper branches.

A number of nuclear gene families displayed this kind phylogenetic behavior in the empirical studies. The *Adh* gene served as a good marker for phylogenetic reconstruction at the interspecific level, but displayed extensive paralogy at the interfamilial level (Small and Wendel, 2000b). The MADS-box genes have a very dynamic history of duplication and deletion during the evolution of flowering plants (e.g., Purugganan et al., 1995). However, certain members of the gene family served as useful single-copy markers for phylogenetic reconstruction at the interspecific level, for example, *ASAP3/TM6* and *ASAP1* genes for the Hawaiian silversword alliance (Barrier et al., 1999) and the *pistillata* gene for *Sphaerocardamum* species (Bailey and Doyle, 1999). The chalcone synthase gene (*Chs*) has had a relatively high rate of duplication throughout angiosperm evolution (Clegg et al., 1997; Durbin et al., 2000), but served as a useful marker for resolving relationships among five tribes of the Brassicaceae (Koch et al., 2001).

While it may be generally true that sorting out orthologous relationships is more difficult at high taxonomic levels, there might be exceptions for genes that have a fast cycle of duplication and deletion. If the turnover time of gene duplication and deletion is fast and the deep branches that separate groups of terminal taxa are relatively long, paralogy may become concentrated at the short terminal branches. For example, the 4CL gene was a good nuclear marker for reconstructing relationships of distantly related genera of Pinaceae, but showed paralogous relationships within certain genera (Wang et al., 2000). These theoretical and empirical examples demonstrate that choosing an appropriate nuclear gene is critical to the phylogenetic reconstruction at different taxonomic levels.

B. Lineage Sorting or Deep Coalescence

Lineage sorting or deep coalescence poses another problem for the phylogenetic utility of low-copy nuclear genes. Lineage sorting occurs as a result of the random fixation of ancestral polymorphic alleles in descendent taxa. The way that lineage sorting contributes to topological incongruence between gene trees and the species tree is somewhat analogous to that of paralogy (Maddison, 1997). Unlike the paralogy problem, the chance of lineage sorting increases as divergence time between the studied taxa decreases. Therefore, lineage sorting poses the most challenging problem for phylogenetic inference at the interspecific and intraspecific levels.

Because coalescence time of nuclear genes is four times longer than organellar genes, the latter are less susceptible to lineage sorting or deep coalescence (Moore, 1995). Unfortunately, sequences of chloroplast and mitochondrial genomes evolve too slowly in plants to resolve close relationships. Thus, we face the dilemma that low-copy nuclear genes seem to be the only promising phylogenetic markers that can provide sufficient resolution at the populational level, but they usually undergo lineage sorting or deep coalescence at this level.

However, low-copy nuclear genes can be very useful markers for addressing evolutionary questions around the species boundary even though alleles have not completely coalesced within a taxon. For example, despite the polyphyly of alleles of the *G3pdh* gene from the cassava popula-

tions, the closest relationship of the haplotypes to those of *Manihot esculenta* subsp. *flabellifolia* suggested that cassava was domesticated from the populations at the southern border of the Amazon basin (Olsen and Schaal, 1999). The phylogenies of *Adh1* and *Glb1* genes showed that alleles of maize are intermixed with those of its wild progenitor *Zea mays* ssp. *parviglumis*, but clearly separated from those of *Zea luxurians* (Hilton and Gaut, 1999).

By the same token, low-copy nuclear genes can be useful markers for understanding processes of natural speciation. For example, when species A is derived from a few populations or individuals of a widespread species B, such as founder speciation, a low-copy nuclear phylogeny can trace alleles of species A to the donor alleles of species B without a requirement of coalescence within each species. This approach follows the same idea of identifying the progenitor and derivative relationship based on isozyme electrophoresis (Crawford, 1990), but uses markers that are much more sensitive than isozyme electrophoresis, which only determines nonsynonymous substitutions that change the eletrophoretic migration of proteins.

C. Hybridization

Hybrid speciation, through allopolyploidization or homoploid hybridization, is widely documented in angiosperms and ferns (Grant, 1981; Arnold, 1997). Utilization of low-copy nuclear genes has already enhanced our understanding of allopolyploidy in a number of plant groups, and will play an increasingly important role in the phylogenetic reconstruction of allopolyploidization in plants. There has been considerable theoretical discussion on issues related to the inference of hybrid speciation from gene trees (Rieseberg and Morefield, 1995; Doyle, 1997; Wendel and Doyle, 1998; Sang and Zhang, 1999; Sang and Zhong, 2000). Development of theoretical approaches that deal with this difficult phylogenetic problem, especially in the context of low-copy nuclear gene markers, will continue to be critical for the robustness of plant phylogenetic reconstruction.

Allopolyploidization can be reconstructed by cloning and analyzing two homoeologs of a nuclear gene that are derived from both parents. Yet, the sequence polymorphism could also represent ancestral polymorphic alleles or gene duplication. Here is a simplified hypothetical example to elucidate this problem. There are four ingroup species, A, B, C, and D, and an outgroup species, O. Two distinct types of sequences C_1 and C_1', are cloned from a nuclear gene, 1, of species C, while only one type of sequence, A_1, B_1, D_1, and O_1, is found in each of the remaining species, A, B, D, and O, respectively. Phylogenetic analysis indicates that C_1 is a sister group of A_1, and C_1' forms a sister group of D_1 (Figure 2a).

Because species C has two distinct types of sequences that are closely related to those of species A and D, one possibility is that C is a hybrid between A and D (Figure 2b). Alternatively, the sequence polymorphism in species C represents either duplicated loci or ancestral polymorphic alleles. In this case, paralogy or lineage sorting has to be invoked. Assuming that the true phylogeny of the ingroup is A(B(C,D)), the gene tree is illustrated in Figure 2c. The gene duplication or allelic polymorphism giving rise to C_1 and C_1' must have occurred in the common ancestor of all ingroup species and was followed by three independent deletions of one gene copy or extinction of an allele from species, A, B, and D.

Obviously, a nuclear gene tree (e.g., Figure 2a) alone cannot distinguish between the alter-

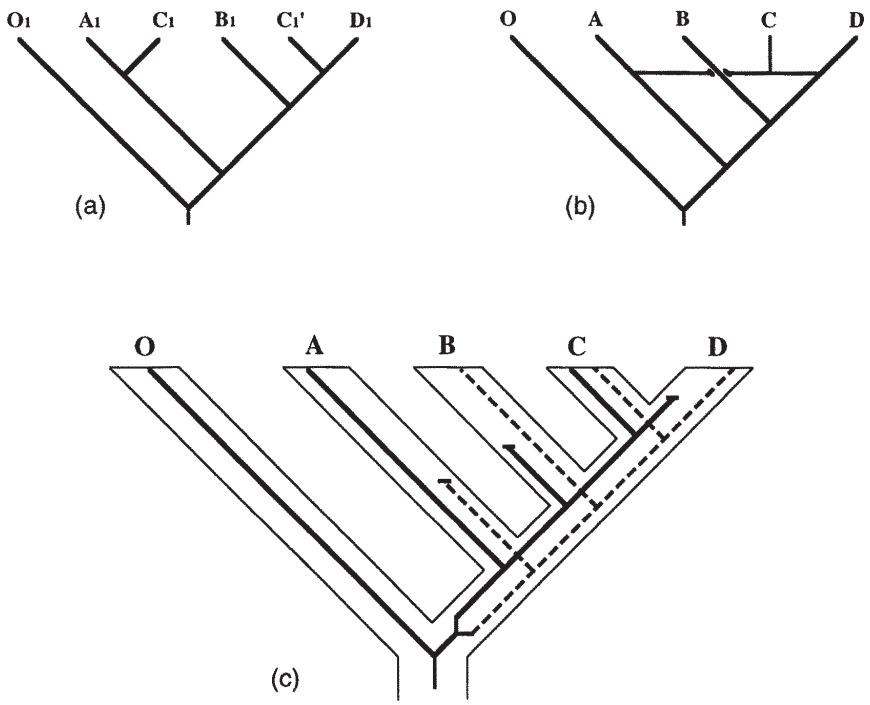

FIGURE 2. Gene trees and species tree of four ingroup taxa, A, B, C, and D, and an outgroup taxon O. (a) Tree of gene 1. A_1, B_1, D_1, and O_1 are sequences of gene 1 of species A, B, D, and O, respectively; C_1 and C_1' are sequences of gene 1 cloned from species C. (b) Species tree inferred from the gene tree based on the hypothesis that species C is an allopolyploid, and C_1 and C_1' represent homoeologous loci. (c) Species tree outlined by thin solid lines and phylogeny of gene 1 illustrated by thick solid and broken lines contained in the species tree. In this hypothesis, C_1 and C_1' represent either ancestral polymorphic alleles or duplicated copies of the gene. Subsequent to occurrence of polymorphic alleles or gene duplication, one allele or gene copy was extinct from species A, B, and D.

native hypotheses, that is, hybridization vs. gene duplication or ancestral polymorphism. A strong hypothesis of hybridization should be derived from the establishment of a correlation among multiple gene trees. For example, if two distinct sequences are cloned at additional unlinked nuclear loci of species C and have congruent positions between the gene trees, the hypothesis of allopolyploidization is supported further. It requires a single hypothesis that species C is an allotetraploid to explain the genome-wide sequence polymorphism (Gaut and Doebley, 1997; Ku et al., 2000).

It is, however, not impossible for a species to be more likely than its relatives in maintaining polymorphic ancestral alleles if, for example, it has a much larger effective population size than the related species (Pamilo and Nei, 1988; Kreitman, 1991; Hudson, 1992). Nevertheless, if multiple alleles are cloned from a single individual of the species at multiple unlinked loci, the probability for these alleles to be ancestral polymorphisms is low. Because lineage sorting is a random process, chances for diverged ancestral alleles to be preserved at

multiple unlinked nuclear loci of the same individual are rather small.

In a relatively ancient allotetraploid, however, one of the homoeologs may become a pseudogene and eventually be deleted from the allotetraploid genome as a result of the reduction of genetic redundancy (Ge et al., 1999). Random fixation of one of the homoeologs across nuclear loci leads to different sister relationships between the hybrid and the parents, that is, the allotetraploid will, by chance, form a sister group with either the paternal or maternal parent on a given nuclear gene tree. This subsequently creates incongruent positions of the hybrid among low-copy nuclear gene trees, and potentially between cpDNA and nrDNA trees as well. Likewise, a homoploid hybrid may also randomly fix nuclear genes from one of the parents through segregation or genome reorganization (Rieseberg, 1997; Ferguson and Sang, 2001). Under these circumstances, hybrid speciation has to be inferred through the comparison of topological incongruence between gene trees. To do so we should first determine that hybridization, rather than paralogy or lineage sorting, is the cause of the topological incongruence.

This is, however, a very difficult theoretical problem with no satisfactory solutions found to date. Here I describe briefly a recent theoretical study (Sang and Zhong, 2000) to exemplify some issues surrounding testing the hybridization hypothesis based on incongruent gene trees. In a simple case involving three ingroup species, species B has incongruent positions between the two gene trees (Figure 3a, b). If this incongruence is caused by hybridization, the species tree, with B being the hybrid between A and C, is illustrated in Figure 3c. The hybrid species B fixed the sequence of gene 1 of species C, and gene 2 of species A. Alternatively, paralogy or lineage sorting of one gene may be the cause of the incongruence. Assume that gene tree 1 represents the species tree and gene 2 has undergone duplication and deletion or lineage sorting (Figure 3d). Two gene copies or alleles of gene 2 arose in the common ancestor of species A, B, and C. Subsequently, one of them is maintained only in A and B, and the other is maintained only in C.

Under the hybridization hypothesis, $T_i = T_1$, and $T_k = T_1$, then, $T_i = T_k$. Under paralogy or lineage sorting hypothesis, $T_i = T_1$ and $T_k = T_p$. Because $T_p > T_1$, then $T_k > T_i$. Therefore, we can test hypotheses of hybridization vs. paralogy or lineage sorting by testing $T_k = T_i$, or $T_k - T_i = 0$. Defining $\Delta(x, y) = T_x - T_y$, we can test:

$$\Delta(k, i) = \begin{cases} 0, & \text{if hybridization;} \\ (a > 0), & \text{if paralogy or lineage sorting for gene 2.} \end{cases}$$

Under the molecular clock hypothesis, we have (d represent sequence divergence):

$$\Delta(k,i) = \left[\frac{2d(A_2C_2)}{d(O_2A_2) + d(O_2C_2)} - \frac{2d(A_1C_1)}{d(O_1A_1) + d(O_1C_1)} \right] T_0,$$

Let $\Delta(k, i) = \Delta_0 T_0$, we can test $\Delta_0 = 0$.

This model, however, has a number of limitations. It relies on the existence of a relatively accurate molecular clock. Designing a powerful statistical test for the theoretical model is not straightforward. Additional effort is needed to continue to explore effective approaches to test hybridization vs. paralogy or lineage sorting hypotheses.

IV. PRACTICAL ISSUES

A. Selection of Nuclear Genes

An ideal gene should maximize phylogenetic information but minimize homoplasy at the taxonomic level studied. If the phylo-

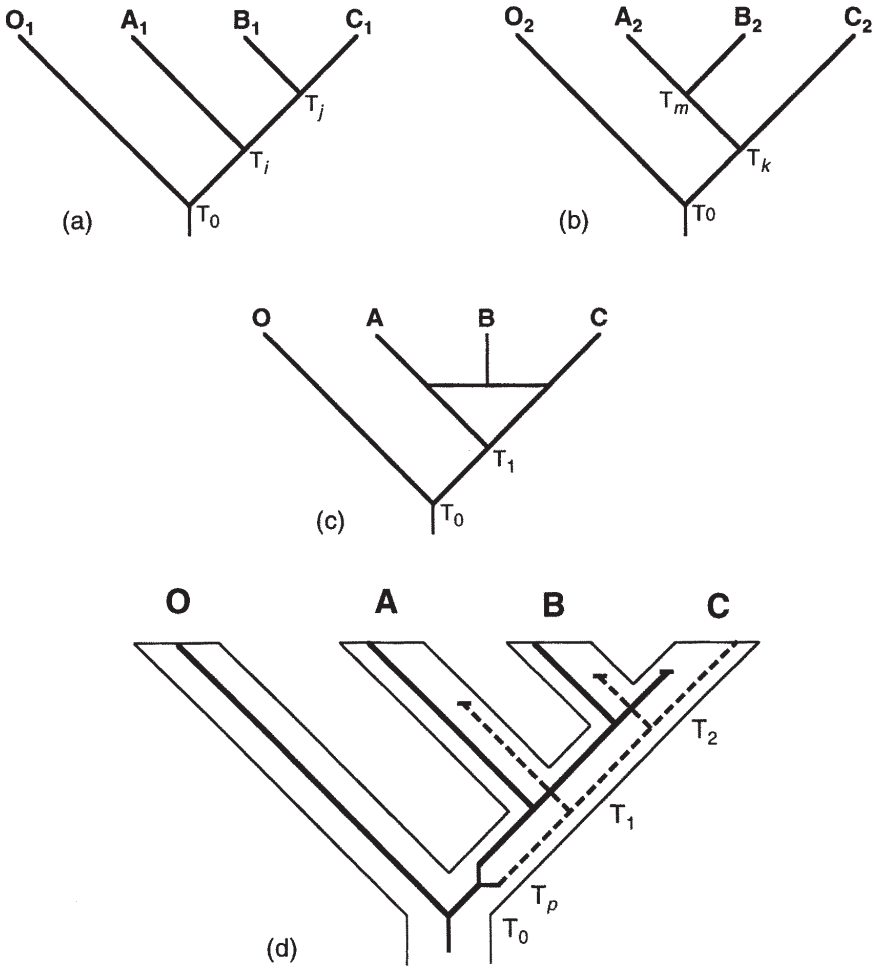

FIGURE 3. Gene trees and species trees of three ingroup taxa A, B, and C, and an outgroup taxon O. A_1, B_1, C_1, and O_1 are sequences of gene I, and A_2, B_2, C_2, and O_2 are sequences of gene 2 from species A, B, C, and O, respectively. T_i, T_j, T_k, and T_m represent divergence times between genes A_1 and (B_1, C_1), B_1 and C_1, C_2 and (A_2, B_2), and A_2 and B_2, respectively. T_0, T_1, and T_2 represent times of speciation. T_p represents time of gene duplication or occurrence of polymorphic alleles. (a) Tree of gene 1. (b) Tree of gene 2. (c) Species tree inferred based on the hypothesis that B is a hybrid species. (d) Species tree (outlined by thin solid lines) inferred on the basis that gene 2 has undergone paralogy or lineage sorting. Phylogeny of gene 2 is illustrated by contained thick lines with solid and broken lines representing two duplicated loci or ancestral alleles.

genetic reconstruction is at or above the family level, it is usually impossible to align intron sequences of a nuclear gene. Thus, gene or gene regions that are evolutionarily conserved and lack introns are preferred for the maximal utilization of sequence data.

At the intergeneric level, introns may or may not be aligned depending on genes and sizes of introns. For example, between the genera of Pinaceae, the majority of intron sequences of the 4CL gene was readily aligned by eyes, whereas few intron sequences of the CAD gene could be aligned without ambiguity (Wang et al., 2000; Tank et al., unpublished). Large introns usually accommodate more mutations of insertion and deletion and thus should be avoided in phylogenetic studies at the intergeneric level. The *Adc* gene, which does not contain introns, provided good resolution for the intergeneric relationships in the Brassicaceae (Galloway et al., 1998). The *Adh* and *waxy* genes are also suitable markers for phylogenetic studies at the intergeneric level as long as gene regions containing large introns are not chosen for sequencing (e.g., Mason-Gamer et al., 1998; Gaut et al., 1999). At or below the interspecific level, however, gene regions that contain a high proportion of intron sequences are preferred. In particular, large introns tend to provide better resolution for relationships among recently diverged species or within species (Tank and Sang, 2001).

In addition to the consideration of sequence divergence, it is critical to choose a gene that is less likely to yield paralogous relationships for the studied group. This depends largely on the dynamics of gene duplication and deletion and its relation to speciation in the studied group. Presumably, a gene family with a large number of gene members is more likely to have undergone many cycles of duplication and deletion, which makes it difficult to sort out orthology from paralogy (Clegg et al., 1997). Thus, genes with a smaller copy number, ideally a single copy, are usually preferred.

However, it is not always possible to predict the duplication/deletion dynamics of a candidate gene in the studied group simply based on copy numbers. The copy number of a low-copy nuclear gene varies from group to group and may not accurately reflect the history of gene duplication and deletion. For example, a diploid *Oryza* species has two *Adh* genes, whereas a diploid *Gossypium* species has up to seven *Adh* loci (Ge et al., 1999; Small and Wendel, 2000a). The *Vicilin* gene is present as a single-copy in *Theobroma* and *Herriania* of Sterculiaceae, but present as a small gene family consisting of five members in the genus *Lens* of Fabaceae (Saenz de Miera and Perez de la Vesa, 1998). *G3pdh* may be a single-copy gene in some dicots (Olsen and Schaal, 1999), but has several copies in maize (Martinez et al., 1989).

The *waxy* gene provided robust phylogenetic hypotheses for Poaceae and Rosaceae, but yielded extensive paralogous relationships in *Paeonia* (Sang, unpublished). Even though *Gpat* is likely a single-copy gene in plant species from several eudicot families and in the majority of *Paeonia* species, it has undergone duplication and deletion during the diversification of *Paeonia* which led to paralogous relationships between sections of the genus (Tank and Sang, 2001).

Because the dynamics of gene duplication and deletion tends to vary among plant groups, it is difficult to develop universal phylogenetic markers from low-copy nuclear genes. Therefore, preliminary studies are required to identify the most suitable markers for specific plant groups of interest (see later discussion).

B. Designing PCR Primers

It is unlikely that there will be universal PCR primers for the majority of low-copy nuclear genes used in plant phylogenetic stud-

ies. At different taxonomic levels, one may choose to amplify different gene regions that have suitable rates of sequence divergence. The difficulty in designing universal PCR primers also stems from practical problems such as the lack of gene sequences or conserved gene regions across a wide taxonomic range. For example, nuclear genes of basal angiosperms are characterized less extensively compared with eudicots and monocots. PCR primers that can amplify a nuclear gene from a basal angiosperm species may have to be designed in exon regions that are conserved between angiosperm and gymnosperms. Because protein-coding nuclear genes have relatively high rates of synonymous substitution (Gaut, 1998), PCR primers designed based on sequences of such distantly related taxa usually contain a large number of degenerate sites, consequently reducing PCR specificity and efficiency.

Even in the case of studying low-copy nuclear genes of eudicots or monocots, there is a tradeoff between designing universal primers and effective primers for the group of interest. Primers that are designed to amplify both eudicots and monocots may inevitably contain many more degenerative sites than ones that amplify each group separately. Although universal primers have greater potential to be useful in a wider taxonomic range, they tend to cause lower specificity and efficiency of PCR. Often, annealing temperatures have to be decreased in order to make the highly degenerate primers work, which can cause high levels of nonspecific amplification and even PCR failure.

For phylogenetic studies at high taxonomic levels, it is necessary to design primers that work across a wide taxonomic range. To amplify low-copy nuclear genes within a genus, however, more specific primers that are conserved only among the gene sequences of the close relatives may be preferred to ensure a highly effective PCR. For example, the PCR primers used to amplify *Adh* genes for phylogenetic studies within the genus *Oryza* were designed based on aligned sequences between *Adh1* and *Adh2* genes of *Oryza sativa* and *Zea mays*. Because the *Adh1* and *Adh2* genes were duplicated before the diversification of the grass family, these primers should be conserved enough to amplify all *Adh* genes that were duplicated subsequently.

Primers for amplifying *Adh* genes from *Paeonia*, however, were designed based on *Adh* sequences that are conserved across four eudicot families, Brassicaceae, Fabaceae, Solanaceae, and Rosaceae, because there had been no published *Adh* sequences from *Paeonia* or close relatives. The primers, with only two degenerate sites, yielded strong specific amplification of *Adh* genes from Paeoniaceae as well as Lamiaceae (Sang et al., 1997b; Williams, personal communication). To design primers to amplify *Adh* genes from both eudicots and monocots, however, a large number of degenerate sites have to be included.

The lack of universal gene markers and universal primers makes the procedure of phylogenetic reconstruction using low-copy nuclear genes somewhat different from those using chloroplast or nuclear ribosomal genes. Certain low-copy nuclear genes, such as *Adh*, *waxy*, and *ncpGS*, may become popular gene markers even though they are not universal markers. As these genes are tested in more and more groups, multiple sets of primers will become available and eventually cover the majority of plant groups. At present, plant systematists have to put some effort into primer design if appropriate primers are unavailable for the group of interest.

C. Cloning

Because distinct loci and alleles may be amplified by PCR, it is necessary to clone

the PCR products before sequencing. Adding the extra step of cloning to the regular procedures of molecular systematic studies raises the question whether it is feasible to use low-copy nuclear genes as routine phylogenetic markers in plants. I address this question and discuss effective ways to conduct cloning.

The decision on the number of clones to pick is a critical point that determines the amount of lab work. Picking and culturing clones, as well as purifying plasmid DNA, require many working hours in the lab. The number of clones to screen depends on the number of loci amplified by PCR. Due to PCR selection (Wagner et al., 1994), different loci may not be represented equally in the PCR products. For example, if two loci are amplified and one is amplified twice as strongly as the other, ideally we expect to identify both loci by randomly screening three PCR clones. However, due to sampling errors, many more clones are usually screened before the less frequent one is identified. It is, however, impossible to predict the minimal number of clones to be screened in order to identify all loci of a gene. Empirically, both *Adh1* and *Adh2* genes were identified by screening seven clones for the majority of diploid species of *Paeonia* and *Oryza*.

Occasionally, one of the *Adh* genes was not found from a diploid peony species after screening more than twenty PCR clones. In such cases, paralogue-specific primers should be designed. The *Adh2* gene was successfully amplified from *P. anomala* and *P. tenuifolia* with the *Adh2*-specific primers, while they were not identified after screening more than 20 clones from the PCR products of the general *Adh* primers (Sang et al., 1997b). Furthermore, uses of paralogue-specific primers are particularly helpful when each paralogous gene exists as a single copy in the studied taxa. In this case, cloning becomes unnecessary unless a locus is clearly heterozygous based on the results of direct sequencing of PCR products.

Paralogue-specific primers are also useful when trying to isolate homoeologous loci from allotetraploid species. If the general PCR primers were used to amplify the *Adh* genes from an allotetraploid peony, a large number of clones were screened to identify all four homoeologous loci of the *Adh1* and *Adh2* genes. This may have resulted from a combination of PCR selection and sampling errors in picking clones. When gene-specific primers were used, two runs of PCR and cloning had to be conducted separately for the *Adh1* and *Adh2* genes. Nevertheless, we found that this actually reduced the amount of lab work to identify all homoeologous loci of both *Adh* genes from allotetraploid species because a much smaller number of clones were screened in comparison with using general PCR primers.

Gene-specific primers should meet some important requirements. They should be able to amplify all orthologs of the gene from all studied taxa but without amplifying its paralogs. Therefore, the primers should be designed based on aligned ortholog sequences from the most diverged taxa of the studied group. For example, the *Adh2* specific primer was located in the region where *Adh2* sequences of the most diverged *Paeonia* species were identical, but differed from the *Adh1* gene by both indels and the nucleotide at the 3′ end of the primer (Sang et al., 1997b).

Isolation of clones with incorrect inserts can also lead to extra lab work. Incorrect inserts are usually short fragments resulting from nonspecific PCR amplification. If a PCR is highly efficient and yields little nonspecific amplification, we found that usually more than 80% of the clones contained correct inserts. However, an average PCR could result in more than a half of clones with

incorrect inserts. In this case, a step of gel purification of PCR products before cloning can be very helpful. This includes running the PCR products through a 1% agarose gel, excising the band that contains the correct fragment, and recovering the DNA fragment with a GeneClean kit (Bio101). Using the TOPO TA cloning kit (Invitrogene), we found that this additional step did not reduce the transformation efficiency of cloning, but could significantly increase the percentage of clones with the correct insert.

For the initial study of a nuclear gene with two gene members, such as the *Adh* gene, we usually isolate at least 10 clones for each PCR. Assuming that all 10 clones contain the correct insert, they can be screened by restriction digestion with frequently cutting enzymes. We found that using three four-cutters could usually distinguish clones that differ by 1% sequence divergence. This level of sequence divergence should normally allow identification of different loci. To screen sequence variation at finer scales, clones can be sequenced with one of the PCR primers.

D. Preliminary Study

Owing to the lack of universal low-copy nuclear gene markers, it is necessary to conduct a preliminary study to determine which genes are appropriate for the studied group. Similar to what has been suggested for nuclear ribosomal and chloroplast genes, this will help evaluate whether the level of sequence variation of a particular gene is suitable for the phylogenetic reconstruction (Soltis and Soltis, 1998). A more important function of the preliminary study of low-copy nuclear genes is to estimate the copy number and the dynamics of gene duplication and deletion of the candidate genes.

Gene duplications that can potentially result in paralogy problems must have occurred before diversification of at least a portion of the ingroup taxa. If sequences cloned from each sampled taxon form a monophyletic group, it is unlikely that gene duplication has occurred prior to or during the diversification of the ingroup. If the gene has been duplicated, distinct sequences will be cloned from all or a portion of the ingroup taxa and should fall into distinct clades that correspond to the duplicated loci. If the preliminary analysis indicates that a gene has not been duplicated extensively in the studied group, and the resulting phylogeny of the sampled taxa is largely congruent with the previous phylogenetic hypotheses, the gene is most likely useful in the studied group.

Additionally, copy number of a low-copy nuclear gene can be estimated by Southern blotting (Small and Wendel, 2000a). The estimate from Southern blotting may help detect possible omission of gene members from the sampled PCR clones. However, we must keep in mind that results of Southern blotting are sensitive to the specificity of hybridization probes and the stringency of hybridization and wash conditions. Low specificity of probes and/or low stringency of the Southern blotting may lead to an overestimate of copy number.

Although it is always advisable to conduct Southern blotting, it is not always practical to do this experiment in systematic studies. Southern blotting of low-copy nuclear genes requires a large quantity of well-preserved DNA. In most cases, it may be feasible to conduct Southern blotting only for a few sampled ingroup taxa given the limited availability of DNA as well as laboratory resources in a plant systematic study.

If the preliminary study suggests that the gene is present as single-copy, it may not be necessary to conduct cloning for the

remaining ingroup taxa unless heterozygosity is detected from the results of direct sequencing. Even when a gene has been duplicated, each gene copy may be treated as a single-copy gene as long as the gene duplication occurred prior to the diversification of the entire ingroup and can be amplified with specific PCR primers. A justified omission of the cloning step is always preferred in a systematic study.

On the other hand, if the preliminary study suggests that the gene has undergone such extensive duplication and deletion that it is impossible to sort out orthology from the paralogous relationships, the gene should not be chosen as a phylogenetic marker. It is tricky, however, to identify paralogy when a gene is determined to be single-copy but historically has undergone duplication and deletion in the studied group (Doyle and Davis, 1998). When a gene was duplicated and deleted during the diversification of the ingroup but left only one copy in each taxon, estimation of gene copy number by PCR-cloning and/or Southern blotting will not help predict the potential paralogy problem. Therefore, it is necessary to compare the preliminary gene tree with previous phylogenetic hypotheses. If significantly incongruent relationships are found between the gene trees and the previous phylogenetic hypotheses, caution must be exercised to decide whether this low-copy nuclear gene should be used in the phylogenetic study.

V. CONCLUSIONS AND PROSPECTS

As low-copy nuclear gene markers become increasingly accessible, we inevitably face a series of challenging questions: (1) When are low-copy nuclear genes needed in addition to cpDNA and nrDNA? (2) How many low-copy nuclear genes are necessary for an accurate phylogenetic reconstruction? (3) How can multiple gene trees be integrated to give the best possible estimate of the species tree? Obviously, satisfactory answers to these questions must wait until more data become available. The following paragraphs summarize the relevant discussion in the previous sections of this article, and attempt to stimulate thoughts on the future phylogenetic studies using low-copy nuclear genes.

Compared with cpDNA and nrDNA, phylogenetic analyses of low-copy nuclear genes require a larger amount of high-quality DNA and usually the additional step of cloning. If a phylogeny can be accurately reconstructed based on cpDNA and nrDNA sequences, one should not choose to sequence low-copy nuclear genes. Thus, the answer to the question of whether a low-copy nuclear gene phylogeny is needed depends on the judgement of whether cpDNA and nrDNA can provide sufficient and accurate phylogenetic information.

The chloroplast genome has served as a remarkable source of data for plant phylogenetic reconstruction. Phylogenetic utility of cpDNA, however, is largely limited by its slow rates of evolution and uniparental inheritance (Olmstead and Palmer, 1994). As a result, cpDNA alone is not sufficient for phylogenetic reconstruction at low taxonomic levels and/or for the plant groups with history of hybridization.

Nuclear ribosomal DNA is a widely used phylogenetic marker in plants as well as other organisms. However, concerted evolution of nrDNA makes it an unreliable nuclear marker for reconstructing relatively ancient allopolyploidization. Furthermore, the ITS region may not provide a sufficient amount of phylogenetic information to resolve close interspecific relationships. Although the ETS region of nrDNA tends to be more informative than ITS (Baldwin and Markos, 1998; Linder et al., 2000), it suf-

fers from the same limitation as ITS in the reconstruction of allopolyploidization. Moreover, the ETS phylogeny does not provide an independent assessment of the species tree because it is tightly linked to ITS.

At this point, it is clear that low-copy nuclear genes are most useful at the interspecific and intraspecific levels where cpDNA and/or nrDNA cannot provide adequate resolution. They are particularly effective for reconstruction of allopolyploidization. At any taxonomic level, if cpDNA and nrDNA phylogenies are poorly resolved, weakly supported, and/or incongruent with each other, utility of low-copy nuclear genes should be considered.

The number of low-copy nuclear genes to be studied depends on the specific needs for the additional gene phylogenies. If low-copy nuclear gene sequences are needed to simply improve resolution of the cpDNA and/or nrDNA phylogeny, as few as one gene may be enough. When the low-copy nuclear gene phylogeny is congruent with the cpDNA and nrDNA phylogenies, a combined analysis of the gene sequences should improve the resolution of the phylogenetic reconstruction.

When the low-copy nuclear gene tree is topologically incongruent with the cpDNA and/or nrDNA trees, additional nuclear genes are likely needed depending on the cause of the incongruence. Correct inference of a species tree from incongruent gene trees has raised a series of challenging theoretical questions that have attracted considerable attention (e.g., Doyle, 1992; Maddison, 1997; Page, 1998; Slowinski and Page, 1999). Here I do not attempt to review these theoretical studies, and instead bring up a couple of issues that are relevant to the low-copy nuclear gene phylogenies.

If paralogy of nuclear genes is solely responsible for the topological incongruence between gene trees, the correct species tree is likely to be inferred by comparing the multiple gene trees. Because gene duplication/deletion is very unlikely to be parallel between unlinked nuclear loci, the probability that the same paralogous relationship occurs between independent loci is small. Therefore, a majority-rule consensus tree of a few low-copy nuclear gene trees may well represent the species phylogeny. It is, however, an open theoretical question of how many nuclear gene trees are needed to arrive at a correct reconstruction of the species tree.

Inference of the species tree from gene trees becomes more complicated when lineage sorting is involved. When branches are sufficiently long (through many generations) or narrow (with small effective population size) (Maddison, 1997), the likelihood that the majority-rule consensus of a relatively large number of nuclear gene trees reflects the species tree is high (Pamilo and Nei, 1988). However, it is unreliable to infer the species tree from a majority-rule consensus when certain branches are short and wide even though a large number of nuclear genes are examined (Pamilo and Nei, 1988).

As we approach the species boundary, our chance to recover the species tree will decrease due to deep coalescence. Nevertheless, a noncoalescent gene genealogy could provide insights into the processes of population differentiation and mechanisms of speciation. Therefore, we must recognize values of nuclear gene trees even though they do not show clear interspecific relationships with all alleles coalesced within species. These gene trees can be invaluable in addressing questions concerning evolutionary processes at the populational level, although they may not be taxonomically meaningful. We simply have to accumulate examples of phylogenetic studies at this level before we can fully assess the phylogenetic value of low-copy nuclear genes.

If topological incongruence is caused by hybridization, the species tree cannot be

inferred correctly from consensus or combined analysis. Comparison of two or more nuclear gene trees may be necessary when one attempts to reconstruct homoploid hybridization (Ferguson and Sang, 2000) or ancient allopolyploidization (Ge et al., 1999; Wendel, 2000). Theoretical models and statistics that distinguish hybridization and paralogy or lineage sorting need to be developed further. Reconstruction of the true organismal phylogeny can be extremely challenging when multiple factors, including hybridization, paralogy, and lineage sorting, are responsible for the incongruence of gene trees. An accurate reconstruction of complex plant phylogenies requires both suitable gene markers and appropriate analytical approaches.

Beyond the immediate goal of reconstructing a robust species tree, phylogenies of low-copy nuclear genes will play an important role in addressing fundamental evolutionary questions. Because nuclear genes determine the vast range of phenotypes that are responsible for adaptation, a nuclear gene phylogeny should shed light on the evolution of morphological and physiological traits that are controlled by these genes. Phylogenetic analyses of homeotic and regulatory genes can potentially open a new avenue to the study of evolution of developmental mechanisms (Doyle, 1994; Frohlich and Meyerowitz, 1997; Bharathan et al., 1999; Purugganan and Suddith, 1999; Becker et al., 2000; Lawton-Rauh et al., 2000; Shu et al., 2000).

Ultimately, low-copy nuclear gene phylogenies will serve as a crucial element bridging our understanding of evolution of genotype and phenotype. For example, phylogenetic analyses of the regulatory region and protein-coding region of the teosinte branched1 locus (*tb1*), the major gene involved in maize evolution, yielded strikingly different phylogenies between the two portions of the gene (Wang et al., 1999). On the gene tree of the coding region, maize sequences fell into multiple clades that are mixed with those of the other two subspecies *Zea mays* ssp. *parviglumis* and *Zea mays* ssp. *mexicana*. On the tree generated from the regulatory region, however, all maize sequences together with a few sequences of ssp. *parviglumis* formed a strongly supported clade. These results suggested that ssp. *parviglumis* was the wild progenitor of maize, and the selection of desired crop morphology was centered on the regulatory region of the *tb1* gene during maize domestication.

As rapid progress is made toward functional genomics of *Arabidopsis* and rice, genes that control morphological, physiological, and ecological traits will be identified. Comparison of the genes between these distantly related model plants will allow us to target orthologous genes that are most likely responsible for phenotypic evolution of flowering plants. Phylogenetic and molecular evolutionary analyses of developmentally important genes will add a new dimension to systematic and evolutionary studies of plant diversity.

ACKNOWLEDGMENTS

I thank Diane Ferguson and David Tank for useful discussion, and Jeff Doyle, Dan Crawford, David Tank, Alan Prather, Rachel Williams, and Jessie Keith for valuable comments on the earlier versions of the manuscript. This work was supported by the National Science Foundation.

REFERENCES

Arnold, M. L. 1997. *Natural Hybridization and Evolution*, New York: Oxford University Press.

Bailey, C. D. and J. J. Doyle. 1999. Potential phylogenetic utility of the low-copy nuclear gene *pistillata* in dicotyledonous plants: Comparison to nrDNA ITS and *trnL* intron in *Sphaerocardamum* and other Brassicaceae. *Mol. Phylogenet. Evol.* **13:** 20–30.

Baldwin, B. G. and S. Markos. 1998. Phylogenetic utility of the external transcribed spacer (ETS) of 18S-26S rDNA: Congruence of ETS and ITS trees of *Calycadenia* (Compositae). *Mol. Phylogenet. Evol.* **10:** 449–463.

Baldwin, B. G., M. J. Sanderson, J. M. Porter, M. F. Wojciechowski, C. S. Campbell, and M. J. Donoghue. 1995. The ITS region of nuclear ribosomal DNA — a valuable source of evidence on angiosperm phylogeny. *Ann. Missouri Bot. Gard.* **82:** 247–277.

Barrier, M., B. G. Baldwin, R. H. Robichaux, and M. D. Purugganan. 1999. Interspecific hybrid ancestry of a plant adaptive radiation: Allopolyploidy of the Hawaiian silversword alliance (Asteraceae) inferred from floral homeotic gene duplication. *Mol. Biol. Evol.* **16:** 1105–1113.

Becker A, K. U. Winter, B. Meyer, H. Saedler, and G. Theissen. 2000. MADS-box gene diversity in seed plants 300 million years ago. *Mol. Biol. Evol.* **17:** 1425–1434.

Bharathan, G., B. J. Janssen, E. A. Kellogg, and N. Sinha. 1999. Phylogenetic relationships and evolution of the KNOTTED class of plant homeodomain proteins. *Mol. Biol. Evol.* **16:** 553–563.

Brown, J. R. and W. F. Doolittle. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. USA* **92:** 2441–2445.

Clegg, M. T., M. P. Cummings, and M. L. Durbin. 1997. The evolution of plant nuclear genes. *Proc. Natl. Acad. Sci. USA* **92:** 7791–7798.

Crawford, D. J. 1990. *Plant Molecular Systematics: Macromolecular Approaches*, New York: John Wiley & Sons.

Cronn, R. C., R. L. Small, and J. F. Wendel. 1999. Duplicated genes evolve independently after polyploid formation in cotton. *Proc. Natl. Acad. Sci. USA* **96:** 14406–14411.

Denton, A. L., B. L. McConaughy, and B. D. Hall. 1998. Usefulness of RNA polymerase II coding sequences for estimation of green plant phylogeny. *Mol. Biol. Evol.* **15:** 1082–1085.

Doyle, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst. Bot.* **17:** 144–163.

Doyle, J. J. 1994. Evolution of a plant homeotic multigene family: toward connecting molecular systematics and molecular developmental genetics. *Syst. Biol.* **43**: 307–328.

Doyle, J. J. 1997. Trees within trees: genes and species, molecules and morphology. *Syst. Biol.* **46:** 537–553.

Doyle, J. J. and J. I. Davis. 1998. Homology in molecular phylogenetics: a parsimony perspective. Pages 101–131 In: *Molecular Systematics of Plants. II. DNA Sequencing* (D, Soltis, P. Soltis, and J. Doyle, Eds.), Boston: Kluwer Academic.

Doyle, J. J. and J. L. Doyle. 1999. Nuclear protein-coding genes in phylogeny reconstruction and homology assessment: some examples from Leguminosae. In: *Molecular Systematics and Plant Evolution* (P. Hollingsworth, R. Bateman, and R. Gornall, Eds.), London: Taylor and Francis.

Doyle, J. J., J. L. Doyle, and A. H. D. Brown. 1999. Incongruence in the diploid B-genome species complex of *Glycine* (Leguminosae) revisited: Histone H3–D alleles versus chloroplast haplotypes. *Mol. Biol. Evol.* **16:** 354–362.

Doyle, J. J., V. Kanazin, and R. C. Shoemaker. 1996. Phylogenetic utility of histone H3 intron sequences in the perennial relatives of soybean (Glycine: Leguminosae). *Mol. Phylogenet. Evol.* **6:** 438–447.

Doyle, J. J., J. L. Doyle, A. H. D. Brown, and B. E. Pfeil. 2000. Confirmation of shared and divergent genomes in the *Glycine tabacina* polyploid complex (Leguminosae) using histone H3–D sequences. *Syst. Bot.* **25:** 437–448.

Durbin, M. L., B. McCaig, and M. T. Clegg. 2000. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. *Plant Mol. Biol.* **42:** 79–92.

Emshwiller, E. and J. J. Doyle. 1999. Chloroplast-expressed glutamine synthetase (*ncpGS*): Potential utility for phylogenetic studies with an example from *Oxalis* (Oxalidaceae). *Mol. Phylogenet. Evol.* **12:** 310–319.

Evans, R. C., L. A. Alice, C. S. Campbell, E. A. Kellogg, and T. A. Dickinson. 2000. The granule-bound starch synthase (GBSSI) gene in the Rosaceae: Multiple loci and phylogenetic utility. *Mol. Phylogenet. Evol.* **17:** 388–400.

Eyre-Walker, A. R., L. Gaut, H. Hilton, D. Feldman, and B. S. Gaut. 1998. Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl. Acad. Sci. USA* **95:** 4441–4446.

Ferguson, D. and T. Sang. 2001. Speciation through homoploid hybridization between allotetraploids in peonies (*Paeonia*). *Proc. Natl. Acad. Sci. USA* **98:** 3915–3919.

Ford, V. S. and L. D. Gottlieb. 1999. Molecular characterization of *PgiC* in a tetraploid plant and its diploid relatives. *Evolution* **53:** 1060–1067.

Frohlich M.W. and E. M. Meyerowitz. 1997. The search for flower homeotic gene homologs in basal angiosperms and gnetales: A potential new source of data on the evolutionary origin of flowers. *Int. J. Plant Sci.* **158:** S131–S142.

Funk, V. A. 1985. Phylogenetic pattern and hybridization. *Ann. Missouri Bot. Gard.* **72:** 681–715.

Galloway, G. L., R. L. Malmberg, and R. A. Price. 1998. Phylogenetic utility of the nuclear gene arginine decarboxylase: an example from Brassicaceae. *Mol. Biol. Evol.* **15:** 1312–1320.

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

Gaut, B. S. and J. F. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. USA* **94:** 6809–6814.

Gaut, B. S., A. S. Peek, B. R. Morton, and M. T. Clegg. 1999. Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaceae). *Mol. Biol. Evol.* **16:** 1086–1097.

Ge, S., T. Sang, B.-R. Lu, and D.-Y. Hong. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc. Natl. Acad. Sci. USA* **96:** 14400–14405.

Gottlieb, L. D. and V. S. Ford. 1996. Phylogenetic relationships among the sections of *Clarkia* (Onagraceae) inferred from the nucleotide sequences of *PgiC*. *Syst. Bot.* **21:** 45–62.

Graham, S. W. and R. G. Olmstead. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am. J. Bot.* **87:** 1712–1730.

Grant, V. 1981. *Plant Speciation,* 2nd ed. New York: Columbia University Press.

Hanson, M. A., B. S. Gaut, A. O. Stec, S. I. Fuerstenberg, M. M. Goodman, E. H. Coe, and J. F. Doebley. 1996. Evolution of anthocyanin biosynthesis in maize kernels: The role of regulatory and enzymatic loci. *Genetics* **143:** 1395–1407.

Hillis, D. M. 1995. Approaches for accessing phylogenetic accuracy. *Syst. Biol.* **44:** 3–16.

Hilton H. and B. S. Gaut. 1999. Speciation and domestication in maize and its wild relatives: Evidence from the globulin-1 gene. *Genetics* **150:** 863–872.

Hudson, R. R. 1992. Gene trees, species trees, and the segregation of ancestral alleles. *Genetics* **131:** 509–512.

Iwabe, N., K. Kumam, M. Hasegawa, S. Osawa, and T. Miyata. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* **86:** 9355–9359.

Koch, M. A., B. Haubold, and T. Mitchell-Olds. 2001. Molecular systematics of the Brassicaceae: Evidence from coding plastid *matK* and nuclear *Chs* sequences. *Am. J. Bot.* **88:** 534–544.

Kolukisaoglu, H. U., S. Marx, C. Wiegmann, S. Hanelt, and H. A. W. Schneider-Poetsch. 1995. Divergence of the phytochrome gene family predates angiosperm evolution and suggests that *Selaginella* and *Equisetum* arose prior to *Psilotum*. *J. Mol. Evol.* **41:** 329–337.

Kreitman, M. 1991. Detecting selection at the level of DNA. Page 204–221 In: *Evolution at the Molecular Level* (R. K. Selander, A. G. Clark, and T. S. Whittam, Eds.), Sunderland, Mass: Sinauer Associates.

Ku, H.-M., T. Vision, J. Liu, and S. D. Tanksley. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA* **97:** 9121–9126.

Lawson, F. S., R. L. Charlebois, and J.-A. R. Dillon. 1996. Phylogenetic analysis of Carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. *Mol. Biol. Evol.* **13:** 970–977.

Lawton-Rauh, A. L, E. R. Alvarez-Buylla, and M. D. Purugganan. 2000. Molecular evolution of flower development. *Trends Ecol. Evol.* **15:** 144–149.

Li, W-H. 1998. *Molecular Evolution*, Sunderland, Mass: Sinauer Associates.

Linder, C. R., L. R. Goertzen, B. V. Heuvel, J. Francisco-Ortega, and R. K. Jansen. 2000. The complete external transcribed spacer of 18S-26S rDNA: amplification and phylogenetic utility at low taxonomic levels in Asteraceae and closely allied families. *Mol. Phylogenet. Evol.* **14:** 285–303.

Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.* **46:** 523–536.

Martinez, P., W. Martin, and R. Cerff. 1989. Structure evolution and anaerobic regulation of a nuclear gene encoding cytosolic glyceraldehyde-3–phosphate dehydrogenase from maize. *J. Mol. Biol.* **208:** 551–565.

Mason-Gamer, R. J., C. F. Weil, and E. A. Kellogg. 1998. Granule-bound starch synthase: Structure, function, and phylogenetic utility. *Mol. Biol. Evol.* **15:** 1658–1673.

Mathews, S. and M. J. Donoghue. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286:** 947–950.

Mathews, S. and R. A. Sharrock. 1996. The phytochrome gene family in grasses (Poaceae): a phylogeny and evidence that grasses have a subset of the loci found in dicot angiosperms. *Mol. Biol. Evol.* **13:** 1141–1150.

Mathews S, R. C. Tsai, and E. A. Kellogg. 2000. Phylogenetic structure in the grass family (Poaceae): evidence from the nuclear gene phytochrome B. *Am. J. Bot.* **87:** 96–107.

McDade, L. A. 1995. Hybridization and phylogenetics. In: *Experimental and*

Molecular Approaches to Plant Biosystematics (P.C. Hoch and A. G. Stephenson, Eds.), Monograph of Systematic Botany, Missouri Botanical Garden.

Moore, W. S. 1995. Inference of phylogenies from mtDNA variation: mitochondiral-gene tree versus nuclear-gene trees. *Evolution* **49:** 718–726.

Nickrent, D. L., C. L. Parkinson, J. D. Palmer, and R. J. Duff. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* **17:** 1885–1895.

Olmstead, R. G. and J. D. Palmer. 1994. Chloroplast DNA systematics: a review of methods and data analysis. *Am. J. Bot.* **81:** 1205–1224.

Olsen, K. M. and B. A. Schaal. 1999. Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. *Proc. Natl. Acad. Sci. USA* **96:** 5586–5591.

Page, R. D. M. 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics* **14:** 819–820.

Pamilo, P. and M. Nei. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5:** 568–583.

Pryer, K. M., H. Schneider, A. R. Smith, R. Cranfill, P. G. Wolf, J. S. Hunt, and S. D. Sipes. 2001. Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature* **409:** 618–622.

Purugganan, M. D., S. D. Rounsley, R. L. Schmidt, and M. F. Yanofsky. 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulation gene family. *Genetics* **140:** 345–356.

Purugganan, M. D. and J. I. Suddith. 1999. Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. *Genetics* **151:** 839–848.

Qiu, Y. L., J. H. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. D. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402:** 404–407.

Raubeson, L. A. and R. K. Jansen. 1992. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science* **255:** 1697–1699.

Rieseberg, L.H. 1997. Hybrid origins of plant species. *Annu. Rev. Ecol. Syst.* **28:** 359–389.

Rieseberg, L. H. and J. D. Morefield. 1995. Character expression, phylogenetic reconstruction, and the detection of reticulate evolution. In: *Experimental and Molecular Approaches to Plant Biosystematics* (P. C. Hoch and A. G. Stephenson, Eds.), Monograph of Systematic Botany, Missouri Botanical Garden.

Saenz de Miera, L. E. and M. Perez de la Vesa. 1998. A comparative study of vicilin genes in *Lens*: negative evidence of concerted evolution. *Mol. Biol. Evol.* **15:** 303–311.

Sang, T., D. J. Crawford, and T. F. Stuessy. 1997a. Chloroplast phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *Am. J. Bot.* **84:** 1120–1136.

Sang, T., M. J. Donoghue, and D. Zhang. 1997b. Evolution of alcohol dehydrogenase genes in peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. *Mol. Biol. Evol.* **14:** 994–1007.

Sang, T. and D. Zhang. 1999. Reconstructing hybrid speciation using sequences of low-copy nuclear genes: hybrid origins of five *Paeonia* species based on *Adh* gene phylogenies. *Syst. Bot.* **24:** 148–163.

Sang, T., and Y. Zhong. 2000. Testing hybridization hypotheses based on incongruent gene trees. *Syst. Biol.* **49:** 422–434.

Schaal, B. A. and K. M. Olsen. 2000. Gene genealogies and population variation in plants. *Proc. Natl. Acad. Sci. USA* **97:** 7024–7029.

Shu, G. P., W. Amaral, L. C. Hileman, and D. A. Baum. 2000. LEAFY and the evolution of rosette flowering in violet cress (*Jonopsidium acaule*, Brassicaceae). *Am. J. Bot.* **87:** 634–641.

Simmons, M. P., C. C. Clevinger, V. Savolainen, R. H. Archer, S. Mathews, and J. J. Doyle. 2001 Phylogeny of the Celastraceae inferred from phytochrome B gene sequence and morphology. *Am. J. Bot.* **88:** 313–325.

Slowinski, J. B. and R. D. M. Page. 1999. How should species phylogenies be inferred from sequence data? *Syst. Biol.* **48:** 814–825.

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 phylogeny reconstruction in a recently diverged plant group. *Am. J. Bot.* **85:** 1301–1315.

Small, R. L., and J. F. Wendel. 2000a. Copy number lability and evolutionary dynamics of the *Adh* gene family in diploid and tetraploid cotton (*Gossypium*). *Genetics* **155:** 1913–1926.

Small, R. L. and J. F. Wendel. 2000b. Phylogeny, duplication, and intraspecific variation of *Adh* sequences in new world diploid cottons (*Gossypium* L., Malvaceae). *Mol. Phylogenet. Evol.* **16:** 73–84.

Soltis, D. E. and P. S. Soltis. 1998. Choosing an approach and an appropriate gene for phylogenetic analysis. In: *Molecular Systematics of Plants. II. DNA Sequencing* (D, Soltis, P. Soltis, and J. Doyle, Eds.), Boston: Kluwer Academy Publications.

Soltis, P. S., D. E. Soltis, and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402:** 402–404.

Soltis, D. E., P. S. Soltis, M. W. Chase, M. E. Mort, D. C. Albach, M. Zanis, V. Savolainen, W. H. Hahn, S. B. Hoot, M. F. Fay, M. Axtell, S. M. Swensen, L. M. Prince, W. J. Kress, K. C. Nixon, and J. S. Farris. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbc*L, and *atp*B sequences. *Bot. J. Linn. Soc.* **133:** 381–461.

Tank, D. and T. Sang. 2001. Phylogenetic utility of the glycerol-3–phosphate acyltransferase gene: Evolution and implications in *Paeonia* (Paeoniaceae). *Mol. Phylogenet. Evol.* **19:** 421–429.

Wagner A., N. Blackstone, P. Cartwright, M. Dick, B. Misof, P. Snow, G. P. Wagner, J. Bartels, M. Murtha, and J. Pendleton. 1994. Surveys of gene families using polymerase chain reaction: PCR selection and PCR drift. *Syst. Biol.* **43:** 250–261.

Wang, R. L., A. Stec, J. Hey, L. Lukens, and J. Doebley. 1999. The limits of selection during maize domestication. *Nature* **398:** 236–239.

Wang, X.-Q., D. C. Tank, and T. Sang. 2000. Phylogeny and divergence time in Pinaceae: evidence from three genomes. *Mol. Biol. Evol.* **17:** 773–781.

Wendel, J. F. 2000. Genome evolution in polyploids. *Plant Mol. Biol.* **42:** 225–249.

Wendel, J. F. and J. J. Doyle. 1998. Phylogenetic incongruence: window into genomes history and molecular evolution. In: *Molecular Systematics of Plants. II. DNA Sequencing* (D, Soltis, P. Soltis, and J. Doyle, Eds.), Boston: Kluwer Academy Publications.

Wendel, J. F., A. Schnabel, and T. Seelanan. 1995. Bi-directional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proc. Natl. Acad. Sci. USA* **92:** 280–284.

Whitlock, B. A. and D. A. Baum. 1999. Phylogenetic relationships of *Theobroma* and *Herrania* (Sterculiaceae) based on sequences of the nuclear gene *Vicilin*. *Syst. Bot.* **24:** 128–138.