Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 29 (2003) 435–455
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes

C. Donovan Bailey,[a,*] Timothy G. Carr,[b] Stephen A. Harris,[a] and Colin E. Hughes[a]

[a] *Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK*
[b] *Department of Ecology and Evolutionary Biology, Corson Hall, Cornell University, Ithaca, NY 14853, USA*

Received 20 January 2003; revised 15 August 2003

Abstract

Many early reports of ITS region (ITS 1, 5.8S, and ITS 2) variation in flowering plants indicated that nrDNA arrays within individuals are homogeneous. However, both older and more recent studies have found intra-individual nrDNA polymorphism across a range of plant taxa including presumed non-hybrid diploids. In addition, polymorphic individuals often contain potentially non-functional nrDNA copies (pseudogenes). These findings suggest that complete concerted evolution should not be assumed when embarking on phylogenetic studies using nrDNA sequences. Here we (1) discuss paralogy in relation to species tree reconstruction and conclude that a priori determinations of orthology and paralogy of nrDNA sequences should not be made based on the functionality or lack of functionality of those sequences; (2) discuss why systematists might be particularly interested in identifying and including pseudogene sequences as a test of gene tree sampling; (3) examine the various definitions and characterizations of nrDNA pseudogenes as well as the relative merits and limitations of a subset of pseudogene detection methods and conclude that nucleotide substitution patterns are particularly appropriate for the identification of putative nrDNA pseudogenes; and (4) present and discuss the advantages of a tree-based approach to identifying pseudogenes based on comparisons of sequence substitution patterns from putatively conserved (e.g., 5.8S) and less constrained (e.g., ITS 1 and ITS 2) regions. Application of this approach, through a method employing bootstrap hypothesis testing, and the issues discussed in the paper are illustrated through reanalysis of two previously published matrices. Given the apparent robustness of the test developed and the ease of carrying out percentile bootstrap hypothesis tests, we urge researchers to employ this statistical tool. While our discussion and examples concern the literature on plant systematics, the issues addressed are relevant to studies of nrDNA and other multicopy genes in other taxa.

Keywords: nrDNA; Internal transcribed spacer; Pseudogene; Phylogeny; Bootstrap hypothesis testing

1. Introduction

Sequences of the internal transcribed spacer region (ITS 1, 5.8S, and ITS 2) of nuclear ribosomal DNA (nrDNA) have been a staple source of data for the study of lower level phylogenetic relationships among plant taxa for more than ten years (e.g., Baldwin, 1992; Baldwin et al., 1995). In fact, the ITS region is one of the most widely applied molecular markers in current angiosperm systematics (e.g., Hershkovitz et al., 1999; Álvarez and Wendel, 2003). Some early reports of ITS variation provided results that were consistent with the existence of homogeneous nrDNA arrays within individuals (e.g., Ainouche and Bayer, 1997; Baldwin et al., 1995) presumably resulting from concerted evolution (gene conversion and unequal crossing over (Arnheim, 1983)). The effects of complete concerted evolution have been clearly documented between homeologous loci in some allotetraploid taxa of *Gossypium* (Wendel et al., 1995) and *Paeonia* (Sang et al., 1995). As a result, intra-individual polymorphism has generally been considered to be the exception rather than the rule for nrDNA (Mayol and Rosselló, 2001). Nevertheless, some studies have identified the occurrence of intra-individual nrDNA polymorphism in a range of taxa including

[*] Corresponding author. Present address: Department of Biology, New Mexico State University, P.O. Box 30001, Dept 3AF, Las Cruces, NM 88003-8001, USA.
E-mail address: dbailey@nmsu.edu (C. Donovan Bailey).

1055-7903/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2003.08.021

non-hybrid diploids and allopolyploids (e.g., Baker et al., 2000; Buckler et al., 1997; Campbell et al., 1997; Denduangboripant and Cronk, 2000; Doyle et al., 1990; Fuertes-Aguilar et al., 1999; Gaut et al., 2000; Gernandt and Liston, 1999; Hartmann et al., 2001; Hughes et al., 2002; Jobst et al., 1998; Kita and Ito, 2000; Kuzoff et al., 1999; Learn and Schaal, 1987; Linder et al., 2000; Mayol and Rosselló, 2001; Muir et al., 2001; O'Kane et al., 1996; Sang et al., 1995; Suh et al., 1993; Vargas et al., 1999; Widmer and Baltisberger, 1999). It has also become clear that polymorphic individuals often contain potentially non-functional nrDNA copies (pseudogenes) in addition to functional copies (e.g., Buckler et al., 1997; Hartmann et al., 2001; Hughes et al., 2002; Kita and Ito, 2000; Mayol and Rosselló, 2001; Muir et al., 2001; Yang et al., 1999). Clearly complete concerted evolution can no longer be assumed when embarking on studies utilizing nrDNA sequences (ITS, ETS, 5.8S, 18S, or 26S) for phylogenetic analysis of plant taxa.

The implications and importance of nrDNA polymorphism and pseudogenes for phylogenetic analyses of plant nrDNA sequences were discussed in the groundbreaking work of Buckler et al. (1997) and more recently by Mayol and Rosselló (2001). However, several major issues relating to the characterization of nrDNA polymorphism and pseudogenes, and their implications for species level phylogeny reconstruction remain ambiguous. These issues involve how to determine orthology and paralogy of nrDNA sequences, how to define and detect nrDNA pseudogenes, and the desirability of using pseudogenes in studies of phylogenetic relationship. For example, views on the use of pseudogene sequences in phylogenetic analysis of species and higher taxa include both deliberate a priori exclusion (e.g., Yang et al., 1999) and explicit inclusion (e.g., Buckler et al., 1997; Hughes et al., 2002).

Here we review and attempt to clarify the issues surrounding nrDNA polymorphism, pseudogenes, and species tree reconstruction. In the first section, we examine the relationship between intra-individual polymorphism in nrDNA and levels of paralogy, arguing that gene tree reconstruction is essential for understanding the potential complexity of nrDNA evolution and for determining orthology and interspecific paralogy. We distinguish between "shallow paralogs" and "deep paralogs" in order to clarify how nrDNA polymorphism and paralogy can affect the inference of species trees from gene trees. We then illustrate why orthology and paralogy of sequences from different individuals should not be inferred from the potential functionality of those sequences, arguing again that gene tree analyses are essential.

In subsequent sections we focus on nrDNA pseudogenes more closely, particularly noting that the identification of pseudogenes can act as a test of nrDNA sampling and that a priori exclusion of pseudogenes from gene tree analyses is generally unjustified. Definitions of pseudogenes and methods of detecting pseudogenes applicable to nrDNA are discussed. We conclude that expression is a poor criterion for identifying nrDNA pseudogenes and that patterns of nucleotide substitution are more appropriate in the context of phylogenetic systematics. We explore the relevance, reliability, and effectiveness of pseudogene detection methods which examine patterns of nucleotide substitution and present a formalized tree-based approach. Examples illustrating these issues and emphasizing the importance of sampling are provided through the reanalysis of nrDNA ITS data sets from *Lophocereus* Britton & Rose (Hartmann et al., 2001) and Brassicaceae (Yang et al., 1999) using a tree-based approach that relies on bootstrap hypothesis testing. Both of these data sets include putatively functional and non-functional copies. We show that the tree-based method has a number of general advantages. Compared to other approaches, it was more powerful at detecting pseudogenes, it revealed complex substitution patterns across gene trees that suggested a much broader range of evolutionary mechanisms, and it was useful for detecting errors such as long branch attraction.

2. Nuclear rDNA polymorphism, orthology, paralogy, and species trees

The reconstruction of phylogenetic relationships among species and higher taxa (species-trees) using data from multicopy sequences, such as nrDNA, depends upon a clear understanding of sequence relationships (gene trees: e.g., Avise, 1989; Doyle, 1992; Goodman et al., 1979; Pamilo and Nei, 1988; Sanderson and Doyle, 1992). If sequences are mistaken as orthologous (derived from speciation events), when they are in fact paralogous (derived from gene duplication events), relationships among species can be inferred incorrectly. While orthology is, by definition, an interspecific relationship among sequences, paralogy can occur at many levels—among sequences within an individual, among individuals within a species, and among species. The complex organization of nrDNA in combination with these varying levels of paralogy has led to confusion about both the level of paralogy that can interfere with species tree reconstruction and the appropriate methods for identifying orthologs and paralogs.

The organization of nrDNA complicates the characterization of intra-individual polymorphism. Individual nrDNA sequence units (the contiguous 18S, ITS 1, 5.8S, ITS 2, 26S, and IGS sequence) are present in multicopy arrays, with hundreds to thousands of each unit present within an array. The larger and more active arrays are called nucleolar organizer regions (NORs). Multiple nrDNA arrays within individuals have been identified for many plant taxa (reviewed in Hamby and Zimmer, 1992). As a result, intra-individual variation in nrDNA can occur within or between these multicopy arrays when concerted evolution is incomplete. Whether polymorphisms occur within or between sets of nrDNA arrays raises interesting questions in molecular evolution. However, it is simply the presence of more than one sequence type within an individual(s), rather than the distribution of sequence types among arrays, that potentially complicates the inference of species trees from gene trees.

Intra-individual nrDNA polymorphism is indicative of incomplete concerted evolution and suggestive of sequence paralogy at all these levels (intra-individual to interspecific). However, intra-individual polymorphism may also be due to heterozygosity at a locus or to homeology (orthologous sequences secondarily combined into the same genome through hybridization: e.g., Wendel et al. (1995)). Any duplicate copy of a DNA sequence (e.g., the contiguous 18S, ITS 1, 5.8S, ITS 2, 26S, and IGS sequence) found at a different position in the same genome is paralogous to any other such copy. Nevertheless, paralogy restricted to individuals or single species is not necessarily of concern in organismal phylogeny reconstruction.

To determine if intra-individual paralogs are maintained and shared with other species, and therefore a potential problem for species tree reconstruction, gene tree analyses are required. That is, gene trees are essential for determining if paralogy is solely intraspecific or also interspecific. Unfortunately, upon finding intra-individual polymorphism in nrDNA and suspecting paralogy, some researchers have inferred interspecific sequence paralogy (e.g., functional and pseudogene sequences are paralogous) without specific and primary reference to a gene tree. This omission has complicated the understanding of nrDNA paralogy in the context of phylogeny reconstruction. Failure to distinguish the difference between methods that are useful for inferring intra-individual sequence homology (heterozygosity and intra-individual/intraspecific paralogy) from those that can be used to infer orthology and interspecific sequence paralogy (i.e., gene trees only), has led to confusion in the characterization of the level of paralogy that can interfere with organismal phylogeny reconstruction using multicopy sequences. This seems to be particularly true for studies based on nrDNA sequence data.

The complexity of nrDNA paralogy in phylogenetic analyses of species can be clarified by distinguishing two categories of historical paralogs. The boundary between these categories is relative to the phylogenetic question being addressed. For the purposes of this discussion, the boundary between populations (or individuals) and species is used to distinguish "shallow paralogy" from "deep paralogy." Shallow paralogy is duplication/divergence subsequent to the most recent speciation event (Fig. 1A) resulting in a species having more than one locus that is orthologous to a locus in other species (Doyle and Davis, 1998). Sequences representing shallow paralogs should be monophyletic on a gene tree (excluding issues of branch attraction) and therefore shallow paralogs should not adversely affect species tree reconstruction (e.g., Baker et al., 2000; Denduangboripant and Cronk, 2000; Hughes et al., 2002). Fig. 1A depicts a hypothetical scenario in which extensive duplication of loci has occurred within one species but not within others. Assuming that individuals from species D are diploid and that D is homozygous at each locus for the sequence under study, we can say that at least three paralogs exist within species D, but that each of these sequences is orthologous to the sequences from species A, B, and C. Any, or all, of the sequences from species D could be used to accurately reconstruct relationships between species A, B, C, and D.

In contrast, deep paralogs result from duplications and divergence prior to speciation events (Figs. 1B and C) and therefore span two or more species (the equivalent of divergent paralogs, sensu Baldwin et al. (1995)).

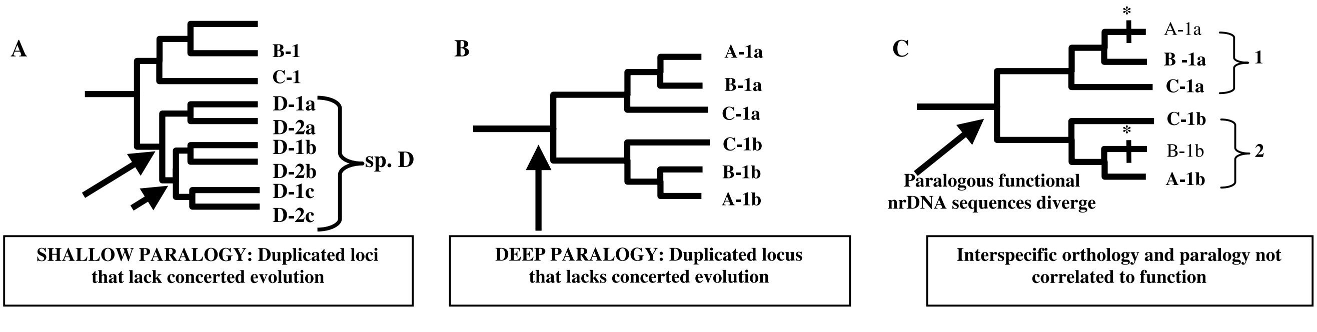
A
B-1
C-1
D-1a
D-2a
D-1b
D-2b
D-1c
D-2c
sp. D
SHALLOW PARALOGY: Duplicated loci that lack concerted evolution
B
A-1a
B-1a
C-1a
C-1b
B-1b
A-1b
DEEP PARALOGY: Duplicated locus that lacks concerted evolution
C
*
A-1a
B -1a
C-1a
1
*
C-1b
B-1b
A-1b
2
Paralogous functional nrDNA sequences diverge
Interspecific orthology and paralogy not correlated to function

Fig. 1. Paralogy and orthology in nrDNA gene trees. Upper case letters, numbers, and lower case letters designate species, accessions within a species, and cloned sequences from an accession, respectively. The arrows indicate divergence of nrDNA duplicate copies. (A) An example of shallow paralogy in accessions of species D. Although each accession of species D has multiple ITS sequence types, species D is monophyletic on the gene tree. (B) An example of deep paralogy in which each accession is represented in the two clades of nrDNA sequences. (C) An example of the deep paralogy and orthology possible within and between functional and pseudogene nrDNA sequences. The arrow indicates the point at which two functional nrDNA sequence copies diverged in the ancestor to species A, B, and C. Each asterisk marks a shift to a lack of functional constraint, one in each NOR. Putative pseudogene and functional sequences are in italicized and plain font, respectively.

Deep paralogs may be identified by the polyphyletic occurrence of sequences from a species on an nrDNA gene tree, but unidentified paralogy is not necessarily the only cause of polyphyletic species. When deep paralogy is present, the potential for generating an erroneous species tree from unknowingly sampling a mixture of sequences from each clade is high (e.g., Sanderson and Doyle, 1992; Slowinski and Page, 1999). Clear examples of deep paralogy in nrDNA not complicated by known homeology have been identified in at least two angiosperm genera (*Lophocereus*: Hartmann et al. (2001); and *Quercus*: Vazquez et al. (2000); Muir et al. (2001)).

3. Orthology, paralogy, and functionality

Our argument in this section is that orthology and paralogy should not be inferred on the basis of sequence functionality. Because pseudogene copies of nrDNA can be easily amplified, researchers are faced with the question of their relationship to the functional copies from which they were likely derived, specifically, whether they represent deep or shallow paralogs. It is easy to assume that pseudogenes will be interspecific paralogs to functional copies, but this need not be the case. Gene tree analyses are essential for determining the orthology and interspecific paralogy of sequences, irrespective of functionality.

For example, Buckler et al. (1997) distinguished putatively functional and non-functional sequences (pseudogenes), which they called class I and class II paralogs, as deep (divergent) paralogs that complicate phylogenetic analyses of species relationships. They identified these classes by examining factors correlated with functionality (substitution rates/patterns and free energies). Characterization of these functional classes as paralogs may be justified for intra-individual paralogy. That is, functional and non-functional sequences within individuals of a species are likely to be paralogous to one another. However, sequence history, and hence orthology and interspecific paralogy, cannot be established using the techniques for distinguishing functional/non-functional sequences. Using these functional classes to infer interspecific paralogy is inappropriate, since pseudogenes can be paralogous or orthologous to functional sequences in an analysis of nrDNA.

The hypothetical gene tree in Fig. 1C illustrates why functionality alone should not be used to infer the orthology or paralogy of sequences taken from more than one species. Within each orthologous set of sequences pseudogenes were detected. The pattern of relationships among sequences suggests two parallel losses of function following the divergence of sequences in *Group 1* and *Group 2*. The ancestral lineage leading to species A, B, and C was presumably polymorphic for two divergent functional nrDNA sequence types (paralogs). Following the divergence of species A and B, a sequence in *Group 1* lost functionality along the branch leading to A, while in *Group 2* loss of functionality occurred along the branch leading to B, leaving both species A and B with one functional and one pseudogene nrDNA sequence type. Within each clade, functional sequences are orthologous to pseudogene sequences while some functional copies between groups are paralogs. In addition, the inadvertent use of one functional sequence type from species C in combination with the single functional types from A and B would mislead phylogeny reconstruction.

Buckler et al. (1997) suggested that class II "paralogs," which were designated as potential pseudogenes, introduce what we call "deep paralogs" (or "divergent paralogs" in Baldwin et al. (1995)) into an analysis, thereby complicating phylogeny reconstruction (also discussed by Mayol and Rosselló, 2001). While it is certainly possible that class I and class II sequences will be paralogous to one another in a gene tree including multiple species, this should not be assumed. As shown, the class I and class II paralog categorization is potentially misleading in the context of phylogeny reconstruction. Furthermore, empirical data do not support the idea that class I and class II "paralogs" necessarily represent deep paralogs (e.g., Baker et al., 2000; Buckler et al., 1997; Denduangboripant and Cronk, 2000; Hughes et al., 2002; Kita and Ito, 2000). The inclusion of both class I and class II sequences in gene tree analyses is critical for delimiting the boundaries between shallow and deep paralogy. Although Buckler et al. (1997) strongly advocated the importance of identifying and including polymorphisms in gene trees, the assumption of functional "paralogy" appears to have led some researchers to intentionally exclude non-functional sequences prior to gene tree analyses. Disregarding potential pseudogenes may result in under-sampled gene trees, greatly decreasing the potential to fully understand the orthology and paralogy of sequences present in an analysis and the ability to accurately infer species relationships.

In our view, the reasons pseudogene sequences can introduce potential problems may have nothing to do with sequence history (i.e., orthology and paralogy). Rather, concerns about mixing functional and non-functional nrDNA copies in a phylogenetic analysis have more to do with concerns about differing patterns of substitution leading to long branch attraction (Felsenstein, 1978), short branch repulsion (Siddall, 1998), or alignment difficulties, than with the phylogenetic history of the sequences involved (see also Kita and Ito, 2000).

4. When are nrDNA polymorphisms likely to be of interest?

All known examples of interspecifically maintained nrDNA paralogous polymorphism occur among closely related species or, in a few cases, between genera. To the best of our knowledge, shared polymorphisms in more distantly related taxa have not been directly uncovered in studies of plant nrDNA. For example, there is no evidence of deep maintained paralogs between conifers and *Arabidopsis*. For pseudogene related polymorphisms this is hardly surprising given that extensive divergence would eventually destroy primer sites in unconstrained pseudogene sequences making such sequences difficult or impossible to amplify. However, for functional types it would seem that maintained deep paralogy is either uncommon or has been hidden by potentially intermittent episodes of complete concerted evolution.

Ancestral polymorphisms can still mislead higher-level phylogenetic nrDNA studies. However, the evidence necessary to identify such problems may have been lost over extensive nrDNA history as a result of concerted evolution. If multiple nrDNA sequences found in each individual are likely to be monophyletic on a gene tree that only includes more distantly related taxa, then there is little to be gained by extensively searching for intra-individual polymorphism. When nrDNA sequences from distantly related taxa are used to infer higher-level relationships (e.g., Nickrent and Soltis, 1995; Soltis et al., 1997), evidence for gene tree-species tree conflict will most likely be provided by reference to other sources of data (e.g., conflict with phylogenies generated with other data sources).

In contrast, maintained polymorphisms of both functional and non-functional sequence types have been found in numerous studies comparing closely related species and have even been found within a few plant families (e.g., Winteraceae). At the species level there are clearly tangible benefits to be gained from identifying polymorphisms as they provide direct evidence that gene tree-species tree conflicts are possible (e.g., Buckler et al., 1997; Kita and Ito, 2000; Mayol and Rosselló, 2001).

Detailed discussion of the range of techniques that can be used to detect nrDNA polymorphism is beyond the scope of this paper. What is clear is that detecting the full extent of nrDNA polymorphism may not be straightforward and may require sensitive techniques. These include PCR amplification under a variety of conditions (Buckler et al., 1997), extensive cloning (e.g., Baker et al., 2000; Buckler et al., 1997; Hughes et al., 2002), the use of restriction enzyme assays (e.g., Hughes et al., 2002; Lim et al., 2000), varying PCR primer combinations, and targeting low-copy or difficult to amplify nrDNA sequences with specific amplification primers (Rauscher et al., 2002). By implication, studies relying solely on direct sequencing may fail to detect potentially informative variation. Several of these issues are reviewed in more detail in Álvarez and Wendel (2003).

5. Why are nrDNA pseudogenes of interest?

In a subset of the cases in which intra-individual polymorphisms in nrDNA have been identified, the potential functionality of sequences has also been assessed. These studies suggest that putatively functional and non-functional sequences are both commonly amplified when nrDNA polymorphisms are encountered (e.g., Buckler et al., 1997). Furthermore, the dynamics of amplification (involving copy number, secondary structure, and primer site conservation) can result in the preferential amplification of either functional or pseudogene sequences (e.g., Buckler et al., 1997).

One fundamental reason for determining if sequences come from pseudogenes is that this acts as a test of whether sampling of an individual's nrDNA is complete. Every individual with a known pseudogene sequence must be polymorphic for nrDNA sequences because functional copies are required for transcription. If no functional copy is identified from an individual, then a gene tree constructed from the sequences is certainly under-sampled. The identification of one functional copy does not mean, however, that sampling is complete.

The potential impacts of under-sampling nrDNA diversity depend on whether functional and non-functional copies from species are monophyletic on a gene tree (see above and Fig. 1). Assessing monophyly on a gene tree is complicated by the possibility that limited sampling may influence the overall inter-individual and/or interspecific gene tree topology. Examinations of sequence functionality have detected under-sampling of ITS diversity in studies of *Desmanthus* (Hughes et al., unpublished data) and *Leucaena* (Hughes et al., 2002), and under-sampling is also apparent upon reexamination of previously published *Quercus* (Mayol and Rosselló, 2001), *Lophocereus* (Hartmann et al., 2001), Brassicaceae (Yang et al., 1999), and *Aconitum* (Kita and Ito, 2000) data sets.

Potential pseudogene sequences have also been identified in order to exclude them a priori from phylogenetic analysis (e.g., Yang et al., 1999, numerous personal communications). The rationales underlying a priori exclusion are usually unstated but may include (1) assumptions that pseudogenes are deeply paralogous to functional loci in a way that interferes with the inference of species trees, (2) assumptions that pseudogenes can only be "shallow paralogs" and therefore unimportant, and (3) concerns about elevated rates of evolution in pseudogenes. As we have shown above, however, a priori exclusion on the basis of rationales 1 and 2 is unjustified given the possible complexity of paralogy and orthology relations in nrDNA. Both assumptions need to be tested by increasing sampling and/or constructing gene trees including functional and pseudogene sequences. Elevated rates of evolution in pseudogenes certainly prompt concerns about sequence alignment, long-branch attraction (Felsenstein, 1985), and short-branch repulsion (Siddall, 1998). However, problems with alignment and branch attraction/repulsion are basic issues relating to the potential utility of *any* sequence used for phylogeny reconstruction, regardless of functionality. Not all pseudogene sequences will necessarily be subject to any or all of these problems (e.g., Hughes et al., 2002), nor are functional sequences necessarily immune to them.

In contrast to a priori exclusion based on paralogy, Richardson et al. (2001) excluded pseudogene sequences from a matrix used to estimate divergence times, and in this circumstance exclusion may be justified. If the exclusion of pseudogenes does not change the inferred relationships among the functional sequences, then excluding pseudogenes to estimate divergence seems appropriate. Of course this does not mean that functional (or pseudogene) copies alone will necessarily follow a molecular clock. While we located only two examples of explicit pseudogene exclusion in the literature (Richardson et al., 2001; Yang et al., 1999), informal discussions with numerous researchers lead us to believe the practice of exclusion is widespread.

Because the identification of potential nrDNA pseudogene sequences is essential for assessing gene tree sampling and inferring species relationships, methods for accurately determining sequence functionality are important. However, even a cursory examination of the range of techniques and criteria that have been used to distinguish pseudogene sequences from functional copies (see below), suggests that researchers are either defining pseudogenes in different ways or using inappropriate identification techniques.

6. Definitions of nrDNA pseudogenes

Standard definitions of "pseudogene" are difficult to apply to nrDNA. Consider the following common definitions: "a silent, non-functional DNA sequence," "non-functional genes related in sequence to functional genes," and "sequences which resemble the functional genes with which they are associated, but which differ at a number of base pair sites and are not transcribed because they have internal "stop" codons" (all in Futuyma, 1998). These definitions, or aspects of them, are problematic when multicopy sequences undergoing concerted evolution are considered. This is especially true when such sequences occur in multicopy arrays and/or do not code for proteins, as in nrDNA. For nrDNA, two related problems arise: (1) the meaning of "functional" is not as straightforward as in protein-coding genes, and (2) criteria which may be neither necessary nor sufficient for identifying nrDNA pseudogenes are included in the definition.

The first problem arises because many of these nrDNA copies (loci) are non-functional in the sense that they are not transcribed, yet because they are influenced by concerted evolution they may retain the exact characteristics (i.e., the nucleotide sequence) of actively transcribed (functional) copies. In nrDNA, a small subset of repeats are expressed at a given time and some repeats may not be expressed (Lim et al., 2000). Presumably functional nrDNA repeat types can exist in methylated/condensed unexpressed forms throughout the ontogeny of an individual (Leitch et al., 1992; Lim et al., 2000). While these might be considered pseudogenes by someone studying expression of nrDNA, from the standpoint of molecular systematics, it makes little sense to differentiate them from transcribed copies since they contain the same sequence.

Identifying necessary and sufficient criteria for determining nrDNA pseudogenes would help in forming a useful definition. Nucleotide diversification and expression patterns have generally been considered simultaneously when attempting to define and identify nrDNA pseudogenes as "non-functional sequences" (e.g., Buckler et al., 1997; Kwon et al., 1991; Muir et al., 2001). However, changes in expression and sequence diversification can be independent of one another (for coding or non-coding DNA), suggesting that one need not consider both of these factors when defining the term "pseudogene."

While sequence expression, or lack thereof, may be indicative of functional (or functioning) and pseudogene loci, there is little reason to assume that pseudogenes will not be expressed or that potentially functional copies will necessarily be expressed. Many unexpressed nrDNA sequences, which may contain identical nucleotide sequences to expressed "functional" types as a result of concerted evolution, would be designated as "pseudogenes" if expression were taken as a necessary attribute of a pseudogene sequence. Indeed, expression may be a poor criterion for determining pseudogenes derived from protein-coding genes. Expressed pseudogenes have been documented (e.g., Choi et al., 2001; Hirotsune et al., 2003). We even find evidence for an expressed nrDNA pseudogene in the data of Hartmann et al. (2001, discussed below).

Thus, for those concerned with pseudogenes in the context of phylogenetic analysis of sequence data, the key aspect of pseudogenes is that "they are subject to no functional constraints" (Li and Graur, 1991) and so their nucleotide substitution pattern should reflect this. In the context of phylogeny reconstruction, pseudogenes (nrDNA or protein coding) can be defined as sequences whose nucleotide divergence pattern has not been constrained by function irrespective of expression patterns.

7. Detecting pseudogenes

How to detect nrDNA pseudogenes remains a perplexing issue. This is partly because of problems with how pseudogenes have been defined (above) and partly due to the continuum of change found between obvious functional copies and obvious pseudogenes (i.e., sequences that have lost functional constraint but have not extensively diverged from functional copies).

The methods that have been used to detect pseudogene nrDNA sequences rely on examining attributes of sequences that are presumably correlated with gene function/lack of function. One group of methods attempts to detect whether sequences are transcribed. These include the assessment of direct transcription (e.g., Buckler et al., 1997; Hartmann et al., 2001; Muir et al., 2001), DNA condensation (e.g., Leitch et al., 1992), degree of methylation (e.g., Lim et al., 2000; Torres-Ruiz and Hemleben, 1994), and copy number (e.g., Hartmann et al., 2001). These methods are not discussed in detail because expression patterns may not be good indicators of functional constraint in nrDNA (see above).

Other methods look at whether substitution patterns are consistent with functional constraint. These methods examine nucleotide diversification/divergence (e.g., Buckler and Holtsford, 1996; Buckler et al., 1997; Hershkovitz et al., 1999; Hughes et al., 2002; Yang et al., 1999), insertion–deletion events (e.g., Baldwin, 1992; Hartmann et al., 2001; Hershkovitz et al., 1999; Hughes et al., 2002; Kwon et al., 1991), sequence free energy (e.g., Buckler et al., 1997; Mayol and Rosselló, 2001), secondary structure (e.g., Denduangboripant and Cronk, 2000; Gernandt and Liston, 1999; Hartmann et al., 2001; Mayol and Rosselló, 2001), and methylation induced substitution patterns (e.g., Buckler and Holtsford, 1996; Buckler et al., 1997; Lim et al., 2000; Mayol and Rosselló, 2001). The applicability of these methods to detecting nrDNA pseudogenes, as defined here, is discussed below.

8. Methods of detection

8.1. Nucleotide diversification

Patterns of nucleotide substitution have been used as evidence to distinguish between functionally constrained and unconstrained nrDNA sequences (e.g., Buckler and Holtsford, 1996; Buckler et al., 1997; Hershkovitz et al., 1999; Hughes et al., 2002; Mayol and Rosselló, 2001; Yang et al., 1999). In most cases, this approach has relied on pairwise methods to detect degrees of divergence based on the assumption that functional sequences are under strong selective constraints that limit their substitution rates (Graur and Li, 2000; Kimura, 1983). Patterns of nucleotide sequence divergence provide powerful and reliable evidence of whether sequences represent functional or pseudogene types. Each nucleotide substitution is a discrete entity that can be easily tallied and compared. However, some methods of examining nucleotide substitution patterns are more rigorous than others.

Two types of pairwise comparisons can be used to detect pseudogene sequences. One type involves comparison of conserved regions or entire sequences between a presumed functional sequence (or consensus sequence) and another sequence (e.g., Kita and Ito, 2000; Yang et al., 1999). Such a simplistic approach fails to consider overall rates of sequence change, and in some implementations it fails to provide a non-arbitrary quantitative criterion for deciding what constitutes considerable divergence. In addition, this approach may fail if a matrix is composed mostly of pseudogenes, unless known functional sequences from outside the matrix are used as comparators.

A more rigorous approach, which incorporates the comparison of conserved and relatively unconstrained regions between two sequences, takes into account relative divergence. Sequences that may have undergone an overall increase in substitution rate, but that maintain a conserved 5.8S relative to ITS 1 and ITS 2, can be distinguished from those that have a comparable rate of change across the 5.8S, ITS 1, and ITS 2 (i.e., pseudogenes). For the ITS region, it is generally agreed that putatively functional copies maintain a conserved 5.8S sequence while the ITS 1 and ITS 2 sequences diverge more quickly (e.g., Buckler et al., 1997). Conversely, for pseudogene sequences, which lack functional constraint, we expect that the putatively conserved regions (e.g., 5.8S) will have substitution rates equal to those in the relatively unconstrained regions (e.g., ITS 1 and ITS 2), all other things being equal.

Relative rate tests have been used to detect pseudogenes in both types of comparisons (e.g., Buckler et al., 1997; Muir et al., 2001; Yang et al., 1999). The reasoning underlying the use of relative rate tests is that a pseudogene sequence should exhibit an increased rate of nucleotide substitution throughout its length relative to functional copies. Although quantitative, relative rate tests are not particularly sensitive tests for detecting pseudogenes, especially since increases in evolutionary rate may have causes other than loss of functional constraint. When single regions or entire sequences are compared, it is unknown whether an increase in evolutionary rate occurs in a particular region (e.g., ITS 1) or throughout a sequence (e.g., ITS 1 and 5.8S). This difficulty arises because strict implementation of the second type of comparison (directly comparing typically conserved and unconstrained regions) is impossible with a relative rate test (although not with other possible pairwise methods). The best that can be done is to determine if the ITS region shows an increased rate of evolution relative to other ITS regions and if the 5.8S also shows an increased rate relative to other 5.8S regions. In addition, the use of relative rate tests to identify pseudogenes assumes that functional sequences are already known. If one or both of the presumed functional sequences used as comparators are pseudogenes, the results of the test can be misleading for identifying pseudogene sequences. Finally, the sensitivity of some relative rate tests to the sequences that are chosen as comparators, further complicates inference.

Ultimately pairwise approaches cannot take into account the overall pattern of nucleotide substitution that can be identified using a phylogenetic tree (e.g., Wenzel and Siddall, 1999). Pairwise methods do not reveal the pattern of divergence, only a fixed value of divergence between two sequences or sets of sequences. Furthermore, pairwise approaches can suffer because sequences are not independent but are related to varying degrees. As the degree of relatedness varies, so does the degree of non-independence, yet this can be difficult to account for in a pairwise framework. In contrast, tree-based approaches partition similarity among sequences into independent units based on the genealogy, thus allowing a suite of independent tests to be constructed. Furthermore, unlike relative rate tests, a tree-based approach allows direct comparison of rates of evolution in different gene regions along a branch. As a result, it does not rely on obtaining presumed functional copies for comparison.

An example of the benefit that can be gained from using a phylogenetic approach was inadvertently provided by Buckler et al. (1997). These authors used Kimura distances and observed less variation in the 5.8S region than would be expected among putative pseudogenes from nrDNA data sets representing *Gossypium*, *Nicotiana*, *Tripsacum*/*Oryza*, Winteraceae, and *Zea* taxa. One of their possible explanations for this phenomenon was "when two functional nrDNA arrays diverge, the ITS regions will diverge faster than the 5.8S. Subsequently, if one array becomes transcriptionally inactive its 5.8S will begin to diverge rapidly (Baldwin et al., 1995: p. 827)." This statement is easiest to interpret in the context of a tree. A functional copy may become a pseudogene copy anywhere along a branch. If the shift is not at the very earliest stage of nucleotide divergence along a branch, the ITS sequence may diverge somewhat before the 5.8S is released from constraint. In this case ITS will show greater divergence than the 5.8S along that branch. While this statement is certainly true, it also identifies an inherent limitation of pairwise approaches for assessing nrDNA divergence patterns correlated with maintained function. Such values reveal comparatively little information, particularly if two terminals are well separated on a phylogenetic tree (Wenzel and Siddall, 1999). By taking into account the pattern of nucleotide substitution across a phylogenetic tree the possible explanation could have been tested. If an ancestral branch has this pattern and subsequently derived branches show a typical pseudogene substitution pattern then the hypothesis is not refuted. However if ancestral and descendant branches both show this pattern the hypothesis is not supported.

Comparisons of conserved and relatively unconstrained regions along branches of a phylogenetic tree can provide better inferences about functionality based on patterns of nucleotide substitution than the corresponding pairwise approaches. Gene trees have been used to quantify methylation induced substitutions, shared substitution patterns, and to investigate the distribution of nrDNA pseudogenes (e.g., Buckler and Holtsford, 1996; Buckler et al., 1997; Lim et al., 2000), but they have rarely been used to infer the presence of pseudogenes based on comparisons of nucleotide substitution patterns in putatively constrained and unconstrained regions assessed for individual branches on a tree (e.g., Hughes et al., 2002).

Results from the two pairwise and tree-based methods can differ dramatically, as illustrated below (see Results). This is particularly true when the simple pairwise approach, examining only one region (in this case the 5.8S), is compared to the relative divergence approach (comparing two or more regions) using either pairwise comparisons or a tree-based approach. However, differences are also possible between the results of pairwise comparisons and phylogenetic methods both using the relative divergence approach. In this case, discrepancies are likely when a single sequence is selected as a "functional type" for pairwise comparisons. The interpretation of all such pairwise comparisons can be far more difficult than interpretation of results from a phylogenetic approach.

8.2. Length variation

Indels in highly conserved regions of nrDNA sequences (e.g., 5.8S, 18S, or 26S) have also been taken as indicators of potential pseudogenes (e.g., Baldwin, 1992; Hughes et al., 2002; Kwon et al., 1991). For example, the loss or gain of 5.8S sequence has been used to infer that a sequence lacks functional constraints because putatively functional 5.8S regions are conserved to 163–164 bp in most angiosperms (Baldwin et al., 1995). However, the observation that functional 5.8S regions are 163–164 bp long does not entail its converse, that sequences within the "functional" length range necessarily represent functional copies. Length alone provides little information about the underlying nucleotide sequence, and this method of pseudogene detection is only relevant if the sequences contain indels that place them outside of known functional ranges. In contrast to the highly conserved length of the 5.8S region in angiosperms, the 18S and 26S regions have greater length variation (Hershkovitz et al., 1999). The use of length variation for nrDNA pseudogene identification in these regions should be approached with even more caution than its use in the 5.8S region.

Indels in relatively unconstrained regions such as ITS 1 and ITS 2 have also been used in addition to, or in place of, 5.8S indels as indicators of putative pseudogenes (e.g., Hartmann et al., 2001; Kita and Ito, 2000; Mayol and Rosselló, 2001). While changes (substitutions or indels) in a few very small (<26 bp) potentially constrained ITS regions (Buckler and Holtsford, 1996; Gernandt and Liston, 1999; Mayol and Rosselló, 2001) may provide some information about nrDNA sequence functionality, in general, there is considerable length variation among functional copies of the two ITS segments Hershkovitz et al. (1999). The lengths of putatively functional ITS sequences in angiosperms are known to range from 187–298 bp for ITS 1 and 187–252 bp for ITS 2 (Baldwin et al., 1995). ITS regions from presumably functional nrDNA loci range to over 3000 bp in non-flowering seed plants (Liston et al., 1996). This level of variation largely precludes general reliance on ITS 1 or ITS 2 indels for pseudogene identification.

8.3. Secondary structure and free energy

Secondary structure of the 5.8S region (e.g., Hartmann et al., 2001), the ITS 1 and ITS 2 regions (e.g., Denduangboripant and Cronk, 2000; Gernandt and Liston, 1999; Hartmann et al., 2001; Mayol and Rosselló, 2001), and/or free energies of these regions (e.g., Buckler et al., 1997; Mayol and Rosselló, 2001) have also been used to distinguish potentially functional from non-functional sequences. Both of these factors can indeed represent measures of functional constraint and have the potential to provide even more detailed assessment of conservation of particular correlated nucleotide patterns than sequence data alone (Mai and Coleman, 1997). However, estimates of secondary structure and free energy are dependent on folding models derived from select organisms so that their general applicability across taxonomic groups is debatable. This is particularly true for the relatively unconserved primary sequence of ITS 1 and ITS 2. For example, Baldwin et al. (1995) showed little conservation of ITS 1 secondary structure across five plant families. They noted greater conservation for ITS 2 aligned from just three asterid families but found that radically different folding patterns were only marginally less likely than those selected.

Given the potential uncertainty of accurately estimating secondary structure and free energies across broad taxonomic groups, we view these criteria as secondary sources of information on potential functionality that can be used to corroborate or refute conclusions drawn from patterns of nucleotide substitution and perhaps indels. Estimates of secondary structure and free energies drawn from more highly conserved regions, such as 5.8S, are likely to provide more reliable indicators of function than those drawn from more variable regions (e.g., ITS 1 and ITS 2). This view is supported by empirical results from human mitochondrial and nuclear ribosomal 12S paralogs. Olsen and Yoder (2002) demonstrated that the extensive models for human 12S sequences could not distinguish mitochondrial derived 12S pseudogenes from their functional nuclear paralogs. If models for rDNA fail for such a well-studied organism, their universal utility must remain questionable. Such models may be of primary use when a non-functional sequence type has only slightly diverged from functional types. Chance substitutions in key conserved regions may indicate a pseudogene locus before sufficient substitutions have accumulated for other methods (e.g., relative divergence) to detect such changes (Mai and Coleman, 1997). However, pseudogenes will not be detected if random substitutions have failed to hit a few "key" sites, even if the overall pattern of substitution is inconsistent with functional constraint.

8.4. Methylation induced substitution patterns

The general use of methylation patterns to distinguish functional from non-functional nrDNA types straddles the boundary between methods associated with substitution patterns in the broad sense (presented in this section) and those that deal with expression (not addressed—see above). Since methylation patterns have been correlated with specific categories of substitutions (e.g., CpG and CpNpG; Gardiner-Garden et al., 1992), methylation-induced substitution patterns, calculated from a phylogenetic tree, have been used as a source of information for pseudogene identification (e.g., Buckler and Holtsford, 1996; Buckler et al., 1997; Lim et al., 2000). However, specific substitution patterns, irrespective of their underlying causal factors, are the evidence of whether or not a sequence has diverged under functional constraint. Methylation provides one potential explanation for a pattern of nucleotide substitution inconsistent with functional constraint.

9. Example analyses

9.1. Methods

9.1.1. Data sets and phylogenetic analysis

Two previously published nrDNA ITS sequence matrices that included putatively functional and non-functional sequences (Hartmann et al. (2001) for *Lophocereus* (Cactaceae) and Yang et al. (1999) for Brassicaceae) were selected to: (1) provide examples of some of the issues discussed above; (2) to illustrate application of a tree-based approach to pseudogene detection; and (3) to highlight the critical importance of extensive intra-individual, intraspecific, and interspecific sampling in nrDNA studies. The sequences are available in GenBank (Accession Nos. included in Fig. 2). The *Lophocereus* matrix included all the sequences presented in the original publication (which excluded publication of the ITS 2 region; Hartmann et al., 2001), whereas the Brassicaceae matrix included the sequences analyzed by Yang et al. (1999) plus two sequences that were added to illustrate the importance of taxon sampling (*Berteroella maximowiczii* O. E. Schulz, AF137573; *Arabis jacquinii* G. Beck., AJ232919) and a third to represent an appropriate outgroup (*Cleome lutea* E. Mey., AF137588).

Sequences were provisionally aligned using ClustalX ver. 1.8 (Thompson et al., 1997) and then adjusted by eye in WinClada (Nixon, 1999). In the phylogenetic analyses, all characters were scored as unordered and equally weighted. Parsimony analyses were conducted using NONA (Goloboff, 2000) spawned via WinClada (Nixon, 1999) using 1000 random addition sequences, tree bisection and reconnection, holding 50 trees per replication, and attempting to swap to completion (h/50; mult*1000; max*). One thousand strict consensus bootstrap replicates (Davis et al., 1998) each comprising 10 random addition sequences and holding 100 trees (h/100; mult*10) were spawned from WinClada into NONA.

9.1.2. Statistical test for pseudogenes based on substitution patterns

Our test for the presence of nrDNA pseudogenes is derived from the definition of pseudogene accepted in

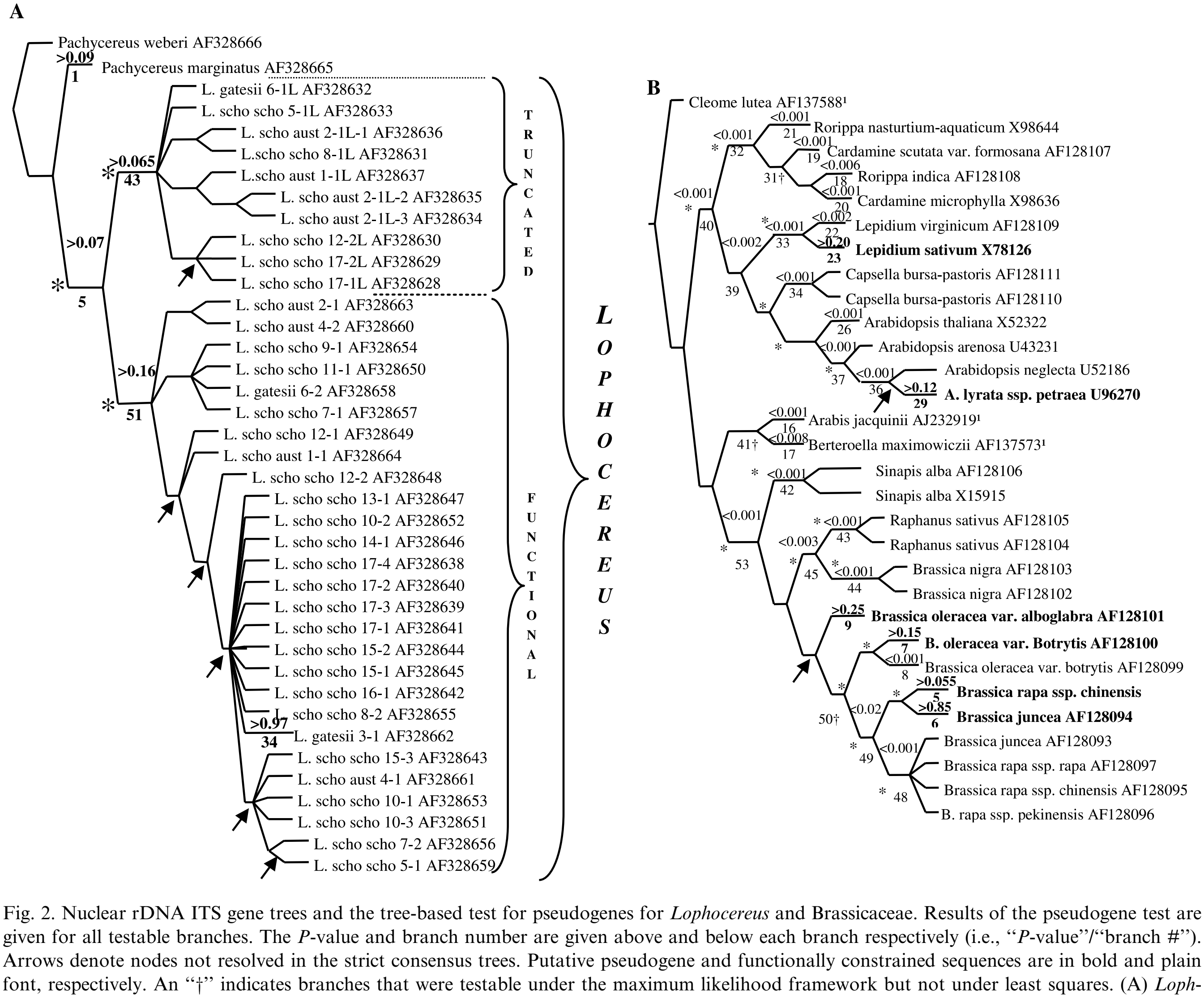
A
Pachycereus weberi AF328666
>0.09
1
Pachycereus marginatus AF328665
L. gatesii 6-1L AF328632
L. scho scho 5-1L AF328633
L. scho aust 2-1L-1 AF328636
L.scho scho 8-1L AF328631
L.scho aust 1-1L AF328637
L. scho aust 2-1L-2 AF328635
L. scho aust 2-1L-3 AF328634
L. scho scho 12-2L AF328630
L. scho scho 17-2L AF328629
L. scho scho 17-1L AF328628
>0.065
43
>0.07
5
TRUNCATED
L. scho aust 2-1 AF328663
L. scho aust 4-2 AF328660
L. scho scho 9-1 AF328654
L. scho scho 11-1 AF328650
L. gatesii 6-2 AF328658
L. scho scho 7-1 AF328657
>0.16
51
L. scho scho 12-1 AF328649
L. scho aust 1-1 AF328664
L. scho scho 12-2 AF328648
L. scho scho 13-1 AF328647
L. scho scho 10-2 AF328652
L. scho scho 14-1 AF328646
L. scho scho 17-4 AF328638
L. scho scho 17-2 AF328640
L. scho scho 17-3 AF328639
L. scho scho 17-1 AF328641
L. scho scho 15-2 AF328644
L. scho scho 15-1 AF328645
L. scho scho 16-1 AF328642
L. scho scho 8-2 AF328655
>0.97
34
L. gatesii 3-1 AF328662
L. scho scho 15-3 AF328643
L. scho aust 4-1 AF328661
L. scho scho 10-1 AF328653
L. scho scho 10-3 AF328651
L. scho scho 7-2 AF328656
L. scho scho 5-1 AF328659
FUNCTIONAL
LOPHOCEREUS
B
Cleome lutea AF137588¹
<0.001
21
Rorippa nasturtium-aquaticum X98644
<0.001
32
<0.001
19
Cardamine scutata var. formosana AF128107
<0.006
31†
18
Rorippa indica AF128108
<0.001
20
Cardamine microphylla X98636
<0.001
40
<0.002
22
Lepidium virginicum AF128109
<0.001
33
>0.20
23
Lepidium sativum X78126
<0.001
34
Capsella bursa-pastoris AF128111
Capsella bursa-pastoris AF128110
39
<0.001
26
Arabidopsis thaliana X52322
<0.001
Arabidopsis arenosa U43231
37
<0.001
36
Arabidopsis neglecta U52186
>0.12
29
A. lyrata ssp. petraea U96270
<0.001
16
Arabis jacquinii AJ232919¹
<0.008
17
Berteroella maximowiczii AF137573¹
41†
<0.001
42
Sinapis alba AF128106
Sinapis alba X15915
<0.001
<0.001
43
Raphanus sativus AF128105
Raphanus sativus AF128104
<0.003
45
<0.001
44
Brassica nigra AF128103
Brassica nigra AF128102
53
>0.25
9
Brassica oleracea var. alboglabra AF128101
>0.15
7
B. oleracea var. Botrytis AF128100
<0.001
8
Brassica oleracea var. botrytis AF128099
50†
<0.02
>0.055
5
Brassica rapa ssp. chinensis
>0.85
6
Brassica juncea AF128094
49
<0.001
Brassica juncea AF128093
Brassica rapa ssp. rapa AF128097
48
Brassica rapa ssp. chinensis AF128095
B. rapa ssp. pekinensis AF128096

Fig. 2. Nuclear rDNA ITS gene trees and the tree-based test for pseudogenes for *Lophocereus* and Brassicaceae. Results of the pseudogene test are given for all testable branches. The *P*-value and branch number are given above and below each branch respectively (i.e., "*P*-value"/"branch #"). Arrows denote nodes not resolved in the strict consensus trees. Putative pseudogene and functionally constrained sequences are in bold and plain font, respectively. An "†" indicates branches that were testable under the maximum likelihood framework but not under least squares. (A) *Lophocereus* gene tree (one of >52,000 equally most parsimonious trees; $L = 148$, CI = 0.83, RI = 0.96) from the reanalysis of Hartmann et al. (2001). All testable branches had a $P > 0.05$ generally supporting pseudogene null hypothesis. An asterisk denotes nodes with 100% strict consensus bootstrap support and the *functional* and *truncated* clades reflect those defined by Hartmann et al. (2001). (B) Brassicaceae gene tree (one of four equally most parsimonious trees; $L = 660$; CI = 0.58, RI = 0.76). "¹" denotes sequences added to the Yang et al. (1999) matrix and bold accessions identify putative pseudogenes. (See text. Asterisks denote nodes supported by at least 70% strict consensus bootstrap support.) Long branch attraction is implicated in the resolution of node 41 (see text).

this paper and is based on the expectation that in a pseudogene the rate of evolution in the 5.8S region will equal the rate in the ITS 1 and ITS 2 regions, all other things being equal, while in a functional nrDNA locus the rate of evolution in the 5.8S region will typically be less than that in the ITS 1 and ITS 2 regions. We conduct this test in the hierarchical framework of a gene tree by comparing substitution rates (K's, branch lengths) of the 5.8S and ITS regions along each branch (given some restrictions, see below). That is, if we fail to reject the null hypothesis that $K_{\text{ITS}} = K_{5.8\text{S}}$ for any branch, we conclude that the sequences involved in that branch show a "pseudogene signature" and are therefore putative pseudogenes. This is a sensitive test in the sense that it can potentially detect recently evolved pseudogenes for which only the most derived branches of a tree would show an increase in the rate of 5.8S evolution. Furthermore, the assumption that pseudogenes do not revert to functional loci entails the prediction that the descendant branches of any branch with the pseudogene signature should also show the pseudogene signature. Examining descendant branches acts as a test of the conclusion that sequences contributing to an ancestral branch are pseudogenes since phenomena like long branch attraction, poor taxon sampling, and positive natural selection on the 5.8S region could "artificially inflate" $K_{5.8\text{S}}$ but are less likely to have effects from ancestral to descendant branches.

Our test procedure involves the following steps (described in more detail below). (1) Split sequences into separate ITS and 5.8S matrices. (2) On a particular tree (e.g., a most parsimonious tree, a maximum likelihood tree), estimate the branch lengths based on the ITS data and the 5.8S data separately. (3) Bootstrap both the ITS matrix and the 5.8S matrix. (4) Use the data sets resulting from bootstrapping to estimate the ITS lengths and the 5.8S lengths for each branch on the tree used in step 2. For each branch there is now an estimate of branch length for each region (from step 2) and a distribution around that estimate (from step 4). The distribution can be used to construct confidence intervals and to conduct hypothesis tests. (5) Use confidence intervals to determine which branches have sufficient length to be testable. (6) For each testable branch, use the bootstrap distributions of ITS length and 5.8S length generated in step 4 to test the null hypothesis that $K_{\text{ITS}} = K_{5.8\text{S}}$.

To obtain estimates of the number of substitutions/site ($\widehat{K}_{\text{ITS}}$ and $\widehat{K}_{5.8\text{S}}$, $\widehat{\ } =$ 'estimated') for each branch on a gene tree, sequences were split into their ITS (ITS 1 & ITS 2 combined if both available) and 5.8S regions with each region analyzed separately. Pairwise genetic distances (among ITS sequences and 5.8S sequences) were estimated using the maximum likelihood model (see Kishino and Hasegawa, 1989) with default Tr/Tv = 2.0 settings implemented in the DNADIST program of PHYLIP (Felsenstein, 1993). Simulation studies show negligible differences among estimates of pairwise K (and hence among estimates of branch length) by many models when K is $\leqslant 0.5$ (Li, 1997). Our estimates of pairwise K for both data sets fall well within this range as only three values were greater than 0.5 (greatest was 0.5534). Also, our own comparisons based on other models (and other reasonable Tr/Tv ratios including equal Tr/Tv) available in PHYLIP support the conclusion that the choice of model has a negligible effect on estimates of K when sequences are relatively closely related (T. Carr, unpublished results). The FITCH program of PHYLIP was used to generate least squares estimates of branch length for each branch on one of the most parsimonious trees ("user tree" option in effect). This method provides estimates of the rate of substitution per site in the ITS region and the 5.8S region for each branch. In all analyses, no negative branch lengths were allowed as these are without biological meaning. This lower bound on branch length can greatly influence the shapes of the distributions of branch length used to test the pseudogene null hypothesis and may necessitate a fully non-parametric approach (see Dopazo, 1994).

Because small branch lengths result in tests with low power (leading to potentially false conclusions of the presence of pseudogenes) and because $\widehat{K}_{\text{ITS}} = \widehat{K}_{5.8\text{S}} = 0$ reflects the absence of data with which to test the pseudogene hypothesis, we chose an a priori criterion for which branches were testable. A branch was considered testable if the 95% confidence interval (CI) of either the ITS branch length or the 5.8S branch length did not *effectively* include 0. If the lower bound of the 95% CI around a branch length ($\widehat{K}$) multiplied by the number of base pairs unambiguously optimizing (by parsimony—unambiguous optimization in WinClada) to that branch indicated less than 1 substitution along that branch, the 95% CI was considered to effectively include 0. If the 95% CI's around *both* the ITS and 5.8S estimates of branch length effectively included 0 for a branch, that branch was excluded from further testing.

Ninety-five percent CI's were estimated using the percentile bootstrap method and the generally more accurate percentile-*t* bootstrap method (Chernick, 1999; Mooney and Duval, 1993). For the percentile method, we took 1000 bootstrap replicates of each original ITS and 5.8S data set, generated the branch lengths for each replicate on the single tree being investigated (using the methods described above), and plotted the distribution of the 1000 branch lengths for each branch (Dopazo, 1994). The bounds of the 95% CI's were the upper and lower 0.025 percentage points of each distribution.

In the percentile-*t* method, each bootstrap estimate of branch length (denoted $\widehat{K}^*$, * = bootstrap) is standardized by finding $t^* = (\widehat{K}^* - \widehat{K})/\sigma_{\widehat{K}^*}$. The distribution of t^*

is then used to calculate the critical points of the CI by $[\widehat{K} - t^*_{(1-\alpha/2)} \cdot \sigma_{\widehat{K}}, \widehat{K} - t^*(\alpha/2) \cdot \sigma_{\widehat{K}}]$ in a manner analogous to parametric CI estimation using student's *t* (Chernick, 1999; Mooney and Duval, 1993). Estimating the $\sigma_{\widehat{K}}$ for the ITS region and for 5.8S region is obtained through the initial bootstrap resampling. The method is more computationally intensive than the percentile method because estimating $\sigma_{\widehat{K}^*}$ (the standard deviation of each bootstrap estimate of $\widehat{K}$) typically requires bootstrapping the bootstrap. From each of the 1000 bootstrap samples originally generated above, we took 50 additional bootstrap samples, used each of these to estimate branch lengths using the methods already described, and calculated $\sigma_{\widehat{K}^*}$ for each branch for each of the 1000 original bootstraps. Because this method is computationally intensive, we only estimated percentile-*t* CI's for those branches whose 90% CI's in the percentile method did not effectively include 0. This suite of branches included 4 (1 from *Lophocereus* and 3 from Brassicaceae) whose 95% CI's effectively included 0 under the percentile method. These branches are important indicators of potential differences between the percentile and percentile-*t* methods in this application. The 90% CI criterion did not reduce the total amount of bootstrapping performed, but greatly reduced the amount of data manipulation required subsequent to bootstrapping.

For the testable branches, we used non-parametric bootstrap hypothesis testing (Hall and Wilson, 1991; Manly, 1997) to test the null hypothesis that $\Delta\widehat{K} = \widehat{K}_{\text{ITS}} - \widehat{K}_{5.8\text{S}} = 0$. This is a procedure akin to bootstrap CI estimation but involving evaluation of a test statistic against a null distribution that is generated in part by bootstrap resampling. Bootstrapping provides a means of estimating the variation around $\widehat{K}$ for a branch without the additional assumptions of other methods (e.g., a Poisson distribution of branch lengths across the tree; Dopazo, 1994). As in estimating CI's around branch lengths by bootstrapping, the test is applied to a particular tree held constant throughout.

The test procedure involves the following steps (* = estimate deriving from bootstrap resampling) (after Timmer et al., 1999):

(1) For each testable branch, estimate $\widehat{K}_{\text{ITS}}$ and $\widehat{K}_{5.8\text{S}}$ from the original data as described above.

(2) For each of these branches calculate $\Delta\widehat{K} = \widehat{K}_{\text{ITS}} - \widehat{K}_{5.8\text{S}}$.

(3) Use bootstrap resampling to generate 1000 ITS and 1000 5.8S data sets. Each ITS data set is randomly paired with a 5.8S data set, although it is essential that the bootstrap resampling be done independently on each region.

(4) For each of the 1000 pairs of data sets, estimate $\widehat{K}^*_{\text{ITS}}$ and $\widehat{K}^*_{5.8\text{S}}$ for each branch and calculate $\Delta\widehat{K}^* = \widehat{K}^*_{\text{ITS}} - \widehat{K}^*_{5.8\text{S}}$ for each branch.

(5) For each testable branch, obtain the distribution (based on the 1000 resamples) of $\Delta\widehat{K}^* - \Delta\widehat{K}$. The distribution of $\Delta\widehat{K}^* - \Delta\widehat{K}$ is an estimate of the distribution of differences in branch length if the null hypothesis is true (Hall and Wilson, 1991).

(6) For each testable branch, compare $\Delta\widehat{K}$ with the null distribution from 5 and reject the null hypothesis of equal branch lengths for ITS and 5.8S if $\Delta\widehat{K}$ falls outside the upper or lower $\alpha/2$ boundary of the distribution (two-tailed test). Because we were most concerned with making a Type II error (accepting a false null hypothesis—i.e., incorrectly concluding "pseudogene"), we set α equal to 0.10 to increase the power of the test. Additionally, we made no correction to maintain a set tree-wise α.

We also conducted the hypothesis test using pivoted statistics. Test statistics are pivoted when they are divided by an estimate of scale. In this case, we obtained the distribution of $\Delta\widehat{K}^* - \Delta\widehat{K}/\sqrt{(\sigma^2_{\widehat{K}^*_{\text{ITS}}} + \sigma^2_{\widehat{K}^*_{5.8\text{S}}})}$ and examined the test statistic $\Delta\widehat{K}/\sqrt{(\sigma^2_{\widehat{K}_{\text{ITS}}} + \sigma^2_{\widehat{K}_{5.8\text{S}}})}$ in relation to this distribution. Each estimate of $\sigma^2_{\widehat{K}_{\text{ITS}}}$ and $\sigma^2_{\widehat{K}_{5.8\text{S}}}$ had already been obtained through the initial bootstrap resampling, and each estimate of $\sigma^2_{\widehat{K}^*_{\text{ITS}}}$ and $\sigma^2_{\widehat{K}^*_{5.8\text{S}}}$ had already been obtained by bootstrapping this initial bootstrap (as explained above). Pivoting should generally make small improvements in the accuracy of a test and so becomes important when resulting *P* values are close to α (Hall and Wilson, 1991). Comparing the quantitative and qualitative results from pivoted and non-pivoted statistics provides important empirical information on whether the extra computational effort involved in pivoting is worthwhile in this application.

10. Results and discussion

10.1. *Lophocereus*

The tree-based reanalysis of the available *Lophocereus* ITS 1 and 5.8S data (Hartmann et al., 2001), failed to reject the pseudogene null hypothesis for the branches subtending both major clades of nrDNA sequences (branches 43, 51, and 52 in Fig. 2A and Table 1) as well as for the one outgroup (*Pachycereus*) branch tested (branch 1 in Fig. 2A and Table 1). As branches 43 and 51 are descendants of branch 52, they act as tests of the prediction that the descendants of a branch showing the pseudogene signature should also show the pseudogene signature if the sequences involved are from pseudogenes. Only one other branch (branch 34) was long enough to be tested, and it represents an interesting case of significantly greater rate of evolution in the 5.8S region

Table 1
Statistical analysis for the presence of pseudogenes in sequences from *Lophocereus*

Branch	Branch length estimate[a]	$\widehat{K}_{\mathrm{ITS}}$	$\widehat{K}_{5.8\mathrm{S}}$	$\widehat{K}_{\mathrm{ITS}}/\widehat{K}_{5.8\mathrm{S}}$	$P^{\mathrm{b,c}}$	Pseudogene?
1	LS	0.05854	0.02688	2.18	>0.09	Y
	ML	0.05705	0.02497	2.28	>0.13	
34	LS	0.00293	0.03456	0.085	>0.97	?[d]
	ML	0.00652	0.03830	0.17	>0.96	
43	LS	0.13815	0.07623	1.81	>0.065	Y
	ML	0.14701	0.08625	1.70	>0.10	
51	LS	0.11122	0.07307	1.52	>0.16	Y
	ML	0.10737	0.8130	1.32	>0.27	
52	LS	0.15101	0.09034	1.67	>0.07	Y
	ML	0.16604	0.08266	2.01	>0.055	

Note. Branch lengths estimated by both the least squares (LS) and maximum likelihood (ML) methods. The branches included qualify as "testable" since the 95% CI (assessed using the percentile method [for ML] or the percentile-*t* method [for LS] based on 1000 initial bootstrap replicates) for either the ITS branch length or the 5.8S branch length does not effectively include 0. Two-tailed test used with $\alpha = 0.1$. Results shown for least squares are those based on pivoting the null distribution and test statistic for each branch. Results shown for maximum likelihood are those based on the percentile method. *P* values <0.05 reflect a $\widehat{K}_{\mathrm{ITS}}$ significantly greater than $\widehat{K}_{5.8\mathrm{S}}$, *P* values >0.95 reflect a $\widehat{K}_{\mathrm{ITS}}$ significantly less than $\widehat{K}_{5.8\mathrm{S}}$.

[a] LS, least squares and ML, maximum likelihood.

[b] *P*-value given is technically $1 - P$. This simply aids interpretation and does not change the results.

[c] Given 1000 bootstrap replicates, smallest *P*-value possible to assess is <0.001.

[d] In the case of node 34, $K_{5.8\mathrm{S}}$ is significantly greater than K_{ITS}. If the sequences involved throughout the *Lophocereus* group were not pseudogenes, this might be interpreted as a case of positive natural selection on the 5.8S region. In this particular case the observed pattern could be explained by partial concerted evolution between 5.8S regions in which many 5.8S substitutions accrued.

than in the ITS 1 region. As this is a descendant branch of nodes 52 and 51, it falsifies the prediction that descendant branches of putative pseudogenes should also show the pseudogene signature. While this result for branch 34 might be explained by positive selection on the 5.8S region of a putatively functional copy, this hypothesis is difficult to accept given the pseudogene pattern observed at multiple ancestral nodes. An alternative hypothesis is that a partial concerted evolution event between the 5.8S region along this branch with another copy type (functional or non-functional) introduced several substitutions. While this hypothesis is admittedly ad hoc, partial concerted evolution among copy types in nrDNA is known.

Although not tested explicitly, consideration of conglomerated data available within the large clades (i.e., those branches too short to test individually) also supported the acceptance of the pseudogene null hypothesis for both major clades. That is, the numbers of variable sites in the 5.8S and ITS region are what one would expect if the sequences were pseudogenes. Together these results strongly suggest that Hartmann et al. (2001) did not include any functional sequences in their *Lophocereus* ingroup (nor in the outgroup).

In contrast to the conclusions drawn from our results, Hartmann et al. (2001) concluded that one clade in the nrDNA gene tree represented a functional lineage ("*functional*") and the other a pseudogene lineage ("*truncated*"). Their sequence functionality hypotheses were based primarily on three factors. First, *truncated* sequences had a 150 bp deletion in ITS 2 that was absent in *functional* types. Second, a transcription assay of one individual recovered a *functional* sequence type. Finally, a series of PCR reactions targeted at amplifying a full-length *truncated* type sequence failed. However, none of these factors particularly support the conclusions of functionality (also see above discussion about length). Although the 150 bp indel difference between *functional* and *truncated* suggests the presence of pseudogenes, without reference to patterns of change on the resulting trees, full lengths of ITS 2 in both types, or at least estimates of secondary structure for the region in question (ITS 2), it is not possible to distinguish a deletion in *truncated* from an insertion in *functional*. Failing to make this distinction leaves open the question of whether *truncated* are too short or *functional* are too long. Accounting for the overall lengths of ITS 1 and ITS 2 relative to known functional classes (Baldwin et al., 1995) actually suggests the sequence lengths for the *functional* class are outside the range of known functional types. If anything, ITS 1 and ITS 2 of *truncated* are more consistent with functional lengths than the *functional* clade sequences. In addition, failure to obtain a particular PCR product is not a reliable method for detecting pseudogenes, particularly when the same tests were not applied to the *functional* class of sequences. Finally, as we have already noted, expression need not relate to patterns of sequence evolution.

As an additional check of our conclusions, we calculated free energies (ΔG values) for the ITS 1 regions of *functional* and *truncated* sequences following the methods of Buckler et al. (1997). There was no appreciable difference in free energies between these two "classes" of ITS 1 sequences (average ΔG values for *truncated* and *functional* were -55.31 ± 6.9 and -54.8 ± 5.8, respectively). The essentially identical ΔG for the two groups is not consistent with one group being functional and the other non-functional.

In the *Lophocereus* example, pseudogene substitution patterns are evident for both *truncated* and *functional* lineages and provide strong evidence that both of these are composed entirely of pseudogenes. This indicates that the gene tree may be severely under-sampled, given the lack of functional sequence types. It appears that pseudogenes were preferentially amplified in this study. Importantly, "*functional*" sequence types may represent the first documented example of expressed non-functional nrDNA sequences.

10.2. Brassicaceae

The Brassicaceae ITS matrix (Yang et al., 1999) was primarily re-examined to compare the information provided by simple pairwise approaches of pseudogene detection to the tree-based approach advocated here. Among the sequences in the original Yang et al. (1999) matrix, five were identified as potential pseudogene sequences based on our implementation of Yang et al.'s (1999) pairwise comparison of a consensus 5.8S region examining simply the number of nucleotide differences in comparison to a consensus sequence (*Brassica juncea* AF128094, *B. rapa* ssp. *chinesis* AF128098, *B. oleracea* var. *botrytis* AF128100, *B. oleracea* var. *alboglabra* AF128101, and *Lepidium sativum* X78126). While the 5.8S region of the sequence from *L. sativum* differed from a consensus sequence by the same number of nucleotides as the sequence from *B. rapa* ssp. *chinensis*, Yang et al. (1999) did not classify it as a pseudogene because the sequence came from GenBank, making them unsure of its reliability. Applying Yang et al.'s (1999) pairwise approach to the two new ingroup sequences suggested that both *A. jacquinii* and *B. maximowiczii* were represented by putative pseudogenes.

Yang et al. (1999) also conducted relative rate tests and found significant increases in rate of nucleotide substitution (for entire sequences) in *B. juncea* AF128094, *B. rapa* ssp. *chinesis* AF128098, *B. oleracea* var. *botrytis* AF128100, *B. oleracea* var. *alboglabra* AF128101, and *Rorippa nasturtium-aquaticum* X98644, considering all but the latter pseudogenes (again, because the sequence from *R. nasturtium-aquaticum* was obtained from GenBank). We implemented the same relative rate test (Wu and Li, 1985) on the sequences of *B. maximowiczii* and *A. jacquinii* using the computer program K2WuLi v. 1.0 (Jermiin, 1996). *C. lutea* AF137588 was used as the reference sequence and both sequences were tested against six other sequences (not shown to be pseudogenes in our tree-based test). For both sequences, some comparisons suggested that they were pseudogenes, while others did not. However, the bulk of the evidence from the relative rate tests suggests that the sequence from *A. jacquinii* is not a pseudogene (5 of 6 comparisons) and that the sequence from *B. maximowiczii* is a pseudogene (4 of 6 comparisons).

The tree-based reanalysis of the complete Brassicaceae data set strongly supported the conclusion of the simple pairwise approach and the relative rate test that *B. rapa* ssp. *chinesis* AF128098, *B. juncea* AF128094, *B. oleracea* var. *botrytis* AF128100, *B. oleracea* var. *alboglabra* AF128101, and *L. sativum* X78126 are represented by sequences from pseudogenes (branches 5, 6, 7, 9, 23 respectively, Fig. 2B, Table 2). The tree-based approach strongly rejected the conclusion from the relative rate test that the sequence from *R. nasturtium-aquaticum* (X98644) is a pseudogene (branches 21, 32, and 40, Fig. 2B, Table 2). In addition, the tree-based approach identified the sequence from *A. lyrata* ssp. *petrea* U96270 (branch 29, Fig. 2B, Table 2) as a putative pseudogene, although in this case we are concerned that the test fails to reject the pseudogene null hypothesis due to low statistical power (see below).

In contrast to the simple pairwise approach and the relative rate test, however, the tree-based approach rejected the conclusion that the sequences from either *A. jacquinii* or *B. maximowiczii* are pseudogenes, but did so in an interesting fashion. The branch uniting *A. jacquinii* and *B. maximowiczii* (branch 41) showed the pseudogene signature, but the pseudogene null hypothesis was strongly rejected on the branch's two descendant branches (branches 16 and 17, the terminal branches for *A. jacquinii* and *B. maximowiczii*), calling into question the conclusion that branch 41 results from the inclusion of pseudogene sequences. The branch lengths found in the tree-based analysis suggested that long branch attraction was the cause of this pattern. Branch 41 uniting *A. jacquinii* and *B. maximowiczii* had only 10 unambiguously optimized substitutions while the terminal branches (branches 16 and 17) had 19 and 52 substitutions. Long branch attraction was further investigated by the sequential removal of each of these two individual terminals and reanalysis. The removal of *B. maximowiczii* had no influence on the overall topology and *A. jacquinii* alone did not show the substitution pattern indicative of a pseudogene ($\widehat{K}_{\mathrm{ITS}} = 0.08932$, $\widehat{K}_{5.8\mathrm{S}} = 0.02483$, $P < 0.003$). The removal of *A. jacquinii* caused *B. maximowiczii* to be resolved as basal to all other Brassicaceae and in this position it also did not show a pseudogene signature ($\widehat{K}_{\mathrm{ITS}} = 0.16266$, $\widehat{K}_{5.8\mathrm{S}} = 0.01840$, $P < 0.001$). A phylogenetic analysis of more than 200 Brassicaceae ITS sequences (available in GenBank) provided further evidence that long branch attraction likely caused the grouping of *A. jacquinii* and *B. maximowiczii* in our analysis (Bailey, unpublished data). The results of this

Table 2
Statistical analysis for the presence of pseudogenes in sequences from the Brassicaceae

Branch	Branch length estimate[a]	$\widehat{K}_{\mathrm{ITS}}$	$\widehat{K}_{5.8\mathrm{S}}$	$\widehat{K}_{\mathrm{ITS}}/\widehat{K}_{5.8\mathrm{S}}$	P[b,c]	Pseudogene?
5	LS	0.04385	0.01847	2.37	>0.055	Y[d]
	ML	0.04097	0.01897	2.16	>0.05	
6	LS	0.02765	0.04536	0.61	>0.85	Y
	ML	0.03165	0.04487	0.71	>0.75	
7	LS	0.06409	0.04382	1.46	>0.15	Y
	ML	0.06322	0.04392	1.44	>0.15	
8	LS	0.02141	0.00000	—	<0.001	
	ML	0.02157	0.00000	—	<0.005	
9	LS	0.10878	0.09070	1.20	>0.25	Y
	ML	0.10448	0.09178	1.14	>0.30	
16	LS	0.08592	0.00665	12.92	<0.001	
	ML	0.06962	0.00620	11.23	<0.005	
17	LS	0.14828	0.00555	26.72	<0.008	
	ML	0.16143	0.00603	26.77	<0.001	
18	LS	0.03031	0.00597	5.08	<0.006	
	ML	0.02884	0.00609	4.74	<0.01	
19	LS	0.03641	0.00000	—	<0.001	
	ML	0.03901	0.00000	—	<0.001	
20	LS	0.04029	0.00611	6.59	<0.001	
	ML	0.04216	0.00608	6.93	<0.001	
21	LS	0.07919	0.00608	13.02	<0.001	
	ML	0.07825	0.00609	12.85	<0.001	
22	LS	0.02429	0.00000	—	<0.002	
	ML	0.02998	0.00000	—	<0.005	
23	LS	0.03831	0.02541	1.51	>0.20	Y
	ML	0.02981	0.02508	1.19	>0.40	
26	LS	0.03722	0.00000	—	<0.001	
	ML	0.03958	0.00000	—	<0.001	
29	LS	0.01673	0.00603	2.77	>0.12	Y?[e]
	ML	0.01691	0.00609	2.78	>0.08	
31	LS	—	—	—	—	
	ML	0.01198	0.00000	—	<0.05	
32	LS	0.03996	0.00000	—	<0.001	
	ML	0.05235	0.00000	—	<0.001	
33	LS	0.08316	0.00612	13.59	<0.001	
	ML	0.08526	0.00609	14.00	<0.001	
34	LS	0.06254	0.00000	—	<0.001	
	ML	0.07547	0.00000	—	<0.001	
36	LS	0.02596	0.00000	—	<0.001	
	ML	0.02221	0.00000	—	<0.01	
37	LS	0.04248	0.00000	—	<0.001	
	ML	0.03199	0.00000	—	<0.001	
38	LS	—	—	—	—	
	ML	0.02937	0.00000	—	<0.005	
39	LS	0.02129	0.00000	—	<0.002	
	ML	0.02165	0.00000	—	<0.025	
40	LS	0.03538	0.00000	—	<0.001	
	ML	0.03309	0.00000	—	<0.005	

Table 2 (*continued*)

Branch	Branch length estimate[a]	$\widehat{K}_{\mathrm{ITS}}$	$\widehat{K}_{5.8\mathrm{S}}$	$\widehat{K}_{\mathrm{ITS}}/\widehat{K}_{5.8\mathrm{S}}$	P[b,c]	Pseudogene?
41	LS	—	—	—	—	
	ML	0.02149	0.01858	1.16	>0.40	N[f]
42	LS	0.04543	0.00000	—	<0.001	
	ML	0.04876	0.00000	—	<0.001	
43	LS	0.02950	0.00000	—	<0.001	
	ML	0.03751	0.00000	—	<0.001	
44	LS	0.01935	0.00000	—	<0.001	
	ML	0.01610	0.00000	—	<0.015	
45	LS	0.01251	0.00000	—	<0.003	
	ML	0.01332	0.00000	—	<0.05	
48	LS	0.01070	0.00000	—	<0.001	
	ML	—	—	—	—	
49	LS	0.01416	0.00000	—	<0.002	
	ML	0.01568	0.00000	—	<0.002	
50	LS	—	—	—	—	
	ML	0.01516	0.00000	—	<0.02	
52	LS	—	—	—	—	
	ML	0.01371	0.00000	—	<0.025	
53	LS	0.04704	0.00000	—	<0.001	
	ML	0.05398	0.00000	—	<0.001	

Note: Branch lengths estimated by both the least squares (LS) and maximum likelihood (ML) methods. The branches included qualify as "testable" since the 95% CI (assessed using the percentile method [for ML] or the percentile-*t* method [for LS] based on 1000 initial bootstrap replicates) for either the ITS branch length or the 5.8S branch length does not effectively include 0. Two-tailed test used with $\alpha = 0.1$. Results shown for least squares are those based on pivoting the null distribution and test statistic for each branch. Results shown for maximum likelihood are those based on the percentile method. P values <0.05 reflect a $\widehat{K}_{\mathrm{ITS}}$ significantly greater than $\widehat{K}_{5.8\mathrm{S}}$, P values >0.95 reflect a $\widehat{K}_{\mathrm{ITS}}$ significantly less than $\widehat{K}_{5.8\mathrm{S}}$.

[a] LS, least squares and ML, maximum likelihood.

[b] P-value given is technically $1 - P$. This simply aids interpretation and does not change the results.

[c] Given 1000 bootstrap replicates, smallest P-value possible to assess is <0.001.

[d] The result for branch 5 without pivoting the LS estimates found $\widehat{K}_{\mathrm{ITS}}$ significantly greater than $\widehat{K}_{5.8\mathrm{S}}$, rejecting the pseudogene null hypothesis ($P < 0.04$). This was the only branch for which pivoting made a qualitative difference to the conclusion.

[e] Although the pseudogene null hypothesis cannot be falsified for branch 29, KITS is ~2.8 times greater than K5.8S. We suspect that this test has low power resulting from a small number of substitutions along this branch and the fact that one of them, perhaps by chance, occurs in the 5.8S region.

[f] Branch 41 results from long branch attraction. See Results.

analysis demonstrate that our intentional addition of two sequences resulted in an under sampled gene tree. Their inclusion misled inference based on the pairwise comparison method of detecting pseudogenes but ultimately did not mislead inference from the tree-based approach. In addition, the use of the tree-based method provided evidence for long branch attraction, an issue of concern to molecular systematists.

Although Yang et al. (1999) were correct in stating that the overall topology of their tree was not altered by excluding the pseudogenes (i.e., the branching order of the functional types remains the same when pseudogenes are excluded), the inclusion of the pseudogenes does identify two potential gene tree/species tree conflicts. First, *B. oleracea* var. *alboglabra* is only represented by a pseudogene sequence. The missing functional type is another clear example of gene tree under-sampling. Second, the functional and pseudogene sequences from *B. rapa* ssp. *chinensis* are polyphyletic on the gene tree (on all four parsimony trees), which suggests underlying deep paralogy, hybridization, or lineage sorting problems. Ultimately, the exclusion of these two pseudogene sequences from the original matrix is unjustified. This level of matrix pruning results in a potentially problematic matrix appearing unproblematic.

10.3. Bootstrap CI estimation and hypothesis testing

In general, the substantial extra computational effort required to calculate bootstrap percentile-*t* CI's and pivot the bootstrap-based null distribution and test statistic made little difference to the quantitative and qualitative results. For these two data sets, computationally less intensive percentile methods appear to be adequate for the applications developed here.

In comparison to the percentile CI method, the percentile-*t* CI method most often increased both the upper and lower 95% CI bounds on branch length (in 29 out of 34 cases), and generally resulted in wider 95% CI's (in 30 out of 34 cases). Quantitatively, these changes in the width of 95% CI's when the percentile-*t* method was used ranged from −29% to +35%. However, both methods identified the same set of testable branches (95% CI of either ITS or 5.8S branch length not effectively including 0), so these quantitative differences made no qualitative differences to our conclusions. Because it typically produces CI's with slightly greater lower and upper bounds than the percentile method, the percentile-*t* method may appear to be more desirable in determining testable nodes, especially when branch lengths are small. However, hypothesis tests involving such branches will have very low power (see below) and should be undertaken with caution. Moreover, four additional branches for which the 90% percentile CI's did not include 0 were all rejected for testing based on both percentile and percentile-*t* 95% CI's. This result further suggests that the extra accuracy provided by the percentile-*t* method has little qualitative effect for the applications developed here.

Across the two data sets there was only one qualitative change to the results when the null distribution and test statistic were pivoted. In the Brassicaceae data matrix, branch 5 involving only *B. rapa* ssp. *chinensis* AF128098 shows the pseudogene signature when pivoting is applied ($P \sim 0.06$) but otherwise does not ($P \sim 0.03$; under a two-tailed test with $\alpha = 0.05$ instead of 0.10, the conclusion from both methods would be that the branch shows the pseudogene signature). Prior to pivoting, the results for this branch were the most uncertain (*P*-value closest to α), and it is in such cases that pivoting is expected to make a difference (Hall and Wilson, 1991). Quantitatively, changes in *P*-values of the test statistic due to pivoting ranged from −0.017 to +0.049 in comparison to the non-pivoted cases. While having relatively little qualitative effect on our results, a change of ~0.05 (5%) is not trivial, especially near α. Still, changes to *P*-values due to pivoting which shift results from rejecting a null hypothesis to failing to reject a null hypothesis should be treated cautiously when tests are suspected of having low statistical power. In testing for pseudogenes, power will generally be of greater concern than changes to *P*-values due to pivoting (see below).

Furthermore, it is important to bear in mind that our null hypothesis, equal rates of evolution in the ITS and 5.8S regions in a pseudogene, assumes that "all other things are equal" between the regions. If factors known to influence the neutral mutation rate (e.g., base composition) differ between the two regions, then another null hypothesis, such as $K_{\mathrm{ITS}} = 1.5K_{5.8\mathrm{S}}$, might be more appropriate. The existence of such differences between the regions can be assessed beforehand and their influence on the rate of evolution accounted for by refining the null hypothesis. Indeed, the null hypothesis could even be refined for particular clades or branches, although changes in the factors influencing the neutral mutation rate would probably have to be large from clade to clade or branch to branch to make such refinements worthwhile.

10.4. Statistical power

In the test for pseudogenes developed in this paper, the conclusion that a sequence is from a pseudogene rests on failing to reject a null hypothesis. This raises concerns about Type II errors (failing to reject a false null hypothesis) and the power of the particular statistical test used to examine the null hypothesis. While it would be possible to use simulation studies applying differing rates of evolution to the ITS and 5.8S regions (e.g., 1:1, 2:1, 3:1) to generate power functions (power vs. effect size for a particular number of substitutions) for the bootstrap-based test, such an undertaking is beyond the scope of this paper. However, we can point out that the bootstrap-based test developed here outperforms χ^2 tests examining estimated and expected (under the pseudogene null) numbers of substitutions in the 5.8S and ITS regions along a branch (T. Carr unpublished data). That is, under the same test conditions the bootstrap-based test rejects the pseudogene null hypothesis when the χ^2 test accepts it, especially at low branch lengths. χ^2 tests have notoriously low power at small effect sizes and low *N*, and we interpret the difference in rate of rejection as representing greater power in the bootstrap-based test.

However, the bootstrap-based test can still suffer from low power when short branch lengths are involved. For example, for branch 29 in the Brassicaceae matrix, the rate of evolution in the ITS region is ~2.8 times as high as that in the 5.8S region, but the test does not reject the pseudogene null (Table 2). Unfortunately, this is a terminal branch, and so we cannot check the conclusion by testing descendant branches. The failure to reject the null hypothesis in this case likely represents a Type II error due to low power resulting from a small number of substitutions along this branch and the fact that one substitution, perhaps by chance, occurs in the 5.8S region. Generally, the pseudogene null hypothesis was not rejected when $\widehat{K}_{\mathrm{ITS}}/\widehat{K}_{5.8\mathrm{S}}$ ratios were 2.3 or less (Tables 1 and 2), indicating that for the matrices and trees investigated here, the test was not powerful enough to distinguish 2:1 from 1:1.

While the most satisfactory strategy for increasing power is to obtain additional sequence data from other typically constrained and unconstrained regions, there are other possible methods. These include: (1) setting a higher α (e.g., $\alpha = 0.10$ here, although this had almost no effect on the qualitative conclusions from the data sets examined); (2) not correcting the α-level per test to maintain a particular tree-wise α; (3) making the test one-tailed (although this latter strategy makes it impossible to detect cases where $K_{5.8S} > K_{ITS}$ [e.g., branch 34, Table 1]); (4) limiting testing to branches with some minimum number of estimated substitutions (perhaps 8–10) in either 5.8S or ITS regions and/or increasing the CI that cannot effectively include 0 for a branch to be testable (e.g., from 95 to 99%); (5) removing branches to create longer branches (corrections to the α-level might be necessary when this strategy produces multiple non-independent hypothesis tests); and (6) combining branches. This latter strategy seems particularly promising as bootstrap estimates of K_{ITS} and $K_{5.8S}$ from each branch at issue could be combined until some minimum K was met and then used to form bootstrap distributions of K_{ITS} and $K_{5.8S}$ across the branches. One could then test the null hypothesis that, for example, $(K_{ITS,Branch\ 1} + K_{ITS,B2} + K_{ITS,B3}) - (K_{5.8S,Branch\ 1} + K_{5.8S,B2} + K_{5.8S,B3}) = 0$. There are a number of possible ways to combine branches, but in general complete sets of ancestor-descendant branches should be maintained. Increasing the number of bootstrap replicates might also be advisable when branches are combined. The more branches that are removed or combined, however, the more hierarchical information is lost and the closer the method becomes to a pairwise approach.

10.5. Branch length estimation and sensitivity to changes in tree topology

Because our test could be sensitive to different methods of estimating branch lengths, we examined this possibility by comparing results from least squares estimates of branch length to those from direct maximum likelihood estimation (using DNAML in PHYLIP with $Tr/Tv = 2.0$) for the same tree. While many of the estimates of branch length were similar, there were a few branches where maximum likelihood estimates differed from least squares estimates by up to ±0.015 substitutions/site. Under the percentile method for CI estimation, both methods of estimating branch length identified almost the same sets of testable branches (Tables 1 and 2). The differences mostly had to do with which short branches were identified as testable, and because of concerns about low statistical power, the results from these branches should be interpreted with caution anyway. Interestingly, the branch involved in long branch attraction (branch 41) was testable under direct maximum likelihood estimation but not under least squares. This difference has to do with the fact that the branch has a wide 95% CI, perhaps in part because it is erroneous, and that the maximum likelihood estimates of branch length were more than 0.010 substitutions/site greater than least squares estimates for both ITS and 5.8S, perhaps because the least squares method can spread change out across a tree.

Despite differences in the estimates of branch length, the ratios of $K_{ITS}/K_{5.8S}$ were similar for each method and the qualitative results from the hypothesis tests are almost exactly the same when the percentile method is used (branch 5 of the Brassicaceae discussed above is the one exception). Quantitatively, the likelihood estimates (and the tests based around those estimates) tended to provide more support for pseudogene branches while just as strongly rejecting the pseudogene null hypothesis for other branches (Tables 1 and 2).

Estimates of branch length can also change based on tree topology, so our test for detecting the presence of pseudogenes could be sensitive to changes in topology With the exception of the case of long branch attraction in the Brassicaceae tree, we found no affect of a range of changes in tree topology on the qualitative results. Changes in topology can also interact with the method used to estimate branch lengths, and some estimation methods will be more sensitive than others to changes in topology. However, because one usually needs to choose a fully resolved tree to estimate branch lengths, many of the issues surrounding choice of topology will involve branches of effectively 0 length (at least when split into ITS and 5.8S partitions). The arrangement of such branches should have little effect on the qualitative results, regardless of which method of estimating branch length is chosen.

11. Conclusions

The existence of intra-individual nrDNA polymorphism raises a series of important issues that need to be considered when reconstructing phylogenetic relationships, particularly among closely related species. Here we have focused on clarifying issues related to nrDNA paralogy and pseudogenes. Previous discussions of nrDNA evolution unjustifiably assumed that pseudogenes are interspecific deep paralogs of functional loci in studies that include more than one species (Buckler et al., 1997; Mayol and Rosselló, 2001). This assumption conflates sequence history and sequence function, and ultimately overlooks the possible complexity of nrDNA sequence evolution. Without the context of a phylogenetic hypothesis, it should not be assumed that functional and non-functional sequences included in phylogenetic analyses of taxa are necessarily paralogous to one another or that functional copies are necessarily orthologous.

While patterns of nrDNA expression are inappropriate for identifying nrDNA pseudogenes, nucleotide diversification patterns provide powerful evidence for determining if sequences are functionally constrained. The interpretation of these patterns is best accomplished using tree-based approaches rather than pairwise methods. Here we have used ITS region sequences to illustrate the use of nucleotide substitution patterns for inferring whether or not sequence change is consistent with functional constraint on nrDNA. Additional data from each entire nrDNA sequence and the subsequent comparison of other putatively conserved regions (18S and 26S) and relatively unconstrained regions (e.g., IGS) could provide characters for increased statistical support, which may be limited based on the relatively short ITS region sequences alone. Furthermore, this general approach for assessing functional constraint among nrDNA sequences is applicable to any sequence type in which highly conserved and more variable regions have been conclusively delimited and where codon information is neither observed nor applicable.

It is abundantly clear that the accurate identification of nrDNA pseudogenes, paralogy, and orthology, are dependent on extensive interspecific, intraspecific, and intra-individual sampling. The crucial importance of sampling is one reason that pseudogene sequences should not be ignored in phylogenetic analyses of taxa or genes. They can provide critical information on the adequacy of sampling included in a study, and they may also supply important data relating to DNA sequence diversification (paralogy, orthology, and nucleolar dominance) and interspecific hybridization. It is only through extensive sampling that reasonable hypotheses relating the factors that affect the evolution of particular nrDNA groups and their potential influence on gene tree and species tree interpretations may be established.

The general bootstrap-based statistical hypothesis test developed in this paper can be applied to trees and branch lengths estimated by any method. We found no qualitative differences in conclusions when least squares or maximum likelihood methods of estimating branch lengths were used. Nor, with the exception of a case of long branch attraction, did we find evidence that changes in tree topology had any qualitative effect on the conclusions, although our investigation of this question was not extensive. Furthermore, the substantial additional computational effort involved in bootstrapping the bootstrap (to estimate σ_{K^*} for each bootstrap resample) in order to apply the percentile-*t* method of CI estimation and to pivot the null distribution and test statistic did not substantially change the qualitative results in comparison to less computationally intensive percentile methods. However, the quantitative changes observed as a result of pivoting test statistics suggest that pivoting is valuable when percentile-based *P*-values are within ± 0.05 of $\alpha/2$ (or α if the test is one-tailed). While strategies for increasing the power of the test exist (see Section 10.4), collecting additional sequence data from other constrained and unconstrained regions is most likely to provide satisfactory resolution of uncertain results when the test is suspected of having low power. Given the apparent robustness of the test and the ease and speed of carrying out percentile bootstrap-based hypothesis tests (without pivoting) on modern computers, we urge researchers to employ this statistical tool. In fact, bootstrap-based hypothesis tests will be useful for a range of questions in molecular evolution and phylogenetics.

Finally, given the issues discussed here, along with those put forward by Álvarez and Wendel (2003), one could easily conclude that nrDNA has no place in phylogeny reconstruction of plant taxa. Nevertheless, nrDNA sequence data continue to be the most widely used nuclear encoded sequence in plant systematics. This is in part because nrDNA represents the only universally amplifiable nuclear encoded region currently available and because potential alternatives from low-copy number DNA sequences are often exceptionally difficult to work with as well as complicated by sequence paralogy (Bailey et al., in press; Sang, 2002). Undoubtedly the use of low-copy number alternatives to nrDNA will increase with greater availability of genomic sequence and EST data (Álvarez and Wendel, 2003), but until such information is obtained for a broader selection of taxa and a greater understanding of potential copy number issues is available, nrDNA is likely to continue to play a prominent role in phylogeny reconstruction of plants.

Acknowledgments

The authors thank Jeff Doyle, Helga Ochoterena, and Mark Simmons for providing comments prior to manuscript submission as well as Jonathan Wendel for sending a prepublication copy of Álvarez and Wendel (2003). We are grateful to reviewer Eric Roalson and an anonymous reviewer for helpful comments on the manuscript. We are also grateful to another anonymous reviewer for suggesting comparisons with likelihood estimates of branch length and examination of how changes in tree topology might affect the outcome of the statistical test. Support for this study was provided by funds from the Leverhulme Trust and the Royal Society.

References

Ainouche, M.L., Bayer, R.J., 1997. On the origins of the tetraploid *Bromus* species: insights from ITS sequences of nrDNA. Genome 40, 730–743.

Álvarez, I., Wendel, J.F., 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29, 417–434.

Arnheim, N., 1983. Concerted evolution in multigene families. In: Nei, M., Koehn, R. (Eds.), Evolution of Genes and Proteins. Sinauer, Sunderland, MA, pp. 38–61.

Avise, J.C., 1989. Gene-trees and organismal histories: a phylogenetic approach to population biology. Evolution 43, 1192–1208.

Bailey, C.D., Hughes, C.E., Harris, S.A., in press. Using RAPDs to identify DNA sequence loci for species level phylogeny reconstruction: an example from *Leucaena* (Fabaceae). Syst. Bot.

Baker, W.J., Hedderson, T.A., Dransfield, J., 2000. Molecular phylogenetics of subfamily Calamoideae (Palmae) based on nrDNA ITS and cpDNA *rps16* intron sequence data. Mol. Phylogenet. Evol. 14, 195–217.

Baldwin, B.G., 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the Compositae. Mol. Phylogenet. Evol. 1, 3–16.

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

Buckler, E.S.I., Holtsford, T.P., 1996. *Zea* ribosomal repeat evolution and substitution patterns. Mol. Biol. Evol. 13, 623–632.

Buckler, E.S.I., Ippolito, A., Holtsford, T.P., 1997. The evolution of ribosomal DNA: divergent paralogues and phylogenetic implications. Genetics 145, 821–832.

Campbell, C.S., Wojciechowski, M.F., Baldwin, B.G., Alice, L.A., Donoghue, M.J., 1997. Persistent nuclear ribosomal DNA sequence polymorphism in the *Amelanchier* agamic complex (Rosaceae). Mol. Biol. Evol. 14, 81–90.

Chernick, M.R., 1999. Bootstrap Methods: A Practitioner's Guide. John Wiley & Sons, Inc., New York, NY.

Choi, D., Yoon, S., Lee, E., Hwang, S., Yoon, B., Lee, J., 2001. The expression of pseudogene cyclin D2 mRNA in the human ovary may be a novel marker for decreased ovarian function associated with the aging process. J. Assist. Reprod. Gen. 18, 110–113.

Davis, J., Simmons, M.P., Stevenson, D.W., Wendel, J.F., 1998. Data decisiveness, data quality and incongruence in phylogenetic analysis: an example from the monocotyledons using mitochondrial *atpA* sequences. Syst. Biol. 47, 282–310.

Denduangboripant, J., Cronk, Q.C.B., 2000. High intraindividual variation in internal transcribed spacer sequences in *Aeschynanthus* (Gesneriaceae) implications for phylogenetics. Proc. R. Soc. Lond. B 267, 1407–1415.

Dopazo, J., 1994. Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. J. Mol. Evol. 38, 300–304.

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

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

Doyle, J.J., Doyle, J.L., Brown, A.H.D., 1990. Analysis of a polyploid complex in *Glycine* with chloroplast and nuclear DNA. Aust. Syst. Bot. 3, 125–136.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Felsenstein, J., 1978. Cases in which parsimony and compatibility methods will be positively mis-leading. Syst. Zool. 27, 401–410.

Felsenstein, J., 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, WA.

Fuertes-Aguilar, J., Rosello, J.A., Feliner, G.N., 1999. Nuclear ribosomal DNA (nrDNA) concerted evolution in natural and artificial hybrids of *Armeria* (Plumbaginaceae). Mol. Ecol. 8, 1341–1346.

Futuyma, D.J., 1998. Evolutionary Biology, third ed. Sinauer, Sunderland, MA.

Gardiner-Garden, M., Sved, J.A., Frommer, M., 1992. Methylation sites in angiosperm genes. J. Mol. Evol. 34, 219–230.

Gaut, B.S., Tredway, L.P., Kubik, C., Gaut, R.L., Meyer, W., 2000. Phylogenetic relationships and genetic diversity among members of the Festuca-Lolium complex (Poaceae) based on ITS sequence data. Plant Syst. Evol. 224, 33–53.

Gernandt, D.S., Liston, A., 1999. Internal transcribed spacer region evolution in *Larix* and *Pseudotsuga* (Pinaceae). Am. J. Bot. 86, 711–723.

Goloboff, P., 2000. NONA: a tree searching program. Available from www.cladistics.com.

Goodman, J., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G., 1979. Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28, 132–168.

Graur, D., Li, W.H., 2000. Fundamentals of Molecular Evolution. Sinauer, Sunderland, MA.

Hall, P., Wilson, S.R., 1991. Two guidelines for bootstrap hypothesis testing. Biometrics 47, 757–762.

Hamby, R.K., Zimmer, E.A., 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. In: Soltis, P.S., Soltis, D.E., Doyle, J.J. (Eds.), Molecular Systematics of Plants. Chapman & Hall, New York, NY, pp. 50–91.

Hartmann, S., Nason, J.D., Bhattacharya, D., 2001. Extensive ribosomal DNA genic variation in the columnar cactus *Lophocereus*. J. Mol. Evol. 53, 124–134.

Hershkovitz, M.A., Zimmer, E.A., Hahn, W.J., 1999. Ribosomal DNA sequences and angiosperm systematics. In: Hollingsworth, P.M., Bateman, R.M., Gornall, R.J. (Eds.), Molecular Systematics and Plant Evolution. Taylor & Francis, London, pp. 268–326.

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., Yoshiki, A., 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature 423, 91–96.

Hughes, C.E., Bailey, C.D., Harris, S.A., 2002. Divergent and reticulate species relationships in *Leucaena* (Fabaceae) inferred from multiple data sources: insights into polyploid origins and nrDNA polymorphism. Am. J. Bot. 89, 1057–1073.

Jermiin, L.S., 1996. K2WuLi v. 1.0. Available from http://jcsmr.anu.edu.au/dicb/humgen/lars/k2wulisub.htm.

Jobst, J., King, K., Hemleben, V., 1998. Molecular evolution of the internal transcribed spacers (ITS 1 and ITS 2) and phylogenetic relationships among species of the family Cucurbitaceae. Mol. Phylogenet. Evol., 9.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

Kishino, H., Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in the Hominoidea. J. Mol. Evol. 29, 170–179.

Kita, Y., Ito, M., 2000. Nuclear ribosomal ITS sequences and phylogeny in East Asian *Aconitum* subgenus *Aconitum* (Ranunculaceae), with special reference to extensive polymorphism in individual plants. Plant Syst. Evol. 225, 1–13.

Kuzoff, R.K., Soltis, D.E., Hufford, L., Soltis, P.S., 1999. Phylogenetic relationships within *Lithophragma* (Saxifragaceae) hybridization, allopolyploidy and ovary diversification. Syst. Bot. 24, 598–615.

Kwon, O.Y., Ogino, K., Ishikawa, H., 1991. The longest 18S ribosomal RNA ever known: nucleotide sequence and presumed secondary structure of the 18S rRNA of the pea aphid, *Acyrthosiphon pisum*. Eur. J. Biochem. 202, 827–833.

Learn, G.H., Schaal, B.A., 1987. Population subdivision for ribosomal DNA repeat variants in *Clematis fremontii*. Evolution 41, 433–438.

Leitch, A.R., Mosgöller, W., Shi, M., Helsop-Harrison, J.S., 1992. Differential patterns of rDNA organisation in nuclei of wheat and rye. J. Cell Sci. 101, 751–757.

Li, W.H., 1997. Molecular Evolution. Sinauer, Sunderland, MA.

Li, W.H., Graur, D., 1991. Fundamentals of Molecular Evolution. Sinauer, Sunderland, MA.

Lim, K.Y., Kovarík, A., Matýãsek, R., Bezdek, M., Lichtenstein, C.P., Leitch, A.R., 2000. Gene conversion of ribosomal DNA in *Nicotiana tabacum* is associated with undermethylated, decondensed and probably active gene units. Chromosoma 109, 161–172.

Linder, C.R., Goertzen, L.R., Heuvel, B.V., Francisco-Ortega, J., Jansen, R.K., 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.

Liston, A., Robinson, W.A., Oliphant, J.M., Alvarez-Buylla, E.R., 1996. Length variation in nuclear ribosomal DNA internal transcribed spacer region of non-flowering seed plants. Syst. Bot. 21, 109–120.

Mai, J.C., Coleman, A.W., 1997. The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. J. Mol. Evol. 44, 258–271.

Manly, B.F.J., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology, second ed. Chapman & Hall, London.

Mayol, M., Rosselló, J.A., 2001. Why nuclear ribosomal DNA spacers (ITS) tell different stories in *Quercus*. Mol. Phylogenet. Evol. 19, 167–176.

Mooney, C.Z., Duval, R.D., 1993. Bootstrapping: a nonparametric approach to statistical inference. Sage University Paper series on Quantitative Applications in the Social Sciences 07-095, Sage Publications, Inc., Newbury Park, CA.

Muir, G., Fleming, C.C., Schlötterer, C., 2001. Three divergent rDNA clusters predate the species divergence of *Quercus petraea* (Matt.) Liebl.,*Quercus robur* L. Mol. Biol. Evol. 18, 112–119.

Nickrent, D.L., Soltis, D.E., 1995. A comparison of angiosperm phylogenies from nuclear 18S rDNA and *rbcL* sequences. Ann. Mo. Bot. Gard. 82, 208–234.

Nixon, K.C., 1999. WinClada (Beta) version 09 Published by author. Ithaca, New York Shareware.

O'Kane Jr., S.L., Schaal, B.A., Al-Shehbaz, I.A., 1996. The origins of *Arabidopsis suecica* (Brassicaceae) as indicated by nuclear rDNA sequences. Syst. Bot. 21, 559–566.

Olsen, L.E., Yoder, A.D., 2002. Using secondary structure to identify ribosomal numts: cautionary examples from the human genome. Mol. Biol. Evol. 19, 93–100.

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

Rauscher, J.T., Doyle, J.J., Brown, A.H.D., 2002. Internal transcribed spacer repeat-specific primers and the analysis of hybridization in the *Glycine tomentella* (Leguminosae) polyploid complex. Mol. Ecol. 11, 2691–2702.

Richardson, J.E., Pennington, R.T., Pennington, T.D., Hollingsworth, P.M., 2001. Rapid diversification of a species-rich genus of neotropical rainforest trees. Science 293, 2242–2245.

Sanderson, M.T., Doyle, J.J., 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy and confidence. Syst. Biol. 41, 4–17.

Sang, T., 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. 37, 121–147.

Sang, T., Crawford, D.J., Stuessy, T.F., 1995. Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: Implications for biogeography and concerted evolution. Proc. Natl. Acad. Sci. USA 92, 6813–6817.

Siddall, M.E., 1998. Success of parsimony in the four-taxon case: long branch repulsion by likelihood in the Farris Zone. Cladistics 14, 209–221.

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

Soltis, D.E., Soltis, P.S., Nickrent, D.L., Johnson, L.A., Hahn, W.J., Hoot, S.B., Sweere, J.A., Kuzoff, R.K., Kron, K.A., Chase, M.W., Swensen, S.M., Zimmer, E.A., Chaw, S.M., Gillespie, L.J., Kress, W.J., Sytsma, K.J., 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. Ann. Mo. Bot. Gard. 84, 1–49.

Suh, Y., Thien, L.B., Reeve, H.E., Zimmer, E.A., 1993. Molecular evolution and phylogenetic implications of internal transcribed spacer sequences of ribosomal DNA in Winteraceae. Am. J. Bot. 80, 1042–1055.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

Timmer, J., Lauk, M., Vach, W., Lücking, C.H., 1999. A test for a difference between spectral peak frequencies. Comput. Stat. Data Anal. 30, 45–55.

Torres-Ruiz, R.A., Hemleben, V., 1994. Pattern and degree of methylation in ribosomal RNA genes of *Cucurbita pepo* L. Plant Mol. Biol. 26, 1167–1179.

Vargas, P., McAllister, H.A., Morton, C., Jury, S.L., Wilkinson, M.J., 1999. Polyploid speciation in *Hedera* (Araliaceae): phylogenetic and biogeographic insights based on chromosome counts and ITS sequences. Plant Syst. Evol. 219, 165–179.

Vazquez, M.L., Doyle, J.J., Nixon, K.C., 2000. Paralogous ITS loci in Mexican red oak species (*Quercus* section *Lobatae*) and their implications in phylogeny reconstruction. American Society for Plant Taxonomists annual meeting—online abstracts, p. 189. Available from http://www.botany.org/bsa/portland.

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

Wenzel, J.W., Siddall, M.E., 1999. Noise. Cladistics 15, 51–64.

Widmer, A., Baltisberger, M., 1999. Molecular evidence for allopolyploid speciation and a single origin of the narrow endemic *Draba ladina* (Brassicaceae). Am. J. Bot. 86, 1282–1289.

Wu, C.-I., Li, W.-H., 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA 82, 1741–1745.

Yang, Y.W., Lai, K.N., Tai, P.Y., Ma, D.P., Li, W.H., 1999. Molecular phylogenetic studies of *Brassica*, *Rorippa*, *Arabidopsis* and allied genera based on the internal transcribed spacer region of 18S–25S rDNA. Mol. Phylogenet. Evol. 13, 455–462.