Syst. Biol. 54(5):697–709, 2005
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500221028

How Much Data are Needed to Resolve a Difficult Phylogeny? Case Study in Lamiales

Alexandra H. Wortley,[1] Paula J. Rudall,[2] David J. Harris,[3] and Robert W. Scotland[1]

[1] *Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK; E-mail: robert.scotland@plants.co.uk (R.W.S.)*
[2] *Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, UK*
[3] *Royal Botanic Garden, Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, UK*

Abstract.— Reconstructing phylogeny is a crucial target of contemporary biology, now commonly approached through computerized analysis of genetic sequence data. In angiosperms, despite recent progress at the ordinal level, many relationships between families remain unclear. Here we take a case study from Lamiales, an angiosperm order in which interfamilial relationships have so far proved particularly problematic. We examine the effect of changing one factor—the quantity of sequence data analyzed—on phylogeny reconstruction in this group. We use simulation to estimate a priori the sequence data that would be needed to resolve an accurate, supported phylogeny of Lamiales. We investigate the effect of increasing the length of sequence data analyzed, the rate of substitution in the sequences used, and of combining gene partitions. This method could be a valuable technique for planning systematic investigations in other problematic groups. Our results suggest that increasing sequence length is a better way to improve support, resolution, and accuracy than employing sequences with a faster substitution rate. Indeed, the latter may in some cases have detrimental effects on phylogeny reconstruction. Further molecular sequencing—of at least 10,000 bp—should result in a fully resolved and supported phylogeny of Lamiales, but at present the problematic aspects of this tree model remain. [Angiosperms; *matK*; molecular sequence data; *ndhF*; parsimony; *rbcL*; resolution; support.]

Computer algorithm-based phylogenetic analysis, often now using molecular genetic sequence data, has become the most important tool for elucidating phylogenetic relationships between species. The outcome of any phylogenetic analysis depends upon the data available, the taxa sampled, and the method of analysis. Here we investigate the first of these factors—the data analyzed—for one difficult part of the tree of life. We focus on determining how much, and what type of, sequence data would be needed to resolve a problematic phylogenetic question.

Relative to other groups of organisms, significant advances have been made in inferring the phylogeny of angiosperms, resulting in an improved ordinal classification of the group (APG, 1998; APGII, 2003). Despite the considerable progress at this level, the relationships of a number of angiosperm families remain unclear. For instance, in Lamiales, one of the largest orders of flowering plants, a number of well-supported families have been recognized, but interfamilial relationships are unclear and poorly supported (Olmstead et al., 2001; Bremer et al., 2002).

Here we investigate one potential strategy for improving phylogenetic inference in Lamiales. Using simulated sequence data, we assess the likelihood that a well-supported phylogeny can be resolved using molecular data, and how much data would then be needed.

Lamiales contains ca. 22,000 species, many of them of economic, horticultural, or scientific importance. These are grouped into ca. 23 families, including the mint family (Lamiaceae), figwort family (Scrophulariaceae), and *Acanthus* family (Acanthaceae). Most previous attempts at phylogeny reconstruction in the group have focused at the level of families, such as Acanthaceae, Bignoniaceae, or Gesneriaceae (e.g., Scotland et al., 1995; Smith et al., 1997; Spangler and Olmstead, 1999). More recently, broadening taxonomic coverage and increasing amounts of molecular data have begun to improve resolution within Lamiales (e.g., Olmstead et al., 2001), but even analyses of up to six molecular partitions have failed to provide fully supported resolution of all interfamilial relationships (e.g., Albach et al., 2001; Bremer et al., 2002). Although family-level clades tend to be consistently resolved and supported, the relationships between them are unstable, resulting from a concentration of short branches towards the base of the tree (Olmstead et al., 2001; Bremer et al., 2002).

Kim (1997) defined the term *tree model* to describe the topology, branch lengths, and aspects of character evolution associated with a phylogeny. The type of tree model seen in Lamiales, with short or zero-length internal branches, low decay values, and poor bootstrap support, is widespread in angiosperm systematics—it is displayed by at least 25 out of the 45 orders recognized in the recent APG classification (APGII, 2003)—and often attributed to rapid evolutionary radiations (e.g,. Chase et al., 2000; Philippe et al., 2000; Anderberg et al., 2002). A number of factors have been invoked to account for tree models of this type, including conflict between characters due to homoplasy within a sequence, insufficient sequence length, poor taxon sampling, rate variation across characters or taxa, compositional bias, conflict between gene trees due to paralogy and hybridization (Renner and Chanderbali, 2000), the presence of ("rogue") taxa with particularly unstable positions that may reduce support levels in the tree as a whole (Sanderson and Shaffer, 2002), or the historical signal of a rapid radiation (Fishbein et al., 2001).

Two of these factors most amenable to investigation are sequence length (character sampling) and taxon sampling; both have been widely discussed (e.g., Graybeal, 1998; Mitchell et al., 2000; Pollock et al., 2002; Hillis et al., 2003). In Lamiales, existing sequence data sets (only the chloroplast-encoded markers *rbcL* and *ndhF* are presently available for a broad range of taxa) already span the root node of all major families (as recommended by Olmstead et al., 2001; Prendini, 2001). Although adding more taxa might aid in the positioning of certain isolated genera, this article focuses on resolving the internal nodes between families. For this purpose, adding taxa would only increase the density of sampling within monophyletic groups, as determined by family-level studies (e.g., Wagstaff et al., 1998; Spangler and Olmstead, 1999; Young et al., 1999; Beardsley and Olmstead, 2002; Schwarzbach and McDade, 2002), and would therefore be unlikely to help resolve the problematic interfamilial relationships. Thus, in Lamiales, we chose to investigate how increased sampling of characters for existing taxa can improve phylogenetic inference.

We focused on increasing the lengths of nucleotide sequences that were otherwise realistic in their nature and rate of evolution. If this strategy improves the inference of phylogeny in Lamiales, it would suggest that its problematic tree model is due to data availability (a soft polytomy) rather than a real rapid radiation (a hard polytomy). If the strategy fails, then further investigations will be required to determine whether Lamiales is an example of a hard polytomy or whether another factor (such as those listed above) is responsible for the tree model seen. The best strategy for phylogenetic inference probably varies depending upon the group in question, but the results obtained for Lamiales may suggest ways to investigate similar questions in other taxa.

Empirically investigating the utility of different data sets in phylogeny reconstruction can be expensive and time consuming (but c.f. Soltis et al., 1998). Furthermore, the accuracy of the inferred phylogeny cannot be tested, because the underlying evolutionary relationships are not known. An alternative way to approach such questions is by studying simulated data sets (e.g., Lecointre et al., 1994; Hillis, 1995; Huelsenbeck, 1995; Berbee et al., 2000).

Simulation studies can be used a priori to investigate the most likely approach to yield a resolved, supported, and accurate phylogeny, and how much data might be needed for this. Existing problems in resolving the phylogeny of Lamiales result from a series of short internal branches. For molecular sequence data, the length of a branch is proportional to the number of nucleotide substitutions that have occurred along it (Swofford et al., 1996), which is a product of the time elapsed between nodes, the rate of nucleotide substitution, and the number of nucleotide sites sampled. Branch lengths can therefore be increased by sampling a larger number of sites, or sites that have experienced a greater rate of substitution.

Increased sampling is usually modeled by generating single partitions of increasing length (e.g., Lecointre et al., 1994; Philippe et al., 1994; Huelsenbeck et al., 1996). For real data sets, sampling is increased by combining sequences from several molecular markers. Combining genes may actually improve support relative to increasing the length of a single partition (Pennington, 1996; Mitchell et al., 2000), i.e., simultaneous analysis of separate sequence partitions might generate greater improvements in phylogeny reconstruction than would be expected from the number of additional nucleotides they provide. Therefore, as well as determining the absolute length of sequence that would be needed to resolve the phylogeny of Lamiales, we investigate how the effect differs when sequences are combined.

Finally, simulated sequences are used to estimate the practical impact of sequencing a particular additional marker (*matK*) for combination with the *rbcL-ndhF* data set. *MatK* was chosen as a potentially useful marker because it has previously been employed in combination with *rbcL* and *ndhF* to investigate interfamilial relationships in other orders of plants, including a similar tree model in Saxifragales (Fishbein et al., 2001; Bremer et al., 2002; Xiang et al., 2002; Hilu et al., 2003). This method of estimating the impact of a molecular data set prior to sequencing could be a valuable technique for planning systematic investigations. Here it is used to determine the most appropriate strategy for inferring Lamiales phylogeny with maximum accuracy, resolution, and support.

METHODS

Taxon Sampling

Thirty-seven exemplar species were sampled in Lamiales, up to three from each major family (after Yeates, 1995) (Appendix 1). Three out-group taxa were sampled from the related orders Boraginales, Gentianales, and Solanales. Sequence data for two genes were taken from the same species, where possible, or from closely related species.

Molecular Sequencing

A starting matrix of molecular sequences was assembled for the chloroplast genes *rbcL* and *ndhF*. New sequences were generated for *Thomandersia hensii, T. laurifolia* (unplaced in Lamiales), and *Synapsis ilicifolia* (Schlegeliaceae), using a modified CTAB extraction protocol method (Doyle and Doyle, 1987). The remainder were assembled from Olmstead et al. (2001), with additional sequences from Scotland (unpublished data), Olmstead and Reeves (unpublished data), and GenBank (http://www.ncbi.nlm.nih.gov/entrez/). GenBank accession numbers for new sequences used in this study are shown in Appendix 1. For a few genera, the two genes were sequenced from different species; for *Martynia* only an *ndhF* sequence was available. Sequence alignment was achieved by eye (Simmons and Ochoterena, 2000). The aligned matrix was submitted to TreeBASE (http://www.treebase.org/treebase) under the accession number M2231.

Simulation

Starting phylogenies.—Fully resolved unrooted tree topologies were generated for Lamiales from *ndhF* and*rbcL* using the neighbor-joining method in PAUP* 4.0b10 (Swofford, 2002). Branch lengths were estimated using uncorrected distances. It was assumed that the tree models reconstructed using *rbcL* and *ndhF* were representative of the underlying evolutionary tree model of

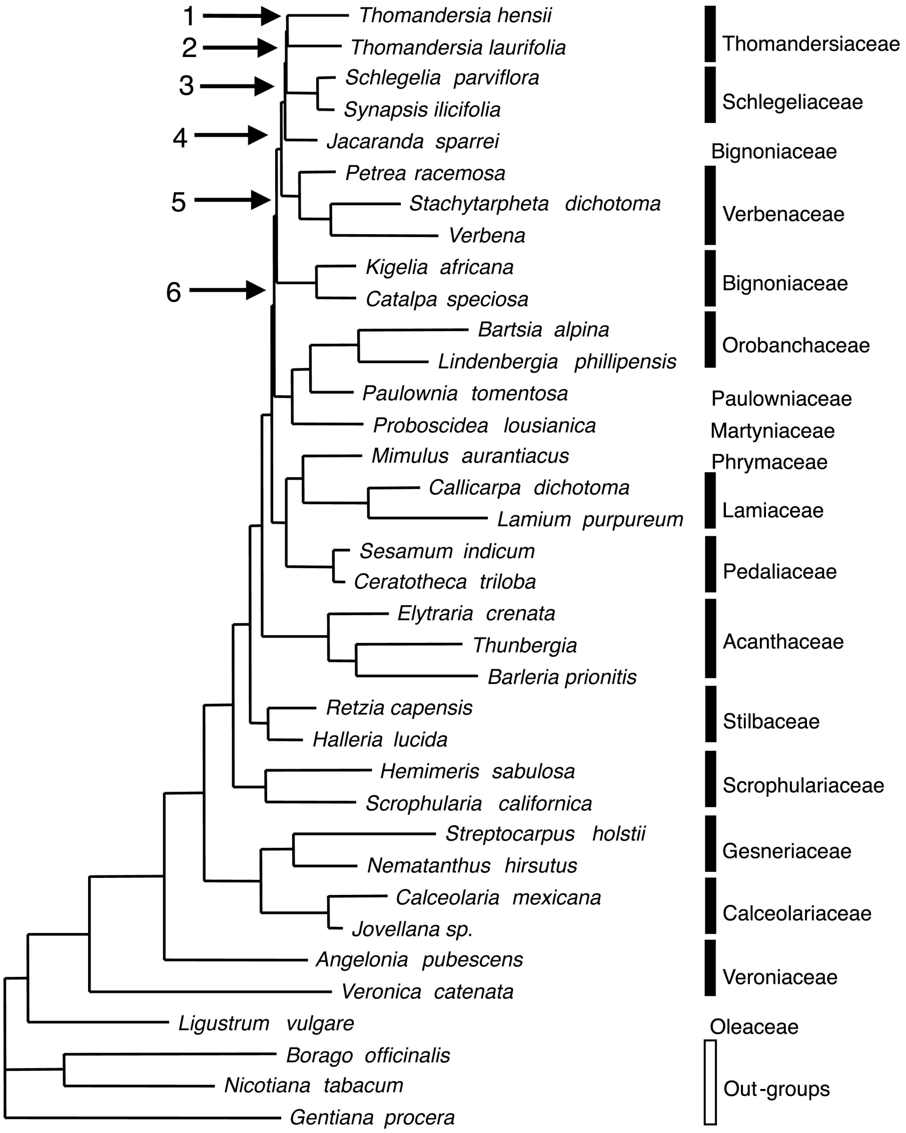

FIGURE 1. Starting phylogeny of Lamiales based on neighbor-joining analysis of combined *ndhF* and *rbcL* sequences, with relative branch lengths from uncorrected data. Key nodes highlighted.

Lamiales, because very similar relative branch lengths were obtained whether the phylogeny was reconstructed using the neighbor-joining method, parsimony, or maximum likelihood with a more complex model as generated from the sequences, using *rbcL* or *ndhF*, and because this form of tree model is very prevalent amongst other orders of angiosperms. A starting tree was also designed to represent the phylogeny that might be reconstructed from *matK*, based on a topology provided by a combined *rbcL-ndhF* analysis (Fig. 1). Employing this topology assumes minimal conflict between the three partitions and therefore means the results will, if anything, overestimate the potential usefulness of *matK*. Branch lengths were generated for the *matK* tree by doubling the branch lengths of the *rbcL* tree (no comparative data were available for *matK* and *ndhF*), because the average number of nucleotide substitutions per site for *matK* is consistently about twice that of *rbcL* (e.g., Fishbein et al., 2001; Lopez et al., 2002).

Sequence simulation.—To determine the most realistic evolutionary model and parameters for sequence simulation, the *ndhF* and *rbcL* data sets were tested against 56 possible models using ModelTest 3.06 and ModelBlock 3 (Posada and Crandall, 1998). For *rbcL*, the most suitable model, according to the hierarchical likelihood-ratio test, was the general reversible model, GTR+I+Γ, with relative base frequencies A = 0.2723, C = 0.1927, G = 0.2357, T = 0.2993, relative substitution rates A-C = 1.4734, A-G = 2.7252, A-T = 0.3801, C-G = 0.9138, C-T = 3.4541, G-T = 1, and gamma distribution shape parameter 0.8023. For *ndhF*, the model selected was also GTR+I+Γ, with relative base frequencies A = 0.2949, C = 0.1476, G = 0.1606, T = 0.3969, relative substitution rates A-C = 2.1962, A-G = 3.2938, A-T = 0.2243, C-G = 2.5359, C-T = 3.2724, G-T = 1, and gamma distribution shape parameter 1.0267. A model and parameters for simulation of *matK* sequences were taken from an existing data set for *Diploxylon* (Lopez et al., 2002). This was assumed to be applicable to Lamiales because the evolutionary model for *rbcL* in *Diploxylon* was very similar to that of *rbcL* in Lamiales. The evolutionary model selected for *matK* was HKY85, with transition/transversion ratio 1.68. Base frequencies were also taken from Lopez et al. (2002): A = 0.31, C = 0.18, G = 0.18, T = 0.33.

Using these parameters, replicate sequence data sets were simulated by parametric bootstrapping (Nei, 1991; Adell and Dopazo, 1994; Wollenberg and Atchley, 2000) on an iMac using Seq-Gen 1.2.5.1 (Rambaut and Grassly, 1997). Sets of replicates were simulated, each with a different combination of substitution rate and sequence length, and using the parameters for either *rbcL* or *ndhF* (Appendix 2). Further sets were generated to investigate the effects of combining two different partitions. These data sets were interleaved in all possible *rbcL-ndhF* combinations (Appendix 3). Data sets were also simulated to the specifications of the real *matK, ndhF*, and *rbcL* data, and were interleaved to investigate the potential of combining *matK* with the existing two-gene data set (Appendix 4). In each case, 100 replicate sets were generated to identical specifications.

Phylogenetic Analysis

Parsimony analysis was conducted on the two sequence partitions alone and in combination, using PAUP* 4.0b10 for Macintosh (Swofford, 2002). Characters were unordered and equally weighted. Gaps were coded as missing data, and zero-length branches collapsed so that only branches with unambiguous support were retained. Two search strategies were employed, in an attempt to find all most parsimonious (MP) trees and all possible islands (complying with Maddison, 1991; Catalan et al., 1997). The first strategy utilized 1000 replicate heuristic searches, with tree-bisection-reconnection (TBR) branch-swapping and MULPARS on, saving all trees. The second strategy employed 10,000 searches, with TBR branch-swapping and MULPARS on, saving only two trees per replicate. Trees were rooted in a basal polytomy using the predefined out-group. For simulated molecular sequences, searches of 1000 random-addition replicates were conducted, retaining a maximum of two trees at each step and saving a maximum of 100 trees.

Comparison of Resulting Phylogenies

The trees generated using real data were compared in terms of tree length, consistency index (CI, corrected for invariable sites), retention index (RI), resolution, and support. Resolution was measured by the consensus fork index (CFI), the number of nodes in strict consensus divided by the maximum possible number of nodes (Colless, 1980). Internal support was estimated by a nonparametric bootstrap search with 1000 replicates, each comprising 100 random addition heuristic search cycles, TBR branch-swapping and MULPARS off, saving no more than two trees per cycle (after De Bry and Olmstead, 2000). The total number of nodes achieving 50%, 75%, 90%, or 95% bootstrap values was counted, as a measure of support across each tree.

The replicate trees generated from simulated sequences were inspected automatically using a C^{++} program (A. South, personal communication), and were also compared to their respective starting trees according to the parameters above. Accuracy was also measured, as the number of MP trees containing each of several key nodes from the starting tree (labeled on Fig. 1), and the presence or absence of those nodes in strict consensus. The nodes were chosen to encompass a range of hierarchical levels, by taking successively nested nodes, and were subtended by some of the shortest branches in the tree, thereby being pertinent to the resolution of problematic interfamilial relationships in Lamiales. The bootstrap support value for each of the nodes was compared.

Results

Molecular Data.—Parsimony analysis of the *ndhF* sequences produced seven MP trees 2760 steps long, with CI 0.44 and RI 0.56. Analysis of the *rbcL* sequences produced 44 MP trees 826 steps long, with CI 0.43 and RI 0.57. Few conflicts in topology between the *ndhF* and *rbcL* trees were well supported: in only three instances were two incongruent clades both supported by a bootstrap value greater than even the low threshold of 50%. These were *Lindenbergia-Proboscidea* (54% in the *rbcL* tree) and *Lindenbergia-Bartsia* (100% in the *ndhF* tree); *Elytraria-Thunbergia* (97% *rbcL*) and *Barleria-Thunbergia* (51% *ndhF*); and *Borago-Nicotiana* (91% *rbcL*) and *Gentiana-Nicotiana* (70% *ndhF*). In all three cases, one topology was better supported than the other. This suggested there was no "hard incongruence" (Farris et al., 1994) between the *rbcL* and *ndhF* data sets and they were therefore combined into a single data set for further analysis.

This baseline two-gene molecular data set for 37 taxa contained 3578 aligned characters, of which 1371 were variable and 749 parsimony informative. Thirty-one MP trees were recovered, 3620 steps long (the second search strategy found 28 of these trees), with CI 0.43 and RI 0.35. A strict consensus tree is shown in Figure 2.

Effect of Molecular Sequence Length

Using simulated data sets, resolution increased with the length of sequence analysed (Fig. 3A). This effect was most marked with short lengths of sequence. High resolution (taken as a CFI of at least 0.95) was achieved using a total of approximately 10,000 base pairs (bp) of sequence data, for data sets based on *ndhF*-type sequences. A similar result was found for *rbcL*-type data sets. Accuracy, measured as the mean percentage of MP trees containing each of six nodes, increased with the length of sequence analysed (Fig. 3B). For *ndhF*-type sequence data sets, node 1 (comprising two species of *Thomandersia*) was found in all MP trees when at least 500 bp of sequence were analyzed, regardless of substitution rate. Deeper nodes required a greater length of sequence to be

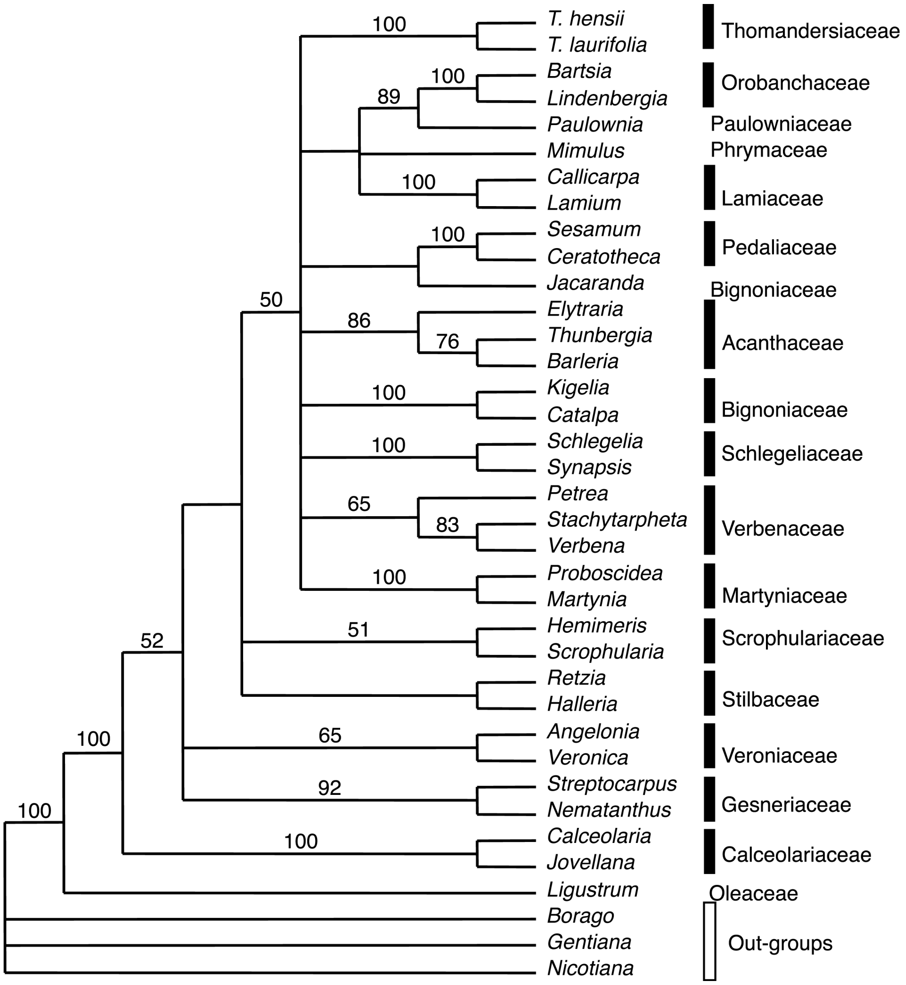

FIGURE 2. Strict consensus of 31 MP trees obtained from parsimony analysis of the two-gene molecular data set for 37 taxa in Lamiales. Numbers above branches are bootstrap values greater than 50%. Families after Olmstead et al. (2001).

accurately resolved: nodes 3 and 4 were found in all trees with 10,000 bp; nodes 5 and 6 required 20,000 bp to be accurately reconstructed almost all of the time. Node 2 was the least likely to be accurately resolved: 20,000 bp were required to recover it in 90% of MP trees. According to the second measure, the number of strict consensus trees in which each node was found, accuracy followed a very similar pattern.

Support, measured as the number of nodes attaining a certain bootstrap value, also increased with the length of sequence analysed. For *ndhF*-type data, 31 out of a maximum 34 nodes achieved at least 50% support when 10,000 bp were analyzed, regardless of substitution rate (Fig. 3C). No further increase in support was found with sequence lengths longer than 10,000 bp. The number of nodes achieving 90% bootstrap support was fewer, and was more responsive to increasing sequence length at a lower nucleotide substitution rate. *RbcL*-type data sets gave similar results. When support values for individual nodes were compared, with *ndhF*-type sequences, node 1 achieved 100% bootstrap support with 1000 bp. Nodes 3 and 4 achieved at least 70% support when 5000 bp were analyzed and 90% with 10,000 bp. Nodes 5 and 6 achieved at least 80% support with 10,000 bp and about 95% with 20,000 bp. Node 2 was the poorest supported, achieving less than 90% using 20,000 bp, even when sequences with a doubled nucleotide substitution rate were used. For this node the relationship between sequence length and support reached an asymptote: 90% was the maximum support achieved. *RbcL*-type sequences provided at least 95% bootstrap support for all six nodes with 20,000 bp of data (Fig. 3D). All nodes except 4 and 5 achieved at least 70% support with 5000 bp of sequence. Node 1 required longer sequences than under the *ndhF* model to achieve the same level of support: 5000 bp for 100%.

Effect of Substitution Rate

The effect of substitution rate was tested for two lengths of sequence, 1000 and 2000 bp, approximately the length of real single gene partitions. Resolution increased with substitution rate, but the relationship soon became asymptotic (Fig. 4A): when 2000 bp were used,

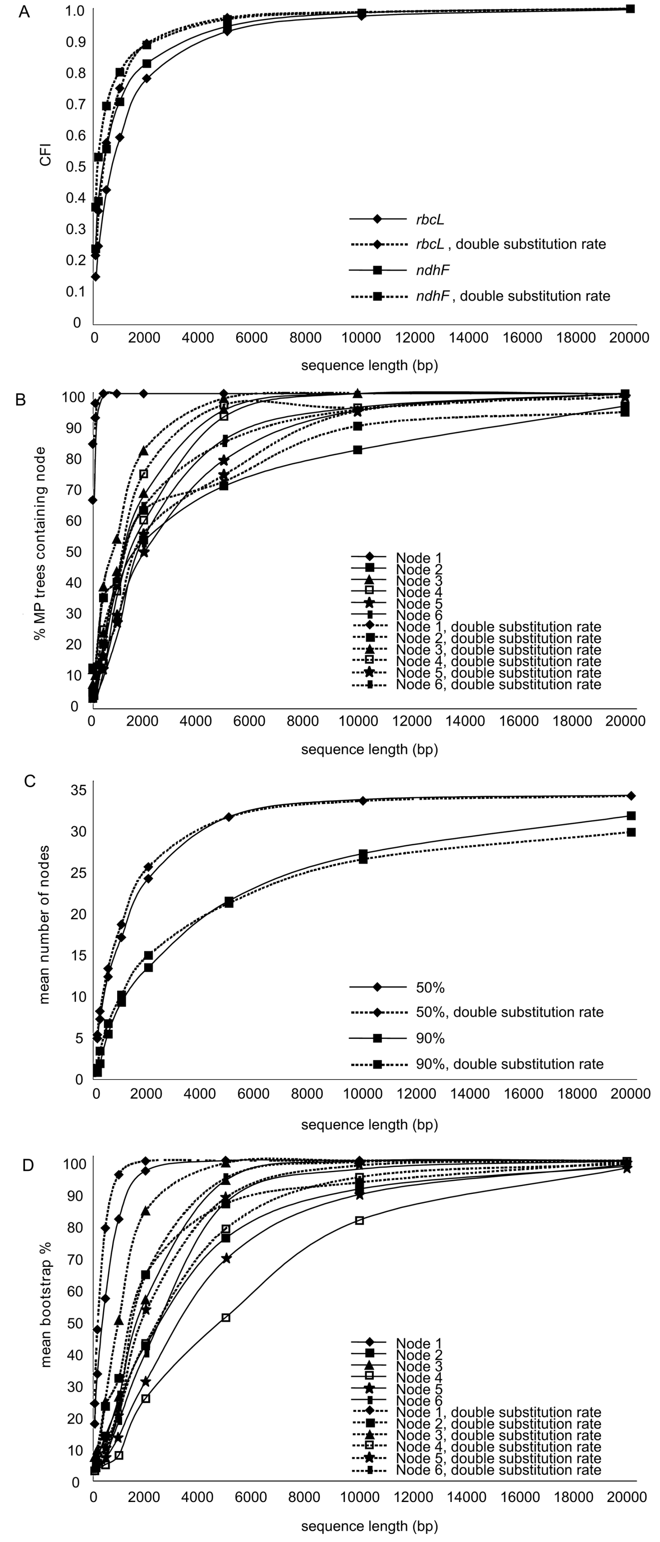

FIGURE 3. Effect of sequence length on phylogenies reconstructed using molecular data simulated at normal, and twice-normal, substitution rates. A, Resolution, measured by the CFI, for *ndhF*- and *rbcL*-type sequences. B, Accuracy, measured as the mean percentage of MP trees containing nodes 1 to 6, for *ndhF*-type sequences. C, Support, measured as the number of nodes attaining a bootstrap value of at least 50% or 90%, for *ndhF*-type sequences. D, Support, measured as the mean bootstrap support value attained by nodes 1 to 6, for *rbcL*-type sequences.

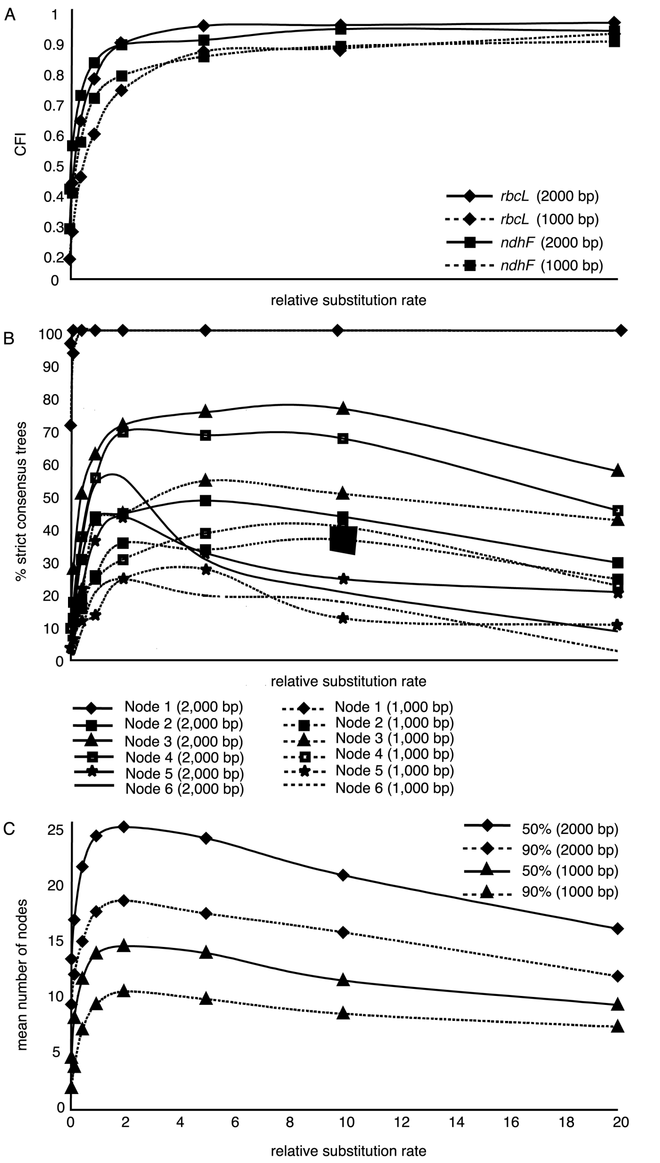

FIGURE 4. Effect of increasing substitution rate on phylogenies reconstructed using 1000 and 2000 aligned bp of simulated molecular data. A, Resolution, measured by the CFI, for *ndhF*- and *rbcL*-type sequences. B, Accuracy, measured as the percentage of strict consensus trees containing nodes 1 to 6, for *ndhF*-type sequences. C, Support, measured as the mean number of nodes attaining a bootstrap value of at least 50% or 90%, for *ndhF*-type sequences.

a maximum CFI of 0.95 was achieved at substitution rates of approximately double those in real *rbcL* and *ndhF* sequences. When 1000 bp were used, maximum resolution was achieved at approximately five times the real substitution rate of these genes. Accuracy showed a rapid increase followed by a gradual decline as substitution rate was increased. For *ndhF*-type sequences, node 1 was resolved accurately in 100% of strict consensus trees with substitution rates of at least 0.2 times the real rate (with 2000 bp, or 0.5 times with 1000 bp; Fig. 4B). For all other nodes, accurate reconstruction was not achieved in more than 80% of trees. Maximum accuracy occurred at between once and twice the original substitution rate (with 2000 bp, or twice and five times with 1000 bp) for *ndhF*-type sequences. For *rbcL*-type sequences, highest accuracy occurred at between 2 and 5 times the original substitution rate (with 2000 bp) or 5 and 10 times (with 1000 bp).

Support values showed a rapid increase and then a gradual decline as nucleotide substitution rate was increased, with both lengths of sequence and both types of partition tested. For *ndhF*-type data sets, strongest support was achieved at a substitution rate about double that of real sequences (Fig. 4C). For *rbcL*-type, it was found at about five times the real rate. With *ndhF*-type data, only node 1 received high bootstrap support; for all other nodes, support peaked at between twice and five times the real substitution rate and thereafter declined. Support values with the *rbcL*-type data were generally higher than with *ndhF* and peak support occurred at a faster substitution rate (twice normal for node 1, 10 times for other nodes).

Effect of Combining Gene Partitions

Figure 5A shows the resolution achieved using combinations of *rbcL*- and *ndhF*-type sequences in varying proportions. Greatest resolution was achieved when the entire length of sequence was simulated according to an *ndhF*-type model. All combined data sets, of the same total length, achieved lower resolution than this. Accuracy was also greater with a single partition than with most combined data sets of the same total sequence length, in this case when the entire length of sequence was simulated according to an *rbcL*-type model. Support showed a slightly different pattern (Fig. 5B shows this for the number of nodes achieving 95% bootstrap support). Although these data points are not all independent, the general indication is that combined data sets provided a better-supported phylogeny than single gene partitions of the same length, especially at total lengths below about 7000 bp. Above this length, single-partition *rbcL*-type sequences provided the greatest number of supported nodes. At 50% bootstrap values (not shown), combined partitions provided greater support than single partitions for a smaller range of total sequence lengths, between 2500 and 4000 bp.

Effect of Adding matK

The resolution obtained in phylogenetic analysis of the *matK*-type simulated data set alone was greater than that obtained with *rbcL* but less than that with *ndhF* (Fig. 6A). Combining the *matK*-type data set with *rbcL* and *ndhF* increased resolution significantly ($P < 0.0001$,

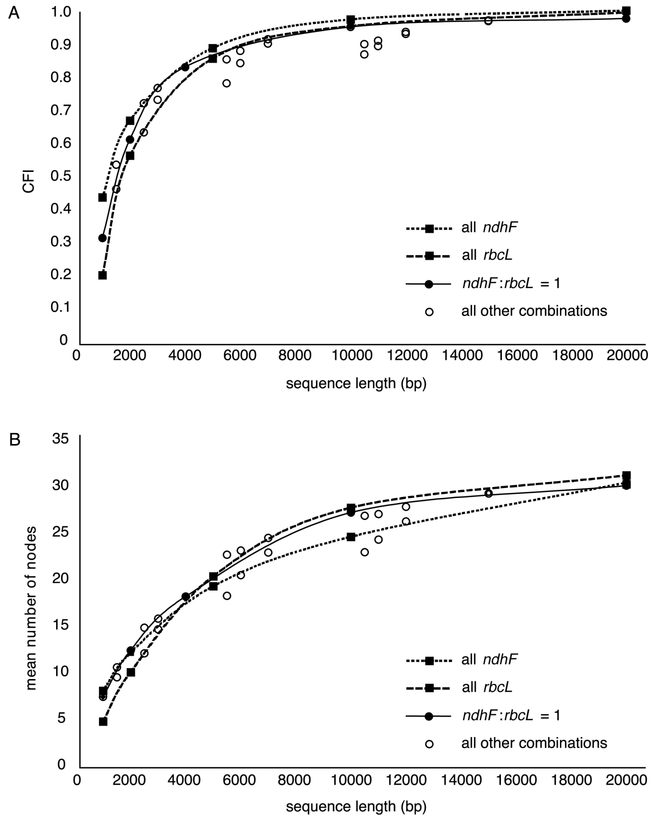

FIGURE 5. Effect of combining *ndhF* and *rbcL* data sets in different proportions. A, Resolution, measured by the CFI, against total length of aligned sequence analyzed. B, Support, measured as the mean number of nodes attaining a bootstrap value of at least 95%, against total length of aligned sequence analyzed.

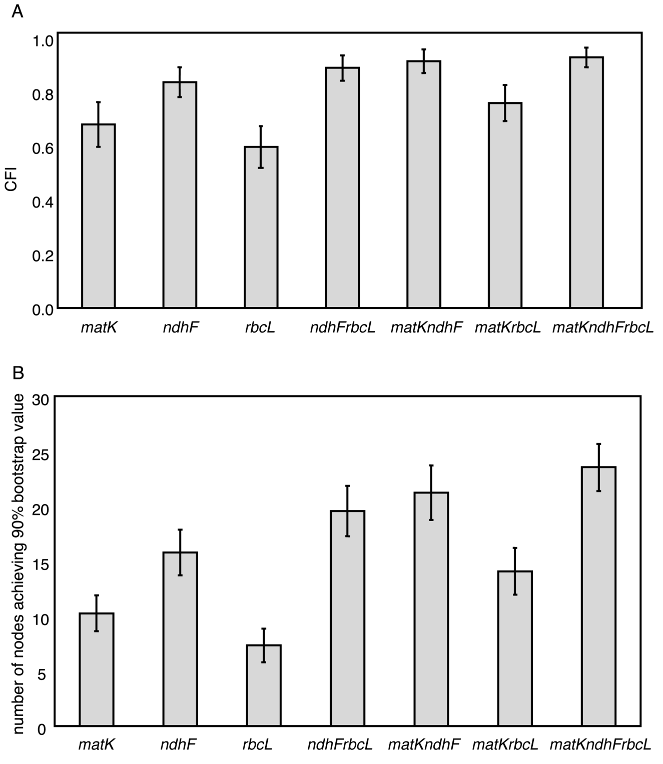

FIGURE 6. Effect of adding simulated *matK* sequences to the existing data set of *ndhF* and *rbcL* sequences. A, Resolution, measured by the CFI. B, Support, measured as the number of nodes attaining a bootstrap value of at least 90%. Error bars show standard deviation of 100 replicate data sets.

two-sample *t*-test) over that of *rbcL* and *ndhF* combined. The three-gene data set provided a significantly greater number of supported nodes than all other combinations ($P < 0.0001$; Fig. 6B shows the number of nodes achieving 90% bootstrap support).

Discussion

Lamiales exemplifies a widespread type of tree model in angiosperms, in which interfamilial relationships are difficult to resolve using molecular sequence data. In this example, phylogenies based on *ndhF* and *rbcL* sequences were broadly corroborative, showing no strongly supported incongruencies. However, a baseline molecular phylogeny generated using the combined sequence data set (Fig. 2) reproduced the same type of topology as previous analyses (Olmstead et al., 2001; Bremer et al., 2002), failing to resolve any well-supported interfamilial relationships beyond the basal branches of the order. This study contributes to estimating how much, and what type of, sequence characters would be required to improve the inference of phylogeny in Lamiales. Because all the chloroplast loci studied are linked, we assumed that conflict between gene trees does not present a problem for inferring the species tree in this group.

The question was investigated using data simulated on model trees. Although these trees may not be a true representation of evolutionary relationships among real organisms, they are most likely a realistic approximation to the type of branch lengths that would be found in the true phylogeny of Lamiales. Thus, although the phylogenies reconstructed from simulated sequences are also not necessarily accurate, they can be used to indicate the conditions required to reconstruct the true tree.

Effect of Sequence Length and Substitution Rate

All measures of resolution, support, and accuracy showed a rapid, then slowing improvement as increasing lengths of sequence data were analyzed. The asymptotic effect on resolution is due to the fact that, as clades are resolved, a decreasing number remain unresolved, and therefore the total number of newly resolved clades decreases for each further increment in character sampling. Moreover, once a certain level of sampling is reached, only very recalcitrant clades will remain to be resolved. With increasing nucleotide substitution rate, resolution also followed an asymptotic pattern, whereas support and accuracy showed a rapid increase followed by a slow decline.

Optimum Conditions for Phylogeny Reconstruction

The data sets that provided greatest resolution, highest accuracy, and strongest support were the four longest partitions simulated on an *rbcL*-type model. Greatest improvements in support, resolution, and accuracy were achieved by increasing the number of sites rather than the nucleotide substitution rate. This is shown in Figure 7 for *ndhF*-type data sets: for the same total number of substitutions, all nodes are more likely to be accurately reconstructed when 2000 bp are used than they are with 1000 bp; i.e., doubling the sequence length had a greater effect than doubling the substitution rate. In addition, too high a substitution rate can actually decrease support and accuracy (Fig. 4). This is in contrast with the results of Hillis (1998).

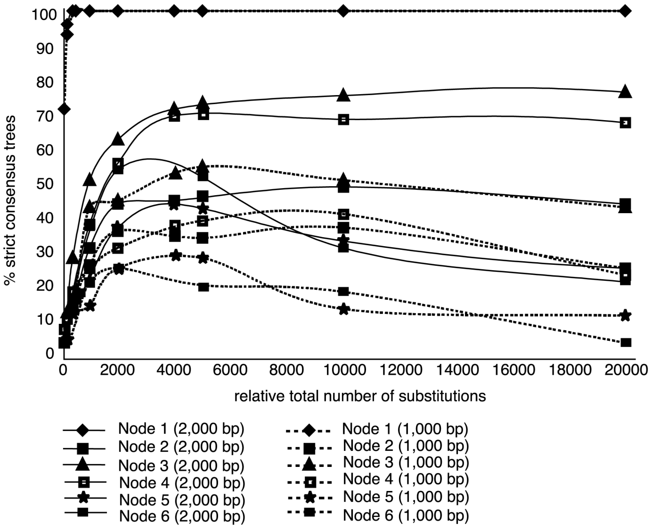

FIGURE 7. Relationship between accuracy, measured as the percentage of strict consensus trees containing nodes 1 to 6, and relative total number of nucleotide substitutions in phylogenetic analysis, for simulated *ndhF*-type sequences of 1000 and 2000 aligned bp.

One possible explanation for the superiority of longer, rather than faster-evolving, sequences for phylogenetic inference is that faster rates of nucleotide substitution may result in multiple substitutions (e.g., Moritz et al., 1987; Yang, 1998). The most useful data sets for supported, accurate phylogenetic inference in Lamiales were those comprising a large number of slowly-evolving sites, minimizing the likelihood of multiple substitutions at the same site. Node 1 does not conform to the same pattern, perhaps because it is the most recent node in the phylogeny and therefore the least likely to have experienced multiple substitutions at the same nucleotide sites.

The exact length of sequence required to infer the phylogeny of Lamiales depends upon the levels of support, accuracy, and resolution demanded. It is unreasonable to expect any phylogenetic analysis to provide 100% accurate, supported resolution (but c.f. Rokas et al., 2003), but good resolution (CFI > 0.9) can be achieved with relatively small amounts of sequence data (ca. 5000 bp). Extrapolating from Figure 3B, for *ndhF*-type data, approximately 19,000 bp of sequence are required to accurately resolve node 2 at normal substitution rates (employing conventional confidence limits, accuracy is taken as finding correctly reconstructed clades in 95% of MP trees), whereas only 300 bp are needed for node 1, and 5000 to 10,000 bp for the remaining nodes. For *rbcL*-type data, node 4 requires around 12,000 bp, other nodes between 1000 and 10,000 bp. Taking acceptable bootstrap support to be 75%, and extrapolating from Figure 3C, approximately 8600 bp of *rbcL*-type sequence are needed to support node 4, with 800 to 6000 bp for the other five nodes. For *ndhF*-type sequences, 12,000 bp are needed to support node 2, and 200 to 7500 bp for other nodes.

Currently, the nucleotide sequence data available for a wide range of Lamiales taxa are limited to *rbcL* (1402 bp) and *ndhF* (2176 aligned bp in this analysis). Although sequences of additional partitions such as *rps2* are available for a smaller range of families (e.g., Olmstead et al., 2001), the total sequence data available for large-scale phylogenetic studies in Lamiales amount to less than 3600 bp at present, of which about 750 were parsimony informative. To resolve most nodes with the support and accuracy described above requires, ca. 10,000 bp or an average of about 2,000 parsimony-informative sites, about three times as much as is currently available, but an amount that is achievable over the next few years, given the increasing ease of extracting and sequencing DNA data (Doolittle, 1999; Rokas et al., 2003; Sanderson and Driskell, 2003). In contrast, the largest morphological data set ever analyzed for Lamiales (Wortley, unpublished data) contains 105 parsimony-informative characters, and this figure is unlikely to grow at such a rate as molecular data.

Effect of Combining Gene Partitions

This estimate assumes it would be feasible to generate 10,000 bp of sequence in a single partition. In practice, increasing the number of nucleotide sites available for analysis is achieved by combining several data sets. Combining separate molecular data sets might be beneficial, because two gene partitions may provide optimal resolution and support at slightly different hierarchical levels. Here, *rbcL* and *ndhF*-type sequences differed in the hierarchical level at which they provided best resolution. To resolve node 1 accurately in 95% of strict consensus trees required only 500 bp of *ndhF*-type but 2000 bp of *rbcL*-type sequence data; *ndhF*-type sequences were best for accurately resolving node 1, *rbcL*-type for node 2. This suggests, as is often assumed, that simultaneous analysis can improve support and resolution by at least as much as the analysis of the individual datasets would suggest (Hillis, 1987; Reed and Sperling, 1999); i.e., “the whole will exceed the sum of the parts.” We used the simulated sequences to investigate whether combining molecular sequence data from different gene partitions does have such a synergistic effect.

In fact, combining two partitions reduced the effectiveness of phylogenetic analysis in Lamiales. All combined data sets gave lower resolution than single-partition analyses of the same length (Fig. 5A). Most combined data sets also accurately reconstructed fewer nodes than the same length single partitions. The effect of data combination on support values was equivocal: combined data may support a greater or lesser number of nodes than a single partition, depending upon the total length of sequence analysed (Fig. 5B). Thus, increasing data by combining partitions may necessitate a much larger sequencing project than the 10,000 bp suggested above, perhaps because the combined partitions contain contradictory signals. In addition, real data sets may display complex features that are not accurately replicated by simulation, which could confound phylogenetic inference further.

Effect of Adding matK

We extended our results based on existing data sets (*ndhF* and *rbcL*) to predict the effect of sequencing an additional molecular marker, *matK*. This provided, at best, marginal improvements in resolution, and no significant improvement in accuracy. The biggest gains were seen in support values, which increased for all nodes except node 2. This is consistent with the fact that adding *matK* to the existing data produces a total data set of around 4612 bp, which is within the range of lengths found to increase support (Fig. 3C), but shorter than would be needed to significantly improve accuracy (Fig. 3B). Because *matK* sequences were simulated upon a starting phylogeny whose topology was derived from the *ndhF* and *rbcL* data sets, conflict between data sets should be minimal in this case. Yet adding simulated *matK* sequences to the existing data set was not as successful as expected, suggesting the prospects for resolving a supported phylogeny for Lamiales by adding new molecular datasets may be limited. Alternatively, because *matK* was assumed to have evolved along a phylogeny with the same branch lengths as that of *rbcL* and *ndhF*, it might increase support to those branches that are already resolved but not improve resolution of those that are extremely short in the two-gene molecular morphology. Therefore, in reality, adding *matK* might provide fewer improvements in support but greater improvements in resolution than seen here, depending upon whether it tends to conflict with the existing data or to follow a similar pattern.

Phylogeny of Lamiales

The success of phylogenetic inference can be measured in terms of resolution, support, and accuracy. Although molecular sequences are not the only data available for inferring phylogenetic relationships, they are probably the most efficient in the case of Lamiales and other similar groups. Informative morphological characters are few and far between and have been shown to have little significant effect on phylogeny reconstruction in Lamiales, although this may not be the case in other groups (Graham et al., 1998; Wortley, unpublished data). The most promising approach to resolving accurate, supported interfamilial relationships in Lamiales using molecular sequence data would be to sequence at least 2000 parsimony-informative characters, corresponding to a sequence data set of 10,000 bp or more. This equates to approximately six genes, three times the size of the existing two-gene data set. In practice, combining genes may further increase the total number of nucleotides needed to resolve an accurate, supported phylogeny. This is consistent with the recent study of Rokas et al. (2003), which suggested that at least 20 unlinked genes may need to be sequenced in order to overcome incongruence and reconstruct a supported, accurate phylogeny.

Until larger amounts of molecular sequence data become available, the trees previously published (e.g., Olmstead et al., 2001), remain the best available representation of the phylogeny of this group. The problematic aspects of the Lamiales tree model remain, as seen in the baseline molecular phylogeny from real data presented here (Fig. 2). There are few well-supported branches (taken as those with at least 75% bootstrap support) at interfamilial level. Acanthaceae, Calceolariaceae, Gesneriaceae, Lamiaceae, Martyniaceae, Orobanchaceae, Pedaliaceae, Schlegeliaceae, and Thomandersiaceae are all well supported by the bootstrap test. A core Lamiales clade comprising Acanthaceae, Bignoniaceae, Lamiaceae, Martyniaceae, Pedaliaceae, Orobanchaceae, Schlegeliaceae, Thomandersiaceae, Verbenaceae, *Mimulus*, and *Paulownia* is consistently recovered but with little support. The only other well-supported branches are those subtending Lamiales as a whole, Lamiales excluding *Ligustrum* (Oleaceae), and the grouping of *Paulownia* with Orobanchaceae. The position of some relatively basal groups such as Gesneriaceae and Veronicaceae is still unstable; others, such as Carlemanniaceae and Tetrachondraceae (Stevens, 2001; APGII, 2003), have not been sampled here. Three recently erected families (Scrophulariaceae sensu stricto, Stilbaceae, and Veronicaceae, all containing members of what was once Scrophulariaceae sensu lato) are also only poorly supported.

Our case study of Lamiales suggests that large-scale, focused molecular sequencing projects and subsequent concatenation of multiple markers should be able to resolve this problematic tree model. Whether this inference can be applied to similarly problematic angiosperm orders, such as Apiales, Cornales, Ericales, Laurales, and Malpighiales (Plunkett et al., 1996; Soltis et al., 1999; Anderberg et al., 2000; Renner and Chanderbali, 2000; Xiang et al., 2002), remains to be seen. The resolution of problematic phylogenies will depend upon the origin of the problem—in cases of hard polytomies such a tree model represents a solution, and adding more data will not improve resolution. However, in other plant groups, and other organisms, increasingly large multigene and whole-genome molecular phylogenetic investigations are fast becoming the norm and are generally resulting in better-resolved phylogenies (e.g., Bapteste et al., 2002; Matsuoka et al., 2002; Goremykin et al., 2003; Rokas et al., 2003). Thus, although the predicted, comprehensive solution to the entire tree of life (Doolittle, 1999) is a distant objective of molecular systematics (Sanderson et al., 2003), an accurate, supported, and resolved phylogeny of well-studied parts of the tree will soon be achievable if current rates of progress continue.

Acknowledgements

AHW would like to acknowledge Dr. Andrew Rambaut for help with simulating and analyzing simulated sequences, and Professor Richard Olmstead for provision of unpublished sequences and much informative discussion. The authors would also like to thank associate editor Peter Linder, Sean Graham, and an anonymous reviewer for many helpful comments on the manuscript.

References

Adell, J. C., and J. Dopazo. 1994. Monte Carlo simulation in phylogenies: An application to test the constancy of evolutionary rates. J. Mol. Evol. 38:305–309.

Albach, D. C., P. S. Soltis, D. E. Soltis, and R. G. Olmstead. 2001. Phylogenetic analysis of asterids based on sequences of four genes. Ann. Miss. Bot. Garden 88:163–212.

Anderberg, A. A., C. Rydin, and M. Kallersjö. 2002. Phylogenetic relationships in the order Ericales *s.l*: Analyses of molecular data from five genes from the plastid and mitochondrial genomes. Am. J. Bot. 89:677–687.

Anderberg, A. A., B. Stahl, and M. Kallersjö. 2000. Maesaceae, a new primuloid family in the order Ericales *s.l* Taxon 49:183–187.

APG. 1998. An ordinal classification for the families of flowering plants. Ann. Miss. Bot. Garden 85:531–553.

APGII. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot. J. Linn. Soc. 141:399–436.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Nat. Acad. Sci. USA 99:1414–1419.

Beardsley, P. M., and R. G. Olmstead. 2002. Redefining Phrymaceae: The placement of *Mimulus*, tribe Mimuleae, and *Phryma*. Am. J. Bot. 89:1093–1102.

Berbee, M. L., D. A. Carmean, and K. Winka. 2000. Ribosomal DNA and resolution of branching order among the Ascomycota: How many nucleotides are enough? Mole. Phylogenet. Evol. 17:337–344.

Bremer, B., K. Bremer, N. Heidari, P. Erixon, R. G. Olmstead, A. A. Anderberg, M. Kallersjö, and E. Barkhordarian. 2002. Phylogenetics of asterids based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. Mol. Phylogenet. Evol. 24:274–301.

Catalan, P., E. A. Kellogg, and R. G. Olmstead. 1997. Phylogeny of Poaceae subfamily Pooideae based on chloroplast *ndhF* gene sequences. Mol. Phylogenet. Evol. 8:150–166.

Chase, M. W., M. F. Fay, and V. Savolainen. 2000. Higher-level classification in the angiosperms: New insights from the perspective of DNA sequence data. Taxon 49:685–704.

Colless, D. H. 1980. Congruence between morphometric and allozyme data for *Menidia* species: A reappraisal. Syst. Zool. 29:288–299.

De Bry, R. W., and R. G. Olmstead. 2000. A simulation study of reduced tree-search effort in bootstrap resampling analysis. Syst. Biol. 49:171–179.

Doolittle, W. F. 1999. Phylogenetic classification and the universal tree. Science 284:2124–2128.

Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure from small quantities of fresh leaf tissues. Phytochem. Bull. 19:11–15.

Farris, J. S., M. Kallersjö, A. G. Kluge, and C. J. Bult. 1994. Testing significance of incongruence. Cladistics 10:315–319.

Fishbein, M., C. Hibsch-Jetter, D. E. Soltis, and L. Hufford. 2001. Phylogeny of Saxifragales (Angiosperms: Eudicots): Analysis of a rapid, ancient radiation. Syst. Biol. 50:817–847.

Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol. Bio. Evol. 20:1499–1505.

Graham, S. W., J. R. Kohn, B. R. Morton, J. E. Eckenwalder, and S. C. H. Barrett. 1998. Phylogenetic congruence and discordance among one morphological and three molecular data sets from Pontederiaceae. Syst. Biol. 47:545–567.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47:9–17.

Hillis, D. M. 1987. Molecular versus morphological approaches to systematics. Ann. Rev. Ecol. and Syst. 18:23–42.

Hillis, D. M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 47:3–8.

Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3–8.

Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwicki. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124–126.

Hilu, K. W., T. Borsch, K. Muller, D. E. Soltis, P. S. Soltis, V. Savolainen, M. W. Chase, C. E. Powell, L. A. Alice, R. C. Evans, H. Sauquet, C. Neinhuis, T. A. B. Slotta, J. G. Rohwer, C. S. Campbell, and L. W. Chatrou. 2003. Angiosperm phylogeny based on *matK* sequence information. Ame. J. Bot. 90:1758–1776.

Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

Huelsenbeck, J. P., D. M. Hillis, and R. Jones. 1996. Parametric bootstrapping in molecular phylogenetics: Applications and performance. Pages 19–45 *in* Molecular zoology: Advances, strategies and protocols (J. D. Ferraris and S. R. Palumbi, eds.). Wiley-Liss, New York.

Kim, J. 1997. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. 47:43–60.

Lecointre, G., H. Philippe, H. L. Van Le, and H. Le Guyader. 1994. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. Mol. Phylogenet. Evol. 3:292–309.

Lopez, G. G., K. Kamiya, and K. Harada. 2002. Phylogenetic relationships of *Diploxylon* pines (subgenus *Pinus*) based on plastid sequence data. Int. J. Plant Sci. 163:737–747.

Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. 40:315–328.

Matsuoka, Y., Y. Yamazaki, Y. Ogihara, and K. Tsunewaki. 2002. Whole chloroplast genome comparison of rice, maize, and wheat: Implications for chloroplast gene diversification and phylogeny of cereals. Mol. Biol. Evol. 19:2084–2091.

Mitchell, A., C. Mitter, and J. C. Regier. 2000. More taxa or more characters revisited: Combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuiodea (Insecta: Lepidoptera). Syst. Biol. 49:202–224.

Moritz, C., T. E. Dowling, and W. M. Brown. 1987. Evolution of animal mitochondrial DNA: Relevance for population Biol. and Systematics. Annu. Rev. Ecol. Syst. 18:269–292.

Nei, M. 1991. Relative efficiencies of different tree-making methods for molecular data. Pages 90–128 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford University Press, New York.

Olmstead, R. G., C. W. De Pamphilis, A. D. Wolfe, N. D. Young, W. J. Elisons, and P. A. Reeves. 2001. Disintegration of the Scrophulariaceae. Am. J. Bot. 88:348–361.

Pennington, R. T. 1996. Molecular and morphological data provide resolution at different levels in *Andira*. Syst. Biol. 45:496–515.

Philippe, H., A. Chenuil, and A. Adoutte. 1994. Can the Cambrian explosion be inferred through molecular phylogeny? Development 1994:(Suppl):S15–S25.

Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, and H. Le Guyader. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly-evolving positions. Proc. R. Soc. Lond. Ser. B 267:1213–1221.

Plunkett, G. M., D. E. Soltis, and P. S. Soltis. 1996. Clarification of the relationships between Apiaceae and Araliaceae based on *matK* and *rbcL* data. Ame. J. Bot. 84:565–580.

Pollock, D. D., D. J. Zwicki, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.

Posada, D., and K. A. Crandall. 1998. Model Test: Testing the model of DNA substitution. Bioinformatics 14:817–818.

Prendini, L. 2001. Species or supraspecific taxa as terminals in cladistic analysis? Groundplans versus exemplars revisited. Syst. Bot. 50:290–291.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Reed, R. D., and F. A. H. Sperling. 1999. Interaction of process partitions in phylogenetic analysis: An example from the swallowtail butterfly genus *Papilio*. Mol. Biol. Evol. 16:286–297.

Renner, S. S., and A. S. Chanderbali. 2000. What is the relationship among Hernandiaceae, Lauraceae, and Monimiaceae, and why is this question so difficult to answer? Int. J. Plant Sci. 161(Suppl.):S109–S119.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Sanderson, M. J., and A. C. Driskell. 2003. The challenge of constructing large phylogenetic trees. Trends Plant Sci. 8:374–379.

Sanderson, M. J., and H. B. Shaffer. 2002. Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. 33:49–72.

Schwarzbach, A. E., and L. A. McDade. 2002. Phylogenetic relationships of the mangrove family Avicenniaceae based on chloroplast and nuclear ribosomal DNA sequences. Syst. Bot. 27:84–98.

Scotland, R. W., J. A. Sweere, P. A. Reeves, and R. G. Olmstead. 1995. Higher-level Syst.s of Acanthaceae determined by chloroplast DNA sequences. Ame. J. Bot. 82:266–275.

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

Smith, J. F., J. C. Wolfram, K. D. Brown, C. L. Carroll, and D. S. Denton. 1997. Tribal relationships in the Gesneriaceae: Evidence from DNA sequences of the chloroplast gene *ndhF*. Ann. Miss. Bot. Garden 84:50–66.

Soltis, D. E., M. E. Mort, P. S. Soltis, C. Hibsch-Jetter, E. A. Zimmer, and D. R. Morgan. 1999. Phylogenetic relationships of the enigmatic angiosperm family Podostemaceae inferred from 18S rDNA and *rbcL* sequence data. Mol. Phylogenet. Evol. 11:261–272.

Soltis, D. E., P. S. Soltis, M. Mort, M. W. Chase, V. Savolainen, S. B. Hoot, and C. M. Morton. 1998. Inferring complex phylogenies using parsimony: An empirical approach using three large DNA datasets for angiosperms. Syst. Biol. 47:32–42.

Spangler, R. E., and R. G. Olmstead. 1999. Phylogenetic analysis of Bignoniaceae based on the cpDNA gene sequences *rbcL* and *ndhF*. Ann. Miss. Bot. Garden 86:33–46.

Stevens, P. F. 2001. Angiosperm Phylogeny website; http://www.mobot.org/MOBOT/research/APweb/.

Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., G. L. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics (D. M. Hillis, C. Morowitz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

Wagstaff, S. J., L. Hickerson, R. E. Spangler, P. A. Reeves, and R. G. Olmstead. 1998. Phylogeny in Labiatae *s.l.*, inferred from cpDNA sequences. Plant Syst. Evol. 209:265–274.

Wollenberg, K. R., and W. R. Atchley. 2000. Separation of phylogenetic and functional associations in biological sequences using the parametric bootstrap. Proc. Nat. Acad. Sci. USA 97:3288–3291.

Xiang, Q. Y., M. L. Moody, D. E. Soltis, C. Z. Fan, and P. S. Soltis. 2002. Relationships within Cornales and circumscription of Cornaceae—*matK* and *rbcL* sequence data and effects of outgroups and long branches. Mol. Phylogeneti. Evol. 24:35–57.

Yang, Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

Yeates, D. K. 1995. Groundplans and exemplars: Paths to the tree of life. Cladistics 11:343–357.

Young, N. D., K. E. Steiner, and C. W. De Pamphilis. 1999. The evolution of parasitism in the Scrophulariaceae/Orobanchaceae: Plastid gene sequences refute an evolutionary transition series. Ann. Miss. Bot. Garden 86:876–893.

First submitted 18 May 2004; reviews returned 27 August 2004; final acceptance 10 November 2004
Associate Editor: Peter Linder

APPENDIX 1. Taxa sampled.

Taxon	DNA source (reference), voucher or collection	GenBank accession numbers	
		rbcL	*ndhF*
Lamiales			
Acanthaceae			
Barleria prionitis L.	Chase et al. (1993), Scotland et al. (1995)	L01886	U12653
Elytraria crenata Vahl	Scotland et al. (1995)	AF188127	U12657
Thunbergia alata Boj.	Scotland et al. (1995)	—	U12667
T. usambarica Lindau	Chase et al. (1993)	L12596	—
Bignoniaceae			
Catalpa speciosa Warder	Olmstead et al. (1992)	L11679	L36397
Jacaranda sparrei A.H.Gentry	Spangler and Olmstead (1999)	AF102647	AF102631
Kigelia africana (Lam.) Benth.	Spangler and Olmstead (1999)	AF102648	AF102632
Calceolariaceae			
Calceolaria mexicana Benth.	Olmstead et al. (2001)	AF123669	AF123769
Jovellana sp.	Olmstead et al. (2001)	AF123666	AF123684
Gesneriaceae			
Nematanthus hirsutus (Mart.) Wiehler	Olmstead and Reeves (1995)	L36446	L36404
Streptocarpus holstii Engl.	Olmstead et al. (2001)	L14409	L36415
Lamiaceae			
Callicarpa dichotoma (Lour.) K. Koch	Olmstead et al. (1993)	L14393	L36415
Lamium purpureum L.	Olmstead et al. (1993)	U75702	U78694
Martyniaceae			
Martynia annua L.	Olmstead and Reeves (1995)	—	AF190906
Proboscidea lousianica (Mill.) Thell.	Chase et al. (1993)	L01946	AF123690
Oleaceae			
Ligustrum vulgare L.	Olmstead et al. (1992)	L11686	AF130164
Orobanchaceae			
Bartsia alpina L.	Young et al. (1999)	AF190903	AF123678
Lindenbergia philippensis Benth.	Young et al. (1999)	AF123664	AF123686
Paulowniaceae			
Paulownia tomentosa (Thunb.) Steud.	Olmstead and Reeves (1995)	L36447	L36406
Pedaliaceae			
Ceratotheca triloba (Bernh.) Hook. f.	Olmstead and Reeves (1995)	AY919277	AY919281
Sesamum indicum L.	Olmstead et al. (1993)	L14408	L36413
Phrymaceae			
Mimulus aurantiacus Curtis	Young et al. (1999)	AF026835	AF188186
Schlegeliaceae			
Schlegelia parviflora (Oerst.) Monach.	Olmstead and Reeves (1995)	L36448	L36410
Synapsis ilicifolia Griseb.	Ekman 19085	AY919278	AY919282
Scrophulariaceae			
Hemimeris sabulosa L.	Young et al. (1999)	AF123668	AF123682
Scrophularia californica Cham. and Schltdl.	Olmstead and Reeves (1995)	L36449	L36411
Stilbaceae			
Halleria lucida L.	Olmstead et al. (2001)	AF026828	AF188185
Retzia capensis Thunb.	Bremer et al. (1994)	Z29669	AF14776
Thomandersiaceae			
Thomandersia hensii De Wild. and T. Durand	*Carroll* 1008	AY919279	AY919284
Thomandersia laurifolia (T. Anderson ex Benth.) Baill.	*Cable* 4012	AY919280	AY919285
Verbenaceae			
Petrea racemosa Kunth	Olmstead and Reeves (1995), Wagstaff and Olmstead (1997)	U28879	AY919283
Stachytarpheta dichotoma Vahl.	Olmstead et al. (2001)	U32161	L36414
Verbena bonariensis L.	Olmstead et al. (1993)	L14412	—
V. bracteata Lag. and Rodr.	Olmstead and Reeves (1995)	—	L36418
Veronicaceae			
Angelonia pubescens Benth.	Olmstead et al. (2001)	AF123672	AF123675
Veronica catenata Pennell	Olmstead and Reeves (1995)	L36453	L36419
Outgroups			
Borago officinalis L.	Olmstead et al. (1992)	L11680	L36393
Gentiana procera Holm	Olmstead et al. (1993)	L14398	L36400
Nicotiana tabacum L.	Olmstead et al. (1993)	Z00044	Z00044

APPENDIX 2. Simulated molecular data sets and parameters.

Data set	Topology and parameters as	Relative no. of substitutions per site	Sequence length (bp)	No. of replicate matrices	Starting seed
1	*ndhF*	0.1	2000	100	15448376
2	*ndhF*	0.2	2000	100	47709730
3	*ndhF*	0.5	2000	100	78399216
4	*ndhF*	1	2000	100	79842335
5	*ndhF*	2	2000	100	52711674
6	*ndhF*	5	2000	100	2486395
7	*ndhF*	10	2000	100	40019257
8	*ndhF*	20	2000	100	50255930
9	*rbcL*	0.1	2000	100	84624646
10	*rbcL*	0.2	2000	100	19788860
11	*rbcL*	0.5	2000	100	67529069
12	*rbcL*	1	2000	100	24128070
13	*rbcL*	2	2000	100	82906566
14	*rbcL*	5	2000	100	46522473
15	*rbcL*	10	2000	100	18348872
16	*rbcL*	20	2000	100	25410028
17	*ndhF*	0.1	1000	100	73024013
18	*ndhF*	0.2	1000	100	89971506
19	*ndhF*	0.5	1000	100	31302977
20	*ndhF*	1	1000	100	47677141
21	*ndhF*	2	1000	100	58334587
22	*ndhF*	5	1000	100	25115300
23	*ndhF*	10	1000	100	78492019
24	*ndhF*	20	1000	100	91255506
25	*rbcL*	0.1	1000	100	41461212
26	*rbcL*	0.2	1000	100	11702582
27	*rbcL*	0.5	1000	100	52191925
28	*rbcL*	1	1000	100	98032014
29	*rbcL*	2	1000	100	93555486
30	*rbcL*	5	1000	100	29566716
31	*rbcL*	10	1000	100	95852382
32	*rbcL*	20	1000	100	87387046
33	*ndhF*	1	100	100	49319467
34	*ndhF*	1	200	100	87626389
35	*ndhF*	1	500	100	99908045
36	*ndhF*	1	1000	100	18054075
37	*ndhF*	1	2000	100	30201192
38	*ndhF*	1	5000	100	27046649
39	*ndhF*	1	10,000	100	46273839
40	*ndhF*	1	20,000	50	76508394
41	*rbcL*	1	100	100	10078655
42	*rbcL*	1	200	100	3605281
43	*rbcL*	1	500	100	28595320
44	*rbcL*	1	1000	100	55770326
45	*rbcL*	1	2000	100	41513186
46	*rbcL*	1	5000	100	22136409
47	*rbcL*	1	10,000	100	22808617
48	*rbcL*	1	20,000	50	24561367
49	*ndhF*	2	100	100	38322705
50	*ndhF*	2	200	100	88423710
51	*ndhF*	2	500	100	30393937
52	*ndhF*	2	1000	100	19995511
53	*ndhF*	2	2000	100	97044384
54	*ndhF*	2	5000	100	94042089
55	*ndhF*	2	10,000	100	8494918
56	*ndhF*	2	20,000	50	49726076
57	*rbcL*	2	100	100	80927635
58	*rbcL*	2	200	100	11440646
59	*rbcL*	2	500	100	79718335
60	*rbcL*	2	1000	100	34909131
61	*rbcL*	2	2000	100	63707149
62	*rbcL*	2	5000	100	89050894
63	*rbcL*	2	10,000	100	89970723
64	*rbcL*	2	20,000	50	23612088

APPENDIX 2. Simulated molecular data sets and parameters. *(Continued)*

Data set	Topology and parameters as	Relative no. of substitutions per site	Sequence length (bp)	No. of replicate matrices	Starting seed
65	*rbcL*	1	500	100	88936737
66	*rbcL*	1	1000	100	48334848
67	*rbcL*	1	2000	100	17673847
68	*rbcL*	1	5000	100	95982113
69	*rbcL*	1	10,000	100*	61466965
70	*ndhF*	1	500	100	88191445
71	*ndhF*	1	1000	100	49317736
72	*ndhF*	1	2000	100	70621389
73	*ndhF*	1	5000	100	93790617
74	*ndhF*	1	10,000	100*	38780095
75	*matK*	1	1034	100	68743963
76	*ndhF*	1	2176	100	7104129
77	*rbcL*	1	1402	100	90779718

*Only 50 replicate matrices were used when data sets 69 and 74 were combined, to maintain a manageable file size for analysis.

APPENDIX 3. Combined simulated data sets.

Combined data set	*rbcL* Data set	*rbcL* Sequence length (bp)	*ndhF* Data set	*ndhF* Sequence length (bp)	Total sequence length
A	65	500	70	500	1000
B	65	500	71	1000	1500
C	65	500	72	2000	2500
D	65	500	73	5000	5500
E	65	500	74	10,000	10,500
F	66	1000	70	500	1500
G	66	1000	71	1000	2000
H	66	1000	72	2000	3000
I	66	1000	73	5000	6000
J	66	1000	74	10,000	11,000
K	67	2000	70	500	2500
L	67	2000	71	1000	3000
M	67	2000	72	2000	4000
N	67	2000	73	5000	7000
O	67	2000	74	10,000	12,000
P	68	5000	70	500	5500
Q	68	5000	71	1000	6000
R	68	5000	72	2000	7000
S	68	5000	73	5000	10,000
T	68	5000	74	10,000	15,000
U	69	10,000	70	500	10,500
V	69	10,000	71	1000	11,000
W	69	10,000	72	2000	12,000
X	69	10,000	73	5000	15,000
Y	69	10,000	74	10,000	20,000

APPENDIX 4. Combined *matK*, *ndhF*, and *rbcL* data sets.

Combined data set	*matK* Data set?	*ndhF* Data set?	*rbcL* Data set?	Total sequence length
matK	Y	N	N	1034
ndhF	N	Y	N	2176
rbcL	N	N	Y	1402
ndhF-rbcL	N	Y	Y	3578
matK-ndhF	Y	Y	N	3210
matK-rbcL	Y	N	Y	2436
matK-ndhF-rbcL	Y	Y	Y	4612
matK-real data	Y	(Real)	(Real)	4612