Syst. Biol. 55(5):818–836, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150600999150

Phylogenetic Supermatrix Analysis of GenBank Sequences from 2228 Papilionoid Legumes

MICHELLE M. MCMAHON[1,2] AND MICHAEL J. SANDERSON[1,3]

[1]*Section of Evolution and Ecology, University of California, Davis, One Shields Avenue, Davis, California 95616, USA; E-mail: mcmahonm@email.arizona.edu (M.M.M.); sanderm@email.arizona.edu (M.J.S.)*
[2]*Current Address: Department of Plant Sciences, University of Arizona, Tucson, Arizona 85721, USA*
[3]*Current Address: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA*

Abstract.— A comprehensive phylogeny of papilionoid legumes was inferred from sequences of 2228 taxa in GenBank release 147. A semiautomated analysis pipeline was constructed to download, parse, assemble, align, combine, and build trees from a pool of 11,881 sequences. Initial steps included all-against-all BLAST similarity searches coupled with assembly, using a novel strategy for building length-homogeneous primary sequence clusters. This was followed by a combination of global and local alignment protocols to build larger secondary clusters of locally aligned sequences, thus taking into account the dramatic differences in length of the heterogeneous coding and noncoding sequence data present in GenBank. Next, clusters were checked for the presence of duplicate genes and other potentially misleading sequences and examined for combinability with other clusters on the basis of taxon overlap. Finally, two supermatrices were constructed: a "sparse" matrix based on the primary clusters alone (1794 taxa × 53,977 characters), and a somewhat more "dense" matrix based on the secondary clusters (2228 taxa × 33,168 characters). Both matrices were very sparse, with 95% of their cells containing gaps or question marks. These were subjected to extensive heuristic parsimony analyses using deterministic and stochastic heuristics, including bootstrap analyses. A "reduced consensus" bootstrap analysis was also performed to detect cryptic signal in a subtree of the data set corresponding to a "backbone" phylogeny proposed in previous studies. Overall, the dense supermatrix appeared to provide much more satisfying results, indicated by better resolution of the bootstrap tree, excellent agreement with the backbone papilionoid tree in the reduced bootstrap consensus analysis, few problematic large polytomies in the strict consensus, and less fragmentation of conventionally recognized genera. Nevertheless, at lower taxonomic levels several problems were identified and diagnosed. A large number of methodological issues in supermatrix construction at this scale are discussed, including detection of annotation errors in GenBank sequences; the shortage of effective algorithms and software for local multiple sequence alignment; the difficulty of overcoming effects of fragmentation of data into nearly disjoint blocks in sparse supermatrices; and the lack of informative tools to assess confidence limits in very large trees. [Alignment; Fabaceae/Leguminosae; Papilionoideae; phylogeny; phyloinformatics; supermatrix.]

Large sequence databases such as GenBank (Benson et al., 2005) are a rich repository of phylogenetically relevant information at the molecular level. As of April 2006, GenBank contained 57 million sequences for 151,000 species. Electronic access to these databases via the Internet facilitates the timely deposition of sequence data and makes those data widely available. Although the rapid doubling time of the database as a whole is now legendary, the pace of *taxonomic* sampling observed in GenBank has also been dramatic: the number of species with at least one sequence in GenBank has doubled since 2001. Whereas the former can be attributed in large part to industrial-scale whole-genome sequencing, the latter clearly testifies to the intense worldwide interest in systematics and biodiversity.

Most attempts to exploit very large amounts of sequence data for phylogenetic inference have adopted either a "phylogenomics" approach (Eisen and Fraser, 2003), in which whole genomes are used to construct a "supermatrix" of concatenated loci, from which trees are built (e.g., Rokas et al., 2003; Lerat et al., 2003), or a "phyloinformatics" approach, in which sequences are extracted from the databases irrespective of the original source (e.g., as part of genome projects or not: Driskell et al., 2004; Wolf et al., 2004; Philippe et al., 2004, 2005). Phylogenomic studies have typically combined 50 to 200 orthologous loci from whole genomes and assembled large matrices with little missing data, generally for few taxa. Although some of these studies have inferred trees with remarkably little uncertainty attached, others have opened up new controversies, as in the case of angiosperm phylogenies constructed from whole chloroplast genomes (e.g., Goremykin et al., 2003; Leebens-Mack et al., 2005), a case that illustrates the important effects of taxon sampling.

On the other hand, "phyloinformatics" approaches can take advantage of additional taxon sampling available in sequence databases, at the cost of considerable heterogeneity and sampling problems. This heterogeneity has been a prime motivation for the development of supertree strategies (reviewed in Bininda-Emonds, 2004), which combine evidence from different loci by combining the trees built from those loci. A number of quite complete, taxon-rich supertrees have been published for various clades (e.g., carnivores: Bininda-Emonds et al., 1999; legumes: Wojciechowski et al., 2000; dinosaurs: Pisani et al., 2002), and several have been used for evolutionary or ecological studies (e.g., Grotkopp et al., 2004; Moles et al., 2005). Alternatively, the loci can be combined into one very large super *matrix* from which a tree is inferred directly. Because the data are not obtained by targeted genome projects or other coordinated community-wide sequencing efforts, taxon sampling among loci tends to be very heterogeneous, leading to supermatrices that typically have much more missing data than those in phylogenomic studies (Driskell et al., 2004; Wolf et al., 2004; Philippe et al., 2004, 2005).

Phyloinformatic approaches may also be differentiated from increasingly common "multigene" studies that target a selected set of taxa and loci in an effort to resolve relationships at a particular level in a phylogeny. Although this is an approach with a proven track record (e.g., the spectacular success revising understanding of angiosperm relationships based on *rbcL*, *atpB*, and 18S rDNA [and now several other loci: Qiu et al., 2005], culminating in a revised classification; Angiosperm Phylogeny Group II, 2003), it becomes increasingly infeasible in large sets of densely sampled taxa. By sampling from sequence databases, it may be possible to combine more efficiently information relevant to both deep and shallow nodes in the tree of life, assuming that such data have been deposited by workers interested in problems at those different levels.

The goal of this paper is to test the limits of the phyloinformatics approach by attempting to build a credible supermatrix and phylogenetic tree of all species of papilionoid legumes in GenBank. In this paper we pursue a supermatrix rather than supertree approach. Papilionoid legumes are a good candidate for this exercise. They are an important group of angiosperms that comprise some 14,000 species nested within the Leguminosae (Lewis et al., 2005) and account for about 70% of its species diversity. They have been studied by both systematists and biologists focusing on the many model legume systems. The phylogenies of its largest genera (*Astragalus*, *Indigofera*, *Dalea*, *Medicago*, *Trifolium*, etc.) have been the subject of molecular systematic work; the backbone of the tree has been extensively studied with broad samples across a large number of genera (Pennington et al., 2000; Kajita et al., 2001; Wojciechowski et al., 2004; for a brief review, see Doyle and Luckow, 2003); and several studies at an intermediate scale have focused on "tribal" level relationships (e.g., Amorpheae: McMahon and Hufford, 2004; Robinieae: Lavin and Sousa S., 1995; Millettieae: Hu et al., 2000; Mirbelieae and Bossiaeeae: Crisp and Cook, 2003; Vicieae and Trifolieae: Steele and Wojciechowski, 2003; Loteae: Allan et al., 2003). Finally, their taxonomy has recently been comprehensively surveyed (Lewis et al., 2005).

The scale of analysis spans roughly 12,000 sequences and 2200 taxa. Exploitation of sequence databases at this scale for phylogenetic purposes faces a number of methodological and computational obstacles (Sanderson and Driskell, 2003; Delsuc et al., 2005). First, strategies for assembling data from a large database into collections of manageable data sets have been little explored (Sanderson and Driskell, 2003; Driskell et al., 2004). Second, several problems that have long received attention from the phylogenetics community, such as tree-building and sequence alignment, face novel challenges from the biased and heterogeneously distributed data in the databases (e.g., patterns of missing data). One of the clearest examples of this is the problem of aligning sequence data from diverse loci or partitions that contain both closely and distantly related taxa simultaneously. Finally, the scale of the data, especially the large number of taxa, poses a serious challenge to many existing protocols for phylogenetic analysis (Guindon and Gascuel, 2003; Tamura et al., 2004; Salamin et al., 2005; Vinh and von Haeseler, 2005). This paper examines all of these general problems in the specific context of reconstructing relationships of one clade.

Much of this paper describes protocols for acquiring and processing large quantities of phylogenetically relevant sequence data. Earlier versions of some of these (Driskell et al., 2004) have been adapted to the specific problems encountered in this project; other protocols are new, especially those involving the very significant alignment problems encountered in data that are dominated by noncoding DNA sequences. Driskell et al. (2004) reconstructed deep phylogenies of green plants and metazoan animals based on amino acid sequence data, which was easier to automate. The present study includes 20 times as many taxa, which necessitated additional modifications to alignment and phylogenetic inference protocols. Evaluating the results of our analyses, either in terms of conventional estimates of support or by comparing results to previously published work, presented new challenges.

METHODS

Supermatrix Construction

Data.—We downloaded the GenBank flat files (Release 147, 15 April 2005: gbpln files only) for green plants from NCBI (ftp.ncbi.nih.gov/genbank) and extracted all sequence records of molecule type "DNA" in the LOCUS line for organisms belonging to the clade Papilionoideae. These files do not include very large collections from high-throughput genome sequencing or EST projects, nor do they include cDNAs, but they do include the vast majority of sequences known from non-model organisms. We further limited the records to include only those of 5kb length or less to exclude; e.g., genomic sequences from whole chromosomes that had not been shunted by NCBI to other divisions.

Sequence clustering and alignment.—Assembly of an appropriate collection of phylogenetically related sequences is a hard problem, well studied in the context of construction of protein "families," but poorly studied in the more typical phylogenetic setting of heterogeneous mixtures of coding and noncoding data of different lengths and degree of overlap. The latter can occur because of evolutionary insertions and deletions or, more problematically, because of primer choice and mosaic patterns of local homology (e.g., intron loss or domain shuffling). Because multiple sequence alignment algorithms generally assume that sequences differ only through processes of substitutions and small insertions or deletions, large-scale differences in length and degree of overlap can lead to severe problems (Lassmann and Sonnhammer, 2002, 2005a). This involves the general problem of local multiple sequence alignment (Gusfield, 1997), for which solutions are less well established than global alignment (Lassmann and Sonnhammer, 2002). Tools that have been developed relatively recently include those that attempt to align locally

homologous segments, (DIALIGN 2: Morgenstern, 1999; DIALIGN-T: Subramanian et al., 2005), to sidestep the "multiple" part of the alignment problem entirely by aligning pairwise to a reference sequence (Blast Align: Belshaw and Katzourakis, 2005), to combine local and global alignment procedures (e.g., T-Coffee: Notredame et al., 2000; POA: Grasso and Lee, 2004), or to employ hidden Markov models to align sequences to profiles (HMMer: Eddy, 1998). Other methods are designed solely to align proteins (e.g., Kalign: Lassmann and Sonnhammer, 2005b; HMMer 2: Eddy, 1998, http://hmmer.wustl.edu/) and were not applicable to our data.

Figure 1 illustrates the general problem we encountered in trying to use standard alignment software with sequences of heterogeneous length. Our most troubling data sets originated from sequenced regions for which there has been much variation in primer sets, such as the nuclear ribosomal ITS region and the chloroplast *trnK* intron. These are also among the most frequently sequenced regions and were therefore critical to the study. Preliminary experimentation with several alignment programs demonstrated that global alignment algorithms (such as that found in Clustal W; Thompson et al., 1994), which are not designed for this problem, encountered difficulty in assigning partial sequences to the correct subregion (Fig. 1). For our data, the local alignment program DIALIGN (Morgenstern, 1999) performed best in this regard, although it was significantly slower than any other program. Our assessment of the quality of alignments was largely based on this gross level, i.e., whether sequences annotated as a certain subregion were indeed placed in that subregion. This is not an easy process to automate, however, because sequence annotations are highly variable and difficult to parse. An alternative approach is to assess intraalignment or interalignment consistency and use this as a basis of comparison (Poirot et al., 2003; Lassmann and Sonnhammer 2005a). For the example shown in Figure 1, the results of an interalignment consistency analysis using MUMSA (Lassmann and Sonnhammer 2005a) were similar to our gross-scale assessment (see Fig. 1), with DIALIGN and Clustal W receiving the highest and lowest scores, respectively, for this particularly difficult set of sequences.

None of these programs alone offered satisfactory solutions, so we developed two alternative heuristic strategies. A schematic of these strategies, as well as other analysis steps, is shown in Figure 2. The first step was to assemble sequences into sets of homologs, or "clusters," for eventual multiple alignment. We used all-against-all local similarity searches using the NCBI BLAST program *blastn* (Altschul et al., 1990) to compare each sequence to every other sequence with a cutoff maximum E-value of 1.0E-10 and the low-complexity filter DUST turned off, which prevents the breakage of long homologs into short fragments based on intervening low complexity runs of bases. This strategy identifies pairs of sequences that have one or more regions of statistically significant homology within them. We then used single-linkage clustering to put all sequences together that hit against at least one other member of the cluster (cf., Dondoshansky, 2002). However, additional factors are critical to deciding whether or not to assemble sequences into a cluster of homologs, especially the length, number, and degree of overlap of the local regions of homology in the pair.

To characterize these, we developed two statistics describing the set of BLAST hits between any two sequences. In Figure 3 the typical case is illustrated. One or

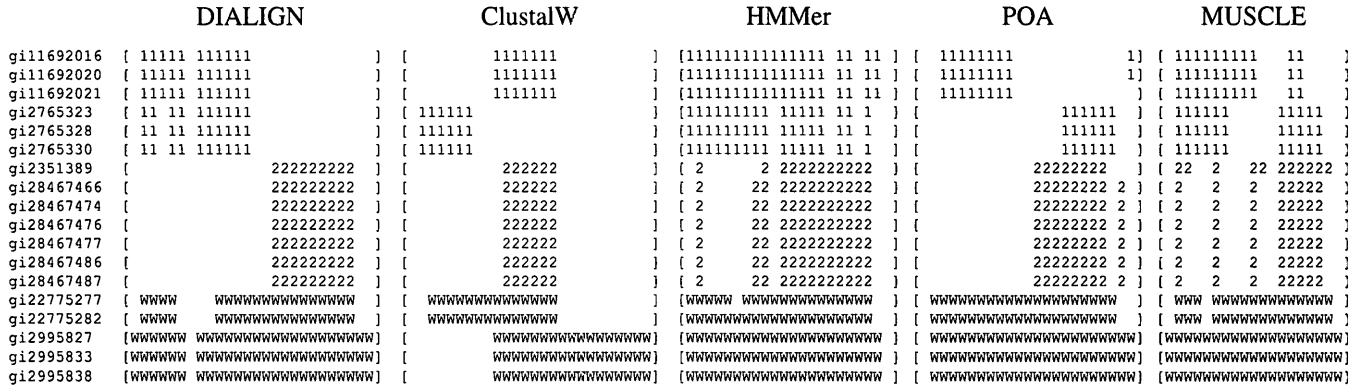

FIGURE 1. Example of performance of alignment programs Clustal W 1.83 (Thompson et al., 1994), DIALIGN 2.2 (Morgenstern, 1999), MUSCLE 3.6 (Edgar, 2004), POA 2 (Grasso and Lee, 2004), and HMMer 1.8.5 (Eddy, 1998) on the same collection of sequences exhibiting only local homologies. Input sequences consisted of 123 papilionoid GenBank accessions comprising parts or all of the nuclear ribosomal internal transcribed spacer (ITS) region, including ITS1, 5.8S rRNA gene, ITS 2, and small parts of the flanking 18S and 26S rRNA genes. A subset of sequences is shown. Programs were run with default options. Output shows a symbol if there are any bases in the alignment in a binned length of size 50 bases. Sequences containing data from only ITS1 or only ITS2 are indicated as 1 or 2; sequences that span most or all of the region are indicated as W. The diagram shows overall structure of the programs' homology assessment, although its fine structure obviously includes additional gaps. Total length of the five alignments differs because of the difference in numbers of gaps inserted. Note the poor performance of Clustal W, expected because the sequences lack global homology. The multiple overlap score (MOS) calculated by MUMSA (Lassmann and Sonnhammerm, 2005a) is given for each alignment, which is related to the proportion of base homologies found in that alignment that are consistent across alignments. Sequence data set available at http://ginger.ucdavis.edu/Benchmark.htm and http://systematicbiology.org.

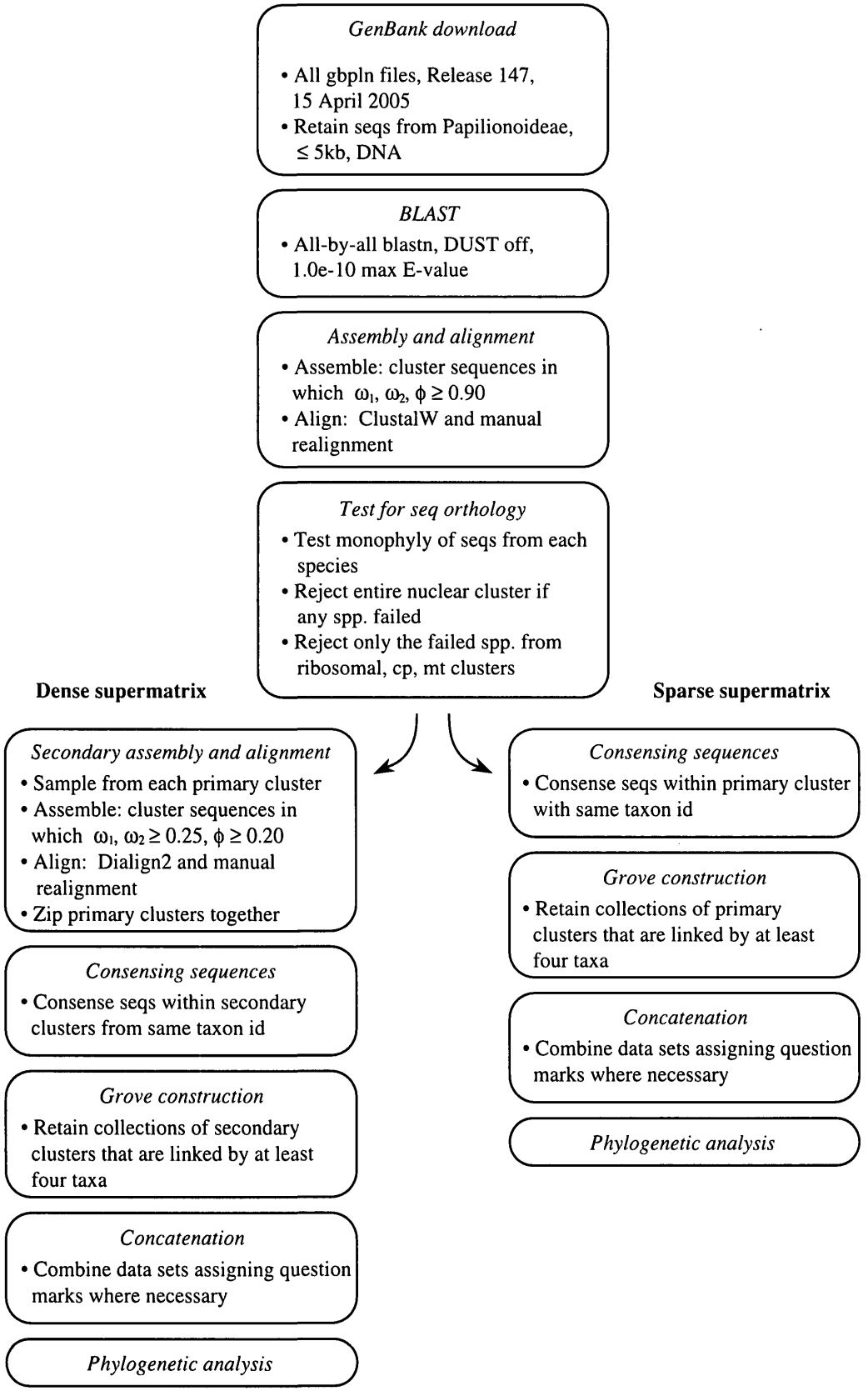

FIGURE 2. Flowchart of the data analysis steps used to assemble the sparse and dense supermatrices. Steps proceed from top to bottom.

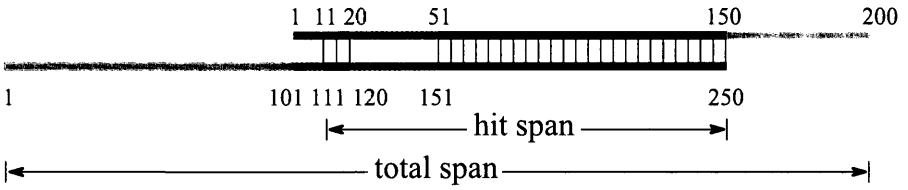

FIGURE 3. Schematic diagram of our criteria used to establish homology between pairs of sequences following BLAST searches. Example shows two overlapping sequences in which BLAST has reported two "hits" (regions of sufficient local homology above the user-supplied cutoff expect value, indicated by black regions). Hit span is the length from beginning of first hit to end of last hit. The hit fraction, ω, is the fraction of the sequence's length involved in BLAST hits. The hit span fraction, ϕ, is the fraction of the total span taken up by the hit span. See Methods for more precise definitions. Adjustment of these two parameters during assembly of pairwise BLAST hits into clusters greatly affects how heterogeneous the sequences are in terms of lengths, internal structure, and sequence divergence.

more hits occur (black areas) and the remaining sequence is judged non-homologous. Consider the top sequence, sequence 1. Its length, S_1, is the sum of three things: the so-called hit span length, H_1, which is the length of sequence from the beginning of the first hit to the end of the last hit (thus including some interspersed nonhomologous regions), plus a leftmost nonhomologous piece of length L_1, and rightmost nonhomologous piece of length R_1. Similar lengths are defined for the bottom sequence, of length S_2. Note that H_1 and H_2 need not be equal, e.g., interspersed nonhomologous regions may differ in length, or the sequences may contain repeats, which produce complex hit patterns. Define the "hit span fraction" as

$$\phi = \frac{\max(H_1, H_2)}{\max(L_1, L_2) + \max(H_1, H_2) + \max(R_1, R_2)}$$

This provides a quantitative indication of the degree of overhanging nonhomologous ends in the sequences, with values close to 0 reflecting very long nonhomologous ends on one or both sequences. However, values close to 1.0 could occur even if there were two small hits separated by nonhomologous interspersed regions, as long as there were short ragged ends. To monitor this, we construct a "hit fraction." Assume there are k hits, labeled, $h_i^{(j)}$ to indicate the ith hit in coordinates relative to sequence j. The total length of hits on sequence 1, H_1^*, is given by the length of the union of these sequences: $H_1^* = \text{length}(\bigcup_{i=1}^{k} h_i^{(1)})$. That is, if they are disjoint, the lengths are summed; if they overlap, the length is based on the combined sequence. Then the hit fraction is $\omega_1 = H_1^*/S_1$ for sequence 1 and $\omega_2 = H_2^*/S_2$ for sequence 2. The hit fraction gives an indication of how much overall similarity is present between the input sequences. Our indices are similar in concept to the coverage parameters in BLASTCLUST (Dondoshansky, 2002) but differ in that we take the set union from potentially many separate hits. In addition, BLASTCLUST does not provide access to all of the parameters available when using BLAST alone.

Clustering was implemented in two rounds. In the first round, we clustered sequences together if ω_1, ω_2, and $\phi \geq 0.90$. The cutoff value of 0.90 was determined through extensive experimentation with the goal of creating small "well-behaved" clusters that are nearly identical in length and suitable for subsequent alignment by global multiple alignment programs. These "primary" clusters were retained only if they were potentially phylogenetically informative, having four or more species. Because taxon names are notoriously difficult to standardize (Page, 2005), we relied on the taxon identification numbers assigned by NCBI. If the sequence was given a taxon ID that corresponded to an infraspecific taxon, we used the species ID as given in the NCBI taxonomy to determine phylogenetic informativeness.

Sequences within these primary clusters were aligned using Clustal W 1.83 (Thompson et al., 1994). Manual adjustment was necessary for nearly all of the clusters and was conducted in a variety of programs including Se-Al (Rambaut, 1996), Seaview (Galtier et al., 1996), and MacClade (Maddison and Maddison, 2005).

To push the limits of alignment protocols, we also undertook a secondary round of clustering and alignment. To determine which of the primary clusters could be assembled further, we sampled three sequences from each primary cluster and compared this collection of sampled sequences using BLAST as described above. For each pair of sequences, if ω_1 and $\omega_2 \geq 0.2$ and $\phi \geq 0.25$, the sequences were assembled into a cluster. By relaxing the cutoff values in this second round, we linked sequences that overlapped less and/or had greater sequence divergence than in the first round. These "sample" clusters, made by linking samples from the primary clusters, were aligned using DIALIGN (Morgenstern, 1999), a program designed for local (as opposed to global) multiple sequence alignment. The alignments were then edited manually, an important step because DIALIGN can leave several bases unaligned (indicated in the alignment as lower case letters). As in the primary alignments, our manual editing was conducted as conservatively as possible, bringing into alignment only those bases that clearly matched, leaving many gaps when necessary to reduce the possibility of incorrect alignment. We "zipped" together the aligned primary clusters according to the alignments of the sample clusters, producing "secondary" clusters in which the relative alignments of sequences within each primary cluster were not disturbed (similar to profile alignments in Clustal W).

To explore the tradeoff between alignment accuracy and data density, we constructed supermatrices in two

ways (Fig. 2). A sparse supermatrix was constructed from the primary clusters only, skipping the second round of assembly and alignment. Because the primary clusters are sets that are similar in both sequence and length, we expect that the alignment accuracy is highest within these clusters. However, not aligning across the clusters reduces data density because each cluster is treated as a separate "gene," and more missing data are required in the final concatenated matrix.

A "dense" (relatively speaking) supermatrix was constructed by combining the secondary clusters. Because these clusters contain a much more heterogeneous set of sequences, both in terms of sequence divergence and length variation, alignment accuracy may well be pushed to its limits, but the density of data is higher because more sequences from more taxa are placed into homology with one another. Rather than having 41 separate ITS clusters in separate alignments, for example, the dense supermatrix places 37 of them into one multiple alignment.

Orthology.—Prior to final supermatrix assembly, however, primary clusters were examined to see if they contain gene duplications, because multigene families cannot easily be analyzed together with single-copy genes using standard tree-building algorithms (Maddison, 1997) (although see Simmons et al., 2000). To each cluster containing more than one sequence accession per species we applied a phylogenetic test of orthology (modified from Sanderson et al., 2003) in which we used the KH test for parsimony (Kishino and Hasegawa, 1989; Swofford et al., 1996) as implemented in PAUP* (Swofford, 2003) to compare the results from a parsimony search with and without the species constrained to be monophyletic. If the result was significant at the $\alpha = 0.05$ level, we considered the species to be problematic. For the nuclear ribosomal regions and the organellar markers, we removed the problematic species and used the pruned data sets. All nuclear (nonribosomal) data sets that had any problematic species were removed entirely. Our rationale for this is that ribosomal and organellar data, the mainstay of molecular phylogenetics, are assumed to behave as single-copy genes unless proven otherwise. Failing the orthology test typically (with some evident exceptions!) indicated processes such as lineage sorting, hybridization, paraphyly of a species, or simple misidentification of the taxon, and we therefore removed the offending taxon from subsequent analyses. Nuclear sequences, on the other hand, occur in gene families at a much greater frequency (e.g., Arabidopsis Genome Initiative, 2000), and we felt it was safest to regard any evidence of duplication as an indictment of the whole cluster.

Consensus sequences.—Once the duplicated species were pruned from the data sets, we reduced the number of sequences to one per taxon within the primary clusters (sparse matrix) or within the secondary clusters (dense matrix) by making consensus sequences. To construct a consensus sequence, all aligned sequences with the same taxon ID were compared and the majority state selected where they differed, ignoring missing data. In this way, for example, a taxon with separate ITS1 + 5.8 S and 5.8 S + ITS2 sequences would merge into one sequence covering the entire ITS1-5.8S-ITS2 region, using consensus for the region of overlap in the middle. This was motivated not by the goal of resolving disputes between sites but rather to construct the longest sequences possible.

Data set combinability and concatenation.—An important barrier to combining alignments into a supermatrix is the amount of taxon overlap between data sets. If data sets overlap in their taxon sets completely, it is appropriate to consider combining them into a supermatrix. If, however, they share no taxa, then no analysis will produce new information when those data sets are combined, and it is appropriate to keep them separate. The risk of combining such data sets can be extreme, depending on the analysis. For example, if two data sets with no taxonomic overlap are concatenated in a supermatrix and analyzed with parsimony, the strict consensus of the resulting set of trees will be completely collapsed. At best, a maximum agreement subtree may pull out isolated relationships *within* each of the data sets, but no information about relationships *between* data sets can be obtained (Ané et al., 2006). Elsewhere (Driskell et al., 2004; Ané et al., 2006), we have termed collections of minimally combinable data sets "groves." Identification of groves in extreme cases is straightforward: if all taxon sets are equal or are subsets of one another, the entire collection forms a single grove; if all taxon sets are disjoint, each data set is in a separate grove. Intermediate cases such as overlap of a single taxon can be combinable or not, depending on the pattern of overlap (Ané et al., 2006). Therefore, taking a relatively conservative approach to the question of minimal taxon overlap, we required that data sets share at least four taxa to be considered combinable. For each analysis (sparse and dense), the largest set of combinable data sets was then concatenated, adding missing data (as question marks) where necessary to fill out the matrix.

Phylogenetic Analyses

Parsimony analyses of each supermatrix were conducted in PAUP* (Swofford, 2003). We started with simple-addition-sequence or random-addition-sequence starting trees and variously limited the searches to keep only one tree or to keep only 10 to 50 trees that were longer than the shortest known length, with various additional limits on time from one to 150 hours. Gaps were treated as missing data. We also used Perl scripts to implement the parsimony ratchet (Nixon, 1999) directly in PAUP*. The ratchet was set up such that after 30 min of searching on the original matrix, the tree was saved, 10% of the characters were randomly selected and given weight of two, the previous tree was used to start a new search, the best tree from this was saved after 15 min, all characters were given weight one, and the search continued for 200 iterations. Five separate runs were started with simple-addition-sequence trees and another five were started with random-addition-sequence trees, and each was limited to one tree. The tree or trees with the shortest length from all analyses

was used as starting trees for searches that were limited to 5000 trees and were not allowed to swap to completion.

Tree Assessment

Clade support was estimated using the nonparametric bootstrap (Felsenstein, 1985) as implemented in PAUP* (Swofford, 2003). Analyses of 100 pseudoreplicates (92 pseudoreplicates for the sparse matrix) were started with simple addition-sequence trees, and the heuristic searches were limited to 10 trees and 4 h each (dense only) or 1 tree and 24 h each (both matrices). These values were chosen based on extensive experimentation with tree search protocols. We also used a reduced consensus procedure outlined by Wilkinson (1996) to look for support in subtrees of our data that might be lost in the majority-rule consensus of the entire data set. Conventional consensus techniques, including majority-rule consensus, are inadequate for summarizing common information across collections of large trees; they are too sensitive to "rogue" taxa, which are a significant issue in the partially fragmented supermatrices we constructed. (Obtaining maximum agreement subtrees is an additional approach, but we found this problematic because of a bug in PAUP* v4.0b10 that prevented calculation of agreement subtrees for many inputs.) Instead of searching for largest reduced consensus trees in a systematic fashion, we examined the reduced consensus in the most recent and comprehensive phylogenetic analysis across legumes (Wojciechowski et al., 2004), which is an analysis of 330 taxa, 226 of which are found in our tree (225 for the sparse matrix). We took the entire bootstrap profile of trees from our analyses and pruned them down to the taxa found in common between our study and that of Wojciechowski et al. (2004), then constructed the majority-rule consensus of the resulting profile. This gives an assessment of subtree support embedded in the larger analysis—essentially a projection of the bootstrap support from the large trees onto the taxa present in the smaller tree. Note that this is entirely different from merely restricting the phylogenetic *analyses* of the large data set to the smaller set of taxa; instead, it extracts the information present in the analysis of *all* taxa that is relevant to the smaller taxon set.

Finally, we examined the correspondence of the trees with generic delimitations. For each genus that contained more than one species sampled in the data set, we assessed whether the genus was monophyletic or, if not, into how many disjoint maximal clades it was segregated. We did this not because we expect all genera to be monophyletic, but rather that the average level of nonmonophyly provides some sense of the tree's correspondence with conventional taxonomy and may be useful in providing some comparative sense of quality between the two strategies we examined.

All analyses were conducted on a dual Xeon 2.80-GHz CPU with 3 GB of RAM or on a 35-node Linux cluster, in which the head node is a dual Xeon 2.66-GHz CPU with 3 GB RAM and each node is a dual AMD 1.4-GHz CPU with 1 GB RAM.

RESULTS

Assembly, Alignment, and Supermatrix Construction

The GenBank gbpln flat files contained 523,094 records, 17,527 from Papilionoideae. Of these, 4182 were not of molecule type DNA, 1464 were longer than 5 kb, leaving 11,881 that were accepted for further analysis, representing 2416 taxa (including subspecific taxa). The first round of clustering, using stringent requirements for ω_1, ω_2, and ϕ, produced 3893 primary clusters. Of these, 3261 contained only a single sequence, and 131 contained four or more species. Together, these 131 phylogenetically informative clusters contained 6416 sequences and 2264 taxa (54% and 94% of the original sequences and taxa, respectively).

Primary cluster sizes ranged from 4 to 365 sequences. Following alignment with Clustal W and manual editing, the primary clusters had a mean density (proportion of total cells without gaps) of 91%. Notable exceptions included a cluster consisting of 308 sequences from the chloroplast tRNA-Leu(UAA) gene intron, which was replete with gaps and had a density of only 58%. Most of the primary clusters required relatively little manual editing, although areas of tandem repeats frequently required adjustment. None required substantial shifts, indicating that the large-scale homologies were assessed correctly (as opposed to what we had seen when using Clustal W with clusters that had not been screened using ω_1, ω_2, and ϕ) (Fig. 1).

Primary clusters were classified into five broad categories based on GenBank annotation: mitochondrial (3 clusters), chloroplast (41), nuclear ribosomal (53), other nuclear (32), and transposable elements (2). Testing each cluster for orthology by examining monophyly of multiple sequences within species (for those species that had two or more sequences), we found that the results differed greatly between the nuclear markers and the chloroplast + ribosomal markers (Table 1). Across the nuclear (nonribosomal) clusters, an average of 33% of the duplicated species failed the test of orthology. We accepted only those clusters in which all species passed (or there were no duplicated species). Across the chloroplast and ribosomal clusters, average failure rate was 7.5% and 7.4%, respectively. For these, we used the clusters after removing any species that failed. The mitochondrial clusters were small, contained no duplicated species, and we accepted them all. Both transposable element clusters consisted of sequences from Ty1-copia, a retrotransposon generally found in highly heterogeneous superfamilies within plant genomes (Kumar and Bennetzen, 1999). Therefore, we excluded both of the transposable element clusters.

"Sparse" supermatrix.—After removal of problematic clusters or species from within the clusters, 117 orthologous clusters remained. Based on the criterion of overlap of four or more taxa, the largest "grove" contained 72 clusters, one grove contained 12 clusters, three groves had 3 clusters each, and the remaining 24 clusters were isolated. The grove with the most clusters also had the most taxa, and this set of clusters was used to make the

TABLE 1. Summary statistics for primary clusters.

Source[a]	Clustering and alignment Primary clusters[b]	Orthology/paralogy detection: Clusters with duplicated species	Average proportion duplicated species	Average proportion failed monophyly	"Orthologous" clusters[c]
Chloroplast	41	23	0.12	0.075	41
Mitochondrion	3	0	0	0	3
Nuclear	32	27	0.36	0.330	20
Nuclear-ribosomal	53	38	0.21	0.074	53
Transposable elements	2	0	0	0	0
Total	131	88			117

[a]Source genomes were determined from annotations after all data were assembled and aligned.
[b]Only those clusters with four or more species were included.
[c]For organellar and nuclear ribosomal data, these included all clusters for which four or more species remained after removing any species that failed the orthology test. For nuclear data, these included only those clusters in which there were no species that failed the orthology test.

sparse supermatrix by concatenation. It contained 1794 taxa (1688 species) and 53,977 characters, representing 4437 of the original sequences. The matrix was 2.7% filled (96.6% missing data, 0.7% gaps). Table 2 summarizes the supermatrix assembly.

"Dense" supermatrix.—For the dense supermatrix, the second round of assembling and aligning started with the 117 orthologous clusters, uniting them into 47 larger secondary clusters, with most of the linking occurring among the chloroplast and ribosomal clusters (Table 2). The largest sample data set had 37 sequences (i.e., 37 primary clusters shared some sequence similarity with each other). The program DIALIGN aligned these sample datasets quite well with respect to overall structural homology; e.g., the ITS1, ITS2, and whole-ITS sequences appeared to be in their proper positions. However, many gaps were inserted by the program to accommodate high levels of sequence divergence among the samples. Manual adjustment of the alignments was minimal, partly because there was considerable variation on a fine scale within the alignments and manual editing would not improve the alignments significantly. The final data sets (i.e., the secondary clusters, formed by combining the primary clusters according to the alignments within the sample data sets) had, on average, 12.7% missing data (gaps).

Evaluating the secondary clusters' combinability (with respect to taxonomic overlap), we found that all 47 shared at least one taxon with at least one other data set. Two small data sets shared only one taxon with any other data set, one data set shared at most two, and five shared at most three taxa. Therefore, requiring a minimum of four shared taxa resulted in eight isolated data sets and a large grove of 39 data sets. The large grove was used to construct the dense supermatrix by concatenating all sequences in these clusters. The final "dense" supermatrix contained 2228 taxa (2102 species, and 93% of the taxa in the original download) and 33,168 characters, representing 5615 of the original sequences. The matrix was 4.3% filled (89.0% missing data, 6.7% gaps). See Table 2 for summary statistics on the assembly of the dense supermatrix. A list of all loci contained in this supermatrix is presented in Table 3.

Both supermatrices are available as supplemental material (http://systematicbiology.org) or from our website (http://ginger.ucdavis.edu).

Phylogenetic Analysis and Tree Assessment

Several CPU-months of heuristic searches were employed. Significant differences in the behavior of heuristic searches in the two matrices necessitated considerable experimentation with search strategies. We report only the most salient results here.

The sparse supermatrix had 9429 parsimony-informative characters, and the shortest trees found had

TABLE 2. Primary and secondary clusters combined in the "sparse" and "dense" supermatrices.

Source[a]	Sparse supermatrix assembly: Clusters in large grove[b]	Species[c]	Taxa[d]	Sequences[e]	Dense supermatrix assembly: Secondary clusters	Clusters in large grove[b]	Species[c]	Taxa[d]	Sequences[e]
Chloroplast	27	989	1026	1625	13	13	1221	1263	2113
Mitochondrion	2	12	12	12	3	3	17	17	17
Nuclear	13	63	69	196	17	12	72	78	260
Nuclear-ribosomal	30	1259	1339	2604	14	11	1636	1731	3225
Total	72	1688	1794	4437	47	39	2102	2228	5615

[a]As in Table 1.
[b]The large grove was the largest set of clusters that were linked by sharing at least four taxa.
[c]Species were determined using the NCBI taxonomy tree.
[d]Including species, subspecies, and varieties.
[e]Sequences represented in the final supermatrix as consensus sequences.

TABLE 3. Secondary clusters combined in the "dense" supermatrix.

"Gene"[a]	Taxa	Sequences	Clusters
ITS	1648	2888	37
trnL	705	852	14
matK	517	561	10
rbcL	260	274	1
ETS	92	92	1
ETS	90	98	3
ndhF	89	92	2
psbA-trnH	79	79	3
trnS-trnG	51	55	1
atpB-rbcL	35	49	4
rps16	34	34	1
18S	26	26	1
ITS1	24	31	1
trnT-L	22	27	1
ITS2	21	27	1
trnL	17	26	1
legcyc 1	16	44	2
phyE	16	16	2
H3	14	18	2
ETS	13	14	1
rps3-rpl16	13	24	1
trnS-trnG	12	18	1
rps11-rpl36	12	22	1
lec1	9	14	1
26S	9	9	2
ITS1	9	9	1
cox2	8	8	1
rga	7	30	1
bbi	7	16	1
5S	7	14	1
Sat5	7	13	1
shst2	6	7	1
SGlu2	6	25	1
ITS1	5	17	1
B-III	5	7	1
nadh4	5	5	1
nad3-rps12	4	4	1
aai2	4	58	1
rga	4	12	1

[a]Gene (or sequence region) names were determined from record annotations *after* data assembly and alignment of secondary clusters. See Supplemental Material (available online at http://systematicbiology.org) for more complete names.

length 42,816. The shortest score was found in a random-addition-sequence run in which the search was limited to five trees and 48 h. The second best score (42,837) was found during 1 of 10 parsimony ratchet runs (each taking 150 h), in which the resulting trees ranged up to 43,356 steps at worst, averaging 302 steps longer than the best. Across the random-addition-sequence runs, results also ranged widely up to 43,432, averaging 307 steps longer than the best tree. The simple-addition-sequence search (limited to one tree) swapped to completion in 61 h but found a score of 42,878. In general, tree search progress was slower and more variable in the sparse matrix than in the dense matrix.

The dense supermatrix had 7199 parsimony-informative characters. The shortest score, 56,093 steps, was found during one of the parsimony ratchet replicates. Across 10 parsimony ratchet runs, each taking 150 h, trees were found with lengths from 56,093 to 56,662, with an average result of 140 steps longer than the shortest tree. The second best score (56,095) was found using a heuristic search without the ratchet, starting with a simple-addition-sequence starting tree and limited to one tree. This search swapped to completion in 25.5 h. Ten searches with random-addition-sequence starting trees each swapped to completion on average within 23 h, when limited to one tree, and found trees that had lengths from 56,133 to 56,521, averaging 165 steps longer than the shortest known trees. Additional search efforts were less successful: searches in which we retained more than a few trees never swapped to completion and failed to make much progress in finding shorter trees.

Using the one or few trees with the best scores as starting trees, we expanded the sets to 5000 equally parsimonious trees (available as supplemental material [http://systematicbiology.org] and on our website [http://ginger.ucdavis.edu]). Strict consensus trees of these tree sets had fair levels of resolution: for the sparse matrix, the strict consensus had 938 nodes (out of a possible 1793), and the dense strict consensus tree had 1292 nodes (out of a possible 2227). The 50% majority rule trees were much more resolved: 1775 nodes and 2189 nodes for sparse and dense, respectively. Most values in the majority rule trees were very high (e.g., for the dense matrix, the 99% majority rule consensus had 2129 nodes, so only 60 nodes had values between 50% and 98%), indicating that the 5000 equally parsimonious trees were extremely similar.

The overall structure of the strict consensus tree was evaluated by locating, where possible, clades that corresponded to those recognized in the legume phylogenetics literature (Wojciechowski et al., 2004; Lewis et al., 2005). Many large recognized clades were at least approximately recovered in the analysis of the dense matrix (Fig. 4), although in several cases a small number of taxa were placed outside of the clades in which they likely belong. It was far more difficult to locate, even approximately, the previously recognized clades in the strict consensus tree made from the sparse matrix (tree not shown).

Bootstrap analyses of the dense supermatrix were conducted first by limiting each of 100 replicates to 4 h and then by limiting each to 24 h. In the first analysis, the bootstrap consensus had 1079 nodes (with bootstrap proportion [BP] > 50%), of which 571 were well supported (BP ≥80%). Adding time to the searches affected the results generally by improving the bootstrap scores. In the second analysis, 1117 nodes were resolved (BP >50%) and 623 were well supported (BP ≥80%). The bootstrap (92 replicates) on the sparse matrix had only 715 nodes with greater than 50% support (359 nodes with BP ≥ 80%) and a vast basal polytomy containing 910 lineages.

In the reduced bootstrap consensus analysis there were very clear differences between the sparse and dense analyses. The pruned bootstrap majority rule tree for the sparse matrix was highly unresolved. Of the six sizable "major" clades (having more than ten taxa) highlighted in the *matK* tree, only the Mirbelioids was supported with BP > 50%. The pruned bootstrap analysis of the dense matrix (Fig. 5), however, was quite well resolved and supported *all* 12 of the major clades recognized by

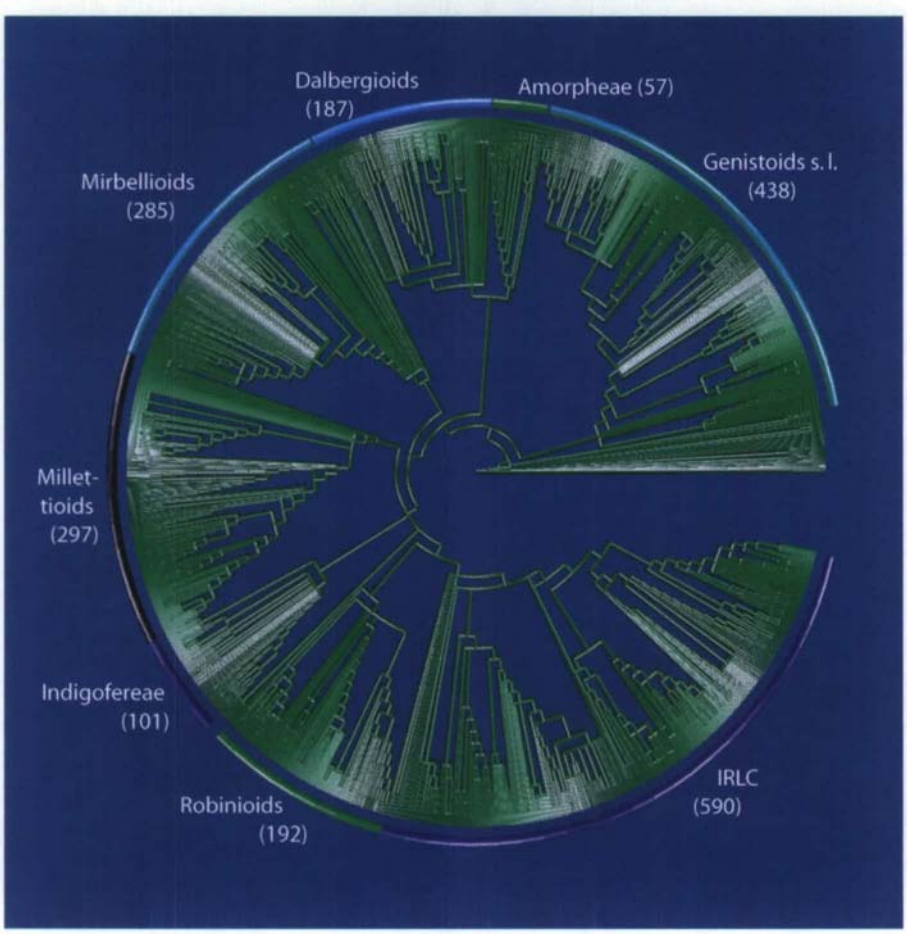

FIGURE 4. Strict consensus of 5000 equally parsimonious trees based on the dense supermatrix. Tree of 2228 taxa of papilionoid legumes is displayed using the program Paloverde (Sanderson, 2006); branches are partially white for visual effect only. Terminal taxon names have been removed, but arcs around circumference show eight large named clades used in the literature of legume phylogenetics (Wojciechowski et al., 2004). Numbers in parentheses are the number of terminal taxa in named clade. The tree is rooted at a basal polytomy that includes all members of the Swartzioid clade (*Swartzia*, *Bobgunnia*, *Ateleia*, *Cyathostegia*, and *Bocoa* Wojciechowski et al., 2004) and two other lineages. IRLC defined in text.

Wojciechowski et al. (2004), with the single exception of the misplacement of *Ormosia* outside of one of those clades.

A final indication of the relative quality of the two trees was obtained by examining the degree of nonmonophyly of the genera that had at least two sampled species in these analyses (Table 4). Genera were broken into more disjoint clades in the sparse analysis than in the dense analysis, reflecting in part some level of failure in assembly and phylogenetic reconstruction. Of course, it is unclear what fraction of these genera is truly monophyletic (it is surely less than 100%), but the highly dispersed position of disjoint pieces of many genera in the sparse tree suggests that "real" nonmonophyly of genera is only part of the story.

DISCUSSION

This paper reports results from a series of computationally intensive analyses of ~12,000 sequences and ~2200 species, involving numerous data processing steps (Fig. 2). The complexity of the methodology necessitates some further discussion. First, we discuss the phylogenetic results in the context of present knowledge of papilionoid legume relationships to explore the credibility of the supermatrices and their resulting phylogenies. Second, we discuss in some detail a long list of issues and problems that these analyses revealed about phyloinformatics approaches to supermatrix construction, some of which will require substantial additional research to obtain workable solutions. Finally, we compare our approach to others that have been taken and

TABLE 4. Nonmonophyly of genera.

	Sparse analysis	Dense analysis
Number of generic names	320	345
Number of genera with more than one species in data set ("nonmonotypes"[a])	121	127
Number of origins of clades in these "nonmonotype" genera	525 (4.33:1)	468 (3.68:1)
Number of origins of clades in most fragmented genus	35 (*Astragalus*)	22 (*Genista*)

[a]These are monotypic genera with respect to the data set only. Some have additional species that were not in the data set.

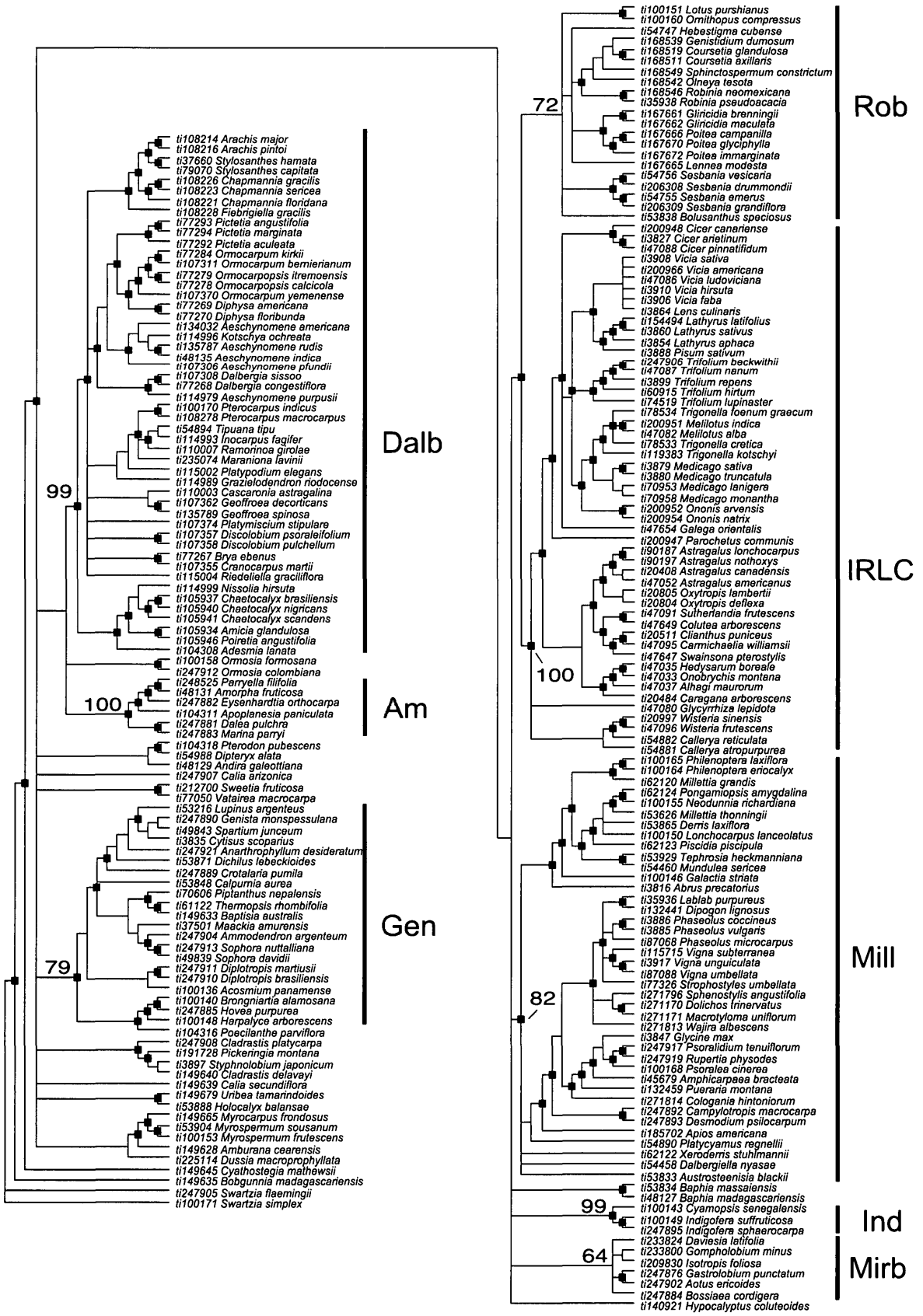

FIGURE 5. Reduced bootstrap consensus tree (Wilkinson, 1996) based on bootstrap analysis of the dense supermatrix, followed by pruning resulting trees to 226 taxa found in both the dense supermatrix and the *matK* tree of Wojciechowski et al. (2004). Clades with black boxes are supported at BP >75%. Eight named clades of Figure 4 are shown with their exact bootstrap proportion (Am = Amorpheae; Dalb = Dalbergioids; Gen = Genistoids s. l.; Ind = Indigofereae; IRLC = inverted-repeat-lacking clade; Mill = Millettioids; Mirb = Mirbelioids; Rob = Robinioids).

end with some speculation about the prospects for semi-automated supermatrix construction.

The Phylogeny of Papilionoid Legumes

Prior to this analysis the state of Papilionoid phylogenetics was a mixture of robust results at various hierarchical levels but little integration across levels. A backbone phylogeny of several hundred genera, based on a single highly informative plastid gene, *matK*, had been constructed (Wojciechowski et al., 2004) and largely accepted by the community of legume systematists (Lewis et al., 2005; building on previous work based on other genes; e.g., *rbcL*: Kajita et al., 2001; *trnL*: Pennington et al., 2000; Lavin et al., 2001). Complementing this were a large number of intra- and intergeneric studies based on nuclear and plastid spacers, also single-gene analyses, and also in many cases fairly well supported. However, no comprehensive multigene studies across the clade had been undertaken, and no "high-resolution" study attempting to bring together the taxon samples already available in the literature had been done with the exception of a supertree analysis of one large clade of papilionoids, the IRLC (inverted-repeat-lacking clade; see our Fig. 4; Wojciechowski et al., 2000).

The results of our two supermatrix construction strategies were mixed, with the sparse analysis performing relatively poorly compared to the dense analysis. The sparse analysis produced a strict consensus tree with numerous anomalies, including a large wastebasket clade in one part of the tree containing a hodgepodge of taxa from dispersed clades identified in previous papilionoid studies. This probably stems from the structure of overlap—or lack of overlap—between blocks of data in the supermatrix, a general phenomenon discussed further below. It may also be accounted for by failure of heuristic searches to sort out relationships under these conditions when numerous local rearrangements are equally parsimonious.

On the other hand, the structure of the strict consensus tree from the dense supermatrix is much less anomalous (Fig. 4). First, it lacks pathological subtrees found in the sparse analysis. Second, it recovers the essential elements of the well-supported previous studies of the backbone papilionoid phylogeny (Wojciechowski et al., 2004), while adding in many genera and a very large number of species (~1900) not sampled in that study. The reduced bootstrap tree (Fig. 5) is very similar to the backbone *matK* tree. The composition of the eight named clades indicated there and in Figure 4 accords well with that given in Lewis et al. (2005), and their relationships to each other agree in the main with the strongly supported relationships indicated in the *matK* analysis (Wojciechowski et al., 2004). Finally, the strict consensus of the full taxon set recovers important elements of phylogenies of lower level groups while placing them in phylogenetic context to each other and the larger scale tree of papilionoids. The sparse analysis generally broke up genera to a greater degree than did the dense analysis (Table 4). Some of these instances must be due to true nonmonophyly, but some are obvious artifacts.

A tale of two clades: Amorpheae and the Astragalean clade.—Rather than attempt to discuss the entire tree in detail, we comment briefly on the two clades with which the authors have significant previous experience, Amorpheae and the Astragalean clade. Amorpheae is a moderately small (ca. eight genera and 250 species) clade for which McMahon and Hufford (2004) inferred a well-supported phylogeny, which agrees very closely with the results of our dense analysis. This is not surprising given that almost all of the data concerning these taxa were in fact collected by these authors, but it is reassuring that these strong results were not "lost" in the process of data assembly, at least in the dense analysis. The sparse analysis recovered the same relationships as the dense analysis among species of Amorpheae, but 38 unrelated taxa were placed within *Dalea*, a phenomenon similar to the large wastebasket clade described above.

The Astragalean clade is a species-rich group of temperate herbaceous plants dominated by the huge angiosperm genus, *Astragalus*, and includes about 15 other genera. Its phylogeny has been reconstructed using chloroplast RFLP data and sequence data from nrITS, *matK*, and the chloroplast *trnL* intron (reviewed in Wojciechowski et al., 2000), using both conventional single gene matrix reconstructions and supertree approaches (Wojciechowski et al., 2000). Within *Astragalus*, cytogenetic evidence (Spellenberg, 1976) and all molecular phylogenetic work supported the monophyly of a New World aneuploid group, called Neo-Astragalus (Sanderson and Doyle, 1993; Liston and Wheeler, 1994; Wojciechowski et al., 1993, 1999).

Results from the dense supermatrix analysis agree in most respects with this emerging picture of the phylogeny of the Astragalean clade. As with earlier analysis, a "Coluteoid" clade, including the *Astragalus sinicus* group, the large genus *Oxytropis*, and the remaining *Astragalus* each form clades. A gratifying result was the placement of a number of recently deposited sequences of Asian *Astragalus*, none of which had previously been integrated into a broad phylogenetic analysis of the Astragalean clade.

The one significant anomaly in the dense supermatrix with respect to the Astragalean clade is the polyphyly of Neo-Astragalus. Examination of the alignments revealed the source of this result. One collection of Neo-Astragalus taxa represented by separate ITS1 and 2 sequences, which had clustered into two separate primary clusters, did not properly align to a cluster consisting of other sequences of Neo-Astragalus represented by a complete ITS1-5.8S-ITS2 primary cluster. Although the separate primary alignments constructed using global alignment via Clustal W appeared fine, the product of the DIALIGN/zipped local alignment procedure, the secondary cluster, was not. Inspection of the alignment revealed several false synapomorphies owing to an obvious shift among the blocks of sequences coming from the original primary clusters. When this alignment (pruned to just *Astragalus* exemplars) was

re-edited by eye, the monophyly of Neo-Astragalus was once again supported.

As with Amorpheae, results for the sparse supermatrix were more problematic. One indication is the fragmentation of many large genera. In the dense analysis, *Astragalus* was dominated by one very large clade, one small clade (the *A. sinicus* group) and three isolated terminal taxa. In the sparse analysis, *Astragalus* was broken into 35 different lineages, including 14 clades and 21 isolated terminals.

Methodological Issues

Errors in GenBank accessions.—A mundane but remarkably significant source of problems in our analyses were various mistakes in GenBank sequence accessions. Previous work had indicated a fairly high rate of mistaken gene annotation, so we relied entirely on similarity search algorithms like BLAST in the clustering stage. One of the most surprising examples of misannotation was a set of five ITS sequences that had been entered into the database reverse complemented (several *Crotalaria* and *Piptanthus*, e.g., gi:14716940). Default BLAST settings detect these as homologies, which initially were shunted off to alignment programs, causing no end of problems. The fix was simple, but the example illustrates the perils of relying on annotations.

Unfortunately, many mistakes involve incorrect identifications of species, for which there is no easy solution. First indication of these often occurred in the orthology detection step, when two accessions from the same species for a presumptively single-copy gene might appear to be phylogenetically distantly related and thus classified as paralogs. The paraphyly might be real, of course, either because the species is paraphyletic, or there is lineage sorting, hybridization, or introgression. However, the orthology test was set up fairly stringently and let pass cases where accessions from the same species were "close" to monophyly, failing only those in which accessions were removed by several branches from each other. Up to about 7% of species containing multiple sequences of plastid and nuclear ribosomal sequences failed this test (Table 1), and we interpret most of these as mistaken identifications.

Alignment.—Alignment presented one of the most severe obstacles to supermatrix construction in the context of the heterogeneous nucleotide data found in GenBank. The fundamental difference between our phyloinformatics approach and traditional molecular phylogenetic studies is that the latter are often composed of data gathered by one group of investigators targeting one or more homologous regions of approximately the same length (researchers using the same PCR primer pairs, for example). This generates sequences that can be aligned well by robust "global" multiple sequence alignment methods such as the progressive alignment used in Clustal W (Thompson et al., 1994). When significant sequence divergence is combined with considerable sequence length heterogeneity, however, global alignment algorithms begin to fail in spectacular fashion (Lassmann and Sonnhammer, 2002, 2005a; see our Fig. 1), and "local" multiple sequence alignment methods are necessary (e.g., Subramanian et al., 2005). Unfortunately, although several have been developed in the last decade, none are as proven in their own domain as global methods. Our experience with several methods, including DIALIGN (Morgenstern, 1999; Subramanian et al., 2005), POA (Grasso and Lee, 2004), and others, suggested that none was simultaneously capable of handling the level of length variation, sequence divergence, and large number of sequences such as was found in the 1700+ ITS sequences for papilionoid legumes. Their only virtue was that they consistently outperformed global alignment procedures.

In response we developed a heuristic approach reminiscent of seed alignment strategies used to identify large protein clusters from databases (e.g., Pfam: Sonnhammer et al., 1997). This constructed blocks of highly homologous and length-homogeneous sequences—the primary clusters—took small samples from these, and then attempted to use local alignment to piece these blocks together based on the samples. Overall, because of the generally superior outcome obtained from the dense analysis using this heuristic local alignment strategy, we conclude that this approach is promising. It enabled, for example, the assembly of a very large ITS alignment, a large set of data that was not brought together in the sparse analysis when the primary clusters were kept separate. However, a glance at some of these alignments reveals that some are very "gappy." DIALIGN adds a large number of gaps into noncoding sequence alignments in an attempt to conservatively avoid bringing nonhomologous bases into the same column.

Not surprisingly, there were clear cases in which this strategy simply did not work—where DIALIGN simply went too far—such as with the apparent polyphyly of Neo-Astragalus described earlier. In attempting to combine separate and partial ITS 1 and 2 regions with complete sequences, the combination of sampling exemplars from the primary alignments and DIALIGN's inability to correctly align these exemplars generated clearly faulty regions of poor alignment that were easily diagnosed by eye (once the visual clutter was reduced by removing a large number of sequences). As usual, the consequence of shifts among blocks of sequences in an alignment was to introduce false synapomorphies into the phylogenetic analysis, driving apart a formerly well-supported clade. This was almost immediately recognizable because the new clades were comprised of the same class of sequence, such as ITS 2 only.

Quality assessment of multiple sequence alignments is a difficult problem when the underlying homology is unknown (Pei and Grishin, 2001; Thompson et al., 2001; Lassmann and Sonnhammer 2005a). Approaches include the identification of unambiguously aligned blocks (GBlocks: Castresana, 2000), or, similarly, relatively conserved columns (Al2co: Pei and Grishan, 2001). Objective functions have been developed to allow comparison of alignments produced by various parameter sets or alignment algorithms (e.g., Notredame et al., 1998; Thompson

et al., 2001). Another promising tactic is to compare across alignments, assuming that the most frequently found homologies are indeed correct (Lassmann and Sonnhammer, 2005a). This approach also incorporates the observation that the performance of different alignment techniques is data set dependent (Lassmann and Sonnhammer, 2005a). However, assessment, like alignment itself, is most well studied for data sets that differ significantly from ours in two important respects. First, many of these techniques are aimed specifically at successful protein identification, and second, benchmark data sets on which the techniques are tested do not have the severe length heterogeneity found in many of ours. We find it encouraging that, for the sample data set depicted in Figure 1, the cross-alignment scoring system in MUMSA (Lassmann and Sonnhammer, 2005a) ranked the programs similarly to our preliminary assessment. However, more work is needed to fully understand the effects of highly length-heterogeneous nucleotide data on multiple sequence alignment.

Data set structure and fragmentation.—Although it was first described in the context of supertree construction (Sanderson et al., 1998), an important issue for supermatrix construction is the structural relationship among different blocks of data in the final matrix. For supertrees the issue is how much taxon overlap is necessary in the input trees for there to be new inferences about relationships not found in any of the input trees, a kind of "cross-talk" among the input trees. In a supermatrix, the same issue arises. Consider a simple example: a matrix in which there are sequence data for gene 1 in taxa A-D and for gene 2 in taxa E-H, but nothing else. The supermatrix looks like two nonoverlapping blocks of data in two of the corners with blocks of questions marks in the other two corners. A parsimony analysis can correctly infer relationships within the blocks but not between them (assuming sufficiently informative data within the blocks). However, in the collection of equally parsimonious trees based on the whole matrix, all possible cross-relationships will appear and the strict, semistrict, and majority-rule consensus trees will be unresolved. The information from the separate blocks is still buried in the collection of trees and can be retrieved using maximum agreement subtrees, for example, but it is obscured.

Even *some* shared information between blocks may not be enough to overcome this problem. In the supertree case, it is insufficient for two trees to share one taxon in common and expect any cross-relationships to be inferred. These considerations led to the idea (Sanderson et al., 1998) that an overlap of two was necessary, a criterion used in one empirical study (Driskell et al., 2004). Recent work has demonstrated for supertrees (Ané et al., 2006) that this is probably sufficient but not always necessary if more than two input trees are used. The analogous results for supermatrices are not yet known, but our supermatrix construction herein illustrates some of the issues playing an important role in the outcome. We have used the term "grove" for a set of trees that can provide cross-information in a supertree analysis (Ané et al., 2006). In a supermatrix perhaps a better term is a "block."

A telltale sign of the presence of multiple blocks in a supermatrix is the collapse of large parts of the consensus trees due to the presence of certain collections of taxa. This is confounded in large analyses, unfortunately, by the inability of heuristic search strategies to find, or retain, all of the equally parsimonious trees at any stage in the analysis. Thus, a set of rogue taxa that really can be placed anywhere in the tree as long as their relationships to each other are preserved may or may not be evident if key trees have simply not been stored. In suspect cases, one can check by manually moving the suspected rogue taxa around the tree (preserving their relationships to each other), and seeing if the parsimony score remains the same. If it does, there is evidence that this set of taxa forms a block isolated from others.

In the sparse analysis many of the problematic results, such as the large "wastebasket clade," can be attributed to the lack of overlap structure between the primary clusters, many of which appear to be acting as independent blocks. This occurred despite efforts to ensure a minimum level of overlap between clusters. Every primary and secondary cluster was required to overlap with at least one other by four taxa, meaning that between, say, clusters of sequence from genes 1 and 2, there were at least four species with *both* sequences. Evidently, this is not enough to prevent some pathological results, which is not entirely surprising because there might be no phylogenetic information in those key sequences even if they are present (because of lack of variation, for example).

The most instructive case of data fragmentation we uncovered involved species of *Glycine* dispersed incorrectly around the dense analysis tree. In this case, issues of insufficient taxonomic overlap between data sets combined with simple mistakes in the GenBank accession to produce substantially incorrect results. At the base of the strict consensus tree from the dense supermatrix, we found two polytomies that involve many taxa thought to derive from basal nodes, as expected, but we also found five species of *Glycine*. *Glycine* has always been considered to be related to other members of the Millettioid clade (Doyle et al., 2003), just where its other 16 species are placed in our tree. So how did these five taxa move into a position so far from the other *Glycine* species? Insufficient tree search effort does not appear to be the cause. To test this, we ran a brief search in which they were constrained to be near the other *Glycine* species, using the rest of the relationships from one of the MP trees as a starting point. This resulted in more steps than in the original MP tree, not less, which we would expect if it was simply incomplete tree searching that placed them near the base. Alignment is also not to blame. The five problematic taxa are each found in six small clusters and the large ITS cluster. The small data sets were easily inspected and their alignments appeared reasonable. Likewise, the ITS data set, when pared down, also appeared to be aligned reasonably for these taxa.

Most data sets that involve the taxa in question contain only *Glycine* (*H3*, *rps11-rpl36*, *rps3-rpl16*; *trnS-trnG* has

one *Neonotonia*, a recent segregate; Lackey, 1977). These data cannot be placing the problem taxa at the base of the tree. Instead, these data will agree with any placement, as long as the relationships within the genus are maintained. Therefore, the potentially misleading data must be within the two data sets that have non-Millettioid taxa: *atpB-rbcL* and ITS. A quick heuristic search and bootstrap of the *atpB-rbcL* data set (35 taxa, 1135 characters) shows strong support for the monophyly of *Glycine* to the exclusion of *Genista* and *Vigna*, inconsistent with the final placement of the five problem species of *Glycine*, so these data are also not likely to be the cause.

The single primary cluster that contains *Glycine stenophita* has sequences only from *Glycine* and one other taxon: *Myrospermum sousanum*. *Myrospermum sousanum* is a member of the Sophoreae, a tribe reasonably expected to be found in the basal grade of the tree. *Myrospermum sousanum* is well represented in the supermatrix (*trnL*, *PhyE*, *matK*, *rbcL*, and ITS), and four of these sequences are most similar to sequences from other Sophoreae. However, the ITS sequence for *M. sousanum* is very similar to *G. stenophita* (BLAST search: 99% identity, 1 gap). This suggests that a *Glycine* ITS sequence has been mistakenly labeled as *M. sousanum* and deposited in GenBank. To test this, we removed this one *M. sousanum* ITS sequence from the supermatrix and ran an abbreviated tree search. The resulting trees restore all *Glycine* species to the Millettioid clade, as expected. Thus, a sequence mislabeled *Myrospermum* combined with limited data overlap between *Glycine* species and other taxa, and strong data overlap between the *Myrospermum* and other taxa, conspired to scatter *Glycine* in "unexpected" directions.

Search strategies in large data sets.—The large size and sparseness of both supermatrices has a strong impact on the running times of heuristic search strategies for obtaining parsimony solutions. Deterministic and randomized heuristics all required upwards of a week of processor time before showing evidence of lack of continued improvement in tree length. No doubt shorter trees can be found with more extended efforts. The parsimony ratchet (Nixon, 1999) found the shortest trees among all searches run on the dense matrix, but this occurred in just one of its replicates. Deterministic heuristics in PAUP* outperformed other ratchet replicates. Of special concern was the likelihood that vast collections of equally parsimonious trees were not found with any of our search strategies. Not only are there a large number of 2000+ taxon trees of similar length that evidently exist, the search procedures did not get around to sampling much of this space in the first place, because all equally parsimonious trees were derived from rearrangements from just one run—the ratchet run in the case of the dense matrix. It would be ideal to approach the same optimality score from different random starting points, but the searches required so much processor time and ended at such suboptimal scores that this was simply not possible.

Improvements in algorithms and tree search heuristics aimed at large data sets have been reported (Ronquist, 1998; Goloboff, 1999; Huson et al., 1999; Ganapathy et al., 2003), and current algorithms have been tested with large simulated data sets (Salamin et al., 2005). However, the fragmented, sparse structure of our supermatrices suggests a need for heuristics tailored to this structure. For example, if a matrix contained two independent blocks, there would be no reason to perform rearrangements of taxa across blocks in trying to find better trees: such rearrangements would always produce new trees with the same parsimony score (the *Glycine* example showed abundant evidence of this). To avoid such computationally expensive mistakes, explicit methods for characterizing this block structure are needed.

Confidence limits in very large trees.—Little is known about how to assess confidence in large phylogenies (Salamin et al., 2003). Several studies have examined the impact of heuristic search effort on confidence limits (Debry and Olmstead, 2000; Mort et al., 2000; Salamin et al., 2003; Müller, 2005), generally concluding that more exhaustive searches increase bootstrap support, although some search strategies seem to exhibit this effect more than others (Müller, 2005). In addition, theory and limited empirical work (Sanderson and Wojciechowski, 2000) suggest that bootstrap confidence levels will decline for a given clade as more taxa are sampled from it. This can happen because the hypothesis becomes increasingly specific relative to the null hypothesis of nonmonophyly—there are increasingly many ways for the collection to be nonmonophyletic versus monophyletic, and monophyly is rejected merely if one taxon drops out of the clade in a majority of the replicates, even if it is a different taxon each replicate. Both of these issues are important in data sets of the size assembled here.

Bootstrap proportions for both sparse and dense supermatrices were quite low across the tree and the majority rule bootstrap trees were highly unresolved. We suspected that although large clades have low bootstrap support, evidence for relationships of pruned (reduced) subtrees might be much higher. The differences between the reduced bootstrap trees and the original ones were striking. For example, although the bootstrap proportion for all 590 species in the IRLC taken together as a clade was <50% on the bootstrap tree from the original dense supermatrix, the support for this clade when reduced to the 51 species in the pruned collection was 100% (Fig. 5). Overall, there was high concordance between the bootstrap levels observed on the reduced tree and those seen in the tree of Wojciechowski et al. (2004). However, this only emerged after heuristic searches were extended to 24 h per replicate. Preliminary, less exhaustive, searches identified fewer clades in the bootstrap majority rule trees with much lower bootstrap proportions, a pattern seen in earlier studies (e.g., Mort et al., 2000) but quite dramatic in our data.

These results are reassuring in the sense that the signal present in subsets of the data is retained in the collection of bootstrap trees derived from the supermatrices, but they do not help to characterize support at that larger scale. The bootstrap proportion just does not scale well, and other hypothesis tests about trees may be more

instructive in large data sets (Sanderson, 1989; Page, 1996).

Alternative Strategies

The most widely used strategy for reconstructing deep (though not necessarily taxon-rich) phylogenies is to fill in a complete multigene supermatrix by targeting a defined list of taxa and loci. An important feature of most of these matrices is that they contain loci that can be aligned. They do not typically include rapidly evolving loci that are useful in small subtrees but are very difficult to align across the tree as a whole. Indeed, it is difficult to imagine any strategy that can produce dense supermatrices when loci of widely varying rates are combined and "aligned." Ironically, even "phylogenomics" approaches exemplify this issue. Analyses based on complete genomes (Rokas et al., 2003; Lerat et al., 2003; Ciccarelli et al., 2006) only combined ~100 to 200 loci or fewer in the same matrix, with alignment problems and lack of homology (e.g., absence of orthologs or high levels of divergence between orthologs) presumably limiting the number of taxa scored as having a particular locus.

Other strategies have been widely discussed. Compartmentalization (Mishler, 1994; Zanis et al., 2003) extracts strongly supported subclades, replacing them with terminal taxa possessing synthesized ancestral states for the clades, but as yet has not been widely used. Supertree methods (Bininda-Emonds, 2004; Wilkinson et al., 2005), which build trees from a collection of smaller trees, have engendered a rich theoretical literature, a not insignificant number of case studies, and considerable controversy, especially as a competitor to supermatrix approaches (Gatesy et al., 2004). Our view on the relative merits of the two competing approaches after this exercise is uncommitted. If anything, the similarities between the issues raised by construction of sparse supermatrices and supertrees suggest that there are important commonalities in the two approaches.

Prospects for Large Supermatrix Construction and Analysis

Algorithmic improvements.—Enumeration of the issues raised above highlights the need for significant improvements to algorithms for several problems. Multiple local sequence alignment stands out as one of the most formidable challenges to effective assembly of phylogenetic data sets at large scales. The problem of assembly of segments of sequence of different lengths and divergence levels has received little attention but is taking on increasing significance in comparative genomics, so advances in this area are to be expected. The appropriate problem to solve in phyloinformatics approaches to tree-building may be slightly different, however, than that for the case when whole genomes are available. There may well be tradeoffs in attempting to assemble data sets that maximize number of taxa, number of sites, or minimize number of missing bases. Moreover, the structure of the data matrix assembled has strong implications on the quality of phylogenetic results that can be obtained, and it may be appropriate to incorporate notions of this structure early in the data assembly process.

Efficient tree-building parsimony heuristics that take account of the increasingly common sparseness of large data sets are also still needed. Given the interest in model-based approaches to phylogenetics, an obvious question is whether and what future these methods have when scaled to the level of thousands of taxa. Problems of long branch attraction, which confound parsimony analysis, undoubtedly will remain in large-scale phylogeny efforts, and overcoming them will be a persistent challenge. However, even model-based methods with promising running time behavior, such as RAxML (Stamatakis et al., 2005), are heuristic mixtures of fast parsimony steps and slow likelihood-based steps, and any such "impure" method may lose the desirable properties of consistency characteristic of likelihood-based approaches (Felsenstein, 2004).

Automated error detection.—One of the most important limitations of any informatics approach is the frequency of undetected errors in the data—a problem with any kind of data, of course, but one that is exacerbated as the familiarity of the investigator with the original data lessens. It should be possible to detect some errors from the data themselves, either because of failure of consistency checks, incongruence, or failures of more elaborate tests such as implemented in our tests for orthology. Nonetheless, there will remain a strong need for trusted annotations attached to the data used in these analyses. The presence of voucher specimen information in the GenBank record, though it does not guarantee a correct taxonomic identification, bespeaks a concern for avoiding mistakes in identification and provides a mechanism for ultimately checking on it. Other kinds of annotation errors, such as those regarding sequence "features" (positions of introns, exons, etc.) are more or less important depending on the reliance of steps in the data-mining pipeline on those annotations. We have deliberately sidestepped almost all of this information, relying on algorithm-driven techniques to assemble and align sequences into data matrices. This is not necessarily the most efficient technique, but it does avoid certain pitfalls due to mistaken annotations. A long-term strategy might be to incorporate feature annotations by passing them through error detection routines of our own devising, analogous to the orthology tests. This approach may connect with efforts to develop phylogenetically driven comparative genomics annotation schemes (e.g., Hughes et al., 2005).

How much automation will be possible?—This raises the general question of how much automation is possible in large-scale tree reconstruction. Alignment procedures implemented here were not sufficiently good to leave completely automated. Numerous—and tedious—manual edits were necessary even in the primary alignments. Our sense is that this will come as no surprise to the molecular phylogenetics community, but it presents a cautionary note for the prospects of removing human intervention from this analysis pipeline entirely. On a more positive note, what can be automated is

extensive, including downloading sequences, running BLAST searches, assembling matrices, constructing first cuts at alignments, checking for orthology, and building supermatrices. There then remains a great deal of heuristic tree construction, which can also be automated, but additionally benefits from the investigator adaptively modifying search procedures in response to performance on the particular data set at hand.

Conclusion: The Shape of Supermatrices to Come

The pace of data acquisition and its easy accessibility raise the possibility of near "real-time" phylogenetic syntheses that span both existing and newly deposited data sets. Systematists beginning a new study of a clade or returning to one studied previously have a strong motivation to understand the existing relevant phylogenetic information contained in the databases. Other biologists who wish to pair their own data with phylogenies to undertake ecological, physiological, or other comparative studies also have a stake in finding (or building) the most comprehensive phylogenetic trees available. Though not all phylogenetically informative data are confined to sequence databases, they are the most accessible source of comparative data for the most taxa in the tree of life at present.

Phyloinformatics methods for the construction of large supermatrices are at an early stage (Driskell et al., 2004; Delsuc et al., 2005; Philippe et al., 2005). This paper extends these methods in the direction of taxonomic richness by including 93% of the species of papilionoid legumes found in GenBank, some 2200 of them. It is not the largest phylogenetic tree ever built (see e.g., Källersjö et al., 1998, Hibbett et al., 2005), but it is unusual in its combination of deep and shallow phylogenetics at such a large scale. Despite significant problems, the resulting phylogeny represents a reasonably good synthesis of available phylogenetic knowledge from molecular sequence data. We find it particularly useful as an indicator of the shape of supermatrices to come: the extraordinary problems of data sampling, alignment, and homology assessment that characterized our analysis at these different scales drive supermatrix construction in the direction of sparseness, both between loci and within alignments (Philippe et al., 2004). The vision of large, complete data matrices with little missing data is just that—a vision. It is approximated best by protein-coding data that are easily aligned, but at low taxonomic levels these data sets are frequently uninformative about relationships, whereas deeper in the tree, many exhibit important structural changes, such as gains, losses, and rearrangements of introns, exons, and entire functional domains (e.g., Reyes et al., 2004; Kim et al., 2006). If the ultimate goal of phylogenetics is construction of large, high-resolution trees, then the problem to be solved is integrating data from very different loci in a single analysis.

Acknowledgments

We thank Marty Wojciechowski, Amy Driskell, Gordon Burleigh, Cécile Ané, Brian O'Meara, Taum Hanlon, and Vincent Savolainen, Mark Simmons, and Nicolas Salamin for comments. This research was supported by the National Science Foundation. Title of final section respectfully modified from Wilkinson et al. (2005).

References

Allan, G. J., E. A. Zimmer, W. L. Wagner, and D. D. Solokoff. 2003. Molecular phylogenetic analyses of tribe Loteae (Leguminosae): Implications for classification and biogeography. Pages 371–393 *in* Advances in legume systematics, Part 10 (B. B. Klitgaard and A. Bruneau, eds.). Royal Botanic Gardens, Kew.

Altschul, S., W. Gish, W. Miller, E. W. Myers, and D. Lipman. 1990. A basic local alignment search tool. J. Mol. Biol. 215:403–410.

Ané, C., O. Eulenstein, R. Piaggio-Talice, and M. J. Sanderson. 2006. Groves of phylogenetic trees. Pages 1–31 *in* Technical Report 1123. Department of Statistics, University of Wisconsin, Madison, Wisconsin.

Angiosperm Phylogeny Group II. 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.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.

Belshaw, R., and A. Katzourakis. 2005. BlastAlign: A program that uses blast to align problematic nucleotide sequences. Bioinformatics 21:122–123.

Benson, D., I. Karsch-Mizrachi, D. Lipman, J. Ostell, and D. Wheeler. 2005. GenBank. Nucleic Acids Res. 33:D34–D38.

Bininda-Emonds, O. R. P. 2004. Phylogenetic supertrees. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Bininda-Emonds, O. R. P., J. L. Gittleman, and A. Purvis. 1999. Building large trees by combining phylogenetic information: A complete phylogeny of the extant Carnivora (Mammalia). Biol. Rev. 74:143–175.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.

Crisp, M. D., and L. G. Cook. 2003. Phylogeny and embryo sac evolution in the endemic Australasian Papilionoid tribes Mirbelieae and Bossiaeeae. Pages 253–268 *in* Advances in legume systematics, Part 10 (B. B. Klitgaard and A. Bruneau, eds.). Royal Botanic Gardens, Kew.

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

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Gen. 6:361–375.

Dondoshansky, I. 2002. BLASTCLUST, version 6.1. National Center for Biotechnology Information, Bethesda, Maryland.

Doyle, J. J., J. L. Doyle, and C. Harbison. 2003. Chloroplast-expressed glutamine synthetase in *Glycine* and related Leguminosae: Phylogeny, gene duplication, and ancient polyploidy. Syst. Bot. 28:567–577.

Doyle, J. J., and M. A. Luckow. 2003. The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Pl. Phys. 131:900–910.

Driskell, A. C., C. Ané, J. G. Burleigh, M. M. McMahon, B. O'Meara, and M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172–1174.

Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

Edgar, R. C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:1–19.

Eisen, J. A., and C. M. Fraser. 2003. Phylogenomics: Intersection of evolution and genomics. Science 300:1706–1707.

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

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Galtier, N., M. Gouy, and C. Gautier. 1996. Seaview and phylo_win: Two graphic tools for sequence alignment and molecular phylogeny. Comp. Appl. Biosci. 12:543–548.

Ganapathy, G., V. Ramachandran, and T. Warnow. 2003. Better hill-climbing searches for parsimony. Lect. Notes Bioinformatics 2812:245–258.

Gatesy, J., R. H. Baker, and C. Hayashi. 2004. Inconsistencies in arguments for the supertree approach: Supermatrices versus supertrees of Crocodylia. Syst. Biol. 53:342–355.

Goloboff, P. A. 1999. Analyzing large data sets in reasonable times: Solutions for composite optima. Cladistics 15:415–428.

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

Grasso, C., and C. Lee. 2004. Combining partial order alignment and progressive multiple sequence alignment increases the alignment speed and scalability to very large alignment problems. Bioinformatics 20:1546–1556.

Grotkopp, E., M. Rejmánek, M. J. Sanderson, and T. L. Rost. 2004. Evolution of genome size in pines (*Pinus*) and its life-history correlates: Supertree analyses. Evolution 58:1705–1729.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Gusfield, D. 1997. Algorithms on strings, trees and sequences. Cambridge University Press, New York.

Hibbett, D. S., R. H. Nilsson, M. Snyder, M. Fonseca, J. Costanzo, and M. Shonfeld. 2005. Automated phylogenetic taxonomy: An example in the homobasidiomycetes (mushroom-forming fungi). Syst. Biol. 54:660–668.

Hu, J. M., M. Lavin, M. F. Wojciechowski, and M. J. Sanderson. 2000. Phylogenetic systematics of the tribe Millettieae (Leguminosae) based on chloroplast *trnK/matK* sequences and its implications for evolutionary patterns in papilionoideae. Am. J. Bot. 87:418–430.

Hughes, J. R., J.-F. Cheng, N. Ventress, S. Prabhakar, K. Clark, E. Anguita, M. De Gobbi, P. de Jong, E. Rubin, and D. R. Higgs. 2005. Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. Proc. Natl. Acad. Sci. USA 102:9830–9835.

Huson, D. H., S. M. Nettles, and T. J. Warnow. 1999. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. J. Comp. Biol. 6:369–386.

Kajita, T., H. Ohashi, Y. Tateishi, C. D. Bailey, and J. J. Doyle. 2001. *RbcL* and legume phylogeny, with particular reference to Phaseoleae, Millettieae, and allies. Syst. Bot. 26:515–536.

Källersjö, M., J. S. Farris, M. W. Chase, B. Bremer, M. F. Fay, C. J. Humphries, G. Petersen, O. Seberg, and K. Bremer. 1998. Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. Pl. Syst. Evol. 213:259–287.

Kim, S., P. S. Soltis, K. Wall, and D. E. Soltis. 2006. Phylogeny and domain evolution in the APETAL2-like gene family. Mol. Biol. Evol. 23:107–120.

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

Kumar, A., and J. L. Bennetzen. 1999. Plant retrotransposons. Annu. Rev. Genet. 33:479–532.

Lackey, J. A. 1977. *Neonotonia*, a new generic name to include *Glycine wightii* (Arnott) Verdcourt (Leguminosae, Papilionoideae). Phytologia 37:209–212.

Lassmann, T., and E. L. L. Sonnhammer. 2002. Quality assessment of multiple alignment programs. FEBS Lett. 529:126–130.

Lassmann, T., and E. L. L. Sonnhammer. 2005a. Automatic assessment of alignment quality. Nucleic Acids Res. 33:7120–7128.

Lassmann, T., and E. L. L. Sonnhammer. 2005b. Kalign—an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6:298.

Lavin, M., R. T. Pennington, B. B. Klitgaard, J. I. Sprent, H. C. de Lima, and P. E. Gasson. 2001. The dalbergioid legumes (Fabaceae): Delimitation of a pantropical monophyletic clade. Am. J. Bot. 88:503–533.

Lavin, M., and M. Sousa S. 1995. Phylogenetic systematics and biogeography of the tribe Robinieae (Leguminosae). Syst. Bot. Mon. 45:1–165.

Leebens-Mack, J., L. Raubeson, L. Cui, J. Kuehl, M. Fourcade, T. Chumley, J. Boore, R. Jansen, and C. dePamphilis. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. Mol. Biol. Evol. 22:1948–1963.

Lerat, E., V. Daubin, and A. Moran. 2003. From gene trees to organismal phylogeny in prokaryotes: The case of the gamma-Proteobactera. PLoS Biol. 1:1–9.

Lewis, G., B. Schrire, B. Mackinder, and M. Lock. 2005. Legumes of the world. Royal Botanic Gardens, Kew.

Liston, A., and J. A. Wheeler. 1994. The phylogenetic position of the genus *Astragalus* (Fabaceae): Evidence from the chloroplast genes *rpo*C1 and *rpo*C2. Biochem. Syst. Ecol. 22:377–388.

Maddison, D. R., and W. P. Maddison. 2005. MacClade, version 4. Sinauer Associates, Sunderland, Massachusetts.

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

McMahon, M., and L. Hufford. 2004. Phylogeny of Amorpheae (Fabaceae: Papilionoideae). Am. J. Bot. 91:1219–1230.

Mishler, B. 1994. Cladistic analysis of molecular and morphological data. Am. J. Phys. Anthropol. 94:143–156.

Moles, A., D. Ackerly, C. Webb, J. Tweddle, J. Dickie, and M. Westoby. 2005. A brief history of seed size. Science 307:576–580.

Morgenstern, B. 1999. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211–218.

Mort, M. E., P. S. Soltis, D. E. Soltis, and M. L. Mabry. 2000. Comparison of three methods for estimating internal support on phylogenetic trees. Syst. Biol. 49:160–171.

Müller, K. F. 2005. The efficiency of different search strategies in estimating parsimony jackknife, bootstrap, and Bremer support. BMC Evol. Biol. 5:58.

Nixon, K. C. 1999. The Parsimony Ratchet, a new method for rapid parsimony analysis. Cladistics 15:407–414.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.

Notredame, C., L. Holm, and D. G. Higgins. 1998. COFFEE: An objective function for multiple sequence alignments. Bioinformatics 14:407–422.

Page, R. D. M. 1996. On consensus, confidence and "total" evidence. Cladistics 12:83–92.

Page, R. D. M. 2005. A taxonomic search engine: Federating taxonomic databases using web services. BMC Bioinformatics 6:48–55.

Pei, J. M., and N. V. Grishin. 2001. AL2CO: Calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712.

Pennington, R. T., B. B. Klitgaard, H. Ireland, and M. Lavin. 2000. New insights into floral evolution of basal Papilionoideae from molecular phylogenies. Pages 233–248 *in* Advances in legume systematics, Part 9 (P. S. Herendeen and A. Bruneau, eds.). Royal Botanic Gardens, Kew.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22:1246–1253.

Philippe, H., E. Snell, E. Bapteste, P. Lopez, P. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: Impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.

Pisani, D., A. M. Yates, M. C. Langer, and M. J. Benton. 2002. A genus-level supertree of the Dinosauria. Proc. R. Soc. Lond. B 269:915–921.

Poirot, O., E. O'Toole, and C. Notredame. 2003. Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. Nucleic Acids Res. 31:3503–3506.

Qiu, Y. L., O. Dombrovska, J. Lee, L. B. Li, B. A. Whitlock, F. Bernasconi-Quadroni, J. S. Rest, C. C. Davis, T. Borsch, K. W. Hilu, S. S. Renner, D. E. Soltis, P. S. Soltis, M. J. Zanis, J. J. Cannone, R. R. Gutell, M. Powell, V. Savolainen, L. W. Chatrou, and M. W. Chase. 2005. Phylogenetic

analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. Int. J. Pl. Sci. 166:815–842.
Rambaut, A. 1996. Se-Al: Sequence alignment editor. Available at http://evolve.zoo.ox.ac.uk/.
Reyes, J., M. Muro-Pastor, and F. Florencio. 2004. The GATA family of transcription factors in Arabidopsis and rice. Plant Phys. 134:1718–1732.
Rokas, A., B. Williams, N. King, and S. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.
Ronquist, F. 1998. Fast Fitch-parsimony algorithms for large data sets. Cladistics 14:387–400.
Salamin, N., M. W. Chase, T. R. Hodkinson, and V. Savolainen. 2003. Assessing internal support with large phylogenetic DNA matrices. Mol. Phyl. Evol. 27:528–539.
Salamin, N, T. R. Hodkinson, and V. Savolainen. 2005. Towards building the Tree of Life: A simulation study for all angiosperm genera. Syst. Biol. 54:183–196.
Sanderson, M. J. 1989. Confidence limits on phylogenies: The bootstrap revisited. Cladistics 5:113–130.
Sanderson, M. J. 2006. Paloverde: an OpenGL 3D phylogeny browser. Bioinformatics 22:1004–1006.
Sanderson, M. J., and J. J. Doyle. 1993. Chloroplast DNA relationships in North American *Astragalus*. Syst. Bot. 18:395–408.
Sanderson, M. J., and A. C. Driskell. 2003. The challenge of constructing large phylogenetic trees. Trends Pl. Sci. 8:374–379.
Sanderson, M. J., A. C. Driskell, R. H. Ree, O. Eulenstein, and S. Langley. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. 20:1036–1042.
Sanderson, M. J., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: Assembling the trees of life. Trends Ecol. Evol. 13:105–109.
Sanderson, M. J., and M. F. Wojciechowski. 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). Syst. Biol. 49:671–685.
Simmons, M. P., C. D. Bailey, and K. C. Nixon. 2000. Phylogeny reconstruction using duplicate genes. Mol. Biol. Evol. 17:469–473.
Sonnhammer, E. L. L., S. R. Eddy, and R. Durbin. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins 28:405–420.
Spellenberg, R. 1976. Chromosome numbers and their cytotaxonomic significance for North American *Astragalus* (Fabaceae). Taxon 25:463–476.
Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.
Steele, K. P., and M. F. Wojciechowski. 2003. Phylogenetic analyses of tribes Trifolieae and Vicieae, based on sequences of the plastid gene, *matK* (Papilionoideae: Leguminosae). Pages 355–370 *in* Advances in legume systematics, Part 10 (B. B. Klitgaard and A. Bruneau, eds.). Royal Botanic Gardens, Kew.
Subramanian, A. R., J. Weyer-Menkhoff, M. Kaufmann, and B. Morgenstern. 2005. DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics 6:66.
Swofford, D. L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates, Sunderland, Massachusetts.
Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.
Tamura, K., M. Nei, and S. Kumar. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. USA 101:11030–11035.
Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Thompson, J. D., F. Plewniak, R. Ripp, J. C. Thierry, and O. Poch. 2001. Towards a reliable objective function for multiple sequence alignments. J. Mol. Biol. 314:937–951.
Vinh, L. S. and A. von Haeseler. 2005. Shortest triplet clustering: Reconstructing large phylogenies using representative sets. BMC Bioinformatics 6:92.
Wilkinson, M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol. Evol. 13:437–444.
Wilkinson, M., J. A. Cotton, C. Creevey, O. Eulenstein, S. R. Harris, F. J. Lapointe, C. Levasseur, J. O. McInerney, D. Pisani, and J. L. Thorley. 2005. The shape of supertrees to come: tree shape related properties of fourteen supertree methods. Syst. Biol. 54:419–431.
Wojciechowski, M. F., M. Lavin, and M. J. Sanderson. 2004. A phylogeny of legumes (Leguminosae) based on analyses of the plastid matK gene resolves many well-supported subclades within the family. Am. J. Bot. 91:1846–1862.
Wojciechowski, M. F., M. J. Sanderson, B. G. Baldwin, and M. J. Donoghue. 1993. Monophyly of aneuploid *Astragalus*: Evidence from nuclear ribosomal DNA internal transcribed spacer sequences. Am. J. Bot. 80:711–722.
Wojciechowski, M. F., M. J. Sanderson, and J.-M. Hu. 1999. Evidence on the monophyly of *Astragalus* and its major subgroups based on nuclear ribosomal DNA ITS and chloroplast DNA trnL intron data. Syst. Bot. 24:409–437.
Wojciechowski, M. F., M. J. Sanderson, K. P. Steel, and A. Liston. 2000. Molecular phylogeny of the "temperate herbaceous tribes" of papilionoid legumes: a supertree approach. Pages 277–298 *in* Advances in legume systematics, Part 9 (P. S. Herendeen, and A. Bruneau, eds.). Royal Botanic Gardens, Kew.
Wolf, Y., I. Rogozin, and E. Koonin. 2004. Coelomata and not ecdysozoa: Evidence from genome-wide phylogenetic analysis. Gen. Res. 14:29–36.
Zanis, M. J., P. S. Soltis, Y. L. Qiu, E. Zimmer, and D. E. Soltis. 2003. Phylogenetic analyses and perianth evolution in basal angiosperms. Ann. Miss. Bot. Gard. 90:129–150.

First submitted 5 January 2006; reviews returned 15 March 2006; final acceptance 9 June 2006
Associate Editor: Vincent Savolainan