

**Bacterial Symbiosis and Plant Host Use Evolution in Dryophthorinae (Coleoptera,  
Curculionidae): A phylogenetic study using parsimony and Bayesian analysis**

A Thesis

Presented to the Department of Biology

Harvard University

By Brian O'Meara [omeara@post.harvard.edu]

In Partial Fulfillment of the Requirements

For the Degree of

Bachelor of Arts with Honors

April 2001

Thesis Director: Professor Brian D. Farrell

**Table of Contents:**

Abstract: p. 3

Introduction: p. 4

Methods: p. 11

Results: p. 22

Discussion: p. 30

Conclusion: p.36

References: p.38

Tables: p. 44

Table 1: Primers used in this study

Table 2: Tests of alternate topologies (Bayesian, KH, Templeton, winning sites)

Figures: p. 48

Figure captions

Figure 1: Bacteria phylogeny

Figure 2: Monocot phylogeny

Figure 3: Total evidence parsimony phylogeny, with Bremer and bootstrap values

Figure 4: COI phylogeny, branchlengths by parsimony, with bootstrap values

Figure 5: EF-1a phylogeny, branchlengths by parsimony, with bootstrap values

Figure 6: 28S phylogeny, bootstrap consensus and values

Figure 7: Clock-optimized total evidence tree, using COI

Figure 8: Bayesian phylogeny, with Bayesian posterior probabilities

Figure 9: Ln likelihood score versus MCMCMC generation

Figure 10: Strict consensus of most probable Bayesian tree with MP trees

Figure 11: Plant host mapped on (A) parsimony and (B) Bayesian trees

Figure 12: Symbiont clade mapped on (A) parsimony and (B) Bayesian trees

Appendices: p. 64

Appendix I: Using Bayesian analysis for phylogenetics

Appendix II: Dryophthorine sequences used in this study (not in pdf version)

**Abstract**

This work presents a phylogenetic analysis of the weevil subfamily Dryophthorinae (Coleoptera: Curculionidae). Three genes (cytochrome oxidase I, elongation factor 1-alpha, and 28S) were sequenced and analyzed using parsimony and Bayesian techniques. The results were examined in comparison with the phylogeny of the weevils' bacterial endosymbionts and plant hosts. The analysis showed that symbiosis originated no more than once within this group of weevils, which is consistent with results from other insect-bacteria systems. However, when the results of this and the bacterial phylogeny are compared, at least five shifts of symbiotic partners are found. This result is robust based on the sequence data (probability of fewer steps is less than 0.001). The evolution of hostplant use was also investigated. The group was found to be ancestrally palm-feeding and to have originated about 70 million years ago.

## **Introduction**

### *Animal and Plant Microbial Intracellular Symbioses*

While the importance of intracellular symbioses (where the symbiont occurs within host cells) in both terrestrial and marine ecosystems has been appreciated for many decades (Paracer and Ahmadjian 2000; Buchner, 1965), the advent of molecular techniques has galvanized new research into symbioses of multicellular plants and animals with bacteria, fungi, and algae. In particular, PCR and sequencing permit the identification of microbial symbionts and their gene products in species for which only morphological descriptions existed before, resulting in the investigation of questions concerning the number of symbionts present, their mode of transmission, whether or not these occur in other related species, their evolutionary origins, and consequences for both host and symbiont (Moran and Telang, 1998; Charles et al., 1997; Charles et al., 1995; Campbell et al., 1992).

Molecular research has thus opened up an astonishingly wide range of phenomena that both inform our understanding of symbioses and bring insights from these interactions to other areas in evolutionary biology. For example, among recent notable findings include combined molecular and fossil evidence that mycorrhizal associations of Glomales date to 600 MYA, and thus possibly facilitated the colonization of land by green plants (Redecker et al. 2000), and the bacterial symbionts of aphids confirm the

theoretical expectations of Muller's ratchet of accumulation of slightly deleterious mutations when population sizes are as severely curtailed as they are in these symbionts (Moran, 1996) .

To date, most intracellular symbioses studied fall into one of two categories: those which are strictly vertically transmitted once the initial symbiosis forms, such as in aphids (Moran and Telang 1998), or those with strictly horizontal transfer, as in *Vibrio* with luminescent squid (McFall-Ngai and Ruby, 2000). According to Douglas (1994), "all intracellular symbionts in insects are vertically transmitted."

#### *Insect-bacterial symbioses*

Much of the most detailed work in animal-bacterial symbioses involves insects, and in particular, the symbioses of aphids, carpenter ants, tsetse flies and weevils in subfamily Dryophthorinae (Moran and Telang 1998). Physiological experiments have demonstrated the role of symbionts in host nutrition in aphids (Sandstrom and Moran, 1999) and dryophthorine weevils (Grenier et al., 1997), and molecular phylogenetic studies of have confirmed parallel phylogenesis between insects and hosts in aphids (Moran and Telang 1998), tsetse flies (Chen et al. 1999) and carpenter ants (Sauer et al. 2000).

Symbioses between insects and bacteria are quite widespread, and seem to occur in insect groups that feed on nutritionally deficient food such as phloem, blood or wood where the ability of symbionts to produce essential amino acids would be useful. The symbionts are often housed in cells in the haemocoel, fat body, or midgut caeca (Douglas 1989). They have been reported in wood decomposers and detritus feeders such as carpenter ants (Formicidae) and beetles (Throscidae, Nosondendridae, Bostrychidae, Lyctidae, Anobiidae, Silvanidae and Cerambycidae), cockroaches (Blattidae), lice (Phthiraptera); bloodfeeders such as bedbugs (Cimicidae) and tsetse flies (Glossinidae); and some specialist herbivore beetles (Chrysomelidae, Bruchidae, Curculionidae; homopterans (Aphididae; Cicadidae; Diaspididae); Douglas 1989; Moran and Telang 1998).

Our best knowledge of insect-bacteria symbioses comes from the aphid-*Buchnera* symbiosis. In this system, careful analysis of the phylogenies of the symbiont and the host have allowed researchers to deduce the age of the symbiosis, rate of gene loss by the bacteria, the effects of small population size on the beetle's molecular evolution, and amount of horizontal transfer of the symbiont (Moran and Telang, 1998, and refs therein). This also builds on work determining the genes involved in the symbiosis, the development of the symbiosis through an aphid's lifetime, transmission of the symbiont, and many other factors (i.e., Buchner, 1965).

*The system: The bacteria*

The second-best known insect-bacteria symbiosis is that between dryophthorine weevils and their endosymbiotic, intracellular bacteria. These intracellular bacteria are found only in specialized cells known as bacteriocytes, where they survive with no host-derived membrane around them (Charles et al., 1997, Nardon et al., 1998) and are transmitted maternally via the ovarioles. Recent work on the phylogeny of the weevil symbionts (Figure 1) reveals three distinct clades within the  $\alpha$ -3 proteobacteria, near groupings containing other insect symbionts, as well as *Escherichia coli* and numerous pathogens (Heddi et al. unpublished). The bacteria induce formation of an organ known as a bacteriome in which they are housed (Charles et al., 1997), found around the gut in the larvae and at the ovaries and the midgut caeca in adults (Charles et al., 1997; Nardon et al., 1998). Aposymbiotic weevils can be produced by heating to 38 °C, allowing performance of symbiont-containing and aposymbiotic weevils to be compared. Aposymbiotic weevils lose the ability to fly, and reproduction is impaired unless nutrients normally supplied by the bacteria (pantothenic acid, riboflavin, phenylalanine, and proline) are provided in the diet (Charles et al. 1997; Grenier et al. 1994). Weight gain per day was 45% higher, and development time was 20% faster, in the symbiotic strain than in the lab-created aposymbiotic strain of *Sitophilus granarius* (Delobel and

Grenier, 1993). The weevil controls symbiont number, as shown by mating experiments between strains selected for high and low symbiont number (Nardon et al., 1998).

The symbiosis has apparently been lost in *Sitophilus linearis*, which also has switched to eating legumes (Charles et al., 1995). This species can not be raised on grain, unlike its congeners which have been studied (Delobel and Grenier, 1993). Similarly, some Egyptian strains of *Sitophilus granarius* are reported to be naturally aposymbiotic (Koch, 1967; Buchner, 1965), though other individuals in the species do have symbionts.

*The system: Weevils in the subfamily Dryophthorinae (Curculionidae)*

The weevil subfamily Dryophthorinae is interesting in its own right, even without the symbiosis. The group comprises 140 genera, totaling 1,010 species (Thompson, 1992). This group is widely known for its economic importance as pests of monocots. Three species alone, *Sitophilus oryzae*, *S. zeamais*, and *S. granarius*, consume \$35 billion dollars worth of the world's grain annually (D. Shuman USDA-ARS, pers. comm.). During World War I, infestations of grain stores in Australia were so severe that at one site alone, one billion weevils (i.e., over one ton) were swept up daily and destroyed (Zimmerman 1993).

The seed feeding habit is fairly unusual in this group, however. Most of the Dryophthorinae are stem borers as larvae and adults (Zimmerman 1993). Hosts include

palms (Aracaceae), bromeliads (Bromeliadae), bananas (Zingiberales), yucca (Liliales), sugarcane (Poales), and other monocots (Zimmerman, 1993, many others — see references) The phylogeny of their hosts, after Bremer (2000), is shown in Figure 2. The weevils often target wounded plants (Vaurie, 1970). Primary centers of diversity seem to be Africa and the Indo-Pacific, with tropical America as a secondary center of diversification (Zimmerman, 1993). Members of this group have become fairly widespread due to introductions. These introduced species include many severe economic pests and can also disturb native plant species. For example, a recent introduction of the Mexican/Central American *Metamasius callizona* now endangers native species in the bromeliad genus *Tillandsia* (the state-protected *T. utriculata*, and *T. paucifolia* and *T. fasciculata*) in Florida (Frank and McCoy 1995).

The relatively well-studied physiology and natural history of the interactions between the Dryophthorinae and their endosymbionts and hostplants provide a natural opportunity for the integration possible with a well-supported phylogeny of these groups. To accomplish this, I have been engaged in an over two year collaboration with researchers in the Nardon lab at INSA-Lyon in France, who have provided me with a 16S phylogeny estimate of the symbionts, plus numerous beetle samples. Here I present the results of this collaboration. I have sequenced cytochrome oxidase I, elongation factor 1-alpha, and 28S from the weevils. This thesis integrates the preliminary results from their

analysis with the results from my phylogenetic investigation of the beetles, which uses parsimony tests and Bayesian analysis to provide a statistical basis for the conclusions drawn about evolution in this system.

## Methods

*DNA acquisition and extraction:* Weevil specimens were obtained from colleagues in France, the US, and Australia. DNA extractions were done using a salting out protocol (Normark et al. 1999) or using a Qiagen Tissue Kit (Qiagen catalog number 29304). Additional beetle DNA template was obtained from the INSA-Lyon lab for taxa otherwise difficult to acquire.

*DNA amplification:* Mitochondrial protein-coding gene cytochrome oxidase I (COI), nuclear protein-coding gene elongation factor 1-alpha (abbreviated EF-1a), and nuclear rRNA-coding gene 28S were used for this analysis. Primers were standard Farrell lab sequences (Table 1). The optimized PCR mix was 39.35  $\mu$ l water, 5  $\mu$ l 10x buffer, 2  $\mu$ l  $MgCl_2$  solution, 1  $\mu$ l DNA template, 1  $\mu$ l of forward primer, 1  $\mu$ l of reverse primer, 0.4  $\mu$ l dNTPs, and 0.2  $\mu$ l Taq. Qiagen Taq Polymerase kits (Qiagen catalog number 201203) were used for the buffer,  $MgCl_2$ , and Taq enzyme. In some PCR amplification reactions, 5  $\mu$ l of Genereleaser (BioVentures) was used in place of an equal volume of water. Reactions were run on MJ Research PTC-200 peltier thermal cyclers. Seven different amplification programs were used. Three of the programs were simple "hot start" programs: 95 °C for one minute, then 40 cycles of 95 °C for 30 s, the annealing temperature for 60 s, then 72 °C for 90 s, finishing with a 72 °C final extension step. The

annealing temperature for each program was 42 °C, 44 °C, and 47 °C, respectively. For some templates, programs were modified to extend the annealing and extension steps to 120 s and 150 s, respectively. This doubled reaction yield. The seventh program was a touchdown program. After a 95 °C for 60 s denaturing step, cycles of 95 °C for 30 s, an annealing temp for 60 s, and a 72 °C for 60 s take place. Initial annealing temperature was 52 °C. Each cycle would drop the annealing temperature down by 2 °C. The cycle with the final annealing temperature of 42 °C was repeated 19 times, followed by a 72 °C for five minute final extension step. The touchdown program was primarily used for EF-1a; the 42 °C programs were used for COI, and the 45-47 °C programs were used for 28S.

*Sequencing:* Sequencing was carried out mostly on an ABI 370a machine, though a few samples were run on an ABI 3100. Reactions were run following ABI protocols. Dye terminators were used for the 370a machine, sometimes using halfTERM (GenPak, Ltd) to dilute the reactions from half-reactions to quarter-reactions. Plates were poured using Sequagel (National Diagnostics) or Long Ranger (FMC Bioproducts). Big Dye was used on the 3100 machine.

*Sequence editing and alignment:* Sequences were edited in Sequencher 3.0 and 3.1.1 (Gene Codes Corporation). COI and the coding regions of EF-1a were aligned by eye.

The intron was aligned with ClustalX 1.6 (Thompson et al., 1994) using the default 15/6.66 gap opening/extension ratio.

ClustalX 1.6 was used for the 28S alignment, according to the technique of Maddison et al. (1999). Gap opening/extension costs were 50/5, 20/5, 15/6.66, 15/3, 12/7, 10/5, 10/2, 8/3, 7/2, 5/1, 3/2, and 3/0.5. Taxa missing internal stretches of sequence were aligned in multiple parts. These sequences were reunited in MacClade 4.0 (Maddison and Maddison, 2000) after the alignment was complete.

A two-step process was used to choose the final alignment. On the first pass, the three obviously best alignments were chosen by eye by examining the character matrix in MacClade, without referring to taxon names. These three alignments had been created using gap opening/extension costs of 10/2, 10/5, and 20./5. Each of these three alignments was then scored for number of hypervariable regions, the number of bases in hypervariable regions, the total number of parsimony informative characters, and the number of non-hypervariable parsimony informative characters. The 20/5 alignment, was selected, as it had the highest number of parsimony informative characters and lowest number of hypervariable characters. The taxon names were then covered and obvious errors in the matrix were corrected by eye. Subsequent analyses excluded the 28S hypervariable regions, as well as the intron of EF-1a, due to uncertainty in primary homology assessments in those regions. The sequences appear in Appendix II [not in pdf].

*ILD test/taxon selection* : The ILD test (also known as partition homogeneity test) is used to test for incongruence between different partitions of the data (Farris et al., 1995). In this case, the data were partitioned by gene. A 500-replicate ILD search, with 5 random addition sequence heuristic searches per replicate, TBR branch swapping, was performed, using PAUP\* version 4.0b6 (PAUP\* versions 4.0b6 to 4.0b8 were used for this analysis — Swofford 1998). Only taxa with all three genes were used in this and later ILD analyses (including taxa without certain partitions would bias the ILD non-conservatively). As there was some conflict found (see results), pairwise comparisons were run to find which gene was causing the most conflict. Once this gene was identified, the possibility that one taxon was causing the problem was tested by using a batch file to automate a taxon jackknifing ILD test. Testing over all three partitions, the file removed one taxon, performed an ILD test with the remaining taxa, replaced that taxon, removed another, and so forth for all taxa in the analysis. One sequence was found to be causing the incongruence. As there were strong grounds for thinking that this sequence had been misidentified, it was removed from further analysis. The ILD test for all genes together was then run again. All taxa with at least two genes sequenced were used in this analysis. The outgroup sequence, *Tanysphyrus lemnae*, comes from Erirrhinae, the sister group of Dryophthorinae (Marvaldi et al., 2001).

*Total evidence searches and support:* Parsimony heuristic searches under a variety of strategies were attempted in PAUP\*. The first search used 500 random taxon addition sequences and TBR branch swapping, saving multiple trees per replicate. A second search used 50,000 random trees instead of random taxon addition, saving only one tree per replicate. A third search used the parsimony ratchet (Nixon, 1999), as implemented in PaupRat (Sikes and Lewis, 2000) in a 500 replicate search.

Bootstrap proportions for the total evidence tree were calculated in a 1000 bootstrap replicate, 50 random taxon addition sequence per rep, TBR branchswapping search. Bootstrap values should not be interpreted literally as probability of support for a clade: as has been shown in numerous studies, these values are only roughly correlated with the clade's probability of actually existing (Hillis and Bull, 1993). The Bayesian methods used later in this thesis also calculate these values.

Another way to examine support of nodes is Bremer support (Bremer 1988 and its descendant, partitioned Bremer support). With this method, the length of the tree with and without certain nodes is compared for different partitions of the data. The difference in treelength is related to the amount of support that data partition (COI first codon position, for example) gives to that node. In this analysis, the program TreeRot v2 (Sorenson, 1999) was used to generate a batch file, with character partitions as follows: COI 1<sup>st</sup>

codon position, COI 2<sup>nd</sup> codon position, COI 3<sup>rd</sup> codon position, EF-1a 1<sup>st</sup> codon position, EF-1a 2<sup>nd</sup> codon position, EF-1a 3<sup>rd</sup> codon position, and 28S non-hypervariable region.

Each search at a given node was a heuristic search with 250 random taxon additions and TBR branch swapping. Multiple most parsimonious trees are often found in the search using all the data, but they often have different scores when only one character partition is included. In such a case, the support values from different trees are averaged to give the support value for that node and partition. The values may not all be integers for this reason. Since the program TreeRot introduced rounding error when parsing the output file, Bremer support values were calculated manually within Microsoft Excel.

*Searches by locus:* Separate analyses for each locus were run with 500 random taxon addition sequence TBR heuristic searches. Bootstrap values for gene trees were calculated in a 500 bootstrap replicate, 50 random taxon addition sequence per rep, TBR branchswapping search.

*Likelihood:* Full likelihood searches are prohibitively time intensive for this 27 taxon, three locus data set. Nineteen tree rearrangements took 38 minutes to perform under an HKY model of likelihood in PAUP\*. At this rate, it would take over 32 computer-years to do the same number of rearrangements as was done in the 500 random

taxon addition sequence parsimony search. Completing one random taxon addition replicate would take over 23 days.

Likelihood could be used for branchlength optimization of a parsimony tree. ModelTest (Posada and Crandal, 1998) was used to determine the appropriate likelihood model for each locus and for the combination of all three loci. Clock and non-clock model tree scores were compared using likelihood ratio tests for the combined loci, for COI alone, and for EF-1a alone. The clocklike gene was used to optimize the branchlengths of the likelihood maximum parsimony tree.

The calibration point, the oldest known dryophthorine fossil, from the Upper Miocene (>33 million years ago), comes from Zherikhin (2000). This was used as a point estimate for the minimum age of the base of the group, with other node ages calculated from that. The tree is also consistent with the oldest known *Sitophilus* fossil from the Upper Miocene-Lower Turolian (Zherikhin 2000). The calibration rate of 1.5% per million years for beetle COI from Farrell (2001) was also used to calibrate the clock, using the divergence between *Sitophilus zeamais* and *S. granarius* of 13.3% as the best estimate of true (no multiple hit) divergence, which was then used to convert likelihood branchlengths into percent divergence estimates, then ages.

*Bayesian MCMCMC analysis:* Bayesian analysis was used in this study, one of the first to do so (for another paper using the technique, see Larget and Simon, 1999). Appendix I defines terms, explains the technique, and discusses its advantages and disadvantages. The program MrBayes (Huelsenbeck, 2000), which uses the “Metropolis-coupled Markov chain Monte Carlo” (MCMCMC) search strategy, was used for Bayesian analysis. Eighteen searches (requiring over 100 computer-hours on Macintosh G4 450 MHz machines) were run. Four chains were used per run, with a temperature of 0.2 and sampling every 100 generations. The total number of generations was 925,500; number of generations per run ranged from 34,900 to 68,400; average number of generations was 51417. Only samples from after the chain reached stationarity were used. A general time reversible (GTR) model (number of substitution types=6) with gamma-distributed site rates (four categories) was used with default, flat (uninformative) priors and no specified tree topology. Posterior probabilities were calculated by getting a majority rule consensus of the Bayesian tree samples within PAUP\*. The Bayesian tree was compared to the parsimony trees using the Shimodaira-Hasegawa test under a GTR + gamma likelihood model (Shimodaira and Hasegawa, 1999) and using the Kishino-Hasegawa, Templeton, and winning-sites tests under parsimony (Swofford et al., 1996).

*Sitophilus relationships:* The relationships within *Sitophilus* are of great interest due to the global economic importance and the intensive studies of symbiosis in this genus. Exhaustive searches using just five *Sitophilus* species and including the intron of EF-1a and the hypervariable regions of 28S were done under the parsimony and maximum likelihood criteria. To evaluate whether some trees were significantly better, Kishino-Hasegawa, Templeton, and winning-sites tests were performed on all trees for parsimony, while Kishino-Hasegawa and Shimodaira-Hasegawa (1000 RELL replicates) were done on trees for likelihood. The likelihood model used was HKY85 (number of substitution parameters = 2). A bootstrap run of 500 bootstrap replicates, with 50 random taxon additions per replicate, was also performed under the parsimony criterion.

*Hypothesis testing:* Confidence in hypotheses about relationships was assessed using Bayesian and parsimony based methods. Bayesian results can be used to calculate the total posterior probabilities of hypotheses by finding the proportion of Bayes trees which meet the constraint of the hypothesis. This evaluates the absolute probability of the hypothesis being true given the data and model of DNA substitution, not the relative support for two different hypotheses, though the Bayesian probabilities of the two hypotheses could be compared. There are also ways to test hypotheses under parsimony. For example, tests like the Kishino-Hasegawa test, the Templeton-Wilcoxon signed ranks

test, and the winning sites test are at least intended to determine when certain trees are significantly rejected based on other trees (Swofford et al., 1996). One way to test a hypothesis, generally a statement of monophyly of some group, is to find the best trees with and without this clade as a constraint and comparing these trees. Tests of monophyly of beetles grouped by genus, hostplant, or symbiont clade use a simple two-node constraint, while constraints for loci tree tests are perfectly-resolved trees. In this study, the hypotheses to examine are monophyly of the genera, monophyly of the beetles possessing a symbiont from one of the three clades recovered by Heddi and associates (Figure 1), significant difference between the Bayesian and the parsimony trees, and monophyly of beetles consuming a particular hostplant. These were all stored as backbone constraints and 500 random taxon addition heuristic searches with TBR branch swapping were done for trees which differed in the constraint from the most parsimonious tree. These trees were then compared to the most parsimonious total evidence tree with the higher likelihood score using PAUP\*'s implementation of the Kishino-Hasegawa, Templeton (Wilcoxon signed ranks), and winning-sites (sign) tests (Swofford et al., 1996). The Kishino-Hasegawa test was not quite the appropriate test to use, as the test assumes the trees being compared were chosen a priori, which was generally not the case (Goldman et al, 2000). The Bayesian p-value was calculated by determining the proportion of trees in the Bayesian searches which met the constraint.

*Character optimization:* Heddi and colleagues provided a 16S neighbor joining phylogeny of the endosymbionts which included information on the host of each symbiont.. Beetles were coded for a multistate character of having 1) no symbionts, 2) clade one symbionts, 3) clade two symbionts, or 4) clade three symbionts. This character was then mapped on the parsimony and Bayesian trees in MacClade 4 (Maddison and Maddison, 2000).

Information on host use came from the literature (see references) and from labels in the Museum of Comparative Zoology collection. This, too, was mapped in MacClade and optimized on the Bayesian and parsimony trees.

## Results

There were two sequences for the multi-copy gene EF-1a which had one more intron relative to the other sequences. These were excluded to prevent paralogy, as the different copies of EF-1a can be distinguished by their number of introns (Normark et al, 1999). There were three other sequences that did not extend into this region, so the presence of this intron could not be directly determined. However, the known two-intron copies had the motif AGCGA at position 474-478 whereas the known one-intron copies had the motif TCNWS at those positions. Two of the unknown intron copies had a TCCAS sequence at these positions, the same as the one-intron copies. The third sequence had a NNTAC sequence (the first two bases were not certain from the original sequence), which is also the same as the one-intron copy. Thus, only two EF-1a sequences came from a different copy and needed to be excluded.

The initial difference between partitions of the data was found to be significant (unpermuted data tree length 3698, best permuted data tree length was 3728,  $p = 1 - (499/500) = 0.002$ ). Three pairwise ILDs were then run to compare pairs of loci. There was no conflict between COI and EF-1a ( $p = 0.174$ ), but there was between 28S and COI ( $p = 0.002$ ) and 28S and EF-1a ( $p = 0.002$ ). This indicated that 28S had a different history than the other two genes or that there was some other problem with the data (perhaps a misaligned sequence or mislabeled taxon). In the ensuing taxon jackknife ILD

test, removal of all but one taxon did not change the ILD result from the unchanged taxon set (all  $p = 0.002$ ). However, when taxon 7 was excluded the ILD  $p$ -value went up to a nonsignificant 0.100. Thus the conflict was pinpointed to the 28S sequence of this taxon, labeled *Sitophilus zeamais*. This was the only sequence which I did not sequence myself — it came from an earlier study done by our lab on the phylogeny of phytophagous beetles and was coded as RPR02. When used in this analysis, the sequence was incorrectly recoded as *Sitophilus zeamais*. The ILD test after the removal of this 28S sequence was not significant ( $p=0.100$ ).

The number of parsimony informative characters for each locus were 501 of 1301 for COI, 208 out of 877 for EF-1a, and 135 out of 776 for 28S. The 500 addition sequence replicate search strategy on the total evidence tree on all taxa with at least two genes found two trees (Figure 3) of length 4110 for 355 of the random additions (71 percent of the time). The search strategy using 50,000 random trees found these same two trees of score 4110 (12420 times for one tree, 8773 for the other — saving one tree per replicate precluded finding both trees in a given replicate. The next shortest tree was of score 4113). The parsimony ratchet also found these two trees. The trees differ only in reversing the placement of *Rhynchophorus cruentatus* and *Rhynchophorus phoenicis*.

The partitioned Bremer support (Figure 3) shows wide variation in different codons and different loci. There are no trends when Bremer support values are graphed

against time (based on a clock tree) or against bootstrap values (results not shown). Only COI second codon positions were free of conflict with the most parsimonious tree on every branch. The average value of COI first codon position support was negative, though they only had negative values on 7 of the 24 nodes, the same number as did 28S. COI third codon positions had negative values at 9 nodes, EF1 first codon positions had just one negative value, and EF1 second and third codon positions had three negative values each.

Trees for individual loci were found in a 500 random addition sequence replicate parsimony heuristic search with TBR branch swapping. One most-parsimonious COI tree was found in 173 of 500 replicates and was of length 2673 (Figure 4). One EF-1a tree was found in all 500 replicates and was of length 822 (Figure 5). Thirty-nine 28S trees were found of length 556: three came from one TBR island (hit 26 times), 36 came from another (hit 464 times) (Figure 6). Of all comparisons between single locus trees and the total evidence parsimony tree, the only consistent significant difference under parsimony tests occurred between the COI and the total evidence tree in explanatory power for all the data (Table 2: IV.2). The EF-1a and total evidence trees only significantly differed under the winning sites test for all the data (Table 2: VI.2:  $p=0.02$ ). No single locus tree was significantly better than the total evidence tree at explaining the evolution of that single locus (Table 2: V.2, VII.2, and IX.2).

Modeltest found that the best likelihood model for COI was TRN with invariable sites and gamma distribution (nst=6 base=est rmat=est rclass=(a b a a c a) rates=gamma shape=est pinv=est), while GTR + I + gamma was chosen for the combination of all three loci, for EF-1a, and for 28S (nst=6 base=est rmat=est rclass=(a b c d e f) rates=gamma shape=est pinv=est). The clock was rejected for the combined data and for EF-1a, but not rejected for COI (-lnL score 11523.76542 with clock, 11510.7136 without clock, 26 degrees of freedom,  $p = .457$ ). COI was used to optimize branchlengths of the maximum parsimony tree of greater likelihood.

The clock tree (Figure 7) provides estimates of the minimum possible ages, as a single fossil was used as a calibration point for the base. The tree is consistent with the first known *Sitophilus* fossil, which comes from the Upper Miocene-Lower Turolian, five to six million years ago (Zherikhin, 2000). Had this fossil been used instead, minimum ages would be reduced by about 60 percent (age of the group would be 12.7 million years old). A COI calibration in beetles of 1.5% divergence per million years (Farrell, 2001) were also used to calibrate the clock tree. Numbers on the branches are the length of that branch in millions of years.

The Bayesian search result is shown in Figure 8. The searches reached stability after about 10,000 generations, as shown by the likelihood score reaching an asymptote (Figure 9). The searches resulted in 7455 tree samples after the initial burn-in samples

were discarded. Figure 9 shows the progress of the search for trees by generation. These samples contain 459 unique topologies. Negative ln likelihood scores ranged from 20997.904 to 20838.422. The topology with best likelihood score was found most often as well, as would be expected in a thorough Bayesian analysis. The mean score was 20890.686, with standard deviation of 32.178. Posterior probabilities of topologies, calculated by the frequency of the topology in the sampled trees, ranged from 0.000134 for topologies found once to 0.17 for the most likely topology. The Bayesian tree with posterior probabilities for each clade was calculated by loading all the tree samples into PAUP\* and creating a majority rule consensus tree.

The Bayesian and parsimony total evidence trees do not, in general, show significant disagreement. The p-values of all the parsimony based tests comparing the Bayesian and parsimony trees are greater than 0.35. The value of the Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999) comparing the Bayesian with the two parsimony trees using a GTR plus gamma likelihood model found one comparison barely significant (0.021) and one insignificant (0.055). The strict consensus of the two MP trees with the best-supported Bayesian topology is shown in Figure 10 to show areas of incongruence.

The results in Table 2 show varying support for a priori hypotheses about monophyly of genera. Monophyly was supported for *Sitophilus* (Table 2: I.2:

probabilities of *Sitophilus* being paraphyletic were all below 0.05) but was rejected for *Sphenophorus* (Table 2: I.5: p-values were all below 0.00013).<sup>1</sup> Though not found in the most parsimonious trees, monophyly of *Metamasius* could not be rejected under the parsimony tests ( $p > 0.1$ ), though it was not observed in the Bayesian trees. The most unusual result from the hypothesis testing is from the test of the monophyly of *Rhynchophorus* (Table 2: I.4). The search enforcing the genus' monophyly found two trees of length 4123, 199 and 46 times respectively, in a 500 random addition heuristic search. These trees differed from each other at nine out of 25 nodes. The tree found more often differed from the total evidence unconstrained trees at one to two nodes and showed no significant difference under the parsimony tests, while the other tree differed at nine or more nodes and did show significant differences (parsimony tests:  $p < 0.02$ ). The mixed result shows that there is at least one parsimonious tree that keeps *Rhynchophorus* monophyletic but does not strongly conflict with the unconstrained tree, so the monophyly of *Rhynchophorus* cannot be rejected.

The exhaustive *Sitophilus* search was inconclusive. Tree lengths ranged from 517 to 530 for parsimony, 6754.21676 to 6791.91536 for likelihood. None of the nodes had over 46.3 percent bootstrap support under parsimony. Two of the trees were barely

---

<sup>1</sup> Note that monophyly of *Sitophilus* was tested as probability of the clade *not existing* while monophyly of *Sphenophorus* was tested as the probability of the clade *existing* because the opposite constraints were

significantly worse than the best one under the likelihood Kishino-Hasegawa test ( $p=0.0454$  and  $p=0.0498$ ), none were significantly worse under the likelihood Shimodaira-Hasegawa test or any of the parsimony tests. These results suggest that the topology within the *Sitophilus* genus is not robust.

Hostplants of the beetles are mapped on the maximum parsimony and Bayesian trees using MacClade ( Figures 11A and 11B). Group association shows great amounts of homoplasy. Results of tests of monophyly of herbivores of particular plant groups are shown in tests I.8 to I.14 of Table 2. Monophyly of herbivores (both generalists and specialists) of Poales (all herbivores and just stem-borers), Arecaceae, Zingiberales, and Liliales were strongly rejected (all parsimony tests  $p<0.007$ ). Presence of a single shift away from monocots is also rejected (parsimony tests  $p<0.0001$ ). The only group for which monophyly could not be rejected was weevils which feed on Bromeliads (Table 2: I.10:  $p>0.16$ ), though this was not recovered in the parsimony or Bayesian trees.

The symbiont clade present in each beetle, as indicated in Figure 1, was mapped on the maximum parsimony and Bayesian trees in MacClade (Figures 12A and 12B). Beetles with symbiont clade one were monophyletic on both trees, though the possible

---

found on the unconstrained most parsimonious trees. Had the opposite constraints been used, the parsimony tests would be comparing identical trees, which provides very little information.

paraphyly of this group cannot be rejected (Table 2: III.2)<sup>2</sup>. Beetles containing symbiont clade two are paraphyletic on the best trees. Monophyly of these taxa is strongly rejected (Table 2: III.3,  $p < 0.0001$ ). Beetles with symbiont clade three follow the same pattern (Table 2: III.4). Symbiosis as an ancestral character of the group is strongly supported. The only topology which could create ambiguity in this reconstruction would place *Sitophilus linearis* as basal to the rest of the Dryophthorinae. This topology is strongly rejected (Table 2: II.2,  $p < 0.0001$ )<sup>3</sup>.

---

<sup>2</sup> Only taxa for which the presence/absence and identity of the symbionts has been determined were included in this test and other tests in this section.

<sup>3</sup> Constraint tree forced *Sitophilus linearis* to be sister to the outgroup, which was included for this reason, though its symbiont status is unknown.

## Discussion

The phylogeny estimates and clock optimization allow evaluation of the direction and rate of evolution of Dryophthorinae with their symbionts and hostplants. The phylogenies show ancestral feeding on Areaceae, followed by shifts onto other monocot or dicot hosts (Figure 11). The approximately 70 million year old age of the group postdates that of palms by about 10 million years (Cronquist, 1981). Except for a very recent shift from palms to cycads (in the South African genus *Phaecorynes*) all host shifts took place between 20 and 50 million years ago, during the Eocene-Late Oligocene cooling and drying period (Farrell, pers. comm.). While the shifts to Sonoran cacti in the two species of *Cactophagous*, and use of legume and dipterocarp seeds by single species of *Sitophilus* have obviously not been accompanied by substantial diversity, shift to Asteraceae in *Rhodobaenus* is at least accompanied by a radiation of 70 species, though it is not yet clear whether this represents elevated diversification rates. Many host shifts occurred independently on multiple branches (monophyly was strongly rejected for beetles on most hostplant families). These shifts seem to have no correlation with the hostplant phylogeny, as shown by a nonsignificant test using TreeMap (Page, 1995). This suggests that some other factor, perhaps host range or biochemical similarity, may be playing a role determining host shifts (Strong et al., 1984).

An interesting comparison is between the number of shifts between plant groups and plant tissues. The beetles shift hosts from monocots to other plants five times. The beetles shift from stem boring to seed feeding (in *Sitophilus*) only once. This is further evidence for a recurring observation of tissue use being less evolutionarily labile than host use (Farrell et al. 2001).

The phylogeny of the weevils (under parsimony and Bayesian searches) shows one loss of symbiosis, which otherwise occurs in all sampled species. Based on the bacteria phylogeny alone (Figure 1), there must be three origins of symbiosis in the bacteria. This in itself would not be terribly surprising: symbioses have originated several times in bacteria (Moran and Telang, 1998), though they generally follow vertical transmission once founded (i.e., Sameshima, 1999). The beetle phylogeny must show at least three steps when bacteria are mapped on as a multistate character by clade. However, the observed phylogeny requires at least five steps. The statistical methods used here demonstrate that this paraphyly is well supported (p values for monophyly of the beetles possessing clade two or clade three of the symbionts are all less than 0.001). Bacterial clade one, the group of bacteria found in *Sitophilus*, appears to be acquired once by the weevils (Figure 12), while the other two clades of symbionts must each show paraphyly, indicating multiple acquisitions or horizontal transfer within this group. This study can also date the origin of several of these symbioses. For example, *Diocalandra*

and *Sitophilus* share a common ancestor about 52.6 million years ago and have two different symbionts, suggesting that for each genus the association with a particular bacteria clade is this age or younger. The entire group, with its multiple origins/shifts of beetle clade-bacteria clade associations, is approximately 68.7 million years old. This contrasts with the well-known aphid symbiosis, which is 250 million years old and shows no horizontal transfer or multiple origins (Moran and Telang, 1998).

One robust result of this analysis is an observed loss of symbiosis. If symbiosis is scored as a presence/absence character, it is obvious that there is a loss of symbiosis in *Sitophilus linearis* (parsimony-based tests and Bayesian posterior probabilities strongly reject trees with *S. linearis* at the base). Interestingly, this species is the only one known to have switched onto legumes. Do legumes provide the amino acids which the symbiont normally provides to other weevils, rendering the symbiont unnecessary? To answer this, we need to know the amino acid profiles of legumes, palms, and grasses, as well as the profile of amino acids produced by the symbionts.

According to Moran and Telang (1998), "No study has yet produced evidence contradicting the hypothesis of parallel phylogenesis for bacteriocyte-associated endosymbiosis in insects." This study shows strong evidence of just that, despite our initial expectation of vertical transmission of symbionts in Dryophthorinae. We would predict vertical transmission for several reasons. First, we know that the symbionts are

transmitted inside the egg, at least for *Sitophilus*. Second, the association seems fairly highly developed, with beetles exerting control over symbiont number (Nardon et al., 1998), for example. Interestingly, though, there are some strains of wild *Sitophilus* species which lack symbionts, indicating some intraspecific variation. It is not known why the beetles have lost their symbionts.

These observations, and the results of this study, suggest two possible hypotheses for the multiple evolution of these associations. One is that beetles lose their symbionts, are symbiont free for some time, and then acquire new symbionts. If this were true, then beetles should be found in nature without symbionts, as is the case for *Sitophilus linearis* as well as some strains of *Sitophilus* (Koch, 1967). This may have happened on the branch leading to *Sitophilus*, as the basal *Sitophilus linearis* is symbiont-free and the rest of the genus has a different symbiont clade than their nearest relative, *Diocalandra*. Another, closely related, hypothesis is that one strain of symbionts is displaced by another strain, probably through competition, though perhaps due to drift (small size of symbiont inoculum to egg would increase the importance of drift rather than competition). If this were the case, we should find some insect individuals with two strains (intrahost competition) or two insects in the same species but each with a different symbiont strain (interhost). Intrahost competition would lead to more “selfish” symbionts, more likely to have a higher reproductive rate to outcompete other symbionts

in the same host. Interhost competition would result in symbionts adapted to aid their host's survival, though symbionts could also evolve under this scenario to feminize males and create cytoplasmic incompatibility, as in *Wolbachia* (Werren, 1997). Interhost competition could also be observed under the first hypothesis of loss, then gain, if gains happened before all the beetle individuals lost their symbionts.

So far, the question of where and how exactly the beetles acquire their new partners (or, for the bacterial perspective, how the bacteria acquire new beetle hosts) has been unexplored. According to the bacterial phylogeny (Figure 1), the nearest relatives of the *Sitophilus*-associated bacteria (clade 1) are tsetse fly symbionts. The nearest relatives of symbiont clade two are vertebrate pathogens, followed by nematode and whitefly symbionts. The nearest relatives of symbiont clade three are psyllid symbionts, followed by carpenter ant symbionts. It may be that close relatives in the phylogeny are always eukaryote-associated because these clades themselves were ancestrally eukaryote-associated. However, the choice of taxa included in the bacteria phylogeny biases this conclusion, as all the included taxa come from eukaryote associates, most of which are mammal- or insect-associated. Other possible ancestors, such as palm pathogens, were not included. For more rigorous conclusions, more bacteria from the weevils' habitat should be sequenced, and these sequences should be analyzed under methods besides neighbor-joining. The bacteria tree used in this analysis seems fairly robust, though, in its

relevant result of three distinct bacterial clades (bootstraps 95 percent or better for the clades).

This still leaves the questions of where the symbionts come from. Perhaps they are weevil pathogens which evolve to be beneficial. Perhaps they are plant pathogens which survive in the weevil gut. This study alone can rule out some possible causes of the observed pattern. For example, geography does not explain the occurrence of paraphyletic associations. The beetles which have symbiont clades two or three are all New World, except for *Diocalandra*, which is Old World and has symbiont clade two. The association also is not necessarily linked to plant host. Clade two symbionts are found in beetles using palms and bromeliads, while clade three symbionts are found in beetles using palms, bromeliads, Zingiberales, yucca, and grasses. This might hint that clade three symbionts allow their hosts to use more plant taxa, but since there are only three examples for each clade, further investigation into the weevil/symbiont associations is warranted.

## Conclusion

The evolution of associations between weevil species in the subfamily Dryophthorinae and their symbionts in the  $\alpha$ -3 proteobacteria has obviously followed a complex history. Though there are only three symbiont clades, there are five steps when these clades are mapped on the weevil tree, indicating horizontal transfer within Dryophthorinae or convergent evolution of the symbiosis. While these weevils initially attacked palms, and today are still largely monocot feeders, their associations with various monocot groups does not follow the phylogeny of these plants.

This group provides a good complement to the known aphid symbiosis. In aphids, we see the results of a single, vertically-transmitted, 250 million year old infection. The aphid system provides information on the development of a symbiosis through time but little knowledge on origins of the symbiosis. In Dryophthorinae, there are several origins of symbiotic associations between beetle and bacterial taxa in the last 70 million years. There is apparently variation within the group for lack of symbionts (*Sitophilus linearis* and some *Sitophilus granarius*). This system, featuring a dynamic equilibrium of gains and losses of symbionts, holds the promise of shedding light on the origin, maintenance, and loss of insect-bacteria symbioses.

**Acknowledgements**

I would like to thank Brian Farrell for giving me the opportunity for doing this research and for his advice, mentoring, discussion, and editorial assistance. The Heddi, Charles, and Nardon lab provided the bacteria phylogeny as well as many weevil samples. Charlie O'Brien, Bob Anderson, J.H. Frank, and many others provided weevil samples and identification of specimens. I would like to thank the members of the Systematics Discussion Group, whose conversations have helped develop my understanding of phylogenetics. John Huelsenbeck provided a copy of MrBayes. Jeff Chung carefully taught me lab techniques. The other Farrell lab members also provided crucial help. Finally, I would like to thank Andrea Sequeira for her help and guidance.

## References

- Aksoy, S., A. A. Pourhosseini, and A. Chow. 1995. Mycetome endosymbionts of tsetse flies constitute a distinct lineage related to Enterobacteriaceae. *Insect Molecular Biology* 4:15-22.
- Aksoy, S., X. Chen, and V. Hypsa. 1997. Phylogeny and potential transmission routes of midgut-associated endosymbionts of tsetse (Diptera: Glossinidae). *Insect Molecular Biology* 6:183-190.
- Baumann, P. 1998. Symbiotic associations involving microorganisms. *BioScience* 48:254-255.
- Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 44:795-803.
- Bremer, K. 2000. Early Cretaceous lineages of monocot flowering plants. *Proceedings of the National Academy of Sciences* 97:4707-4711.
- Buchner, Paul. 1965. Endosymbiosis of animals with plant microorganisms. Interscience Publishers, New York. Translated by Bertha Mueller.
- Cronquist, A. 1981. *An Integrated System of Classification of Flowering Plants*. Columbia University Press, New York.
- Douglas, A. E. 1989. Mycetocyte symbiosis in insects. *Biological Reviews of the Cambridge Philosophical Society* 64:409-434.
- Douglas, A.E. 1994. *Symbiotic Interactions*. Oxford University Press, Oxford.
- Farrell, B.D; Sequeira, A.S., O'Meara, B, Normark, B.B, Chung, J and Jordal, B. H. 2001. The evolution of agriculture in beetles (Curculionidae: Scolytinae and Platypodini).
- Farris, J. S., M. Källersjö, A. G. Kluge, and C. Bult. 1995. Testing significance of incongruence. *Cladistics* 10:315-319.
- Felsenstein, J. 1988. Phylogenies from molecular sequences: Inferences and reliability. *Annual Review of Genetics* 22: 521-565.
- Goldman, N., J.P. Anderson, and A.G. Rodriggo. 2000. Likelihood-based tests of topologies in phylogenetics. *Systematic Biology* 49:652-670.
- Hillis, D.M. and J.J.Bull. 1993. An empirical test of bootstrapping as a method of assessing confidence in phylogenetic analysis. *Systematic Biology*. 42: 182-192.
- Koch, A. 1967. Insects and their endosymbionts. Pp. 1-106 in S. M. Henry, ed. *Symbiosis: Its Physiological and Biochemical Significance*. Vol. II. Academic Press, Inc., New York.

- Larget, B. and D.L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*. 16:750-759.
- Maddison, D. and W. Maddison. 2000. *Macclade 4*. Sinauer Associates, Sunderland, Massachusetts.
- Maddison, Baker, and Ober, 1999. Phylogeny of carabid beetles as inferred from 18S ribosomal DNA (Coleoptera : Carabidae) *Systematic Entomology*, 24: 103-138.
- Marvaldi, A.E., A.S. Sequeira, and B.D. Farrell. 2001. Molecular and morphological phylogenetics of weevils (Coleoptera, Curculionoidea): Do niche shifts accompany diversification? *Systematic Biology*. In review.
- McFall-Ngai, M.J. and E.G. Ruby. 2000. Developmental biology in marine invertebrate symbioses. *Current Opinion in Microbiology* 3:603-607.
- Moran, N. A., and A. Telang. 1998. Bacteriocyte-associated symbionts of insects. *BioScience* 48:295-304.
- Nixon, K. C. 1999. The Parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics* 15: 407-414.
- Normark, B. B., B. H. Jordal, and B. D. Farrell. 1999. Origin of a haplodiploid beetle
- Page, R.D.M. 1995. *TreeMap*. Distributed by the author.
- Paracer, S. and V. Ahmadian. 2000. *Symbiosis: An introduction to biological associations*. Second edition. Oxford University Press, New York.
- Posada, D. and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14 (9): 817-818
- Redecker, D., R. Kodner, and L.E. Graham. 2000. Glomalean fungi from the Ordovician. *Science* 289:1920-1921.
- Rosner, B. 2000. *Fundamentals of Biostatistics*, 5<sup>th</sup> edition. Duxbury Thompson Learning, Pacific Grove, California.
- Sameshima, S., E. Hasegawa, O. Kitade, N. Minaka, and T. Matsumoto. 1999. Phylogenetic comparison of endosymbionts with their host ants based on molecular evidence. *Zoological Science* 16:993-1000.
- Sikes, D.S, and P.O. Lewis, 2000. *PaupMacRat*. Distributed by the authors.
- Sorenson, M.D. 1999. *TreeRot*, version 2. Boston University, Boston, MA.
- Strong, D.R., J.H. Lawton, and R. Southwood. 1984. *Insects on Plants*. Harvard University Press, Cambridge.
- Swofford, D. L. 1998. *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. *In Molecular Systematics*, second edition. Hillis, D.M., C. Moritz, and B.K. Mable, eds. Sinauer Associates, Sunderland, MA. Pages 407-514.

- Thompson, R. T. 1992. Observations on the morphology and classification of weevils (Coleoptera, Curculionoidea) with a key to major groups. *Journal of Natural History* 26: 835-891.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673-4680.
- Werren, J. H. 1997. Biology of *Wolbachia*. *Annual Review of Entomology* 42:587-609.
- Zherikhin, V.V. 2000. Tertiary brachycerid weevils (Coleoptera: Brachyceridae) from the collections of Muséum Nationale d'Histoire Naturelle, Paris, with a review of other fossil *Brachyceridae*. *Paleontological Journal* 34:S333-S343.

**References for general Drophthorine information (hostplants, ranges), as well as information on the symbiosis:**

- Campbell, B. C., T. S. Bragg, and C. E. Turner. 1992. Phylogeny of symbiotic bacteria of four weevil species (Coleoptera: Curculionidae) based on analysis of 16S ribosomal DNA. *Insect Biochem. Molec. Biol.* 22(5): 415-421.
- Charles, H., A. Heddi, J. Guillaud, C. Nardon, and P. Nardon. 1997. A molecular aspect of symbiotic interactions between the weevil *Sitophilus oryzae* and its endosymbiotic bacteria: over-expression of a chaperonin. *Biochemical and Biophysical Research Communications* 239: 769-774.
- Charles, H., H. Ishikawa, and P. Nardon. 1995. Presence of a protein specific of endocytobiosis (symbionin) in the weevil *Sitophilus*. *C. Royal Academy of Sciences, Paris, Life Sciences* 318: 35-41.
- Charles, H. and P. Nardon. 1999. Intracellular symbiotic bacteria within insects. In Seckbach, J. (ed.) *Enigmatic microorganismes and life in extreme environments*. Netherlands: Kluwer pp.651-660.
- Charles, H., G. Condemine, C.Nardon, and P. Nardon. 1997. Genome size characterization of the principal endocellular eymbiotic bacteria of the weevil *Sitophilus oryzae*, using pulsed field gel electrophoresis. *Insect Biochemistry and Molecular Biology* 27: 345-350.
- De la Chapelle, B., J. Guillaud, and P. Nardon. 1992. Glycine conversion to sarcosine in the aposymbiotic weevil *Sitophilus oryzae* L. *Symbiosis* 12:261-274.

- Delobel, B. and A. M. Grenier. 1993. Effect of non-cereal food on cereal weevils and tamarind pod weevil (Coleoptera: Curculionidae). *Journal of Stored Product Research* 29:7-14.
- Delobel, B., Y. Rahbé, C. Nardon, J. Guillaud, and P. Nardon. 1993. Biochemical and cytological survey of tyrosin Frank, J.H., and M.C. Thomas. 1994. *Metamasius callizona* (Chevrolat) (Coleoptera: Curculionidae), an immigrant pest, destroys bromeliads in Florida. *Canadian Entomological Society* 126: 673-682.
- Frank, J.H., and E.D. McCoy. 1995. Invasive adventive insects and other organisms in Florida. *Florida Entomological Society* 78: 1-15.
- Giblin-Davis, R. M., J. E. Peña and R. E. Duncan. 1994. Lethal pitfall trap for evaluation of semiochemical mediated attraction of *Metamasius hemipterus sericeus* (Coleoptera: Curculionidae). *Florida Entomological Society* 77 :247-255.
- Giblin-Davis, R. M., J. E. Peña and R. E. Duncan. 1996. Evaluation of entomogenous nematodes and chemical insecticides for control of *Metamasius hemipterus sericeus* (Olivier) (Coleoptera:Curculionidae). *Journal of Entomological Science* 31: 240-251.
- Grenier, A. M., M. Mbaiguinam, and B. Delobel. 1997. Genetical analysis of the ability of the rice weevil *Sitophilus oryzae* (Coleoptera: Curculionidae) to breed on split peas. *Heredity* 79:15-23.
- Grenier, A.M., C. Nardon, and P. Nardon. 1994. The role of symbiotes in flight activity of *Sitophilus* weevils. *Entomologia-Experimentalis-et-Applicata* 70: 201-208.
- Grenier, A.M. and P. Nardon. 1994. The Genetic Control of Ovariole Number in *Sitophilus oryzae* L. (Coleoptera: Curculionidae) is Temperature Sensitive. *Genetics Selection Evolution, Paris* 26: 413-430.
- Heddi, A., H. Charles, C. Khatchadourian, G. Bonnot, and P. Nardon. 1998. Molecular Characterization of the Principal Symbiotic Bacteria of the Weevil *Sitophilus oryzae*: A Peculiar G+C Content of an Endocytobiotic DNA. *Journal of Molecular Evolution* 47: 52-61.
- Kern, W.H., and P.G. Koehler. 1998. Rice Weevil, *Sitophilus oryzae* (Coleoptera: Curculionidae). Publication ENY-261 of the University of Florida Cooperative Extension Service, June 1998. Found at <http://hammock.ifas.ufl.edu/text/ig/25151.html>
- Lyon, W. Granary and Rice Weevils. Publication HYG-2088-97 of Ohio State University Extension. Found at <http://www.ag.ohio-state.edu/~ohioline/hyg-fact/2000/2088.html>.

- Murphy, S. T. and B. R. Briscoe. 1999. The red palm weevil as an alien invasive: biology and the prospects for biological control as a component of IPM. PEST CABWeb: Biocontrol News and Information 20: 1, 35N-46N.
- Nardon, P., A. M. Grenier, and A. Heddi. 1998. Endocytobiote control by the host in the weevil *Sitophilus oryzae*, Coleoptera, Curculionidae. Symbiosis 25: 237-250.
- Nardon, P., C. Nardon, B. Delobel, Y. Rahbe, and J. Guillaud. 1992. Characteristics and development of the tyrosine-rich protein granules in the adipose tissue of the curculionid beetle *Sitophilus oryzae*. Tissue and Cell 24(2): 157-170.
- Nardon, P., and A. M. Grenier. 1989. Endocytobiosis in Coleoptera: biological biochemical and genetical aspects. Pp. 175-216 in Werner Schwemmler, ed. Insect Endocytobiosis: Morphology, Physiology, Genetics, Evolution. CRC Press, Boca Raton, Florida.
- Nardon, P., and A. M. Grenier. 1991. Serial endosymbiotic theory and weevil evolution. Pp. 153-169 in L. Margulis and R. Fester, eds. Symbiosis as a source of evolutionary innovation : speciation and morphogenesis. MIT Press, Cambridge.
- Nardon, P., and A. M. Grenier. 1993. Symbiose et évolution. Annales de la Société Entomologique de France 29:113-140.
- Nardon, P., and C. Nardon. 1998. Morphology and cytology of symbiosis in insects. Annales de la Société Entomologique de France 34:105-134.
- Nardon, P. 1999. Reproduction and development: Main successes and perspectives Annales de la Societe Entomologique de France 35:54-58.
- Nardon, P., A.M. Grenier, and A. Heddi. 1998. Endocytobiote control by the host in the weevil *Sitophilus oryzae*, Coleoptera, Curculionidae. Symbiosis 25:237-250.
- Nardon, P., Nardon, C., Delobel, B., Rahbe, Y., and Guillaud, J. (1992) Tissue and Cell 24: 157-170. Characteristics and Development of the Tyrosine-Rich Protein Granules in the Adipose Tissue of the Curculionid Beetle *Sitophilus oryzae*.
- O'Brien, C. W., M. C. Thomas, and J. H. Frank. 1991. A new weevil pest of *Tillandsia* in South Florida. Journal of the Bromeliad Society 40(5): 203-205, 222.
- Peña, J. E., R. M. Giblin-Davis and R. Duncan. 1995. Impact of indigenous *Beauveria bassiana* (Balsamo) Vuillemin on banana weevil and rotten sugarcane weevil (Coleoptera: Curculionidae) populations in banana in Florida. Journal of Agricultural Entomology 12: 163-167.
- Perez, A.L., Y. Campos, C.M. Chinchilla, A.C. Oehlschlager, G. Gries, R. Gries, R.M.

- Giblin-Davis, G. Castrillo, J.E. Peña, R.E. Duncan, L.M. Gonzalez, H.D. Pierce, Jr., R. McDonald and R. Andrade. 1997. Aggregation pheromones and host kairomones of West Indian Sugarcane Weevil, *Metamasius hemipterus sericeus*. *Journal of Chemical Ecology* 23: 869-888.
- Pintureau, B., A. M. Grenier and P. Nardon. 1991. Polymorphism of esterases in three species of *Sitophilus* (Coleoptera: Curculionidae). *J. Stored Prod. Res.* 27(3): 141-151.
- Satterthwait, A. F. 1931. Key to known pupae of the genus *Calendra*, with host-plant and distribution notes. *Annals of the Entomological Society of America* XXIV(1):143-172.
- Sosa, O. 1995. The West Indian cane weevil and the sugarcane rootstalk borer weevil-likely pests of sugarcane in Florida. *Sugar. J.* 58: 27-29.
- Sosa, O., J. Shine and P. Tai. 1997. West Indian cane weevil (Coleoptera: Curculionidae): a new pest of sugarcane in Florida. *Journal of Economic Entomology* 90: 634-638.
- Vaurie, P. 1966. A revision of the Neotropical genus *Metamasius* (Coleoptera: Curculionidae, Rhynchophorinae). Species groups I and II. *Bulletin of American Museum of Natural History* 131: 213- 337.
- Vaurie, P. 1967. The nawradii species group of *Rhodoabaenus* (Coleoptera, Curculionidae, Rhynchophorinae). *American Museum Novitates* 2310: 1-36.
- Vaurie, P. 1968. A new genus of weevils from South America (Coleoptera, Curculionidae, Rhynchophorinae). *American Museum Novitates* 2338: 1-14.
- Vaurie, P. 1970. Weevils of the tribe Siplini (Coleoptera, Curculionidae, Rhynchophorinae) part 1. The genera *Rhinostomus* and *Yuccaborus*. *American Museum Novitates* 2419: 1-57.
- Vaurie, P. 1970. Weevils of the tribe Siplini (Coleoptera, Curculionidae, Rhynchophorinae) part 2. The genera *Mesocordylus* and *Orthognathus*. *American Museum Novitates* 2441: 1-78.
- Vaurie, P. 1971. Weevils of the tribe Siplini (Coleoptera, Curculionidae, Rhynchophorinae) part 3. The genus *Sipalinus*. *American Museum Novitates* 2463: 1-43.
- Woodruff, R.E. and R.M. Baranowski. 1985. *Metamasius hemipterus* (Linnaeus) recently established in Florida (Coleoptera: Curculionidae). Florida Dept. Agric. and Consumer Serv. Division of Plant Industry, Entomology Circular No. 272. 4 pp.
- Zimmerman, E. C. 1993. Australian Weevils. Volume III. CSIRO.

**Table legends:**

Table 1: Primers used for sequences in this study.

Table 2: Hypothesis test results. All trees are based on all three loci unless otherwise noted. In cases where the number of trees found was few, the results for each are reported. Where multiple topologies were found, tests were done on all of them, but only the mean result  $\pm$  one standard deviation was reported. In comparisons with most parsimonious total evidence trees, comparisons with the second total evidence tree, the one with higher likelihood score, were reported, though scores barely differ from those calculated with the first total evidence tree (results not shown). Double lines separate groups of trees used in a given comparison — all trees in the first section (test family I) were compared with the MP total evidence tree, all trees in the last section were compared with the 28S gene tree, for example. Note that for less than completely resolved constraints, such as monophyly of *Metamasius* (two nodes specified as polytomies), the Bayesian probability is for *all* trees that could have that group monophyletic, not the probability of the most parsimonious topology with that group.

Table 1: Primers used in this project

**COI**

S1460: TACAATTTATCGCCTAAACTTCAGCC  
S1514: ACCAATCATAAAAATATTGG  
S1541: TGAICYGGAATASTAGGANCATC  
S1718: GGAGGATTTGGAAATTGATTAGTTCC  
S1847: GGAGCAGGAACAGGTTGAAC  
S1859: GGAACIGGATGAACWGTTTAYCCICC  
A1969: CCTTTAGGTCGTATATTAATTAC  
S1991: GTAATTAATATACGACCTAAAGG  
S2183: CAACATTTATTTTGATTTTTTGG  
S2191: GAAGTTTATATTTTAATTTTACCRG  
A2191: CCCGGTAAAATTTAAAATATAAACTTC  
A2442: GCTAATCATCTAAAACTTTAATTCCWGTWG  
S2442: CCAACAGGAATTTAAAATTTTAGATGATTAGC  
A2542: GTAATATCAATTGATRAATTAGC  
A2771: GGATARTCAGARTAACGTCGWGGTATWC  
A2963: AGGRAGTTCATTATAIGAATGTTC  
A3014: TCCAATGCACTAATCTGCCATATTA

**EF-1A**

EFA785: ARAGCTTCRTGRTGCATTTTC  
EFS149: GARAARGARGCNCARGARATGGG  
EFSI: GTCGGTGTCAACAAAATGG  
EFSII: GGTTACAATCCNGCTGCTG  
EFSIII: CTCTTATTGAYGCTTTGGATGC  
EFSIV: GCCAACATCACCCTGAAG  
EFAlI: CAGCAGCNGGATTGTAACC  
EFAlII: GCATCCAAAGCRTCAATAAGAG  
EFAlIV: GGTGGGAGAATRG CRTCCAAAG  
EFAlV: CCACCAATTTTGTAGACATC  
EFAlVI: CATTTCAACAGACTTTACTTC  
EFAlVII: GGGTGGGTTGTTCTTYGAGTC  
EFA923: ACGTTCTTCACGTTGAARCCAA  
EFA1106: GTATATCCATTGGAAATTTGACCNGGRTGRTT

**28S**

28SS3660: GAGAGTTMAASAGTACGTGAAAC  
28SS1: GACCCGTCTTGAAMCAMGGA  
28SA1: TCCKGKTTCAAGACGGGTC  
28SA335: TCGGARGGAACCAGCTACTA  
28SA160: CGCCTCTTCTCGCAATGAGA  
28SA247: CCTGACTTCGTCCTGACCAGGC

Table 2: Hypothesis testing

Test #	Tree type	Bayesian posterior probability	Tree number	Length (parsimony)	p-value		
					Kishino-Hasegawa	Templeton	Winning-sites
I.1	MP total evidence, all taxa	<0.00013	1	4110	best	best	best
			2	4110	best	best	best
I.2	<i>Sitophilus</i> <u>not</u> a clade	<0.00013	1	4151	0.0082**	0.0080**	0.0241*
			2	4151	0.0044**	0.0043**	0.0115*
I.3	<i>Metamasius</i> as clade	<0.00013	1	4215	0.1692	0.1714	0.1636
			2	4215	0.1317	0.1349	0.1017
I.4	<i>Rhynchophorus</i> as clade	<0.00013	1	4123	0.0067**	0.0067**	0.0106*
			2	4123	0.3754	0.3776	0.3776
I.5	<i>Sphenophorus</i> as clade	<0.00013	1	4215	<0.0001***	<0.0001***	<0.0001***
			2	4215	<0.0001***	<0.0001***	<0.0001***
I.6	All Bayesian topologies	1	1-459	4230 ± 6	0.223 ± 0.133	0.224 ± 0.132	0.214 ± 0.146
I.7	Bayesian topology with best posterior prob.	0.17	1	4123	0.3883	0.3927	0.3506
I.8	Poales-feeders as clade	<0.00013	1-2	4373	<0.0001***	<0.0001***	<0.0001***
I.9	Arecaceae-feeders as clade	<0.00013	1	4334	<0.0001***	<0.0001***	<0.0001***
I.10	Bromeliad-feeders as clade	0.00013	1	4115	0.1656	0.1655	0.2668
			2	4115	0.3842	0.3841	0.4862
I.11	Zingiberales-feeders as clade	<0.00013	1-5	4211	<0.0001***	<0.0001***	<0.0001***
I.12	Liliales-feeders as clade	<0.00013	1-3	4150	<0.004**	<0.004**	<0.005**
I.13	Poales-borers as clade	<0.00013	1-8	4158	<0.006**	<0.007**	<0.006**
I.14	Monocot-feeders as clade	<0.00013	1	4405	<0.0001***	<0.0001***	<0.0001***
II.1	MP total evidence, symbiont tested taxa plus outgroup	1	1	1720	best	best	best
II.2	<i>Sitophilus linearis</i> at base	<0.00013	1	1756	0.0002***	0.0002***	0.0003***
III.1	MP total evidence, symbionts tested taxa	0.444	1	1520	best	best	best
III.2	Symbiont clade 1 not a beetle clade	0.00241	1	1531	0.1725	0.1724	0.2148
III.3	Symbiont clade 2 as beetle clade	<0.00013	1-3	1595	<0.0001***	<0.0001***	<0.0001***
III.4	Symbiont clade 3 as beetle clade	<0.00013	1-3	1588	<0.0001***	<0.0001***	<0.0001***

Table 2 (cont): Hypothesis testing

Test #	Tree type	Bayesian posterior probability	Tree number	Length (parsimony)	p-value		
					Kishino-Hasegawa	Templeton	Winning-sites
IV.1	MP total evidence, taxa without COI pruned, but length on all loci	<0.00013	1	3989	best	best	best
			2				
IV.2	COI tree, but length on all loci	<0.00013	1	4077	0.0001***	<0.0001***	0.0002***
V.1	COI tree, length on COI only	<0.00013	1	2673	best	best	best
V.2	MP total evidence, taxa without COI pruned, length on COI only	<0.00013	1	2707	0.0683	0.1033	0.2002
			2				
VI.1	MP total evidence, taxa without EF-1a pruned, but length on all loci	<0.00013	1	3847	best	best	best
			2	3847	best	best	best
VI.2	EF-1a tree, but length on all loci	<0.00013	1	3880	0.079	0.0801	0.0212*
VII.1	EF-1a tree, length on EF-1a only	<0.00013	1	822	best	best	best
VII.2	MP total evidence, taxa without EF-1a pruned, length on EF-1a only	<0.00013	1	839	0.113	0.123	0.1013
			2	833	0.2833	0.2999	0.2679
VIII.1	MP total evidence, taxa without 28S pruned, but length on all loci	<0.00013	1	4012	best	best	best
			2	4012	best	best	best
VIII.2	28S tree, but length on all loci	<0.00013	1-39	4047 ± 11	0.082 ± 0.122	0.084 ± 0.123	0.114 ± 0.155
IX.1	28S tree, length on 28S only	<0.00013	1-39	556	best	best	best
IX.2	MP total evidence, taxa without 28S pruned, length on 28S only	<0.00013	1	564	0.2384	0.2384	0.3239
			2	570	0.0521	0.0527	0.0673

**Figure legends:**

Figure 1: 16S neighbor-joining tree from Heddi et al. (unpublished). Bootstrap values above nodes.

Figure 2: Monocot phylogeny, after Bremer (2000).

Figure 3: Total evidence (COI, EF-1a, 28S) parsimony tree, length 4110. This is one of the two most parsimonious trees: the starred branch collapses in the strict consensus. Bootstrap values are above nodes, partitioned Bremer values are below nodes.

Figure 4: The single most parsimonious COI tree. Bootstrap values are above the nodes. The branchlengths are parsimony-optimized lengths using COI only.

Figure 5: The single most parsimonious EF-1a gene tree. Bootstrap values are above the nodes. The branchlengths are parsimony-optimized lengths using EF-1a only.

Figure 6: 28S gene tree. This is strict consensus of the 39 parsimony trees. Bootstrap values are above the nodes.

Figure 7: Clock tree: This is the tree from Figure 3 after taxa without COI have been pruned. The likelihood model is TRN+gamma+I (see text). Branchlengths in millions of years appear above the branch. Three values are shown as three different calibrations were used. Calibration A was the oldest known dryophthorine fossil, which sets the minimum age of the group at 33 million years (Zherikhin, 2000). Calibration B used the estimate of 1.5 percent COI divergence per million years of Farrell (2001). The 13.3 percent uncorrected divergence between *Sitophilus zeamais* and *Sitophilus granarius* was used to convert the likelihood branchlengths into percent divergences. This comparison was chosen because it was large enough to be logically comparable to other branchlengths of the tree but least likely out of the significant divergences to be affected by multiple hits.

Figure 8: This is the Bayesian phylogeny. Values above nodes are the posterior probability of the clades (the probability of the clade given the data and the substitution model). The tree was computed by bringing all the tree samples from the Bayesian analysis after the burn-in period into PAUP\* and computing the majority rule consensus. Since the Bayesian search used an MCMCMC algorithm, the probability of a clade is the proportion of times it was found in the search, after the burn-in period.

Figure 9: Ln likelihood score versus MCMCMC generation. The scores reach quasi-stationarity (stop going up) at about 10,000 generations, which was thus chosen as the length of the burn-in period.

Figure 10: This is a strict consensus of the two most parsimonious trees and the Bayesian tree (Figure 9). This shows the agreement between the trees.

Figure 11: This is the optimization of plant host on the (A) strict consensus of the maximum parsimony trees and the (B) Bayesian tree.

Figure 12: This is the optimization of symbiont clade (mapped as a four state character: no symbiont, clade one symbiont, clade two symbiont, and clade three symbiont). States for beetles which have not been examined for symbionts were left as uncertainties. (A) is the strict consensus of the parsimony trees, (B) is the Bayesian tree.

Figure 1: Endosymbiont 16S neighbor-joining phylogeny (Courtesy Heddi et al.)

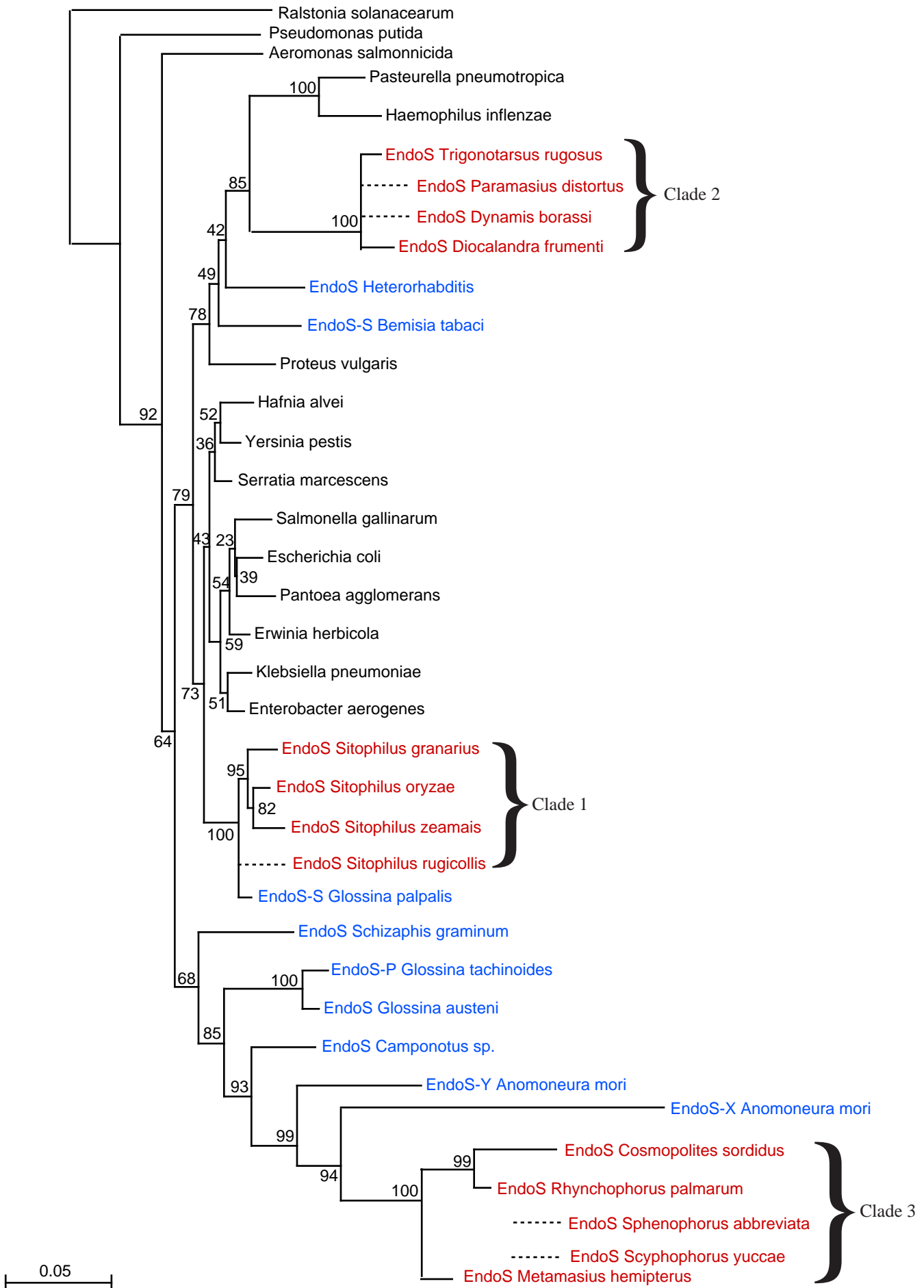
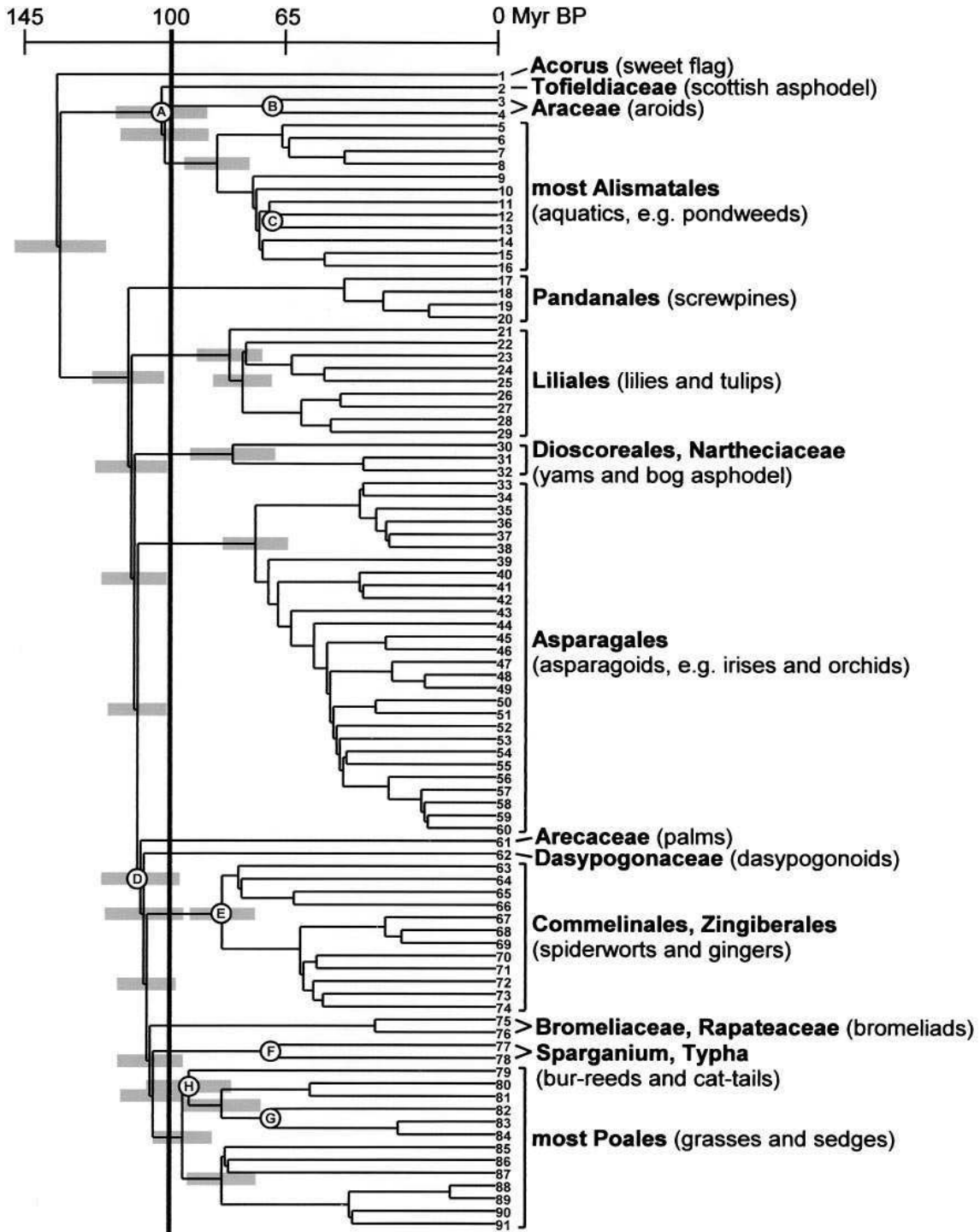


Figure 2: Monocot phylogeny with ages (figure and legend taken from Bremer 2000)



Phylogeny of monocots with major clades that date back to the Early Cretaceous (>100 Myr B.P.). The nodes are arranged approximately by the age estimated from the branch lengths. Reference nodes A–H were used for calculating the change rates. The reference nodes have a minimum age given by the fossils, and nodes B, C, F, and G are positioned accordingly. For nodes A, D, E, and H, the estimated age is greater, and they are thus arranged according to the age obtained from the mean branch length divided by the change rate (see text for calculation of change rates: node A  $75/0.73 = 103$  Myr, node D  $85/0.73 = 116$  Myr, node E  $61/0.73 = 84$  Myr, and node H  $78/0.83 = 94$  Myr). Confidence intervals (95%) for the datings of the basal nodes are indicated by gray bars.

Figure 3 Total Evidence Parsimony Tree

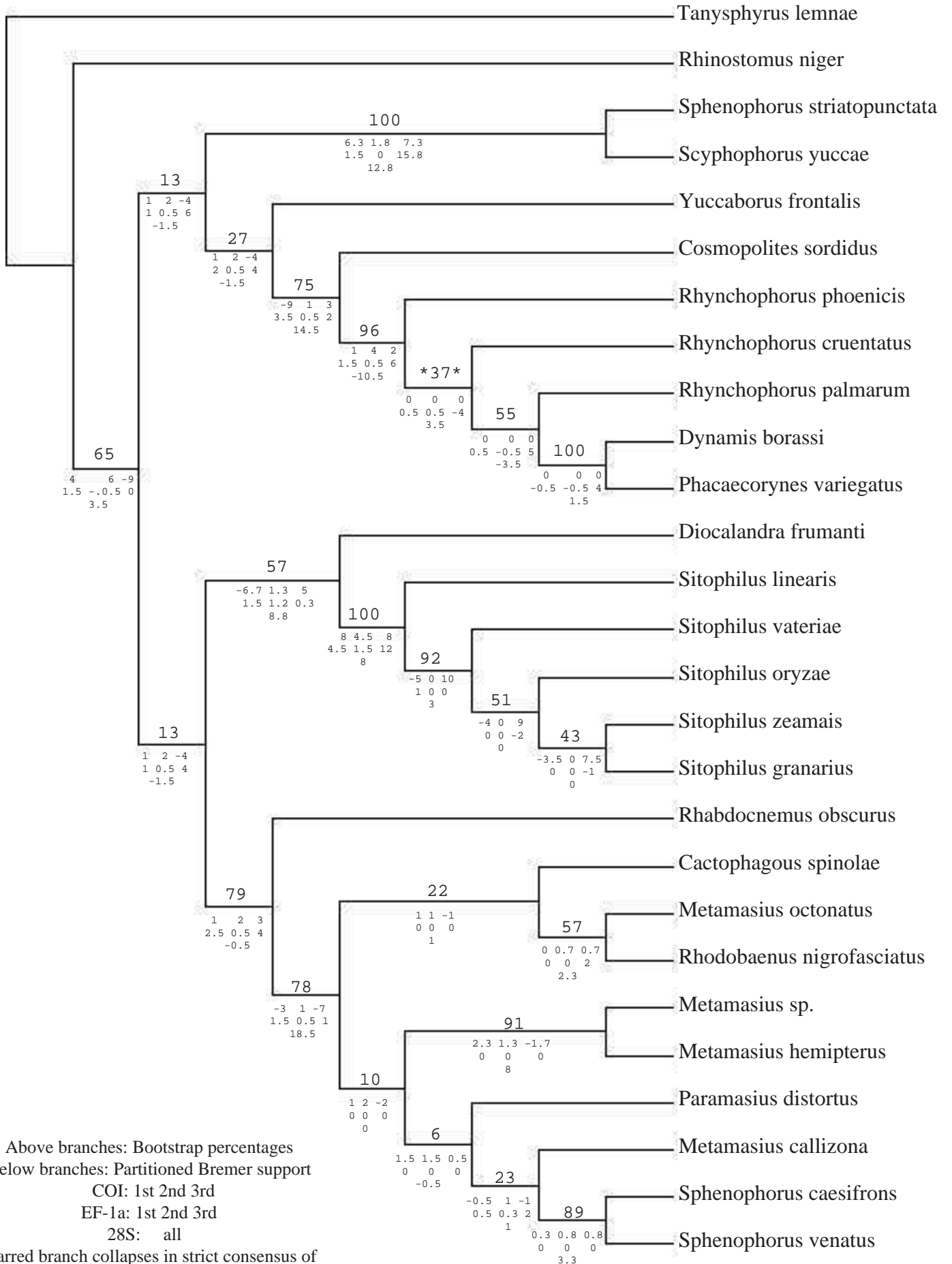
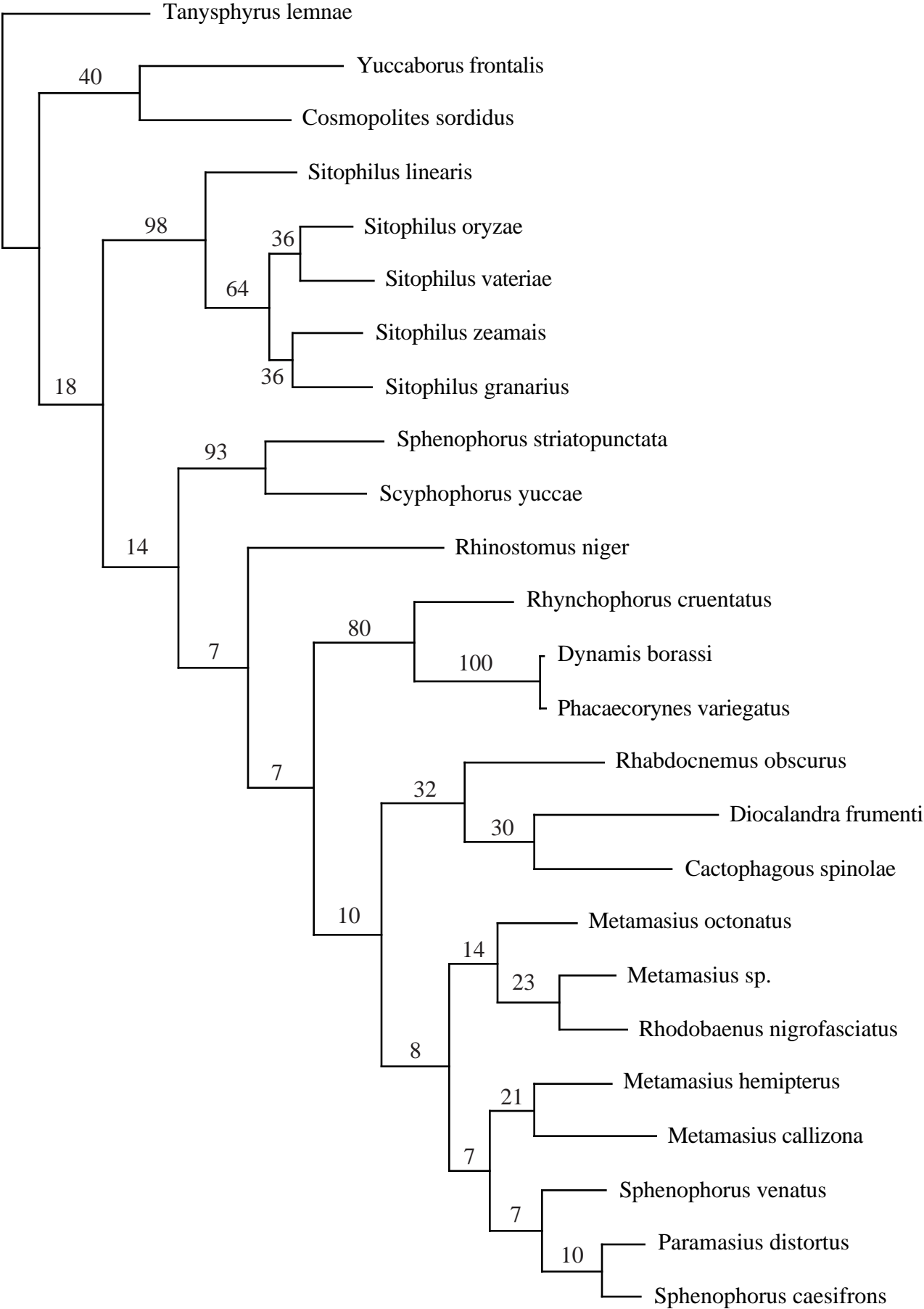
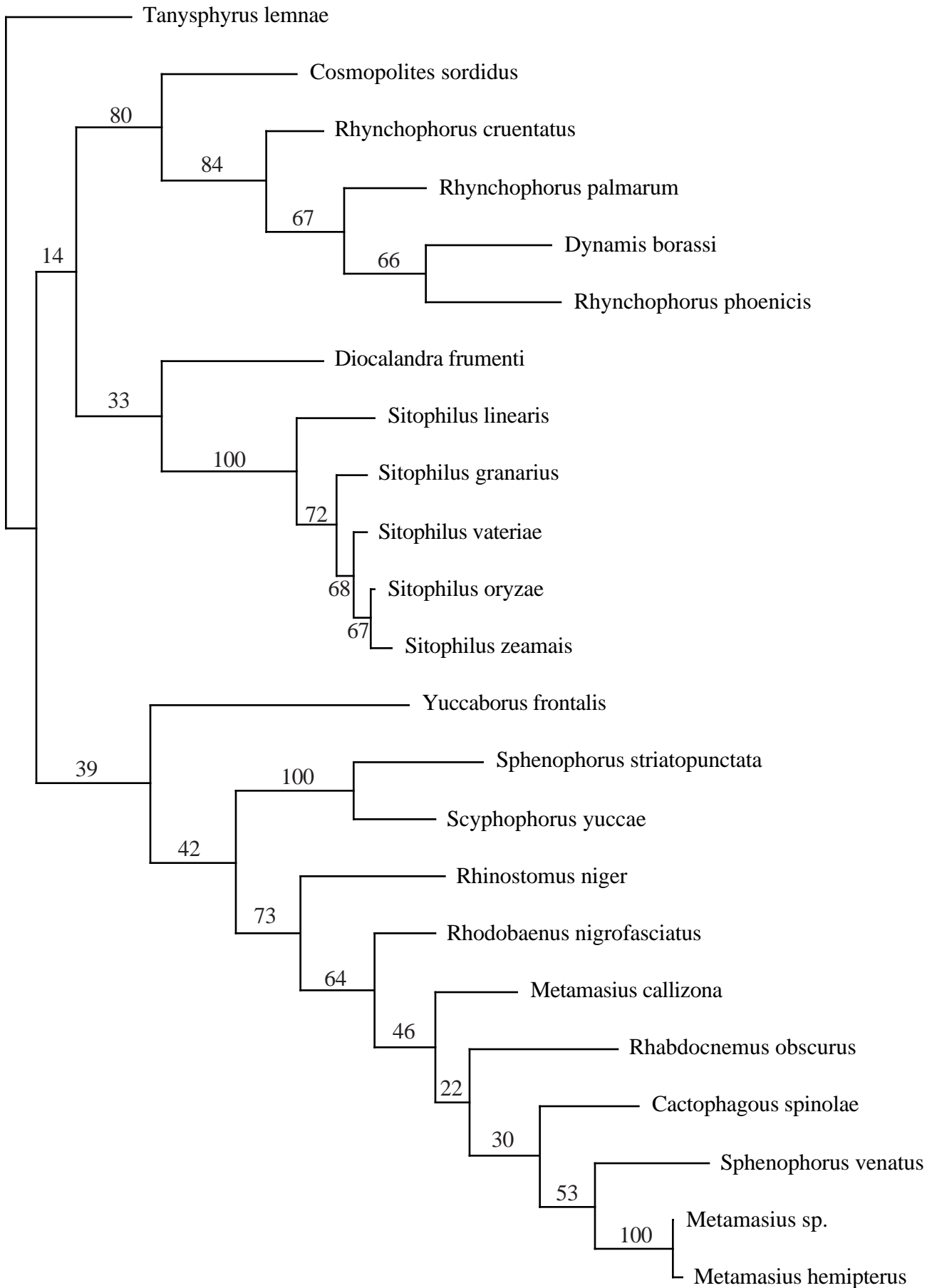


Figure 4: COI Parsimony Tree with Bootstrap Proportions



— 50 changes

Figure 5 EF-1a Parsimony Tree with Bootstrap Proportions



— 10 changes

Figure 6: 28S Parsimony Bootstrap Consensus Tree

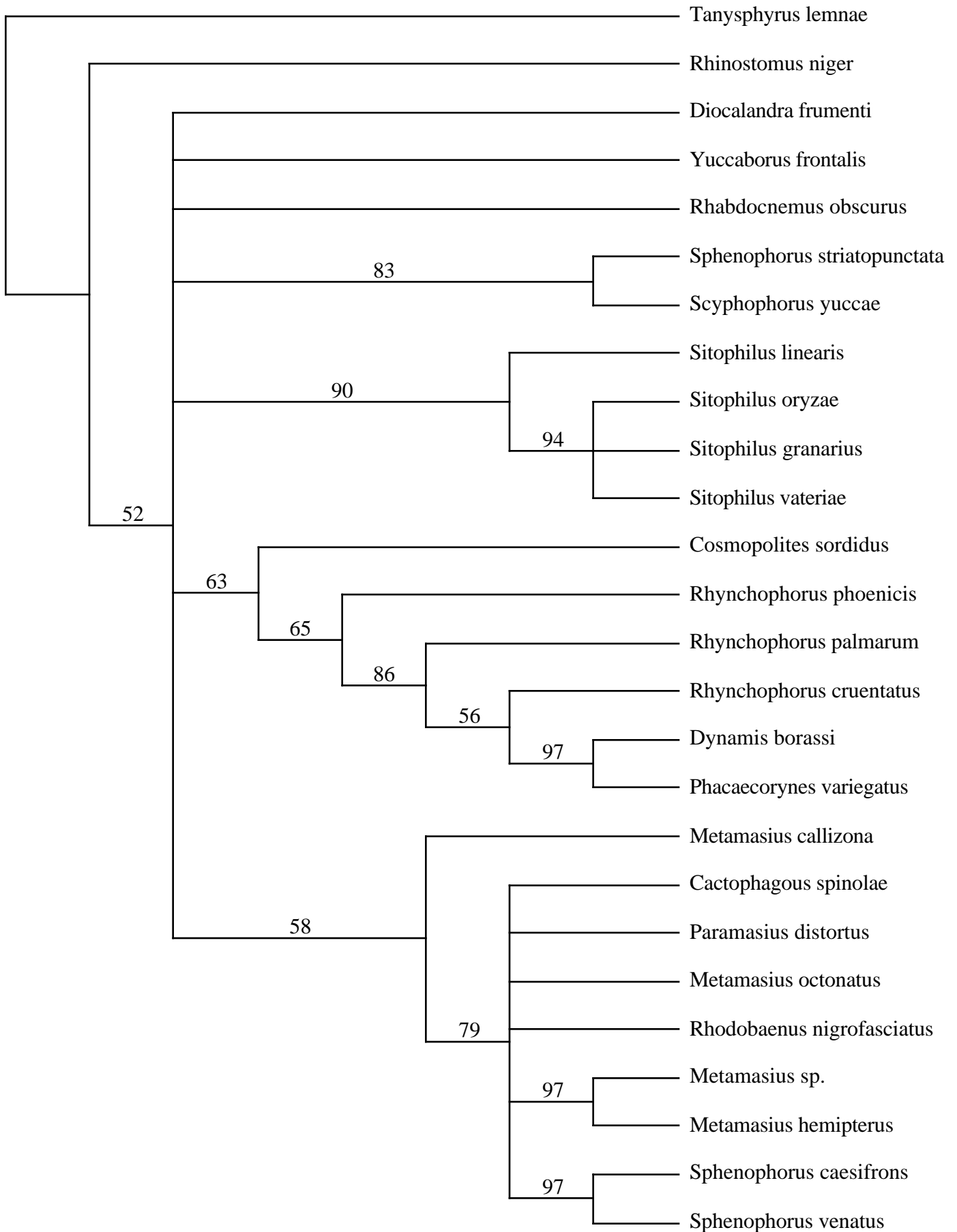
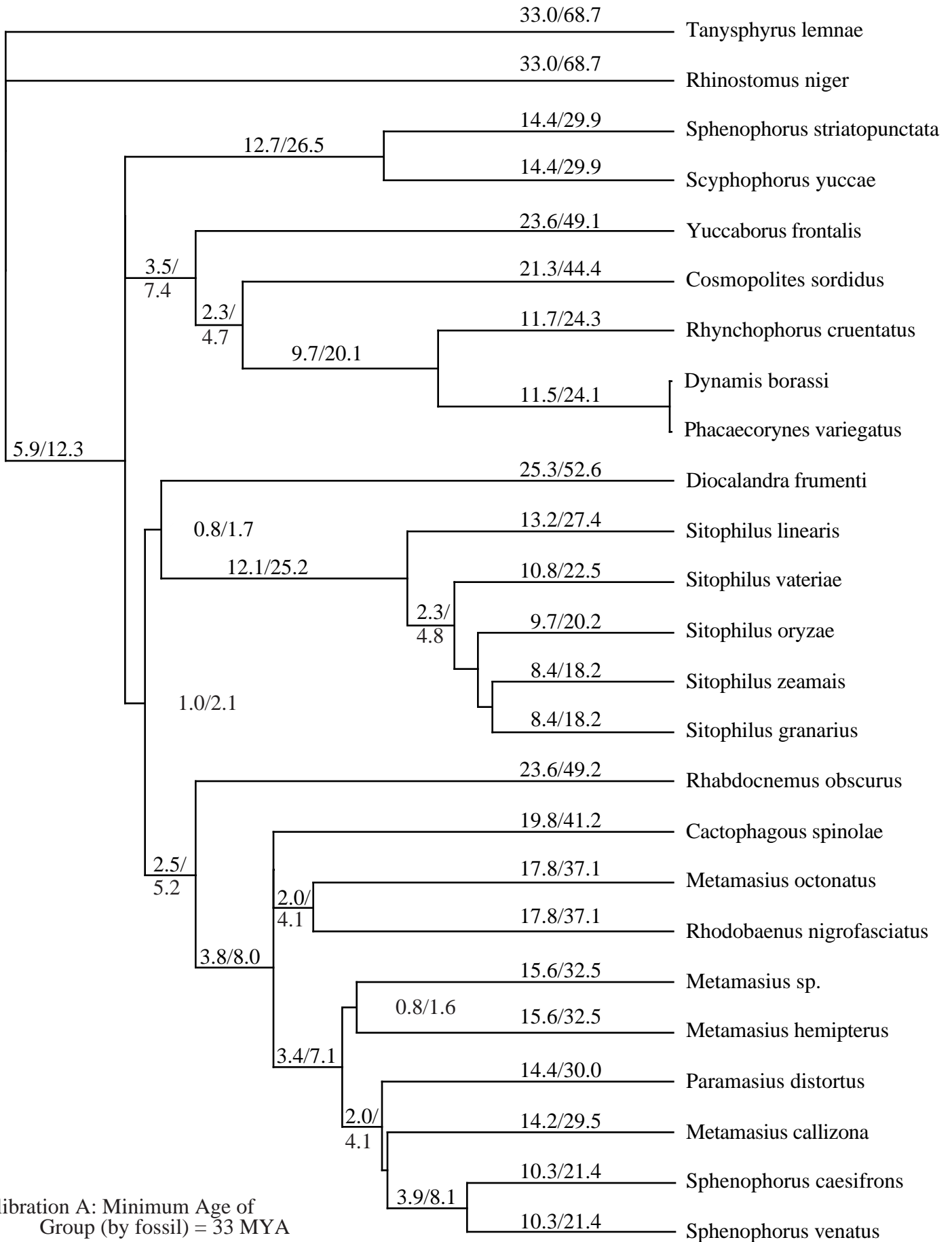


Figure 7: COI clock tree. Numbers are age by fossil/%divergence calibration in millions of years



Calibration A: Minimum Age of Group (by fossil) = 33 MYA  
 Calibration B: 1.5% divergence/MY

Figure 8: Bayesian Tree w/ Posterior Probabilities of Clades

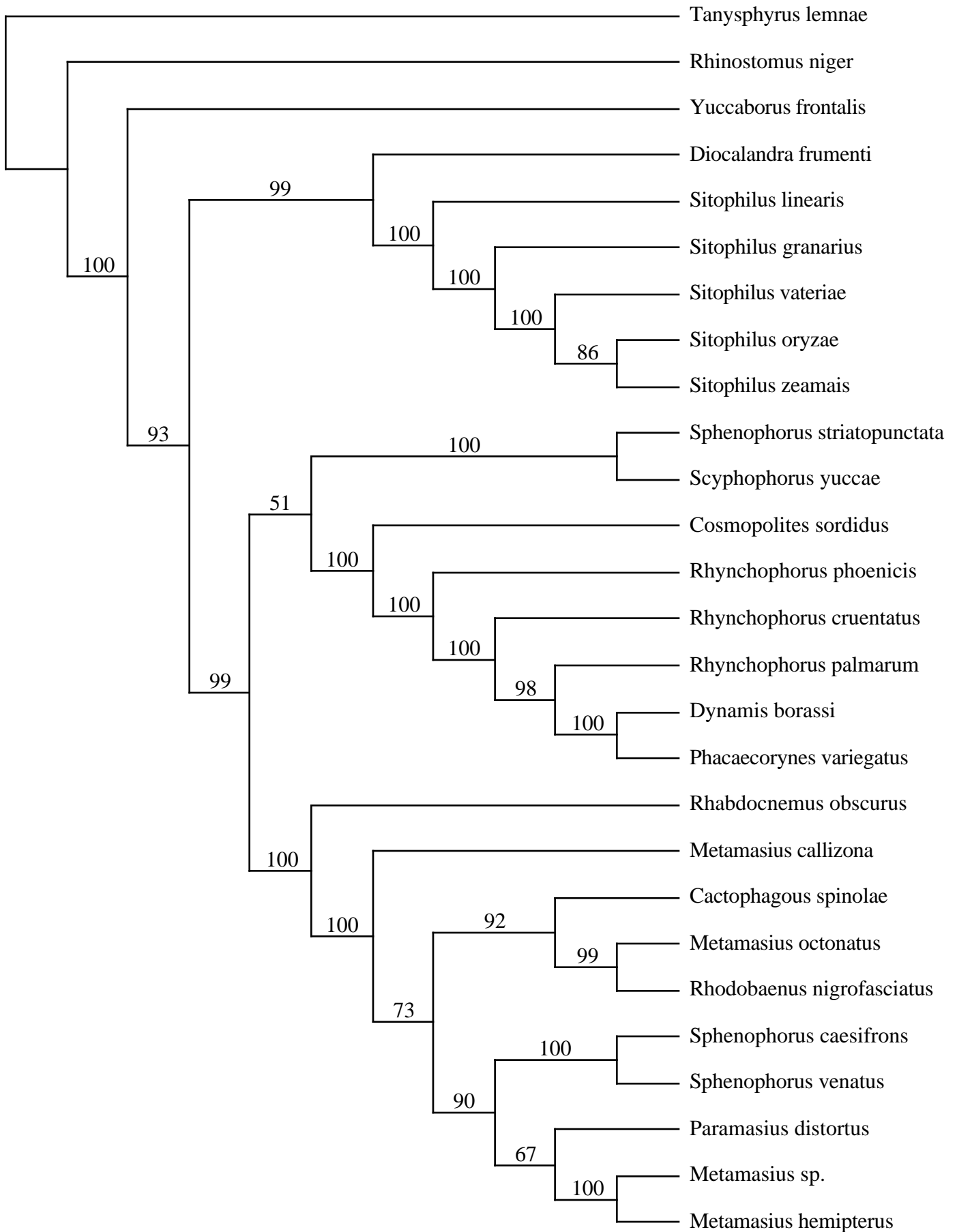


Figure 9: Ln likelihood vs. generation for 18 Bayesian runs

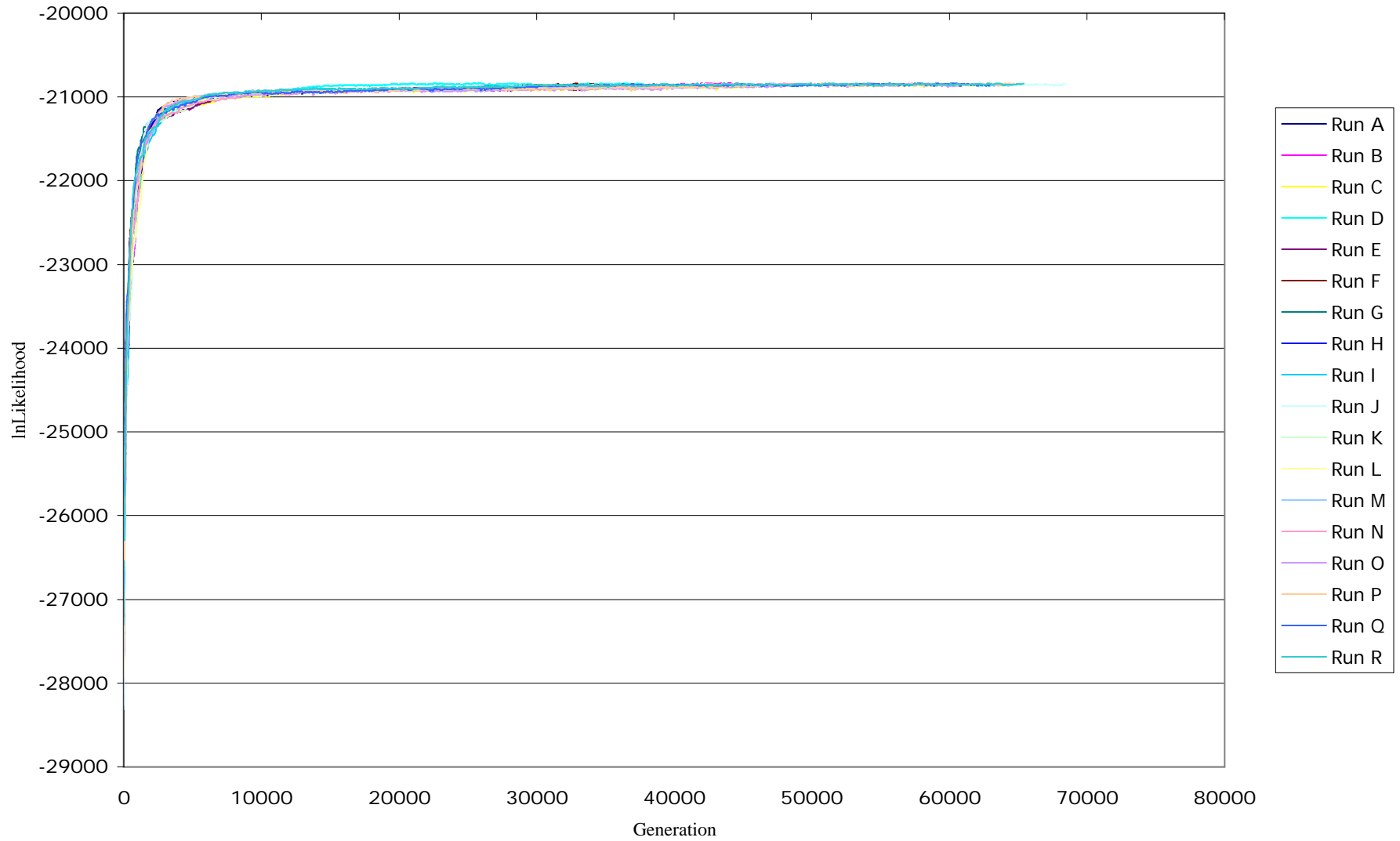


Figure 10: Strict Consensus of Parsimony Trees and Bayesian Tree

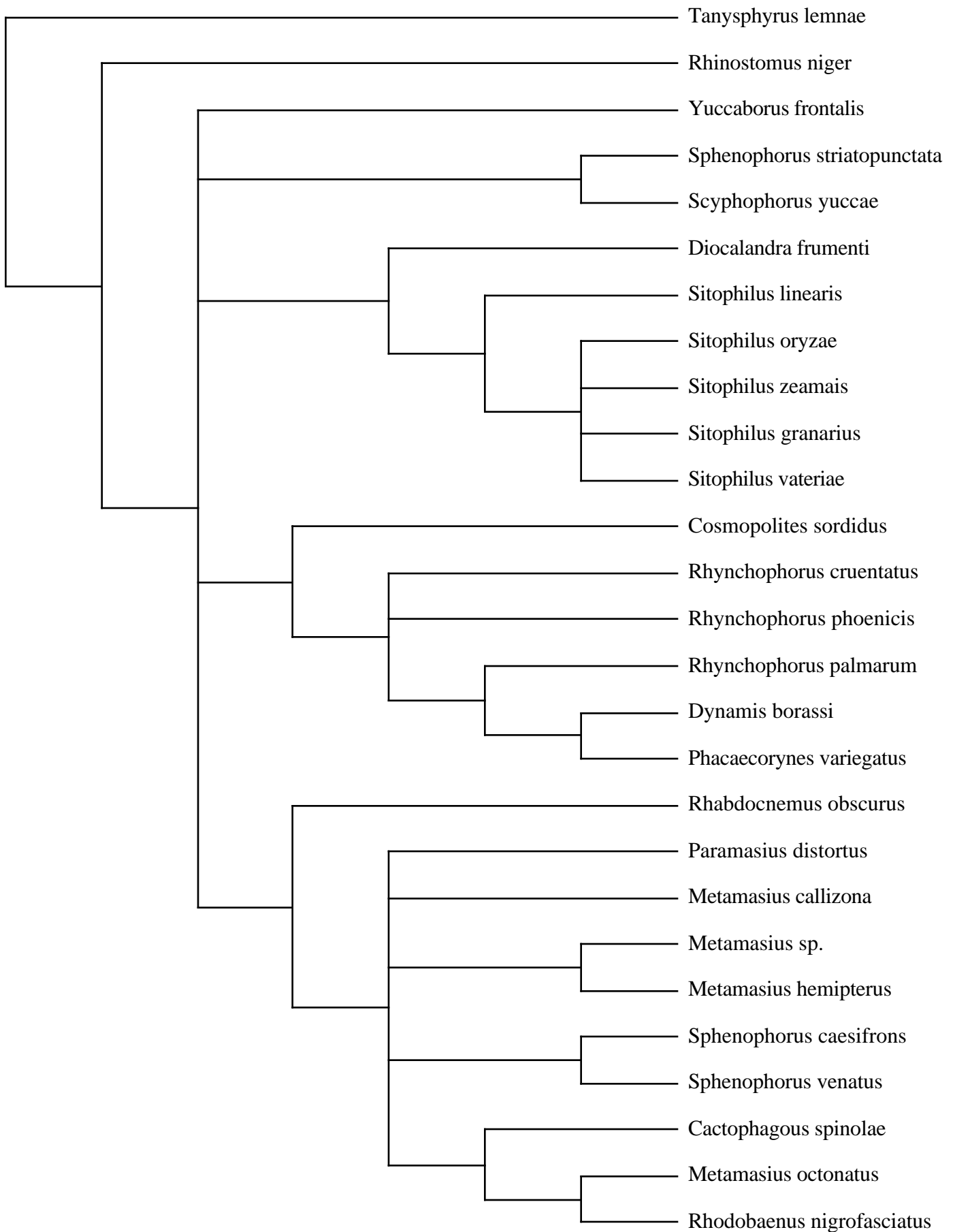


Figure 11a: Host plant mapped on MP tree

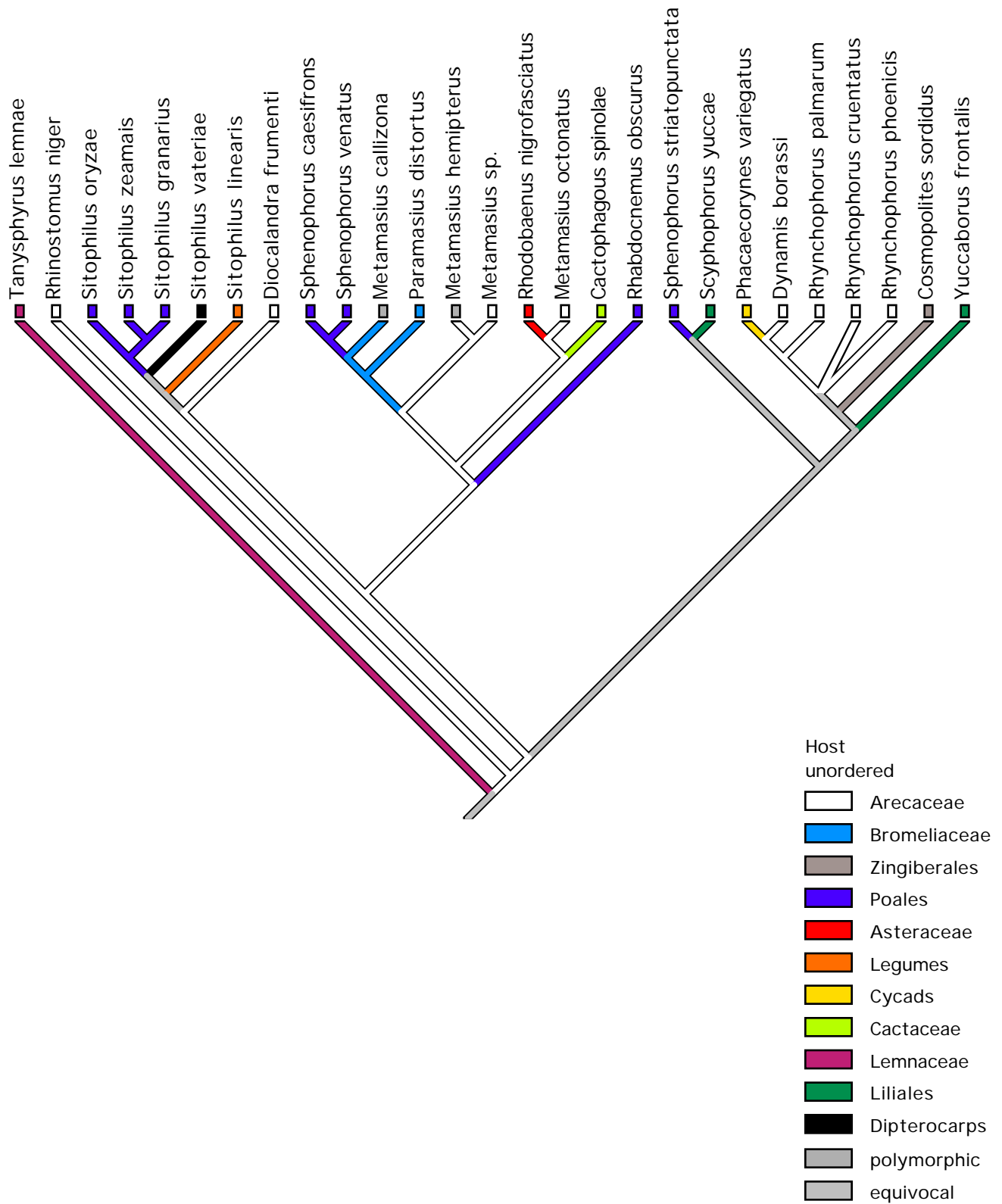


Figure 11b: Host plant mapped on Bayesian tree

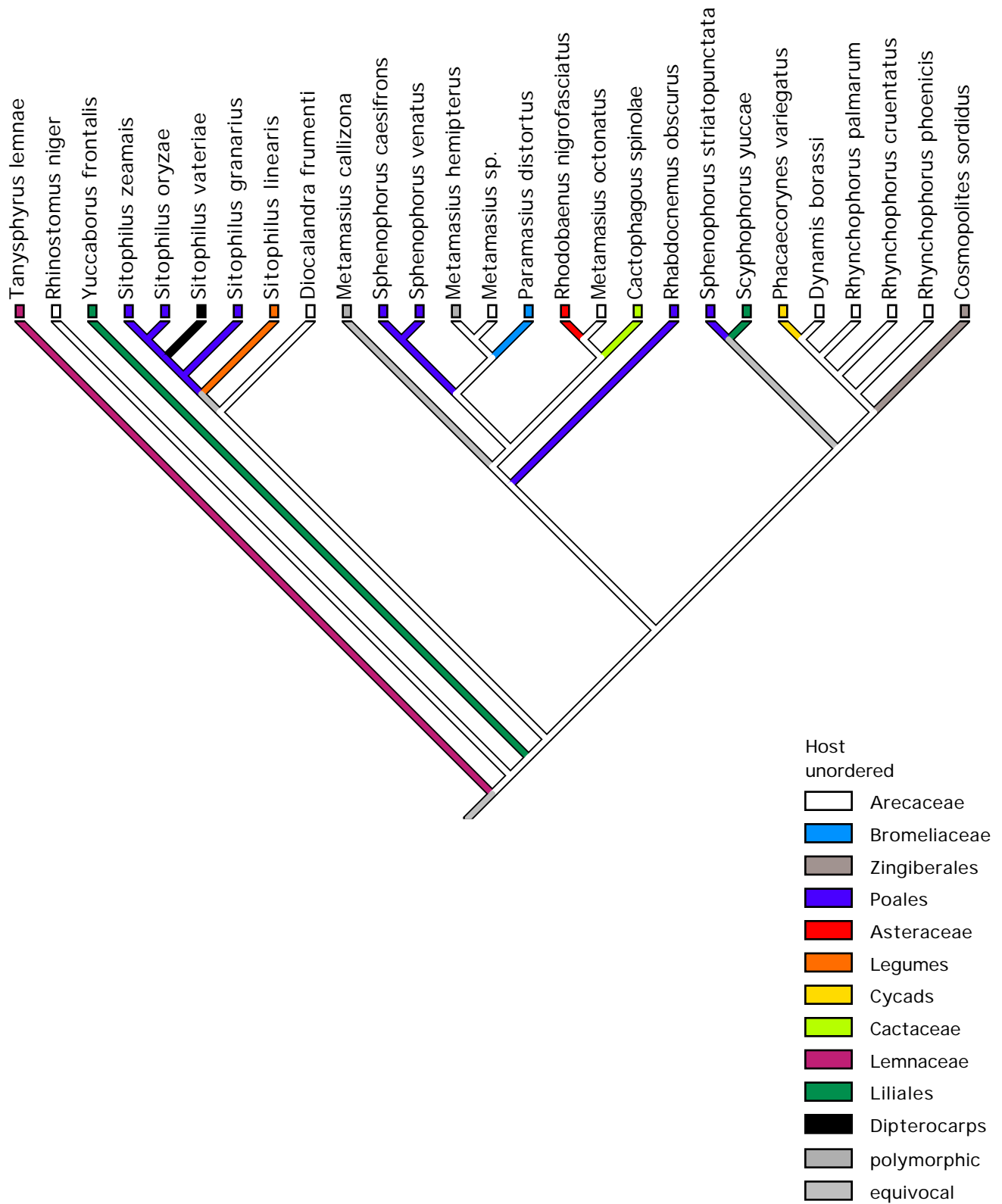


Figure 12a: Symbiont clades mapped on MP trees

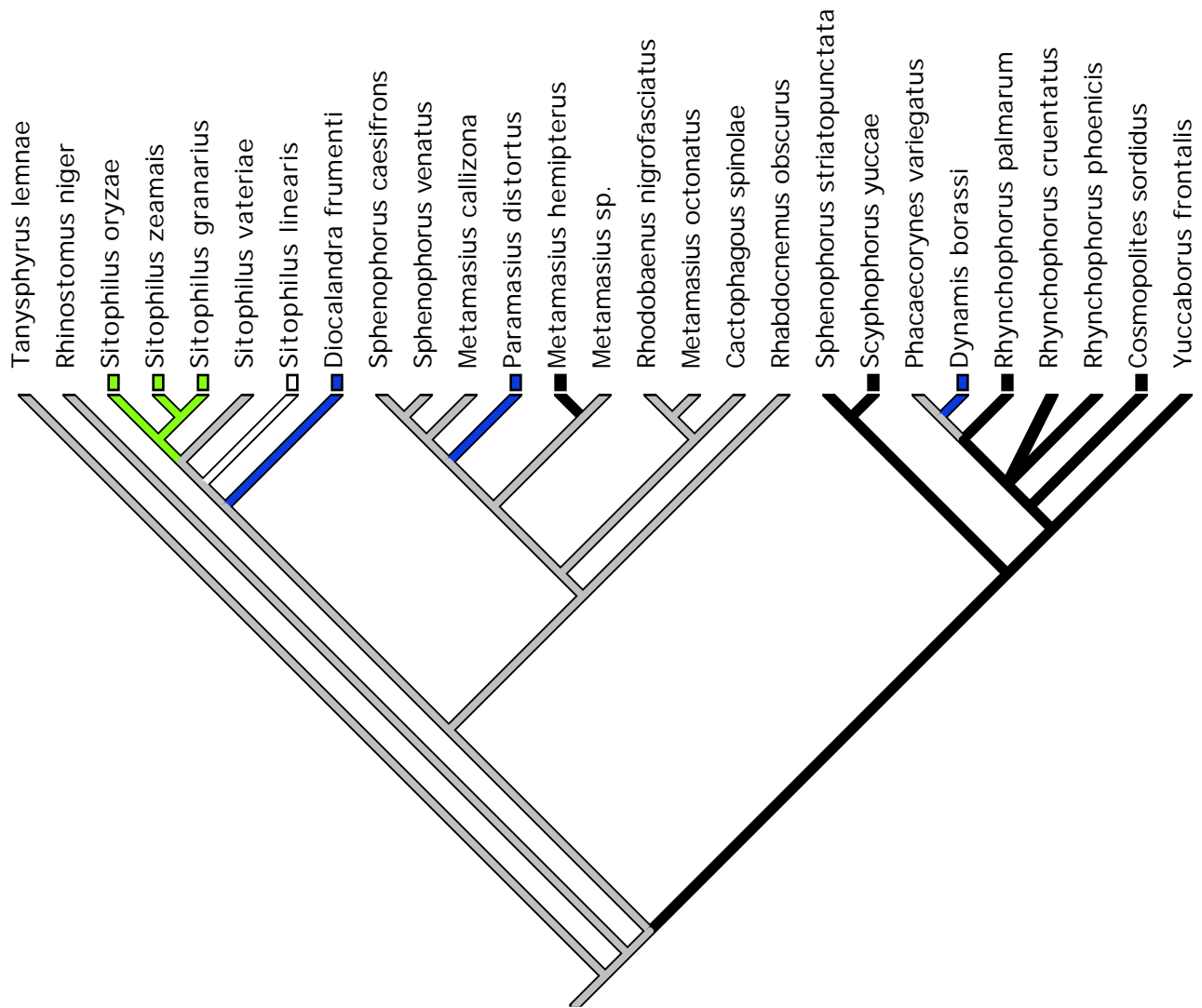
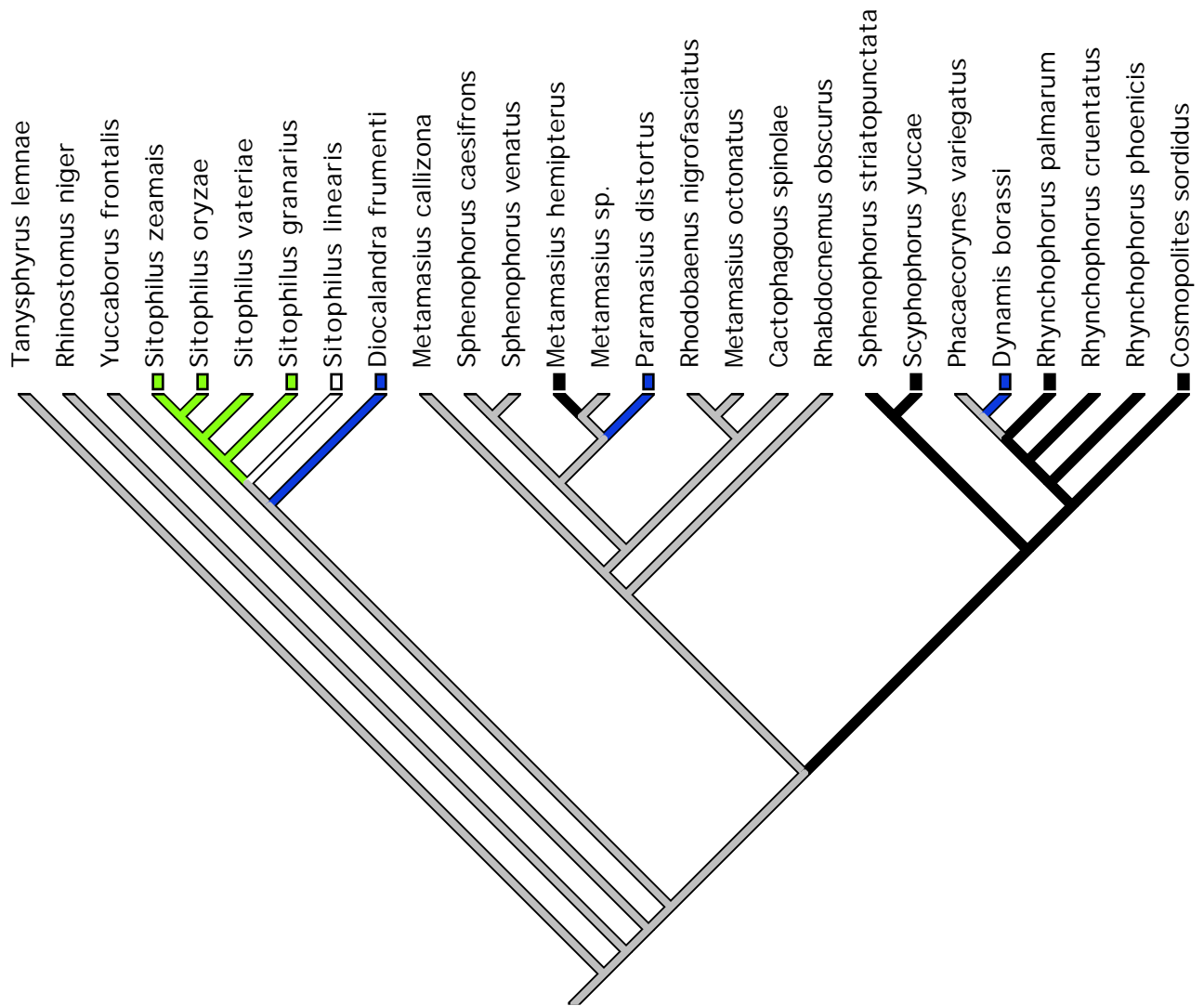


Figure 12b: Symbiont clades mapped on Bayesian tree



Symbiont  
unordered

- None
- Clade 1
- Clade 2
- Clade 3
- equivocal

### Appendix I: Bayesian and MCMCMC analysis

As this is one of the first times this technique has been used in phylogenetic analysis, Bayesian analysis will be explained. This method is based on Bayes' rule, which relates the probability of B given A to the probability of A given B and the probability of B itself. The formula is  $\Pr(B_i|A) = \Pr(A|B_i) \times \Pr(B_i) \div [\Pr(A|B_j) \times \Pr(B_j)]$ , where the sum is from  $j=1$  to  $k$ ,  $\Pr(B_i|A)$  is the probability of  $B_i$  given A (Rosner, 2000). For phylogenetic analysis,  $\Pr(T_i|X)$ , the probability of the  $i$ th tree given dataset X, is  $\Pr(X|T_i) \times \Pr(T_i) \div [\Pr(X|T_j) \times \Pr(T_j)]$ , where the sum is for all trees.  $\Pr(X|T_i)$  is the likelihood value, L (the probability of the data given the tree). Note that likelihood scores (L) are usually reported as  $-\ln L$ , as L is very low (for example, a reported negative  $\ln$  likelihood score of 2000 represents a probability of the data given the tree,  $\Pr(X|T_i)$ , of  $2.57 \times 10^{-869}$ ). The probability of the tree,  $\Pr(T_i)$ , is the reciprocal of the total number of trees. By solving this formula, we can find the probability of every tree given the data.

One way to do this would be to compute the likelihood score of every tree. However, as there are  $5.8436 \times 10^{31}$  possible bifurcating trees for the taxa analyzed, this is rather impractical (Swofford et al, 1996). Bayesian analysis is often done with Markov-chain Monte Carlo (MCMC) methods instead (Huelsenbeck, 2000; Larget and Simon, 1999). Basically, this works by calculating a likelihood score for a tree, proposing a new tree (often derived in some way from the old — branch swapping, change of the gamma

parameter, etc.), and getting a score for the new tree. Only one of the two trees is held for the next step of swapping. Which tree is chosen is partly random and partly based which has the better likelihood score. If the new tree is much better than the old one, there is a high probability of using the new one, but there is less probability of accepting the new one if the trees have similar scores. It is also possible to go to a tree with a worse likelihood score. The tree accepted is then used for the next step of the chain. Most searches start from a significantly suboptimal tree, then the trees get better until the likelihood score reaches an asymptote. The chain wanders around this value for the remainder of the search, sometimes going up in likelihood, sometimes going down. The initial period when the chain is going up to the plateau is the “burn-in” period. Samples from this time are generally not used, as they are trees which were sampled only because they were near the starting tree, not due to any special goodness. However, the trees in the plateau are sampled according to their posterior probabilities: better trees are hit more often by the search, since the probability of moving in their direction is higher. As this search strategy may have trouble crossing steep valleys, a further refinement, called Metropolis-coupled Markov chain Monte Carlo (MCMCMC) is used. This uses multiple chains running in parallel. Some of these chains are “heated,” meaning that the posterior probability is raised by a power less than one (technically, it is raised by  $(1/(1+T(i-1)))$ , where  $T$  is the chain’s “temperature” and  $i$  is the number of the chain). This flattens out

the tree space for some chains, making it easier for the search to cross valleys.

Occasionally, two chains are made to exchange some states, allowing each to move to a different part of tree space.

Bayesian techniques have the advantage of combining a tree search with a search for support, merging the traditional searches of a thorough search for a tree with a less thorough bootstrap search for node support. This results in significant speed advantages over likelihood. A Bayesian search results in the posterior probability of every parameter which is varied in the search, from topology to proportion of invariant sites, to be calculated. It also allows easy tests of other phylogenetic hypotheses. For example, the posterior probability of the monophyly of symbiont clade one can be determined by filtering the total group of all Bayesian topologies for the presence or absence of this clade and calculating the proportion of trees matching the constraint. This was quite easily done in PAUP\*, first loading a constraint tree to use as a filter. This technique will aid in the adoption of a statistical, hypothesis testing mindset in phylogenetics, as was implemented in this thesis (i.e., Table 2).

Bayesian analysis using the method of MCMCMC does have some problems. First, the posterior probabilities are dependent on the likelihood model. An incorrect model will give incorrect posterior probabilities, which, while perhaps not resulting in an incorrect topology, make it inappropriate to interpret the probabilities as the true

probabilities of the observed value of the parameter given the data. Presumably, a more parameter-rich model will better approximate reality and thus give more accurate posterior probabilities. However, the more complex model has a cost in increased computation time. In this analysis, a fairly complex model (GTR [nst=6], gamma-distributed rates, estimated base frequencies, clock not enforced) was used, which is similar to the model chosen by ModelTest for likelihood, though without invariant sites. A philosophy of how to pick a likelihood model for Bayesian analysis has apparently not yet developed, though the likelihood ratio test approach seems to work. Alternatively, researchers could use the most complex model they had time to run, using the model chosen through likelihood ratio tests as the lower bound of complexity. Time requirements for this Bayesian analysis were not terribly severe, requiring about 100 computer-hours of time (compare with the estimate of years for a likelihood search). The use of multiple runs allows the analysis to be run in parallel to some degree. However, each individual run has a minimum number of generations required as a burn-in period, so there is a minimum time required (an hour or so, in this study). Deciding the length of the burn-in period is fairly easy as likelihood scores reach a definite plateau, visible by eye, though more rigorous techniques (perhaps computing the slope for a best fit line over a portion of the graph) could be used to determine when the chains reached stationarity.

MrBayes, the program used in this analysis, was not as user-friendly as it could be, though it represents a significant advance in accessibility of Bayesian analysis for empirical researchers. The main difficulty I had with the program was in an attempted clock analysis. First, the program could not accept a perfectly constrained tree as an input. This meant that the program could change the topology at some branches, which makes recovery of branchlengths for these varying branches rather difficult (for example, PAUP\* cannot calculate a majority-rule consensus tree with branchlengths). MrBayes should be able to do this itself, but the “sumt” command, which should load in trees and compute variances, did not work. As a result, only likelihood optimization was used for the clock tree in this study. More experience with the program may allow me to use it for clock optimization in the future, though.

Evaluating the thoroughness of the MCMCMC search is more difficult. How do we know there are not “islands” of trees which the analysis has not hit yet? For a standard heuristic search, as implemented in PAUP\*, the proportion and number of times a certain island was hit gives some suggestion as to a search’s thoroughness: hitting the best island just once in a short search suggests that a longer search could find more; hitting the best island in every random addition sequence replicate suggests that there are few other islands. This information is not visible from a Bayesian analysis, at least as implemented in MrBayes 1.0. In this study, I attempted to perform a thorough analysis by doing many runs,

which will follow different paths due to the statistical nature of the MCMCMC. The program was allowed to find its own starting trees, though it is unclear from the program documentation how these were found, so it is possible that all 18 runs started from the same tree. The temperature of 0.2 was adequate in this analysis to cause frequent mixing between the four chains in each run, which also helps search treespace and cross valleys. One concern which I have is the low posterior probabilities for the vast majority of the tests done on hypotheses not found in the best trees (Table 2). This could represent a truly low probability of these constraints, or it could represent a tree search which was not thorough enough. I believe the former is the case, but the literature on how thorough a search must be does need to be better developed.

I am optimistic about the future uses of Bayesian analysis in phylogenetics. Searches are more efficient than likelihood searches but should still have the benefits of likelihood. The method combines a topology search with a confidence search, further saving time. Finally, Bayesian analysis using the MCMCMC algorithm provides an easy, user-friendly way to test phylogenetic hypotheses by filtering the Bayes samples within PAUP\*.