A Novel Test for Host-Symbiont Codivergence Indicates Ancient Origin of Fungal Endophytes in Grasses

C. L. SCHARDL,1 K. D. CRAVEN,2 S. SPEAKMAN,3 A. STROMBERG,3 A. LINDSTROM,3 AND R. YOSHIDA3

1Department of Plant Pathology, 201F PSB, 1405 Veterans Drive, University of Kentucky, Lexington, KY 40546-0312, USA
2Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA
3Department of Statistics, University of Kentucky, Lexington, KY 40526-0027, USA; E-mail: ruriko@ms.uky.edu (R.Y.)

C.L.S., K.D.C., S.S. and R.Y. contributed equally to this work

Abstract—Significant phylogenetic codivergence between plant or animal hosts (H) and their symbionts or parasites (P) indicates the importance of their interactions on evolutionary time scales. However, valid and realistic methods to test for codivergence are not fully developed. One of the systems where possible codivergence has been of interest involves the large subfamily of temperate grasses (Pooidae) and their endophytic fungi (epichloae). These widespread symbioses often help protect host plants from herbivory and stresses and affect species diversity and food web structures. Here we introduce MRCALink (most-recent-common-ancestor link) method and use it to investigate the possibility of grass-epichloae codivergence. MRCALink applied to ultrametric H and P trees identifies all corresponding nodes for pairwise comparisons of MRCA ages. The result is compared to the space of random H and P tree pairs estimated by a Monte Carlo method. Compared to tree reconciliation, the method is less dependent on tree topologies (which often can be misleading), and it crucially improves on phylogeny-independent methods such as ParaFit or the Mantel test by eliminating an extreme (but previously unrecognized) distortion of node-pair sampling. Analysis of 26 grass species-epichloae species symbioses did not reject random association of H and P MRCA ages. However, when five obvious host jumps were removed, the analysis significantly rejected random association and supported grass-endophyte codivergence. Interestingly, early cladogenesis events in the Pooidae corresponded to early cladogenesis events in epichloae, suggesting concomitant origins of this grass subfamily and its remarkable group of symbionts. We also applied our method to the well-known gopher-louse data set. [Coevolution; endophytes; grasses; phylogenetic trees.]

Symbioses between cool-season grasses (Poaceae subfamily Pooidae) and fungi of genus Epichloae (including their asexual derivatives in the genus Neotyphodium) are very widespread and occur in a broad taxonomic range of this important grass subfamily. These symbioses span the continuum from mutualistic to antagonistic interactions, making them an especially interesting model for evolution of mutualism, and particularly the possible role of codivergence (Jackson, 2004; Piano et al., 2005; Schardl et al., 1997; Sullivan and Faeth, 2004; Tredway et al., 1999). They have major ecological implications, affecting food web structures (Omacini et al., 2001) and species diversity (Clay and Holah, 1999). These endophytes extensively colonize host vegetative tissues without eliciting symptoms or defensive responses, and in many grass-epichloae symbiosis the endophyte colonizes the embryos and is vertically transmitted with exceptional efficiency (Freeman, 1904; Sampson, 1933, 1937). The asexual endophytes rely upon vertical transmission for their propagation, whereas sexual (Epichloae) species are also capable of horizontal transmission via meiotically derived ascospores (Chung and Schardl, 1997). Mixed vertical and horizontal transmission strategies are also common. The tendency for vertical transmission is predicted to select for mutualism (Bull et al., 1991; Herre, 1993). In fact, many vertically transmitted and even some horizontally transmitted epichloae help protect their hosts from herbivores, nematodes, and other stressors (Clay and Schardl, 2002). Grass-epichloae symbioses occur in most tribes of the Pooidae, the highly speciose subfamily of temperate C3 grasses. Most but not all Epichloae and Neotyphodium species are specialized to individual host species, genera, or tribes in the Pooidae (Schardl and Leuchtmann, 2005). Therefore, it is reasonable to hypothesize a history of Pooidae-epichloae coevolution extending back to the origin of this grass subfamily.

Previous analyses of the grass-epichloae system have suggested some codivergence, along with some host species transfers (jumps; Jackson, 2004; Schardl et al., 1997). However, methods for such comparative phylogenetic studies are in need of refinement. Interpreting evidence for codivergence based solely on congruence and reconciliation of tree topologies has significant shortcomings. Although tree reconciliation is useful particularly to assess strict cospeciation, precise mirror phylogenies for coevolving hosts and symbionts can be an overly restrictive expectation (Legendre et al., 2002; Page and Charleston, 1998). For example, incomplete taxonomic sampling as well as lineage sorting of gene polymorphisms can result in topological incongruence between H and P trees, and deviations from strict cospeciation may mask tendencies for phylogenetic tracking. Refined methods are needed to assess significant patterns of codivergence without strict cospeciation. An attractive approach is to assess the correspondence in timing of cladogenesis events (Hafner et al., 1994). Methods that compare H and P pairwise distance matrices, e.g., by the Mantel test (Hafner et al., 1990) or ParaFit (Legendre et al., 2002), would seem to assess the timing of corresponding cladogenesis events directly. However, comprehensive pairwise distance methods suffer from grossly unequal sampling of corresponding cladogenesis events depending on the depth of the nodes representing those events in the actual H and P trees. The problem is illustrated by the simple examples in Figure 1 (and further documented in Discussion). A statistical test for codivergence should evaluate the relationship between corresponding MRCA (most recent common ancestor)
ancient origins of these symbioses. This same method is applied to a well-known gopher-louse data set for comparison purposes.

**METHODS**

**Fungal Endophyte Isolates and Endophyte-Infected Grasses**

The cool-season grasses and their respective fungal endophytes are listed in Table 1. All endophytes examined were from natural infections from which the corresponding natural host plant or leaf material from this plant was available for chloroplast sequence analysis. Representatives of all available *Epichloë* species and nonhybrid *Neotyphodium* species were included. In most cases where an *Epichloë* species infects multiple host genera, isolates from each genus were sampled. The sole exception was *E. typhina*, for which codivergence can be rejected a priori due to its broad host range (Leuchtmann and Schardl, 1998; Craven et al., 2001). Many *Neotyphodium* species are interspecific hybrids and therefore possess multiple genomes of distinct origin. These were usually excluded because of the difficulty in choosing the appropriate genome for analysis. However, four hybrid endophytes associated with *Lolium* species were included because they possess genomes in a clade (designated LAE, for *Lolium*-Associated Endophytes) that was unrepresented among those of known *Epichloë* species. Each of these four species—*Neotyphodium coenophialum*, *Neotyphodium occultans*, *Neotyphodium* sp. FaTG2, and *Neotyphodium* sp. FaTG3—had a distinct history of interspecific hybridization (Tsai et al., 1994; Moon et al., 2000).

**Sequencing of Chloroplast DNA (cpDNA)**

**Non-Coding Regions**

Genomic DNA was isolated from 0.5 to 1.0 g of harvested endophyte-infected plant leaf material by the CTAB method (Doyle and Doyle, 1990) and dissolved in 1 mL of purified water (Milli-Q; Millipore, Bedford, Massachusetts). DNA was quantified by bisbenzimide fluorescence and measured with a Hoefer (San Francisco, California) DyNA Quant 200 fluorometer.

PCR amplification of one intron (*trnL* intron) and two intergenic spacers (*trnT-trnL*, *trnL-trnF*) from cpDNA was performed from total plant DNA with primers described by Taberlet et al. (1991). Reactions were performed in 50-µL volumes containing 15 mM Tris-HCl, 1.5 mM MgCl₂, 50 mM KCl, pH 8.0 in the presence of 200 µM of each dNTP (Panvera, Madison, Wisconsin), 200 nM of each primer (Integrated DNA Technologies, Coralville, Iowa), 1.25 unit AmpliTaq Gold DNA polymerase (Applied Biosystems, Foster City, California), and 10 ng of genomic DNA. Reactions were performed in a PE Applied Biosystems DNA thermal cycler, with a 9-min preheat step at 95°C to activate the enzyme, followed by 35 cycles of 1 min at 95°C, 1 min at 55°C, and 1 min at 72°C. All amplification products were verified by 0.8% agarose gel electrophoresis, followed by visualization with ethidium bromide staining and UV fluorescence. The concentration of amplified products was estimated by comparison with a 100-bp quantitative ladder.

![Diagram](image-url)
(Panvera). The amplified cpDNA products were purified with Qiaquick spin columns (Qiagen, Valencia, California), then sequenced by the Sanger method with a BigDye Terminator Cycle version 1.0 or 3.1 sequencing kit (Applied Biosystems) or CEQ 2000 Dye Terminator Cycle Sequencing kit (Beckman-Coulter, Fullerton, California). The primers used in PCR were also used in sequencing, along with several primers designed for internal sequencing of amplified cpDNA fragments (Table 2). Both DNA strands were sequenced. Products were separated by capillary electrophoresis on an Applied Biosystems model 310 genetic analyzer or on a CEQ 8000

**Table 1.** Hosts and symbionts. All listed taxa, as well as trimmed taxon sets $T_1$ to $T_4$, were assessed for probability of codivergence.

<table>
<thead>
<tr>
<th>Grasses</th>
<th>Endophytes</th>
<th>Included in</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$T_1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$T_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$T_3$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$T_4$</td>
</tr>
<tr>
<td><em>Brachyelytrum erectum</em></td>
<td><em>Epichloë brachyelytri</em> (root) 201561</td>
<td>+</td>
</tr>
<tr>
<td><em>Brachypodium sylvaticum</em></td>
<td><em>Epichloë sylvatica</em></td>
<td>-</td>
</tr>
<tr>
<td><em>Echinopogon ovatus</em></td>
<td><em>Epichloë baconii</em></td>
<td>-</td>
</tr>
<tr>
<td><em>Calamagrostis villosa</em></td>
<td><em>Epichloë baconii</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Agrostis tenella</em></td>
<td><em>Epichloë amarillans</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Avena sativa</em></td>
<td><em>Epichloë festucae</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Holcus mollis</em></td>
<td><em>Epichloë sp. 9924</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Hordeum vulgare</em></td>
<td><em>Neotyphodium sp. FaTG2 4074</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Lolium sp. P4078</em></td>
<td><em>Neotyphodium sp. FaTG3 4078</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Lolium multiflorum</em></td>
<td><em>Neotyphodium occultans</em> 999</td>
<td>-</td>
</tr>
<tr>
<td><em>Lolium multiflorum</em></td>
<td><em>Neotyphodium occultans</em> 999</td>
<td>-</td>
</tr>
<tr>
<td><em>Lolium edwardii</em></td>
<td><em>Neotyphodium typhinum</em> 19</td>
<td>-</td>
</tr>
<tr>
<td><em>Lolium perenne</em></td>
<td><em>Epichloë festucae</em></td>
<td>-</td>
</tr>
<tr>
<td><em>Lolium perenne</em></td>
<td><em>Epichloë typhina</em></td>
<td>-</td>
</tr>
<tr>
<td><em>Lolium perenne</em></td>
<td><em>Epichloë lolii</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Festuca rubra</em></td>
<td><em>Epichloë festucae</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Festuca longifolia</em></td>
<td><em>Epichloë festucae</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Holcus mollis</em></td>
<td><em>Epichloë sp. 9924</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Hordeum vulgare</em></td>
<td><em>Neotyphodium sp. 362</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Hordeum brevisubulatum</em></td>
<td><em>Neotyphodium sp. 362</em></td>
<td>+</td>
</tr>
<tr>
<td><em>Elymus canadensis</em></td>
<td><em>Epichloë typhina</em></td>
<td>-</td>
</tr>
<tr>
<td><em>Glyceria striata</em></td>
<td><em>Epichloë glyceriae</em></td>
<td>-</td>
</tr>
<tr>
<td><em>Achnatherum inebrians</em></td>
<td><em>Neotyphodium gansuense</em> 818</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 2.** Primer list. Oligonucleotide primers for amplification and sequencing of plant cpDNA intron and intergenic regions.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Region</th>
<th>Sequence (5′ → 3′)</th>
<th>Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>B48557&lt;sup&gt;a&lt;/sup&gt;</td>
<td>trnT-trnL spacer</td>
<td>CATACAAATGGCATGCTCT</td>
<td>Downstream</td>
</tr>
<tr>
<td>A49291&lt;sup&gt;a&lt;/sup&gt;</td>
<td>trnT-trnL spacer</td>
<td>TCTACCGATTTCCGCTATC</td>
<td>Upstream</td>
</tr>
<tr>
<td>B49317&lt;sup&gt;a&lt;/sup&gt;</td>
<td>trnL intron</td>
<td>CGAATCTGGTAGACGTACG</td>
<td>Downstream</td>
</tr>
<tr>
<td>A49855&lt;sup&gt;a&lt;/sup&gt;</td>
<td>trnL intron</td>
<td>GGGGATAGAGGGACTTGAAC</td>
<td>Upstream</td>
</tr>
<tr>
<td>B49873&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnL-trnF spacer</td>
<td>GGGGATAGAGGGACTTGAAC</td>
<td>Downstream</td>
</tr>
<tr>
<td>A50272&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnL-trnF spacer</td>
<td>GTTCAAGTCCCTGCTATCC</td>
<td>Downstream</td>
</tr>
<tr>
<td>P37.trnT-trnL 359-340u&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnL-trnF spacer</td>
<td>TTTAAGATTCATTAGCTTTC</td>
<td>Downstream</td>
</tr>
<tr>
<td>P56.trnT-trnL 378-398d&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnT-trnL spacer</td>
<td>CTCXXXXXXCTCTGGCTCTA</td>
<td>Upstream</td>
</tr>
<tr>
<td>P2772.trnT-trnL 306-325d&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnT-trnL spacer</td>
<td>GTAATCGTAAATCTGCGGATA</td>
<td>Downstream</td>
</tr>
<tr>
<td>P57.trnT-trnL 359-340u&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnT-trnL spacer</td>
<td>CCAATCGTAAATCTGCGGATA</td>
<td>Downstream</td>
</tr>
<tr>
<td>P57.trnT-trnL 359-340u&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnT-trnL spacer</td>
<td>CCAATCGTAAATCTGCGGATA</td>
<td>Downstream</td>
</tr>
<tr>
<td>P2772.trnT-trnL 306-325d&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnT-trnL spacer</td>
<td>CCAATCGTAAATCTGCGGATA</td>
<td>Downstream</td>
</tr>
<tr>
<td>P57.trnT-trnL 359-340u&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnT-trnL spacer</td>
<td>CCAATCGTAAATCTGCGGATA</td>
<td>Downstream</td>
</tr>
</tbody>
</table>

<sup>a</sup> Primer used in PCR and sequencing.
<sup>b</sup> Internal primers used in sequencing only.
genetic analyzer (Beckman-Coulter) at the University of Kentucky Advanced Genetic Technologies Center. Sequences were assembled with either Sequence Navigator software (Applied Biosystems) or Phrap (CodonCode, Dedham, Massachusetts). Sequences were entered into GenBank as accession numbers AY450932 to AY450949 and EU119353 to EU119377.

Endophyte DNA Sequences

β-Tubulin and translation-elongation factor 1α gene sequences for the endophytes included in this study were obtained previously (Craven et al., 2001; Moon et al., 2004). Employing a standardized gene nomenclature for Epichloë and Neotyphodium species, these genes are designated tubB (formerly tub2) and tefA (formerly tef1), respectively.

Phylogenetic Tree Reconstruction and Analysis

Sequences were aligned with the aid of PILEUP implemented in SEQWeb Version 1.1 with Wisconsin Package Version 10 (Genetics Computer Group, Madison, Wisconsin). PILEUP parameters were adjusted empirically; a gap penalty of two and a gap extension penalty of zero resulted in reasonable alignment of intron-exon junctions and intron regions of endophyte sequences and of intergenic spacer and intron regions of cpDNA sequences. Alignments were scrutinized and adjusted by eye, using tRNA or protein coding regions as anchor points. For phylogenetic analysis of the symbionts, sequences from tubB and tefA were concatenated to create a single, contiguous sequence of approximately 1400 bp for each endophyte, of which 357 bp was exon sequence and the remainder was intron sequence. For phylogenetic analysis of the hosts, sequences for both cpDNA intergenic regions (trnT-trnL and trnL-trnF) and the trnl intron were aligned individually then concatenated to give a combined alignment of approximately 2200 bp.

Ultrafast trees were inferred with BEAST v1.4.1 (Drummond and Rambaut, 2007) with the general time reversible model with a proportion of invariant sites and a gamma distributed rate variation among sites (GTR+I+G). This model was selected by the software MrModelTest vers. 2.0 (Johan A. A. Nylander, Uppsala University, http://www.csit.fsu.edu/~nylander/mrmodeltest2readme.html; Posada and Crandall, 1998) as the model of nucleotide substitution that best fits the data. Based on published phylogenetic inference for the grass subfamily Pooidaeae (Soreng and Davis, 1998), Brachyelytrum erectum was chosen as the outgroup for the grass phylogenies. The corresponding endophyte, Epichloë brachyelytri, was the outgroup chosen for endophyte phylogenies. Due to the lack of historical dates for interior nodes, BEAST used a fixed substitution rate to reconstruct the trees. This results in branch lengths (and tree height) being measured in substitutions per site. The Markov chain Monte Carlo method used by BEAST was allowed to run for 5,000,000 steps. Every 100th step was recorded and analyzed for height, tree likelihood, and many other components. Preceding these recordings is a burn-in period equal to 10% of the MCMC chain. All data from the burn-in period are discarded and the operators are not optimized during this time, thus preventing operators from optimizing incorrectly on trees that are still considered random at the beginning of each run. This process was done in two independent runs from different tree topologies in order to avoid stacking at a local optimum, and resulted in a sample of 10,000 trees.

The MRCALink algorithm reported herein requires ultrametric trees. However, for illustrative purposes only, phylograms were also inferred and posterior probabilities estimated with MrBayes version 3 (Ronquist and Huelsenbeck, 2003), using a GTR+G model (lset nst = 6, rates = gamma). Four chains (three heated at temp = 0.2) were run for 450,000 generations, saving one out of every 100 trees (mcmc ngen = 450,000, printfreq = 10,000, samplefreq = 100, nchains = 4). The first 2000 trees (200,000 generations) were discarded as burn-in. This was an extremely conservative choice because likelihood values stabilized within 10,000 generations for both data sets. The 50% majority rule consensus trees and posterior support values were determined from the 2500 trees sampled from the remaining 250,000 generations. We ran three independent runs with 200,000 iterations and always got the same consensus tree, as shown in Figures 2 and 3.

Sequence alignments and trees reconstructed via MrBayes and BEAST of plants and endophytes have been submitted to TreeBASE.

The MRCALink Algorithm

The MRCALink algorithm introduced here identifies and stores each corresponding H and P MRCA pair. Crucially, the data for each corresponding MRCA pair are selected only once for subsequent statistical analysis. For example, if we have the pairs of trees in Figure 1, then we pick MRCA pair (7, 7') in the congruent tree four times from the set of all pairs of taxa in H and P. The MRCALink algorithm picks (7, 7') only once instead of four times. Ultrametric H and P trees must be used so that a unique age is estimated for each MRCA as half the patristic distance between any two of its descendant leaves. Trees must be strictly bifurcating for unique identification of valid H and P MRCA pairs. The BEAST program outputs ultrametric and strictly bifurcating trees. Note that the method does not assume an equal number of taxa in H and taxa in P and also does not assume similar substitution rates in H and P. Please see Appendix for the pseudo-code of the algorithm.

Source codes for The MRCALink algorithm and the dissimilarity methods as well as data files are available at http://polytopes.net.MRCALink.

Significance of Codivergence

In this section we will discuss the two statistical methods used to test significance of codivergence between the host and the parasite sets. The first is a dissimilarity estimate between trees in the same tree space. The second test uses ParaFit (Legendre et al., 2002) to analyze the
MRCA pairs sampled by the MRCALink algorithm. The hypotheses are:

Null hypothesis: Trees $T_H$ and $T_P$ are independent.  
Alternative hypothesis: Trees $T_H$ and $T_P$ are not independent.

Test via the dissimilarity method.—We are interested in estimating the probability that the host and symbiont tree have some degree of dependence that may be due to a history of codivergence. To this end, we use the method of total evidence. Suppose we have the set of pairwise distances in $H$ and $P$ or the set of pairwise distances in $H$ and $P$ from the MRCA pairs sampled by the MRCALink algorithm. Let the sum of differences in unique estimated MRCA ages for trees $A$ and $B$ be $S(A, B)$. The null hypothesis is that our $T_H$ and $T_P$ are independent, so we can generate a distribution of $S$ for pairs of unrelated random trees with the same number of leaves and root-to-tip normalized distances (i.e., we normalize the heights of $T_H$ and $T_P$ to 1) as $T_H$ and $T_P$. Then we compare our $S(T_H, T_P)$ with this distribution. If the $P$-value is significantly low ($<0.05$), we reject the null hypothesis and conclude that there is evidence of codivergence between $T_H$ and $T_P$. To calculate $S(A, B)$ with all pairwise distances, we take the sum of differences between pairwise distances for $A$ and $B$ over all pairwise distances. To calculate $S(A, B)$ with the set of the MRCA pairs sampled by the MRCALink algorithm we take the sum of differences between pairwise distances for $A$ and $B$ over the set of the MRCA pairs sampled by the MRCALink algorithm.

We generate 10,000 random trees with the given branch lengths from the birth and death process (BDP) via evolver from the PAML package (Yang, 1997) for each $T_H$ and $T_P$. For each tree, we used birth rate 0.5, death rate 0.5, and sampling fraction (SF = ratio of sample size to population size) 1, 0.5, 0.001, or 0.0005. We use the BDP model because it is biologically plausible (Aris-Brosou and Yang, 2003; Yang and Rannala, 1997; also see more details in Discussion).

Note the CPU time of this estimation with all pairwise distances is $O(Rn^2)$ and with the MRCALink algorithm it is $O(Rn^4)$, where $R$ is the number of random trees and $n$ is the largest number of taxa in $H$ or $P$ (i.e., $n = \max\{n_1, n_2\}$ where $n_1$ is the number of taxa in $H$ and $n_2$ is the number of taxa in $P$). The CPU time of the dissimilarity method is a polynomial time algorithm in terms of the number of taxa. Also, note that this process is easily distributed. Thus, if we fix $R$, the computation time of this method with all pairwise distances is $O(n^2)$ and the method with the MRCALink algorithm is $O(Rn^4)$.

Testing via ParaFit.—The ParaFit (Legrande et al., 2002) method requires three input files: a relation table designating the hosts and their parasites and the coordinate representation of phylogenetic distances for $H$ and $P$. These are the results of PCA (principle component analysis) on the respective distance matrices. This method has been applied with all $\binom{n}{2}$ pairwise distances between two taxa in $H$ and all $\binom{n}{2}$ pairwise distances between two taxa in $P$. Instead, we use the distance matrices for $H$ and $P$ computed from the set of MRCA pairs via the MRCALink algorithm.

Below is the same distance matrix using only the set of MRCA pairs provided by the MRCALink algorithm as used on the congruent tree in Figure 1. Notice the removal of the distances between the sets of taxa (1,4), (2,3), and (2,4). These represent the redundant information removed by the MRCALink algorithm. The sets of taxa (1,3), (1,4), (2,3), and (2,4) all have the same MRCA. The distance between taxa 1 and 3 is sufficient and is the only one used in this reduced distance matrix.

$$
\begin{pmatrix}
0 & d_{12} & d_{13} & 0 \\
\end{pmatrix}
$$

Following is another distance matrix using only the set of MRCA pairs provided by the MRCALink algorithm as used on the incongruent tree in Figure 1. Notice the removal of the distances between the sets of taxa (2,3) and (2,4). These represent the redundant information removed by the MRCALink algorithm. The MRCALink algorithm has a smaller effect on less congruent trees.

$$
\begin{pmatrix}
0 & d_{12} & 0 & d_{14} \\
\end{pmatrix}
$$

Random Tree Generators for the CPU Time Test

We generated random trees with 200 taxa by BDP with 0.5 birth rate, 0.5 death rate, and 0.0001 sampling fraction because Aris-Brosou and Yang (2003) and Yang and Rannala (1997) suggested that this is realistic model to generate random phylogenetic trees. The heights of the trees are the same as $T_H$ and $T_P$ in the $T_4$ data set. Because the computational time was 4 to 7 h, we only took two sets of random trees for $T_H$ and $T_P$ and then we recorded the CPU time.
RESULTS

Grass Phylogenies

PCR amplification of \textit{trnT-trnL} and \textit{trnL-trnF} intergenic spacers and the \textit{trnL} intron from endophyte-infected host plant genomic DNA yielded products of the sizes expected (approximately 850 to 950 bp, 400 to 450 bp for \textit{trnT-trnL} and \textit{trnL-trnF}, respectively; 350 to 600 bp for the \textit{trnL} intron). The majority-rule consensus tree with average branch lengths from MrBayes search on host cpDNA sequences is shown in Figure 2. In general, the inferred relationships among grass tribes were in good agreement with published grass phylogenies inferred by various genetic criteria (Soreng and Davis, 1998; Kellogg, 2001; Catalán et al., 2004). Host grasses in the tribes Poeae, Agrostideae (syn = Aveneae), and Bromeae all formed monophyletic clades. Two of the three grasses in tribe Hordeae (syn = Triticeae)—\textit{Hordeum brevisubulatum} and \textit{Elymus canadensis}—also grouped in a well-supported clade. However, \textit{ Hordeum europaeus}, currently classified in the Hordeaeae, grouped in the Bromeae clade basal to the \textit{Bromus} species.

Among the grasses in the tribe Poeae, a clear phylogenetic separation between the fine-leaf fescues (\textit{Festuca} subg. \textit{Festuca}) and the broad-leaf fescues (\textit{Lolium} subg. \textit{Schedonorus} = genus \textit{Schedonorus}) was evident (Fig. 2). Among the broad-leaf fescues in this analysis were three hexaploids previously classified as \textit{Festuca arundinacea}, and representing plants from Europe (shown as \textit{L. arundinaceum}), southern Spain (\textit{Lolium} sp. P4074), and Algeria (\textit{Lolium} sp. P4078; Tsai et al., 1994). The latter two plants had closely related cpDNA sequences in a subclade basal to the European plant and species of \textit{Lolium subg. Lolium}. Given this relationship, plants P4074 and P4078 are listed here as an undescribed \textit{Lolium} species. The Poeae clade also had \textit{Holcus mollis} in a basal position, which was the sister to the Agrostideae clade.
The precise branching order of the most deeply rooted grasses was poorly resolved (Fig. 2). The exception was *Brachypodium sylvaticum* (tribe Brachypodieae), which was placed nearest the clade comprising tribes Agrostideae, Poeae, Hordeae, and Bromeae. This result agreed with published studies (Soreng and Davis, 1998; Kellogg, 2001), which also indicate that tribe Brachyelytreae (represented by *Be. erectum*) diverged very early in the evolution of the cool-season grasses. Therefore, we chose *Be. erectum* to outgroup root the grass phylogeny.

**Endophyte Phylogenies**

The combined sequence data set (tubB + tefA) for the epichloae was approximately 1400 bp in length. The majority rule consensus tree from MrBayes search on the combined sequences (Fig. 3) was in accord with the individual gene trees published earlier (Moon et al., 2004). Two large clades of the epichloae were well supported, and included only isolates from corresponding clades of the grasses. One of these included several species associated with either Agrostideae or its sister tribe Poeae or both. This clade included *E. baconii*, *E. amarillans*, *E. festucae*, and its closely related asexual species *N. lolii*, as well as an *Epichloë* sp. isolate from *Holcus mollis*. Also included in this clade was the LAE (*Lolium*-associated endophyte) subclade of sequences from asexual symbionts of *Lolium* species. The endophytes with LAE genomes were all interspecific hybrids with additional genomes from *E. festucae*, *E. typhina*, or *E. bromicola*, species that
have contributed genomes to a large number of hybrid endophytes (Moon et al., 2004). Because the LAE genomes have not been identified in any sexual (Epichloë) species, it is possible that the clade represents an old asexual lineage. If so, the LAE genome is very likely to have been transmitted vertically, because asexual epichloae are only known to transmit vertically in host maternal lineages (Chung and Schardl, 1997; Brem et al., 1999).

Another clade grouped endophytes from sister grass tribes Bromeae and Hordeaeae. This clade included E. elymi, E. bromicola, and asexual endophytes from Bromus pungans, He. europaeus, and H. brevisubulatum.

Epichloë glyceriae and Neotyphodium gansuense, infecting grasses in tribes Meliceae and Stipeae, respectively, were placed at basal positions relative to the other endophyte species. Another basal clade grouped E. typhina from L. perenne with E. sylvatica and two asexual endophytes, Neotyphodium typhinum and Neotyphodium aotearoae.

Comparison of Endophyte and Host Tree Topologies

Several major groups or clades in the endophyte phylogeny corresponded to clades within the host phylogeny (Fig. 4). For example, sister host tribes Agrostideae and Poeae mostly coincided with a similar grouping of their endophytes. Within tribe Poeae, a group containing L. multiflorum, L. arundinaceum, and Lolium sp. plants P4074 and P4078 was mirrored by the branching orders of their respective LAE-clade endophytes. The sister clade relationship of Lolium and Festuca species was reflected by the LAE and E. festucae sister clades, and the basal position of Hol. mollis within tribe Poeae was nearly matched by that of its corresponding symbiont. Similarly, endophytes of the sister tribes Bromeae and Hordeaeae grouped in a clade. Grasses in basal host tribes Brachyelytrae, Stipeae, and Meliceae corresponded to basal endophyte clades E. brachyelytri, N. gansuense, and E. glyceriae, respectively.

Several instances of incongruence between host and endophyte phylogenies were also evident both within and across clades (Fig. 4). Notable cases involved E. typhina and two related asexual endophytes, N. typhinum and N. aotearoae. All three of these endophytes infect grasses in tribes Poeae (E. typhina and N. typhinum) or Agrostideae (N. aotearoae), yet they grouped in a clade that was maximally divergent from the larger clade of endophytes from these grass tribes. Other examples of incongruence involved an E. festucae isolate from Koeleria cristata (tribe Agrostideae), N. lolii (an asexual derivative of E. festuca) from L. perenne, and E. elymi from Bro. pungans.

Analyses of Codivergence

Computation Results.—For each data set, we generated 10,000 ultrametric trees through BEAST and chose the tree that had the maximum likelihood (Fig. 4). From this tree, we obtained corresponding pairs of H (grasses) and P (endophytes) MRCA ages (plotted in Fig. 5). The significance of codivergence was estimated from randomly generated tree pairs with several sampling fractions (Table 3). The lowest sampling fraction is probably the most biologically relevant (Aris-Brosou and Yang, 2003). We denote SF as a sampling fraction. We first estimated P-values via the dissimilarity methods. For the full grass and endophyte trees at SF = 0.0005, we estimated P = 0.123 with the MRCALink algorithm and P = 0.784 with all pairwise distances. Thus, analysis of this data set did not reject the null hypothesis that the host and the parasite trees are independent.

An obvious source of discordance between endophyte and host trees was E. typhina and related endophytes. Both N. lolii and the E. typhina isolate in this study were from L. perenne, but the endophytes were maximally divergent from one another. Previous surveys have indicated that E. typhina has an unusually broad host range and is ancestral to asexual endophytes (such as N. typhinum) in several other grasses (Leuchtmann and Schardl, 1998; Moon et al., 2004; Gentile et al., 2005). Therefore, the first trimmed tree set, T1, had these taxa removed, and the host and endophyte T1 trees were estimated. Analysis of the T1 set gave lower P-values (P < 0.001 with the MRCALink algorithm and P = 0.117 with all pairwise distances; Table 3). Thus, the MRCALink analysis of this dataset supported dependence of the trees, although analysis with all pairwise distances did not. The hypothesis that the entire clade including E. typhina, N. typhinum, E. sylvatica, and N. aotearoae contributed most of the discordance was tested by eliminating all four of these taxa and their hosts in tree set T2. The calculated P-values for T2 (P < 0.001 with MRCALink and P = 0.093 with all pairwise distances) were comparable to those of T1.

Trimmed set T3 had E. typhina and N. typhinum and their hosts removed, as well as other taxa that appeared likely to represent jumps of endophytes between divergent hosts; namely N. lolii and its host L. perenne, the E. festucae–K. cristata symbiont, and the E. elymi–Bro. pungans symbiont. The basis for considering these to be likely jumps was that the symbiont species were much more common on other grass genera: E. festucae on Festuca species, and E. elymi on Elymus species (Craven et al., 2001; Moon et al., 2004). The P-values for the taxa in the

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>SF = 0.0005</th>
<th>SF = 0.001</th>
<th>SF = 0.5</th>
<th>SF = 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>Full</td>
<td>0.784</td>
<td>0.783</td>
<td>0.677</td>
<td>0.374</td>
</tr>
<tr>
<td>MRCA</td>
<td>Full</td>
<td>0.123</td>
<td>0.123</td>
<td>0.081</td>
<td>0.039</td>
</tr>
<tr>
<td>ALL</td>
<td>T1</td>
<td>0.117</td>
<td>0.115</td>
<td>0.035</td>
<td>0.009</td>
</tr>
<tr>
<td>MRCA</td>
<td>T1</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>T1</td>
<td>0.093</td>
<td>0.085</td>
<td>0.027</td>
<td>0.012</td>
</tr>
<tr>
<td>MRCA</td>
<td>T1</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>T1</td>
<td>0.064</td>
<td>0.061</td>
<td>0.017</td>
<td>0.005</td>
</tr>
<tr>
<td>MRCA</td>
<td>T1</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>T1</td>
<td>0.018</td>
<td>0.020</td>
<td>0.005</td>
<td>0.002</td>
</tr>
<tr>
<td>MRCA</td>
<td>T1</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
If codivergence has been an important trend since the origin of the epichloës and the pooid grasses, and if the DNA regions analyzed for these fungi and their hosts have had comparable substitution rates in the regions analyzed, then the estimated heights ($h$) of their respective phylogenetic trees should be comparable. Very different tree heights could be due to a lack of long-term codivergence of $H$ and $P$ or to a large difference in substitution rates for the regions chosen for analysis. The tree heights estimated by BEAST in substitutions per site were as follows: for the full data set, $h(T_H) = 0.048623$ and $h(T_P) = 0.028991$; for $T_1$, $h(T_H) = 0.046534$ and $h(T_P) = 0.028633$; for $T_2$, $h(T_H) = 0.045813$ and $h(T_P) = 0.028924$; for $T_3$, $h(T_H) = 0.048321$ and $h(T_P) = 0.028367$; for $T_4$, $h(T_H) = 0.045188$ and $h(T_P) = 0.027939$. Thus, for all of these data sets the inferred host tree heights were similar, and the inferred endophyte tree heights were similar. For both
hosts and endophytes, almost all of the sequence analyzed was noncoding. The host data set was mainly intergenic sequence from chloroplast DNA, and most of the endophyte data comprised nuclear intronic sequence. The tree height estimates suggested that the substitution rate of the host sequences has been 1.58 to 1.70 times the substitution rate of the endophyte sequences. Thus, the estimated substitution rates were comparable, lending additional support to the hypothesis that the Pooideae and the epichloae originated at approximately the same time.

We tested if the analyses reject suboptimal parasite and host trees with no detectable host jumps. Instead of choosing the tree with maximum likelihood, we chose the three samples with likelihood values closest to the mean of likelihood values from the set of all trees sampled by BEAST. For example, the maximum likelihood tree for the full plant data set that was used in our methods had a negative log likelihood of $-6771.2947$. The suboptimal samples had negative log likelihoods of $-6783.4135$, $-6783.4134$, and $-6783.4076$. We then calculated $P$-values by the dissimilarity method (Table 4) and ParaFit (Table 5).

Application of the dissimilarity methods to the suboptimal trees did not provide evidence to reject the null hypothesis (Table 4). The dissimilarity method with the MRCALink algorithm obtained significant $P$-values for some but not all samples. Computations were conducted on a Dual Core Pentium L2400 1.66 GHz PC machine in IBM Thinkpad laptop X60S with 2 GB RAM running Fedora Core 6 Linux. For the dissimilarity method with all pairwise distances, calculating the $P$-value has the time complexity $O(n^2)$ where $n$ is the largest number of taxa in $H$ or $P$ (i.e., $n=\max(|H|,|P|)$). Thus, this is a polynomial time algorithm in terms of the number of taxa. Analyses of grass-endophyte data sets with the dissimilarity method with all pairwise distances required CPU time of 39.5 s for the full 26 taxon pairs, 33.481 s for $T_1$, 27.2 s for $T_2$, 25.0 s for $T_3$, and 20.1 s for $T_4$. For computational time simulation analysis of our method with randomly generated 200-taxon trees with all pairwise distances, the dissimilarity method took 3.59 h of CPU time.

For the dissimilarity method with the MRCALink algorithm, calculating the $P$-value has the time complexity $O(n^4)$ where $n$ is the largest number of taxa in $H$ or $P$ (i.e., $n=\max(|H|,|P|)$). Thus, this is also a polynomial time algorithm in terms of the number of taxa. Analyses of grass-endophyte data sets with the MRCALink algorithm required CPU time of 2 min 49 s for the full 26 s taxon pairs, 2 min 16 s for $T_1$, 1 min 41 s for $T_2$, 1 min 33 s for $T_3$, and 1 min 9 s for $T_4$. The CPU time of our method with 200 taxa was 7 h.

Note that to estimate the $P$-value using random trees, we can easily distribute computations for both methods. Thus, this estimation method could be applied to host-parasite associations with several hundred taxa.

Results with ParaFit.—We also analyzed the full grass and endophyte data sets $T_1$ through $T_4$ with ParaFit,
Table 4. The $P$-values obtained using the dissimilarity method with suboptimal trees with 26 full and $T_i$ to $T_j$ plant and endophyte data sets (all taxa listed in Table 1) via the Bayesian MCMC method in BEAST. ALL means the dissimilarity method with all pairwise distances, and MRCA means the dissimilarity method with the MRCALink-derived matrix. SF means a sampling fraction. Each sampled tree is assigned a number from 1 to 3 to distinguish it from the others.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>Sample number</th>
<th>SF = 0.0005</th>
<th>SF = 0.001</th>
<th>SF = 0.5</th>
<th>SF = 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>Full</td>
<td>1</td>
<td>0.700</td>
<td>0.686</td>
<td>0.466</td>
<td>0.294</td>
</tr>
<tr>
<td>MRCA</td>
<td>Full</td>
<td>1</td>
<td>0.011</td>
<td>0.011</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>ALL</td>
<td>Full</td>
<td>2</td>
<td>0.474</td>
<td>0.483</td>
<td>0.245</td>
<td>0.119</td>
</tr>
<tr>
<td>MRCA</td>
<td>Full</td>
<td>2</td>
<td>0.064</td>
<td>0.064</td>
<td>0.025</td>
<td>0.014</td>
</tr>
<tr>
<td>ALL</td>
<td>Full</td>
<td>3</td>
<td>0.684</td>
<td>0.683</td>
<td>0.450</td>
<td>0.262</td>
</tr>
<tr>
<td>MRCA</td>
<td>Full</td>
<td>3</td>
<td>0.193</td>
<td>0.190</td>
<td>0.102</td>
<td>0.061</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>1</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>2</td>
<td>0.029</td>
<td>0.033</td>
<td>0.005</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>3</td>
<td>0.006</td>
<td>0.007</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>3</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>1</td>
<td>0.346</td>
<td>0.355</td>
<td>0.190</td>
<td>0.097</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>1</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>2</td>
<td>0.355</td>
<td>0.360</td>
<td>0.184</td>
<td>0.099</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>2</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>3</td>
<td>0.084</td>
<td>0.079</td>
<td>0.022</td>
<td>0.010</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>3</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>1</td>
<td>0.070</td>
<td>0.067</td>
<td>0.020</td>
<td>0.007</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>1</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>2</td>
<td>0.030</td>
<td>0.029</td>
<td>0.007</td>
<td>0.030</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>2</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>3</td>
<td>0.132</td>
<td>0.138</td>
<td>0.050</td>
<td>0.021</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>3</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>1</td>
<td>0.106</td>
<td>0.103</td>
<td>0.039</td>
<td>0.015</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>1</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>2</td>
<td>0.024</td>
<td>0.026</td>
<td>0.007</td>
<td>0.002</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>2</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>3</td>
<td>0.017</td>
<td>0.016</td>
<td>0.006</td>
<td>0.002</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>3</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Results with Gopher-Louse Data Sets

We also tested our methods with a well-known gopher-louse data set. The data set in Hafner et al. (1994) (full data set) contains 17 taxa of lice and 15 taxa of gophers, whereas Huelsenbeck et al. (2003) have trimmed host-parasite pairs representing apparent host jumps: louse species Geomydoecus thomomyns, Geomydoecus actuosi, Thomomydoecus barbarae, and Thomomydoecus minor, and gopher species Thomomys talpoides and Thomomys bottae.

To reconstruct trees, we used BEAST with the GTR+I+G model (Fig. 6). In this analysis, T. talpoides and T. bottae were outgroups in the gopher data set, and T. barbarae and T. minor were outgroups in the louse data set. With the full data sets, application of ParaFit to all pairwise distances gave $P = 0.001$. In contrast, the dissimilarity

Table 5. The $P$-values obtained through ParaFit to all pairwise distances (ALL) and to the MRCALink-derived matrix (MRCA) for the suboptimal trees of 26 full and $T_i$ to $T_j$ plant and endophyte data sets. The suboptimal trees were chosen as samples of the most common likelihood trees. This was done by choosing samples with tree likelihoods close to the mean of all 10,000 tree likelihoods. For each data set, the three sampled suboptimal trees were arbitrarily assigned a number from 1 to 3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>Full</td>
<td>0.146</td>
<td>0.007</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MRCA</td>
<td>Full</td>
<td>0.097</td>
<td>0.014</td>
<td>0.020</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>0.011</td>
<td>0.003</td>
<td>0.006</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>0.023</td>
<td>0.029</td>
<td>0.115</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>&lt;0.001</td>
<td>0.033</td>
<td>0.033</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>0.003</td>
<td>&lt;0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>0.015</td>
<td>&lt;0.001</td>
<td>0.017</td>
</tr>
<tr>
<td>ALL</td>
<td>$T_i$</td>
<td>0.046</td>
<td>0.046</td>
<td>0.042</td>
</tr>
<tr>
<td>MRCA</td>
<td>$T_i$</td>
<td>0.108</td>
<td>0.084</td>
<td>0.066</td>
</tr>
</tbody>
</table>
method with the sampling fraction 0.0005 gave $P = 0.589$, not rejecting the null hypothesis (Table 6). With the trimmed data sets, application of ParaFit to all pairwise distances gave $P = 0.001$ and also the dissimilarity method with the sampling fraction $= 0.0005$ gave $P = 0.012$. However, application of the dissimilarity method with MRCA pairs from MRCALink gave significant $P$-values, evidence to reject the null hypothesis for both the full data and trimmed gopher-louse data sets.

With the full data sets and the trimmed data sets, applying ParaFit on all pairwise distances gave a significance at $P < 0.01$ for some sample fractions (Table 6). Applying ParaFit to the MRCALink-derived matrix for these full and trimmed data sets also gave $P < 0.01$ for all sample fractions tried, indicating rejection of the null hypothesis.

**TABLE 6.** The $P$-values obtained using the dissimilarity method for the gopher and louse data. The full data set includes all hosts and parasites from Hafner et al. (1994), whereas for the trimmed data set data were removed for the gophers and lice from Huelsenbeck et al. (2003). ALL means the dissimilarity method with all pairwise distances and MRCA means the dissimilarity method with the MRCALink-derived matrix.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>SF = 0.0005</th>
<th>SF = 0.001</th>
<th>SF = 0.5</th>
<th>SF = 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>Full</td>
<td>0.589</td>
<td>0.577</td>
<td>0.394</td>
<td>0.248</td>
</tr>
<tr>
<td>MRCA</td>
<td>Full</td>
<td>0.002</td>
<td>0.003</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ALL</td>
<td>Trimmed</td>
<td>0.012</td>
<td>0.015</td>
<td>0.006</td>
<td>0.003</td>
</tr>
<tr>
<td>MRCA</td>
<td>Trimmed</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**DISCUSSION**

Endophyte Codivergences with Hosts

In this study we have taken a novel approach to investigate codivergences between hosts and symbionts (parasites), which differs from others (Jackson, 2004; Legendre et al., 2002; Page and Charleston, 1998) in that it is a more direct comparison of historical cladogenesis events represented by inferred MRCA ages in ultrametric time trees in a way that avoids excessive weighting of deeper nodes compared with shallower nodes. The results indicate that, with relatively few exceptions, evolution of symbiotic epichloë fungi largely tracked evolution of their grass hosts. These symbioses extend across the taxonomic range of the Pooidae, including the basal tribe Brachyelytreae, yet are restricted to this subfamily. Endophytes related to epichloë (such as *Balansia* species) are known from other hosts, but the combination of very benign, often mutualistic, interactions and extremely efficient vertical transmission is known only in the Pooidae-epichloae system (Clay and Schardl, 2002). The conclusion that the system is dominated by codivergence implies that this unusually intimate symbiotic system emerged coincidentally with the emergence of this important grass subfamily.

In an earlier study comparing grass and endophyte evolutionary histories, the topological relationships of host tribes matched those of *Epichloë* species with a mixed mode of transmission (Schardl et al., 1997). However, no asexual lineages were included, and the possibility that some of the strictly sexual, horizontally transmitted...
species might also have a history of codivergence was not assessed. More importantly, the inference of codivergence was based on branching order, not relative timing of cladogenesis events. Although mirror phylogenies are suggestive of codivergence, it is the concomitance of corresponding cladogenesis events that defines codivergence (Hafner et al., 1994). Conversely, unless the codivergences correspond to actual speciation events (that is, isolation of populations into distinct gene pools), lineage sorting effects, species duplications, and incomplete taxon sampling can prevent $H$ and $P$ phylogenies from mirroring each other (Page and Charleston, 1998). Therefore, we undertook the current study to assess codivergence by a more direct assessment of relative ages of corresponding cladogenesis events.

The null hypothesis was that relative ages of corresponding host and endophyte MRCAs were unrelated. If all sampled taxa were included, the null hypothesis was not rejected. This, however, was expected for the full data set for two reasons: (1) topology within the $E. tychina$ clade bears no resemblance to that of the hosts, in keeping with the fact that $E. tychina$ is a broad host-range species; and (2) some of the topological discordances in other clades strongly suggested occasional host jumps. Topological discordances tended to involve rarer associations. For example, $E. festucae$ has only been identified in $K. cristata$ once but is very common in $Festuca$ species. These features of the grass-endophyte system allowed for a rational basis for trimming trees of exceptions to assess support for codivergences in remaining taxa. When the $E. tychina$ clade was removed, the significance of codivergence increased dramatically. Using all pairwise distances for dissimilarity analysis, the null hypothesis was still not rejected, but when the method was applied to the MRCA-Link-derived data set, the null hypothesis was strongly rejected. Trimming other possible host jumps further decreased the $P$-values in all analyses, strongly supporting the conclusion that the relationships between host and endophyte trees had a degree of dependence.

Various factors may cause a tendency for codivergence in evolution of these symbioses. For example, it may be much more likely for an endophyte to colonize a new host species that is closely related to its host of origin than a host that is more distantly related. A less likely (though not mutually exclusive) scenario would be speciation of hosts driven in part by adaptation to different symbionts. For example, considering that benefits of these endophytes are likely to be highly dependent on environmental conditions, geographical separation of the combined host-endophyte systems followed by different selective forces in the separate populations might eventually lead to a circumstance (after the populations merge again) in which sharing of endophytes is less beneficial to the hosts. Analogously to the problem of hybrid disadvantage, a “hybrid symbiont disadvantage” may simultaneously promote evolution of genetic isolation for both the hosts and their endophytes (Thompson, 1987).

However, it is also possible that hosts will tend to benefit from endophytes that have adapted to related hosts but will benefit far less or actually suffer detriment from endophytes adapted to distantly related hosts. So, for example, most $E. tychina$, $E. baconii$, and $E. glyceriae$ strains infecting their hosts cause complete suppression of seed production, thus eliminating a means of dispersal (seeds) as well as a means of genetic diversification (meiotic recombination) that could enhance survivability of host progeny in the face of changing environmental factors. In contrast, endophytes such as $E. festucae$, $E. chlymi$, $E. brachyelytri$, and $E. amarillans$ allow substantial seed production by producing their fruiting structures (stromata) on only a portion of the tillers of the infected plant (Leuchtmann and Schardl, 1998; Schardl and Moon, 2003). The situation with $E. bromicola$ and $E. sylvatica$ is more complicated because expression of host seeds or endophyte stromata depends on host and endophyte genotypes, but these species also have the potential to be far less damaging to their hosts than are $E. tychina$, $E. baconii$, and $E. glyceriae$. It should be much more to the benefit of a host to maintain compatibility with the benign or mutualistic $Epichloë$ species than with the more antagonistic species. Indeed, $E. tychina$ has a broad host range and clearly has not codiverged with those hosts (Leuchtmann and Schardl, 1998), which is why this species and the asexual endophytes most closely related to it were removed in the trimmed data sets for analysis of codivergence. The question is whether the remaining narrower host-range endophytes have tended to codiverge with their hosts, and our results suggest that this is the case.

The possibility that these symbiotic systems emerged with the origin of the grass subfamily is intriguing. Kellogg (2001) postulated an early shift from shady to sunny habitats in subfamily Pooidae. According to this hypothesis, lineages derived following the split from the Brachyelytraeae lineage moved into open habitats where, presumably, competition was less intense but solar radiation and drought were more prevalent. Additionally, such early colonizers may have been more conspicuous to potential herbivores. Protection from herbivory and drought are among the better documented effects of the epichloae (Clay and Schardl, 2002). Kellogg’s hypothesis adds perspective to our results suggesting that codivergence between cool-season grasses and their endophytes originated with the Pooidae. It is reasonable to suggest that these symbioses may have played an important role in this habitat shift and that the mutualistic tendencies that these fungi commonly impart to their hosts (drought tolerance, herbivore resistance, etc.) are a direct reflection of these new selective pressures faced in open habitats. Even the more antagonistic endophytes that severely restrict seed production would probably have exerted a strong effect on structuring emerging grassland communities. Following this habitat transition, these symbionts may have significantly enhanced host fitness, aiding the radiation of a highly successful and speciose grass subfamily.

The MRCA-Link Method

Our method involves direct comparison of MRCA ages rather than analysis of host and symbiont pairwise
distance matrices, which is often used in studies of co- 
divergence (Legendre et al., 2002). The pairwise distance 
approach is attractive in that it seems not to require 
inference of phylogenies, for which lineage sorting of pre-
existing polymorphisms, and imperfect genetic barriers 
between species, may give phylogenies that inaccurately 
represent histories of speciation (Page and Charleston, 
1998). Nevertheless, there is a true phylogeny of species, 
and each pair of leaves (extant taxa) for which a pairwise 
distance is calculated represents the node in that phy-
logeny that is their MRCA. Codivergence implies that 
the MRCA of a host leaf pair occurred at the same time 
as the MRCA of a symbiont leaf pair. If we test all host 
leaf pairs and corresponding symbiont leaf pairs in order 
to derive a statistical test of the null hypothesis that there 
exists no significant relationship between host and symbiont 
tip clades are sampled only once. In this case, the number 
of times a MRCA pair is sampled is \( \binom{n}{2} \), where \( n \) is the number of taxa in 
each tree), whereas the MRCA pairs from corresponding 
tip clades are sampled only once. In this case, the number 
of times a MRCA pair is sampled is \( \binom{n_1}{2} + \binom{n_2}{2} \), where \( n_1 \) and \( n_2 \) are the numbers of descendants from 
the MRCA in \( H \) and \( P \) trees, respectively. To examine 
the more generalized case of \( H \) and \( P \) trees that may 
not be balanced or congruent, 10,000 random \( H \) and \( P \) trees 
were generated with 25 taxa in PAML by the BDP with 
birth rates = 0.5, death rates = 0.5, and mutation rates = 
100. Then, for each pair of \( H \) and \( P \) trees, we identified all 
corresponding leaf pairs and counted how many times 
each MRCA pair was identified. The maximum number 
of times any MRCA pair was sampled averaged 111 in 
the 10,000 simulations.

In this study we introduce the MRCALink algorithm to 
specifically identify valid \( H \) and \( P \) MRCA pairs to com-
pare divergence times and to avoid repeated sampling of 
any MRCA pairs. The aforementioned analysis sug-
gests that this method of sampling nodes is a substantial 
 improvement over the use of complete pairwise distance 
matrices. However, the MRCA method is affected by in-
congruence of \( H \) and \( P \) trees. As shown in Figure 1, the 
nodes (MRCA) of subtrees where such incongruences 
exist will tend to be sampled more than the nodes in con-
gruent subtrees, even though each node pair is sampled 
no more than once. This may be why, compared to the 
use of all pairwise distances, application of MRCALink 
is more sensitive to dependence between the trees being 
compared by the dissimilarity method in which the ML 
tree pair is compared to pairs of randomly generated 
trees of equal length. This may be considered a ben-
fit of the MRCALink method, but further research into 
this behavior and possible modifications of the method 
are required to fully assess how this characteristic of the 
method affects inference about tree dependence.

Our method described in this paper does not find the 
minimum set of non-codiverging taxon pairs. It would 
be interesting to find an efficient algorithm to find the 
minimum set of non-codiverging taxon pairs from \( H \) 
and \( P \).

Dissimilarity Method

The \( P \)-values obtained via the dissimilarity method 
for all pairwise distances tend to be larger than the \( P \)-values 
for MRCA pairs. This is because (1) analysis of all pair-
wise distances will generate a bias in favor of nonco-
vergence; (2) the \( S \) distribution is the sum of absolute 
values of differences between distances for each pair of 
taxa; and (3) from the computation results in Analyses of 
Codivergence, if we take all pairwise distances, then we 
observed that MRCA pairs tend to be more frequently 
sampled in highly correlated trees than in poorly corre-
lated trees. Because MRCA pairs for highly correlated \( T_H \) 
and \( T_P \) seem to be more frequently sampled among all 
pairwise distances than MRCA pairs for less correlated 
trees, the \( S \) value for \( T_H \) and \( T_P \) for all pairwise distances 
is overestimated and thus an estimated \( P \)-value obtained 
via the dissimilarity method with all pairwise distances 
is likely to be higher than the actual \( P \)-value (see Tables 3 
to 6).

Generating random trees by the BDP tends to pro-
duce trees with long interior branches. Yang and Rannala 
(1997) suggest that taking incomplete species sampling 
into account generates more realistic trees. Thus, we used 
\texttt{evolver} to generate random ultrametric trees with spec-
suggest using a sampling fractions in \((0, 0.001)\). How-
ever, Aris-Brosou and Yang (2003) also note that the sam-
ples in the space of trees is known to affect the topology of random 
trees, thus may affect divergence time of nodes. There-
fore, in this study we took four different sampling fractions 
0.0005, 0.001, 0.5, and 1, though we consider the 
smallest sampling fraction to be the most biologically 
relevant (Aris-Brosou and Yang, 2003).

We have used random trees to estimate the measure 
between two trees. Currently we use the BDP with a spec-
ified sampling fraction. When we measure a dissimilarity 
between the host and parasite trees in the space of trees, 
we used a heuristic method because measuring the exact 
distance between two trees in the space of trees requires 
exponential time in terms of the number of taxa (Billera 
et al., 2001). However, Amenta et al. (2007) recently 
developed a method to approximate a distance between 
trees in the space of trees efficiently, which could be a 
reasonable alternative to our method.

ParaFit Method

From results in Table 5 it seems that some of the \( P \)- 
values obtained via \texttt{ParaFit} with MRCA results are 
higher than the \( P \)-values via \texttt{ParaFit} with all pairwise 
distances. This may be due to the use of PCA prior to 
\texttt{ParaFit}. In the process of removing some of the entries 
in each distance matrix, we cut off more small signals in 
the data.
However, we note that $P$-values for suboptimal trees via ParaFit differ from $P$-values for the ML tree and in some instances have much larger $P$-value than $P$-values from the ML trees (e.g., Table 5). Also, there are large differences between $P$-values from the ML trees and suboptimal trees for the plant-endophyte data sets. For the ML trees, ParaFit returns $P$-values of 0.001 for all data sets. However, with some of the suboptimal trees, ParaFit returns higher $P$-values, not rejecting the null hypotheses. These suboptimal trees are sampled from the distribution with the given data and model via MCMC.

Comparing the dissimilarity method and ParaFit method, both estimate the $P$-values by sampling random trees or random matrices, respectively. ParaFit takes two distance matrices, then estimates the $P$-value by the permutation test. The dissimilarity method estimates the $P$-value by sampling random ultrametric trees from the BDP with a given sampling fraction known to be biologically meaningful. Also, ParaFit does not include constraints that these random matrices are distance matrices and coming from trees. On the other hand, the dissimilarity method includes constraints that these random samples are ultrametric trees. Therefore, these biologically constraints added in the dissimilarity method should result in more biologically meaningful $P$-values. Also, it is interesting that the $P$-values computed from suboptimal trees with $T_d$ data set did not give strong evidence for rejecting the hypothesis, but the ML method did. This suggests that the ParaFit method may be very sensitive to parametric assumptions and/or be unstable.

**ACKNOWLEDGMENTS**

We thank Adrian Leuchtmann (ETH Zürich, Switzerland) for providing biological materials, Walter Hollin (University of Kentucky) for technical support, Stephane Aris-Brosou for very useful conversation, and Mark Hafner for providing the gopher-louse data sets. Also, we would like to thank Roderic Page, Francois-Joseph Lapointe, Michael Charleston, Jack Sullivan, and an anonymous referee for very useful comments to improve this paper. This work was supported by National Science Foundation grant DEB-9707427, National Institutes of Health grant KY-INBRE F20 RR16481, U.S. Department of Agriculture CSREES grant 2005-34457-15712, and the Harry E. Wheeler Endowment of the University of Kentucky Department of Plant Pathology. This is publication 08–12–034 of the Kentucky Agricultural Experiment Station, published with approval of the director.

**REFERENCES**


APPENDIX—THE MRCALINK ALGORITHM

Given a set of host taxa \( H \) and a set of symbiont taxa \( P \) ("parasites," in keeping with other literature in the field), there is a map called \( L : H \rightarrow P \) such that a host \( A \in H \) has a parasite or symbiont \( L(A) \in P \). Define MRCA\((A, B)\) to be the node representing the most recent common ancestor (MRCA) of leaves \( A \) and \( B \).

Algorithm (The MRCALink Algorithm)

- **Input** a set of host taxa \( H \), a set of parasite taxa \( P \), a host tree \( T_H \), and a parasite tree \( T_P \).
- **Output** a set of MRCA pairs of host taxa and parasite taxa.

**Algorithm**

1. Assign each node a unique number from 1 to \( n_1 - 1 \) in the host tree and a unique number from 1 to \( n_2 - 1 \) in the parasite tree such that a node \( i \) is ancestral to a node \( k \) with \( i < k \).
2. Let \( U \) be a set of pairs of one leaf from the left and one from the right
   \[ \left\{ (x, y) \mid x \in H, y \in P \right\} \]
   such that a node \( X_i \) is all such pairs of one leaf from the left and one from the right.

\[
\text{for} \ (i \text{ from } n_1 + 1 \text{ to } 2n_1 - 1) \text{ do} \\
\text{Set } X_i = I_i \times R_i, \text{ where } I_i \text{ is the set of all left-descendants of } i, \text{ and } R_i \text{ is the set of all right-descendants of } i. \\
\text{/* This is just another way of saying } X_i \text{ is all such pairs of one leaf from the left and one from the right. */}
\]

\[
\text{while } (X_i \neq \emptyset) \text{ do} \\
\text{Choose } x = \text{MRCA}(a, b) \in X_i \text{ and identify } y_j = \text{MRCA}(L(a), L(b)) \text{ for each distinct } L(a) \text{ and } L(b). \\
\text{Remove } x \text{ from } X_i. \\
\text{for each distinct } y_j \text{ do} \\
\text{if } (\text{MRCA}(x, y_j) \neq U) \text{ do} \\
\text{ } U \leftarrow U \cup \text{MRCA}(x, y_j). \\
\text{ } \} \\
\text{ } } \\
\text{Output } U. \]