A combinatorial test for significant codivergence between cool-season grasses and their symbiotic fungal endophytes

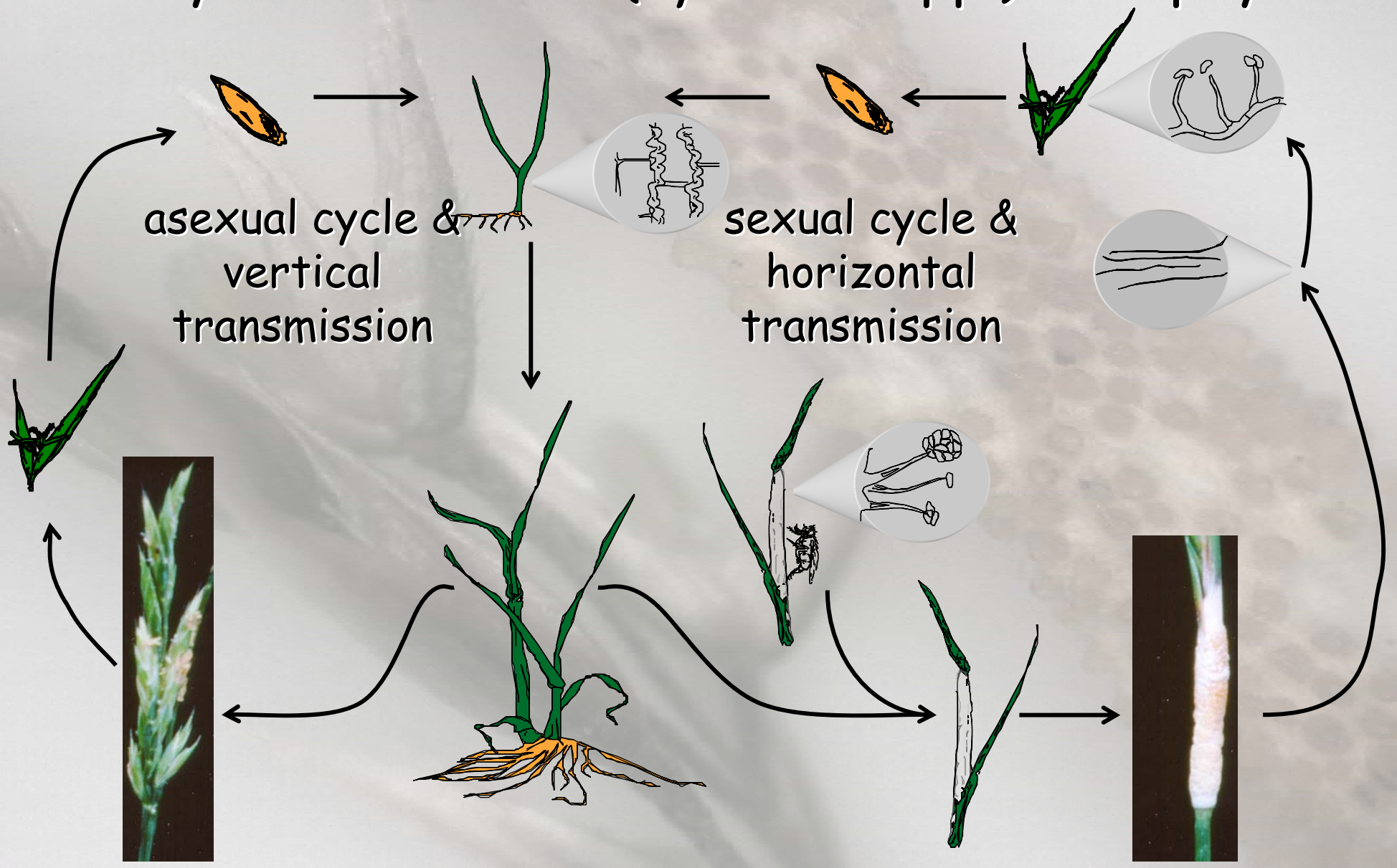
Ruriko Yoshida
Dept. of Statistics University of Kentucky

Joint work with C.L. Schardl, K.D. Craven, A. Lindstrom, A. Stromberg

www.ms.uky.edu/~ruriko

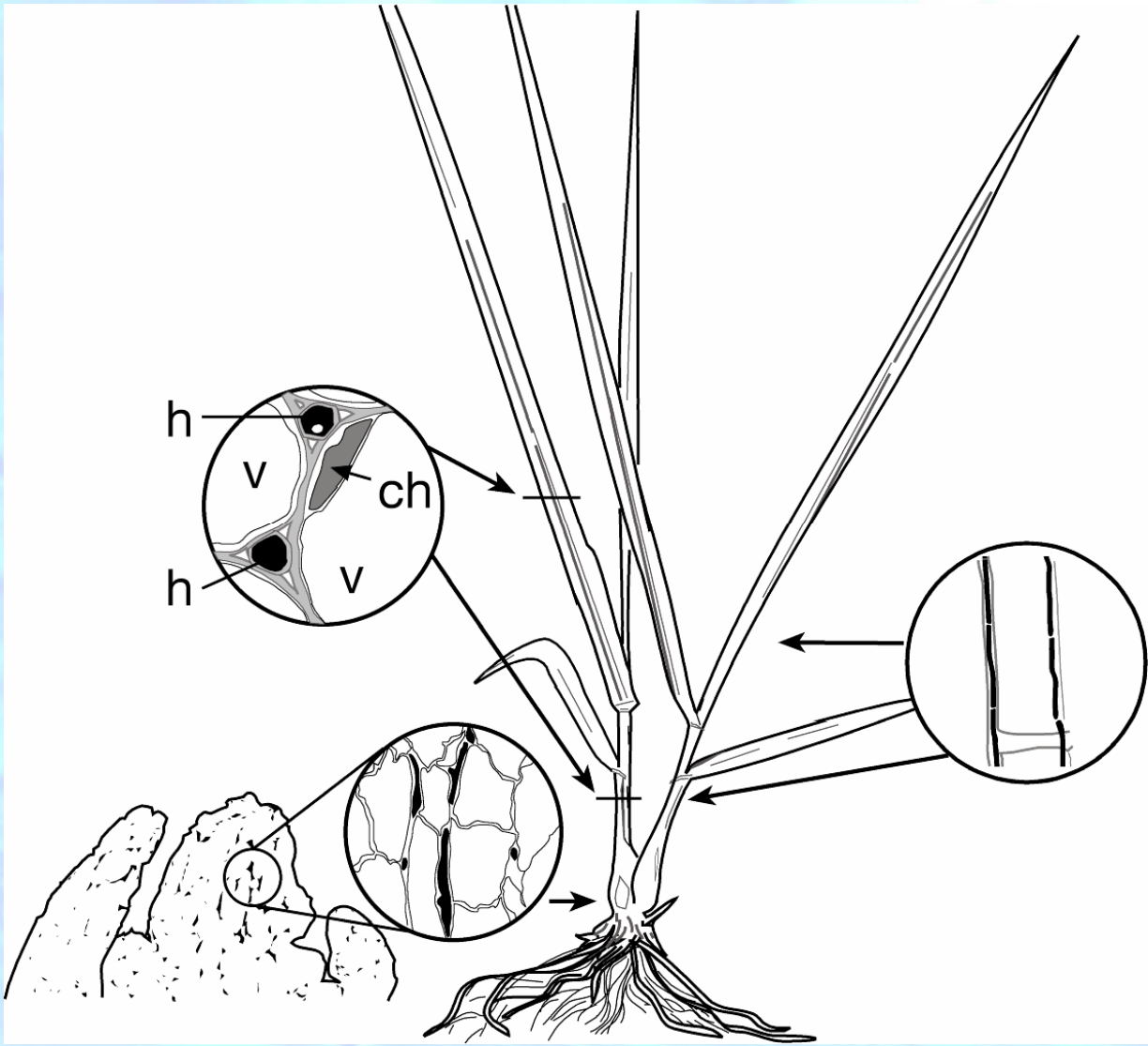
Endophytes species

Life cycles of sexual (*Epichloë* spp.) endophytes

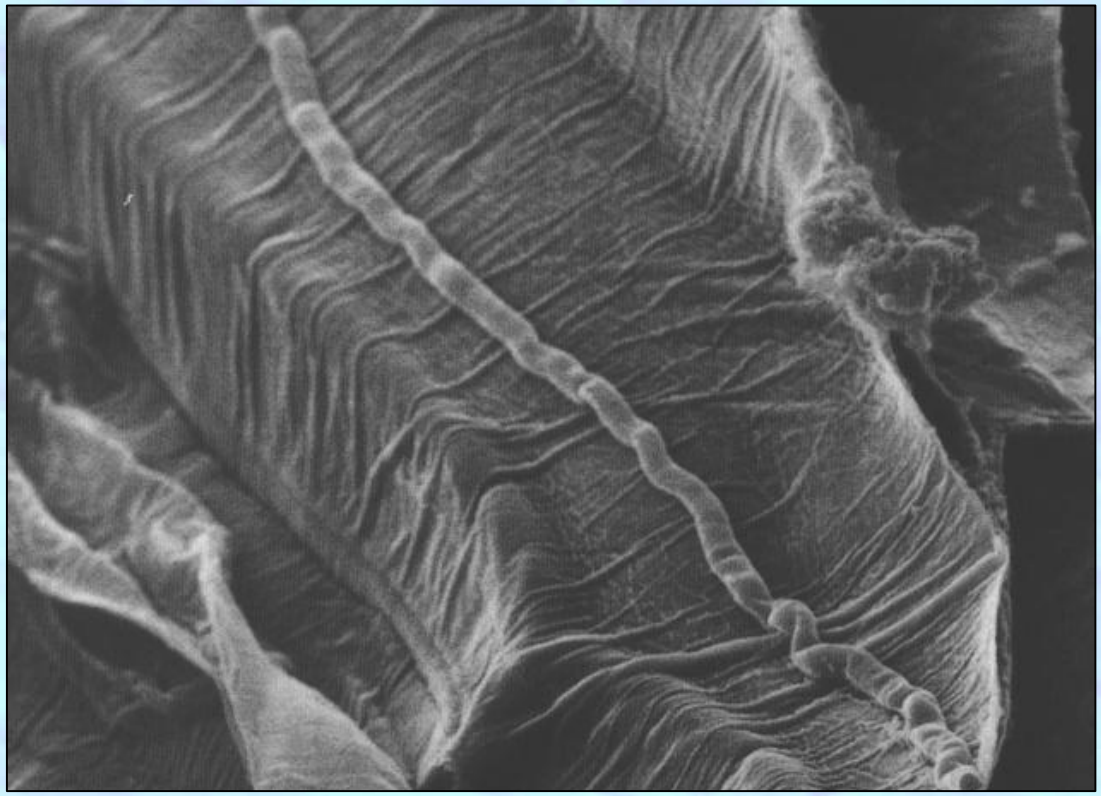


Epichloë/Neotyphodium in a grass plant

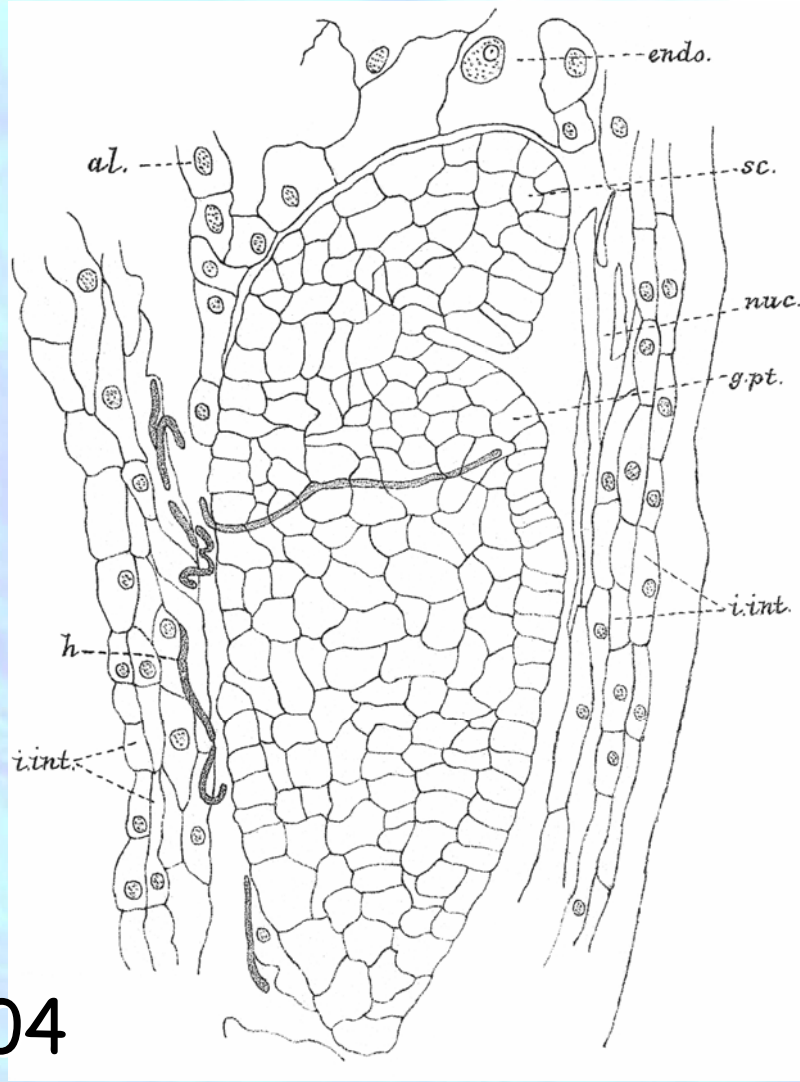
- Symbioses are:
  - Systemic
  - Constitutive
  - Often heritable
  - Often mutualistic



Endophytes:
Constitutive, heritable symbionts

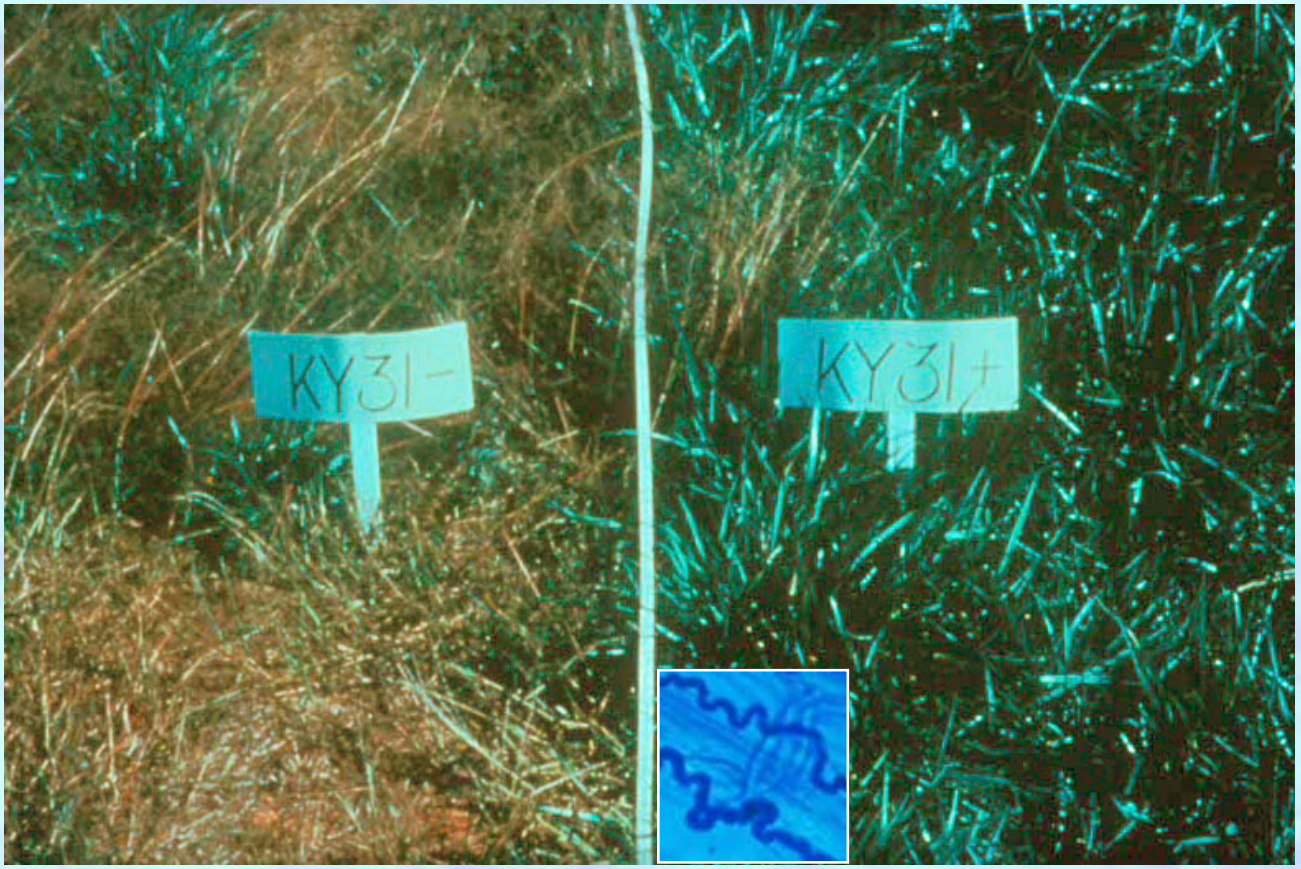


H. Koya, pers. comm.



Freeman 1904

Endophyte effects on host fitness

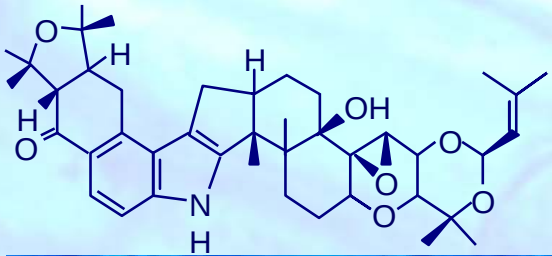


Tall fescue without endophyte

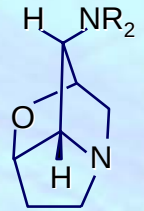
Tall fescue w/ *Neotyphodium coenophialum*

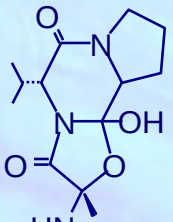
Fruiting structures of *Epichloë* species

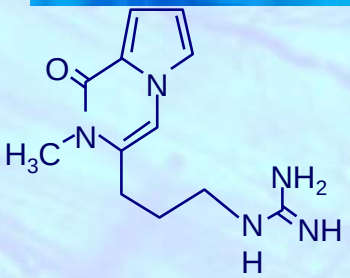
Grass-endophyte mutualisms

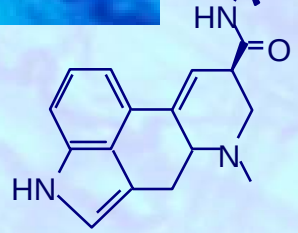


anti-insect
anti-vertebrate
anti-nematode
drought tolerance
etc.









Transmission strategies

Relative importance of vertical vs horizontal transmission.

A. Exclusively or almost exclusively transmitted horizontally. The endophyte will tend to shut down host seed production ("choke disease"), diverting available plant resources to production of infectious spores. Those spores spread to developing seeds of neighboring plants.

B. Exclusive vertical transmission. The host exhibits no disease symptoms due to the endophyte infection. Its seeds develop and germinate normally, but bear the endophyte and thereby transmit it to the next generation.

C. Mixed vertical and horizontal transmission strategy.

Host range

1. Some endophytes are restricted to individual host species. This seems rare for endophyte categories A and C above, but typical of category B.

2. Some are restricted to individual genera.

3. Some are restricted to host tribes.

4. Some are associated mainly with one host tribe, but occasionally can be identified in the sister tribe.

5. Some are present in a phylogenetically broad range of host tribes.

Problem

Question. We would like to analyze how grasses and their endophytes evolved together?

Method.

1. We use phylogenetic trees among grass species and among endophytes species.

2. Compute pairwise distances in the grass tree and in the endophyte tree.

3. Compute **MRCA pairs** of two trees.

4. Estimate the probability of codivergence between two trees and compute their correlations.

Phylogenetic trees

Data. Sequencing of Chloroplast DNA (cpDNA) Non-Coding Regions. 27 species in each group. Sequences were entered into GenBank as accession numbers AY450932–AY450949 and xxxxx-xxxxx.

Based on published phylogenetic inference for the grass subfamily Poöideae (Soreng and Davis 1998), *Brachyelytrum erectum* was chosen as the outgroup for reconstructing the grass phylogenies. The corresponding endophyte, *Epichloë brachyelytri*, was the outgroup chosen for endophyte phylogenies.

We reconstruct phylogenetic trees of grasses (the host tree, T_H) and phylogenetic trees of endophytes (the parasite tree, T_P) via a software `PAUP*` under the GTR+G+I model.

Then we used a software `r8s` to make trees unltrametric (using the least square method).

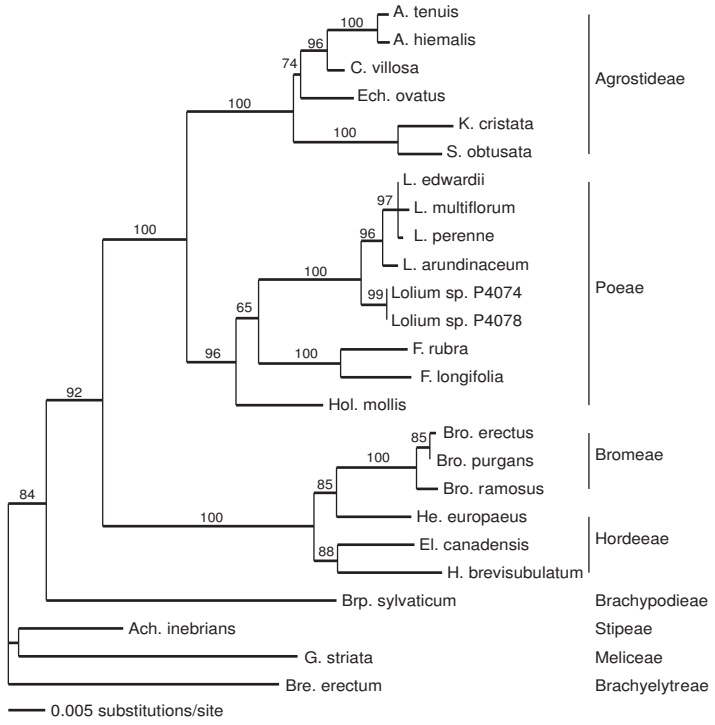


Figure 1: Parametric ML tree estimated from cpDNA intron and intergenic sequences. Numbers above branches indicate bootstrap support percentages (over 50%) obtained by 1000 maximum parsimony searches with branch swapping.

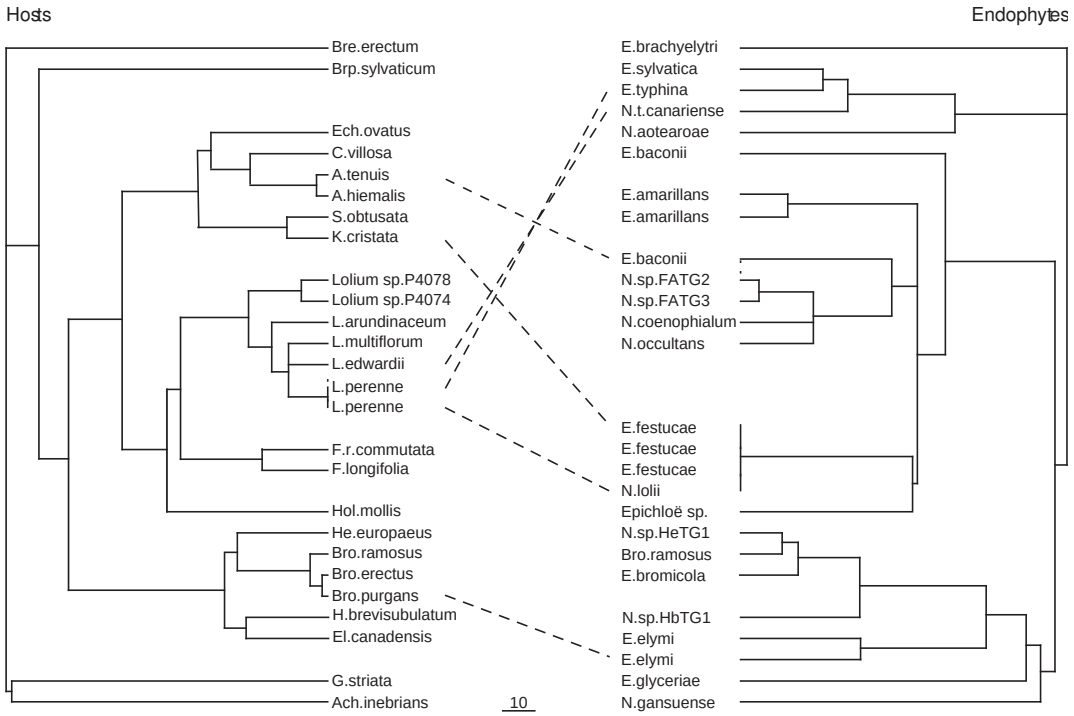


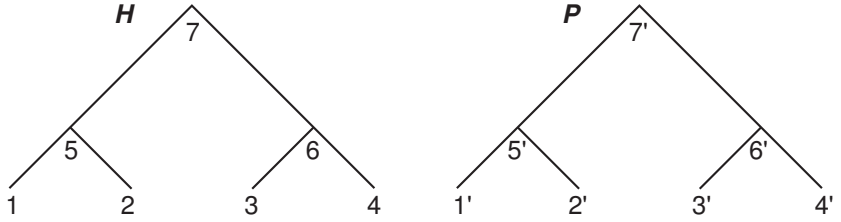
Figure 2: Ultrametric ML time trees for host grasses and their endophytes. Hosts and their endophytes are indicated opposite each other or by connecting dashed lines. Full taxon names are given in Table 1 in our paper.

MRCA pairs

A **MRCA pair** is a pair of a Most Recent Common Ancestor (MRCA) of any pair of host species and a Most Recent Common Ancestor (MRCA) of any pair of parasite species.

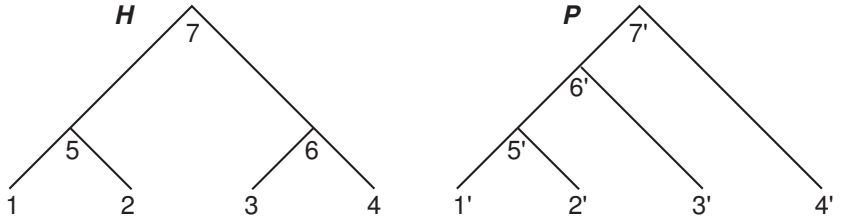
MRCA pairs

Congruent trees



MRCA pair	Pairs of *H* and *P* taxon pairs
(5,5')	((1,2),(1',2'))
(6,6')	((3,4),(3',4'))
(7,7')	((1,3),(1',3')), ((1,4),(1',4')), ((2,3),(2',3')), ((2,4),(2',4'))

Incongruent trees



MRCA pair	Pairs of *H* and *P* taxon pairs
(5,5')	((1,2),(1',2'))
(7,6')	((1,3),(1',3')), ((2,3),(2',3'))
(7,7')	((1,4),(1',4')), ((2,4),(2',4'))
(6,7')	((3,4),(3',4'))

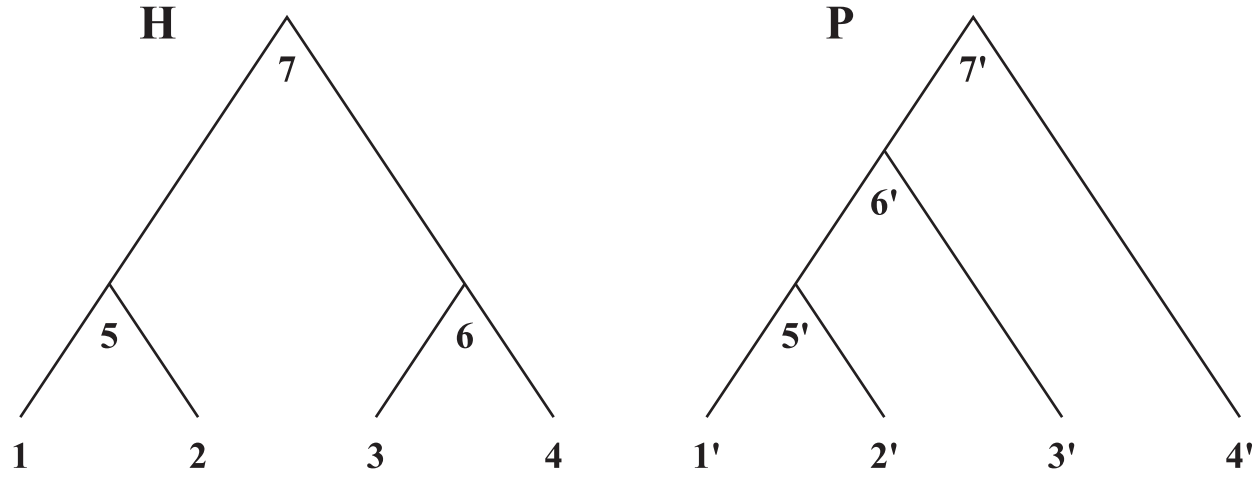
Analysis on codivergence

[Legendre et al 2002] etc used all possible pairs of pairwise distances from the host tree and the parasite tree and used Principal Components Analysis (PCA) to compute their correlations.

A problem of their method is that we possibly pick the same Most Recent Common Ancestor (MRCA) pair multiple times. This causes a bias in the result. In each tip clade a MRCA uniquely relates two taxa. However, a MRCA deeper in the tree relates multiple taxon pairs. For example, for congruent H and P trees the matrix of all pairwise distances of H taxon pairs against all pairwise distances of P taxon pairs represents each corresponding pair of tip clade MRCAs only once, and each corresponding pair of deeper MRCAs multiple times. The **MRCALink algorithm** samples corresponding H and P MRCA pairs only once.

MRCALink algorithm

We will go though the algorithm with an example.



Step 1: Assign each node a unique number such that its number is bigger than its children.

Step 2: **for** each interior node in H, from all possible pairs of offsprings, find corresponding pairs in P.

5: From $5 = (1, 2)$, we find a new MRCA $5' = (1', 2')$ in P.

6: From $6 = (3, 4)$ we find a new MRCA $7' = (3', 4')$.

7: From $7 = (1, 3) = (2, 3) = (2, 4)$, we find new MRCAs $6' = (1', 3')$ and $7' = (1', 4')$.

Thus, we have pairs $(5, 5')$, $(6, 7')$, $(7, 7')$, $(7, 6')$.

Computing the probability of codivergence

Let τ_H be the set of all ultrametric host trees with n taxa and let τ_P be the set of all ultrametric parasite trees with n taxa.

$$S(X, Y, T, t) = \sum_{x \in X, y \in Y} |\text{time}(\text{MRCA}(x)) - \text{time}(\text{MRCA}(y))|,$$

where $T \in \tau_H$, $t \in \tau_P$, X is a set of pairs of taxa in H, and Y is a set of pairs of taxa in P.

Then we estimate the probability

$$P(S(X, Y, T_H, T_P) \leq S(X, Y, T, t) : \forall T \in \tau_H, \forall t \in \tau_P)$$

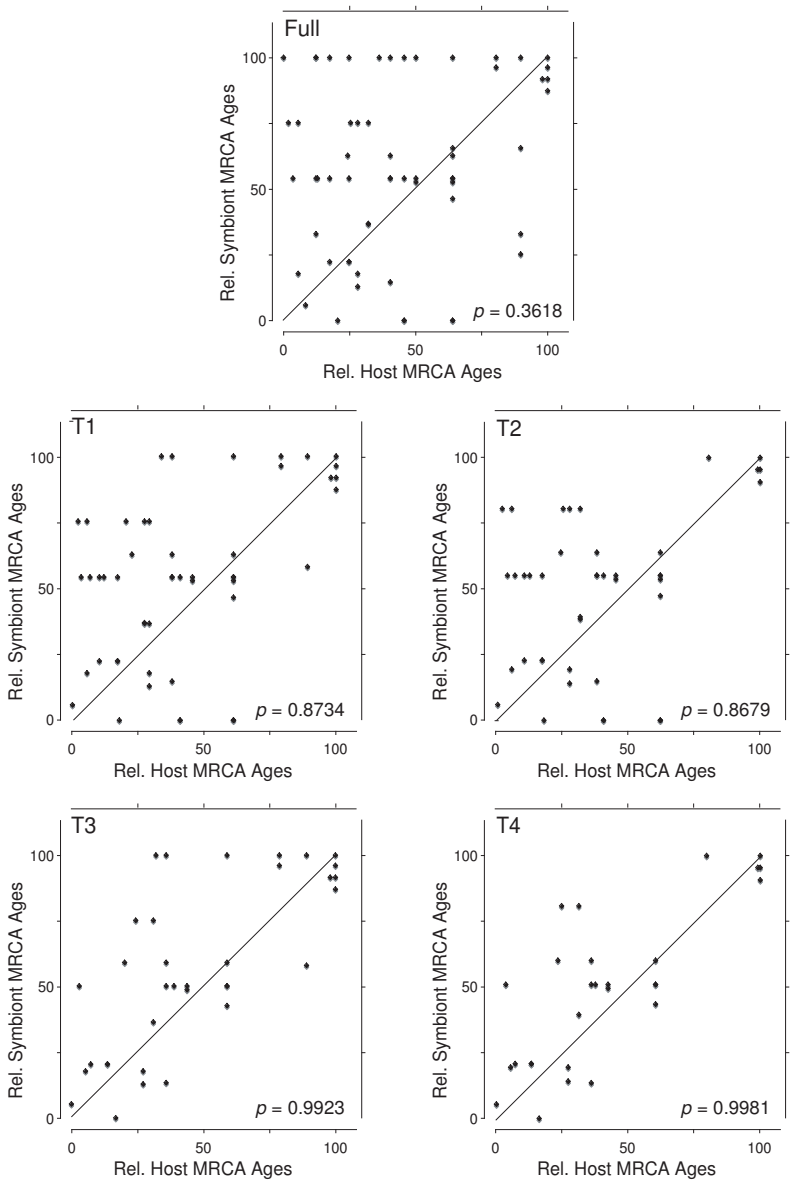
which is the estimated probability of codivergence for T_H and T_P, by randomly generated trees from τ_H and τ_P.

Results

We analyzed 4 pairs of host trees and parasite trees, namely the full tree and T_1–T_4 by removing some of species in the full trees, trimmed trees (T_1–T_4).

For each pair of trimmed trees, we removed some species from the endophytes and corresponding grasses because these endophytes seem to have holizontal or mixed transmission.

Plots of MRCA ages of hosts and their corresponding endophytes identified by the MRCALink algorithm from ultrametric ML trees for the full dataset or trimmed datasets T_1–T_4. Root ages were set to 100 arbitrary units for both trees. The diagonal lines represent expectation for perfect cospeciation and perfect phylogenetic reconstruction.


Full
Rel. Symbiont MRCA Ages
Rel. Host MRCA Ages
p = 0.3618
T1
p = 0.8734
T2
p = 0.8679
T3
p = 0.9923
T4
p = 0.9981
0
50
100

Questions

[Huelsenbeck et al, 2000] showed an algorithm to estimate the MLE for host switching via MCMC. We would like to estimate the MLE of the probability of host switching in endophytes.

All existing methods assume that the host and parasite trees are true trees. Can we reconstruct phylogenetic trees including host switching and also information of relations between hosts and parasites?

How are the host tree and the parasite tree related in the tree space?

More detailed analysis needs to be done.

A preprint is available at arxiv:

`http://arxiv.org/abs/q-bio.PE/0611084`

`MRCALink` will be available soon

Thank you....