The Dilemma of Delineation: A Species Level Analysis of *Penicillium* section *Aspergilloides*

Joseph Hindi, Daniel Raudabaugh, Andrew N. Miller

Department of Molecular & Cellular Biology, College of Liberal Arts and Sciences, University of Illinois at Urbana-Champaign; Department of Plant Biology, College of Liberal Arts and Sciences, University of Illinois at Urbana-Champaign; Illinois Natural History Survey, University of Illinois at Urbana-Champaign, Champaign, IL, USA

INTRODUCTION

In fungal ecology, the ability to properly delineate fungal species is crucial in understanding species level ecological roles (Seifert et al. 2007) and understanding species level community diversity. Species delineation is even more challenging since environmental sequencing has greatly expanded our ability to sequence unculturable and unicellular fungi.

Environmental sequencing utilizes a single-locus, typically the ITS (internal transcribed spacer) region (Seifert et al. 2007) for community species determination and analysis. Unfortunately, current research demonstrates that species delineation can be problematic with fungi because the ITS region does not provide adequate species level resolution for several common genera, including *Aspergillus*, *Fusarium*, *Penicillium* and *Trichoderma* (Seifert et al. 2007, Garcia-Hermoso 2012, Visagie et al. 2014), which together typically represent a large portion of many fungal communities (Houston et al. 1998).

Consequently, different species delineation models have been created and implemented as alternative ways to explore species level boundaries.

PURPOSE

Determine which species delineation model-gene region combination most accurately delineates *Penicillium* species within section *Aspergilloides* using a mock community comprised of 51 known *Penicillium* species, and apply that model to 13 unknown *Penicillium* isolates.

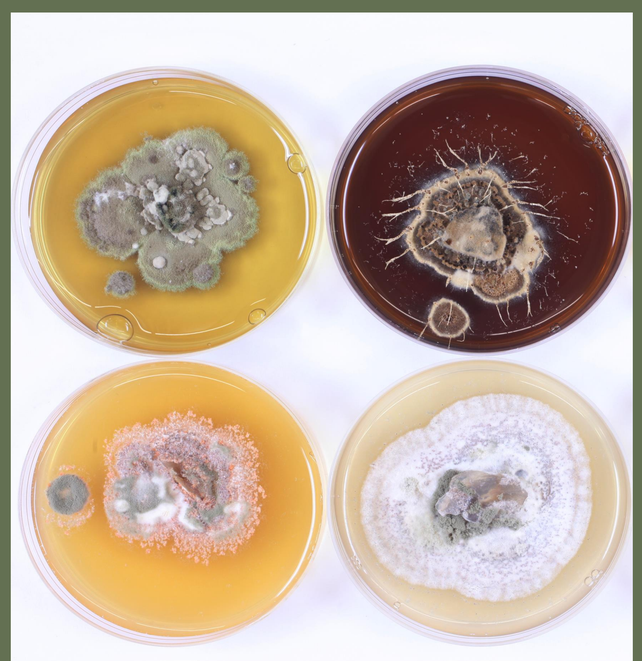

Figure 1. *Penicillium* isolates on malt extract agar.

METHOD

The genus chosen for examination was *Penicllium* because the ITS region in this genus lacks the variability to delineate between species (Seifert et al. 2007, Schoch et al. 2012, Visagie et al. 2014). Due to the limitations of the ITS, three other species level markers, RPB2, CaM, and BenA (Ribosomal Polymerase II, Calmodulin, and β-tubulin, respectively) were evaluated using the same three models.

Mock Community

The mock community was composed of 51 species selected from Houbraken et al. (2014). For each species, sequences were retrieved from the NCBI GenBank database for ITS, RPB2, CaM, and BenA using the associated GenBank accession number. To determine the ability of each model to correctly delineate duplicate species, 15 additional sequences of 5 species were added to the mock community after the singleton analysis. Duplication consisted of replicates of 6, 5, 4, 3, and 2.

Sequence Acquisition, Tree Building, & Modeling

The mock community was composed of 51 species selected from Houbraken et al. (2014). For each species, sequences were retrieved from the NCBI GenBank database for ITS, RPB2, CaM, and BenA using the associated GenBank accession number. Sequences for the sequence similarly was completed using USEARCH 9.2.64. For both tree based models, a maximum likelihood phylogenic tree was generated in SEAVIEW set to the following parameters: GTR model, branch support set to aLRT (SH-like), empirical nucleotide equilibrium frequencies, optimized invariable sites, optimized setting across site rate variation, and best of NNI and SPR for tree searching operations. The starting tree used was set to BIONJ (optimized tree topology) with 5 random starts. The tree was then rooted using the isolate with the longest branch length.

The GMYC model was run in Rstudio using the splits package (Fujisawa et al. 2013). The bPTP model was run using the bPTP.py in python (Zhang et al. 2013). The number of delineated species was recorded for each model. The confidence interval (CL) was recorded for both tree based models. The tree produced from the GMYC and bPTP model were visually examined to review the models ability to group similar species together.

RESULTS

Table 1. Species delineation results for the mock community lacking species replicates.

	USEARCH		GMYC		bPTP	
	Sequence Identity (%)	Cluster (#)	ML Entities	Confidence Interval	ML Entities	Confidence Interval
ITS	97	3	25	12-31	30.72	11-44
	99	9				
RPB2	97	21	7	5-8	10.99	8-25
	99	42				
BenA	97	33	10	9-12	24.69	9-40
	99	47				
CaM	97	31	7	2-20	13.47	7-34
	99	48				

Table 2. Species delineation results for the mock community with species replicates.

	USEARCH		GMYC		bPTP	
	Sequence Identity (%)	Cluster	ML Entities	Confidence Interval	ML Entities	Confidence Interval
ITS	97	3	41	2-42	41.89	2-62
	99	9				
RPB2	97	21	6	2-9	10.99	8-26
	99	45				
BenA	97	34	11	8-12	30	9-49
	99	48				
CaM	97	32	2	2-29	25.88	11-46
	99	52				

Table 3. Species delineation results using unknown *Penicillium* species.

	USEARCH		GMYC		bPTP	
	Sequence Identity (%)	Cluster	ML Entities	Confidence Interval	ML Entities	Confidence Interval
ITS	97	6	2	1-11	5.95	1-13
	99	12				
BenA	97	13	4	1-9	7.74	1-13
	99	13				
CaM	97	13	2	1-11	6.79	1-13
	99	13				

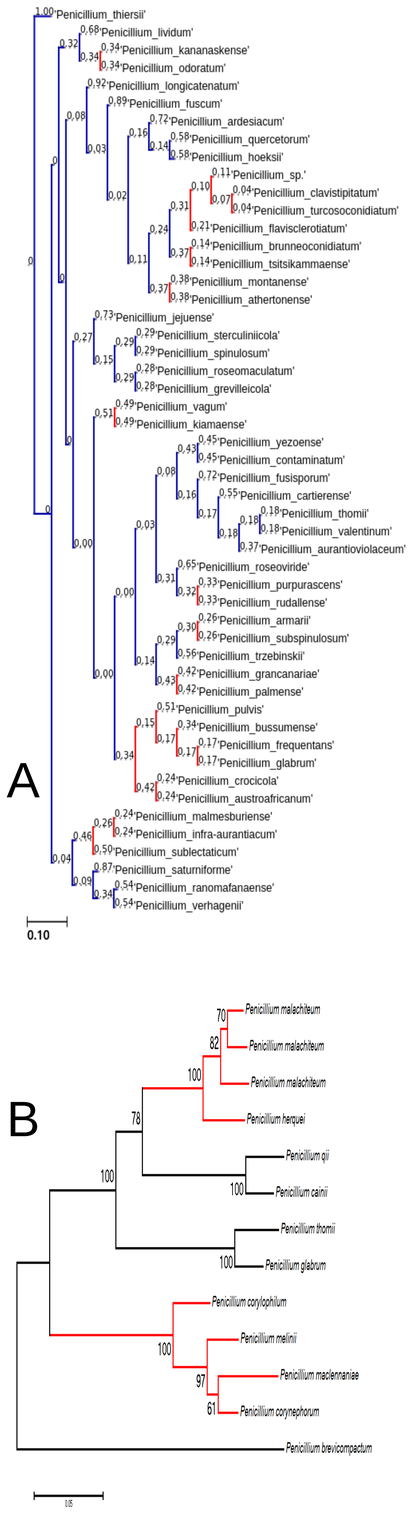

Figure 2. bPTP ITS tree results for A) the mock community and B) the unknown isolates.

CONCLUSIONS

- The bPTP model had a wider confidence interval as compared to the GMYC model, however, the bPTP model was more accurate than the GMYC model in predicting the total number of closely related *Penicillium* species.
- Species delineation trees produced from the bPTP and GMYC illustrated that both models lack the proper sensitivity to accurately group replicated species.
- ITS 99% sequence similarity, BenA, and CaM were more accurate than both tree models in predicting the total number of unknown distantly related species.

This study demonstrates the importance of using multiple species delineation methods and alternative genes when evaluating species level determinations.