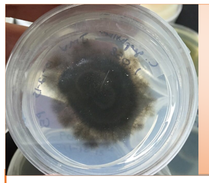
RADseq phylogenetic inferences using *de novo* downstream analyses as compared to single gene sequencing of the ectomycorrhizal fungus *Cenococcum geophilum*

Jessica M. Vélez[1,2], Todd W. Pierson[7], Reese M. Morris[1], Steven J. Lebreux[1], Jessy Labbe[1], Luz M. Serrato-Diaz[5], Rytas Vilgalys[3], Scott J. Emrich[4], Christopher W. Schadt[1,2,6,*]

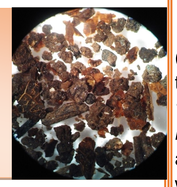
[1]*Oak Ridge National Laboratory,* [2]*The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee,* [3]*Duke University,* [4]*Min H. Kao Dept of Electrical Engineering and Computer Science, University of Tennessee,* [5]*University of Puerto Rico - Río Piedras Campus,* [6]*Department of Microbiology, University of Tennessee,* [7]*Department. of Ecology and Evolutionary Biology, University of Tennessee*

BACKGROUND

- Agricultural security is essential to economic security.
- A significant portion of land is contaminated with heavy metals.
- Biofuel crops such as Poplar could be planted instead.
- Increasing the ability of Poplar to grow in these areas increases the potential for agricultural development and can remediate the land.
- A soil symbiont may be able to increase heavy metal tolerance.
- The ubiquitous and hardy fungus *Cenococcum geophilum* will be studied to determine if this species can increase Poplar heavy metal tolerance.
- The genetic complexity of *C. geophilum* promotes the use of the novel ddRAD-seq technique.



Cenococcum geophilum

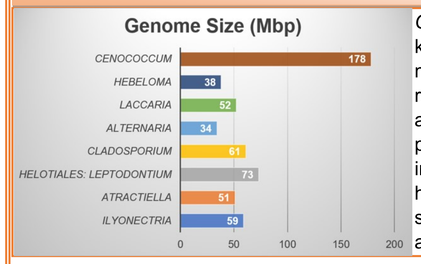


The fungal species *Cenococcum geophilum*, an ubiquitously distributed ectomycorrhizal fungus which is positively associated with plant health, growth, and increased contaminant resistance. This fungus is known to associate with both angiosperm and gymnosperm tree species across 40 genera, representing over 200 total species, and is generally tolerant across a salinity gradient, water stress conditions, and extreme heavy metal contamination conditions. This wide-ranging resistance to stressful conditions in a soil environment is often attributed to a high melanin content, as *C. geophilum* is phenotypically deep black in coloring. This high concentration of melanin has been directly implicated to play a role in increased resistance to heavy metal contaminants.

The genome of *C. geophilum* is among the largest in the fungal kingdom, with a mapped size of 178 Mbp and a total estimated size of up to 203 Mbp. *Cenococcum geophilum* has no documented sexual or asexual spore production, and is considered asexual as a species despite high levels of genetic and physiological diversity. Due to these high levels of intraspecies diversity found among *C. geophilum* isolates, even within isolates from a single tree, *C. geophilum* is proposed to be a species complex, with many studies finding significant variations in cultured isolate characteristics and physiology. This type of cryptic species population implicates the incorporation of morphologically identical members which are separate species on the genetic level. As a result, while *C. geophilum* appears to have a myriad of potential benefits for agricultural applications, this fungus has been difficult to study phylogenetically, as this research is necessarily resource-intensive in order to delineate the phylogenetic relationships observed within the species complex.

SINGLE GENE versus RADseq *DE NOVO* ASSEMBLY

WHY RADSEQ?

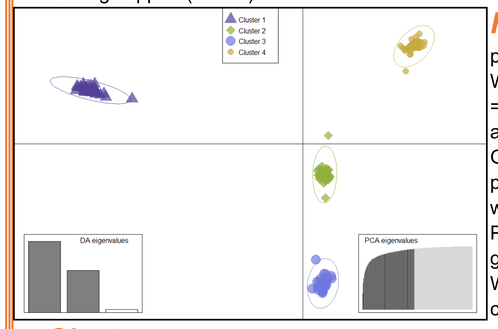


Cenococcum geophilum has one of the largest known genomes in the fungal kingdom, making full-genome studies costly and resource-intensive. RADseq is a low-cost alternative to full genome sequencing which provides robust *de novo* SNP detection, which increases intraspecies resolution. This method has been successfully applied to multiple species. However, RADseq has rarely been applied to fungi.

SINGLE GENE versus RADseq

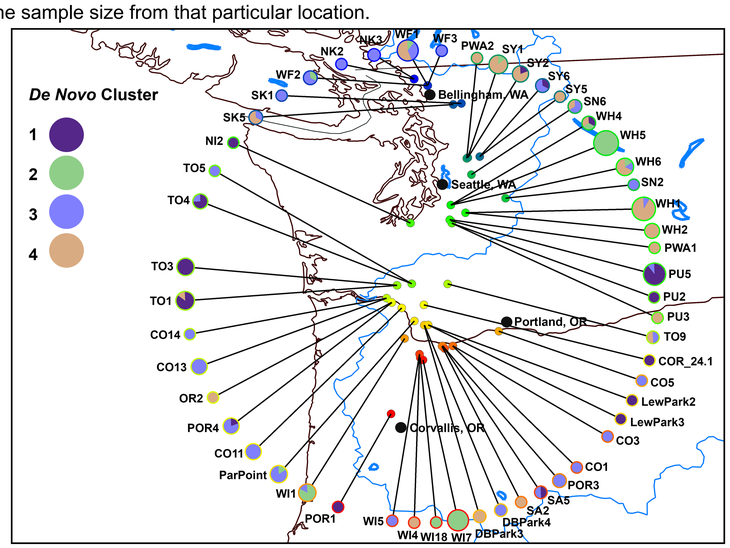
231 unique strains were isolated from 60 sites, with a total geographic range of 283 miles from central Oregon to northern Washington State. A maximum likelihood phylogeny of the Pacific Northwest isolates (left) using the glyceraldehyde 3-phosphate dehydrogenase (GAPDH) gene region revealed 15 strongly resolved (>80%) clades (**numbered**).

Fifteen well-supported clades (>80%) were implicated within the RADseq dataset (right). Generally, there is a negative correlation (r = -0.82) between the *de novo* cluster and the clade assigned based on the GAPDH gene. However, when the largest GAPDH clades, 10 and 11, are examined separately, there is a positive correlation (r = 0.70) between the *de novo* assigned clade and the clade assigned based on the GAPDH gene. Reflecting this, apart from isolate WH2_54.10, all of GAPDH clade 11 are included in the *de novo* clade 9 with strong support (81.1%).



Four phylogenetic clusters are present in the RADseq dataset (right). While there is a negative correlation (r = -0.82) between the *de novo* cluster and the clade assigned based on the GAPDH gene, clade 11 is predominantly assigned to cluster 1 with the exceptions of isolates PU5_60.2, and TO4_90.5, which are grouped into cluster 3, and WH2_54.10, which is grouped into cluster 4.

Clusters have no noticeable patterns of segregation relating to latitudinal values (below), and many individual isolation sites also contained more than one cluster regardless of the sample size from that particular location.



PNW GAPDH RAxML Phylogeny (Left) compared to RADseq De Novo Assembly (Right) RAxML Phylogeny