Jessica Maccaro
GEN220
Dec 18, 2020

## Introduction:

*Ascosphaera* is a diverse genus of fungi that only associates with bee nests and comprises species that can range from commensals to virulent bee pathogens (Anderson et al., 1998). Its type species, *A. apis,* causes the chalkbrood disease in *Apis mellifera (*honey bees), while another known pathogen in the genus, *A. aggregata*, infects the leaf-cutting alfalfa bees *Megachile rotundata*. The infection starts after the ingestion of ascospores, with penetration of the gut epithelium and invasion of the hemocoel with subsequent systemic mycosis. However, the pathogenicity mechanisms are yet to be fully elucidated (Boomsma et al., 2014).

In a collaborative project this summer (with Moreira Salgado F., Argueta Guzmán, M.P, Gnor, L., Klinger, E., Stajich, J., McFrederick, Q.) , we looked at which genes might be under positive selection in the two pathogenic *Ascosphaera spp.* We captured single copy orthologs (using orthofinder) to infer the phylogeny and created codon- aware alignments to determine orthologs under positive selection in our species (using ete3). Two metabolism related genes were found under significant positive selection in the pathogens: D-arabinitol dehydrogenase and NADH dependent xylose reductase.

In this class project, I wanted to understand the broader context for one of these enzymes (D-arabinitol dehydrogenase) across the Onygenales (the order that contains *Ascosphaera).* This order is very diverse in terms of lifestyles, with representatives that vary from saprobes to pathogens. This makes it ideal for the question I was interested in: **If I build a tree with only the D-arabinitol dehydrogenase homologs will the fungi cluster by lifestyle?**

## Methods:

To address this question, I first needed a database with Onygenales proteomes. This comparison had to be done with amino acid sequences because it was so broad- across the order-level. Jason shared with me his database that contained all the Onygenales proteomes and I just made a symlink of the file (allseqs) to my directory.

Next I wanted to search this database with my 5 single copy ortholog sequences of D-arabinitol dehydrogenase from my 5 *Ascosphaera spp.* I started by creating an hmm specifically for these *Ascosphaera* specific sequences with hmmbuild (Copyright (C) 2019 Howard Hughes Medical Institute). Next, I was able to search this against the database using hmmsearch. With the output of this, I was able to visualize in a tab delimited file, the hits with the lowest E values. Since the sequences were ordered from lowest to highest E-values, I was able to just pull the first 50 lines with head. I got a few repeats, most likely because some sequences had both forward and reverse strands, so I removed repeats with awk. I was then able to make a file with these names by using awk and printing the column with the sequence names.

Next, I was hoping to use esl-fetch to index and pull sequences from the database by suppling the names of the top hits. I ran into a problem indexing the database though, because there appeared to be repeats. However, these repeats were not in the sequences I was searching for, so I tried a different method to be able to search a list of names and pull their sequences. I tried using BioPython for this by importing Seq from Bio.Seq. However, I was still running into problems and I also did not want to mix programming languages too much since I thought of a way to do it with shell.

What I needed to do next was restructure the database so that every sequence name was followed by its full sequence on the next line, as a single line. That way I would be able to search a list of names and just pull the line after the name with grep -A 1. I restructured the database just using tr and sed. I first put the whole file into one line. Then I broke up those lines by making a new line before each “>”. Since the sequence names seemed like they ended in an identifying number, I just used regular expressions to create a new line when the numbers ended. The idea here was that when the numbers ended the sequence would begin, which was letters because they were amino acids. However, it created some issues down the line which I will get to in the next paragraph.

So now I had my database formatted in a way conducive to making a for loop that would iterate through a list of names (the names of the top hits) and pull the name plus the line after (the sequence) with grep -A 1. I then concatenated all these sequences into one file, which was ready to align. However, I ran into a couple more problems. (1) Some of the sequences were named differently than others so annotation information ended up on the same line as the sequence for a few. This happened because when I was restructuring this database file, I thought I could just have a new line start after the numbers ended, but in some cases after the numbers ended the annotation information began. There were also some differences in how the annotation information was added to each sequence which made it difficult to clean in an automated way. Therefore, I had to just manually go through and make sure the sequences were formatted correctly and only the sequence itself was on the line underneath the sequence identification. This was fairly easy to do manually because there were less than 50 sequences at this point. (2) Next, I realized that for some reason I ended up with more sequences than my top hit names. I figured this out by just grep -c ">" the sequence file and saw how this varied from the top hit's names amount. I thought this was unusual because they all had specific sequence ID numbers. So, I thought I would just do grep -f with the top hit names again and pull the line after. This did end up removing those spurious sequences. Then I was able to add my 5 *Ascosphaera* sequences that I started with to this file so that they would be included in the downstream alignments and tree building.

Now it was time for alignment and tree building. To align, I used hmmalign and included my specific hmm I made at the beginning for my *Ascosphaera* D-arabinitol dehydrogenase and the file full of sequences of the top homologs. I then reformatted this to be in clustal format using esl-reformat. Then I was able to use iqtree2 to build a tree using maximum likelihood (Bui Quang Minh et al., 2020; Kalyaanamoorthy et al., 2017). I then visualized the trees using FigTreev1.4.4 and color coated them by lifestyle.

**(script and other important files: https://github.com/Jmaccaro/GEN220Project2020)**

## Results:

My top hits across the Onygenales represented a diversity of lifestyles. I summarize the species, their lifestyle and host in Table 1 (using information from: Caballero Van Dyke et al. 2019, Garces et al 2020, Geiser et al. 2006). I represent my tree in a couple different ways to visualize the results. Figure 1A shows the species grouping with unscaled branch lengths for easy of viewing the clustering. Whereas Figure 1B has scaled branch lengths to represent amount of genetic change. Note that there was more than one sequence per species, but groupings remain similar by species even across the different sequence. These difference sequences represent different enzymes that are most similar including: variations of D-arabinitol dehydrogenase, gluconate 5-dehydrogenase, alcohol dehydrogenase, hydroxysteroid dehydrogenase, L-xylulose reductase, mannitol dehydrogenase, and oxidoreductase. You can see that there is still some loose clustering occurring by lifestyle but that it is not monophyletic (Figure 1A).

## Discussion:

Ultimately, there was some grouping by lifestyle. These groupings did not appear monophyletic which could simply be because the top hits I pulled represented different genes in some cases. This seemed like a major flaw, so next time I would make sure I'm comparing the same gene across the order instead of just arbitrarily picking the top 50 hits. After doing this I would still need to do gene-tree species-tree reconciliation to really get at my original question. I wanted to know if the pathogens of this order would group together for the gene I found under positive selection in two *Ascosphaera* pathogens. If this gene was extremely important in pathogenicity, then maybe I would have seen a stronger signal of this. However, my candidate gene was not very strong to start out with because it is just a metabolic gene. Although, the upside of it being a metabolic gene was that it is present in a wide variety of fungal lifestyle, making it ideal for comparative approaches.

I am re-running our initial analysis on the *Ascosphaera* species but removing one of the genomes that was only 20% complete. This increased the amount of orthologs the pathogens shared, and I am hoping that with stronger candidate genes I can attempt this again. Then perhaps, it will be worth it to create a species tree of Onygenales and compare it to the gene trees using Notung or Mantel Tests. Just to add a few more caveats though, even if the species tree is equal to the gene tree it does not mean that the gene is not important for pathogenicity and conversely if the species tree is not equal to the gene tree it still does not imply that the gene is important for pathogenicity. If the species and gene trees matched, at least you can be amazed that this single gene was enough to recapitulate the species tree relationships. If the species tree did not match the gene tree, it would generate some interesting hypotheses about the possibility of horizontal gene transfer driving pathogenicity.

**Table 1**: Summary of species used to create tree and their lifestyle.

| Species (Abbreviation from Tree) | Lifestyle | Hosts |
|---|---|---|
| • *Blastomyces dermatitidis* (BDCG)<br>• *Coccidioides immitis (Cimm)*<br>• *Emergomyces pasteurianus (AJ78)*<br>• *Emmonsia crescens (EMCG)*<br>• *Emmonsia parva (EMPG)*<br>• *Paracoccidioides brasiliensis (Pbra)*<br>• *Ophidiomyces ophiodiicola (SFD)*<br>• *Histoplasma capsulatum (HCBG)* | Systemic Pathogens | Vertebrates |
| • *Ascosphaera apis (AAP)*<br>• *Ascosphaera aggregate (Aagg)* | Systemic Pathogens | Bees |
| • *Ascosphaera acerosa (Aace)*<br>• *Ascosphaera atra (Aatr)*<br>• *Ascosphaera pollinicola (Apol)* | **Unclear**<br>Saprobes or Opportunistic Pathogens | |
| • *Amauroascus mutatus* (AMMUT)<br>• *Amauroascus niger* (AMNIG)<br>• *Byssoonygena ceratinophila (BYCER)*<br>• *Onygena corvina (OCOR)* | Saprobes | |
| • *Microsporum canis (MCYG)*<br>• *Nannizzia gypsea (MGYG)*<br>• *Trichophyton rubrum (Trub)*<br>• *Trichophyton soudanense (H105)* | Skin Pathogens | Vertebrates |
| • *Aspergillus fumigatus (Afum)* | Opportunistic Pathogen (Outgroup) | Vertebrates |

Jessica Maccaro
GEN220
Dec 18, 2020

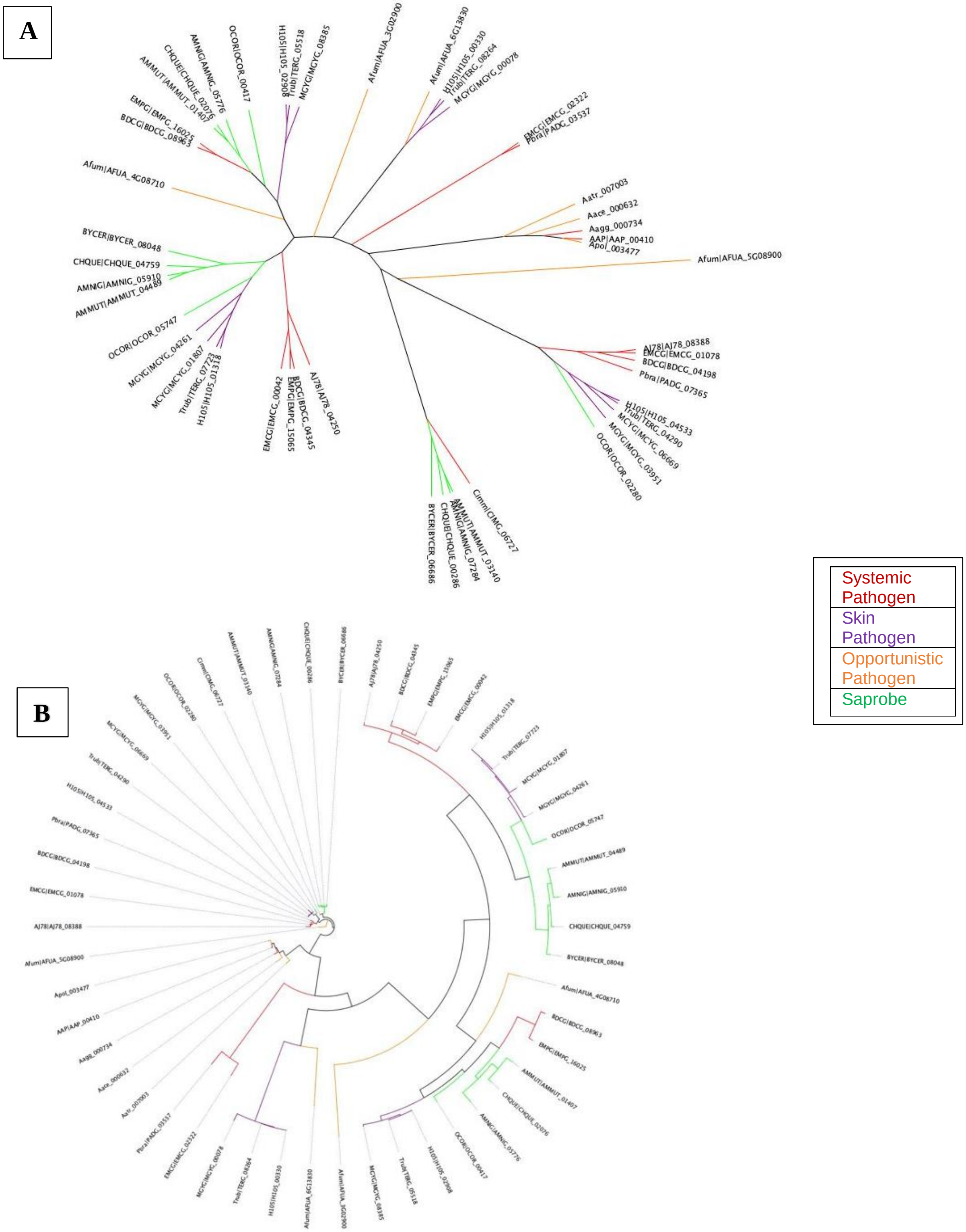

**Figure 1: Phylogeny of D-arabinitol dehydrogenase homologs across Onygenales.** Clades are color coated for each lifestyle as indicated in the key. **(A)** Branch lengths are unscaled for ease of viewing the clustering**. (B)** Branch lengths scaled.

Jessica Maccaro
GEN220
Dec 18, 2020

## References:

Anderson, D.L., Gibbs, A.J., Gibson, N.L., 1998. Identification and phylogeny of spore-cyst fungi (Ascosphaera spp.) using ribosomal DNA sequences. Mycol. Res. 102, 541–547. https://doi.org/10.1017/S0953756297005261

Boomsma, J.J., Jensen, A.B., Meyling, N.V., Eilenberg, J., 2014. Evolutionary Interaction Networks of Insect Pathogenic Fungi. Annu. Rev. Entomol. 59, 467–485. https://doi.org/10.1146/annurev-ento-011613-162054

Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol., in press. https://doi.org/10.1093/molbev/msaa015

David M. Geiser, Cécile Gueidan, Jolanta Miadlikowska, François Lutzoni, Frank Kauff, Valérie Hofstetter, Emily Fraker, Conrad L. Schoch, Leif Tibell, Wendy A. Untereiner & André Aptroot (2006) Eurotiomycetes: Eurotiomycetidae and Chaetothyriomycetidae, Mycologia, 98:6, 1053-1064, DOI: 10.1080/15572536.2006.11832633

Garcia Garces, H., Hamae Yamauchi, D., Theodoro, R.C. et al. PRP8 Intein in Onygenales: Distribution and Phylogenetic Aspects. Mycopathologia 185, 37–49 (2020). https://doi.org/10.1007/s11046-019-00355-6

Marley C Caballero Van Dyke, Marcus M Teixeira, Bridget M Barker. Fantastic yeasts and where to find them: the hidden diversity of dimorphic fungal pathogens. Current Opinion in Microbiology, 2019 Volume 52, Pages 55-63, ISSN 1369-5274, https://doi.org/10.1016/j.mib.2019.05.002.

Subha Kalyaanamoorthy, Bui Quang Minh, Thomas KF Wong, Arndt von Haeseler, and Lars S Jermiin (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature Methods, 14:587–589. https://doi.org/10.1038/nmeth.4285