Using sequences from NCBI GenBank to estimate how many species there are in the fungi genus *Stereum*

Sarah DeLong-Duhon

Reference: Fryssouli, V., Zervakis, G., Polemis, E., & Typas, M. A. (2020). A global meta-analysis of ITS rDNA sequences from material belonging to the genus Ganoderma (Basidiomycota, Polyporales) including new data from selected taxa. MycoKeys, 75(December), 71–143. https://doi.org/10.3897/mycokeys.75.59872

Intro There are estimated to be 2.2 to 2.8 million fungi species on earth, but we have discovered only a small fraction of them (Hawksworth & Lücking 2017, Vu et al. 2014). Even large mushroom-forming fungi, which are easy to find and document, are not very well characterized. Understanding species diversity in fungi is important for many reasons, that encompass both human interests and intrinsic value. Fungi provide important ecosystem services, particularly the recycling of nutrients, but are also useful in medicine and industry. In addition, the unfortunate reality is that unless we discover biodiversity now, we will never know what we have lost as human activity continues to transform the planet.

Background

As an undergraduate, I started a research project to collect and sequence some mushrooms – specifically, of the wood-decay genus *Stereum* – from Iowa. It was difficult to tell the species apart from morphology alone, so I figured I would sequence them and create a phylogenetic tree, using the accepted DNA barcode for fungi, the internal transcribed spacer (ITS). This revealed that one of the most common species in North America, *S. ostrea*, was actually three genetically distinct species that formed a well-supported clade together, and had been previously described but popularly considered synonyms of *S. ostrea* due to their similar morphology. Mapping phenotypic traits on to this phylogeny, however, made it clear that although similar and highly variable, there were a few key characteristics that could be reliably used to differentiate these species. Now, for my master's thesis work, I want to create a global molecular phylogeny for the genus, as *Stereum* occur commonly worldwide. However, publicly available sequences have issues – a high rate of misidentification, poor sequence quality, and lack of metadata being some of them. Therefore, for my master's research I am requesting herbarium samples to examine and sequence, but in the meantime public sequences can still be useful to me in another way.

One thing that is not clear is how many species of *Stereum* there are, how many are described, and how many have yet to be described. To achieve an estimate of species diversity in the genus, I will create a dataset consisting of all publicly available ITS sequences. To process this data I will develop a program to estimate how many phylogenetic species are represented in this dataset. This will allow me to pinpoint avenues of future research, especially for sequences associated with collection location metadata, since *Stereum* species tend to have their own unique geographic distributions and host tree preferences.

Materials and Methods

Figure used as template for project

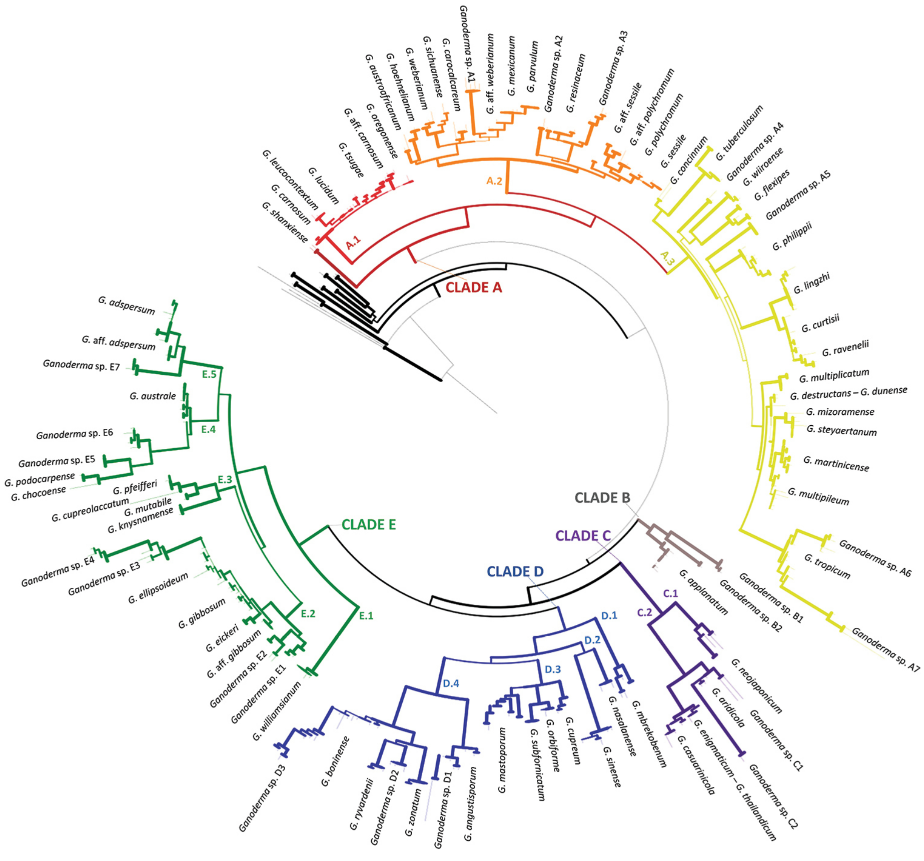

Figure 3. Summary tree of the genus *Ganoderma* inferred from ML analysis, based on ITS sequence data (main dataset, DS; Table 2). Thick lines represent ML bootstrap values (BS) greater than or equal to 65% and Bayesian Posterior Probabilities (BPP) greater than or equal to 0.95. Clades and Clusters within the tree appear as presented in Table 1 and Suppl. material 1: Table S2. Species names correspond to those inferred in this study. Scale bar: 0.01 nucleotide substitutions per site.

In Fryssouli et al. 2020, the authors used almost four thousand internal transcribed spacer (ITS) sequences from public genetic databases such as NCBI GenBank and UNITE to estimate species diversity in a genus of large shelf mushrooms, some of which have a long history of medicinal use. They found 80 *Ganoderma* species, 21 of which are not named or properly/fully identified, separated into three major lineages.

They acquired these sequences by retrieving those labeled as genus *Ganoderma*, as well as those from environmental sequencing and other misidentified entries (though they do not state clearly how they found the latter data). They excluded sequences of short length (less than 350 bases), and sequences mistakenly labeled as *Ganoderma* (what criteria they use for this is not clear, but I would use a sequence similarity cutoff of ~85%). The resulting sequences were compared for similarity and sets of identical sequences were considered amplicon sequence variants (ASV) (so a group of identical sequences were treated like a single sequence) which were preferentially used for phylogenetic analyses – Maximum Likelihood (via RaxML) and Bayesian Inference (via MrBayes).

They differentiated species with thresholds for sequence similarity – that is, ITS sequence similarity above 97-98% are indicative of same-species, which is a widely-adopted threshold for species delimitation in Basidiomycota (Nilsson et al. 2008). They also considered if the terminal subclade was well-supported, and whether any overlap existed between intraspecific genetic distance vs. respective interspecific values compared to the closest related taxon (so essentially, more closely related to each other than the next closest clade).

Data

- This project will use *Stereum* internal transcribed spacer (ITS) sequences acquired from NCBI GenBank in addition to 44 ITS sequences that are included in my manuscript (https://www.biorxiv.org/content/10.1101/2020.10.16.342840v2) that have confirmed identities, to serve as references for sequence similarity cutoffs.
    - My dataset as of **March 17, 2021** (consists of a preliminary collection, to initially avoid the issues of managing a huge amount of data:
        - 44 reference sequences (see above)
            - SGD_stereum_alignment.fasta

        - 207 sequences acquired from GenBank using the multiple BLAST searches (https://blast.ncbi.nlm.nih.gov/Blast.cgi) on select reference sequences representing several species, with a similarity cutoff of ~95%
            - Stereum_all_1_MAFFT.fasta
    - I will add to/improve this dataset later with methods as follows:
        - First, I will search for sequences assigned to the genus *Stereum*, and discard any that are mislabeled (similarity of less than 85% compared to my Stereum sequences) or too short (less than 375 bases).
        - Then, I will use NCBI GenBank BLAST to locate sequences that are similar to those I already have (my usual outgroup *Xylobolus subpileatus* is 83-78% similar, so I think 85% will be a good cutoff).

Methods

- Align all sequences in the dataset using MAFFT (usually I do this via CIPRES but can explore other options)
- Combine identical sequences while preserving their names (maybe I can combine them, like MK269283_MK269284_MK269286)
- Use sequence similarity cutoffs (97%) to sort sequences into groups, which will represent species
- Create a figure to visually illustrate the groups (circular phylogenetic tree and/or K-means clustering)

Reflections My personal challenge with this project is that I have somehow avoided learning programming up to this point. I have a clear idea of what I want to achieve, but run into all kinds of obstacles while figuring out how to do it. Despite that, I am excited to finish this project and learn new skills, even if it takes me a while to get up to speed and start to develop some kind of programming literacy.