**Constructing a robust bioinformatic pipeline to investigate taxonomic relationships within the genus *Amanita***

Earth is experiencing significant losses of biodiversity1–3, but fungi are rarely included in conservation science4 due, in part, to scientists historically assuming most fungi (especially mushrooms) were globally distributed. A local extinction of a cosmopolitan species is less concerning than the extinction of an endemic one. However, over the last few decades, DNA sequence data have changed our understanding of “cosmopolitan fungi”. In fact, most fungal lineages are made up of cryptic, endemic species found within restricted ranges5. Fungal species are often difficult to tell apart morphologically and, historically, more attention was paid to naming plants and animals. Only 5-10% of fungal species are formally described6. This discrepancy leaves us with little useful knowledge with which we can base conservation practices off. A specific example of this challenge is found in the genus *Amanita* in Wisconsin.

The *Amanita* genus encompasses species with a wide range of ecological niches and evolutionary histories, with some considered prize edibles like *A. caesarea*7, while others like *A. phalloides* are deadly8. Of the three subgenera of *Amanita*, two are considered ectomycorrhizal (ECM) and the other is considered saprotrophic9. Although it is a charismatic and famous genus, the biodiversity of the group remains cryptic and poorly understood. There is an abundance of species complexes in *Amanita* – a direct result of historical use of solely morphological data as the basis for taxonomic organization. As DNA sequencing technologies have become available and widespread, *Amanita* taxa going under a scientific name are repeatedly found to have enough genetic difference to be considered different lineages. For example, molecular evidence suggests *Amanita muscaria*, currently understood as a globally distributed species, is likely a complex composed of at least 8 lineages with endemic and relatively narrow ecological niches10.

If we want to be effective in our efforts to conserve the biodiversity of fungi, we must have a firm grasp of the taxonomic organization of the kingdom. Increased efforts to build quality phylogenetic trees is needed across the entire kingdom. I am building a pipeline using a sample set of *Amanita* DNA sequences to have a robust tool with which I can apply to my own data once available. I present here an explanation of my pipeline and a description of my exploration of tools and models I can use to build robust phylogenies for *Amanita*.

**Methods:**

Data:

The data consist of 24 taxa each with sequences for three common loci that code for the nuclear large ribosomal subunit (nrLSU), trans-elongation factor 1 alpha (TEF1α), and RNA polymerase subunit 2 (RPB2) (Table 1). The data are a subset of taxa that were used to create the current accepted phylogeny for the genus of *Amanita*11. Taxon accession numbers for each of the 3 loci were used to create a fasta file with the tool Entrez Direct from NCBI12. To investigate the phylogenetic space of these taxa, I implement an alignment software (CLUSTAL-W13), a maximum parsimony software (phangorn14), a maximum likelihood software (IQ-Tree 215), a Bayesian inference software (MrBayes16), and, finally, a coalescent software (ASTRAL17).

Alignment:

Fasta files were aligned using two methods. The first software used was Clustal-W13, a progressive multiple sequence alignment method that boosts sensitivity through dynamic use of weights. It creates an easily customizable alignment that more accurately represents mutation types and the probability of their occurrence in different areas of the genome. Further, it implements dynamic substitution matrices meaning it changes the substitution matrix used depending on sequence divergence – matrices become more relaxed as sequences become more diverged. Despite many strengths, Clustal-W still deals with local minimum issues as the alignment starts with the most similar sequences first. If there are initial errors in the beginning stages of alignment, they will proliferate as more sequences are added. For this project, I specified DNA sequences and conservative weights of 1 for both the gap opening and gap extension penalties. The alignments were then viewed with Aliview18 and trimmed using trimAl19. If 10% or more had the gap, the gap was removed.

Maximum parsimony:

After manual inspection of alignments, the fasta alignment file was initially used to create a maximum parsimony tree in R using the software package, phangorn14. Phangorn is a tool that can implement distance based, maximum parsimony, and maximum likelihood models to produce phylogenetic trees and can even compare the trees reconstructed from each model. Distance based and parsimony methods are hardly in use anymore as they have been shown to be overly simplistic, often leading to statistical inconsistencies. They rely on the assumption that the rate of evolution is slow which is often not the case. However, parsimony trees can be used to create strong starting trees for maximum likelihood methods rather than just using an entirely random starting tree.

Maximum likelihood:

To estimate maximum likelihood (ML) phylogenies, I used IQtree215. The software is a fast, stochastic algorithm that implements bootstrap resampling to continually reassess nucleotide sites, incrementally reconstruct the tree, and provide branch support values in the form of a bootstrap score. IQtree2 is impressive in that it evaluates different substitution methods to automatically determine the model that will best fit the data. Having a stochastic algorithm also allows it to escape local optima, giving the user more confidence in the ML tree it produces. IQtree2 is more computationally expensive than other methods and still hinges on the assumptions made for all ML methods: (1) the mutation process is the same for every branch of the tree (2) sites evolve independently (3) all sites evolve the same. However, algorithms (including IQtree2) have a way to bypass the third assumption of ML by applying the gamma model of rate heterogeneity. For my analysis, I implemented the general time reversible (GTR) model and the gamma distribution model with a random bootstrap of n=1000 replicates.

Coalescence:

Using IQtree2, I created three robust gene trees for different loci of each specimen (nrLSU, TEF1α, RPB2). As none of these trees can be considered accurate evolutionary dynamics of the species, it is necessary to implement the multispecies coalescent model to gain a better understanding of the taxonomic organization of the species represented by the data. To do this, I used the software ASTRAL17 which uses maximum likelihood gene trees as its input to produce a species tree which is statistically consistent with the multispecies coalescent model. The algorithm is not computationally expensive and scales well for large data sets (100+ taxa and numerous loci). However, it should be noted ASTRAL is likely to be inaccurate if using the general parameters and with smaller data sets like the one used in this project. Furthermore, ASTRAL is not optimized for bootstrap maximum likelihood methods and instead will produce better accuracy if fed trees reconstructed using a best-fit maximum likelihood method. Due to time constraints, I stuck with the IQtree bootstrap trees produced and made one species tree with the preset ASTRAL parameters and one with a bootstrap only performed by resampling gene trees for 1000 replicates.

Bayesian Inference:

For the final investigation of phylogenetic space using this data set, I used the software MrBayes316 to create a gene tree using a Bayesian inference method. Bayesian inference is a popular method for inferring phylogenetic trees as it refines ML by implementing any prior knowledge we have about the data in question. Even if there is no prior knowledge for a particular data set, an uninformative prior can be used. Bayesian inference also gives us an idea of how reliable the results are based on the variance around the “global” maximum the tree search comes to – a feature no other phylogenetic models can replicate. MrBayes is a popular phylogenetic tool as it is highly customizable for its parameter, prior, and model selection – the latter can also be automatically inferred by the algorithm itself. MrBayes also implements a Metropolis-coupled Markov Chain Monte Carlo (MC3) which implements heated chains allowing for a more even traversal of tree space and reduces the likelihood of getting caught in a local maximum. The main drawback to MrBayes and really any Bayesian inference method is they are extremely computationally expensive and can take days to finish. They also assume the chosen prior is accurate and appropriate for the specific data set. I did one run with general parameters: gamma rate variation, 1,000,000 replicates, 3 chains, a burn-in of 10%, and trees sampled every 10 generations. The second run follows the parameters of Wolfe *et al.* 20129: GTR+I+gamma, 10,000,000 replicates, 4 chains, a burn-on of 10%, and trees sampled every 1,000 generations.

**Results:**

Maximum parsimony (MP):

The method of inferring phylogenetic trees using maximum parsimony has been proven statistically inconsistent20. That said, this method can be used as a means of providing Maximum Likelihood models with a more robust tree rather than an entirely random staring tree. The MP tree generated with the nuclear large ribosomal subunit (nrLSU) sequences (fig. 1) is not entirely accurate however most monophyletic groupings are consistent with the other trees generated for the project. The MP generated would be an excellent starting tree for maximum likelihood and Bayesian inference methods

Maximum likelihood (ML):

The three trees generated for the different loci used in the data (fig. 2) highlight the phenomena that historical evolutionary dynamics can be different for genes within the same genome. Still, the more robust monophyletic groupings remain consistent and the same taxa displaying low confidence in their organization (*A. coajizong, A. eijii, A. chiui, A. farinosa, A. elata, A. imazekii*) are the ones who are shuffled into different clades in each of the three trees. The taxa are also known to be found in the more cryptic sections of the genus. The other taxa all fall into consistent monophyletic clades among the trees generated for this project as well as the tree generated by Cui *et al.* 201811. While this gives me a decent level of confidence in the organization of the taxa, I would need to spend more time manipulating parameters and potentially adding more data to consider these trees truly useful. A coalescent approach is applied to the three loci trees to have a better idea of what the true species tree for these taxa may look like.

Coalescent approach:

Using ASTRAL, I took the three trees generated from ML to create a species tree. The first tree I generated (fig 3a) used the preset parameters provided by ASTRAL which is not recommended by the developers. Still, the tree came out with relatively strong support values with only two divergences having less than a 67% support (43% and 33%). The second tree that was generated performed a bootstrap with only gene tree resampling and 1000 replicates. The tree itself infers higher support values suggesting the species tree provided is more robust than the tree generated using preset parameters with none of the support values below 67% and most being above 94% - consistent with the developers’ claim. However, since the model assumes the gene trees are accurate, I am less confident about the support values suggesting the true phylogenetic relationships of the taxa. Again, more data and further model manipulation are required to gain a better understanding of the taxonomic organization of *Amanita*.

Bayesian inference:

Two trees were generated using MrBayes for the nuclear large ribosomal subunit (nrLSU). Using Tacerplot21, I was able to confirm both trees converged and showed high levels of mixing throughout the tree search. The initial tree (fig4a) used more general parameters and still produced a relatively robust phylogeny with mostly good support values however, there were 4 polytomy nodes indicating unresolved phylogenetic conflict. The second tree (fig4b), which used more specific parameters and a larger search of the tree space, was able to resolve two of the polytomies. Support values of the two trees were similar with the initial tree (fig 4a) having slightly higher support however this is likely due to the resolved polytomies having lower support (fig. 4b). Overall, the two runs show the consistency of the MrBayes program and highlight the importance of intelligent parameter selection for phylogenetic inference.

Comparing the second run on MrBayes (fig 4b) with the ML run for the nrLSU (fig. 2a), the two phylogenies were remarkably consistent with the MrBayes run usually giving higher support values. Most monophyletic groups were conserved between the two phylogenies with the main organizational inconsistencies occurring between taxa that are a part of a notoriously cryptic section of the *Amanita* genus (sect. *Vaginatae*).

**Discussion:**

Overall, the given project gives a nice overview of the available tools to use when inferring phylogenies. It shows the robustness of the methods and models that have been created for phylogenetic inference and marks the necessity of time required to thoroughly search the phylogenetic landscape of a given group of organisms. All trees are slightly different from one another showing the differences in the methods implemented but their overall similarity brings a level of confidence needed to have a better understanding of evolutionary history. The tree produced using Bayesian inference and refined parameters is the tree I put the most confidence, indicating to me a need to further investigate the Bayesian inference tools available for phylogenetics.

The next major step for investigating *Amanita* phylogenies is to use what I have learned on genetic data I have gathered from my own specimens. With this data, I can create a robust phylogeny for species of *Amanita* in Wisconsin to gain a better understanding of its local taxonomic organization. Using my own data, I will have better knowledge of the morphological diversity as well giving me the ability to have more confidence in the trees I generate. Further phylogenetic investigation will include closer scrutiny of parameters for different algorithms as well as a focus on generating trees using the software BEAST2 (CITE).

Further work will bring more taxonomic clarity to the genus *Amanita* and allow us to make confident conclusions about the species delineations which can then be used to better understand their distributions. With this knowledge, we can have more confidence informing conservation efforts.

**Figures:**

Table 1: Data from Cui *et al.* 2018

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species name | Location | nrLSU | TEF1α | RPB2 |
| *A.* aff. *brunneofuliginea* | Austria | MH486361 | MH508664 | MH485858 |
| *A. altipes* | Shangri-La, Yunnan, China | JN941158 | KR824810 | JQ031112 |
| *A. aspericeps* | Mount Wuyi, Fujian, China | MH486369 | MH508671 | MH485864 |
| *A. avellaneosquamosa* | Incheon's Central park, South Korea | KJ466482 | KJ481981 | KJ466647 |
| *A. battarrae* | Changdu, Tibet, China | MH486383 | MH508684 | MH485875 |
| *A. brunneofuliginea* | Changbai mountain, Jilin, China | MH486394 | MH508694 | MH485886 |
| *A. caojizong* | Yulong, Yunnan, China | MH486425 | MH508712 | MH485905 |
| *A. chiui* | Yanyuan, Sichuan, China | MH486447 | MH508727 | MH485930 |
| *A. cinereopannosa* | USA | MH486450 | MH508728 | MH485932 |
| *A. citrina* | Germany | MH486457 | MH508733 | MH485937 |
| *A. citrinoannulata* | Lianyungang, Jiangsu, China | MH486458 | MH508734 | MH485938 |
| *A. citrinoindusiata* | Shangri-La, Yunnan, China | MH486469 | MH508745 | MH485948 |
| *A. eijii* | Ninglang, Yunnan, China | MH486484 | MH508761 | MH485963 |
| *A. elata* | Xishuangbanna, Yunnan, China | MH486486 | MH508763 | MH485965 |
| *A. farinosa* | Baoshan, Yunnan, China | JN941154 | MH508772 | JQ031110 |
| *A. flavipes* | Shangri-La, Yunnan, China | MH486507 | MH508783 | MH485979 |
| *A. fulva* | France | MH486554 | MH508825 | MH486021 |
| *A. imazekii* | Tomakomai, Hokkaido, Japan | MH486591 | MH508859 | MH486052 |
| *A. ocreata* | CA, USA | MH486688 | KJ481947 | KJ466607 |
| *A. pantherina* | Czech Republic | KR824782 | KR824825 | KR824789 |
| *A. strobiliformis* | Germany | MH486895 | MH509117 | MH486298 |
| *A. suballiacea* | PA, USA | KJ466484 | KJ481940 | KJ466600 |
| *A. virosa* | Juva, Filand | JX998058 | JX998007 | KJ466664 |
| *L. delicata* | Austria | KT833807 | KT833835 | KT833822 |

Figure 1: Maximum parsimony phylogeny

A diagram of a number of trees

AI-generated content may be incorrect.

Figure 2: Maximum likelihood gene phylogenies

A black background with a black square

AI-generated content may be incorrect.Figure 2a (nrLSU loci)

A black background with a black square

AI-generated content may be incorrect.Figure 2b (RPB2 loci)

A black background with a black square

AI-generated content may be incorrect.Figure 2c (TEF1-α loci)

Figure 3: Coalescent phylogenies

Figure 3a Figure 3b

A black background with a black square

AI-generated content may be incorrect.A black background with red numbers

AI-generated content may be incorrect.

(3a) Phylogeny generated with preset astral parameters.

(3b) Phylogeny generated when bootstrapping is only applied to gene trees with 1000 replicates.

Figure 4: MrBayes phylogenies

Figure 4a Figure 4b

A black background with a black square

AI-generated content may be incorrect.**A black background with a black square

AI-generated content may be incorrect.**

(4a) phylogeny generated with a gamma rate variation, 1,000,000 replicates, 3 chains, a burn-in of 10%, and trees sampled every 10 generations.

(4b) Phylogeny generated with GTR+I+gamma, 10,000,000 replicates, 4 chains, a burn-on of 10%, and trees sampled every 1,000 generations.

**References:**

1. Stuart, S. N. *et al.* Status and Trends of Amphibian Declines and Extinctions Worldwide. *Science* **306**, 1783–1786 (2004).

2. Boonman, C. C. F. *et al.* More than 17,000 tree species are at risk from rapid global change. *Nat Commun* **15**, 166 (2024).

3. Adams, A. M. *et al.* The state of the bats in North America. *Annals of the New York Academy of Sciences* **1541**, 115–128 (2024).

4. 15/4. Kunming-Montreal Global Biodiversity Framework. in (United Nations Environment Programme, Montreal, Canada, 2022).

5. Taylor, J. W., Turner, E., Townsend, J. P., Dettman, J. R. & Jacobson, D. Eukaryotic microbes, species recognition and the geographic limits of species: examples from the kingdom Fungi. *Philos Trans R Soc Lond B Biol Sci* **361**, 1947–1963 (2006).

6. Niskanen, T. *et al.* Pushing the Frontiers of Biodiversity Research: Unveiling the Global Diversity, Distribution, and Conservation of Fungi. *Annual Review of Environment and Resources* **48**, 149–176 (2023).

7. Pegler, D. N. Useful Fungi of the World: Caesar’s mushroom and the Christmas mushroom. *Mycologist* **16**, 140–141 (2002).

8. Vetter, J. Amanitins: The Most Poisonous Molecules of the Fungal World. *Molecules* **28**, 5932 (2023).

9. Wolfe, B. E., Tulloss, R. E. & Pringle, A. The Irreversible Loss of a Decomposition Pathway Marks the Single Origin of an Ectomycorrhizal Symbiosis. *PLOS ONE* **7**, e39597 (2012).

10. Geml, J., Tulloss, R. E., Laursen, G. A., Sazanova, N. A. & Taylor, D. L. Evidence for strong inter- and intracontinental phylogeographic structure in *Amanita muscaria*, a wind-dispersed ectomycorrhizal basidiomycete. *Molecular Phylogenetics and Evolution* **48**, 694–701 (2008).

11. Cui, Y.-Y., Cai, Q., Tang, L.-P., Liu, J.-W. & Yang, Z. L. The family Amanitaceae: molecular phylogeny, higher-rank taxonomy and the species in China. *Fungal Diversity* **91**, 5–230 (2018).

12. J, K. Entrez Direct: E-utilities on the Unix Command Line. National Center for Biotechnology Information (2013).

13. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).

14. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

15. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

16. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees.

17. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548 (2014).

18. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).

19. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

20. Felsenstein, J. CASES IN WHICH PARSIMONY OR COMPATIBILITY METHODS WILL BE POSITIVELY MISLEADING.

21. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology* **67**, 901–904 (2018).