**Robert Petit**

**IBS594 – Phylogenetics Exercise 1**

**Alignment and Maximum Likelihood Phylogenetic Inference**

Data from:

Lutzoni FM. 1997. Phylogeny of Lichen- and Non-Lichen-Forming Omphalinoid Mushrooms and the Utility of Testing for Combinability among Multiple Data Sets. Syst Biol, 46:373-406.

Before you begin, you may want to download ClustalX:

<http://www.clustal.org/clustal2/>

You can use Clustal omega or Clustal W webservers if you prefer.

You will also need to download FigTree:

<http://tree.bio.ed.ac.uk/software/figtree/>

Your goal is to generate and use multiple sequence alignments to infer phylogenies of omphalinoid mushrooms. Start by opening the ClustalX application. This is a friendly user interface for the wildly popular clustal algorithm for multiple sequence alignment, which has a command line interface.

1. Load the sequences in the lutzoni1997.fasta file. These are sequences of the internal transcribed spacer 2 (ITS2) of the ribosomal RNA of mushrooms. It is located between the 5.8S and the 28S ribosomal RNAs. The spacers are spliced before the ribosomal RNAs fold to make up the ribosome in eukaryotic cells. What do the lengths of the different ITS2 sequences in this data set indicate regarding the homology of different nucleotides?

Length Distribution

length count

63 3

64 1

65 1

66 1

68 1

69 2

70 2

71 6

72 11

73 1

74 1

At first glance, the distribution indicates a large portion of our sequences (length=72) are probably fairly homologous to one another. But as the sequence lengths become more different (63 vs 72) we can expect these to have an older common ancestor and thus be less homologous in nucleotide identity.

2. Go to the Alignment header, and click on “Set all parameters to default”. Under the same header click on alignment parameters, and click on “Reset all gaps before alignments”. Now, open “Multiple alignment parameters” in this last header. Focus on the gap opening and gap extension penalties. These are expressed as costs from 0-100. Using the default parameters for these penalties, click on Alignment, Do complete alignment. What is the length of the alignment? What do the asterisks indicate? How are those sites relevant to the alignment? Use the File header to save your alignment in fasta format.

Using the default settings of clustalw, there was an alignment length of 79 bp. I personally used the command line clustalw2, and the output alignments did not contain asterisks. It did on the other hand contain dashes which are symbolic of gaps in the alignment. In these regions we can expect mutations, for example InDels, to have occurred in the history of the Omphalinoid mushrooms.

3. Sensitivity analysis of gap penalties. Change the global gap opening and extension penalties first to 5 and 1, and then to 50 and 15. Save the resulting alignments in fasta format. What is the relationship between alignment length and gap opening penalty? Which one of the 3 alignments is “right”? Explain.

Due to how alignment scores are calculated, the greater the gap opening and extension penalty the smaller we should expect the alignment length to be. With a high enough penalty, opening and extending becomes too costly and is opted against by the alignment algorithm.

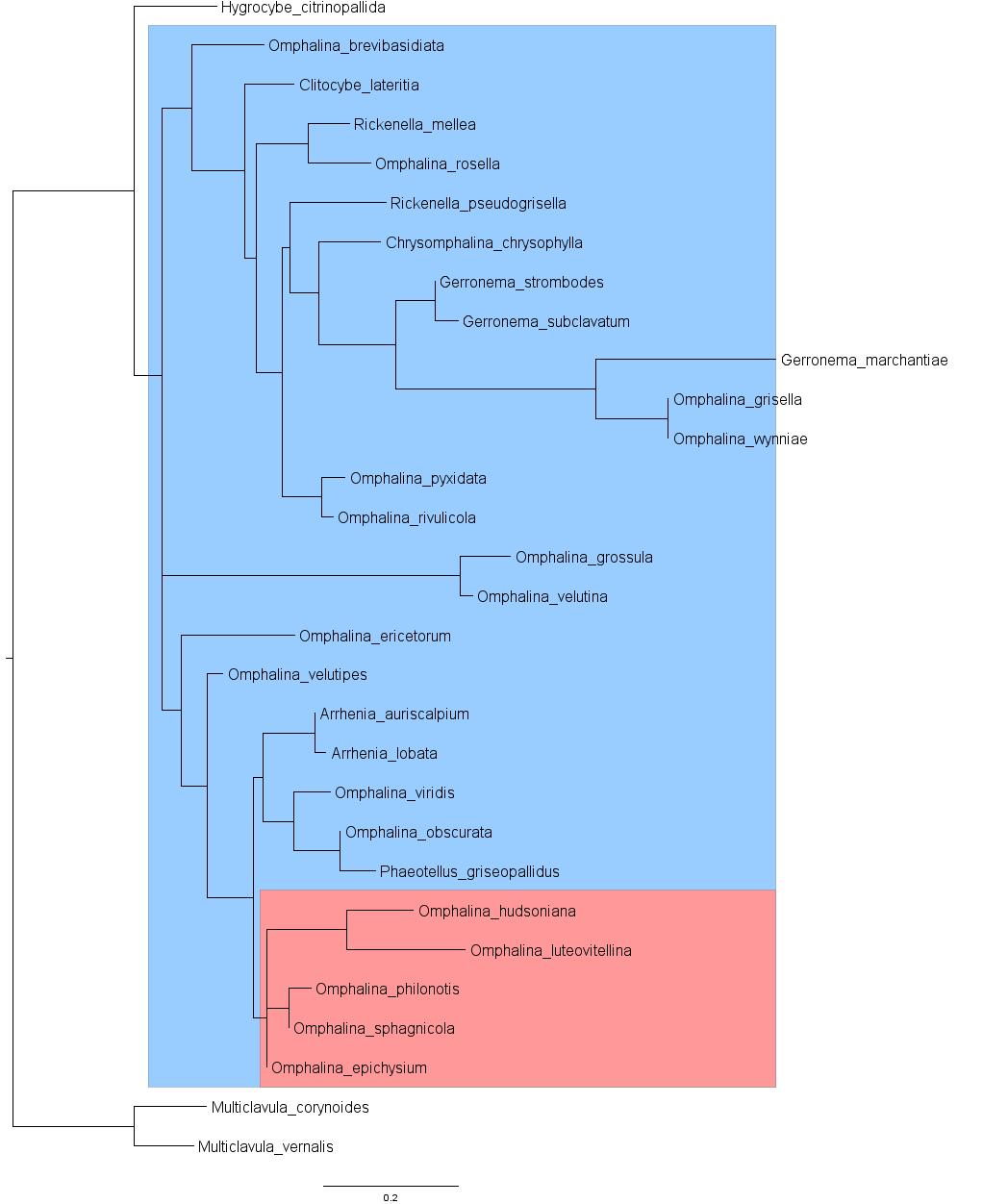
I personally am terrible at eyeballing alignments and determining which the right one is. I believe there should be a systematic protocol for choosing an alignment over another. But viewing the alignments as a tree makes it much easier.

After looking at the trees and the alignments for an hour or so, I personally would chose the alignment based on the default settings (gap open 15, gap extension 6.66). It is worth noting that all three alignments are plagued by multiple polytomies. So at the end of the day, better resolution is likely required to disentangle the evolutionary history of the Omphalinoid mushrooms.

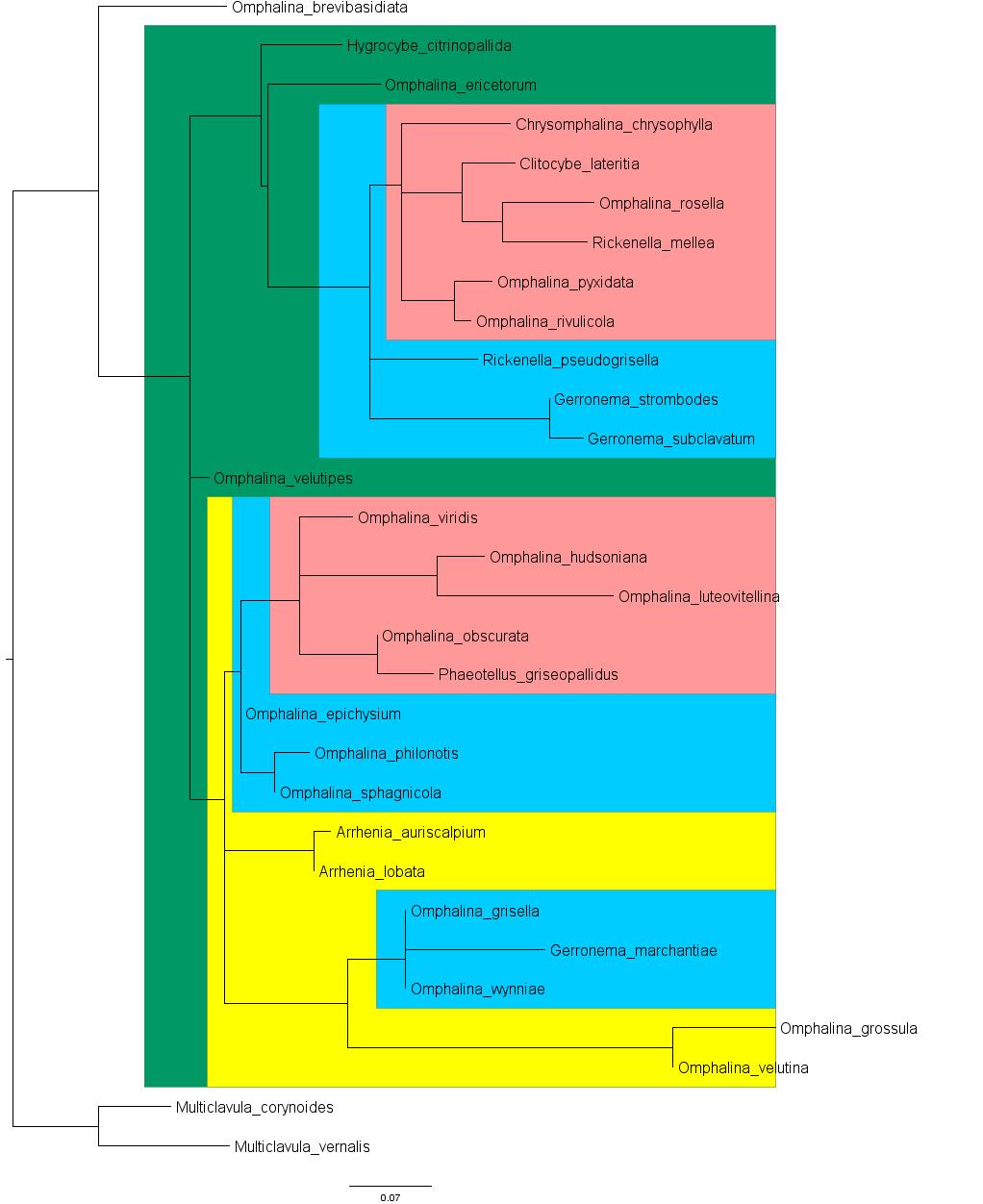
4. Inferring ML phylogenies.

|  |  |  |  |
| --- | --- | --- | --- |
| Gap:open/extension penalty | Length of alignment | Tree length (parsimony) | Tree length (ML) |
| Default (15/6.66) | 79 | 199 | 4.24881 |
| 5/1 | 84 | 165 | 2.98361 |
| 50/15 | 75 | 334 | 7.30270 |

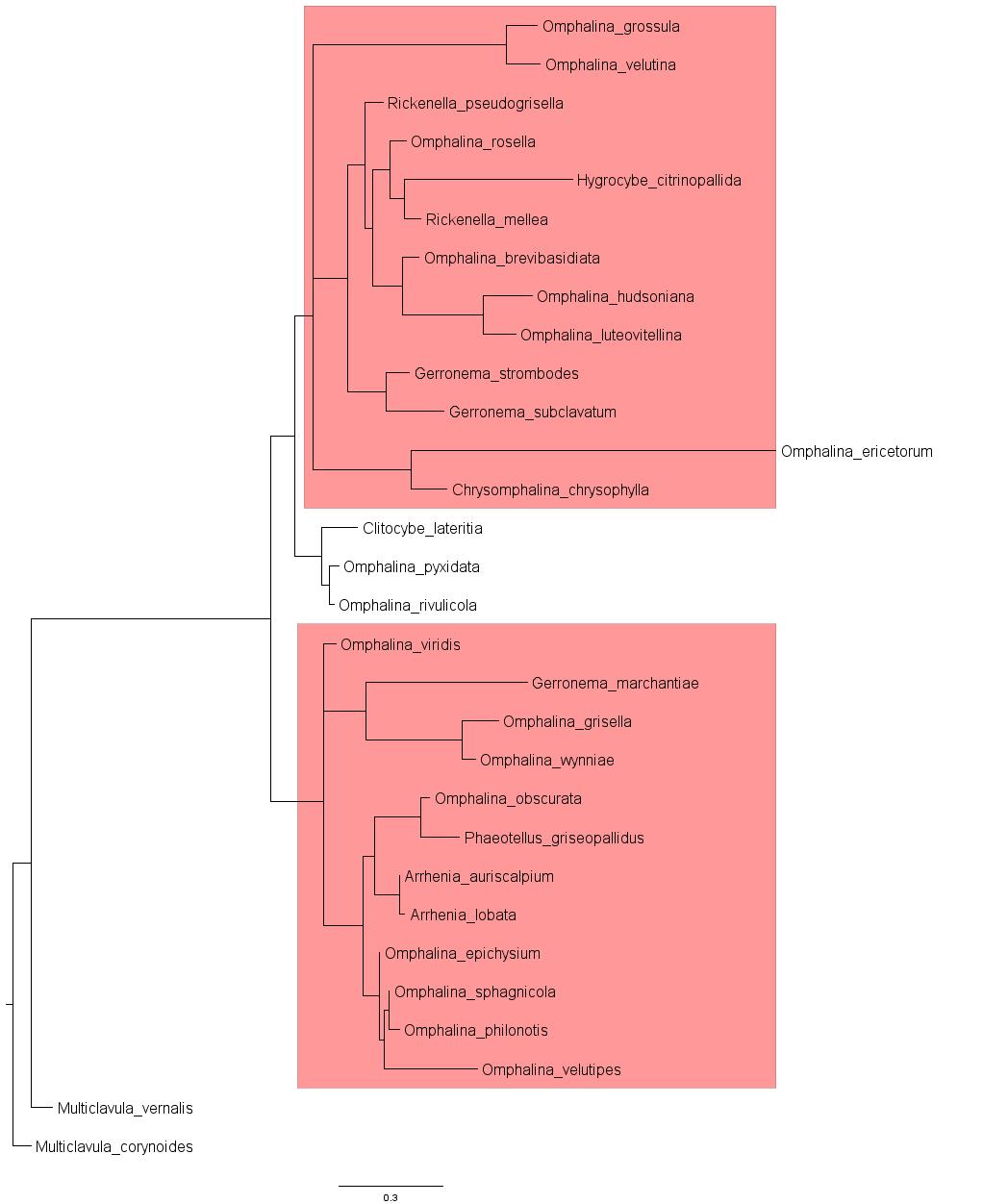
Default Settings (Gap Open 15, Extension 6.66, Polytomy occurs twice)



Gap Open 5, Extension 1 (Polytomy occurs 7 times)



Gap Open 50, Extension 15 (Polytomy occurs twice)



5. Why do phylogenies resulting from alignments using different gap opening and extension penalties differ? How does inferred change in the phylogeny relate to the penalties? Explain. What would you recommend if a colleague asked you what to do given the different results from different parameters? Assume collecting data from another gene is not an option.

I somewhat alluded to this earlier, but the alignments differ because of how the gap penalties affect the alignment score. When aligning bases in two or more sequences the alignment algorithm finds the alignment that produces the best score. When we give a high gap open penalty, more often the algorithm will choose to mismatch versus opening a gap in order to achieve the best alignment score. The opposite is true as well, when given a low gap open penalty the algorithm will choose to open more gaps because it will like produce the best alignment score.

When we look at the three trees, the amount of polytomies is a clear example of how the gap penalties affect tree layouts. When we relaxed the gap penalty (5/1) there are a total of 7 polytomies within the tree. I don’t know the exact number, but with 7 polytomies there are numerous possible resolutions. Now as we make our penalties more stringent (default settings) there are only two polytomies in the tree, so it’s a better result. Then even more stringent (50/15) there are only two polytomies as well. The topology is slightly different between the two (default & 50/15), so they do tell two different stories.

My recommendation to a colleague, given the current results don’t produce a single ‘right’ answer, I would suggest trying a few other methods for creating trees. Maybe even try multiple programs for maximum-likelihood (RAxML comes to mind) because it is well known that they produce different results (for better or worse!). With the small data set it might even be worth trying a Bayesian approach such as BEAST or MrBayes. With these multiple approaches, I believe you can either find an optimal tree or be confident when you say you’ve exhausted all possibilities and just aren’t going to be able to produce a ‘right’ result given the limited data.

You can reproduce my results with the following Git repository: <https://github.com/rpetit3/ibs594-phylogenetics>

**Robert Petit**

**IBS594 – Phylogenetics Exercise 2**

**Comparing Multiple Phylogenetic Algorithms**

1) Write a Methods and Results Section for your work. Include three figures (one for each tree). Make sure you include a caption for each figure, following conventions in published work on what information should be included in the caption. Make sure to include details of what programs you used, what model of evolution you used, how many generations you ran for your Bayesian analysis, etc.

**Methods**

The data used in these analyses were ITS DNA sequences from 12 species. These species include (GenBank accession in parenthesis) *Escovopsioides nivea* CBS 135749 (JQ815078), *Ascomycota sp*. RS054 (EU082786), *Escovopsis lentecrescens* CBS 135750 (JQ815079), *Escovopsis aspergilloides* CBS 423.93 (KF293287), *Escovopsis moelleri* CBS 135748 (JQ815077), *Escovopsis sp.* TF2CWG (FJ948131), *Escovopsis microspora* CBS 135751 (JQ815076), *Escovopsis weberi* ATCC 64542 (KF293285), *Escovopsis weberi* CBS 810.71 (KF293286), *Cladobotryum asterophorum* (EU340835), *Trichoderma viride* (EF568085), and an unknown Escovopsis/Escovopsioides species.

Using the default settings (GAPOPEN=15, GAPEXT=6.66) the ITS sequences were aligned using ClustalW (v2.1). The alignment from ClustalW was then converted from FASTA to Phylip using the shell script convertFasta2Phylip.sh available through RAxML. We predicted a model of evolution for our sequences using jModelTest (v2.1.6). jModelTest was run with the following parameters: -g 4 -i -f -AIC –a. The model with the top score based on Akaike Information Criterion (AIC ) was chosen for downstream maximum-likelihood and Bayesian analysis.

A parsimony tree and maximum likelihood tree was constructed using RAxML (rev. 051fd11). First a parsimony tree was constructed using the default settings of RAxML. We then constructed a maximum-likelihood tree using the parsimony tree as the starting tree and a GTRGAMMAI as our model of selection. The maximum-likelihood tree was bootstrapped 200 times before selecting the best tree. A bayesian tree was constructed using MrBayes (v3.2.2). MrBayes was run for 1 million generations with a burn-in of 25% using a GTRGAMMA model of selection. FigTree (v1.4.2) was used to visualize and annotate each of the three trees. Each of the trees were rooted between *Cladobotryum asterophorum* and *Trichoderma viride*. Also the known *Escovpsis* species were highlighted with green and known  *Escovopsioides*species were highlighted with blue. A Git repository of these data and methods is available at <https://github.com/rpetit3/ibs594-phylogenetics>.

**Results**

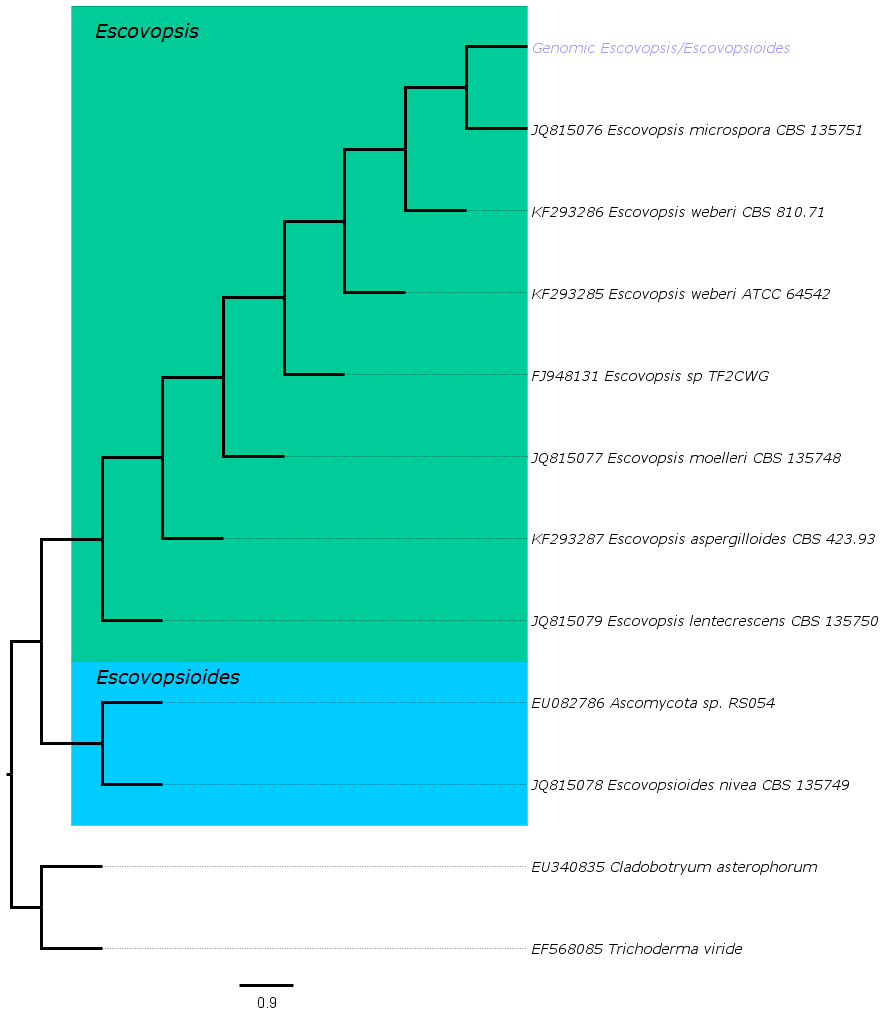
Since the splitting of the *Escovopsis* genus in to *Escovopsis* and *Escovopioides* there are still a species that must be placed into the proper genus. We set out to take an ITS sequence from the former *Escovopsis* genus and determine the proper genus (*Escovopsis* or *Escovopioides*). We compared our unknown ITS sequence against 11 other species of both *Escovopsis* and *Escovopioide*, also a few more distant ancestors.

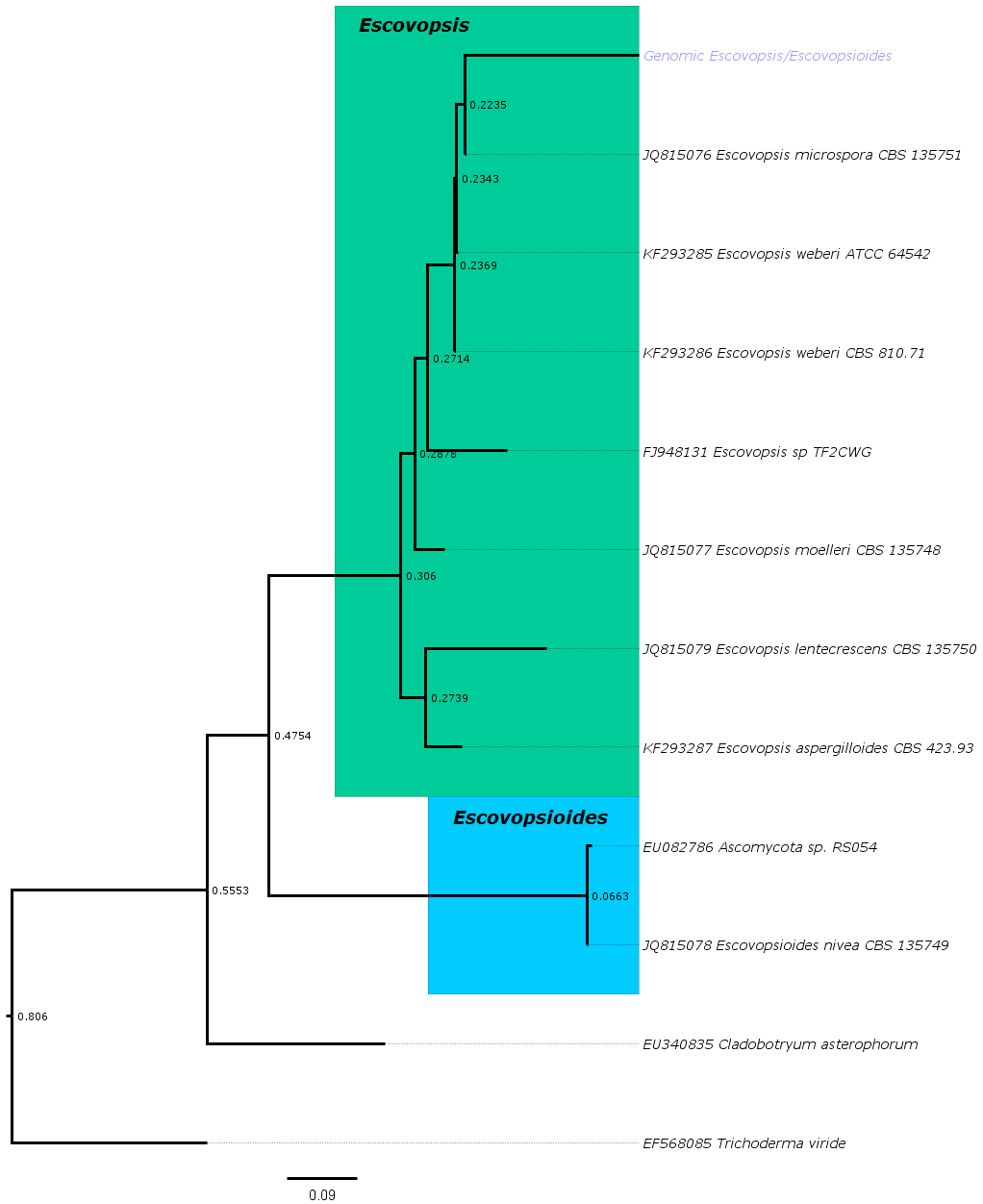
We used three phylogenetic methods to compare our sequences. Based on parsimony (Fig. 1), there is evidence that our unknown sample is likely of the genus *Escovopsis*. This trend continues with the maximum-likelihood tree (Fig. 2) and the Bayesian tree (Fig. 3) as well. In each of the three trees there is a clear monophyletic clade between genus *Escovopsis* and genus *Escovopioides*. Each method also produces evidence that our unknown sample is closely related to *Escovopsis microspora*.

It should be pointed out that our bootstrap values for both the maximum-likelihood and Bayesian trees are rather low. Internal nodes in both of these trees are less than 0.3, much less than the general rule of greater than 0.7 for strong confidence. Although these values do not necessarily mean the trees are incorrect, it does however affect the confidence in them. Most likely this is due to the nature of our data set. We have based these results on only 12 ITS sequences. Extending this analysis to include more genomic regions or accompanying morphology data could improve our confidence in the results.

2) In looking at your sequence alignment, how confident are you in each of your characters being homologous? Is this a problem? If so, how would you suggest someone using these data for a paper proceed?

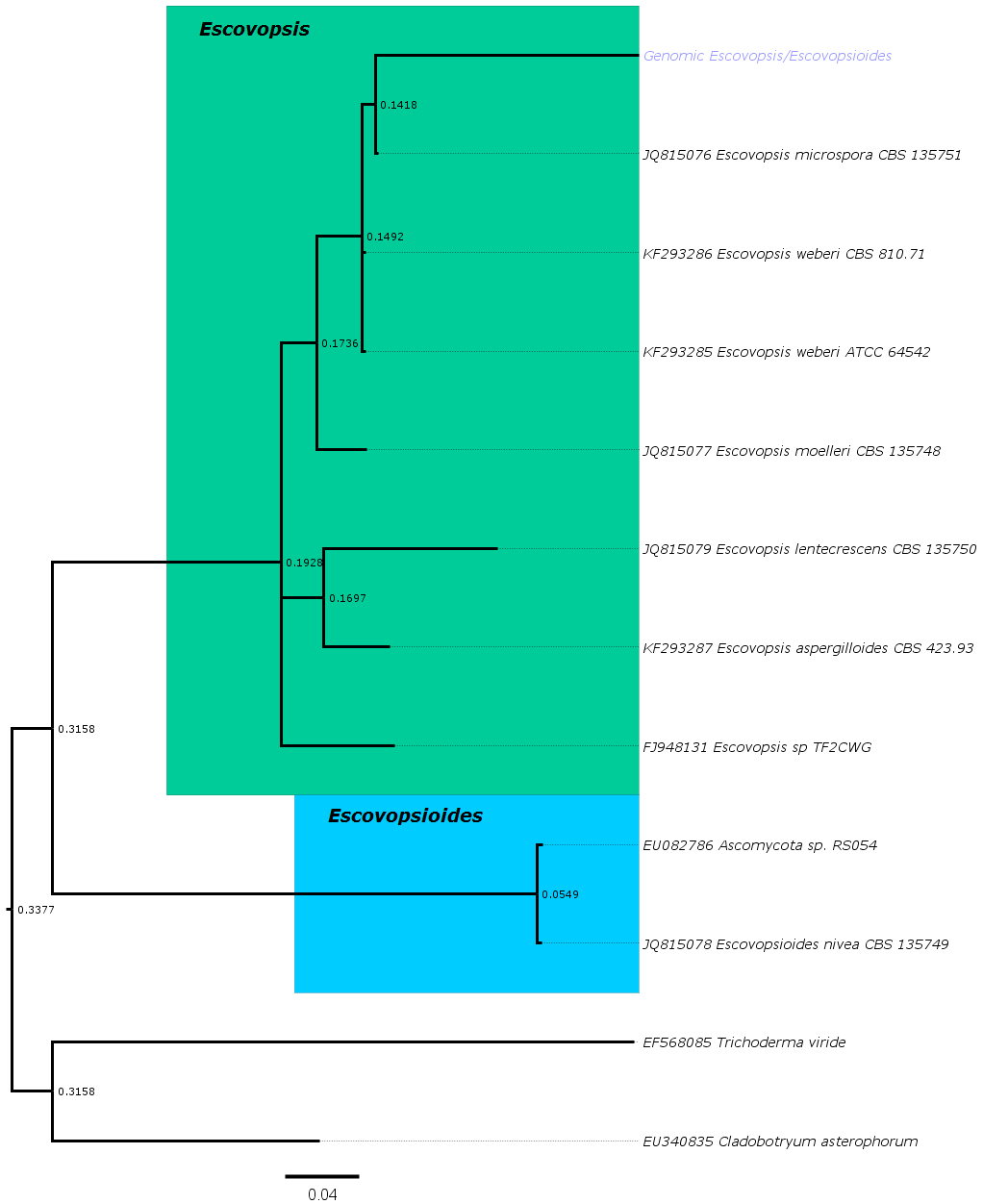
Looking at the alignments there are definitely some more ‘gappy’ (for lack of a better term) regions. There is a central area that is conserved among the sequences. I am confident in the alignments, but I would not base a paper solely on the analysis of these alignments. Surely you could go on a fishing expedition trying to achieve the best alignment by playing with the parameters and even trying different programs. You would probably even get better trees to base the results on. But, at the end of the day, what these results need is supporting evidence. This could be looking at morphological data: What morphological characteristics make our unknown sample more like *Escovopsis* or *Escovopsioide*? Are there morphological characteristics shared between our unknown sample and *Escovopsis microspore*? If the goal was to keep this study sequence driven, extend it to more genomic regions to gain a better resolution. I don’t know much about fungi, but there must be more than ITS that is conserved between the species. I imagine by extending the sequencing to more genomic regions the phylogenies will give much better support.

**Figure 1: Optimal parsimony tree obtained from analysis of ITS DNA sequence data.**

The parsimony tree groups our unknown sample (written in blue) with the genus *Escovopsis* (highlighted in green) and is most closely related to *Escovopsis microspore* CBS 135751*.* 

**Figure 2: Optimal maximum-likelihood obtained from analysis of ITS DNA sequence data.**

The maximum-likelihood tree groups our unknown sample (written in blue) with the genus *Escovopsis* (highlighted in green) and is most closely related to *Escovopsis microspore* CBS 135751. The bootstrap values, after 200 iterations, are displayed at each internal node.

**Figure 3:** **Optimal Bayesian obtained from analysis of ITS DNA sequence data.**

The bayesian tree groups our unknown sample (written in blue) with the genus *Escovopsis* (highlighted in green) and is most closely related to *Escovopsis microspore* CBS 135751. The bootstrap values, after 1 million generations, are displayed at each internal node.