Janet K. Mansaray

Computational Phylogenetics

Gene tree-species tree Reconstruction using the Multispecies Coalescent of *Centropogon subgenus Centropogon*

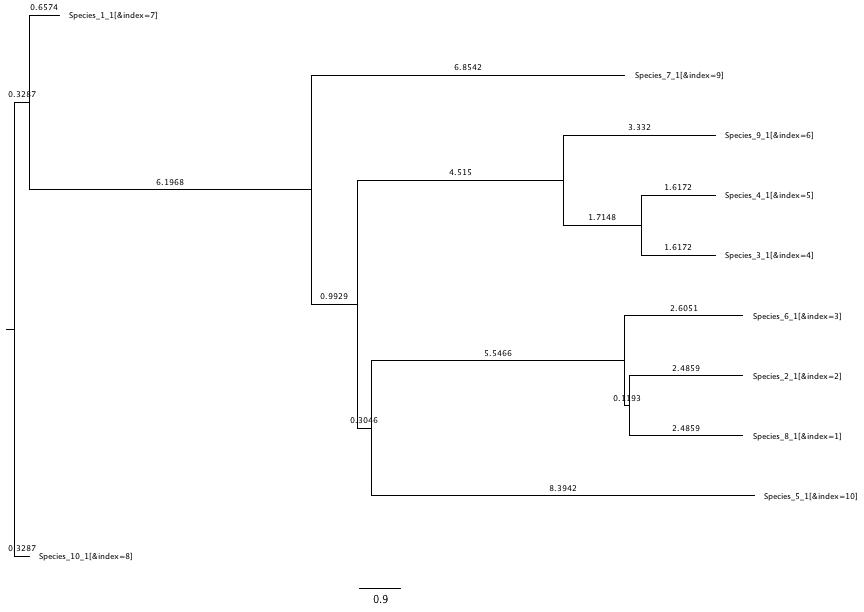
Though subgenus Centropogon also known as the eucentropogonids appears to be monophyletic, the species-level relationships are not clear. For example, *Centropogon granulosus* and *C. solanifolius* are not monophyletic. *Centropogon cornutus* however appears to be a monophyletic species that is well-supported as sister to the other species within the clade (Stein, 1986; Lagomarsino, 2014). Furthermore, within species morphological variation has been observed and the underlying causes of this are not well understood. Therefore, identifying what it means to be a species in this group is essential. In conjunction with an improved phylogeny, a gene tree-species tree reconstruction using a multispecies coalescent could potentially help elucidate ancestral relationships as well as relationships amongst extant species. Thus, for this project, I wanted to ask if we keep the effective population size for each branch of the tree the same, but account for varying transitions/transversions within the gene trees how will this affect the topology of the species tree? In addition, there are non-monophyletic species and high genetic diversity within the clade indicating varying mutations throughout time. The genes chosen for this project were exon1005-7, exon1013-3, exon1025-1, and exon1031-1. The exons have long sequence alignments with largest being 528 characters. The four genes were chosen and represent neutral loci. Overall, the exons represent segments of DNA found within all *Centropogon* species as well as sister taxa. For this project, nine *Centropogon* species were selected at random. Two of the species selected, *C. granulosus* and *C. solanifolius,* are phylogenetically problematic and appear independently outside the subgenus Centropogon as well as within. *Centropogon urubambae, C. congestus, C. alectrolophos, and C. pulcher* are known to be nested within the subgenus Centropogon consistently. *Centropogon baezanus* is another subgeneric clade sister subgenus Centropogon. The tenth species selected is the outgroup, *Siphocampylus scandens*. This was chosen as an outgroup species because it is it in the clade sister to the genus *Centropogon*. *Siphocampylus* is morphologically differentiated from *Centropogon* based on fruit type (*Centropogon* produces berries, *Siphocampylus* produces capsules). Firstly, I simulated a species tree and gene trees in RevBayes using code found using the ?dnMultispeciesCoalscent function. These simulated trees served as a starting point for the project and allowed me to have useful variables to begin with. I could not use the actual species tree for this project because the only one done to date includes other taxa from other subclades and has about 126 tips. I tried to use this tree first but the run time took too long. In addition, I was not using all the species from the actual species tree for this project there would have been discrepancies between my species tree and character matrices. From the simulated tree and using published research about the age of this clade, I was able to make an informed estimate of the root age. I used the RevBayes gene-tree species-tree tutorial as a guide. Firstly, I read in all the gene sequences for the different exons by identifying the exon names and matrices that contained sequence data for each species. I then read in a species tree for *Centropogon* (included sister taxa). I estimated the speciation and extinction rates to set a prior for the diversification and turnover rate. Because the speciation and extinction rates are unknown for this group, I used the default priors which seemed appropriate. I used the birth-death process distribution to construct the topology for the species tree. Then read in each simulated gene tree. An ultrametric tree was created to give the species tree a useful starting values for the topology and rootage. The root age is estimated to be 8.27292. The effective population size e it is meant to represent the number of individuals in a population contributing to the next generation. Since the data set is small, I set the population to ten. In addition, I hypothesized that keeping the number of individuals contributing to the next generation the same across all species would result in more defined phylogenetic relationships. Then the species maps with taxon names were read in to link the gene trees and species trees and used the dnMultiSpeciesCoalescent function to create a gene tree. The HKY85 substition model seemed like the best fit due to what we know about this group. It appears to have experienced rapid speciation around 5-12 million years ago and has evolved morphological traits to adapt to a specialized sicklebill hummingbirds, *Eutoxeres aquila and E. condamine* (Stein, 1986).In addition, there are non-monophyletic species and high genetic diversity within the clade indicating varying mutations throughout time. Then a standard MCMC was run as well as a tree trace to produce a posterior species tree.

**Results**

The gene trees produced similar trees to one another with slight variations after they were rooted with *S. scandens* (see supplemental data). The discrepancies between the trees can possibly be attributed to incomplete sampling. However, the final species tree supported what is known about thesubgenus *Centropogon*. When rooted with *S. scandens*, *C. bazeanus* is monophyletic and outside the subgenus clade as expected. The branch length leading to other *Centropogon* species is 6. 1968. This could potentially due to incomplete sampling and therefore it appears to have undergone more changes to reach the divergence of the other species. The species tree reveals the monophyletic subgenus *Centropogon*. Species\_2\_1-Species\_9\_1, as labeled on the tips, represents the clade (Figure1, Table 1). *Centropogon cornutus* (Species\_5\_1), is its own subclade within the subgenus *Centropogon* and its monophyly is supported in the species tree (Table 1). The branch length is long but this is a common pattern in C. cornutus and other researchers have found the similar results (Lagomarsino, 2014, Antonelli, 2008). This could be due to cryptic species hidden within the species or poor data. The three species (C*entropogon urubambae, C. congestus, and C. alectrolophos*) that consistently nest within the eucentropogonid clade are seen nested here the same way. However, *C. pulcher* appears to be more closely related to *C. granulosus* and *C. solanifolius.* This is not surprising because taxonomically *C. pulcher, C. granulosus* and *C. solanifolius* are group into the same subclade despite the latter two not being monophyletic in large phylogenies with more samples. Interestingly, *C. granulosus* and *C. solanifolius* are sister to another and form their own smaller clade. Morphologically, these two species are very similar with bright red-orange flowers and similar corolla structure (Stein, 1986). They do not form this sort of uniformity in phylogenies with larger samples and often are grouped near each other and throughout the phylogeny. Again, this result could be due to the small sample size and limited amount of genes used. Using Tracer, I analyzed the log data of the MCMC and the trace of the posterior likelihood did not converge fully or create the “fuzzy caterpillar”. An example of my trace file of the posterior likelihood for run one is provided below (Figure 2). The trace for the first run and second run did show more convergence than the trace of the average of the two. I think this can be attributed to the priors I used, but changing the priors did not seem to change the trace files. However, I did not drastically change the priors (only slightly due to time), but it is something in the future I will take note to do. Overall, I think there was not enough samples to conclusively understand ancestral relationships but it was encouraging to see that my results were consistent with published results. Making the effective population size the same and accounting for the rates in transversions and transitions seemed to be a good fit for this data. This is information I can take moving forward as I work to produce an improved phylogeny for the subgenus and for future gene-tree, species-tree reconstructions with larger data sets.

\*Supplemental Data included in MansarayFinalProject folder

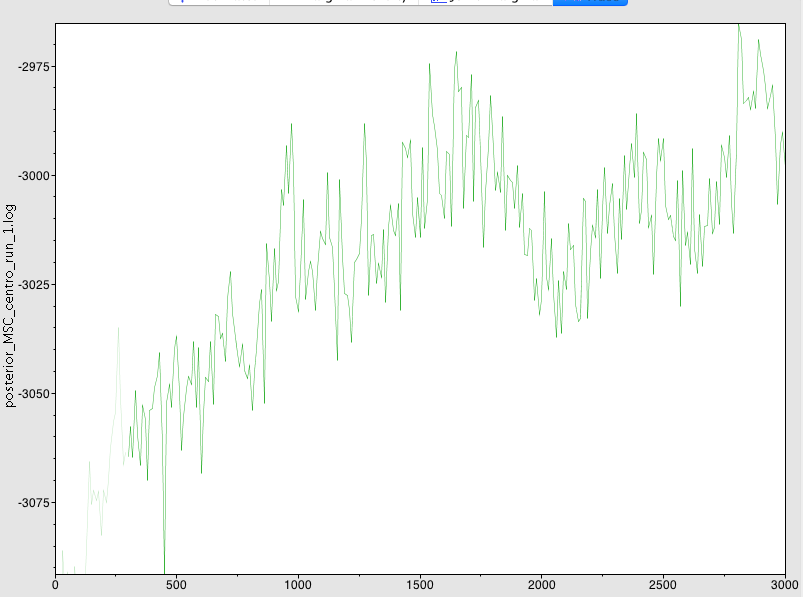
**Figure 1. The final species tree created with the gene trees (clear version in data files).**



**Table 1. The species names used for the project with their respective true species names.**

|  |  |
| --- | --- |
| Species Names on Species Trees | Corresponding Species Name |
| Species\_1\_1 | *C. baezanus* |
| Species\_2\_1 | *C. congestus* |
| Species\_3\_1 | *C. solanifolius* |
| Species\_4\_1 | *C. granulosus,* |
| Species\_5\_1 | *C. cornutus* |
| Species\_6\_1 | *C. urubambae* |
| Species\_7\_1 | *Centropogon baezanus* |
| Species\_8\_1 | *C. alectrolophos* |
| Species\_9\_1 | *C. pulcher* |
| Species\_10\_1 (outgroup) | *Siphocampylus scandens* |

**Figure 3. MCMC Trace file of Posterior Likelihood for Run 1.**



References

Antonelli, A. (2008). Higher level phylogeny and evolutionary trends in Campanulaceae subfam. Lobelioideae e: Molecular signal overshadows morphology. *Molecular Phylogenetics and Evolution*, *46*(1), 1-18.

Lagomarsino, L. P., Antonelli, A., Muchhala, N., Timmermann, A., Mathews, S., & Davis, C. C. (2014). Phylogeny, classification, and fruit evolution of the species‐rich Neotropical bellflowers (Campanulaceae: Lobelioideae e). American journal of botany, 101(12), 2097-2112.

Stein, B. A. (1986). Patterns of diversification in Andean Lobelioideae e: Centropogon subgenus Centropogon. *Amer. J. Bot*, *73*, 787-8.