Merritt Polomsky
EEOB 563
April 15, 2020
Final Project Paper – Draft #1

Final Project Paper – Draft #1

Introduction

The mitochondrial genomes of non-bilaterian animals display intriguing levels of genetic diversity, particularly compared to the high levels of genetic conservation seen in the mitochondrial genomes of bilaterian animals [1]. This diversity extends to gene content, genome organization, mRNA editing, and other factors, including evolution rate. Phylum Porifera, or sea sponges, shows much of this remarkable genetic diversity [1]. Some mitochondrial genomes of sponges have already been published and their diversity documented; this include those of *Amphimedon queenslandica* and *Xestospongia muta* [2]. The Lavrov Lab has generated additional mitochondrial genomes from sponges within the same order as the aforementioned species, the mitochondrial genomes of *Niphates erecta* and *Niphates digitalis*. These two sponges fall within the same genus, but unpublished analysis of gene content and organization show interesting differences. In particular, the mitochondrial genome of *N. erecta* show multiple insertions within protein-coding genes, when compared to *N. digitalis*. Of the 80 insertions found in *N. erecta*, they vary in size from only a few bases to whole stretches over 600 bases long. Many of these insertions are out-of-frame, but it is currently unknown if they become part of the transcript and are subsequently expressed in proteins. Previous research has also shown that *N. erecta* might have a higher rate of evolution, at least compared to *Amphimedon queenslandica* [3], but this relationship has not been shown in relation to another species from Niphates, due to a lack of data.

Merritt Polomsky
EEOB 563
April 15, 2020
Final Project Paper – Draft #1

This project aims to re-evaluate the phylogenetic position of *Niphates erecta* in relation to the new *Niphates digitalis* genome, with additional mitochondrial genomes from *Amphimedon queenslandica*, *Amphimedon compressa*, and *Xestospongia muta* as reference. The goal is to determine if *N. erecta*'s diversity is on a genus level or species specific. In addition, I hope to determine if some of the wild evolution rates seen in previous research is an artifact of these novel insertions, and evaluate whether the insertions show any type of phylogenetic relationship to each other.

Methods

In addition to the *Niphates erecta* and *Niphates digitalis* mitochondrial genomes, three additional species were selected – *Amphimedon queenslandica, Amphimedon compressa,* and *Xestospongia muta*. All of these genomes are well studied and characterized.

From these genomes, five common genes were chosen. These were *rnl*, *rns*, *cob*, *cox1*, *cox2*, and *cox3*. All six genes tend to be highly conserved, due to their necessary functions as either ribosomal subunits (*rnl* and *rns*) or as part of metabolic respiration (*cob*, *cox1*, *cox2*, and *cox3*). The sequences for these genes were extracted, and aligned by gene using the auto function in MAFFTv7 [4]. The goal of aligning by gene was to account for the difference in length of the various sequences and to discourage faulty alignment due to differing sequence lengths. Two different alignments were created – one where *N. erecta's* insertions were kept, and one where the insertions were discarded. It was decided that sequences would not be translated into amino acids for alignment due to the uncertainty around if *N. erecta's* insertions are translated to proteins. Before concatenation, all alignments were manually checked in

SeaView[5], and once concatenated, manually checked again for equal length. Alignments were concatenated by species for final analysis.

Phylogenetic analysis was done using RAxML-NG[6] on the Iowa State University High Performance Clusters. The alignments were checked, parsed, and then the maximum likelihood trees were calculated, with data partitioned by gene and with independent model parameter estimates (GTR+G+FO). Analysis was also run using no partitioning and a basic GTR+G model. From there, a 1000 replicate bootstrap analysis was run on the data, and then applied to the tree. For both the alignment with insertions and the alignment without insertions, bootstrap values converged about 50 replicates.

Results

RAxML-NG analysis returned maximum likelihood values and tree topology for the best tree found. Trees returned for the insertion alignment analysis is found in Figure 1. For the insertion alignment with gene partitioning, the best tree has maximum likelihood score of -35755.04, with an AIC score of 71702.08. Performing 60 tree searches for the insertion alignment (30 random and 30 parsimony-based starting trees) did not impact topology (see Figure 1), and resulted in a final maximum likelihood score of -35754.99 and an AIC score of 71701.99. The analysis was also running using no partitioning and a basic model GTR+G model (20 tree search), and returned a final maximum likelihood score of -35920.53 and an AIC score of 71873. With the simpler model, there were no topology differences. A bootstrap analysis of the insertion tree returned bootstrap values of 100 for all nodes (see Figure 1).

As with the insertion alignment, analysis on sequences without *Niphates erecta* insertions were done with gene partitioning and independents models, increasing the number of start trees to 60, with no partitioning and a simpler model, and by bootstrapping with 1000 replicates (see Figure 2 for all trees). The best tree returned with partitioning returned a final maximum likelihood value of -32836.52 and an AIC score of 65865.05. Increasing the number of start trees to 60 returned a final likelihood score of -32836.75 and an AIC score of 65865.51. There were no topology differences with a different number of start trees. Without gene partitioning and a GTR+G model for all sites, the final maximum parsimony score was -33021.003, with an AIC score of 66074.00. A bootstrap analysis of the no insertional maximum likelihood tree returned bootstrap values of 100 for every node.

All trees, regardless of model, showed *Amphimedon queenslandica* branching off early in the tree, *Niphates digitalis* and *Niphates erecta* clustering with short branches, and an unexpected grouping of *Xestospongia muta* and *Amphimedon compressa* clustering as well. Trees were viewed in FigTree v1.4.4, rerooted with *Amphimedon queenslandica*, and given the same scale of 0.1

Discussion

The goal of this project was to look to re-evaluate the phylogenetic position of *Niphates erecta* in relation to the new *Niphates digitalis* genome, and determine if the *N. erecta’s* previously reported variation was a by-product of their species-specific insertions or a genus wide level diversity. A secondary goal was to determine the effect, if any, *N. erecta’s* novel insertions could have on phylogenetic analysis. Phylogenetic analysis indicates that the variation might be a genus wide pattern of diversity, and not heavily influenced by insertions. In all trees generated in this study, *N. erecta* and *N. digitalis* showed considerable distance from the other species. This was seen regardless of whether insertions were considered or not, and is consistent with the tree generated by Lavrov et al 2019 [3]. That tree showed *N. erecta* with considerably long branches as compared to the many other species. Interestingly, even when the insertions were not considered, *N. erecta* showed high levels of similarity to *N. digitalis*. As these insertions were identified in relation to *N. digitalis'* genome, this level of similarity is unexpected. However, only five genes were considered in this study, and a larger distance might come to light if additional genes were studied.

Unexpectedly, *Niphates digitalis* showed a greater branch length then *Niphates erecta*, regardless of insertions or model. When no insertions are considered (Figure 2), *N. digitalis* has a branch length of 0.0807, while *N. erecta* has a branch length of 0.0375. This indicates that *N. digitalis* has a higher rate of evolution as compared to *N. erecta*, which is surprising considered the novel insertions found in *N. erecta*. When insertions are considered, branch length does not change significantly, to 0.0723 and to 0.0472 respectively. A hypothesis had been that a higher evolution rate could account for the presence of insertions in *N. erecta*, but *N. digitalis* has a higher evolution rate and lacks any of the insertions. A separate, brief analysis of the insertions using the same methods described here showed no homology between the insertions, and BLAST results show the insertions have no close biological matches. The insertions therefore could be a very recent addition to the *N. erecta* mitochondrial genome, or a very recent loss in *N. digitalis*.

Merritt Polomsky
EEOB 563
April 15, 2020
Final Project Paper – Draft #1

In regards to model selection, partitioning the data by gene did not impact tree topology (Figures 1, 2) . As is expected, branch length between trees with and without partitioning were different, and without partitioning, branch lengths were shorter on all branches. However, the trends mentioned above still held. In the trees with partitioning and the more complex models, likelihood scores were always lower and indicative of appropriate model choice. Unsurprisingly, likelihood scores were impacted by the insertion variable. Trees without insertions showed a higher likelihood value (-32836.52 vs -35755.04), but considering the insertions were only in *N. erecta*, is most likely being of increased conservation in the no insertion alignment. The addition of other genes to the analysis might change this model and impact tree topology, so models must be evaluated independently for every tree.

Merritt Polomsky
EEOB 563
April 15, 2020
Final Project Paper – Draft #1

References

1. Lavrov, Dennis V, and Walker Pett. “Animal Mitochondrial DNA as We Do Not Know It: mt-Genome Organization and Evolution in Nonbilaterian Lineages.” *Genome biology and evolution* vol. 8,9 2896-2913. 26 Sep. 2016, doi:10.1093/gbe/evw195
2. Srivastava, Mansi et al. “The Amphimedon queenslandica genome and the evolution of animal complexity.” *Nature* vol. 466,7307 (2010): 720-6. doi:10.1038/nature09201
3. Lavrov, Dennis V., et al. “Phylogenetic Relationships of Heteroscleromorph Demosponges and the Affinity of the Genus Myceliospongia (Demospongiae Incertae Sedis).” *BioRxiv*, Cold Spring Harbor Laboratory, 1 Jan. 2019, www.biorxiv.org/content/10.1101/793372v1.abstract.
4. Katoh, et al. “MAFFT: a Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *OUP Academic*, Oxford University Press, 15 July 2002, academic.oup.com/nar/article/30/14/3059/2904316.
5. Manolo, et al. “SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building.” *OUP Academic*, Oxford University Press, 23 Oct. 2009, academic.oup.com/mbe/article/27/2/221/970247.
6. Kozlov, et al. “RAxML-NG: a Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference.” *OUP Academic*, Oxford University Press, 9 May 2019, academic.oup.com/bioinformatics/article/35/21/4453/548

Merritt Polomsky
EEOB 563
April 15, 2020
Final Project Paper – Draft #1

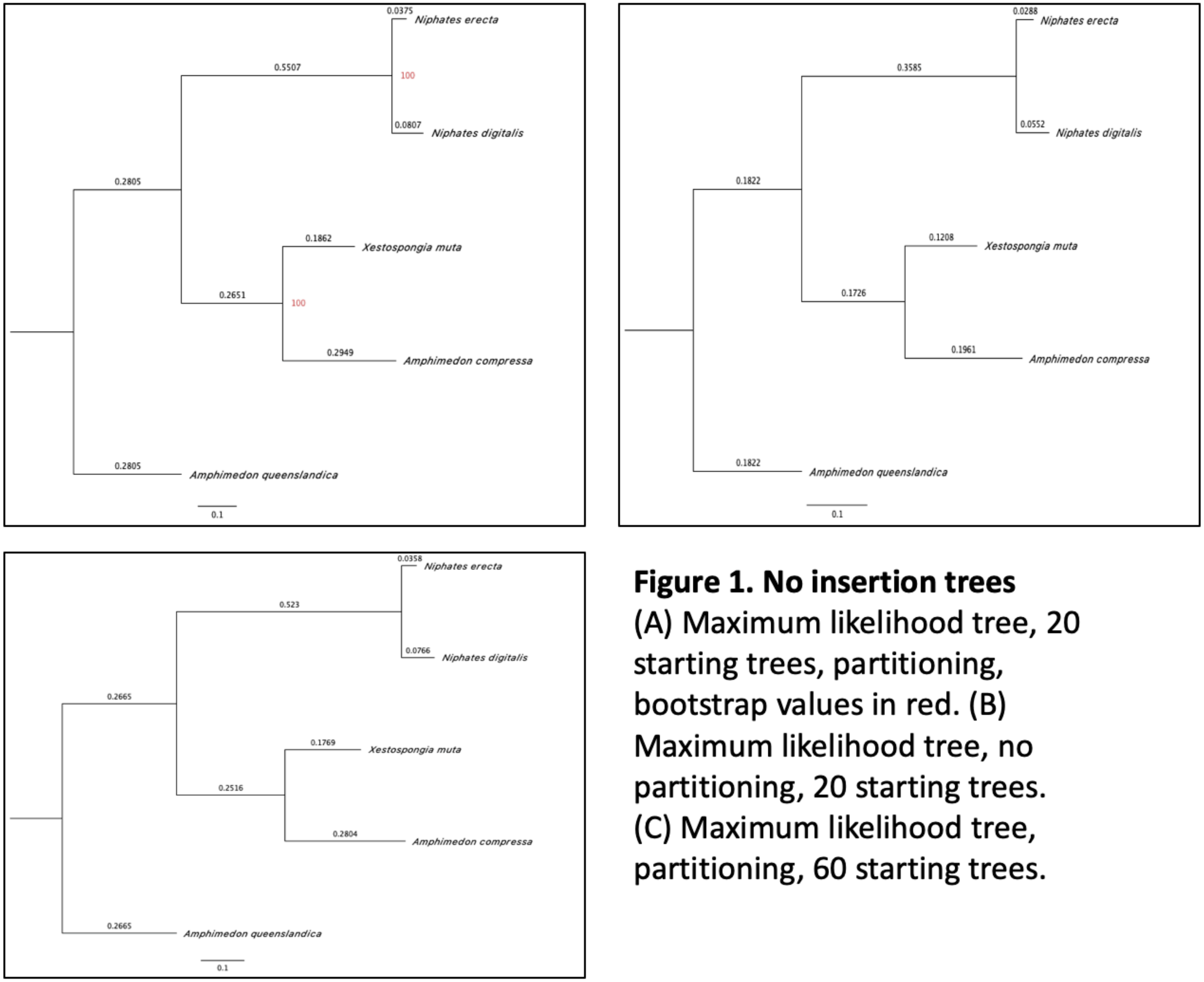

Figure 1. No insertion trees
(A) Maximum likelihood tree, 20 starting trees, partitioning, bootstrap values in red. (B) Maximum likelihood tree, no partitioning, 20 starting trees. (C) Maximum likelihood tree, partitioning, 60 starting trees.

Merritt Polomsky
EEOB 563
April 15, 2020
Final Project Paper – Draft #1

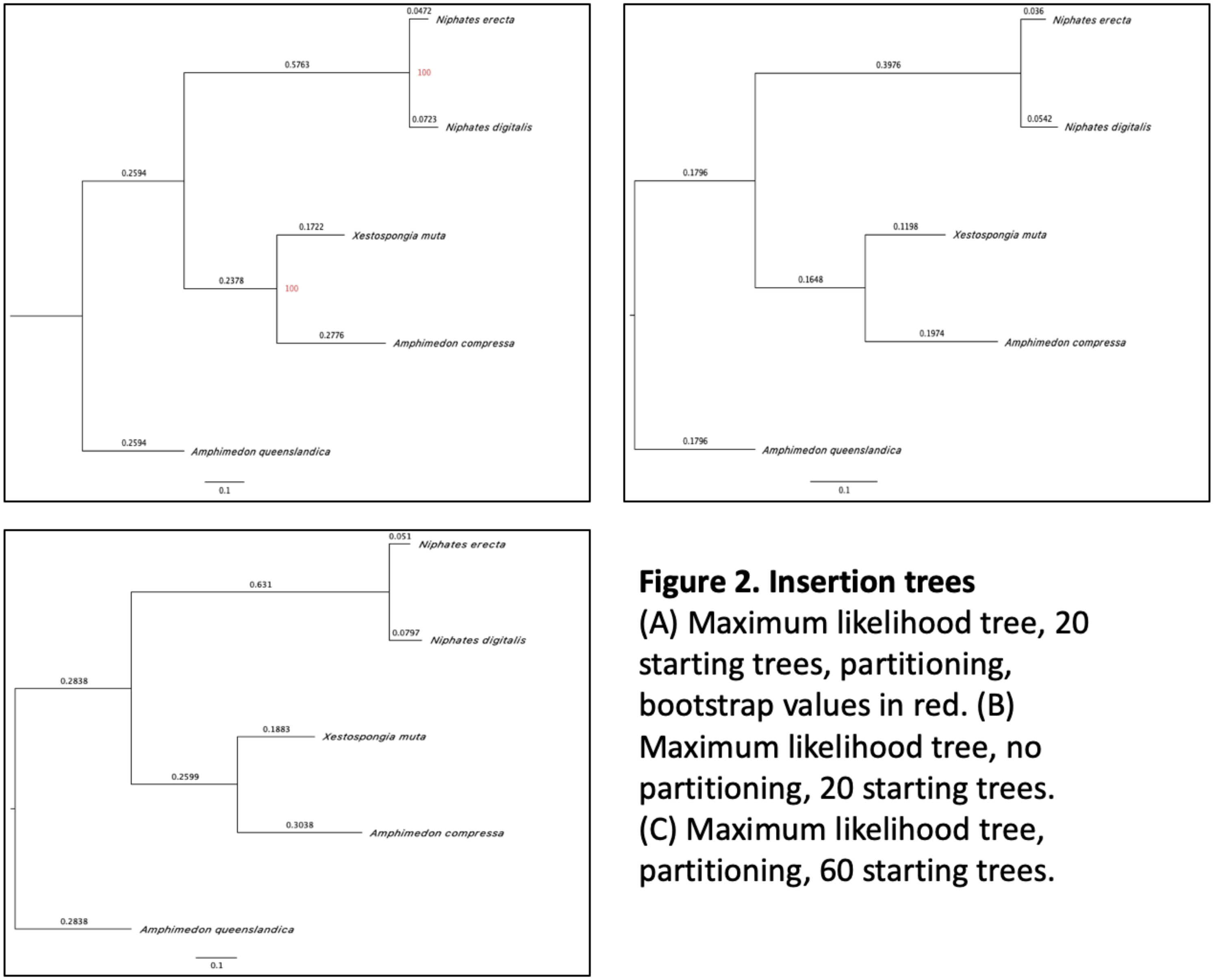

Figure 2. Insertion trees
(A) Maximum likelihood tree, 20 starting trees, partitioning, bootstrap values in red. (B) Maximum likelihood tree, no partitioning, 20 starting trees. (C) Maximum likelihood tree, partitioning, 60 starting trees.