**Phylogenetic Analysis of the genus *Niphates***

**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 These two sponges fall within the same genus, but unpublished analysis of gene content and organization show interesting differences.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 and in intergenic regions, when compared to *N. digitalis*. Of the 80 insertions found in *N. erecta*, they vary in size from only a few bases to 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.

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 and to determine if the high evolution rate is an artifact of these novel insertions.

**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 publicly available on NCBI.

From these genomes, six genes were chosen for analysis - *rnl*, *rns*, *cob*, *cox1*, *cox2*, and *cox3*. All six genes are 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*). Nucleotide sequences were aligned by individual genes using the auto function in MAFFTv7 [4] 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. Amino acid sequences were not used due to the uncertainty of the insertion presence in proteins. 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.

Before tree generation, alignments were assigned partitions based on gene location and run through ModelTest-NG[6] to test for optimal models. ModelTest-NG is a tool for selecting best-fit sequence substitution models, and returns the model that received the highest AIC, BIC, and DT scores for each partition or sequence. These scores evaluate the fit of a model to the data in question. For this experiment, the model with the highest AIC score was chosen for each gene.

Phylogenetic analysis was done using RAxML-NG[7] on the Iowa State University High Performance Clusters. The alignments were checked, parsed, and then the maximum likelihood trees were calculated, with 30 random and 30 parsimony-based starting trees. Data was partitioned by gene and with the selected models suggested by ModelTest-NG. From there, a 1000 replicate bootstrap analysis was run, and then applied to the tree.

**Results**

ModelTest-NG returned the same set of model assignments regardless of the presence of insertions, and this list of models with their AIC values is reported in Table 1.

RAxML-NG returned maximum likelihood values and tree topology for the best tree found. For the insertion alignment with gene partitioning, the best tree has maximum likelihood score of -35782.36 with an AIC score of 71678.73. As a verification of ModelTest-NG results, 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. Between the two models, there were no topology differences. A bootstrap analysis of the insertion tree with partitioning returned bootstrap values of 100 for all nodes (see Figure 1). Bootstrap values converged after 50 replicates.

As with the insertion alignment, analysis on sequences without *Niphates erecta* insertions were done with gene partitioning and ModelTest-NG model assignments, with no partitioning and a simpler model, and by bootstrapping with 1000 replicates (see Figure 2). The best tree with partitioning returned a final maximum likelihood value of -32868.60 and an AIC score of 65849.21. 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 showed *Niphates digitalis* and *Niphates erecta* clustering with short branches to their most recent common ancestor, and a long branch connecting the pair to the rest of the tree. *Amphimedon queenslandica* was the next closest related species. *Xestospongia muta* and *Amphimedon compressa* also clustered together. Trees were viewed in FigTree v1.4.4, rooted to be between the two clusters, and given a set scale of 0.1 for every tree.

**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. Results indicate 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, and with its most recently common ancestor connecting it to *Amphimedon queenslandica*. 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 six 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 0.0568, while *N. erecta* has a branch length of 0.0271. 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 0.0536 and to 0.0358 respectively, but the difference between the two species remains. 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. The insertions therefore could be a very recent addition to the *N. erecta* mitochondrial genome, or a very recent loss in *N. digitalis*.

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, AIC 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 (-32868.60 vs -35782.36), but considering the insertions were only in *N. erecta*, this shift in likelihood is most likely due to either increased sequence conservation or due to a reduction in sequence length. The addition of other genes to the analysis might change these results and impact tree topology, and the presence of insertions and other novel sequences elements need to be taken into account during phylogenetic analysis.

**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. Darriba, et al. “ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models.” *OUP Academic*, Oxford University Press, 21 Aug. 2019, academic.oup.com/mbe/article/37/1/291/5552155.
6. 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.
7. 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

|  |  |  |  |
| --- | --- | --- | --- |
| **Partition** | **Model** | **Insertion AIC** | **No Insertion AIC** |
| *rnl* | GTR+G4 | 24573.5288 | 21307.0438 |
| *rns* | GTR+G4 | 12390.3204 | 10365.8984 |
| *cob* | GTR+I+G4 | 9792.4349 | 9795.1404 |
| *cox1* | HKY+I+G4 | 11373.2144 | 11364.1936 |
| *cox2* | HKY+I | 7410.3945 | 6910.6929 |
| *cox3* | HKY+G4 | 6024.2234 | 6024.2234 |

**Table 1.** The table above shows the sequence evolution models returned by ModelTest-NG for each gene partition in the analysis. The same set of methods was returned regardless of the inclusion or exclusion of insertions. AIC values are reported for both insertion and no insertions.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated