The First Full Genome Phylogenetic Analysis of the Genus *Peromyscus*

# Introduction

Mitochondrial genomes are unique genomic sources for several reasons. First, they are fairly similar across metazoans. Most consist of one large non-coding region containing several regulatory elements, 2 rRNA, 22 tRNA, and 13 protein coding regions, which are of the 80 proteins coded for in oxidative phosphorylation. They also have at least one non-coding region, which contains regulatory elements.

Aside from their cellular functioning, an area where mitochondrial study has been of immense popularity is in phylogenetics. In fact, mitochondria have been the most popular phylogenetic markers since the 1980’s. This is the case for several reasons. First, it’s an easy genetic source to amplify, as there are multiple genomes per cell. Second, it’s a simple model. The genome is circular and contains no introns with only a few duplications. Also, its maternal inheritance means all characters share the same heredity. Third, while divergence can be rapid or conservative for certain mitochondrial genes, others are neutral, accurately reflecting divergence times, and their overall gene synteny is highly consistent across taxa.

The most up to date trees on *Peromyscus* phylogenies were based off of mitochondrial data. Bradley et al. (2007) used whole cytochrome-b genes, Miller and Engstrom (2008) used *Cytb*along with 2 nuclear genes, *Rbp3* and *Ghr*, and Platt et al. (2015) used two introns, *Adh1-I2* and *Fgb-I7*, in addition to *Cytb*and *Rbp3*. All three produced similar trees, but remain unresolved due to lack of support at midnodes (i.e. among species groups). Due to their persistent use of genetic, and not genomic analyses, the use of whole mitochondrial genomes should provide more credible results and clarity to the genus’ phylogenetic relationships. Here, we provide the ML and Bayesian trees of 12 *Peromyscus* species sensu lato and 4 outgroups, whose taxa are representative of *Peromyscus* species groups. Our results show similar groupings as previous trees but with increased nodal support.

# Methods

The data included were 15 recently sequenced mitochondrial genomes, along with a previously sequenced outgroup, of Cricetid rodents. the 15 recent genomes were sequenced on a single Illumina MiSeq lane, and whose genomes were generated using a custom script utilizing two programs, MIRA and Mitobim, a read trimmer and aligner, and a mitochondrial genome assembler. Once completed, these genomes were uploaded to MITOS, a program that annotates the genes and coding positions of each genome. After BLASTing along with manual curation to ensure accuracy of the genes produced, the genes of each were queued in their genomic order to form mitochondrial exomes. Only genes were included so as to prevent uninformative non-coding regions from affecting the output of our trees.

Each exome was aligned via clustalw2, as the length of the sequences were too much data for the memory allotment used by Muscle. The alignments were manually looked over and edited where necessary.

Once alignment was completed, the alignment file output was converted into a phylip and nexus format via EBI’s EMBOSS Seqret page. Once completed, the phylip file was uploaded and run through the JModelTest program. The data were given negative log likelihood scores for all model tests provided. Once finished, the “Show Results Table,” option was chosen, and the model with the lowest log likelihood score provided was chosen for both the ML and Bayesian analyses.

The ML tree was computed via RAxML. The phylip file aforementioned as the alignment file and the *Akodon montensis* mitochondrial genome, included in the phylip file, as the analysis’ outgroup. Additional parameters of note were the number of bootstrap replicates, which was 1,000, and that those were run using the “Rapid bootstrap analysis” command (-f a). The latter command also output a best-tree output file.

The Bayesian tree was computed using Mr. Bayes. The previously generated nexus file was used to run this analysis, and additions were added to the start and end of the file. In the data block, ntax was 16 because the mitochondrial genomes of 16 species were included, datatype was DNA because that’s what mitochondrial genomes consist of, and nchar was 16070 because that was the longest genome included in the provided nexus file. For the Bayes block, The initial end was to show the stop of the data in the nexus file, and lead into the next command, which was to begin running Mr. Bayes. The first command was to designate an outgroup, *A. montensis*, the second to establish the character set of the data. 16,070 was used because all the genomes included are at least that length, and the positions 1-3 included to establish codon positions among the genes included in each exome. These character sets were then combined into partitions. Under setting parameters and priors of parameters, defaults for running data via the GTR+I+G model were included. For parameters for the Monte-Carlo Markov Chain, nruns was made 4 and nchains 5 to use 20 nodes with which to run our data quickly, ngen made 5,000,000 to gain enough generations to ensure the significance of our data, samplefreq made 100, and printfreq made 1,000 to gain up to date checks on the run as it was happening. Finally, the sumt command called for a summary of all trees and their parameters to evaluate the data once generated.

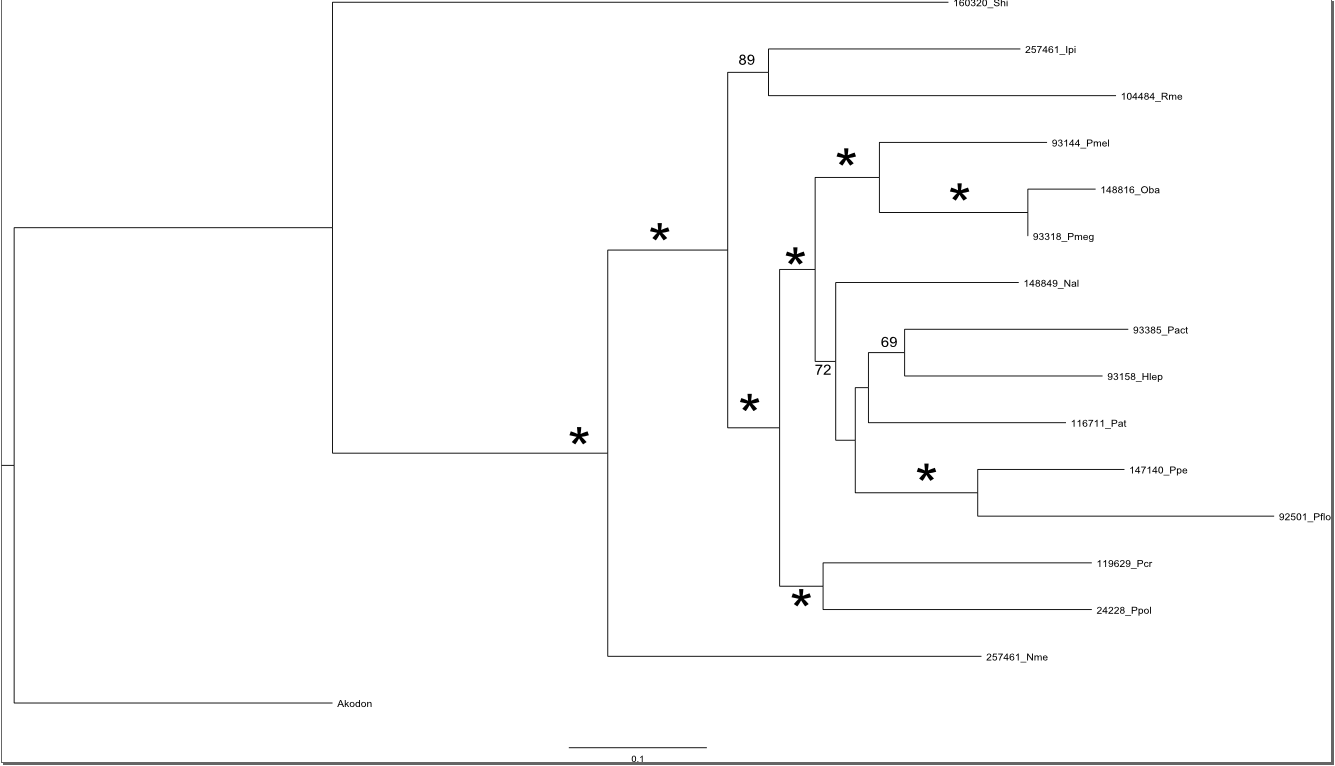
The custom scripts for MIRA and MitoBim, RAxML, and Mr. Bayes were uploaded to GitHub, the genomes uploaded to NCBI, and the alignment files (along with the files above) appended to the end of this paper.

# Results

Given the close relations among the taxa as well as overarching characteristics of mitochondrial genomes, each exome was very similar, and thus no issues with alignment were thought to have arisen. The alignments were clean, save some gapping errors at both ends, the existence of myriad SNPs, and a couple long insertions. All in all, the process was successful, and given the straightforwardness of the data included, any aligner would have been nearly if not more successful.

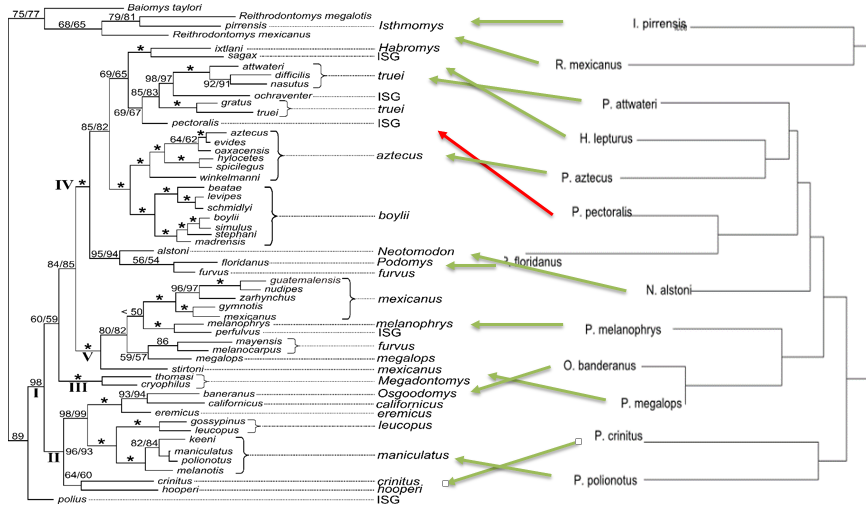
Our choice of model, GTR+I+G, was an appropriate choice given the size and complexity of our data. This model is the most parameter rich, and is actually included in the RAxML program, unlike all non-GTR models.

Figure 1: Best ML tree output by RAxML. Asterisks indicate significant bootstrap support, while missing bootstrap values indicate <50% support.

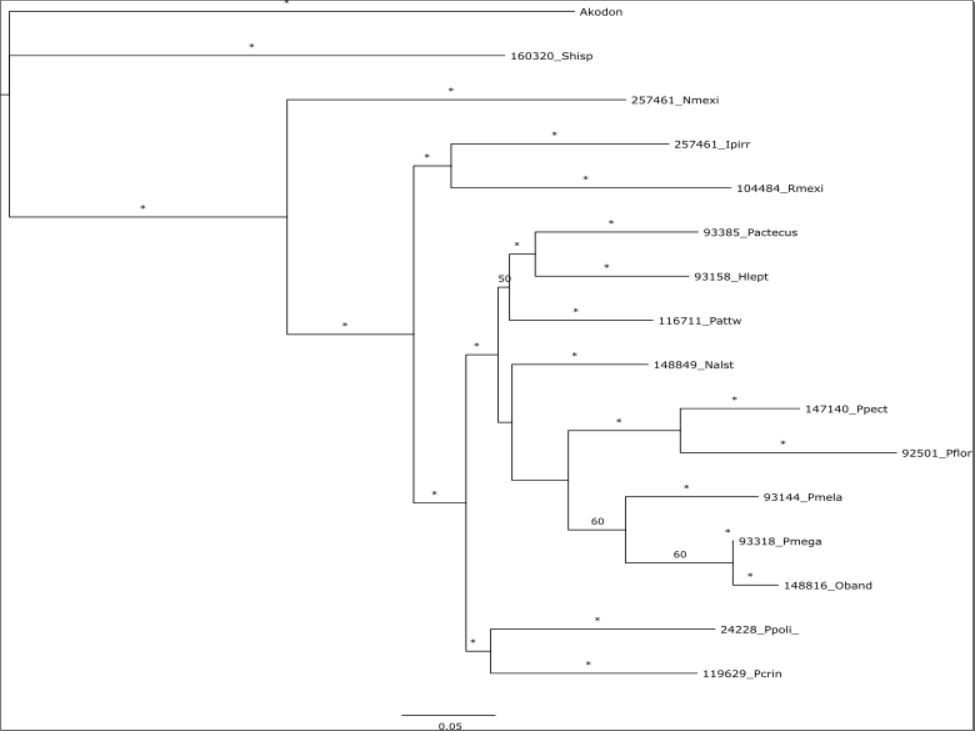


Our Maximum Likelihood tree, presented in Figure 1, paints a fairly similar picture as previous *Peromyscus* phylogenies, while overall showing greater nodal support. Figure 2 shows our tree compared with Bradley et al.’s, and illustrates changes in tree arrangement. First, our ML tree mirrors previous trees in that *S. hispidus, N. Mexicana, I. pirrensis,* and *R. mexicanus*are outgroups to the paraphyletic *Peromyscus* clade. The node separating *N. Mexicana* from the *Peromyscus* clade has 100% bootstrap support. The node separating the *Isthmomys* species group from *Peromyscus* shows 100% support, but the grouping itself only has a bootstrap support of 89%. While not significant, it is much greater than Bradley et al.’s support of 68/65, and might easily achieve support with more taxa in and around the subgenus. The node separating *S. hispidus* from *N. Mexicana* does not have a bootstrap value, as it is sister to *A. montensis*, which was mentioned as the outgroup in our RAxML commands. At the opposite end of the genus is the *P. crinitus* and *P. polionotus* pairing. While these are in different species groups (*crinitus* and *maniculatus*), they are sister sub-genera in Bradley’s analysis, and this tree recapitulates that relationship with 99% bootstrap confidence. Also similar to previous findings is the paraphyletic relationship of species within sensu lato *Peromyscus*. This tree recapitulates that branching, which distinguishes between a clade including the *maniculatus* and *crinitus* species groups compared to the rest. Finally, there is 100% bootstrap support linking the *megalops* and *melanophrys* species groups. While Bradley et al. showed that *P. perfulvus, mayensis,* and *melanocarpus*, representing the *furvus* and an unnamed species group, were grouped inbetween those two, such taxa were not included in this analysis, and thus still mirror previous results regarding what data were included.

Figure 2: a comparison of the ML tree produced by Bradley et al. with our Figure 1 tree. Arrows point from taxa on figure 1 to their species group in Bradley et al.’s tree. The succession correlates similarly, but *N. alstoni* located near the top of the tree shuffles the order for most mid-taxa. Also, while order may be similar, bootstrap support and branching is different for several clades.



Despite these similarities, the order of species and subgenera frequently do not match what’s been seen previously. First, the *Osgoodomys* species group was linked with *maniculatus* and *crinitus* in Bradley’s tree, but here is sister to the *megalops* group. This is surprising, given the bootstrap support that slated *O. banderanus* was significantly linked to its previous position within *Peromyscus*. Second, the previously sister sub-genera of *Neotomodon* and *Podomys* are no longer directly grouped. While tree reconfigurations allowed them to be placed together, *N. alstoni* is sister to the clade containing *Po. floridanus,* and that taxon is now sister to *Pe. pectoralis*. That being said, aside from the grouping of *Po. floridanus* and *Pe. pectoralis*, none of these nodes show significant bootstrap support. More mitogenomes representing multiple species within each sub-genus might do more to garner support at these nodes.

Figure 3: Bayesian tree of the mitogenomes. Asterisks are indicative of 100% probability. Values <50% were removed. 

The Bayesian tree ultimately turned out to be nearly identical to our ML tree, save the branching of *A. montensis* and *S. hispidus*. While the branching of this tree is exactly similar, there are differences in branch length and branch support. Like the ML tree, there is significant support at most nodes, but there isn’t significant support at the midnodes for *Peromyscus* taxa besides the *P. polionotus-crinitus* clade. Specifically, each clade of the paraphyly achieved 100% probability, but the two branches delineating from the larger clade only had probably values of 34 and 50. Additionally, the branch leading to *P. megalops* and *O. banderanus* only yielded a value of 60, the branch including those and *P. melanophrys* also 60, and the branch including those along with *Pe. pectoralis* and *Po. floridanus* a value of 40.

The effective sample size for these runs was far greater than 200, giving significance to the runs provided. By looking at the chains for each run, two of them got stuck at a local optimum with a negative log likelihiood of ~102910. However, two other chains were shown to have clearly migrated from this optimum to a better optimum of -102890. The clear jump shown by two of the chains indicates confidence in the consensus tree provided.

# Discussion

Overall, both trees gave identical phylogenetic relationships that mirrored previous analyses to a major extent, and in the process improved support at major nodes. The only two differences were were in branch lengths and support at certain nodes. Significance gained at nodes unsupported in Bradley et al.’s ML tree are found between *I. pirrensis* and *R. mexicanus* in the Bayesian tree, the node including *P. melanophrys* and *megalops* in the ML tree, and support for both nodes of the paraphyletic branching in both.

There were two major differences is phylogenetic relationships in both trees: the close pairing of *Peromyscus pectoralis* and *Podomys floridanus*, which was previously seen between *P. floridanus* and *Neotomodon alstoni*. While *N. alstoni* lies just outside this grouping in the provided analyses, *P. pectoralis* was previously closer to the *aztecus, attwateri,* and *Habromys* species groups than to *Podomys*. This clade had 100% support in both trees. The second major difference was the inclusion of *Osgoodomys banderanus* on the opposite side of the paraphyletic branch away from *P. crinitus* and *polionotus*. In the Bayesian tree, there was 100% support for inclusion on that side of the paraphyly, as well as grouping with the *megalops* species group, but with no significant support inbetween. In the ML tree, however, this clade showed 100% bootstrap support throughout.

While there was improvement in support, it lacked in the midregions, which is specifically where support previously was lacking. This is a frustrating result, as the use of whole genomes was hoped to provide enough conclusive data to resolve such issues. Regardless of differences in and lack of support, however, the exactness of the two trees despite being produced by different runs is striking, and should give confidence in the sensu lato phylogeny for *Peromyscus*, even if the order of species groups remains to be determined. Future studies should look to sequence and analyze more mitogenomes, filling in gaps and creating more data within and around unsupported nodes. While not as successful as hoped, 15 new mitogenomes were created, and further confirmation for previous phylogenetic analyses has been reached.