2018

Rates and patterns of evolution in a duplicated genome in the family Catostomidae

Megann Michelle Schmidt
University of Northern Iowa

RATES AND PATTERNS OF EVOLUTION IN A DUPLICATED GENOME IN THE FAMILY CATOSTOMIDAE

A Thesis Submitted

In Partial Fulfillment

of the Requirements for the Designation

University Honors and

Bachelor of Science: Biology - Honors Research Degree

Megann Michelle Schmidt

University of Northern Iowa

May 2018

This Study by: Megann Michelle Schmidt

Entitled: Rates and Patterns of Evolution in a Duplicated Genome in the Family Catostomidae

has been approved as meeting the thesis or project requirement for the Designation University Honors.

_____	_____
Date	Dr. Peter Berendzen, Honors Thesis Advisor, Biology
_____	_____
Date	Dr. Jessica Moon, Director, University Honors Program

ACKNOWLEDGMENTS

First and foremost, I would like to thank Dr. Gary and Myrna Floyd for financial support through their Undergraduate Research Assistantship, and I would also like to thank the University of Northern Iowa Biology Department for their support of undergraduate research. I would like to express my utmost gratitude towards Dr. Peter Berendzen and Zachary Sperstad for their support and guidance throughout this entire project. Additionally, thanks to the Iowa EPSCoR for funding collection of the data set, and thanks to Andrew Simons at the University of Minnesota and Jonathan Armbruster at Auburn University for help in data collection and funding. Lastly, I would also like to acknowledge Alan and Emily Lemmon at the Center for Anchored Phylogenomics, Florida State University for generation of the data set.

ABSTRACT

Whole genome duplication (WGD) is a process in which the entire genome of an organism is duplicated, making redundant genes which are subject to unique evolutionary forces. Various modes of selection create different genetic fates such as retention of ancestral function, development of new function, or loss of function. Because of these differing fates, WGD is hypothesized to be a major driving force behind diversification. In this project, DNA sequences from fish species in the family Catostomidae were examined to observe patterns of evolution following a known WGD. Gene trees were generated for 179 loci to determine the amount of divergence among duplicates, revealing divergence to be more common than conservation. Time calibrated phylogenies were generated revealing the date of initial duplicate divergence within the subfamily Ictiobinae to be approximately 63 million years ago. Further analysis could reveal the evolutionary fate of each loci, providing insight into the ways WGD affects diversification.

TABLE OF CONTENTS

INTRODUCTION

The effect of genome duplication on the subsequent evolution and diversification of groups of species has been a persisting question among molecular and evolutionary biologists. This thesis explores the ways in which whole genome duplication affects the subsequent evolution of the genome by looking into patterns of evolution in the sucker fish family Catostomidae. To conduct this study, three specific goals will be addressed. The first goal is to determine whether the duplicate sequences are more often divergent or conserved. The second goal of this research is to use genomic data and new technologies to test previously studied hypotheses regarding the varying amount of gene conservation in the subfamilies of Catostomidae (Ferris & Whitt, 1980). These previous studies were conducted in 1980, and since then, a variety of tools for genome sequences and data analysis have been developed. Therefore, the aim of this work is to see if similar results are obtained when these new technologies are utilized. The third goal of this research is to determine the date in which alleles belonging to members of subfamily Ictiobinae began to diverge. The results obtained from all three of these studies will allow for the identification of patterns affecting the evolution of Catostomidae and provide insight into how genome duplication affects the diversification and evolution of organisms.

LITERATURE REVIEW

Gene duplication is a process by which a region of DNA containing a gene is duplicated, and it is one of the mechanisms in which new genetic material is generated during molecular evolution. When the entire genome of an organism is duplicated the process is called whole genome duplication (WGD). WGD results in polyploidy; a condition in which the organism has more than two copies of each chromosome in the nucleus of cells. WGD events are highly documented among plants, and WGD is accepted as an important driving force behind plant diversification (Mable, 2003). In animals, WGD resulting in polyploidy is far less common than in plants. However, within animals it is most common in fish and amphibians (Mable, 2011; Zhan, 2014).

There are a number of traits of freshwater fishes and amphibians that increase the potential for unreduced gametes and the likelihood for polyploidy in these groups. Unreduced gametes are gametes that have not undergone complete meiosis, so instead of having half the genetic material of the parent cell, they have the same amount of genetic material as the parent cell. Therefore, when fertilization of unreduced gametes occurs, the resulting cell will have double the amount of genetic material resulting in polyploidy. Traits and behaviors that expose the organism to potential stress during breeding allow for the increased production of unreduced gametes. For example, breeding in freshwater environments and external fertilization, expose the organism to fluctuations in temperature causing more stress that could lead to improper meiosis and unreduced gametes (Mable, 2011). Other traits directly enable the production of unreduced gametes, such as the type of gametogenesis and a high propensity for hybridization (Mable, 2011). A combination of all of these traits appear in freshwater fish and amphibians, leading to the conclusion that they increase the likelihood of genome duplication.

Following WGD events, organisms are left with double the amount of genetic material, which allows varying forms of selection to act upon the duplicates because they may or may not be necessary for the organism to function effectively. Immediately after duplication, all regions of the genome are nearly identical, but depending on how selection acts on the gene, the duplicated copy could result in a number of different fates (Fig. 1; Lynch & Conery, 2000; Wedel, 2000). Nonfunctionalization is the most common fate in which one copy of the gene accumulates mutations due to a lack of selection until it loses its function completely, becoming a pseudogene (Lynch & Conery, 2003; Zhang, 2000). In the case of conservation, there is strong purifying selection acting on the gene causing all copies to retain the initial function. When conservation occurs, there is double the amount of gene, so double the amount of protein product would be produced (Zhang, 2000; Lynch & Conery, 2003). In other circumstances, one copy of the gene could retain its initial function while the other copy acquires a new function, which is referred to as neofunctionalization, or the ancestral function could be divided among the duplicates, which is termed subfunctionalization. In subfunctionalization, each of the sequences is degraded a little bit, so combining the two slightly degraded sequences produces the same amount of protein product as would be expected from the single ancestral gene (Lynch & Conery, 2003; Postlethwait et al., 2004; Sémon & Wolfe, 2007; Zhang, 2000).

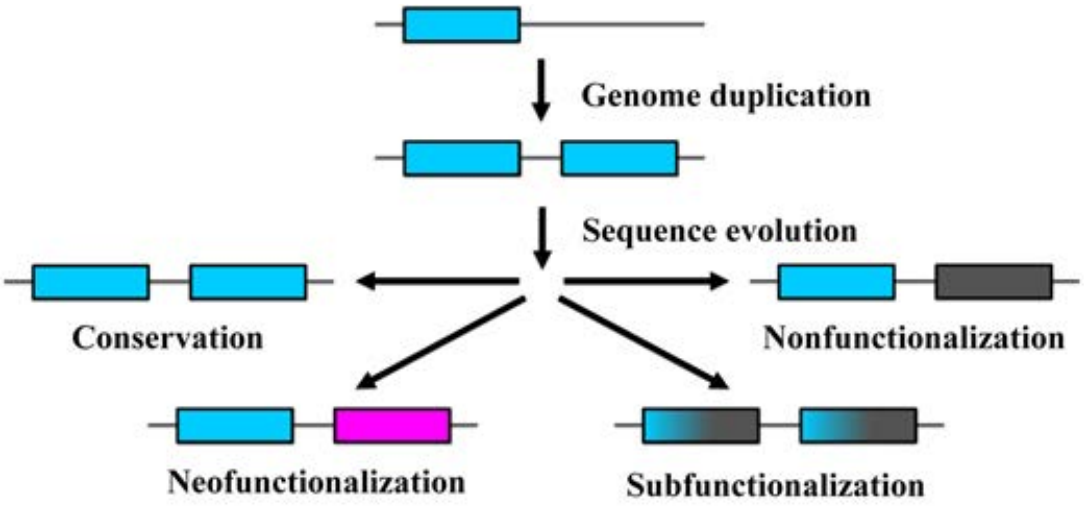

Figure 1: *Functional evolutionary fates of duplicate genes.*

It has been widely hypothesized that WGD events are a major driving force behind diversification (Van de Peer et al., 2009; Zhan et al., 2014; Zhang, 2003). It is thought that WGD events provide new genetic material for mutation, drift, and selection to act on, therefore providing a large contribution to evolution and diversification (Zhang, 2003). Additionally, varying fates of the duplicates in different organisms can lead to new gene functions, which could result in speciation through species-specific adaptation (Roth et al., 2007; Zhang, 2003). An example in which WGD events lead to diversification is the radiation of the teleost fishes into the most species-rich group of vertebrates (Postlethwait et al., 2004; Ravi & Venkatesh, 2008). In this study we will be examining a specific clade of teleost fishes, the Catostomidae, that are polyploid. The goal is to observe patterns of divergence that could reveal the mechanisms leading to diversification.

Catostomidae is a family of sucker fishes that arose from a single tetraploid ancestor that experienced a WGD event. This WGD event rendered the group tetraploid, meaning they have four copies of each chromosome. The group is composed of 78 species, and the majority of them

are distributed in North America. However, one species, *Myxocyprinus asiaticus*, is distributed in eastern Asia, and another, *Catostomus catostomus*, has a holarctic distribution (Chang et al., 2001; Harris & Mayden, 2001). The family contains four subfamilies: Ictiobinae, Cycleptinae, Catostominae, and Myxocyprininae (Ferris & Whitt, 1978; Harris & Mayden, 2001). A number of studies have been conducted on the evolution of the tetraploid genome of the suckers. One of these studies used isozymes to examine trends among the subfamilies of Catostomidae. Isozymes are enzyme proteins that complete the same function but are structurally different somehow, and the study analyzed them using gels, discovering a higher average percent of polymorphic loci in Ictiobinae than in the other clades of Catostomidae (Ferris & Whitt, 1980). In essence, they found that Ictiobinae is the least conserved of all the subfamilies. In this research, we hope to test this conclusion using genomic data that was not available to Ferris and Whitt during their 1980 study.

A variety of new technologies have been developed to aid in the analysis of genomes. Anchored hybrid elements (AHEs) are regions of organismal genomes that are highly conserved among evolutionary distant taxa. Genomic data can be collected from the AHEs and the regions of DNA adjacent to them without any prior knowledge about the genome (Lemmon et al., 2012; Siepel et al., 2005; Stout et al., 2016). These data are very useful when studying evolutionary histories and relationships of species because they can be tailored to fit a specific time scale, allowing researchers to look specifically at the most informative sites of a genome based on their study (Lemmon et al., 2012; Stout et al., 2016). In terms of this thesis, AHE data were generated and analyzed using a variety of phylogenetic software to observe genetic differences across different species in the family.

MATERIALS AND METHODS

Data Set Generation

Genomic DNA from 32 different catostomid species was collected by the Berendzen lab at the University of Northern Iowa as well as collaborating labs at the University of Minnesota and Auburn University. These samples were added to the data of 11 other catostomid species and 11 outgroups, which were obtained by Stout et al. (2016). The list of catostomid taxa included in the data set is listed in Table 1. The genomic DNA was sent to the Center for Anchored Phylogenomics (Lemmon et al., 2012) at Florida State University where genomic sequence data was collected using anchored hybrid enrichment. Anchored hybrid elements are regions of DNA, loci, that are highly conserved among evolutionary distant taxa. These regions were specifically selected because they are tailored to a timescale that will reveal differences among catostomid species. This means that the amount of difference between the sequences is informative enough to allow us to accurately discern the relationships among species of Catostomidae. Each of these regions of the genome were phased for four alleles to take into account the tetraploidy of the group, meaning at each locus, each catostomid species had four different sequences representing each copy of the gene in the genome. DNA sequence data for a total of 179 loci were collected.

Cycleptinae
Cycleptus elongatus
Myxocyprininae
Myxocyprinus asiaticus
Ictiobinae
Carpiodes carpio
Carpiodes cyprinus
Carpiodes velifer
Ictiobus bubalus
Ictiobus cyprinellus
Ictiobus niger
Catostominae
Catostomus ardens
Catostomus bernardini
Catostomus cahita
Catostomus catostomus
Catostomus clarkii
Catostomus columbianus
Catostomus commersonii
Catostomus discobolus
Catostomus insignis
Catostomus latipinnis
Catostomus leopoldi
Catostomus macrocheilus
Catostomus occidentalis
Catostomus platyrhynchus
Catostomus plebius
Catostomus wigginsi
Chasmistes liorus
Deltistes luxatus
Erimyzon oblongus
Hypentelium etowanum
Hypentelium nigricans
Hypentelium roanokense
Minytrema melanops
Moxostoma anisurum
Moxostoma arriommum
Moxostoma cervinum
Moxostoma duquesnei
Moxostoma erythrurum
Moxostoma lachneri
Moxostoma poecilurum
Moxostoma rupiscartes
Moxostoma valenciennsi
Thoburnia atripinnis
Thoburnia rhothoeca
Xyrauchen texanus

Table 1: *Catostomid taxa included in the study organized by subfamily. Taxa highlighted in yellow were obtained from Stout et al. (2016).*

Gene Tree Estimation

From the sequence data, maximum likelihood trees were generated for each locus. A maximum likelihood analysis generates the most likely phylogenetic tree based on site wise comparison of nucleotides. It compares sequences using a substitution model, which assumes that is it more likely for a base pair to stay the same than switch. Therefore, the sequences that are the most similar are the most closely related on the tree (Zwickl, 2006). The trees were generated using GARLI 2.01 (Genetic Algorithm for Rapid Likelihood Inference; Zwickl, 2006) on the CIPRES Science Gateway (http:www.phylo.org/index.php). The analyses were set to run with a maximum of 0.5 hours run time and with two independent search replicates to be performed during a program execution. The substitution model used in each analysis was a general time reversible (GTR) model with gamma distribution and four rate categories. Additionally, no bootstrap repetitions were used. Outgroups were designated after the analysis was complete using FigTree (Rambault, 2016). These parameters allowed for the 179 analyses to be completed in a timely manner. Time was the main consideration in setting of these parameters because only the relationships in these gene trees were being examined, not the branch lengths.

The gene trees obtained from the maximum likelihood analysis were visualized using FigTree (Rambault, 2016) and examined in two ways. First, each individual tree was examined to see which species, if any, had all four alleles included in a monophyletic clade. Monophyly refers to a group of organisms (or sequences) that have descended from a common evolutionary ancestor, and in terms of this analysis, it means all four alleles from a single species were more closely related to one another than they were to any of the other sequences. If a species did have all four alleles in a monophyletic clade, it was recorded as “conserved” for that locus. If the four alleles were not monophyletic, that species was recorded as divergent for that locus. Figures 2

and 3 show examples of trees in which *Ictiobus niger* was considered either divergent or conserved.

The second way in which the gene trees were examined was to identify the loci in which the sequences of Ictiobinae fell into two monophyletic clades. In terms of this analysis, two monophyletic clades meant that all the alleles of every ictiobine species fell into one of two distinct clades that did not contain sequences from any non-ictiobine taxon. The loci identified to have two monophyletic clades of Ictiobinae were used in further analyses to determine a date in which the alleles of ictiobine species began to diverge from one another. The specific group Ictiobinae was selected to narrow the scope of our analysis because it is widely accepted as a monophyletic clade within the Catostomidae. Figures 2 and 3 show examples of trees in which Ictiobinae was considered to be in two monophyletic clades or conserved.

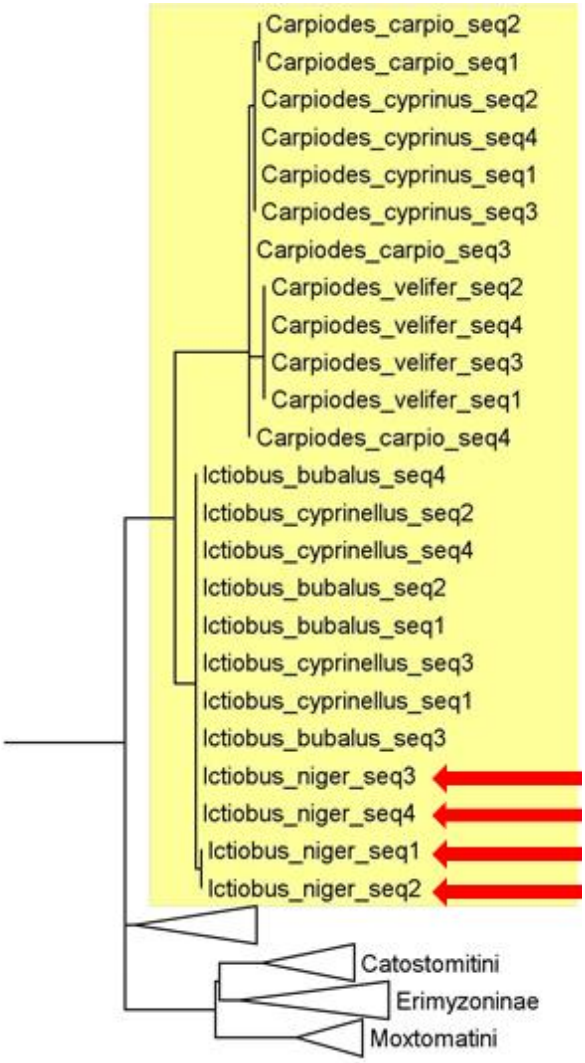

Figure 2: *An example of a gene tree in which Ictiobus niger and the clade of Ictiobinae were considered conserved in terms of this research. The clade Ictiobinae is highlighted in yellow, and the sequences of Ictiobus niger are marked with red arrows.*

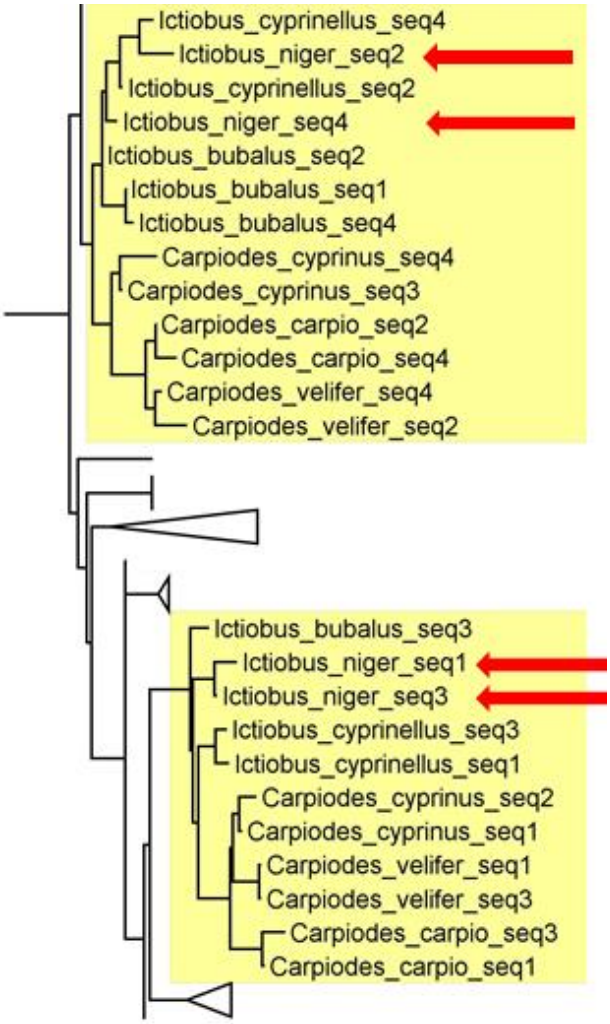

Figure 3: *An example of a gene tree in which Ictiobus niger was considered divergent and Ictiobinae has fallen into two clades. The two clades of Ictiobinae are highlighted in yellow, and the sequences of Ictiobus niger are marked with red arrows.*

Divergence Dating

The loci that were determined to have divergent lineages within the Ictiobinae were analyzed to determine the date in which the two clades diverged. To do this, trees were generated using a bayesian inference analysis in BEAST2 (Drummond & Rambaut, 2007; Suchard &

Rambaut, 2009) implemented on the CIPRES portal. Bayesian inference analysis proposes many random trees from which it samples to determine the most highly supported topological hypothesis (Drummond & Rambaut, 2007; Suchard & Rambaut, 2009). The bayesian trees generated in this study were time-calibrated phylogenies, meaning known fossil dates were used to calibrate the tree. Calibrating the data with fossils constrains the specific node dates to at least the date of the fossils because each of these clades could not have arisen any time later than their oldest known fossil. BEAST2 is capable of examining the amount of base pair substitutions relative to that timeframe to obtain a substitution rate. From there, it uses that substitution rate to date the unknown nodes (Drummond & Rambaut, 2007; Suchard & Rambaut, 2009).

The sequence files were modified a number of different ways before they could be uploaded into BEAST2 and run in an appropriate amount of time with sufficient effective sample size (ESS) values. All outgroups but one, *Cyprinus carpio*, were removed from the fasta files using MEGA. In order for the sequence files to be uploaded into Beauti 2.4.5, they were converted to a nexus format, also on MEGA. Beauti 2.4.5 produces an XML file which configures all the test parameters for the program BEAST2. The analyses were set to run using a GTR substitution model with empirical base frequencies. Additionally, the site heterogeneity model used was Gamma + Invariant Sites with four gamma categories. The clock used was an uncorrelated relaxed clock with a log normal distribution.

There were three fossil calibrations used for every locus. The first fossil calibration included all of the ingroup taxa and was calibrated to 61.7 million years ago (MYA) because that is the date of the oldest known catostomid fossil (Hirt et al., 2017). Calibrating this node restricts the date for the emergence of this group to be older than 61.7 MYA because the group must have arisen sometime before its oldest known fossil. The other fossil calibrations used were for

Ictiobinae. Each of these clades was calibrated to 33.9 MYA, the date of the oldest known Ictiobinae fossil (Mayden, 1992). The relationships obtained from the maximum likelihood trees were used to determine which specific sequences fell into which clade, and each of those clades was calibrated with the 33.9 MYA fossil. Each analysis was set to run for 10 million generations and logged every 10,000. This means the program sampled trees 10 million times, changing nodes and branch lengths every time selecting from branch lengths that results in a higher posterior probability of the tree. A consensus tree was generated for each locus, which is a single most probable tree based off of the trees produced in the 10 million different generations. The consensus trees were analyzed to determine the date of the node representing the most recent common ancestor between the two clades of Ictiobinae. By dating this node, we are able to date when the duplicate sequences for these species began diverging from one another.

RESULTS

The maximum likelihood trees generated were analyzed to determine the amount of divergence and conservation observed among different alleles of a single species. When these trees were analyzed, a large amount of divergence was observed. On average, species were divergent at 79% of the loci (Fig. 4).

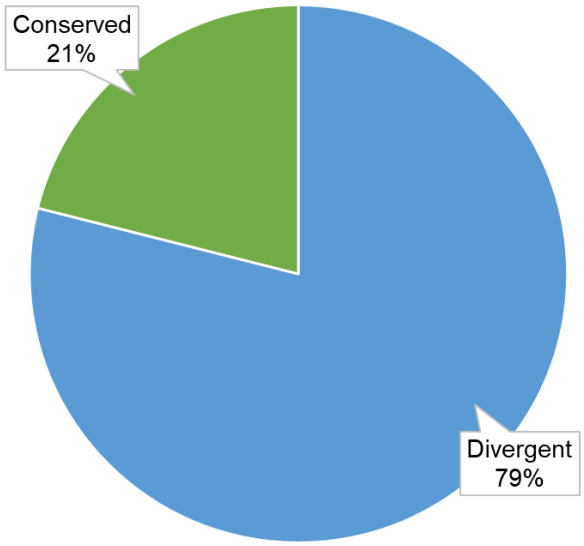

Figure 4: *Pie chart displaying the average amount of divergence and conservation observed for each species within the Catostomidae out of 179 loci.*

The amount of divergence shown by each species was also analyzed by clade to observe trends among the different taxonomic groups within Catostomidae (Fig. 5). The clade Ictiobinae displayed the highest percentage of polymorphic (divergent) loci, while the clade Erimyzoninae displayed the lowest percentage of polymorphic loci, making it the most conserved of all catostomid groups. Analyzing these results by clades allows us to compare (Fig. 5) our results to those obtained by Ferris and Whitt (1980).

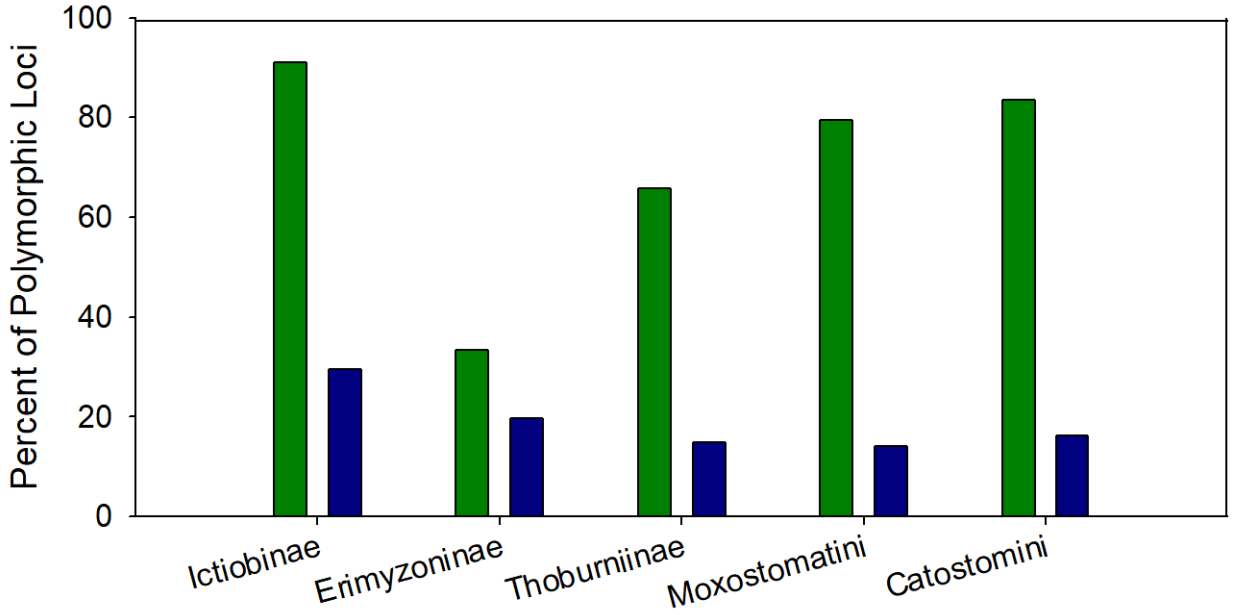

Figure 5: *A bar graph displaying a comparison of the percent of polymorphic loci by clade. The green bars show the results obtained in this study using genomic data, while the blue bars show the results obtained by Ferris and Whitt (1980).*

The maximum likelihood trees were also analyzed to determine which loci have the subfamily Ictiobinae falling into two monophyletic groups. This allowed us to determine which loci to further examine with the generation of time-calibrated phylogenies. Of the 179 loci in the data set, 72 of the loci had Ictiobinae falling into two monophyletic clades. Due to complications in editing the files, only 68 of these loci were used to generated a time calibrated phylogeny. Nine of the 68 loci were unable to produce reliable results with adequate ESS values, so a total of 59 loci were used to obtain time calibrated phylogenies. Figure 6 shows an example time calibrated phylogeny.

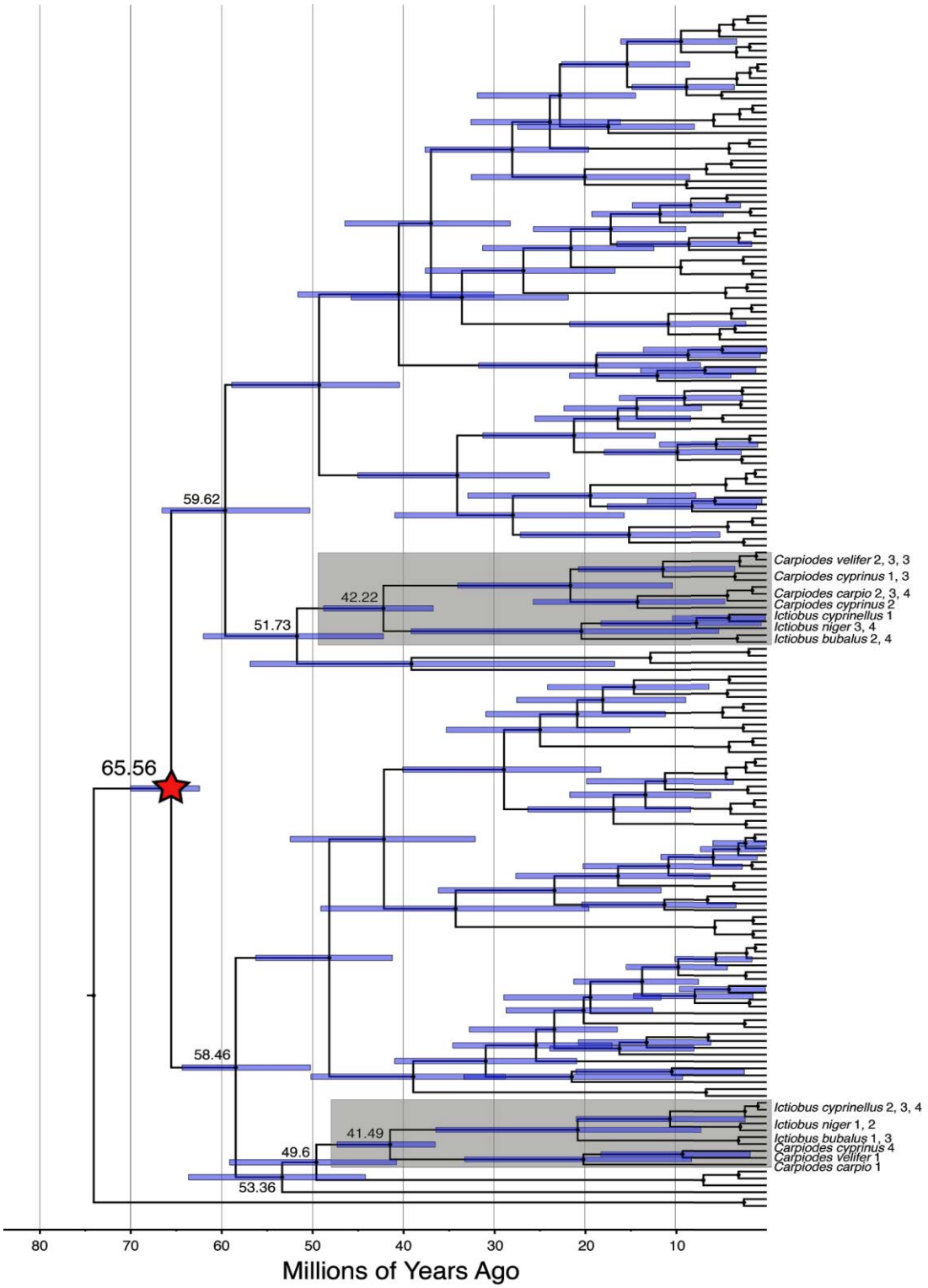

Figure 6: *Example time-calibrated phylogeny. Highlighted in gray are two clades of Ictiobinae. Blue bars show the 95% highest posterior density (HPD) intervals of each node. Numbers are predicted node ages, and node marked with red star is the most recent common ancestor of the divergent clades.*

The consensus trees obtained were examined to determine the date of the node representing the most recent common ancestor of the two diverging clades of Ictiobinae. Diverging alleles of Ictiobinae appear to have a most recent common ancestor between 60 to 70 MYA. This date appears to be consistent among the majority of the loci examined (Fig. 7).

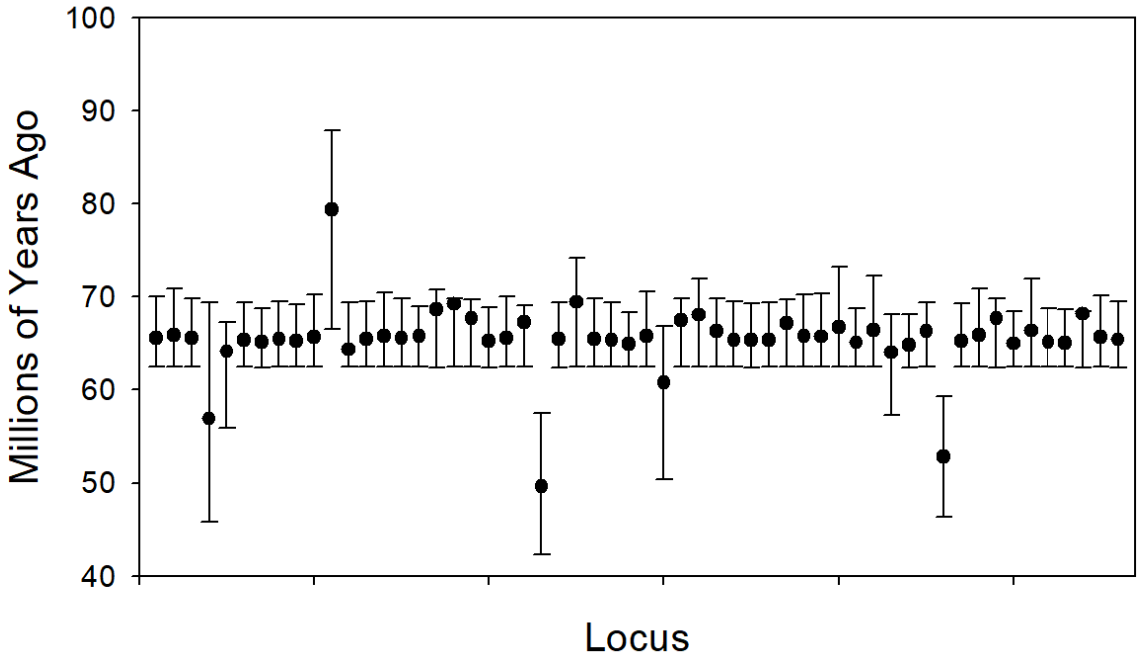

Figure 7: *Age of node representing the most recent common ancestor of diverging clades of Ictiobinae. Circles show node age, and lines show the 95% HPD interval.*

DISCUSSION

Genome duplication has been studied by evolutionary biologists for years, but there is still no clear understanding as to how this mutation affects the evolution and diversity of groups of species. This study hoped to characterize the patterns of divergence within the Catostomidae, a polypoid group. The analysis of the gene trees revealed that there is a large amount of divergence among the duplicated alleles of each species and far less evidence of conservation. At this time no conclusion about how this drives diversification can be drawn because there are a number of other possibilities that need to be considered. First, because we were looking at highly conserved regions of DNA, sequences are generally very similar even among evolutionary distant taxa. This large amount of similarity could have potentially created a false sense of divergence among alleles because at some loci, sequences from different species are so similar that even some outgroups are falling into the clades between sequences of catostomid species. Therefore, instead of observing extensive divergence between the duplicate sequences, we may actually be seeing very little variation indicating greater evidence of conservation among the species.

Another thing to consider when looking at our results (Fig. 4) is the potential fates of duplicate genes displayed (Fig. 1). Just because the alleles are divergent from one another, does not mean they are developing fuel for evolution. The duplicates could be mutating into nonfunctional units of DNA called pseudogenes. This would be termed nonfunctionalization, and in this case, it would not necessarily be providing fuel for evolution because the genes are nonfunctional. Additionally, the genes could potentially be returning to a functionally diploid state in which they only utilize two copies of their DNA even though they have four. Both of these options would mean that genome duplication has not actually provided fuel for

diversification within the Catostomidae, but just resulted in a larger amount of genetic material to manage, essentially junk sequences.

One last point to consider when analyzing our results is allelic variation. Allelic variation is neutral mutations that are different from the parent sequences but still produce the same protein product. This is possible because amino acids can usually be coded for by a couple different combinations of bases. Although these are differences from the ancestor sequence, they do not necessarily mean the sequence is diverging towards diversification because the products are still the same. If the reading frame for each locus is determined, we could eliminate these mutations and get a more accurate value for the amount of and type of divergence.

One of the primary goals of this research was to compare our genomic data with the results of the study by Ferris and Whitt (1980). A much greater amount of divergence is displayed in our data as opposed to Ferris and Whitt's (1980) (Fig. 6), which is to be expected because we were observing DNA sequence data while they observed protein products. Additionally, our results agree with Ferris and Whitt (1980) in that Ictiobinae showed the most polymorphic loci of any clade, but the trends among other subfamilies differ quite a bit. There are a couple possible reasons for this discrepancy.

We were able to visualize other forms of mutation that Ferris and Whitt (1980) were unable to observe. For example, Ferris and Whitt (1980) were unable to observe nonfunctionalization because they were using functional protein products. When nonfunctionalization occurs one of the copies becomes a pseudogene gene, which is no longer functional. In our analysis we were able to observe this potential fate because we were looking at specific sequences whether they were functional or not. However, Ferris and Whitt (1980) could

not have possibly observed any sequences undergoing nonfunctionalization because they were only using functional proteins, and pseudogenes do not produce functional proteins.

Additionally, looking at DNA sequences as opposed to proteins allowed us to observe neutral mutations. When visualizing proteins on a gel Ferris and Whit could only see mutations that affected the size or shape of the protein. Ferris and Whitt (1980) were unable to see neutral mutations by looking at the proteins because the proteins were not different, we were able to observe them by examining the specific sequences. This could also explain why other trends among the clades of Catostomidae differ between our results and those obtained by Ferris and Whitt (1980) (Fig. 6).

Node ages of the most recent common ancestor of the two clades of Ictiobinae were obtained from the time calibrated phylogenies (Fig 7). The majority of the loci had node ages falling between 60 to 70 MYA with only a few exceptions. This node age would mean that the sequences began diverging shortly after the whole genome duplication event that rendered them tetraploid. Results from other studies have found that the probability for divergence to occur is highest immediately after a duplication event, so our results appear to corroborate those findings. However, there are some outliers in our data. A few of the loci do not begin diverging until after the majority of the other loci. These most likely produce a protein product that is highly conserved, so mutations were not able to accumulate as quickly on them as in the others, making the node dates younger. The most puzzling locus is the one that diverged earlier than the rest. However, it is not of great concern because the 95% confidence interval is very large and lies partially within the range of our other data.

CONCLUSION

The effect of genome duplication on the subsequent evolution and diversification of groups of species has been a persisting question among molecular and evolutionary biologists. The goal for this thesis was to characterize the patterns of divergence in the polyploid group, Catostomidae. This was completed by answering three specific questions regarding the amount of divergence between duplicate sequences, patterns among subfamilies of Catostomidae, and the date in which divergence between duplicates began. Beginning with the first research question, a large amount of divergence was observed between duplicated alleles and their ancestor sequences within the Catostomidae. This divergence could potentially support the hypothesis that genome duplication leads to diversification, but further analysis would need to be completed to rule out extreme conservation, nonfunctionalization, and allelic variation. Additionally, genomic data supports the claim that Ictiobinae appears to be more variable than the other clades within the Catostomidae, but the trends regarding other clades differ. This difference is most likely due to the fact that we are able to detect different and smaller changes using genomic data than Ferris and Whitt (1980) were able to observe using protein data. Lastly, the dates when we begin to see divergence among loci within the Ictiobinae are similar, with only a few outliers. The divergence for most alleles began 60 to 70 MYA, suggesting that the majority of the genes began diverging before Ictiobinae diverged from the remainder of the Catostomidae and shortly after the WGD event that defines the group. Overall, the results of this study provide further insights into the ways whole genome duplication affects evolution and diversification and can be used in future studies as a resource when trying to determine how genome duplication has affected other families or species in general.

REFERENCES

Chang, M. M., Miao, D., Chen, Y., Zhou, J., & Chen, P. (2001). Suckers (Fish, Catostomidae) from the Eocene of China account for the family's current disjunct distributions. *Science in China Series D: Earth Sciences*, *44*(7), 577-586.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1), 214.

Ferris, S. D., & Whitt, G. S. (1978). Phylogeny of tetraploid catostomid fishes based on the loss of duplicate gene expression. *Systematic Zoology*, *27*(2), 189-206.

Ferris, S. D., & Whitt, G. S. (1980). Genetic variability in species with extensive gene duplication: the tetraploid catostomid fishes. *The American Naturalist*, *115*(5), 650-666.

Harris, P. M., & Mayden, R. L. (2001). Phylogenetic relationships of major clades of Catostomidae (Teleostei: Cypriniformes) as inferred from mitochondrial SSU and LSU rDNA sequences. *Molecular Phylogenetics and Evolution*, *20*(2), 225-237.

Hirt, M. V., Arratia, G., Chen, W. J., Mayden, R. L., Tang, K. L., Wood, R. M., & Simons, A. M. (2017). Effects of gene choice, base composition and rate heterogeneity on inference and estimates of divergence times in cypriniform fishes. *Biological Journal of the Linnean Society*, *121*(2), 319-339.

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic biology*, *61*(5), 727-744.

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*(5494), 1151-1155.

Mable, B. K. (2004). 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biological Journal of the Linnean Society*, *82*(4), 453-466.

Mable, B. K., Alexandrou, M. A., & Taylor, M. I. (2011). Genome duplication in amphibians and fish: an extended synthesis. *Journal of Zoology*, *284*(3), 151-182.

Mayden, R. L. (1992). *Systematics, historical ecology, and North American freshwater fishes*. Stanford university press.

Postlethwait, J., Amores, A., Cresko, W., Singer, A., & Yan, Y. L. (2004). Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends in Genetics*, *20*(10), 481-490.

Rambault, A. (2016). FigTree (Version 1.4.3) [software]. Available from http://tree.bio.ed.ac.uk/software/figtree/.

Ravi, V., & Venkatesh, B. (2008). Rapidly evolving fish genomes and teleost diversity. *Current opinion in genetics & development*, *18*(6), 544-550.

Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D., & Liberles, D. A. (2007). Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *308*(1), 58-73.

Sémon, M., & Wolfe, K. H. (2007). Consequences of genome duplication. *Current opinion in genetics & development*, *17*(6), 505-512.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., & Weinstock, G. M. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, *15*(8), 1034-1050.

Stout, C. C., Tan, M., Lemmon, A. R., Lemmon, E. M., & Armbruster, J. W. (2016). Resolving Cypriniformes relationships using an anchored enrichment approach. *BMC evolutionary biology*, *16*(1), 244.

Suchard MA & Rambaut A (2009) Many-Core Algorithms for Statistical Phylogenetics. *Bioinformatics, 25*, 1370-1376.

Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature reviews. Genetics*, *10*(10), 725.

Wendel, J. F. (2000). Genome evolution in polyploids. In *Plant molecular evolution* (pp. 225-249). Springer Netherlands.

Wilson, M. V. (1980). Oldest known Esox (Pisces: Esocidae), part of a new Paleocene teleost fauna from western Canada. *Canadian Journal of Earth Sciences*, *17*(3), 307-312.

Zhan, S. H., Glick, L., Tsigenopoulos, C. S., Otto, S. P., & Mayrose, I. (2014). Comparative analysis reveals that polyploidy does not decelerate diversification in fish. *Journal of evolutionary biology*, *27*(2), 391-403.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in ecology & evolution*, *18*(6), 292-298.

Zwickl, D. J. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.