2018

Reconstructing the phylogeny and characterizing the patterns of molecular evolution of the tetraploid freshwater suckers (Cypriniformes: Catostomidae)

Zachary Evan Sperstad
University of Northern Iowa

Copyright ©2018 Zachary Evan Sperstad
Follow this and additional works at: https://scholarworks.uni.edu/etd

Part of the Biology Commons

Let us know how access to this document benefits you

Recommended Citation
Sperstad, Zachary Evan, "Reconstructing the phylogeny and characterizing the patterns of molecular evolution of the tetraploid freshwater suckers (Cypriniformes: Catostomidae)" (2018). Theses and Dissertations @ UNI. 604.
https://scholarworks.uni.edu/etd/604

This Open Access Thesis is brought to you for free and open access by the Graduate College at UNI ScholarWorks. It has been accepted for inclusion in Theses and Dissertations @ UNI by an authorized administrator of UNI ScholarWorks. For more information, please contact scholarworks@uni.edu.
RECONSTRUCTING THE PHYLOGENY AND CHARACTERIZING THE PATTERNS OF MOLECULAR EVOLUTION OF THE TETRAPLOID, FRESHWATER SUCKERS (CYPRINIFORMES: CATOSTOMIDAE)

An Abstract of a Thesis

Submitted

in Partial Fulfillment

of the Requirements for the Degree

Master of Science

Zachary Evan Sperstad

University of Northern Iowa

July 2018
ABSTRACT

The Catostomidae, colloquially known as the suckers, is a family of freshwater fish endemic to North America and Asia. This family is hypothesized to have evolved sometime before or during the Paleocene (56-66 Mya) from a single tetraploid ancestor, which is thought to be the product of a hybridization event between two closely related, diploid cypriniforms. Currently, there are 79 recognized, extant species, some of which are difficult to discriminate between in the field. Despite the numerous studies that have aimed to reconstruct the evolutionary history of this family, little consensus exists for the relationships of the subfamilies within the Catostomidae, with practically every combination of subfamilial relationships having been proposed in the past. Additionally, and of importance to our understanding of the evolution of the catostomids, little is still known about the consequences of whole genome duplication on molecular evolution, especially for polyploid animals. In this study, we sought to reconstruct the evolutionary history of the Catostomidae as well as characterize the patterns of molecular evolution of lineages within this family. Two nucleotide sequence, genome-scale data sets were generated with the aim to reconstruct the evolutionary history of the Catostomidae as well as characterize patterns of molecular evolution of their polyploid genomes. These data sets, an unphased data including one sequence for each taxon and a phased data with the number of sequences per taxon representative of their ploidy level, included 179 and 267 loci, respectively. From the reconstruction of the evolutionary history of the family, we recovered a topology which places *Myxocyprinus asiaticus* as the sister taxon to all other extant catostomids and *Cycleptus elongatus* as the sister taxon to an Ictiobinae +
Catostominae clade. Additionally, we found that *Catostomus* was recovered as paraphyletic, with *Deltistes luxatus*, *Chasmistes liorus*, and *Xyrauchen texanus* forming strongly supported sister species relationships with species within *Catostomus*. In the second chapter, we found that the ictiobines, cycleptines, and myxocyprinines tended to have more polymorphic alleles than taxa within Catostominae. We also found that rates of molecular evolution were significantly greater within catostomine lineages than all other catostomid lineages.
RECONSTRUCTING THE PHYLOGENY AND CHARACTERIZING THE PATTERNS OF MOLECULAR EVOLUTION OF THE TETRAPLOID, FRESHWATER SUCKERS (CYPRINIFORMES: CATOSTOMIDAE)

A Thesis
Submitted
in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Zachary Evan Sperstad
University of Northern Iowa
July 2018
This Study by: Zachary Evan Sperstad

Entitled: RECONSTRUCTING THE PHYLOGENY AND CHARACTERIZING THE PATTERNS OF MOLECULAR EVOLUTION OF THE TETRAPLOID, FRESHWATER SUCKERS (CYPRINIFORMES: CATOSTOMIDAE)

has been approved as meeting the thesis requirement for the 

Degree of Master of Science in Biology


Date    Dr. Peter Berendzen, Chair, Thesis Committee

Date    Dr. James Demastes, Thesis Committee Member

Date    Dr. Mark Sherrard, Thesis Committee Member

Date    Dr. Patrick Pease, Interim Dean, Graduate College
ACKNOWLEDGMENTS

Foremost, I would like to thank my graduate advisor, Dr. Peter Berendzen, for his willingness to accept me as his graduate student. Over the past few years, Dr. Berendzen has been an endless supply of wisdom and support. I recognize Dr. Berendzen as being an important factor in my acceptance into the Ecology, Evolution, and Behavior PhD program at the University of Minnesota. In essence, I am the scientist I am today because of Dr. Berendzen.

To my graduate committee, Dr. Mark Sherrard and Dr. James Demastes, I would like to thank you for your time and energy helping me work through the production of this thesis. Additionally, I would like to thank Dr. Theresa Spradling and Dr. Kenneth Elgersma for their advice and insight regarding analyses and topics reported herein.

I would like to thank the following people and institutions for providing tissues for this study: Alexandra Snyder, University of New Mexico Museum of Southwestern Biology; Brian Sidlauskus, Oregon State University Ichthyological Collection; David Werneke, Auburn University Fish Tissue Collection; and Phil Harris: University of Alabama Ichthyological Collection.

Megann Schmidt, an undergraduate student in the Biology Department, played a very important role in the fruition of this thesis. Her assistance with data analysis and brain-storming aided greatly in this research. Without her help, Chapter 2 of this thesis would be practically nonexistent.

Lastly, I would like to sincerely thank my family and friends for their endless love and support, without which I would not have been able to further my education by
pursuing a graduate degree. I would like to specifically acknowledge my graduate school colleagues (Charles, Jenna, Clarissa, Corinne, Hannah, Navinder, Laura, Brett, and Sean) for pushing me to become a better student, teacher, and human. I would like to thank Katie Croson for being my biggest supporter and dearest friend. Furthermore, I thank my parents, John and Beth, as well as my siblings, Danielle, Ryan, and Kaelyn, for the encouragement, interest, time, and energy they have put into my academic journey.
TABLE OF CONTENTS

LIST OF TABLES ............................................................................................................... vii

LIST OF FIGURES ......................................................................................................... viii

CHAPTER 1. RESOLVING THE EVOLUTIONARY HISTORY OF THE FRESHWATER SUCKERS USING A GENOME-SCALE DATA SET (CYPRINIFORMES: CATOSTOMIDAE) ................................................................. 1

Introduction .................................................................................................................. 1

Materials and Methods ................................................................................................ 6

Taxon Sampling ........................................................................................................... 6

Data Collection .......................................................................................................... 6

Partitioning/Substitution Model .............................................................................. 7

Phylogenetic Analyses ............................................................................................ 8

Coalescent-Based Method of Phylogeny Reconstruction ...................................... 9

Topological Comparison ......................................................................................... 9

Phylogenetic Informativeness and Data Filtration .............................................. 10

Robinson-Foulds Distances ................................................................................... 13

Results ......................................................................................................................... 14

Phylogenetic Analyses of the Total Data Set ...................................................... 14

Topological Comparison ......................................................................................... 18

Phylogenetic Informativeness and Data Filtration .............................................. 18

Robinson-Foulds Distances ................................................................................... 22

Discussion .................................................................................................................... 23
Relationships among the Catostomidae .................................................................27
Conclusion ..................................................................................................................32

CHAPTER 2. PATTERNS OF MOLECULAR EVOLUTION WITHIN A FAMILY OF
TETRAPLOID, FRESHWATER FISH (CYPRINIFORMES: CATOSTOMIDAE) ......35

Introduction ..................................................................................................................35
Materials and Methods .................................................................................................41
Data Collection and Taxon Sampling ...................................................................41
Gene Tree Estimation ............................................................................................41
Estimating the Extent of Polymorphic Loci ...........................................................42
Branch Length Comparisons ..................................................................................42
Comparison of Alleles with Greater than Average Branch Lengths .....................43

Results .........................................................................................................................44
Data Collection and Taxon Sampling ...................................................................44
Estimating the Extent of Polymorphic Loci ...........................................................44
Branch Length Comparisons ..................................................................................46
Comparison of Alleles with Greater than Average Branch Lengths .....................50

Discussion ....................................................................................................................52
Estimating the Extent of Polymorphic Loci ...........................................................52
Branch Length Comparisons ..................................................................................54
Comparison of Alleles with Greater than Average Branch Lengths .....................55

Conclusion ...................................................................................................................56

REFERENCES ..................................................................................................................58

APPENDIX: SUPPLEMENTARY MATERIAL ................................................................65
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summary Statistics of Nodal Support Values for Subfamilial Relationships Recovered from the Data Set Filtration Analyses</td>
</tr>
<tr>
<td>2</td>
<td>Tukey’s Honestly Significant Post-hoc Test for a Pairwise Comparison of Generic Branch Lengths of the Catostominae</td>
</tr>
<tr>
<td>A1</td>
<td>Voucher Specimens Included in the Study</td>
</tr>
<tr>
<td>A2</td>
<td>R Script Used for Obtained Robinson-Foulds Distances</td>
</tr>
<tr>
<td>A3</td>
<td>Tukey’s Honestly Significant Post-hoc Test for a Pairwise Comparison of Generic Branch Lengths of the Catostomidae</td>
</tr>
<tr>
<td>A4</td>
<td>R Script Used for Removing Branch Lengths from Gene Tree Files</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Previously Proposed Hypotheses for the Phylogeny of the Catostomidae ..........4</td>
</tr>
<tr>
<td>2</td>
<td>Profiling Phylogenetic Informativeness .............................................11</td>
</tr>
<tr>
<td>3</td>
<td>Topologies Recovered from the Phylogenetic Analyses .............................15</td>
</tr>
<tr>
<td>4</td>
<td>Phylogenetic Trees Recovered from the Data Set Filtration Approach ............20</td>
</tr>
<tr>
<td>5</td>
<td>Robinson-Foulds Distance of Gene Trees to the Species Trees Recovered ..........22</td>
</tr>
<tr>
<td>6</td>
<td>Robinson-Foulds Distances of Gene Trees to Each Other ............................23</td>
</tr>
<tr>
<td>7</td>
<td>The Phylogeny of the Catostomidae Supported by this Study ........................28</td>
</tr>
<tr>
<td>8</td>
<td>The Potential Fates of Redundant Gene Copies after a WGD Event .................37</td>
</tr>
<tr>
<td>9</td>
<td>Polymorphic Loci Found within the Lineages of the Catostomidae ..................45</td>
</tr>
<tr>
<td>10</td>
<td>A Comparison of Branch Lengths between the Genera of the Catostomidae .........47</td>
</tr>
<tr>
<td>11</td>
<td>A Comparison of Branch Lengths between Genera within Catostominae .............48</td>
</tr>
<tr>
<td>12</td>
<td>A Heat Map of Branch Lengths of Gene Trees from this Study .....................51</td>
</tr>
<tr>
<td>13</td>
<td>A comparison of the abundance of alleles per genus within the Catostomidae ....52</td>
</tr>
</tbody>
</table>
CHAPTER 1
RESOLVING THE EVOLUTIONARY HISTORY OF THE FRESHWATER SUCKERS
USING A GENOME-SCALE DATA SET (CYPRINIFORMES: CATOSTOMIDAE)

Introduction

Over the past two decades, radical advances have been made in the field of phylogenetic systematics. Improvements in DNA sequencing technologies and techniques have enabled the reconstruction of phylogenies using data sets with hundreds to thousands of loci and taxa (Molloy and Warnow 2017), a practice referred to as phylogenomics (Philippe and Blanchette 2007). The prospect of increasing the number of phylogenetically informative sites within a data set, adding more taxa to break up long branches, and the quick turn-around for analyses using innovative methodologies was thought to be the answer to resolving unclear and poorly-supported nodes in the Tree of Life (Pyron et al. 2014; Linkem et al. 2016).

The number of studies using genomic data sets and phylogenomic techniques is increasing rapidly and have both corroborated longstanding hypotheses as well as proposed new, unexpected relationships. Some of these large-scale studies include the reconstruction of Aves (Jarvis et al. 2014; Prum et al. 2015), Angiospermae (Léveillé-Bourret et al. 2017), and Acanthamorpha (Eytan et al. 2015). Although these studies have generated interesting new perspectives and supported past suppositions, some nodes on these trees remain poorly supported, suggesting that increasing the amount of genetic data and taxa alone is not the solution to resolving contentious nodes. One possible explanation is groups that have been popularly targeted for phylogenomic studies are often groups that are hypothesized to have experienced rapid lineage accumulation events.
(Kozak et al. 2006). This process can result in the recovery of gene trees in which distantly related taxa appear more closely related to each other than more closely related taxa. Other biological processes, such as horizontal gene transfer, gene/genome duplication, hybridization, and substitution saturation, also have the potential to produce gene tree topologies that are discordant with the species tree topology (Baum and Smith 2013; Linkem et al. 2016; Molloy and Warnow 2017) and thus, selecting loci inattentively may lead to erroneous conclusions.

To address the issues of using phylogenomic data sets (the accumulation of noisy data, discordant gene trees, etc.), data are being more closely examined, utilizing a multitude of different tree reconstruction methods and/or preferentially selecting data to create data set subsets (Chakrabarty et al. 2017; Shen et al. 2017). Shen et al. (2017) suggested that data sets should be examined to identify sites and loci that disproportionately influence a contentious node to determine which data have the strongest phylogenetic signal and are best used to elucidate relationships among taxa. Furthermore, Shen et al. (2017) showed that removing only a few particularly informative nucleotide sites or loci can radically reduce the relative likelihood of a phylogeny or support an alternative topology, making the discrimination between high-quality and low-quality loci imperative.

The process of data set filtration, however, should be guided by a measure of appropriateness of resolving a node to ensure loci or sites are not being selected haphazardly. Criteria for the selection of loci or sites within loci for phylogenetic reconstruction have included phylogenetic informativeness (Townsend 2007; Dornburg et
al. 2017), gene tree estimation error (Molloy and Warnow 2017), and the amount of missing data from a data set (i.e. selecting loci that include data for all or nearly all taxa included in a study; Molloy and Warnow 2017). Despite the many methods that have been proposed recently that aim to discriminate between informative and noisy data, the most appropriate method of data set scrutiny remains unclear (Arcila et al. 2017).

Herein, we present the reconstruction of the evolutionary history of the Catostomidae using a genome-scale data set. In this study, we use a data set generated using anchored hybrid enrichment (Lemmon et al. 2012) and a data set filtration approach that selects loci based on phylogenetic informativeness (López-Giráldez et al. 2013; Dornburg et al. 2014; Dornburg et al. 2017) in an attempt to provide a robust phylogenetic hypothesis of the family. This study represents the first time the phylogeny of the Catostomidae has been reconstructed using a genomic data set.

The Catostomidae, Suckers, is a family of freshwater fishes within Cypriniformes endemic to North America and Asia. This family is hypothesized to have evolved from a single, tetraploid ancestor resulting from a whole genome duplication event sometime during or before the Paleocene (56-66 Mya; Hirt et al. 2017). There are currently 79 recognized, extant species in this family, and although monophyly of the group has not been questioned, many of the relationships remain elusive (Fig. 1). Despite the incongruences between topological hypotheses for this family, there are several clades that are consistently recovered in phylogenetic analyses. Two of these clades are the subfamilies Ictiobinae, containing the deep-bodied genera *Carpiodes* and *Ictiobus*, and the slender-bodied Catostominae, which contains the tribes Catostomini (containing the
genera *Catostomus, Chasmistes, Deltistes, and Xyrauchen*), Erimyzonini (containing the genera *Erimyzon* and *Minytrema*), Thoburniini (containing the genera *Hypentelium* and *Thoburnia*), and Moxostomatini (containing the genus *Moxostoma*). Two additional subfamilies, Cycleptinae and Myxocyprininae, have been proposed; however, they have not been consistently recovered as monophyletic groups.

**Figure 1** Previously proposed hypotheses for the phylogeny of Catostomidae. Only subfamilial and generic names are used to emphasize how these phylogenies compare to our topological hypotheses.
The occurrence of hybridization (Becker 1983) and tetraploidy (Chen and Mayden 2012) in this family has burdened phylogenetic reconstruction of its evolutionary history using molecular data (Bart et al. 2010; Chen and Mayden 2012). To avoid complications related to total diversity coverage among alleles, DNA amplification, and the sequencing of paralogs, the reconstruction of the catostomid phylogeny has been largely based on mitochondrial DNA (mtDNA) and phenotypic characters (Chen and Mayden 2012). This is problematic as mitochondrial data are prone to saturation and may not accurately recover phylogenetic history, particularly given the possible role of hybridization in catostomids (Bart et al. 2010).

Miller (1959) and Smith (1992) proposed phylogenies for Catostomidae based on morphological data. These hypotheses identified the subfamily Catostominae but disagreed on relationships of the remaining taxa (Fig. 1). The catostomid phylogeny has been reconstructed numerous times using mtDNA (Harris and Mayden 2001; Doosey et al. 2010; Clements et al. 2012). These studies also grouped catostomines into a single clade; however, relationships between the subfamilies of the Catostomidae have remained discordant, rearranged in nearly every possible combination. The use of nuclear DNA (nDNA) and isozymatic (coded by nuclear loci) data to resolve the phylogeny of the Catostomidae has been limited (Ferris and Whitt 1978; Bart et al. 2010; Chen and Mayden 2012; Clements et al. 2012) and came to conflicting conclusions on the topological arrangement of taxa within this family, especially for subfamilial relationships.
Our study included 43 catostomids as well as 11 cypriniform out-group species. We inferred phylogenies using 267 anchored loci, obtained through anchored hybrid enrichment (AHE; Lemmon et al. 2012; Stout et al. 2016). These data were analyzed using three methods of phylogeny reconstruction: maximum likelihood, Bayesian inference, and a coalescent-based gene tree summary method. The aim of this study is to produce a ubiquitously well-supported phylogeny, paying particular attention to resolving the subfamilial relationships of this family by filtering our data using a method of profiling phylogenetic informativeness.

**Materials and Methods**

**Taxon Sampling**

We included 43 in-group taxa and 11 cypriniform out-group taxa in the analyses (Appendix: Table 3). Data for the out-group taxa and 11 in-group taxa were obtained from Stout et al. (2016). Data for the remaining 32 catostomids were generated in this study using the probe kit designed by Stout et al. (2016). We selected species based on the availability of high quality tissue samples and tried to ensure that at least one species from every genus within the Catostomidae was included, increasing taxon sampling for species rich genera.

**Data Collection**

Data for the present study were generated at the Center of Anchored Phylogenomics ([www.anchoredphylogeny.com](http://www.anchoredphylogeny.com)) at Florida State University. Genomic
DNA was fragmented using a Covaris E220 Focused-ultrasonicator with Covaris microTUBES. A protocol, modified from Meyer and Kircher (2010), was used in library preparation and indexing. Then, indexed samples were pooled in equivalent amounts, which was subsequently enriched using the Vertebrate v.1 kit (Agilent Technologies Custom SureSelect XTd). The Vertebrate v.1 kit targets 512 conserved regions, most of which are located within exons. Sequencing of the enriched pool for this study was done at the Translational Science Laboratory in the College of Medicine at Florida State University using a 1 PE100 Illumina HiSeq 2000 lane.

An AHE pipeline, developed by Lemmon et al. (2012) was used to assess paralogy of loci. In short, this pipeline uses an individual as a reference based on its capture efficiency. Following, consensus sequences for each locus were aligned to the sequence of the reference individual. Homologs are subsequently established algorithmically by searching for individuals with the greatest sequence similarities to the reference individual. Once the homolog set is established, these sequences are removed for the candidate pool of sequences. This process is repeated iteratively until all homolog sets have been constructed.

**Partitioning/Substitution Model**

267 loci were concatenated into one contiguous sequence, summing to 399,329 base pairs. The most appropriate partitioning scheme for the loci and best substitution model for each locus was determined using PartitionFinder v2.0 (Lanfear et al. 2016). To increase the speed of the analysis, we used PartitionFinder’s RAxML command line
option (--raxml; Stamatakis 2006). We used the BIC metric to identify the partitioning scheme used in subsequent analyses following the recommendations of the PartitionFinder2 user manual (Lanfear et al. 2016).

Phylogenetic Analyses

We analyzed the concatenated data through a maximum likelihood approach using RAxML-HPC2 v8.2.10 (Stamatakis 2014), available on the CIPRES Science Gateway website (Miller et al. 2010). The partitioning/substitution model scheme obtained from PartitionFinder2 was included as a mixed/partitioned model input file (-q). Every partition in this analysis was assigned a variant of the general time reversible (GTR) model of base pair substitution, since the RAxML command line option in PartitionFinder2 only produces a substitution/partitioning scheme with GTR derivatives (Lanfear et al. 2016). We assessed nodal support by performing 100 bootstrap replicates.

We performed Bayesian analyses using MrBayes v3.2.6 (Ronquist and Huelsenbeck 2003) implemented in Geneious v9 (http://www.geneious.com, Kearse et al. 2012). Two runs were used for the entire, partitioned data set with four chains per run. The analyses were run for 100,000,000 Markov chain Monte Carlo (MCMC) generations. The substitution models for each partition, obtained from the PartitionFinder2 analysis, included only derivatives of GTR. Trees were sampled and logged every 2,000 generations. Initial trees obtained were discarded following a 25% burn-in threshold to ensure only trees with posterior probabilities representing the plateau of the distribution
were used to generate clade credibility values. The remaining trees were combined to obtain support values for the nodes on our tree.

**Coalescent-Based Method of Phylogeny Reconstruction**

A phylogenetic hypothesis recovered from a coalescent-based method of species tree reconstruction was obtained using ASTRAL-II (Mirarab and Warnow 2015). This method requires an input file of individual gene trees. Gene trees included were generated using RAxML on the CIPRES Science Gateway. Models of base pair substitution used for each gene tree reconstruction were assigned using the results of the PartitionFinder analysis mentioned above.

**Topological Comparison**

To ensure phylogenetic hypotheses generated by these analyses were significantly better at explaining the data than previously proposed hypotheses (Fig. 1), we compared the fit of each hypothesis to the data using the maximum likelihood-based Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999) implemented in PAUP* 4.0b10 (Swofford 2003). For these tests, the model of base pair substitution was set to GTR. Across-site rates of substitution were allowed to vary with 4 gamma categories and a shape parameter of 0.5. State frequencies were determined empirically and 1,000 bootstrap replicates were performed. Significance of fit was reported using $P$-values.
Phylogenetic Informativeness and Data Filtration

To resolve the conflict between trees recovered from concatenation and summary method approaches, we built data subsets based on phylogenetic informativeness of individual loci. We followed Dornburg et al. (2017), creating data sets stratified by their ability to resolve nodes at targeted time intervals (Fig. 2). We used PhyDesign (López-Giráldez and Townsend 2011) to analyze the phylogenetic informativeness of individual loci within the data set. This method produces a distribution of phylogenetic informativeness for each locus over a given time scale based on substitution rates of sites within a locus (Townsend et al. 2012). This distribution can then be used to select loci for resolving targeted nodes.

To obtain a phylogenetic informativeness profile for the data set, a concatenated sequence file with a partitioning scheme command block (partitioned by locus) as well as a time-calibrated chronogram were input into PhyDesign. The included chronogram was generated in BEAST2 (Bouckaert et al. 2014) using the GTR+Γ4 model of base pair substitution. Four constraints (fossil calibrations) were included to guide the search through tree space and provide a more accurate estimation of divergence times. These constraints included: a fossil catostomid (~61.7 Mya; Wilson 1980) to constrain the monophyly of all catostomids, a fossil ictiobine (~33.9 Mya; Smith 1992) to constrain an Ictiobus + Carpiodes clade, a fossil Ictiobus spp. (~15 Mya; Cavender 1986) to constrain all species within Ictiobus, and a fossil of a catostomine (~5.3 Mya; Smith1992), which prevented the inclusion of Cycleptus and Myxocyprinus with the well-established Catostominae. The HyPhy program as implemented in PhyDesign was used.
Figure 2 A representation of the phylogenetic informativeness of loci within our data set. A is a chronogram of the taxa included in this study. B is the same chronogram, but with a gray box indicating the time interval in which the early divergence of major, extant catostomid lineages probably occurred. C depicts the distribution of phylogenetic informativeness of loci within our data set over time. D represents the distribution of phylogenetic informativeness of loci within our data set over the time interval, in which the divergence of major, extant catostomid lineages probably occurred. The height of each locus’ informativeness distribution represents that locus’ relative ability for resolving nodes at a given time (e.g. the spikes in informativeness seen close to the y-axis in C represent loci that mutate quickly and are therefore useful when resolving relatively contemporary nodes).
to profile phylogenetic informativeness. We used the relative informativeness of each locus to construct reduced, concatenated data sets, including progressively more loci. The smallest dataset included only the 5 most informative loci from the data set. As we increased the number of loci in the data set, we were selecting from a pool of less informative loci, since the most informative loci had already been used to create the preceding, smaller data set. This method produced 14 new data sets; 5, 10, 25, 50, 75, 100, 133, 167, 192, 217, 242, 257, 262, and 266 loci.

These data sets were analyzed individually using two approaches: maximum likelihood analyses using RAxML and a coalescent-based species tree method using ASTRAL-II (Mirarab and Warnow 2015). Each subset was first analyzed by PartitionFinder2 using the RAxML command line option to obtain subset specific substitution model/partitioning schemes. The new subsets were analyzed by RAxML using their respective partitioning scheme. Each partition was assigned a derivative of the GTR model of base pair substitution and 100 bootstrap replicates were performed per analysis to assess nodal support.

For our gene tree summary method analyses, gene trees were generated for the loci that were included in each data subset. These gene trees were generated by RAxML using derivatives of the GTR model of base pair substitution. Individual gene trees were then pooled into a single tree file and analyzed by ASTRAL-II. We visually compared trees produced by both methods to assess how topologies and support values changed and how each inference method, concatenation and coalescent-based species tree, performed when progressively noisier data were included.
Robinson-Foulds Distances

The unweighted Robinson-Foulds distance metric (Robinson and Foulds, 1981) was used to elucidate how discordant individual gene trees used in the gene tree summary method analysis were to each other as well as how discordant the gene trees were to the recovered species tree(s) using the complete data set. The Robinson-Foulds metric measures the distance (i.e. discordance) of two tree topologies by counting the number of times a clade appears in one tree, but not another tree to which it is being compared. For every instance in which a clade is present in one tree, but not the other, the score of the Robinson-Foulds metric for the comparison is increased by 1. A score of 0 indicates that two trees are identical in their branching pattern, whereas a score of \(2(n - 2)\), where \(n\) is the number of tips in a tree, indicates that two trees are as discordant as they can possibly be.

We chose to use unweighted Robinson-Foulds distances as a metric of gene tree-gene tree/gene tree-species tree discordance following the recommendation of Kuhner and Yamato (2015), due to the presence of particularly short branches near the base of the catostomid phylogeny and the observation of frequent discordance of gene trees within our data set made \textit{a priori}. Measurements of gene tree discordance were obtained using the multiRF function (see Table 4 in the appendix for the R code used for obtaining Robinson-Foulds distances) included in the phytools package (Revell, 2011) using the R programming software (R Development Core Team, 2013).
Results

The complete data set for this study included 267 AHEs with 43 in-group taxa and 11 out-group taxa. After aligning sequences and trimming flanking regions, the concatenated data set included 399,329 nucleotide sites per taxon and a summation of 21,563,766 base pairs. Lengths for individual loci ranged from 118 to 3,273 base pairs with a mean sequence length of 1,495.7 base pairs. The complete data set included 65,090 parsimony informative (PI) sites with the number of PI sites for each locus ranging from 1 to 1,293 and an arithmetic mean of 274.6.

Phylogenetic Analyses of the Total Data Set

The monophyly of Catostomidae was universally supported (BS = 100, PP = 1, ASTRAL = 100; Fig 3). The two subfamilies that appear consistently throughout the literature, Catostominae and Ictiobinae, were also supported in all analyses (BS = 100, PP = 1, ASTRAL = 100; Fig 3). Tribes that have been described for the subfamily Catostominae, Moxostomatini (BS = 100, PP = 1, ASTRAL = 100), Catostomini (BS = 100, PP = 1, ASTRAL = 100), Thoburniini (BS = 100, PP = 1, ASTRAL = 100), and Erimyzonini (BS = 100, PP = 1, ASTRAL = 100), likewise received high support values (Fig. 3).

The maximum likelihood analysis of the concatenated data set produced a novel phylogenetic hypothesis, \((\text{Myxocyprinus (Cycleptus (Ictiobinae plus Catostominae)))})\) (Fig. 3). Bootstrap values for terminal branches were predominantly well supported with scores ranging from 51 to 100 and an arithmetic mean of 91.4 and a median and mode of
Figure 3 The two topologies recovered from our concatenation methods (RAxML and MrBayes) and our gene tree summary method (ASTRAL-II). Colored boxes are used to indicate the taxa that belong in each subfamily of the Catostomidae. Nodal support for our RAxML analysis are indicated the width and pattern of branches on the tree (left). Thick branches represent relationships that received a bootstrap score (BS) ≥ 95, thin branches represent relationships that received a BS = 80-94, and dashed branches represent relationships that received a BS < 80. On our Bayesian tree (left) nodes that did not receive a posterior probability (PP) of 1 were annotated by open circles. Both open circles represent relationships where PP = 0.82. Branch width and length was also used for our ASTRAL tree (right). Thick branches represent relationships that received an ASTRAL score ≥ 95, thin branches represent relationships that received an ASTRAL score = 80-94, and dashed branches represent relationships that received an ASTRAL score < 80.
100. Weak support values (BS < 80) were found exclusively between species within Moxostomatini.

    Notably, not every genus within the Catostomidae was recovered as monophyletic. *Thoburnia* and *Hypentelium* formed a strongly supported clade (tribe *Thoburniini*; BS = 100) with *Thoburnia* paraphyletic with respect to a monophyletic *Hypentelium* clade (BS = 100). Similarly, the genus *Catostomus* was paraphyletic with *Chasmistes liorus* sister to *Catostomus catostomus*, *Xyrauchen texanus* sister to a clade containing several *Catostomus* species (*C. latipinnis*, *C. bernardini*, *C. cahita*, *C. leopoldi*, *C. wigginsi*, and *C. clarkii*), and *D. luxatus* sister to *C. occidentalis*. All instances where a sister species relationship suggested the paraphyly of *Catostomus* were well supported (BS = 100).

    The Bayesian inference analysis produced a phylogenetic hypothesis with the same topology as the maximum likelihood tree. Every node on the Bayesian tree had a posterior probability (PP) of 1 with the exception of the node connecting *Moxostoma arriommum* and *M. cervinum* (PP = 0.82) and the node connecting *M. duquesnei* to a clade containing *M. lachneri* and *M. poecilurum* (PP = 0.82).

    Lastly, the coalescent-based species tree analysis (ASTRAL-II) produced a hypothesis of subfamilial relationships, identical to the hypothesis proposed by Miller (1959; Fig. 1). Although the monophyly of Catostomidae, Ictiobinae, and Catostominae were well supported (ASTRAL = 100), the sister taxon relationship of *Myxocyprinus asiaticus* and *Cycleptus elongatus* (ASTRAL = 61) and the sister taxon relationship between Ictiobinae and Catostominae (ASTRAL = 7) were not well supported.
As in our maximum likelihood and Bayesian analyses, taxonomic groups did not always appear as monophyletic in the ASTRAL species tree. A monophyletic *Hypentelium + Thoburnia* clade was once again well supported (ASTRAL = 100), as was the paraphyly of *Thoburnia* (ASTRAL = 96), crowned by a monophyletic *Hypentelium* clade (ASTRAL = 100). The genera *Chasmistes, Xyrauchen, and Deltistes* also appeared within a well-supported *Catostomus* clade (ASTRAL = 100), corroborating the paraphyly of *Catostomus*. The sister species relationships of non-*Catostomus* species to *Catostomus* species were identical to the relationships found in the maximum likelihood and Bayesian analyses and well supported (ASTRAL = 100).

The relationships between taxa in *Carpiodes* were consistent between the three tree inference approaches used; however, relationships between species within *Ictiobus* were discordant. From the concatenation methods, *I. cyprinellus* was recovered as the sister species to *I. niger* (BS = 85, PP = 1), whereas the summary method recovered *I. cyprinellus* as the sister species to *I. bubalus* (ASTRAL = 91).

Relationships between taxa within Catostominae were not always congruent across analyses. Although the monophyly of *Moxostoma* was well supported in every analysis (BS = 100, PP = 1, ASTRAL = 100), half of the nodes within *Moxostoma* received poor nodal support values (ASTRAL ≤ 78) with a mean value of 74.8. The relationships between taxa in Catostomini were identical to the results from the Bayesian and maximum likelihood analyses, with the exception of a clade containing *C. latipinnis, C. clarkii, C. wigginsi, C. bernardini, C. cahita,* and *C. leopoldi*. In this clade, *C. cahita* and *C. leopoldi* were sister species (ASTRAL = 74) followed by successive sister species
relationships to *C. bernardini* (ASTRAL = 56), *C. wigginsi* (ASTRAL = 100), *C. clarkii* (ASTRAL = 89), and *C. latipinnis* (ASTRAL = 82). Although nodal support values were not exceptionally poor for this clade, the ASTRAL-II analysis was unable to reliably support the relationship for these species.

**Topological Comparison**

The comparison of alternative topological hypotheses (Fig.1) for the catostomid phylogeny revealed that the phylogeny recovered for our maximum likelihood and Bayesian analyses fit the data set significantly better (*p* < 0.05*) than all previously proposed hypotheses, with the exception of Harris and Mayden (2001) and Miller (1959). The Harris and Mayden phylogeny, inferred using large ribosomal subunit sequence data, was approximately 39.8 log likelihood units worse that our ML/Bayesian hypothesis (*p* = 0.106). Our ASTRAL tree and Miller’s tree were approximately 43.3 log likelihood units worse than the best hypothesis (*p* = 0.071). Although we cannot reject these hypotheses given this data set, it is worth noting that the difference in log likelihoods between the best scoring hypothesis (our ML/Bayesian topologies) and our ASTRAL/Miller’s tree was nearly significant.

**Phylogenetic Informativeness and Data Filtration**

The use of concatenation and a gene tree summary method recovered two alternative topologies: one in which Myxocyprininae diverged from all other catostomids at the basal-most node, followed by a secondary divergence of Cycleptinae from the other
catostomids, all of which was crowned by an Ictiobinae+Catostominae clade (concatenation) and a phylogeny in which Myxocyprininae was recovered as the sister taxon to Cycleptinae (gene tree summary method). In order to lend support to one hypothesis over the other, we employed a data set filtration approach to tease out what may be causing the recovery of conflicting trees.

We generated 14 additional data set subsets, which were analyzed using RAxML and ASTRAL-II, recovering 28 additional species trees (Fig. 4). The smallest data set (5 and 10 loci), through the gene tree summary method, recovered a topology, in which (Cycleptinae (Myxocyprininae (Ictiobinae plus Catostominae))) (Fig. 4). Using data sets including 25 to 100 of the most informative loci, a topology was recovered, in which (Myxocyprininae (Cycleptinae (Ictiobinae plus Catostominae))). Trees recovered from the analysis of more inclusive data sets (133-266) converged on a single topology, ((Myxocyprininae plus Cycleptinae) (Ictiobinae plus Catostominae)). Relationships recovered from the use of the gene tree summary method were often poorly supported with the sister subfamilial relationship between Ictiobinae and Catostominae being consistently recovered, although poorly supported.

Trees recovered from the concatenation method did not converge as quickly and, relative to trees recovered from the gene tree summary method, had greater support values. A topology in which (Cycleptinae (Myxocyprininae (Ictiobinae plus Catostominae))) was recovered when analyzing the smallest data set with the 5 most informative loci. Analyzing the data set with the 10 most informative loci recovered a topology, in which (Myxocyprininae (Ictiobinae (Cycleptinae plus Catostominae))). Data
Figure 4 Species trees recovered from the analysis of data sets generated by data filtration. Tips of the phylogenies are annotated by depictions of major lineages within Catostomidae. Branch widths and patterns are used to indicate nodal support values; thick branches represent relationships that received a nodal support value of ≥ 95 (BS) or 0.95 (ASTRAL); thin lines indicate relationships that received nodal support of 80-95 (BS) or 0.8-0.95 (ASTRAL); dashed lines represent relationships that received a score of ≤ 79 (BS) or 0.79 (ASTRAL).
sets containing the 25-217 of most informative loci recovered a topology in which
(Myxocyprininae (Cycleptinae (Ictiobinae plus Catostominae))). The data set containing
242 of the most informative loci recovered a topology, in which ((Myxocyprininae plus
Cycleptinae) (Ictiobinae plus Catostominae)). Thereafter, more inclusive data sets
recovered a topology, in which (Myxocyprininae (Cycleptinae (Ictiobinae plus
Catostominae))) (the same topology that was recovered when 25-217 loci were analyzed).
Nodal support statistics for trees recovered from analyzing data set subsets using
concatenation and the gene tree summary method reconstruction approaches are listed in
Table 1.

Table 1. Summary statistics of nodal support values representing the divergences of subfamilies within Catostomidae. Nodal support values were obtained from trees recovered from the analysis of data set subsets using maximum likelihood (RAxML) and coalescent-based species tree (ASTRAL-II) reconstruction approaches.

<table>
<thead>
<tr>
<th></th>
<th>Concatenation</th>
<th>Gene Tree Summary Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>Q1</td>
<td>56</td>
<td>64</td>
</tr>
<tr>
<td>Q2</td>
<td>99</td>
<td>74</td>
</tr>
<tr>
<td>Q3</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Maximum</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>IQR</td>
<td>44</td>
<td>36</td>
</tr>
<tr>
<td>Range</td>
<td>99</td>
<td>57</td>
</tr>
<tr>
<td>Sum</td>
<td>3301</td>
<td>3236</td>
</tr>
<tr>
<td>Mean</td>
<td>78.6</td>
<td>77</td>
</tr>
<tr>
<td>Median</td>
<td>99</td>
<td>74</td>
</tr>
<tr>
<td>Mode</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Robinson-Foulds Distances

The comparison of the gene trees for each locus included in our data set to the two species trees obtained when analyzing the entire data set revealed that discordance of gene trees to the species trees did not differ considerably based on inferencing approach (Fig. 5); however, a Robinson-Foulds pairwise comparison revealed that individual gene trees were exceptionally discordant to each other (Fig. 6). The distribution of Robinson-Foulds distance for the pairwise comparison of gene tree topologies was left-skewed with many distance scores near the maximum distance score of 104, indicating a great degree of gene tree discordance.

![Figure 5](image)

**Figure 5** A distribution of Robinson-Foulds distances when comparing individual gene trees to the species tree topologies recovered when analyzing the entire data set. Possible values for this analysis ranged from 0 to 104.
Figure 6 A distribution of Robinson-Foulds distances when comparing individual gene trees to other gene trees included in our data set. Possible values for this analysis ranged from 0 to 104 (indicated by the red bar on the right of the graph).

Discussion

It has been notoriously difficult to establish a robust phylogenetic hypothesis for the Catostomidae using molecular data. The challenge stems from the tetraploid origin of the family, the age of the group, and the prevalence of hybridization among the species. We provide the first reconstruction of the phylogeny using a genomic data set, taking these problems into account; however, genomic data sets pose their own sets of problems, and may still result in phylogenies with poorly-supported nodes, conflicting topologies from studies using similar assemblages of taxa, and disparate topologies recovered from different tree reconstruction approaches (Shen et al. 2017). We employed data filtration
approaches to explore the impact of noisy data on our results and to arrive at a robust phylogenetic hypothesis for the Catostomidae.

Analysis of the data set recovered two disparate hypotheses of relationships of the major clades in the family, depending on the method of phylogenetic reconstruction used. The differences in topologies were largely manifested in the phylogenetic position of Cycleptinae as sister to Myxocyprininae or sister to Ictiobinae plus Catostominae. Additionally, we observed minor differences within the clades containing the genera *Catostomus*, *Moxostoma*, and *Ictiobus*.

We employed an approach that combined profiling phylogenetic information with removal of noisy data to interrogate nodes of interest. We integrated the methods of Dornburg et al. (2017) and Shen et al. (2017) by profiling phylogenetic informativeness of loci within the entire data set and reconstructing the species trees of individual data sets stratified by their degree of informativeness. This was done to create data subsets with sufficient phylogenetic signal to resolve nodes representing the split of subfamilies while removing confounding, homoplastic data. Results of the analyses revealed that tree topologies recovered from the gene tree summary method approach converged on a single topology when data sets included 133 or more loci. This topology was identical to the one recovered when analyzing the entire data set through this method (Fig. 4). Topologies recovered from the analysis of smaller data sets using the gene tree summary method include a tree, in which (Cycleptinae (Myxocyprininae (Ictiobinae plus Catostominae)))) (5-10 loci) as well as a topology identical to our maximum likelihood/Bayesian inference tree (using the entire data set), where (Myxocyprininae
The hypothesized sister subfamily relationship of Ictiobinae and Catostominae was corroborated by all species trees recovered from the data set filtration analyses when using the gene tree summary method; however, the nodes connecting the catostomid subfamilies were often poorly supported. The poor support values could be a result of the use of inaccurate substitution models (considering we only used GTR derivatives), which has been demonstrated empirically to reduce the accuracy of phylogenetic reconstruction. This can engender gene tree estimation error, resulting in poorly supported or erroneous gene trees (Lemmon and Moriarty, 2004; Molloy and Warnow 2017). Loci within the entire data set represent a diverse assemblage of gene regions, which likely require a diverse set of substitution models; however, a partitioning scheme/substitution model analysis that considered all models of base pair substitution on a data set of this magnitude was not feasible due to the large computational burden.

Trees from the analysis of data set subsets using the concatenation approach did not converge as quickly, recovering four unique topologies as progressively more loci were included in the analyses until the topologies finally converged when analyzing 257 or more loci (Fig. 4). Unlike the trees recovered from the analysis of data set subsets using the gene tree summary method, trees recovered when using a concatenation approach included sister subfamilial relationships that were supported with confidence. Additionally, the Ictiobinae plus Catostominae sister subfamilial relationship was nearly consistently recovered, with data sets of 25 of more loci recovering this relationship with confidence. One explanation for the erratic nature of topological arrangements when data
sets are analyzed using concatenation relates to the method of data set subset
collection; the smallest of the data sets subsets included in our data set filtration
approach included the most informative loci within our entire data set. As data sets
increased in size, we were selecting from a pool of progressively less informative loci.
Although it is expected that analyses will converge on a single topology as progressively
more data is included in an analysis, this assumes that loci are being added randomly and
not from a pool of loci that are known to be less informative or highly discordant a priori.

In summary, both species tree reconstruction approaches had their own respective
advantages and disadvantages: topological recovery of the concatenation approach was
erratic, yet subfamilial relationships were often supported with confidence. Conversely,
species trees recovered from the gene tree summary method were relatively more
topologically consistent, but no sister subfamilial relationships were supported
confidently. To resolve this issue, we used the Robinson-Foulds distance metric
(Robinson and Foulds, 1981) to elucidate how discordant the individual gene trees used
in the gene tree summary method were to each other as well as how discordant the gene
trees were to the two recovered species tree topologies using the complete data set.

Gene tree discordance, potentially indicative of incomplete lineage sorting, has
been shown to reduce the accuracy of tree reconstruction methods, depending on the
degree of gene tree discordance. Molloy and Warnow (2017) demonstrated that gene tree
summary methods out-compete concatenation methods when incomplete lineage sorting
is moderate to high, but perform poorly when incomplete lineage sorting is very low or
extremely high (Molloy and Warnow 2017). Additionally, Molloy and Warnow (2017)
noted that, when analyzing a simulated dataset, not only do concatenation methods outcompete summary methods when incomplete lineage sorting is extremely high or low, but also perform better than summary methods when gene tree estimation error is high.

Due to the particularly high degree of gene tree discordance within our data set (Fig. 6), following the recommendation of Molloy and Warnow (2017), we support the phylogenetic hypothesis for the evolutionary history of the Catostomidae obtained from our maximum likelihood/Bayesian inference analyses, where (Myxocyprininae (Cycleptinae (Ictiobinae plus Catostominae))) (Fig. 3).

**Relationships among the Catostomidae**

A key objective of this study was to elucidate the relationships of subfamilies within the Catostomidae, a question that has been historically challenging to address. Based on the results of these analyses, we accept the topology recovered from the concatenation method of phylogenetic tree reconstruction as the best estimate of the phylogeny of Catostomidae given these data (Fig. 7). The phylogeny supported by this study is a novel hypothesis for the evolutionary history of the Catostomidae. The taxa within each subfamily of the Catostomidae have remained relatively consistent over time with the exception of *Myxocyprinus* being placed in Cycleptinae in Miller’s (1959) phylogenetic reconstruction. Previous reconstructions have come to highly discordant conclusions, arranging the subfamilies into nearly every permutation possible. We find strong support for the subfamily relationships recovered in our phylogeny, providing a new perspective on the evolution of the Catostomidae. Further, this phylogeny lends
Figure 7 The phylogeny of Catostomidae supported by this study. Thick branches represent nodes that received a bootstrap score (BS) of ≥ 95, thin branches represent nodes that received a BS of 80-94, and dashed branches represent nodes that received a BS of < 80. Nodes that received less than 1 for a posterior probability from our Bayesian analysis were annotated with an open circle (PP = 0.82 for both).
support to studies that have suggested the non-monophyly of genera within Catostomidae (Chen and Mayden 2012).

Several previous reconstructions of the catostomid phylogeny have placed *Myxocyprinus asiaticus* as the sister taxon to all other extant catostomids (Harris and Mayden 2001) or as the sister taxon to *Cycleptus elongatus*, forming a clade that branches at the root node from the other extant catostomids (Miller 1959; Chen and Mayden 2012). Doosey et al. (2010) recovered *Myxocyprinus* as sister taxon to the ictiobines. Our phylogenetic hypothesis for the Catostomidae places *Myxocyprinus* as the sister taxon to all extant catostomids, a relationship that received strong nodal support from our maximum likelihood and Bayesian inference analyses.

The phylogenetic position of *Cycleptus*, like *Myxocyprinus*, has been difficult to resolve. Ferris and Whitt (1978) placed *C. elongatus* as the sister taxon to all other catostomids, but did not include *Myxocyprinus* in their study. Often paired as the sister taxon to *Myxocyprinus* (Miller 1959; Smith 1992; Harris and Mayden 2001; Chen and Mayden 2012; Clements et al. 2012), *Cycleptus* has also been recovered as the sister taxon to Catostominae (Harris and Mayden, 2001), an Ictiobinae + Myxocyprininae clade (Doosey et al. 2010), and an Ictiobinae + Catostominae clade (Ferris and Whitt 1978). The phylogeny recovered from this study challenges the sister relationship of *Cycleptus* and *Myxocyprinus*, supporting instead the sister relationship of *Cycleptus* to a clade containing Ictiobinae and Catostominae. This relationship was well supported (BS = 83; PP = 1) and supports the hypotheses of early divergences of *Cycleptus* and *Myxocyprinus*. 
(Ferris and Whitt 1978; Harris and Mayden 2001). Additionally, this topology results in the monophyly of North American catostomids.

Ictiobinae, a subfamily containing the genera *Ictiobus* and *Carpiodes*, was recovered as a strongly supported clade in this study (BS = 100; PP = 1; ASTRAL 1). Throughout the literature, this clade has been consistently recovered as monophyletic; however, its sister taxon relationship has been contested. Doosey et al. (2010) recovered a phylogeny in which Ictiobinae was the sister taxon to *Myxocyprinus*. Others have recovered Ictiobinae as the sister taxon to Catostominae (Miller, 1959), a clade containing Myxocyprininae and Cycleptinae (Clements et al. 2012; Chen and Mayden 2012), or a clade containing Cycleptinae and Catostominae (Harris and Mayden 2001). Smith (1992), using phenotypic characters, placed Ictiobinae as the sister taxon to all other extant catostomids. In this study, Ictiobinae was recovered as the strongly supported sister taxon to Catostominae (BS = 100; PP = 1), supporting Miller’s (1959) and Ferris and Whitt’s (1978) placement of Ictiobinae on the catostomid phylogeny. Species-level relationships for the ictiobines largely resembled the relationships recovered from previous studies (Smith 1992; Doosey et al. 2010). Discrepancies exist between the topologies recovered by Doosey et al. (2010) and our own topology; however, the topologies recovered from Doosey et al. (2010) conflict with each other on the relationship of the ictiobines, preventing the comparison of our findings. The relationships of species within *Carpiodes* from this study were identical to that of Smith’s (1992) reconstruction. Unlike Smith’s (1992) phylogeny, our study placed *I. cyrinellus*
as the sister species to *I. niger*, whereas Smith recovered *I. cyprinellus* as being more closely related to *I. bubalus* than to *I. niger*.

Catostominae, the largest subfamily of the Catostomidae, was recovered as a strongly supported monophyletic clade, placed as the sister subfamily to Ictiobinae (BS = 100; PP = 1). Reconstructions of the catostomid phylogeny have placed Catostominae as the sister taxon to a clade containing all other subfamilies (Harris and Mayden 2001; Doosey et al. 2010; Clements et al. 2012), the ictiobines (Miller 1959; Ferris and Whitt 1978), Cycleptinae (Harris and Mayden 2001), and a Cycleptinae + Myxocyprininae clade (Smith 1992). With the exception of tribal name discrepancies, no differences were observed between our recovery of Erimyzonini and previous recoveries of this tribe. We recovered a paraphyletic *Thoburnia*, which was crowned by a monophyletic *Hypentelium*. This same pattern was suggested at least twice before by Doosey et al. (2010) and Clements et al. (2012). Additional reconstructions have recovered a polyphyletic *Thoburnia* (Clements et al. 2012) as well as reciprocal monophyly of *Hypentelium* and *Thoburnia* clades (Smith 1992; Harris and Mayden 2001). Although an interesting finding, we are unable to make definitive statements as to whether or not *Thoburnia* should be considered a paraphyletic clade due to incomplete taxon sampling.

*Catostomus* was also recovered as a paraphyletic clade in this study with strongly supported sister species relationships between *Ch. liorus* and *C. catostomus*, *D. luxatus* and *C. occidentalis*, and *X. texanus* to a clade containing *C. latipinnis*, *C. bernardini*, *C. cahita*, *C. clarkii*, *C. wigginsi*, and *C. leopoldi*. The paraphyly of *Catostomus* has been suggested before by Harris and Mayden (2001), Doosey et al. (2010) and Chen and
Mayden (2012). Smith’s (1992) reconstruction of Catostomidae using 157 phenotypic characters is one of the few instances where *Catostomus* has been recovered as a monophyletic clade. Additionally, *Chasmistes*, *Deltistes*, and *Xyrauchen* have unique morphologies, likely a function of their atypical habitats. Historically, these unique morphologies have resulted in these taxa being classified as monotypic genera (Chen and Mayden 2012). Using a genome-scale data set, we recovered a robust phylogeny which strongly corroborates the paraphyly of *Catostomus*. Due to the multiple occasions in which *Catostomus* has been recovered as paraphyletic in previous reconstructions and the strong support of paraphyly from this study, considering the breadth of taxon sampling and amount of data within our data set, we conclude that *Deltistes*, *Chasmistes*, and *Xyrauchen* are junior synonyms of *Catostomus*, not autonomous genera. To regain taxonomic accuracy, we subsume *Deltistes*, *Chasmistes*, and *Xyrauchen* back into *Catostomus*.

**Conclusion**

The data set used in this study to reconstruct the evolutionary history of the Catostomidae provided the unique opportunity to use an assemblage of diverse genes found throughout the genome in inferring the phylogenetic relationships of the catostomids. Before recent advances in DNA sequencing technologies, reconstruction of this family’s phylogeny using genetic data was largely restricted to using mitochondrial DNA sequences. Additional reconstructions have used isozymatic data as well as phenotypic data. Each method has its pitfalls, given the approximate age of this family’s
origin (>61.7 MYA), the prevalence of hybridization and tetraploidy, as well as possible instances of phenotypic convergence and parallelism. Intrinsic properties of anchored hybrid enrichment allowed for the use of a genome-scale data set of conserved, nuclear DNA sequences to reconstruct this family’s phylogeny while taking these issues into consideration.

Attenuation of recalcitrant relationships on the Tree of Life was thought to be the product of transitioning to the use of genome-scale data sets for phylogenetic reconstructions. However, we are seeing that increasing the amount of data alone is not the answer to resolving many of these enigmatic relationships, demonstrated by phylogenies being published with poorly supported nodes and conflicted topologies despite being inferred using genomic data set. Instead, many are proposing that more attention needs to be paid to which data within a data set are being used to resolve challenging nodes, a process often referred to as data filtration. Although various metrics of data set filtration have been proposed, we implemented a method that profiles the distribution of phylogenetic informativeness of loci over time. This allowed us to systematically select loci that would be most appropriate for resolving nodes representing the split of subfamilies within the Catostomidae. This process also revealed that the summary method of species tree reconstruction appeared to be performing poorly relative to the concatenation approach used. This provided support for the selection of the topological hypothesis recovered from our concatenation methods of phylogenetic reconstruction as the best estimation of the phylogeny of the Catostomidae.
By using a genome-scale data set, we recovered a phylogeny with strong nodal support values representing the early divergence order of subfamilies within Catostomidae. This phylogeny also lends strong support to the historically supported clades Ictiobinae, Catostominae, Moxostomatini, Catostomini, Thoburniini, and Erimyzonini. The relationships of taxa within Catostomini have been elusive, with the monophyly of *Catostomus* being both corroborated and refuted by prior reconstructions of this family’s phylogeny. Our study recovered *D. luxatus*, *Ch. liorus*, and *X. texanus* as strongly supported sister species to species within *Catostomus*, validating the paraphyly of *Catostomus*. Due to the magnitude of the data set used, the breadth of taxa within Catostomini included in this study, and the overwhelmingly strong support value for the species-level relationships found within Catostomini, we find that *Deltistes*, *Chasmistes*, and *Xyrauchen* are not appropriate generic names, but are rather synonyms of *Catostomus*. Therefore, we propose that these names be subsumed into *Catostomus* to restore the monophyly of this genus.
CHAPTER 2

PATTERNS OF MOLECULAR EVOLUTION WITHIN A FAMILY OF TETRAPLOID, FRESHWATER FISH (CYPRINIFORMES: CATOSTOMIDAE)

Introduction

Whole genome duplication (WGD) events are hypothesized to play an important role in molecular and phenotypic evolution, speciation, and shaping the architecture of the genome (Meyer and Van de Peer 2005; Volff 2005; Crow et al. 2006). Polyploidy, resulting from WGD events, is markedly prominent in the plant kingdom, but is relatively rare in animals, with the exception of freshwater, ray-finned fishes and amphibians (Mable et al. 2011). This pattern of polyploidy has been thought to result from the increased likelihood of producing unreduced gametes by ectotherms, which can be induced by temperature shock, as well as the aptitude of fish and amphibians to produce viable offspring through hybridization and polyspermy (Mable 2004). These suppositions become even more appropriate, considering virtually all known polyploid fishes and amphibians spawn in freshwater environments, which often fluctuate greatly in temperature over the course of a year and confine spawning events to small areas, increasing the likelihood that more than one sperm will fertilize an egg or that heterospecific gametes will encounter each other and fuse (Mable 2004).

In the mid-twentieth century, it was hypothesized that after a WGD event, redundant copies of genes experienced one of two fates: maintaining functionality or loss of functionality (known today as nonfunctionalization). Early work, which sought to develop a model for predicting the expected frequency at which gene copies are retained within the genome after a WGD event, predicted that the majority of duplicates are
quickly lost to nonfunctionalization (Ohno 1970). Only rarely and under specific conditions were duplicative genes thought to maintain expression and, by doing so, facilitate species diversification, increase molecular and morphological complexity, and provide the raw material, with which to craft biological novelties (Van de Peer et al. 2009). Although the biological philosophers of the 1970’s and 1980’s (e.g. Ohno [1970]; Li et al. [1981]; Nei and Roychoudhury [1973]) established a prodigious foundation on which subsequent research on the fate of duplicative genes could be based, it wasn’t until later that the counterpart of nonfunctionalization, “maintaining expression”, was specified into three of the current categories of redundant gene fates: retention, subfunctionalization, and neofunctionalization (Lynch and Conery 2000; Fig. 8). At present, as a result of advances in DNA sequencing technologies and phylogenetic techniques, researchers are paying an increasing amount of attention to studying the consequences of WGD and molecular evolution of polyploid taxa.

Despite the expansive literature on the rates and patterns of molecular evolution in polyploid plants (Saintenac et al. 2011; Blanc and Wolfe 2004; Walker et al. 2017), far fewer studies have sought to elucidate how molecular evolution of animals is affected by polyploidization. For example, a handful of lineages within Cypriniformes (the carps, minnows, loaches, and suckers) have experienced multiple, independent WGD events within the past 100 million years. Despite their ancient origin, Uyeno and Smith (1972) found that taxa within the Catostomidae, a family nested within Cypriniformes, have ubiquitously retained their tetraploid karyotype, although disomic inheritance has been
Figure 8 The potential fates of redundant gene copies after a WGD event. Segments colored light blue represent alleles performing the ancestral function. Segments colored red represent alleles that have acquired a new function through molecular divergence. Lines without colored segments represent the loss (or partial loss) of functionality of duplicative alleles.
achieved since their origin. This was demonstrated by the finding that catostomids possessed a chromosome number of \(2n = 100\) compared to most other cypriniforms, which have retained a chromosome number of \(2n = 50\). The retention of a tetraploid karyotype of the Catostomidae and their close relationship to diploid species makes the catostomids an ideal system on which to study the patterns of molecular evolution in polyploid animals.

The Catostomidae is a family of freshwater fish, commonly referred to as the suckers. This family is hypothesized to have evolved from a tetraploid ancestor sometime before or during the Paleocene (56-66 MYA; Wilson 1980) following an allopolyploidization event. This family currently includes 79 recognized extant species, all of which are endemic to North America, with the exception of *Myxocyprinus asiaticus* (endemic to the Yangtze River system in China) and *Catostomus catostomus* (found in North America and Siberia). There are four recognized subfamilies within the Catostomidae: Myxocyprininae (genus *Myxocyprinus*), Cycleptinae (genus *Cycleptus*), Ictiobinae (genera *Carpiodes* and *Ictiobus*), and Catostomininae (genera *Moxostoma*, *Minytrema*, *Erimyzon*, *Thoburnia*, *Hypentelium*, and *Catostomus*).

Early work on the catostomids examined retention of duplicate gene expression and enzyme polymorphisms to address questions about the evolution of duplicate genomes in this group. In their study that looked at gene expression of 20 enzymes, Ferris and Whitt (1977) found that an average of 47% of the enzymes examined were expressed as functional duplicates. Additionally, the “morphologically conserved” taxa (that is, taxa that resemble the hypothesized ancestral form [Amyzon-like] sensu Ferris and Whitt
[1977]; genera *Cycleptus*, *Ictiobus*, and *Carpiodes*) expressed 59% of their enzymes in duplicate while the “morphologically divergent/advanced” catostomines expressed only 42% of their enzymes in duplicate. They concluded that the morphologically conserved subfamilies (Ictiobinae and Cycleptinae) tend to retain duplicate gene expression more so than the catostomines. Following this conclusion, Ferris and Whitt (1977) hypothesized that phenotypically advanced, specialist lineages tend to lose duplicate gene expression more often than generalist species and that, after a rapid initial loss of duplicate expression, the unexpressed DNA is physically eliminated from the genome.

Regardless of the disparity in expression patterns between the evolutionary lineages of the Catostomidae, the catostomids retained gene expression for far more duplicate gene copies given their age than what is predicted by evolutionary models that assert that duplicate copies are randomly silenced through time by the accumulation of once “forbidden” mutations. When considering only genes that had maintained duplicate gene expression, it was found that, on average, 20.8% of these pairs were polymorphic (i.e. produced non-identical gene products) at one or both of their loci, suggesting only a small fraction of duplicated genes have undergone neo-, sub-, or nonfunctionalization since the allopolyploidization event that led to the evolution of the catostomids (Ferris and Whitt 1980).

In the present study, we address questions on the patterns of molecular evolution within the genomes of the tetraploid, freshwater suckers, the Catostomidae. In this study, we take a modern approach to Ferris and Whitt (1980), comparing the results of their studies, which used starch gel electrophoresis of enzymatic gene products to estimate the
abundance of genic polymorphisms and our study, using a genome-scale data set of nucleotide sequences. Furthermore, we aim to use our data set to address questions related to patterns of molecular evolution of redundant genes and how base pair substitutions are differentially accumulated within evolutionary lineages of the Catostomidae. The following questions are addressed herein:

- What is the extent of genic polymorphisms within the subfamilies and tribes of the Catostomidae?
  - Do the results of our study corroborate the findings of Ferris and Whitt (1980) that the ictiobines tend to have more polymorphic loci than the catostomines?
- Do evolutionary lineages within Catostomidae show differential accumulation of base pair substitutions?
- How often within our data set do taxa appear to have branch lengths greater than the average branch length at a given locus?

Herein, we present a study, which seeks to characterize the patterns of molecular evolution of a tetraploid genome. We use a data set comprised of 179 anchored hybrid loci obtained through anchored hybrid enrichment (AHE; Lemmon et al. 2012) to address questions related to disparities in the frequency of genic polymorphisms and differences in branch lengths among closely related taxa within polyploid families. In doing so, we hope to establish a starting point in studying the molecular evolution of this polyploid family’s genome, on which further research can expand.
Materials and Methods

Data Collection and Taxon Sampling

The genomic DNA that was used in the preceding chapter was also used for this study. Although sequence data for the taxa included in the following analyses were generated in a similar manner, a modification to the pipeline developed by Lemmon et al. (2012) for constructing homolog sets was used to generate a data set that reflected the ploidy levels of each taxon, a process referred to as phasing. This phasing process resulted in each of the 43 catostomids being represented by four alleles at each locus and each of the eleven outgroup taxa being represented by one (e.g. *Danio rerio*), two (e.g. *Cyprinion semiplotum*), or four (e.g. *Barbus barbus*) alleles. By phasing the data set, the number of loci included in this study was reduced from 267 to 179 and the number of OTUs was increased to 199.

Gene Tree Estimation

A species tree as well as individual gene trees were inferred using a maximum likelihood approach. Best fit models of base pair substitution for each locus were estimated using PartitionFinder 2 (Lanfear et al. 2016). The BIC metric was used to determine the best fitting substitution models (and partitioning scheme for the concatenated species tree) for the data set. Additionally, we used the RAxML command line option (--raxml; Stamatakis 2006) to increase the speed of analyses, since analyses ran without this modifier would have a run time on the order of weeks. Each locus was
imported into CIPRES Science Gateway (Miller et al. 2010) as a FASTA sequence file and analyzed using GARLI v2.01 (Zwickl 2006).

**Estimating the Extent of Polymorphic Loci**

Gene trees recovered from the maximum likelihood analyses were used to estimate the extent to which loci appeared as polymorphic for the subfamilies Ictiobinae, Myxocyprininae, and Cycleptinae as well as tribes within the subfamily Catostominae. In this sense, we define a locus as “polymorphic” for a given taxon if conspecific alleles were not recovered as monophyletic. Conversely, if conspecific sequences were monophyletic, we referred to this pattern as “monomorphic”. To determine the abundance of polymorphic loci for each taxon within our data set, individual gene trees obtained from the GARLI analyses were visualized using FigTree v 1.4.3 (Rambaut 2018). The number of loci that appeared as polymorphic were summed and were expressed in terms of percentages (frequency of taxon-specific, polymorphic loci = (number of polymorphic loci/total number of loci used in this analysis) x 100). The frequency at which the loci of taxa were recovered as polymorphic were averaged to achieve values to characterize the subfamilial and tribal categories mentioned above.

**Branch Length Comparisons**

To assess how base pair substitutions have differentially accumulated in each catostomid lineage, branch lengths were extracted from the concatenated species tree file using an R script comprised of functions (see Table 6 in the appendix for the script used)
from the following R packages: ape (Paradis et al. 2004), phylobase (Hackathon et al. 2017), ade4 (Dray and Dufour 2007), and adephylo (Jombart and Dray 2008). ANOVAs (with subsequent Tukey’s Honestly Significantly Different post-hoc tests) were used to determine if branch lengths differed significantly between the catostomid genera. These tests of significance were conduct in R v3.4.3 (Kite-Eating Tree; R Core Team 2013) using the built-in analysis of variance functions. Significance was determined by a $P$-value < 0.05 (adjusted).

**Comparison of Alleles with Greater than Average Branch Lengths**

In order to compare the abundances of alleles with branch lengths greater than the average branch length at a locus for each taxon, branch lengths were extracted from individual gene tree files using an R script (see Table 7 in the appendix for the script used) comprised of functions from the same four R packages listed in the preceding section. Extracted branch lengths from individual gene trees were then exported and combined into a super-matrix to visualize the disparity in branch lengths between lineages.

After removing branch lengths from our gene tree files and creating the branch length super-matrix, branch lengths were normalized relatively to all other alleles at each locus. By normalizing the data, we could determine how often each taxonomic group had branch lengths that were longer than the average branch length for each locus. The frequency of having longer than average branch lengths was found by summing all instances in which normalized branch lengths exceeded 0 for each taxonomic group.
Values for genera were generated by averaging the number of alleles found to have longer than average branch lengths for taxon belonging to each group.

**Results**

**Data Collection and Taxon Sampling**

The phasing process resulted in a data set of 179 AHE loci. These loci ranged in length from 143 base pairs to 3,039 base pairs with an arithmetic mean of 1,548 base pairs. Within the data set, 76.47% of sites were conserved (i.e. base pairs were identical at homologous sites) across taxa. Additionally, there were 47,517 parsimony informative sites (17.15%). The data set was analyzed using a maximum likelihood approach, recovering 179 gene trees to be used in subsequent analyses.

**Estimating the Extent of Polymorphic Loci**

Quantification and comparison of the prevalence of polymorphic loci between the subfamilies Ictiobinae and Cycleptinae and tribes within Catostominae corroborated Ferris and Whitt’s (1980) finding that the ictiobines tended to have a greater abundance of polymorphic loci than tribes within Catostominae (Fig. 9). However, we found that molecular polymorphisms were far more frequent within Ictiobinae, Cycleptinae, and Catostominae than previously found. With this data set, 91.05% and 72.88% of loci were found to be polymorphic for Ictiobinae and Cycleptinae, respectively. For the catostomines, it was found that 33.33%, 65.76%, 78.41%, and 84.03% of loci were found to be polymorphic for Erimyzonini, Thoburniini, Moxostomatini, and Catostomini,
respectively. Additionally, although unreported by Ferris and Whitt, we found that 77.4% of loci were found to be polymorphic for Myxocyprininae. The average number of polymorphic loci for all catostomids was 78.94%.

\[\text{Figure 9} \quad \text{A comparison of loci found to be polymorphic in the present study using 179 AHE loci and Ferris and Whitt’s (1980) study using 20 enzymatic loci. Values were obtained for both studies by averaging the values of species that fell within each more encompassing clade.}\]
Branch Length Comparisons

Comparison of branch lengths between genera within the Catostomidae revealed substantial differences in mean branch length values as well as variance values for branch lengths. Importantly, non-overlapping branch length distributions revealed a significant difference between the early branching catostomid lineages (*Myxocyprinus*, *Cycleptus*, *Ictiobus*, and *Carpiodes*; see preceding chapter) and the genera of Catostominae (Fig. 10). For the deep-bodied lineages, significant differences were found between all generic pairs, apart from *Myxocyprinus* and *Cycleptus* (*P* = 1; Appendix; Table 5). Within Catostominae (Fig. 11), it was found that significant differences existed between the generic pairs that occupied the tribes Thoburniini (*P* = 4.16 x 10^{-4}; Table 2) and Erimyzonini (*P* = 3.02 x 10^{-4}). The branch lengths of *Thoburnia* (*P* = 0.45) and *Minytrema* (*P* = 0.73) did not differ significantly from *Catostomus* (Table 2); however, the branch lengths of *Moxostoma* (*P* < 0.001), *Erimyzon* (*P* = 1.00 x 10^{-4}), and *Hypentelium* did (*P* = 7.74 x 10^{-4}).
Figure 10 A comparison of branch lengths between the genera of Catostomidae. Whiskers represent the range of branch length values obtained for each genus, with top whiskers indicating maximum branch length values and bottom whiskers representing minimum branch length values. Boxes define approximately 50% of intermediate branch length values. Lines intersecting the boxes represent the sample median for each taxon and the X’s within each box represents the sample mean for each taxon.
Figure 11 A comparison of branch lengths between genera within Catostominae. Whiskers represent the range of values for each genus, with top whiskers representing the maximum value, while bottom whiskers represent the minimum values. Boxes delimitate roughly 50% of the branch length values obtained for each genus. Lines intersecting the boxes represent the sample median for each taxon and the X’s within each box represents the sample mean for each taxon.
**Table 2** A Tukey’s Honestly Significant Difference Post-hoc test table displaying the results of a pairwise comparison of branch lengths between genera within Catostominae. Bold values indicate pairs between which there is a significant difference in branch lengths.

Tukey’s HSD Post-hoc Test for Generic Branch Length Comparisons of the Catostomines

<table>
<thead>
<tr>
<th>Generic Comparison</th>
<th>P-value Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Erimyzon</em>-Catostomus</td>
<td>0.0001001</td>
</tr>
<tr>
<td><em>Hypentelium</em>-Catostomus</td>
<td>0.0007744</td>
</tr>
<tr>
<td>Minytrema-Catostomus</td>
<td>0.7298683</td>
</tr>
<tr>
<td><em>Moxostoma</em>-Catostomus</td>
<td>0.0000000</td>
</tr>
<tr>
<td>Thoburnia-Catostomus</td>
<td>0.4491717</td>
</tr>
<tr>
<td>Hypentelium-Erimyzon</td>
<td>0.3961662</td>
</tr>
<tr>
<td>Minytrema-Erimyzon</td>
<td><strong>0.0003023</strong></td>
</tr>
<tr>
<td>Moxostoma-Erimyzon</td>
<td>0.4085526</td>
</tr>
<tr>
<td><em>Thoburnia</em>-Erimyzon</td>
<td><strong>0.0000224</strong></td>
</tr>
<tr>
<td>Minytrema-Hypentelium</td>
<td><strong>0.0086753</strong></td>
</tr>
<tr>
<td>Moxostoma-Hypentelium</td>
<td>0.9995329</td>
</tr>
<tr>
<td><em>Thoburnia</em>-Hypentelium</td>
<td><strong>0.0004163</strong></td>
</tr>
<tr>
<td>Moxostoma-Minytrema</td>
<td><strong>0.0013512</strong></td>
</tr>
<tr>
<td>Thoburnia-Minytrema</td>
<td>0.9999999</td>
</tr>
<tr>
<td><em>Thoburnia</em>-Moxostoma</td>
<td><strong>0.0000061</strong></td>
</tr>
</tbody>
</table>
Comparison of Alleles with Greater than Average Branch Lengths

Using branch lengths extracted from individual gene tree files, we generated the super-matrix heat map (Fig. 12). Branch lengths within this super-matrix ranged from $1 \times 10^{-8} - 101.15$ substitutions per nucleotide site. After branch lengths were normalized (relative to other branch lengths at a given locus) and alleles summed, it was discovered that *Erimyzon* had the greatest number of alleles with branch lengths greater than the average branch length for a given locus (65.6%; Fig. 13). Despite the pattern observed for overall branch lengths, where taxa within Catostominae had significantly longer branches, the morphologically conserved lineages tended to have a greater abundance of alleles with branch lengths greater than the average branch length (*Myxocyprinus* = 47.9%; *Cycleptus* = 45.8%; *Ictiobus* = 48.6%; *Carpiodes* = 62%). Taxa within *Catostomus* had the fewest number of loci with branch lengths greater than the average ($\bar{x} = 32\%$). The abundance of alleles with branch lengths greater than the average for the remaining catostomids were, for the most part, intermediate (*Minytrema* = 48.8%; *Thoburnia* = 42.5%; *Hypentelium* = 49%; *Moxostoma* = 43.3%).
Figure 12 A heat map produced by the assembly of branch length values obtained for genes tree included in this study. Gene trees were recovered using a maximum likelihood approach. Columns in this heat map represent individual loci while rows represent alleles of catostomids included in this study.
Figure 13 A comparison of the abundance of alleles per genus within the Catostomidae where branch lengths exceed the average branch length at a given locus.

Discussion

Estimating the Extent of Polymorphic Loci

In this study, it was found that the ictiobines had a greater frequency of polymorphic loci (91.05%) than the cycleptines (72.88%), myxocyprinines (77.4%), and tribes within the subfamily Catostominae (Erimyzonini = 33.33%; Thoburniini = 65.76%; Moxostomatini = 78.41%; Catostomini = 84.03%). This observation is consistent with the results obtained by Ferris and Whitt (1980), which estimated the frequency at which
genes that had maintained duplicate gene expression were found to be polymorphic using enzymatic gene product mobility. All taxa included in this study had far higher percentages of polymorphic loci than what Ferris and Whitt observed. This stark contrast is almost certainly a consequence of the use of different inferencing methods, given that Ferris and Whitt’s study used starch gel electrophoresis, which lacks the resolution of DNA sequence data. Following, few changes within our nucleotide sequence could result in a locus for a taxon appearing as polymorphic, whereas an equivalent change in Ferris and Whitt’s isozymes may result in only a slight change in gene product mobility, making loci appear as monomorphic. Additionally, while Ferris and Whitt found that taxa within Catostominae had similar frequencies of polymorphic loci, our data showed that taxa with Erimyzonini had far fewer polymorphic loci than the other catostomines. For example, *Erimyzon oblongus* was found to have 17.2% of its duplicate loci appear as polymorphic, a value comparable to the other catostomines (̄ = 16.9%) and about 49% smaller than the average obtained for Ictiobinae in the study of Ferris and Whitt. In our study, *E. oblongus* was found to have 27.12% of its duplicates be recovered as polymorphic, a value well below the average for the catostomines (̄ = 77.08%) and three times smaller than the value obtained for any of the ictiobines.

While the conclusions drawn from the comparisons of our study and Ferris and Whitt’s study (1980) are thought-provoking, it is worth keeping in mind that differences exist between these studies, which makes taking precaution while interpreting these results imperative. Firstly, Ferris and Whitt used enzymatic gene product data, which is a more conservative estimate of genic polymorphisms than sequence data. Additionally,
Sequence data is a class of data nested within protein data, which makes drawing connections between these studies challenging. Secondly, this study included far more loci (179 AHE’s) than Ferris and Whitt’s (20 isozymes) study. Disparate findings could thus be a consequence of limited genic sampling. Additionally, Ferris and Whitt’s analysis included only 19 catostomids, while our study included 43, evoking biases related to limited taxon sampling. What’s more is that, while most of the ictiobines included in the present study were included in Ferris and Whitt’s study, the catostomines are represented by a variety of different species between these studies. Lastly, DNA sequencing methodologies that aim to construct data sets that are biologically meaningful when studying polyploid taxa are still rather novel, which could lead to erroneous conclusions if data are unrepresentative of the taxa for which they were sequenced.

**Branch Lengths Comparisons**

In this study, it was found that genera within Catostominae had significantly longer species tree branch lengths than the genera *Myxocyprinus*, *Cycleptus*, *Ictiobus*, and *Carpiodes*. It has been observed that factors such as life history and other biological features can correlate with rates of molecular evolution. Some of these factors including: generation time (Thomas et al. 2010) and life longevity (Cordero and Janzen 2013), metabolic rates (Martin and Palumbi 1993), and body size (Hirt et al. 2017). Taxa within the subfamilies Ictiobinae, Cycleptinae, and Myxocyprininae tend to reach larger sizes on average than taxa within Catostominae, potentially resulting in the observed disparity between the longer branch lengths of the catostomines and shorter branch lengths of the
remaining subfamilies. Additionally, there is quite a bit of overlap in the expected lifespan and age at which individuals reach sexual maturity between the catostomids. Outliers from these trends could warrant additional investigation. For example, “Xyrauchen” texanus, a species within Catostominae, has been found to live up to +44 years, far exceeding the maximum recorded age of any ictiobine, cycleptine, or myxocyprinine. This observation seems peculiar given the previous research on correlations between rates of molecular evolution and life history. Analyses comparing these various life history and biological characteristics to branch length values could reveal taxa that are experiencing exceptionally high or low rates of molecular evolution, lending support to differential trends of duplicate gene expression maintenance and loss within the evolutionary lineages of polyploid taxa.

Comparison of Alleles with Greater than Average Branch Lengths

When branch lengths were normalized relative to the other branch lengths at a given locus, it was found that the genera Myxocyprinus, Cycleptus, Ictiobus, and Carpiodes more frequently had branch lengths greater than the average branch length at individual loci. In other words, when the catostomines did have branch lengths greater than the average, they were far greater than the average in order to result in the disparity in overall branch lengths mentioned in the preceding section. This may suggest that many redundant loci that have retained duplicate gene expression since their origin are either experiencing modest directional selection or modest purifying selection in the genomes of catostomids, while a relatively small proportion of loci in the catostomine genome are
experiencing relatively strong direction selection or a lack of selection, which would result in very long branch lengths as a consequence of nonfunctionalization. This observation could add a new facet to Ferris and Whitt’s hypothesis of differential duplicate gene expression retention between the morphologically conserved and morphologically divergent lineages, demonstrating that while generalist, morphologically conserved catostomids do seem to retain duplicate expression more often than the specialized, morphologically divergent lineages, nonfunctionalization may be less prevalent than what was previously thought. Nevertheless, the lack of an open reading frame, high level of sequence conservation, and potential inclusion of multiple gene regions and non-coding sequences makes it difficult to determine with confidence what type of selection is acting on each locus. Further investigation into this topic could reveal important information regarding which and why duplicate genes are experiencing strong purifying or directional selection (or a lack thereof) as well as patterns of differential selective pressures between and among closely-related lineages.

Conclusion

The results of this study add to the current literature and knowledge of the patterns of molecular evolution of polyploid vertebrate genomes. We have provided support that polymorphism of loci, indicative of neo-, sub-, or nonfunctionalization, tend to be more prevalent in Ictiobinae than Catostominae, corroborating the findings of Ferris and Whitt (1980); however, by using a genome-scale, nucleotide sequence data set, it was revealed that genic polymorphisms are far more copious that what was previously
thought for the Catostomidae. Additionally, we have demonstrated that base pair
substitutions have accumulated differentially between the catostomid lineages and that
the catostomines tended to have fewer loci with longer branches than the taxa with
Myxocyprininae, Cycleptinae, and Ictiobinae, yet when the branches of the catostomines
were longer, they were much longer, potentially indicative of strong directional selection
or nonfunctionalization acting on a small proportion of loci.

A tremendous amount of work remains left undone in elucidating the evolutionary
consequences of WGD and little is still known about the patterns and rates of molecular
evolution in this family. This study represents a small step forward in our understanding
of the evolution of this family and, more broadly, of the evolution of polyploid genomes.
A fruitful endeavor in this research area may be to examine which type of selective
pressure and the magnitude at which selection is acting on individual loci. Additional
studies may also capitalize on the results presented here by comparing the rates of
molecular evolution within the various evolutionary lineages on the Catostomidae to
other biological characters of these fishes, such as longevity, age of sexual maturity, and
body size in an attempt to identify outliers that could provide additional information on
why lineages maintain or lose duplicate gene expression differentially.
REFERENCES


Mable BK. 2004. ‘Why polyploidy is rarer in animals than in plants’: myths and mechanisms. Biological Journal of the Linnean Society. 82:453-466.


doi:10.1093/bioinformatics/btv234


Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [PhD dissertation]. [Austin (Texas)]: The University of Texas at Austin.
APPENDIX

SUPPLEMENTARY MATERIALS

Table A1. Catostomid species included in the study. Institutional abbreviations are as follows: JFBM = Bell Museum of Natural History Fish collection, University of Minnesota; AUFT = Auburn University Fish Tissue Collection; MSB = Museum of Southwestern Biology, University of New Mexico; OS = Oregon State Ichthyology Collection; UAIC = University of Alabama Ichthyological Collection.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Specimen Voucher</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cycleptinae</strong></td>
<td></td>
</tr>
<tr>
<td><em>Cycleptus elongatus</em></td>
<td>JFBM (PBB 00-23)</td>
</tr>
<tr>
<td><strong>Myxocyprininae</strong></td>
<td></td>
</tr>
<tr>
<td><em>Myxocyprinus asiaticus</em></td>
<td>JFBM (aquarium trade / UNI)</td>
</tr>
<tr>
<td><strong>Ictiobinae</strong></td>
<td></td>
</tr>
<tr>
<td><em>Carpiodes carpio</em></td>
<td>JFBM (PBB 09-05)</td>
</tr>
<tr>
<td><em>Carpiodes cyprinus</em></td>
<td>JFBM (PBB 09-05)</td>
</tr>
<tr>
<td><em>Carpiodes velifer</em></td>
<td>AUFT 0045</td>
</tr>
<tr>
<td><em>Ictiobus bubalus</em></td>
<td>JFBM (PBB 09-05)</td>
</tr>
<tr>
<td><em>Ictiobus cyprinellus</em></td>
<td>JFBM (PBB 00-19)</td>
</tr>
<tr>
<td><em>Ictiobus niger</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><strong>Catostominae</strong></td>
<td></td>
</tr>
<tr>
<td><em>Catostomus ardens</em></td>
<td>JFBM (PBB 09-01)</td>
</tr>
<tr>
<td><em>Catostomus bernardini</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><em>Catostomus cahita</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><em>Catostomus catostomus</em></td>
<td>AUFT 0402</td>
</tr>
<tr>
<td><em>Catostomus clarkii</em></td>
<td>MSB 49600</td>
</tr>
<tr>
<td><em>Catostomus columbianus</em></td>
<td>JFBM (PBB09-02)</td>
</tr>
<tr>
<td><em>Catostomus commersonii</em></td>
<td>JFBM (AMS 01-17)</td>
</tr>
<tr>
<td><em>Catostomus discobolus</em></td>
<td>MSB 49688</td>
</tr>
<tr>
<td><em>Catostomus insignis</em></td>
<td>MSB 49603</td>
</tr>
<tr>
<td><em>Catostomus latipinnis</em></td>
<td>MSB 49689</td>
</tr>
<tr>
<td><em>Catostomus leopoldi</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><em>Catostomus macrocheilus</em></td>
<td>AUFT 0241</td>
</tr>
<tr>
<td><em>Catostomus occidentalis</em></td>
<td>JFBM (PBB 09-04)</td>
</tr>
<tr>
<td><em>Catostomus play rhynchus</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><em>Catostomus plebius</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><em>Catostomus wigginsi</em></td>
<td>Stout et al. (2016)</td>
</tr>
</tbody>
</table>

(Table Continues)
<table>
<thead>
<tr>
<th>Taxa</th>
<th>Specimen Voucher</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Chasmistes liorus</em></td>
<td>JFBM 45993</td>
</tr>
<tr>
<td><em>Deltistes luxatus</em></td>
<td>OS 18966</td>
</tr>
<tr>
<td><em>Erimyzon oblongus</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><em>Hypentelium etowanum</em></td>
<td>JFBM (PBB 00-10)</td>
</tr>
<tr>
<td><em>Hypentelium nigricans</em></td>
<td>JFBM 44429-10</td>
</tr>
<tr>
<td><em>Hypentelium roanokense</em></td>
<td>JFBM (PBB 01-05)</td>
</tr>
<tr>
<td><em>Minytrema melanops</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><em>Moxostoma anisurum</em></td>
<td>JFBM (PBB 09-05)</td>
</tr>
<tr>
<td><em>Moxostoma arriommum</em></td>
<td>UAIC 12072.01</td>
</tr>
<tr>
<td><em>Moxostoma cervinum</em></td>
<td>JFBM (AMS 08-02)</td>
</tr>
<tr>
<td><em>Moxostoma duquesnei</em></td>
<td>JFBM (RK T135)</td>
</tr>
<tr>
<td><em>Moxostoma erythrurum</em></td>
<td>JFBM (PBB 09-05)</td>
</tr>
<tr>
<td><em>Moxostoma lachneri</em></td>
<td>AUFT 1003</td>
</tr>
<tr>
<td><em>Moxostoma poecilurum</em></td>
<td>JFBM (JJDE MP9)</td>
</tr>
<tr>
<td><em>Moxostoma rupiscartes</em></td>
<td>JFBM (AMS 01-50)</td>
</tr>
<tr>
<td><em>Moxostoma valenciennsi</em></td>
<td>JFBM (PBB 09-05)</td>
</tr>
<tr>
<td><em>Thoburnia atripinnis</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><em>Thoburnia rhothoeca</em></td>
<td>Stout et al. (2016)</td>
</tr>
<tr>
<td><em>Xyrauchen texanus</em></td>
<td>MSB 46722</td>
</tr>
</tbody>
</table>
Table A2. The R script used for obtaining Robinson-Foulds Distances for gene tree-gene tree/gene tree-species tree comparisons.

```r
library(phytools)
library(ape)
library(phangorn)
library(lattice)

L6 <- read.tree("006.tre")
AllTree <- read.tree("RF_Trees.tre")
multiRF(AllTree)
write.csv(multiRF(AllTree), file = "RF_Distance_Matrix.csv")
AllTree_Distances <- read.csv("RF_Distance_Matrix.csv", header = FALSE)
UpperTriangle <- AllTree_Distances[lower.tri(AllTree_Distances)]
write.csv(UpperTriangle, file = "RF_Distances_for_Histogram.csv")
```
Table A3. A pairwise comparison of branch lengths of the Catostomidae using a Tukey’s Honestly Significant Difference Post-hoc Test. Significant values for generic comparisons are bolded.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Difference</th>
<th>Lower</th>
<th>Upper</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catostomus-Carpiodes</td>
<td>0.146</td>
<td>0.131</td>
<td>0.160</td>
<td>0.000</td>
</tr>
<tr>
<td>Cycleptus-Carpiodes</td>
<td>0.088</td>
<td>0.062</td>
<td>0.114</td>
<td>0.000</td>
</tr>
<tr>
<td>Erimyzon-Carpiodes</td>
<td>0.174</td>
<td>0.148</td>
<td>0.201</td>
<td>0.000</td>
</tr>
<tr>
<td>Hypentelium-Carpiodes</td>
<td>0.161</td>
<td>0.142</td>
<td>0.180</td>
<td>0.000</td>
</tr>
<tr>
<td>Ictiobus-Carpiodes</td>
<td>0.060</td>
<td>0.042</td>
<td>0.079</td>
<td>0.000</td>
</tr>
<tr>
<td>Minytrema-Carpiodes</td>
<td>0.137</td>
<td>0.111</td>
<td>0.163</td>
<td>0.000</td>
</tr>
<tr>
<td>Moxostoma-Carpiodes</td>
<td>0.162</td>
<td>0.147</td>
<td>0.178</td>
<td>0.000</td>
</tr>
<tr>
<td>Myxocyprinus-Carpiodes</td>
<td>0.093</td>
<td>0.066</td>
<td>0.119</td>
<td>0.000</td>
</tr>
<tr>
<td>Thoburnia-Carpiodes</td>
<td>0.137</td>
<td>0.117</td>
<td>0.158</td>
<td>0.000</td>
</tr>
<tr>
<td>Cycleptus-Catostomus</td>
<td>-0.057</td>
<td>-0.081</td>
<td>-0.034</td>
<td>0.000</td>
</tr>
<tr>
<td>Erimyzon-Catostomus</td>
<td>0.029</td>
<td>0.005</td>
<td>0.052</td>
<td>0.005</td>
</tr>
<tr>
<td>Hypentelium-Catostomus</td>
<td>0.016</td>
<td>0.001</td>
<td>0.030</td>
<td>0.020</td>
</tr>
<tr>
<td>Ictiobus-Catostomus</td>
<td>-0.085</td>
<td>-0.099</td>
<td>-0.071</td>
<td>0.000</td>
</tr>
<tr>
<td>Minytrema-Catostomus</td>
<td>-0.009</td>
<td>-0.032</td>
<td>0.015</td>
<td>0.975</td>
</tr>
<tr>
<td>Moxostoma-Catostomus</td>
<td>0.017</td>
<td>0.008</td>
<td>0.026</td>
<td>0.000</td>
</tr>
<tr>
<td>Myxocyprinus-Catostomus</td>
<td>-0.053</td>
<td>-0.076</td>
<td>-0.029</td>
<td>0.000</td>
</tr>
<tr>
<td>Thoburnia-Catostomus</td>
<td>-0.008</td>
<td>-0.025</td>
<td>0.009</td>
<td>0.872</td>
</tr>
<tr>
<td>Erimyzon-Cycleptus</td>
<td>0.086</td>
<td>0.054</td>
<td>0.118</td>
<td>0.000</td>
</tr>
<tr>
<td>Hypentelium-Cycleptus</td>
<td>0.073</td>
<td>0.047</td>
<td>0.099</td>
<td>0.000</td>
</tr>
<tr>
<td>Ictiobus-Cycleptus</td>
<td>-0.028</td>
<td>-0.054</td>
<td>-0.001</td>
<td>0.030</td>
</tr>
<tr>
<td>Minytrema-Cycleptus</td>
<td>0.049</td>
<td>0.017</td>
<td>0.081</td>
<td>0.000</td>
</tr>
<tr>
<td>Moxostoma-Cycleptus</td>
<td>0.074</td>
<td>0.050</td>
<td>0.098</td>
<td>0.000</td>
</tr>
<tr>
<td>Myxocyprinus-Cycleptus</td>
<td>0.005</td>
<td>-0.028</td>
<td>0.037</td>
<td>1.000</td>
</tr>
<tr>
<td>Thoburnia-Cycleptus</td>
<td>0.049</td>
<td>0.021</td>
<td>0.077</td>
<td>0.000</td>
</tr>
<tr>
<td>Hypentelium-Erimyzon</td>
<td>-0.013</td>
<td>-0.040</td>
<td>0.013</td>
<td>0.838</td>
</tr>
<tr>
<td>Ictiobus-Erimyzon</td>
<td>-0.114</td>
<td>-0.140</td>
<td>-0.088</td>
<td>0.000</td>
</tr>
<tr>
<td>Minytrema-Erimyzon</td>
<td>-0.037</td>
<td>-0.070</td>
<td>-0.005</td>
<td>0.010</td>
</tr>
<tr>
<td>Moxostoma-Erimyzon</td>
<td>-0.012</td>
<td>-0.036</td>
<td>0.012</td>
<td>0.846</td>
</tr>
<tr>
<td>Myxocyprinus-Erimyzon</td>
<td>-0.082</td>
<td>-0.114</td>
<td>-0.049</td>
<td>0.000</td>
</tr>
<tr>
<td>Thoburnia-Erimyzon</td>
<td>-0.037</td>
<td>-0.065</td>
<td>-0.009</td>
<td>0.001</td>
</tr>
<tr>
<td>Ictiobus-Hypentelium</td>
<td>-0.101</td>
<td>-0.119</td>
<td>-0.082</td>
<td>0.000</td>
</tr>
<tr>
<td>Minytrema-Hypentelium</td>
<td>-0.024</td>
<td>-0.050</td>
<td>0.002</td>
<td>0.104</td>
</tr>
<tr>
<td>Moxostoma-Hypentelium</td>
<td>0.001</td>
<td>-0.014</td>
<td>0.016</td>
<td>1.000</td>
</tr>
</tbody>
</table>

(Table continues)
<table>
<thead>
<tr>
<th>Taxa</th>
<th>Difference</th>
<th>Lower</th>
<th>Upper</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minytrema-Ictiobus</td>
<td>0.077</td>
<td>0.050</td>
<td>0.103</td>
<td>0.000</td>
</tr>
<tr>
<td>Moxostoma-Ictiobus</td>
<td>0.102</td>
<td>0.087</td>
<td>0.117</td>
<td>0.000</td>
</tr>
<tr>
<td>Myxocyprinus-Ictiobus</td>
<td>0.032</td>
<td>0.006</td>
<td>0.059</td>
<td>0.005</td>
</tr>
<tr>
<td>Thoburnia-Ictiobus</td>
<td>0.077</td>
<td>0.056</td>
<td>0.098</td>
<td>0.000</td>
</tr>
<tr>
<td>Moxostoma-Minytrema</td>
<td>0.025</td>
<td>0.001</td>
<td>0.049</td>
<td>0.029</td>
</tr>
<tr>
<td>Myxocyprinus-Minytrema</td>
<td>-0.044</td>
<td>-0.076</td>
<td>-0.012</td>
<td>0.001</td>
</tr>
<tr>
<td>Thoburnia-Minytrema</td>
<td>0.000</td>
<td>-0.027</td>
<td>0.028</td>
<td>1.000</td>
</tr>
<tr>
<td>Myxocyprinus-Moxostoma</td>
<td>-0.070</td>
<td>-0.094</td>
<td>-0.046</td>
<td>0.000</td>
</tr>
<tr>
<td>Thoburnia-Moxostoma</td>
<td>-0.025</td>
<td>-0.043</td>
<td>-0.007</td>
<td>0.001</td>
</tr>
<tr>
<td>Thoburnia-Myxocyprinus</td>
<td>0.045</td>
<td>0.017</td>
<td>0.073</td>
<td>0.000</td>
</tr>
<tr>
<td>Myxocyprinus-Hypentelium</td>
<td>-0.068</td>
<td>-0.095</td>
<td>-0.042</td>
<td>0.000</td>
</tr>
<tr>
<td>Thoburnia-Hypentelium</td>
<td>-0.024</td>
<td>-0.044</td>
<td>-0.003</td>
<td>0.013</td>
</tr>
</tbody>
</table>
Table A4. The R script used for extracting branch lengths from individual gene tree files as well as the species tree file.

```r
library(ade4)
library(phylobase)
library(ape)
library(adephylo)

tree<-read.tree(file = "tree_file.tre")
tree<-read.nexus("tree_file.tre")
d<-distRoot(tree, method = "patristic")
d.matrix<-as.data.frame(d)

write.csv(d.matrix, "tree_branch_lengths.csv")
```