Evolution of Miniaturization and Paedomorphism in Fishes of the Order Cypriniformes

by

Milton Tan

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 6, 2016

Keywords: body size, diversification, phylogenetics
phylogenomics, transcriptome, zebrafish

Approved by

Jonathan W. Armbruster, Chair, Professor of Biological Sciences and Curator of Fishes
Jason E. Bond, Professor and Department Chair of Biological Sciences
Eric Peatman, Associate Professor of Fisheries, Aquaculture, and Aquatic Sciences
Scott R. Santos, Professor of Biological Sciences

Abstract

The order Cypriniformes (carps, minnows and their allies) is a morphologically diverse freshwater fish clade numbering over 4000 species. One species is the model organism zebrafish, *Danio rerio*. Cypriniform fishes have repeatedly undergone extreme reductions in body size, or miniaturization. Some miniatures are particularly small, and are also paedomorphic, retaining larval characteristics into adulthood. The primary goal of this dissertation research is to study the evolution of miniaturization ane paedomorphism in the order Cypriniformes. The dissertation opens with a brief introduction into the motivation for this research, and an outline for the remaining dissertation. The following chapters of the dissertation present various approaches to study this phenomenon from multiple perspectives: patterns of body size evolution, evolutionary relationships of miniature fishes, and functional genomics underlying miniaturization. Chapter 2 presents an empirical study of the dynamics of body size evolution and its relationship to miniaturization in the Danionidae, a clade including the majority of miniature cypriniform species. Not all miniatures are created equal: some of the smallest vertebrates are paedomorphic Cypriniformes (retaining larval characteristics into adulthood). Prior phylogenetic studies conflicted on the relationships between multiple paedomorphic genera – *Paedocypris*, *Sundadanio*, and *Danionella* – with implications for the number of times paedomorphism evolved. Chapter 3 presents a study utilizing phylogenomics to robustly resolve the relationships of these taxa among Cypriniformes. Finally, chapter 4 presents a study to gain insight into the genetic basis of convergent evolution of paedomorphism using comparative transcriptomics.

Acknowledgments

I am especially indebted to my major advisor Jonathan W. Armbruster. This dissertation would have been impossible without Jon's support, who throughout this endeavor has been accepting of my proposed ideas, and gave me the opportunity to work on some of my dream projects with no more assurance of success than my own half-baked plans. I am additionally thankful to my dissertation committee members, Jason E. Bond, Eric Peatman, and Scott R. Santos, for challenging me throughout my dissertation research.

I am also thankful to the present and past Armbruster lab members I was fortunate enough to overlap with, including Malorie M. Hayes, Edward D. Burress, Carla C. Stout, Pamela B. Hart, C. Keith Ray, Shobnom Ferdous, and Lesley S. deSouza, as well as members of the Bond lab, Charles D. R. Stephen, Chris A. Hamilton, Nicole L. Garrison, Rebecca L. Godwin, Juanita Rodriguez, and Vera Opatova, who provided intellectual and social support on this wild ride. Thanks also to Justin C. Havird, Kevin M. Kocot, and Nathan V. Whelan, who have also provided great scientific and intellectual inspiration, motivation, and encouragement. I am also grateful for the pleasure of working with numerous undergraduate mentees, who perhaps beside their own better judgment volunteered to provide assistance in my research, and also provide me a challenging yet rewarding mentorship experience, in particular Kasey C. Benesh. I am also thankful for the staff of the Department of Biological Sciences who helped make sure I could focus on what was important, especially Sandra Abate who made my business her own and Alex

Rittenbaugh who assured I jumped some of the final bureaucratic hurdles. In addition, I am also thankful for numerous others of the Department of Biological Sciences, both faculty and students, too many to name, who have helped provide an enriching intellectual and social environment. I am also thankful to the Auburn University Cell & Molecular Biology Fellowship program, which saw fit to fund three consecutive summers for me to accomplish my dissertation.

Last but in no way least, big thanks go to my family. My sisters Emily, Connie, and Gretchen, and parents Nick Tan and Sue Horng proved to be a constant, and a family and home for me to return to. I am indebted to my parents, who always provided me more than enough support for me to follow my dreams, whenever I saw fit to ask, and who probably never imagined my first aquarium they gifted me would lead me into the fish business.

Table of Contents

List of Tables

Chapter 1

Chapter 3

Chapter 4

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

LITTLE FISH, BIG QUESTIONS

A BRIEF INTRODUCTION TO THE DISSERTATION RESEARCH

Body size is perhaps one of the most salient biological characteristics that varies across the diversity of life. Ranging from classic educational literature titled "Creatures Great and Small" to the use of humans for scale in images of dinosaurs, it is clear that body size has both immense utility and attraction to humans. Body size is tied to a variety of aspects of organismal biology, including their physiological ability to maintain efficient transport of materials, store resources, and their energetic requirements for maintenance and growth; their ecological role as predator or prey; biomechanical constraints related to necessary physical function for support and movement; and life history characteristics such as longevity and fecundity.

Although the very large organisms famously capture the public imagination (e.g. McClain et al. 2015), the continuing discovery of the very small receives not just scientific interest (e.g. Kottelat et al. 2006), but considerable media attention as well. Miniature organisms have received relatively little study. Cope's Rule – the hypothesis that body size tends to increase over time (Bokma et al. 2016) – is an established concept in body size evolution and at its basis seems to oppose the idea of an importance of small body size. It has only been in the last few decades that recognition of miniature taxa has increased (Weitzman & Vari 1988; Hanken & Wake 1993), and studies now focus on miniaturization for its importance as an example of convergent evolution (Rüber et al. 2007; Rundell & Leander 2010), as leading to the origin of diverse clades (Lee et al. 2014), or as sources of phenotypic novelty (Britz & Conway 2016).

Related to miniaturization is the concept of paedomorphism, or the retention of larval characteristics (Hanken & Wake 1993; Rüber et al. 2007). Paedomorphism through developmental truncation has occurred repeatedly in teleosts and often accompanies miniaturization (Britz & Conway 2016), and thus paedomorphism represents a consistent pathway towards the reduction in body size. Paedomorphism can result from different sources of variation in the timing of organismal development (developmental truncation), such as either the early truncation of development or by development occurring at a slower rate (neoteny) (Rüber et al. 2007).

The order Cypriniformes is one of the most diverse clades of freshwater fishes (Mayden & Chen 2010) currently surpassing 4000 species (Eschmeyer et al. 2016). Amongst this diverse group are numerous examples of repeated miniaturization (Rüber et al. 2007). The Cypriniformes may be more likely than many other fish groups to evolve extreme miniature body size (Albert & Johnson 2009). In 2006, one of the smallest fishes in the world, *Paedocypris progenetica*, was discovered, joining the ranks of *Sundadanio* and *Danionella* as miniature, paedomorphic fishes within the Cypriniformes (Kottelat et al. 2006). In the following few years, these three taxa have been the focus of research of a number of papers that inspired this dissertation. The following chapters of dissertation research address the evolution of miniaturization and paedomorphism from multiple perspectives at the intersection of morphological evolution, phylogenetics, and functional genetic evolution with the overall goal of furthering understanding of the patterns and processes underlying the evolution of diversity of miniature fishes.

Rüber et al. (2007) presented the first hypothesis of the evolutionary relationships of the paedomorphic Cypriniformes, recovering them as a part of the family Danionidae, using a single mitochondrial gene, cytochrome b. This was a fairly intuitive finding, because Danionidae includes the bulk of the miniature diversity of Cypriniformes, and Rüber et al. (2007) demonstrated there were at least seven independent transitions to a miniature body size. On the other hand, this study was fairly limited, with body size coded as a binary character (miniature vs. non-miniature), categorizing body sizes across a threshold that has been considered arbitrary since its original conception (Weitzman & Vari 1988). While numerous transitions to a miniature body size is interesting, the arbitrary threshold biases the picture of body size evolution in the Danionidae, and limits insights into more general evolutionary patterns. In Chapter 1, I expand on previous research by analyzing the pattern and rates of body size evolution on an expanded, time-calibrated phylogeny of the Danionidae based on published multilocus phylogenetic data (Tang et al. 2010).

A series of phylogenetic studies followed Rüber et al. (2007), with Britz & Conway (2009), Mayden & Chen (2010), and Britz et al. (2014) addressing the relationships of *Danionella*, *Sundadanio*, and a focus on *Paedocypris*. High conflict among phylogenetic studies prompted a deeper exploration into the phylogenetic relationships of these taxa in relation to Danionidae for my dissertation. With the advent of next-generation sequencing technologies and their application towards collecting phylogenomic data, inferring the evolutionary relationships among taxa no longer need to be limited by the amount of available data. In Chapter 2, using anchored enrichment (Lemmon et al. 2009), I inferred the evolutionary relationships of *Paedocypris*, *Sundadanio*, *Danionella*, and major groups of Danionidae.

Phylogenomics has been hailed as a potential cure for resolving evolutionary relationships (Philippe et al. 2005), but it has also led to the rise of hard conflict (Jeffroy et al. 2006). Although phylogenomic-scale datasets have decreased random noise (i.e. sampling error), non-random noise (i.e. systematic error) has emerged as a new concern. A variety of sources of non-random noise can bias the relationships inferred from phylogenetic analyses. A deeper insight into incongruence can arise from accounting for potential sources of bias and determining their effects on phylogenetic analyses. In Chapter 3, I test for whether a variety of sources of systematic error and taxon sampling affect the recovered relationships of *Paedocypris* from phylogenomic data, test the sensitivity of phylogenomic data to various sources of error, and gain further insight into the phylogenetic relationships of Cypriniformes by comparison of congruence across data subsets.

Genomic scale data can provide insights into evolutionary relationships, but genes also represent the basic units encoding the proteins that underlie organismal function and phenotype. The pattern of gene evolution does not just provide a record of phylogeny, but can contain the signal of the effect of the invisible hand of natural selection. Multiple, independent evolutionary transitions to a certain phenotype additionally provides a replicated natural experiment to test the generality of gene function. The discovery of genes that have independently shifted to a similar selection regime across paedomorphic taxa can demonstrate an important role in these genes in the evolution of these taxa, and provide the potential to gain insights into the evolution of miniaturization and the paedomorphic phenotype. In Chapter 4, I sequence and assemble the transcriptomes of *Paedocypris*, *Sundadanio*, and *Danionella* to discover the genes that are under the same selection regime across these three taxa relative to the other Cypriniformes.

CHAPTER 1

PATTERNS OF BODY SIZE EVOLUTION AND MINIATURIZATION IN THE DANIONIDAE

ABSTRACT

Miniaturization has occurred numerous times among vertebrates. Recently, with the development of phylogenetic comparative methods, miniaturization can be studied with respect to patterns of body size evolution over time. We study the dynamics of body size evolution in the Danionidae, a group of fishes including several independent evolutionary transitions to a miniaturized state. We tested whether a change in the rate of body size evolution was related to the numerous transitions to miniature body size, and found that rates of body size evolution were constant through time. We also tested for whether sustained miniaturization (a directional downward trend in body size over time) has occurred in the Danionidae, and found no support for this hypothesis. We finally tested if rates of body size evolution are dependent on either miniaturized state or body size, and found that rates of body size evolution are decreased in miniature species. Overall, rates of body size evolution in Danionidae do not appear to be increased despite including the highest diversity of miniature fishes among Cypriniformes.

INTRODUCTION

Animals vary across a wide range of body sizes, from miniatures to giants (Kottelat et al. 2006; McClain et al. 2015). The macroevolutionary patterns leading to present-day variation in body size have long interested biologists. The observation of increases in body size over time within many clades has been commonly called as Cope's Rule (Cope 1904; Stanley 1973). It has been suggested that Cope's Rule should be called Depéret's Rule, because Cope never actually originated the hypothesis (Polly and Alroy 1998; Bokma et al. 2016). There is evidence for Depéret's Rule in the paleontological record (eg. Frigot et al. 2014; Heim et al. 2015) and theoretical explanations for this pattern (Stanley 1973). Depéret's Rule may arise because ancestral body sizes of clades typically begin near a lower bound in body size, resulting in preferential increase that can lead to extremely large body sizes over time (Sander et al. 2010; ie. gigantism Frigot et al. 2014), an idea termed Stanley's Rule (Stanley 1973; Gould 1988).

A decrease in size across clades has not been examined as often as increases. Clades undergoing miniaturization, the evolution of extremely small body size (Hanken and Wake 1993), may be less common than those experiencing Depéret's Rule (Stanley 1973; Hanken and Wake 1993). Despite the relative lack of study, miniaturization has occurred multiple times throughout animal evolution, and is a potential phenomenon of interest to study convergent evolution (Rundell and Leander 2010). Miniaturization often results in morphological reduction and simplification, morphological novelty, and increased morphological variability (Hanken and Wake 1993; Miller 1996). Novel body plans that arise as a consequence of miniaturization can lead to the origin of major taxa, such as in snakes, lizards, and bivalves (Hanken and Wake 1993). Small body size may also promote rates of diversification (eg. Hardman and Hardman 2008). Species at the lower bounds of their body size may exhibit constrained evolution where size evolution slows as species reach the limits of their potential adaptive zone. This limit in body size has been hypothesized to form the basis of Depéret's Rule because new clades tend to originate at or near their lower body size boundary, and have only one direction in which they may change or body sizes have been theorized to grow larger over time due to ancestral sizes being smaller than a clade's optimal body size (Stanley 1973; Gould 1988; 1997).

Although miniaturization is found across vertebrates, it is most extreme in poikilotherms such as fishes (Albert and Johnson 2012). The many miniature fish species (Kottelat and Vidthayanon 1993; Conway and Moritz 2006; Rüber et al. 2007) make fishes an appropriate system to study patterns of body size evolution and miniaturization. Small body size in fishes is usually accompanied by reduced development or absence of the lateral line system, scales, and skull and tail bones, and fewer fin rays (Weitzman and Vari 1988; Britz and Conway 2009). The correlation of small body size and reductive characters was used to specify an arbitrary threshold to define miniature fishes as species that reach sexual maturity at less than 20 mm SL and do not exceed 26 mm SL (Weitzman and Vari 1988). Although this arbitrary threshold means some species that exceed 26 mm but exhibit paedomorphic characters typical of miniatures do not strictly qualify as miniatures (Kottelat and Vidthayanon 1993; Calegari et al. 2014), this definition has still been used to document miniature ichthyofauna in freshwaters of North America, Africa, Asia, and South America (Weitzman and Vari 1988; Kottelat and

Vidthayanon 1993; Conway and Moritz 2006; Bennett and Conway 2010; Toledo-Piza et al. 2014).

There have been relatively few studies on the evolution of small body size in fishes using comparative methods, or on rates of body size evolution in miniature fishes (but see Knouft and Page 2003; Hardman and Hardman 2008; Rabosky et al. 2013 for examples). Albert and Johnson (2012) studied fish body size distributions across time, and found that body sizes increased from the origin of fishes in the Cambrian and stabilized in the Devonian at body size distributions similar to modern fishes. They found that certain fish groups (characins, gobies, cyprinids, and poeciliids) may be predisposed for miniaturization, and defined extreme miniatures as species that are more than three standard deviations from the mean for all fishes (< 1.4 cm TL). Another study on the cyprinid family Danionidae used ancestral state reconstruction and coded miniaturization as a binary, discrete character *sensu* Weitzman and Vari (1988) and demonstrated that miniaturization occurred at least seven times (Rüber et al. 2007).

Danionidae (previously Rasborinae) is a species-rich subfamily (>300 species) of Asian and African cyprinids (Tang et al. 2010). Many species are popular in the aquarium trade due to their small size and attractive color patterns, including the zebrafish (*Danio rerio*), a model organism in developmental biology (Howe et al. 2013). While cyprinids may be preadapted to evolve small body sizes (Albert and Johnson 2012), most miniature cyprinids are members of Danionidae (Rüber et al. 2007), making it an excellent group to examine miniaturization. There are three main groups within the Danionidae: Chedrinae, Rasborinae, and Danioninae (Tang et al. 2010; Liao et al. 2011). Most of the miniatures in Danionidae (Table 1) represent proportioned dwarfs, which have some reductions in skeletal development (eg. absent bones of the skull and the caudal skeleton), but adults are similar to adults of non-miniature relatives (Rüber et al. 2007; Britz and Conway 2009). In contrast, all four species of *Danionella* are developmentally truncated miniatures, where adults are paedomorphic, retaining larval characteristics such as transparency, a less ossified skeleton, and larval fin folds as adults (Roberts 1986; Britz 2003; Rüber et al. 2007; Britz et al. 2009; Britz 2009). Additionally, *Danionella* represent some of the smallest vertebrates, with *Danionella translucida* reaching a maximum size of only 12 mm SL (Roberts 1986). The extremely reduced morphology of *Danionella dracula* coupled with evolutionary novelties not found in any other cypriniform demonstrates the role of miniaturization in radically affecting the body plan in *Danionella* (Britz and Conway 2016). Danionidae has also been thought to include the miniature, paedomorphic cypriniforms *Paedocypris* and *Sundadanio* (Tang et al. 2010), though this is not supported by mitogenomic, nuclear phylogenetic, or phylogenomic data (Mayden and Chen 2010 Chapter 2).

Britz et al. (2014) reanalyze previous morphological and molecular data and demonstrate molecular data are indecisive on the relationships of *Paedocypris* and the morphological data is decisive on the monophyly of the paedomorphic taxa forming a clade. Because the phylogenetic relationships for Cyprinoidei are unresolved by all of their morphological datasets, however, it is unclear that *Paedocypris* and *Sundadanio* actually belong to Danionidae despite the apparent certainty that the paedomorphic taxa form a clade. We recognize *Danionella* as a danionid due to consistent support for its inclusion within this clade across phylogenetic and phylogenomic studies (Mayden and

Chen 2010; Tang et al. 2010; McCluskey and Postlethwait 2015), but the relationships of *Paedocypris* and *Sundadanio* among Cyprinoidei are equivocal.

Although miniaturization occurred on at least seven independent occasions within Danionidae (Rüber et al. 2007), Weitzman and Vari (1988) recognized their threshold for miniature fishes was arbitrary. Further insights into how miniaturization evolves can be gained by considering it within the context of body size evolution in general. Rabosky et al. (2013) reported no rate variation in body size evolution occurring within cyprinids, but the scale of the study across ray-finned fishes could obscure rate variation within cyprinids relative to other fish families that showed greater rate variation. This is supported by the discovery of shifts in body size evolution rate within North American cyprinids (Martin and Bonett 2015).

Here, we examine the direction and variation in rate of body size evolution in the Danionidae and how this relates to miniaturization. We compiled a dataset of Danionidae standard lengths and utilize published sequence data to infer a time-calibrated phylogeny for the Danionidae. We use multiple phylogenetic comparative methods to infer variation in rates and direction of body size evolution, and test for dependence of patterns of body size evolution on body size itself. For clarity, we will restrict the term 'miniaturization' to transitions from a non-miniature body size to a miniature body size *sensu* Weitzman & Vari (1988), despite the use of decreases in body size also being termed miniaturization in the literature (Avaria-Llautureo et al. 2012; Lee et al. 2014).

MATERIALS AND METHODS

Taxonomy

We followed Catalog of Fishes for recognition of species and genera as valid (Eschmeyer et al. 2016), though we recognize *Neochela* as a distinct genus, as in Kottelat (2013). We do not recognize the paedomorphic genera *Paedocypris* and *Sundadanio* within Danionidae, which is supported by both multilocus nuclear data and mitogenomic phylogenetic analyses (Mayden and Chen 2010 Mayden pers. comm.) and phylogenomic data (Chapters 2, 3).

Body Size Data

We downloaded maximum size data for species of Danionidae from FishBase using the R package rFishBase (Froese & Pauly 2014; Boettiger et al. 2012) and manual querying of FishBase (last accessed October 15, 2014). We then manually curated the body size dataset by referencing the primary literature. We recorded if a species has been listed as miniature based on previous literature, even if these species have been noted to slightly exceed 26 mm SL (Table 1; Roberts 1986; Kottelat and Vidthayanon 1993; Kottelat and Witte 1999; Britz 2003; Conway 2005; Roberts 2007; Jiang et al. 2008; Britz et al. 2009; Britz 2009; Fang et al. 2009; Kullander and Fang 2009; Conway and Kottelat 2011; Batuwita et al. 2013). We additionally found some species that have not been reported to exceed 26 mm SL and have also not previously been reported as miniature.

Reported size data may be in total length (TL) or SL, and lengths were standardized to SL using ratios, as has previously been done in body size studies in cyprinids (Denys et al. 2014). Descriptions and images of species were used to determine

SL-TL ratios (eg. Hora and Mukerji 1928; Tilak and Husain 1990; Barman 1991; Pethiyagoda et al. 2008). If the maximum SL reported in the literature was greater than the maximum SL that was calculated from a maximum reported TL, the reported maximum SL was used. Body sizes FishBase cited from aquarist literature were not used as these can be unreliable (Pethiyagoda 1991). In some cases, the identities of common species have been clarified or species have been split since the body size reported in the literature cited by FishBase (eg. Kottelat and Pethiyagoda 1990; Siebert 1997; Kottelat 2007; Pethiyagoda et al. 2008; Batuwita et al. 2013; Ng and Kottelat 2013), and so body sizes reported prior to these revisionary works were not used if these data could not be unambiguously assigned to a particular species.

To visualize the distribution of body sizes, body size data and log-transformed body size data were plotted as histograms in R, and skew was calculated using the skewness command in the R package moments (Komsta and Novomestky 2012).

Phylogenetic Analysis

The phylogenetic relationships of the Danionidae have been previously inferred with high taxon sampling using four genes: cytochrome *b* (cyt b), cytochrome oxidase I (COI), recombination activating gene-1 (RAG1), and Rhodoposin (Rh) (Tang et al. 2010). We acquired these sequences from GenBank and augmented these sequences with more recently published danionid sequences (Pramod et al. 2010; Liao et al. 2012; Collins et al. 2012; Kullander 2012). In total, analyses were performed on 284 tips representing 94 outgroup taxa and 190 tips representing Danionidae (Supplementary Material 1). Sequences for each gene were aligned within Geneious 6 by Clustal X and adjusted by eye (Larkin et al. 2007). The gene matrices were concatenated using SequenceMatrix (Vaidya et al. 2011). PartitionFinder v1.1.0 was used to test for the best partitioning scheme and models of substitution out of the models available for BEAST for each subset (divided by gene and codon position), with branch lengths linked, and BIC was used to select the best scheme (Lanfear et al. 2012).

We then performed relaxed clock phylogenetic analysis using BEAST 1.8.2 (Drummond et al. 2012). We constrained monophyly on clades that are well-supported as monophyletic: loaches (Botiidae, Vaillantellidae, Cobitidae, Nemacheilidae, Balitoridae), Catostomidae, Cypriniformes, Cyprinoidei, Characiphysi, and Gonorynchiformes. We did not constrain many clades internal to Cyprinoidei, to allow for the placement of taxa such as *Paedocypris* to be inferred freely among the Cypriniformes (except for exclusion from catostomids and loaches). Relatively few fossil cypriniforms of certain placement are known (Conway et al. 2010). Because of rapid recent developments in the phylogeny of cyprinoid fishes, membership of putative danionid fossils within the clade as currently recognized is uncertain. We used the only calibration point from the Fossil Calibration Database for ostariophysan fishes that has been justified as reliable (Ksepka et al. 2015). This calibration point is based on two fossils, the stem-chanid *Rubiesichthys gregalis* (Gonorynchiformes) which provides a minimum age of 126.3 Ma and the crown otocephalan *Tischlingerichthys* which provides a maximum age estimate of the Ostariophysi as 158.3 Ma (Benton et al. 2014). Given the presence of gonorynchiforms in the Tang et al. (2010) dataset as outgroup taxa, this node corresponds to calibration of the tree height. We set the minimum and maximum age estimates of Ostariophysi as the 2.5% to 97.5% percentiles of the lognormal prior for the age of the tree. MCMC chains were run for 200M generations and for two independent runs. Tracer 1.5 was used to assess convergence of independent runs. Results were manually sampled and burn-in excluded using bash commands, and TreeAnnotator was used to summarize trees and calculate the maximum clade credibility (MCC) tree.

Comparative phylogenetics

Data preparation.— We pruned the MCC phylogeny using ape in R (Paradis et al. 2004) to restrict comparative analyses to only the members of Danionidae for which we had body size data (Chapter 2,3). We pruned tips for species represented by multiple individuals down to a single individual. The resulting phylogeny included 123 species of Danionidae.

Rate shift inference.—To determine rate variation in body size evolution across Danionidae, we used the R package MOTMOT (Models of Trait Macroevolution on Trees) to identify clades or branches where shifts in the rate or direction of evolution occurred, using the traitMedusa 2 algorithm (Thomas and Freckleton 2011). The traitMedusa 2 algorithm tests for shifts in the rates of evolution of trait disparity (trait divergence among taxa) among clades as well as shifts in directional change at certain branches or clades, which are equivalent to shifts in variance and mean, respectively, of a Brownian Motion model (Thomas and Freckleton 2011; Puttick et al. 2014). We set the maximum number of shifts to 10 and set the minimum clade size as 1. Models with varying number of shifts are then compared, and the most likely model penalized by the number of additional parameters using ΔAICc is selected. Because the number of potential shift sites is a function of the number of taxa, following Thomas & Freckleton

(2011), we used 500 simulations of Brownian Motion on the MCC tree to determine the appropriate ΔAICc cut-off for the traitMedusa 2 algorithm as 10.53 to determine significance at $p < .05$.

Ancestral state reconstruction of miniaturization.–We used likelihood-based ancestral state reconstruction to estimate transitions between discrete non-miniature and miniature states. We use this to count the number of transitions to a miniature state (as in Rüber et al. 2007) and for subsequent visualization and analyses. We coded species as miniature or non-miniature based on our review of body size data for the Danionidae (Table 1), then used ace in ape to perform ancestral state reconstruction of this character (Paradis et al. 2004). We fit both the equal-rates (ER) model and the all-rates-different model (ARD), and determined that the ARD model did not fit significantly better than equal rates (ER) between states (log-likelihood ratio test; p=0.07566). Thus we present ancestral states estimated under the ER model and use these estimated states for subsequent analyses.

Ancestral state reconstruction of body size.–We reconstructed ancestral body size as a continuous character to assess the direction of evolution using phylogenetically independent contrasts using the ace command from ape in R (Paradis et al. 2004). We visualized ancestral states across time by mapping ancestral body sizes in traitspace using a traitgram (using the phenogram function implemented in phytools), allowing simple visualization of increases and decreases in body size on the y-axis over relative time along the x-axis (Revell 2013b). We also counted the number of consecutive branches along which body size decrease occurred to identify sustained miniaturization ((as done in Lee et al. 2014), focusing on branches leading to miniature species of Danionidae.

State-dependent rate variation of body size evolution.—We used three methods to study the relationship of character state on the rate of evolution. We tested if rates of body size evolution are influenced by miniaturization as a discrete character state and by body size as a continuous character.

Because miniaturization in fishes is also correlated with reduction in osteology, miniaturization may plausibly correlate with a transition to different biological and developmental constraints on evolution, which ultimately could affect rates of body size evolution. We tested whether the rate of body size evolution depends on miniature body size as a discrete character state using ML.RatePhylo in MOTMOT (O'Meara et al. 2006; Thomas et al. 2006; 2009). As in Thomas & Freckleton (2011), we used results of ancestral state reconstruction (inferred above) to assign all internal branches to either a miniature or non-miniature state, but here we did so for every tree sample. We expect mean body size to differ between miniature and non-miniature states. We tested both the model with a common mean between states and the model with multiple means to determine which had better fit to data. We used the likelihood ratio test to assess if differing rates between branches assigned to miniature and non-miniature states improved over the model with a single rate shared between states.

The direction, magnitude, and rate of body size evolution may be influenced by ancestral size as a continuous function. We implement two approaches to determine if our observed rates differ from that of a null expectation of a Brownian Motion model using simulations. The first of these approaches is a rate-by-state approach based on the ratebystate function in phytools. Ratebystate tests for a correlation between the squared phylogenetically independent contrasts of one trait against the ancestral states of a second trait (Revell 2013a). Because we are interested in the influence on a trait's rate by its own ancestral state, we slightly modified the ratebystate function. We estimated an observed Spearman's Rank correlation on the squared contrasts (using pic in phytools) and ancestral log body sizes at each node (using the pic method of the ace function in ape, as above). To assess the statistical significance of this correlation, we simulated 1000 instances of a single trait (using fastBM implemented in phytools), and for each instance we calculated the correlation of ancestral states and squared contrasts to generate a distribution of Spearman's Rank correlations under the null model. We determined the p-value of our observed correlation by determining the percentile of the correlation relative to the null distribution of correlations.

Second, we tested whether size evolution is influenced by the ancestral size value using ancestor-vs-change plots (Alroy 1998; 2000). For each branch, we plotted the estimated ancestral log body size against the rate of change along the branch. Here, rate of change along each branch is calculated as the amount of change that occurred between an ancestor and a descendent divided by branch length in millions of years, yielding Δln cm/million years (also called darwins, or evolutionary change per million years) (Albert and Johnson 2012). If body size has no influence on body size evolutionary rate, then descendants will be no more likely to be larger or smaller than their ancestors, and thus there will be no relationship between ancestral size and the magnitude or direction of change (ie. an average of zero, a slope of zero, and constant variance of rates of body size change across ancestral body sizes). We used linear regression implemented in R to estimate the slope of the relationship between ancestral state and descendent rates of evolution. To assess if our observed relationship between ancestor-vs-change was outside of the null expectation, we use simulations to generate a null distribution. We simulated 1000 datasets under Brownian Motion (using fastBM implemented in phytools) with the following parameters: 1) root state and σ^2 (using the R package geiger; Pennell et al. 2014) – under Brownian Motion, and 2) minimum and maximum bounds of ln 1.2 cm (.18 ln cm) and ln 30 cm (3.4 ln cm), respectively, based on the observed minimum and maximum of taxa included in our phylogeny. For each simulated dataset, we calculated ancestral body sizes and their descendant rates of body size change for each branch and use linear regression to estimate a slope of the relationship between ancestor-vs-change. Finally, we calculated the percentile of our observed slope relative to this simulated distribution to determine a p-value.

Simple models of trait evolution.—We also used fitContinuous to compare model fit between simple models of trait evolution. We fit Brownian Motion and Ornstein-Uhlenbeck models. The single-stationary peak Ornstein-Uhlenbeck model predicts body size would be attracted towards an adaptive optimum (Butler and King 2004).

RESULTS

Body Size Distribution

We obtained or calculated maximum standard lengths for 323 species of Danionidae. Most of the species in our dataset have data that match what is reported on FishBase (56.4% of the complete dataset; Table 2), but 17.4% do not have size data on FishBase, and 24.8% were different from FishBase (usually differences in FishBase are based on reidentifications, typographical errors in original mansucripts, and new information). 27 species are considered miniature because they do not exceed 26 mm SL or reach sexual maturity below 20 mm SL; three of these species have not been previously reported as miniature (Table 1).

Maximum body sizes in the Danionidae range from as small as 1.1 cm SL in *Danionella translucida* to as large as 38.5 cm SL in *Opsaridium microlepis* for the complete dataset. Body sizes form a continuous, unimodal distribution, with no clear separation between miniature taxa and non-miniature taxa. Untransformed body sizes are right-skewed (skewness = 2.06) with a mean of 7.9 cm SL (Fig. 1a). Log-transformed body sizes are more normal, but skewed somewhat left (skewness = -0.178), with a mean of 6.6 cm SL (1.9 log cm) (Fig. 1b).

123 species were in the phylogenetic analyses (38.1% of the total) and were used in downstream comparative analyses. Of these, a larger proportion of these are miniature species (19 spp.) compared to the complete dataset (15.5% vs. 8.4%), suggesting taxon sampling is biased towards miniature species. When restricted to the 123 species used for comparative analyses, body size ranged from 1.1 cm SL in *Danionella translucida* to 30.0 cm SL in *Raiamas guttatus*. The distribution is less right-skewed (skewness = 1.55) with a mean of 8.0 cm SL (Fig. 1c). The log-transformed distribution also had a slightly smaller left skew (skewness = -0.136) with a mean of 6.3 cm SL (1.8 log cm) (Fig. 1d). Right-skewed distributions for body size and log body size are most common in fishes, though a normal distribution for log-transformed body size is not uncommon (Albert and Johnson 2012).

Phylogenetic Analysis

Our analysis is broadly congruent with prior studies on the relationships of Danionidae (Fig. 2, Supplementary Material 2). As previously recovered, a number of genera are not monophyletic. One exception to congruence with prior studies is that Tang et al.'s (2010) phylogeny and our results differ on the placement of *Paedocypris* and *Sundadanio*. We recover these genera outside of Danionidae, which is congruent with analyses supported by multiple nuclear loci and mitogenomic data (Mayden and Chen 2010), as well as phylogenomic data (Stout et al. *in prep*.).

Comparative Phylogenetics

Rate shift analysis.—The pruned phylogeny includes 123 danionine taxa (Fig. 2). No rate shifts were reconstructed. There was no difference in model fit between the traitMedusa 2 model and Brownian Motion model (Table 3), which is not surprising given the traitMedusa 2 model for this data treatment had no rate shifts. Given that we reconstructed no rate shifts, we do not need to distinguish whether rate shifts in mean or variance of trait evolution (Puttick et al. 2014).

Number of miniaturization events.—For the 19 miniature species that were present in our phylogeny, we estimated ten miniaturization events and no transitions from a miniature to a non-miniature state. For genera that are entirely miniature, such as *Boraras*, *Danionella*, *Microdevario*, these transitions are on their stem branches, suggesting a single transition to miniaturization preceded the origin of each genus. We also reconstructed a transition to miniaturization for the branch leading to the common ancestor of *Danio margaritatus* and *D. erythromicron*, as well as six independent transitions along branches leading to tips, represented by *Brevibora dorsiocellata*, *Rasbora kalbarensis*, *Trigonostigma espei*, *Horadandia atukorali*, *Microrasbora rubescens*, and *Rasbosoma spilocerca*. Previous work identified at least seven transitions were required in the Danionidae (Rüber et al. 2007); with increased taxon sampling, we recovered additional transitions. Not all miniature species of Danionidae are represented in the phylogeny, so the number of miniaturization events is still underestimated; for example, if *Neobola bottegoi* is considered miniature, this represents an additional independent shift to miniaturization since it is the only miniature member of Chedrini.

Ancestral state reconstruction of body size.—We reconstructed ancestral log-body sizes on the chronogram for the Danionidae, and visualized these using a traitgram (Fig. 3). As expected, along all ten branches for which a transition to a miniature state was estimated, there is a decrease in body size. Usually, sustained size decrease leading to miniaturization is restricted to one or two branches. Sustained size decreases across more than two branches are found in just three instances: 1) there are four branches of sustained size decrease leading rootward from the tip representing the miniature species *Rasbosoma spilocerca*; 2) there are four branches of sustained size decrease between the root node of Danionidae and the base of the genus *Danionella*; 3) there are five branches of sustained size decrease between the ancestral node of *Rasbora* leading to the stem branch of *Boraras*. These sustained size decreases are not associated with any specific rate shifts, and therefore could be expected under neutral size evolution.

State-dependent rate variation.—We compared the rates of body size evolution between miniature and non-miniature species using ML.RatePhylo in MOTMOT. As expected, the model with different mean sizes between miniature and non-miniature states fits significantly better than the model with common means (Table 4;

phylogenetically-corrected mean size of miniatures estimated 0.95 ln cm = 2.6 cm; mean size of non-miniatures estimated 1.8 ln cm = 5.8 cm). We found that miniature species have a significantly reduced rate of size evolution (ML rate estimates: miniature = 0.3945, non-miniature = 3.1279; comparison with single rate model: p = 0.000955), supporting the hypothesis that miniaturization is accompanied by a reduced rate of body size evolution.

Rates of body size evolution as measured by squared contrasts were compared to ancestral body sizes for each node (Fig. 4). The correlation of rates of body size evolution to ancestral size is not significantly different from the null distribution (Spearman's rank correlation = 0.063, p = 0.485). Qualitatively, rates of body size evolution are more variable at intermediate body sizes, while rates are smaller at both larger (>20 cm SL = 3.0 log cm) and smaller body sizes (26 mm SL = .96 log cm; Fig. 4). The observation of reduced rates of evolution at small body sizes is supported by the ML.ratePhylo test. Qualitative observation of the distribution of contrasts relative to body size shows that both the smallest and largest taxa may be constrained in their rate of evolution by small body size. Although the correlation is non-significant, it appears that contrasts do vary relative to ancestral body size nonlinearly.

Ancestor-vs-change plots are useful for visualizing if direction and rate of change are dependent on ancestral body size. A linear regression was used to fit a line with intercept = 0.119 ln cm/million years and slope = -0.061 darwins/ln cm (Fig. 5). Mean slopes of our 1000 simulations of body size evolution under Brownian Motion were approximately zero, as expected (mean slope = -0.001), and the observed slope was significantly different from the null distribution (p = 0.001). We observed that small ancestors are slightly more likely to evolve an increase in body size and larger ancestors are slightly more likely to evolve a decrease in body size, with an intermediate optimal size for the Danionidae (ie. an ancestral body size where there is zero expected change) of 7.1 cm SL (1.8 log cm). This hypothetical optimal is near the observed mean size of Danionidae of 7.9 cm SL. Because descendants of species below 7 cm SL tend to be larger than their ancestor, there is no directional trend towards miniature body sizes. There also seems to be constraint on larger body sizes as well, with large-bodied fish more likely to decrease in body size than increase. The biased evolution away from the smallest and largest body sizes provide another visualization supporting the reduced rates found in the rate-by-state analysis. We tested the fit of the Ornstein-Uhlenbeck model (which predicts traits are attracted towards an optimum), and found it fit slightly worse than a Brownian Motion model to these data (ΔAIC = 2.1).

DISCUSSION

Our results provide deeper insight into the evolution of miniaturization (sensu Weitzman and Vari 1988) and its relationship to body size evolution. Miniaturization can affect rates of body size evolution because of specific ecological or physiological constraints (Hanken and Wake 1993; Miller 1996). Alternatively, miniaturization can also provide novel ecological opportunities and release from developmental constraints, which could increase rates of trait evolution (Hanken and Wake 1993; Miller 1996). There is support for constrained evolution at smaller body sizes, although not all tests were consistent. Even though no rate shifts were predicted by the traitMedusa 2 model, when using ML.RatePhylo to test for rate variation dependent on miniature state, we found support for a pattern of rate decrease in smaller body sizes.

We also find smaller and larger body sizes have different rates from intermediate body sizes. The rate-by-state plot shows constrained rates with lower variation at smaller and larger than intermediate body size. The ancestor-vs-change plot shows smaller ancestors tend to evolve larger and larger ancestors tend to evolve smaller. These results do not support an attraction towards smaller body sizes in the Danionidae. In summary, there is evidence that body size evolution has slowed at smaller body sizes, and potentially some indication of reduced rates at larger body sizes.

Miniaturization has also been defined as a decrease in body size over time (Avaria-Llautureo et al. 2012; Lee et al. 2014), but neutral body size evolution does not preclude sustained decreases in body size. While we uncover some sustained miniaturization across consecutive branches in the Danionidae, this size decrease does not appear to be explained by shifts in rates or means of body size evolution. Miniaturization in fishes is not coincident with a directional shift towards smaller body size along any particular branches.

Body size distributions are the result of macroevolutionary processes. It has been suggested that body size distribution in fishes is better explained by diffusion in a power-log scale rather than a log scale (Albert and Johnson 2012). We find this unnecessary to explain the body size distribution in the Danionidae. First, log-transformation is enough to roughly normalize the distribution of body sizes without invoking a power term. Rate-shift analysis inferred zero shifts and did not fit better than a Brownian Motion model, demonstrating good fit for body size evolution on a log-scale diffusing at a constant rate across time. Danionidae differs from many other fish groups for having a relatively normally-distributed (vs. right-skewed) log-transformed body size distribution. The typical right-skewed distribution of log-transformed body size may be related to increasing extinction risk, which increases with size if all other factors are equal (Stanley 1973; Clauset and Erwin 2008). A more symmetric log-transformed size distribution, such as that seen in Danionidae, may indicate that extinction rates in the Danionidae are size-independent (all else equal), and that changes in body size have few selective advantages (Clauset and Erwin 2008).

The lack of increased rates of body size evolution leading to miniature species of Danionidae demonstrates that the high diversity of miniature danionines and the high number of independent transitions to miniaturized state are the result of evolution from species that are already relatively small, and thus require no increase in evolutionary rate. This has implications for our understanding of miniaturization. Certain fish groups may be preadapted to reach a miniature size (Albert and Johnson 2012), but this may not be due to a biased evolvability towards smaller body size, but rather may simply starting from a smaller ancestral body size. It also has implications for our understanding of miniaturization in the context of convergent evolution (Rundell and Leander 2010). Convergence is often thought of as a driven process towards a common phenotype, however the miniature phenotypes herein are well-explained by gradual, neutral processes.

There are few paleontological clues for how body size evolved within the Danionidae because the fossil record of cyprinoids is relatively scant (Conway et al. 2010), and the taxonomic placement of many fossil cyprinoids have been rendered uncertain following recent progress in cypriniform phylogenetics. What was once considered Danionidae also comprises taxa that are morphological similar but evolutionarily distinct (e.g. "Ex-Danioninae" Tang et al. 2010), and thus assignments of fossils to Danionidae should be re-interpreted in the light of new morphological phylogenetic data (Conway 2009; Liao et al. 2011). In the future, accurate assignment of cyprinoid fossils to their respective clades can provide more data to improve time-calibration and analyses of trait evolution. Inclusion of fossil taxa can greatly increase the accuracy of estimation of body size evolution, and is the only way to detect a clade-wide trend in body size evolution (Bokma et al. 2016).

The finding that *Paedocypris* and *Sundadanio* are not part of Danionidae is in contrast to the main study these data are derived from, however support values for the placement of *Paedocypris* and *Sundadanio* as part of Danionidae and Danioninae were low (Tang et al. 2010). Our study differs somewhat by slightly increased taxon sampling of the Danionidae and using a best-fit partitioning of genes and codon positions, which can affect the accuracy of phylogenetic analysis (Zwickl & Hillis 2002; Lanfear et al. 2013). Phylogenetic analysis using relaxed clock models can also increase the accuracy in estimating topology (Drummond et al. 2006), although reanalysis under RAxML also does not result in the published relationships from these data (not presented). The exclusion of *Paedocypris* and *Sundadanio* from Danionidae is confirmed by six nuclear genes and mitogenomic data (Mayden & Chen 2010). Previous exploration of cytochrome b, RAG1, and Rh suggests phylogenetic signal in the placement of these taxa is weak for these genes, and that these data are not decisive on the relationships (Britz et al. 2014). Our results are additionally confirmed by phylogenomic data, both on complete datasets and when systematic error is reduced (Chapters 2, 3).

There are few previous age estimates for the age of Danionidae or its members. Nakatani et al. (2011) reconstructed the divergence of Danionidae (represented by *Danio*) from Cyprinidae (*Cyprinus*) and Psilorhynchidae (*Psilorhynchus*) at approximately 125 Ma. Near et al. (2012) reconstructed the divergence of Danionidae (*Danio rerio*) from Xenocyprididae (*Opsariichthys uncirostris*) at approximately 70 Ma. Chen et al. (2013), with five species of Danionidae representing the major clades, reconstructed the most recent common ancestor of Danionidae as approximately 65 Ma and the divergence of Danionidae from other cyprinoids at approximately 70 Ma. We recover an absolute age estimate for the most recent common ancestor of Danionidae at approximately 107 Ma, within the range of previous studies. Our age estimates may be somewhat older because of the inclusion of mitochondrial genes, which are known to overestimate ages relative to nuclear genes because of saturation (Dornburg et al. 2015). It is the relative age estimates, however, that are important for reconstructing relative rates of body size evolution.

Conclusion

In this study, we explored the dynamics of rates of body size evolution in a group including many miniature fishes. It is important to document the extremes of biodiversity, and highlighting miniature species is of interest to biologists and to the general public. But, we must not let these bias our interpretations of how body size evolves. While Danionidae has a high propensity for miniaturization, this may be due to have a small ancestral body size, rather than an increase in evolutionary rate in body size or a trend towards decreased body size. Rates of body size evolution in Danionidae appear to be constrained in miniature species, and the generality of this phenomenon in miniaturization deserves further exploration.

Acknowledgements

We would like to thank GH Thomas for answering questions on using MOTMOT, and LJ Revell for helpful responses to inquiries on the r-sig-phylo mailing list that contributed to this publication. We would also like to thank two anonymous reviewers for substantive comments and criticisms that increased the quality of this work. We would also like to thank JE Bond, MM Hayes, and CC Stout for helpful discussions of this research and ED Burress, PB Hart, MM Hayes, and CC Stout for reviewing versions of this manuscript. MT graciously thanks the American Society of Ichthyologists and Herpetologists for recognition for an earlier version of this research. This work was supported by Auburn University (Cellular and Molecular Biosciences Peaks of Excellence Research Fellowship to MT); and by National Science Foundation (DEB-0315963, DEB-0107751, DEB-1022720) to JWA.

References

Albert, J. S., and D. M. Johnson. 2012. Diversity and Evolution of Body Size in Fishes. Evol Biol 39:324–340.

Alroy, J. 1998. Cope's Rule and the Dynamics of Body Mass Evolution in North American Fossil Mammals. Science 280:731–734.

Alroy, J. 2000. Understanding the dynamics of trends within evolving lineages. Paleobiology 26:319–329.

Avaria-Llautureo, J., C. E. Hernández, D. Boric-Bargetto, C. B. Canales-Aguirre, B. Morales-Pallero, and E. Rodríguez-Serrano. 2012. Body Size Evolution in Extant Oryzomyini Rodents: Cope's Rule or Miniaturization? PLOS ONE 7:e34654.

Barman, R. P. 1991. A taxonomic revision of the Indo-Burmese species of *Danio* Hamilton Buchanan (Pisces: Cyprinidae). Rec. Zool. Soc. India 137:1–91. Zoological Survey of India.

Batuwita, S., M. de Silva, and U. Edirisinghe. 2013. A review of the danionine genera *Rasboroides* and *Horadandia* (Pisces: Cyprinidae), with description of a new species from Sri Lanka. Ichthyol Explor Freshwaters 24:121–140.

Bennett, M. G., and K. W. Conway. 2010. An overview of North America's diminutive freshwater fish fauna. Ichthyol Explor Freshwaters 21:63.

Benton, M. J., P. C. J. Donoghue, R. J. Asher, M. Friedman, T. J. Near, and J. Vinther. 2014. Constraints on the timescale of animal evolutionary history. Palaentologica Electronica 18.1.1FC:1–106.

Boettiger, C., D. T. Lang, and P. C. Wainwright. 2012. rfishbase: exploring, manipulating and visualizing FishBase data from R. J Fish Biol 81:2030–2039.

Bokma, F., M. Godinot, O. Maridet, S. Ladevèze, L. Costeur, F. Solé, E. Gheerbrant, S. Peigné, F. Jacques, and M. Laurin. 2016. Testing for Depéret's Rule (Body Size Increase) in Mammals using Combined Extinct and Extant Data. Syst Biol 65:98–108. Oxford University Press.

Britz, R. 2003. *Danionella mirifica*, a new species of miniature fish from Upper Myanmar (Ostariophysi: Cyprinidae). Ichthyol Explor Freshwaters 14:217–222. VERLAG DR. FRIEDRICH PFEIL.

Britz, R. 2009. *Danionella priapus*, a new species of miniature cyprinid fish from West Bengal, India (Teleostei: Cypriniformes: Cyprinidae). Zootaxa 2277:53–60.

Britz, R., and K. W. Conway. 2016. *Danionella dracula*, an escape from the cypriniform *Bauplan* via developmental truncation? J. Morphol. 277:147–166.

Britz, R., and K. W. Conway. 2009. Osteology of *Paedocypris*, a miniature and highly developmentally truncated fish (Teleostei: Ostariophysi: Cyprinidae). J. Morphol. 270:389–412.

Britz, R., K. W. Conway, and L. Rüber. 2014. Miniatures, morphology and molecules: *Paedocypris* and its phylogenetic position (Teleostei, Cypriniformes). Zool J Linn Soc 172:556–615.

Britz, R., K. W. Conway, and L. Rüber. 2009. Spectacular morphological novelty in a miniature cyprinid fish, *Danionella dracula* n. sp. Proc. R. Soc. London Ser. B 276:2179–2186.

Butler, M., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. The American Naturalist 164:683–695. JSTOR.

Calegari, B. B., R. E. dos Reis, and R. P. Vari. 2014. Miniature catfishes of the genus *Gelanoglanis* (Siluriformes: Auchenipteridae): monophyly and the description of a new species from the upper rio Tapajós basin, Brazil. Neotrop Ichthyol 12:699–706.

Clauset, A., and D. H. Erwin. 2008. The Evolution and Distribution of Species Body Size. Science 321:399–401.

Collins, R. A., K. F. Armstrong, R. Meier, Y. Yi, S. D. J. Brown, R. H. Cruickshank, S. Keeling, and C. Johnston. 2012. Barcoding and Border Biosecurity: Identifying Cyprinid Fishes in the Aquarium Trade. PLOS ONE 7:e28381.

Conway, K. W. 2005. Monophyly of the genus *Boraras* (Teleostei: Cyprinidae). Ichthyol Explor Freshwaters 16:249. VERLAG DR. FRIEDRICH PFEIL.

Conway, K. W., and M. Kottelat. 2011. *Boraras naevus*, a new species of miniature and sexually dichromatic freshwater fish from peninsular Thailand (Ostariophysi: Cyprinidae). Zootaxa 3002:45–51.

Conway, K. W., and T. Moritz. 2006. *Barboides britzi*, a new species of miniature cyprinid from Benin (Ostariophysi: Cyprinidae), with a neotype designation for *B. gracilis*. Ichthyol Explor Freshwaters 17:73. VERLAG DR. FRIEDRICH PFEIL.

Conway, K. W., M. V. Hirt, L. Yang, R. L. Mayden, and A. M. Simons. 2010. Cypriniformes: systematics and paleontology. Origin and Phylogenetic Interrelationships of Teleosts 295–316.

Conway, K. W., W. J. Chen, and R. L. Mayden. 2008. The "Celestial Pearl danio" is a miniature *Danio* (s.s) (Ostariophysi: Cyprinidae): evidence from morphology and molecules. Zootaxa 1686:1–28.

Cope, E. D. 1904. The primary factors of organic evolution. Open Court Publishing, Chicago, IL.

Denys, G. P. J., P. A. Tedesco, T. Oberdorff, and P. Gaubert. 2014. Environmental correlates of body size distribution in Cyprinidae (Actinopterygians) depend on phylogenetic scale. Ecol Freshw Fish n/a–n/a.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973.

Eschmeyer, W. N., R. Fricke, and R. van der Laan. 2016. Catalog of Fishes: Genera, Species, References.

Fang, F., M. Norén, T. Y. Liao, M. Källersjö, and S. O. Kullander. 2009. Molecular phylogenetic interrelationships of the south Asian cyprinid genera *Danio*, *Devario* and

Microrasbora (Teleostei, Cyprinidae, Danioninae). Zool Scripta 38:237–256.

Frigot, R. A., A. Goswami, B. Andres, R. J. Butler, and R. B. J. Benson. 2014. Competition and constraint drove Cope’s rule in the evolution of giant flying reptiles. Nat Commun 5:1–8. Nature Publishing Group.

Gould, S. J. 1997. Cope's rule as psychological artefact. Nature 385:199–200.

Gould, S. J. 1988. Trends as changes in variance: a new slant on progress and directionality in evolution. J Paleo 62:319–329.

Hanken, J., and D. B. Wake. 1993. Miniaturization of body size: organismal consequences and evolutionary significance. Annu. Rev. Ecol. Syst. 501–519. JSTOR.

Hardman, M., and L. M. Hardman. 2008. The Relative Importance of Body Size and Paleoclimatic Change as Explanatory Variables Influencing Lineage Diversification Rate: An Evolutionary Analysis of Bullhead Catfishes (Siluriformes: Ictaluridae). Syst Biol 57:116–130.

Heim, N. A., M. L. Knope, E. K. Schaal, S. C. Wang, and J. L. Payne. 2015. Cope's rule in the evolution of marine animals. Science 347:867–870.

Hora, S. L., and D. D. Mukerji. 1928. Notes on fishes in the Indian Museum. XVI. On fishes of the genus Esomus Swainson. Records of the Indian Museum 30:41–56.

Howe, K., M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot, M. Muffato, J. E. Collins, S. Humphray, K. McLaren, L. Matthews, S. McLaren, I. Sealy, M. Caccamo, C. Churcher, C. Scott, J. C. Barrett, R. Koch, G.-J. Rauch, S. White, W. Chow, B. Kilian, L. T. Quintais, J. A. Guerra-Assunção, Y. Zhou, Y. Gu, J. Yen, J.-H. Vogel, T. Eyre, S. Redmond, R. Banerjee, J. Chi, B. Fu, E. Langley, S. F. Maguire, G. K. Laird, D. Lloyd, E. Kenyon, S. Donaldson, H. Sehra, J. Almeida-King, J. Loveland, S. Trevanion, M. Jones, M. Quail, D. Willey, A. Hunt, J. Burton, S. Sims, K. McLay, B. Plumb, J. Davis, C. Clee, K. Oliver, R. Clark, C. Riddle, D. Eliott, G. Threadgold, G. Harden, D. Ware, B. Mortimer, G. Kerry, P. Heath, B. Phillimore, A. Tracey, N. Corby, M. Dunn, C. Johnson, J. Wood, S. Clark, S. Pelan, G. Griffiths, M. Smith, R. Glithero, P. Howden, N. Barker, C. Stevens, J. Harley, K. Holt, G. Panagiotidis, J. Lovell, H. Beasley, C. Henderson, D. Gordon, K. Auger, D. Wright, J. Collins, C. Raisen, L. Dyer, K. Leung, L. Robertson, K. Ambridge, D. Leongamornlert, S. McGuire, R. Gilderthorp, C. Griffiths, D. Manthravadi, S. Nichol, G. Barker, S. Whitehead, M. Kay, J. Brown, C. Murnane, E. Gray, M. Humphries, N. Sycamore, D. Barker, D. Saunders, J. Wallis, A. Babbage, S. Hammond, M. Mashreghi-Mohammadi, L. Barr, S. Martin, P. Wray, A. Ellington, N. Matthews, M. Ellwood, R. Woodmansey, G. Clark, J. Cooper, A. Tromans, D. Grafham, C. Skuce, R. Pandian, R. Andrews, E. Harrison, A. Kimberley, J. Garnett, N. Fosker, R. Hall, P. Garner, D. Kelly, C. Bird, S. Palmer, I. Gehring, A. Berger, C. M. Dooley, Z. Ersan-Ürün, C. Eser, H. Geiger, M. Geisler, L. Karotki, A. Kirn, J. Konantz, M. Konantz, M. Oberländer, S. Rudolph-Geiger, M. Teucke, K. Osoegawa, B. Zhu, A. Rapp, S. Widaa, C. Langford, F. Yang, N. P. Carter, J. Harrow, Z. Ning, J. Herrero, S. M. J. Searle, A. Enright, R. Geisler, R. H. A. Plasterk, C. Lee, M. Westerfield, P. J. de Jong, L. I. Zon, J.

H. Postlethwait, C. Nüsslein-Volhard, T. J. P. Hubbard, H. Roest Crollius, J. Rogers, and D. L. Stemple. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503.

Jiang, Y.-E., X. Y. Chen, and J. X. Yang. 2008. *Microrasbora* Annandale, a new genus record in China, with description of a new species (Teleostei: Cyprinidae). Env Biol Fish 83:299–304.

Knouft, J. H., and L. M. Page. 2003. The Evolution of Body Size in Extant Groups of North American Freshwater Fishes: Speciation, Size Distributions, and Cope's Rule. The American Naturalist 161:413–421.

Komsta, L., and F. Novomestky. 2012. moments: Moments, cumulants, skewness, kurtosis and related tests.

Kottelat, M. 2001. Fishes of Laos. WHT Publications.

Kottelat, M. 2007. *Rasbora dies*, a new species of cyprinid fish from eastern Borneo (Teleostei: Cyprinidae). Ichthyol Explor Freshwaters 18:301. VERLAG DR. FRIEDRICH PFEIL.

Kottelat, M. 2013. The fishes of the inland waters of Southeast Asia: a catalogue and core bibliography of the fishes known to occur in freshwaters, mangroves and estuaries. Raffles B Zool 27:1–663.

Kottelat, M., A. J. Whitten, S. N. Kartikasari, and S. Wirjoatmodjo. 1993. Freshwater Fishes of Western Indonesia and Sulawesi. Periplus Editions, Indonesia.

Kottelat, M., and C. Vidthayanon. 1993. *Boraras micros*, a new genus and species of minute freshwater fish from Thailand (Teleostei: Cyprinidae). Ichthyol Explor Freshwaters 4:161–176.

Kottelat, M., and K.-E. Witte. 1999. Two new species of *Microrasbora* from Thailand and Myanmar, with two new generic names for small Southeast Asian cyprinid fishes (Teleostei: Cyprinidae). J. South Asian Nat. Hist. 4:49–56.

Kottelat, M., and R. Pethiyagoda. 1990. *Danio pathirana*, a new species of cyprinid fish endemic to southern Sri Lanka. Ichthyol Explor Freshwaters 4:161–176.

Kottelat, M., R. Britz, H. H. Tan, and K.-E. Witte. 2006. *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. Proc. R. Soc. London Ser. B 273:895–899.

Ksepka, D. T., J. F. Parham, J. F. Allman, M. J. Benton, M. T. Carrano, K. A. Cranston, P. C. J. Donoghue, J. J. Head, E. J. Hermsen, R. B. Irmis, W. G. Joyce, M. Kohli, K. D. Lamm, D. Leehr, J. L. Patané, P. D. Polly, M. J. Phillips, N. A. Smith, N. D. Smith, M. van Tuinen, J. L. Ware, and R. C. M. Warnock. 2015. The Fossil Calibration Database—A New Resource for Divergence Dating. Syst Biol 64:853–859.

Kullander, S. O. 2012. Description of *Danio flagrans*, and redescription of *D. choprae*, two closely related species from the Ayeyarwaddy River drainage in northern Myanmar (Teleostei: Cyprinidae). Ichthyol Explor Freshwaters 23:245–262.

Kullander, S. O., and F. Fang. 2009. *Danio tinwini*, a new species of spotted danio from northern Myanmar (Teleostei: Cyprinidae). Ichthyol Explor Freshwaters 20:223–228.

Lanfear, R., B. Calcott, S. Y. W. Ho, and S. Guindon. 2012. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Mol Biol Evol 29:1695–1701.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

Lee, M. S. Y., A. Cau, D. Naish, and G. J. Dyke. 2014. Sustained miniaturization and anatomical innovation in the dinosaurian ancestors of birds. Science 345:562–566.

Lévêque, C., and J. Daget. 1984. Cyprinidae. Pp. 217–342 *in* Check-list of the freshwater fishes of Africa. Office de la Recherche Scientifique et Technique Outre-Mer and Musée Royal de l'Mrique Centrale, Tongeren, Belgium.

Liao, T. Y., and H. H. Tan. 2011. *Brevibora cheeya*, a new species of cyprinid fish from Malay Peninsula and Sumatra. Raffles B Zool 59:77–82.

Liao, T. Y., and H. H. Tan. 2014. *Brevibora exilis*, a new rasborin fish from Borneo (Teleostei: Cyprinidae). Ichthyol Explor Freshwaters 24:209–215.

Liao, T. Y., J. Arroyave, and M. L. J. Stiassny. 2012. Diagnosis of Asian *Raiamas* (Teleostei: Cyprinidae: Chedrina) with comments on chedrin relationships and previously proposed diagnostic characters for *Opsaridium* and *Raiamas*. Ichthyol Res 59:328–341.

Liao, T. Y., S. O. Kullander, and F. Fang. 2011. Phylogenetic position of rasborin cyprinids and monophyly of major lineages among the Danioninae, based on morphological characters (Cypriniformes: Cyprinidae). J Zoolog Syst Evol Res 49:224–232.

Martin, S. D., and R. M. Bonett. 2015. Biogeography and divergent patterns of body size disparification in North American minnows. Mol Phylogenet Evol 93:17–28. Elsevier Inc.

Mayden, R. L., and W. J. Chen. 2010. The world"s smallest vertebrate species of the genus *Paedocypris*: a new family of freshwater fishes and the sister group to the world"s most diverse clade of freshwater fishes (Teleostei: Cypriniformes). Mol Phylogenet Evol 57:152–175.

McClain, C. R., M. A. Balk, M. C. Benfield, T. A. Branch, C. Chen, J. Cosgrove, A. D.

M. Dove, L. C. Gaskins, R. R. Helm, F. G. Hochberg, F. B. Lee, A. Marshall, S. E. McMurray, C. Schanche, S. N. Stone, and A. D. Thaler. 2015. Sizing ocean giants: patterns of intraspecific size variation in marine megafauna. PeerJ 2:e715.

McCluskey, B. M., and J. H. Postlethwait. 2015. Phylogeny of zebrafish, a "model species," within Danio, a "model genus". Mol Biol Evol 32:635–652.

Miller, P. J. (ed). 1996. Miniature Vertebrates: Implications of Small Body Size. Oxford University Press, New York.

Ng, H. H., and M. Kottelat. 2013. The identity of the cyprinid fishes *Rasbora dusonensis* and *R. tornieri* (Teleostei: Cyprinidae). Zootaxa 3635:62.

O'Meara, B. C., C. Ane, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.

Pennell, M. W., J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn, M. E. Alfaro, and L. J. Harmon. 2014. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 30:2216–2218.

Pethiyagoda, R. 1991. Freshwater fishes of Sri Lanka. Wildlife Heritage Trust of Sri Lanka, Colombo, Sri Lanka.

Pethiyagoda, R., M. Kottelat, A. Silva, K. Maduwage, and M. Meegaskumbura. 2008. A review of the genus *Laubuca* in Sri Lanka, with description of three new species (Teleostei: Cyprinidae). Ichthyol Explor Freshwaters 19:7–26.

Polly, P. D., and J. Alroy. 1998. Cope's Rule. Science 282:47.

Pramod, P., F. Fang, K. Rema Devi, T. Y. Liao, T. J. Indra, K. Jameela Beevi, and S. O. Kullander. 2010. *Betadevario ramachandrani*, a new danionine genus and species from the Western Ghats of India (Teleostei: Cyprinidae: Danioninae). Zootaxa 2519:31–47.

Puttick, M. N., G. H. Thomas, and M. J. Benton. 2014. High rates of evolution preceded the origin of birds. Evolution 68:1497–1510.

Rabosky, D. L., F. Santini, J. Eastman, S. A. Smith, B. Sidlauskas, J. Chang, and M. E. Alfaro. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nat Commun 4:1958.

Rainboth, W. J. 1996. Fishes of the Cambodian Mekong. Food & Agriculture Organization of the United Nations, Rome.

Revell, L. J. 2013a. Investigating whether the rate of one continuous trait is influenced by the state of another (a somewhat *ad hoc* approach).

Revell, L. J. 2013b. Two new graphical methods for mapping trait evolution on phylogenies. Methods Ecol Evol 4:754–759. Wiley Online Library.

Roberts, T. R. 1986. *Danionella translucida*, a new genus and species of cyprinid fish from Burma, one of the smallest living vertebrates. Env Biol Fish 16:231–241. Springer.

Roberts, T. R. 2007. The "celestial pearl danio," a new genus and species of colourful minute cyprinid fish from Myanmar (Pisces: Cypriniformes). Raffles B Zool 55:131–140.

Rundell, R. J., and B. S. Leander. 2010. Masters of miniaturization: Convergent evolution among interstitial eukaryotes. BioEssays 32:430–437.

Rüber, L., M. Kottelat, H. H. Tan, P. K. L. Ng, and R. Britz. 2007. Evolution of miniaturization and the phylogenetic position of *Paedocypris*, comprising the world's smallest vertebrate. BMC Evol Biol 7:38.

Sander, P. M., A. Christian, M. Clauss, R. Fechner, C. T. Gee, E.-M. Griebeler, H.-C. Gunga, J. Hummel, H. Mallison, S. F. Perry, H. Preuschoft, O. W. M. Rauhut, K. Remes, T. Tütken, O. Wings, and U. Witzel. 2010. Biology of the sauropod dinosaurs: the evolution of gigantism. Biol Rev 86:117–155.

Siebert, D. J. 1997. The identities of *Rasbora paucisqualis* Ahl in Schreitmuller, 1935, and *Rasbora bankanensis* (Bleeker, 1853), with the designation of a lectotype for *R. paucisqualis* (Teleostei: Cyprinidae). Raffles B Zool 45:29–37. NATL UNIV SINGAPORE, SCHOOL BIOLOGICAL SCIENCES DEPT ZOOLOGY, KENT RIDGE, SINGAPORE 0511, SINGAPORE.

Stanley, S. M. 1973. An explanation for Cope's rule. Evolution 27:1–26. JSTOR.

Tang, K. L., M. K. Agnew, M. V. Hirt, T. Sado, L. M. Schneider, J. Freyhof, Z. Sulaiman, E. Swartz, C. Vidthayanon, M. Miya, K. Saitoh, A. M. Simons, R. M. Wood, and R. L. Mayden. 2010. Systematics of the subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). Mol Phylogenet Evol 57:189–214. Elsevier Inc.

Thomas, G. H., and R. P. Freckleton. 2011. MOTMOT: models of trait macroevolution on trees. Methods Ecol Evol 3:145–151.

Thomas, G. H., R. P. Freckleton, and T. Székely. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proc. Biol. Sci. 273:1619–1624.

Thomas, G. H., S. Meiri, and A. B. Phillimore. 2009. Body size diversification in *Anolis*: novel environment and island effects. Evolution 63:2017–2030.

Tilak, R., and A. Husain. 1990. Description of a new Cyprinid, *Barilius dimorphicus* (Subfamily: Rasborinae) from Rajaji National Park, Uttar Pradesh. J Bomb Nat Hist Soc 87:102–105.

Toledo-Piza, M., G. M. T. Mattox, and R. Britz. 2014. Priocharax nanus, a new miniature characid from the rio Negro, Amazon basin (Ostariophysi: Characiformes), with an updated list of miniature Neotropical …. Neotrop Ichthyol 12:229–246.

Vaidya, G., D. J. Lohman, and R. Meier. 2011. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27:171–180.

Vidthayanon, C. 2008. Field guide to fishes of the Mekong Delta. Mekong River Commission, Vientiane, Laos.

Weitzman, S. H., and R. P. Vari. 1988. Miniaturization in South American freshwater fishes; and overview and discussion. Proc Biol Soc Wash 101:444–465.

Table 1 (see following page). Miniature species of Danionidae based on a review of the primary literature. Species listed do not exceed 26 mm SL or are known to reach sexual maturity below 20 mm SL. Larger body sizes that originate from aquarist literature or could not be unambiguously assigned to species after taxonomic revision of a species are not listed (see methods for details). References are listed for species that have previously been explicitly described or listed as miniature; if not previously listed as miniature, reference left blank. Species marked with asterisks represent the 19 species included in the phylogeny and comparative phylogenetic analysis.

Species	Max SL (mm)	Reference for size	Reference for miniature status
Chedrinae			
Neobola stellae	23	(Lévêque and Daget 1984)	
Rasborinae			
* *Boraras brigittae*	18	(Kottelat and Vidthayanon 1993)	(Kottelat and Vidthayanon 1993)
* *Boraras maculatus*	20	(Kottelat and Vidthayanon 1993)	(Kottelat and Vidthayanon 1993)
* *Boraras merah*	20	(Kottelat et al. 1993)	(Kottelat and Vidthayanon 1993)
Boraras micros	13.3	(Kottelat and Vidthayanon 1993)	(Kottelat and Vidthayanon 1993)
Boraras naevus	12.7	(Conway and Kottelat 2011)	(Conway and Kottelat 2011)
* *Boraras urophthalmoides*	20	(Vidthayanon 2008)	(Kottelat and Vidthayanon 1993)
* *Brevibora dorsiocellata*	23	(Liao and Tan 2011)	(Liao and Tan 2011)
Brevibora exilis	24.5	(Liao and Tan 2014)	
* *Horadandia atukorali*	19.3	(Batuwita et al. 2013)	(Kottelat and Vidthayanon 1993)
Horadandia brittani	20.4	(Batuwita et al. 2013)	
* *Rasbora kalbarensis*	25	(Kottelat et al. 1993)	(Kottelat and Vidthayanon 1993)
* *Rasbosoma spilocerca*	26	(Kottelat 2001)	(Kottelat and Vidthayanon 1993)
* *Trigonostigma espei*	25	(Rainboth 1996)	
Trigonostigma somphongsi	19	(Kottelat and Vidthayanon 1993)	(Kottelat and Vidthayanon 1993)
Danioninae			
* *Danio erythromicron*	27.6	(Kottelat and Witte 1999)	(Kottelat and Vidthayanon 1993)
* *Danio margaritatus*	21.2	(Roberts 2007)	(Conway et al. 2008)
Danio tinwini	25.6	(Kullander and Fang 2009)	(Kullander and Fang 2009)
* *Danionella dracula*	16.7	(Britz et al. 2009)	(Britz et al. 2009)
* *Danionella mirifica*	14.1	(Britz 2003)	(Britz 2003)
* *Danionella priapus*	16	(Britz 2009)	(Britz 2009)
* *Danionella translucida*	12	(Roberts 1986)	(Kottelat and Vidthayanon 1993)
* *Microdevario gatesi*	23	(Jiang et al. 2008)	(Kottelat and Vidthayanon 1993)
* *Microdevario kubotai*	19	(Kottelat and Witte 1999)	(Kottelat and Vidthayanon 1993)
Microdevario microphthalma	25.7	(Jiang et al. 2008)	
* *Microdevario nana*	15.2	(Kottelat and Witte 1999)	(Kottelat and Vidthayanon 1993)
* *Microrasbora rubescens*	30	(Jiang et al. 2008)	(Kottelat and Vidthayanon 1993)

Table 2. Proportions of body size dataset that matched or did not match FishBase reported sizes (accessed October 16 2014). Except for the totals, all numbers also are given with proportion of the complete dataset or the subset of the dataset used for comparative phylogenetic analysis, depending on the column.

		Complete Dataset	Comparative Dataset
	Total	323 spp.	123 spp.
	Miniatures	27 (8.4%)	19 (15.5%)
	Matched data from FishBase	186 (56.4%)	76 (61.8%)
Differences from FishBase	No data on FishBase at all	57 (17.4%)	5 (4.1%)
	Larger body size was found in the literature	27 (8.4%)	15 (12.2%)
	Calculated SL from TL reported on FishBase smaller than maximum SL reported *	24 (7.3%)	9 (7.3%)
	Species identities have changed **	9 (2.7%)	7 (5.7%)
	FishBase cites aquarium literature	9 (2.7%)	7 (5.7%)
	FishBase has no citation for body size entry	7 (2.1%)	1 (0.8%)
	FishBase's body size does not match the original citation	4 (1.2%)	3 (2.4%)

* FishBase reports a maximum size in TL that is a larger number than the maximum reported SL, but converting this TL to an SL resulted in a smaller body size than the maximum reported SL. We consider this separate because these discrepancies are more cryptic.

** Species identity has changed, so that body sizes attributed previously to it no longer do, or cannot be unambiguously assigned to this species.

Table 3. AICc of fit for various models of trait evolution. Brownian Motion (BM), Ornstein-Uhlenbeck (OU), and drift models were fit using fitContinuous (geiger). The traitMedusa 2 algorithm (MOTMOT) was used to find the best-fit model incorporating rate shifts, and was generally the best-fit model (lowest AICc). ML.ratePhylo (MOTMOT) was used to test the model of differing rates between miniature and non-miniature discrete states, with means either different between states or the same between states, and were only tested on the chronogram.

Brownian Motion	173.959
traitMedusa 2	173.964
Ornstein-Uhlenbeck	176.061
ML.ratePhylo (different means)	135.553
ML.ratePhylo (common means)	171.943

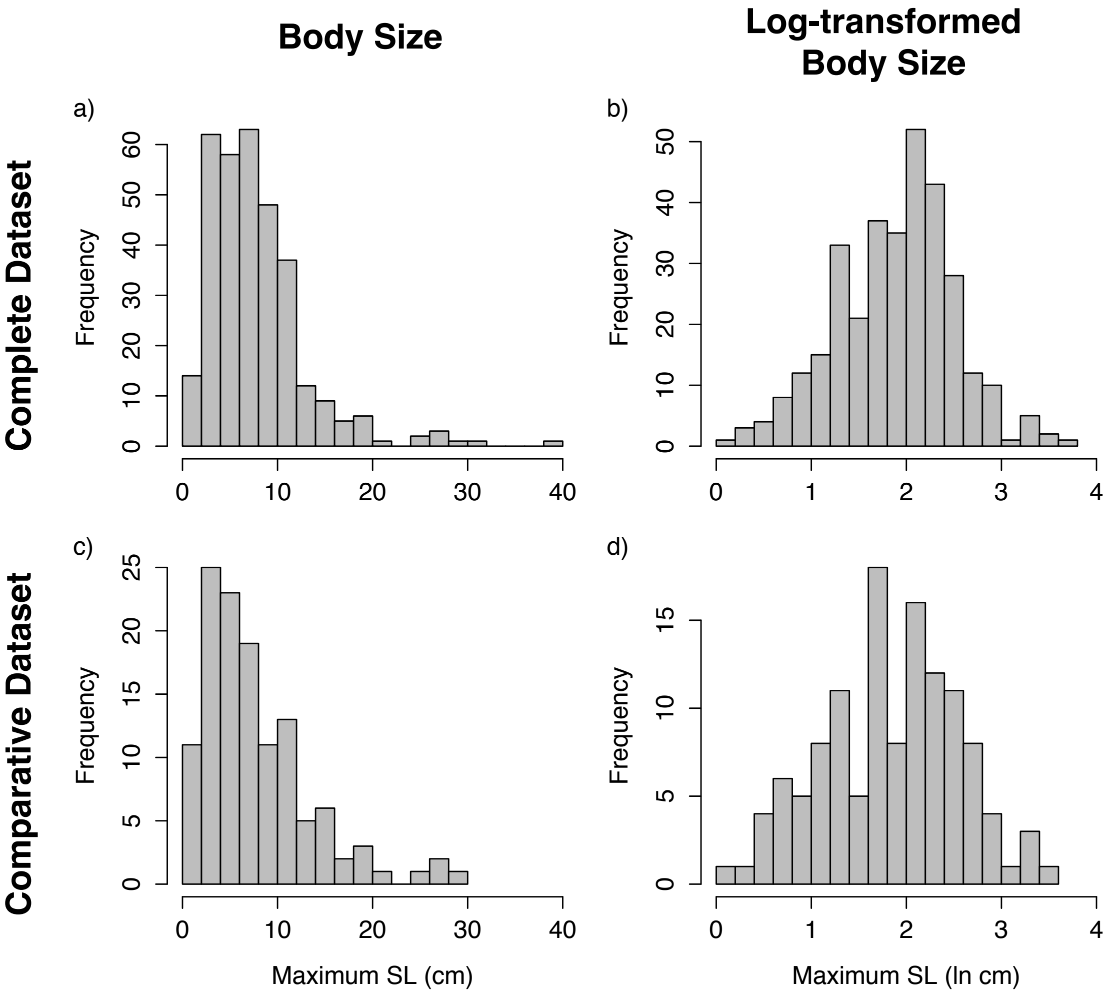


Figure 1. Distribution of maximum size of Danionidae species for the complete dataset (a-b) and the subset of the dataset used in comparative analyses (c-d), with SL in centimeters (a,c), and log-transformed SL (b,d).

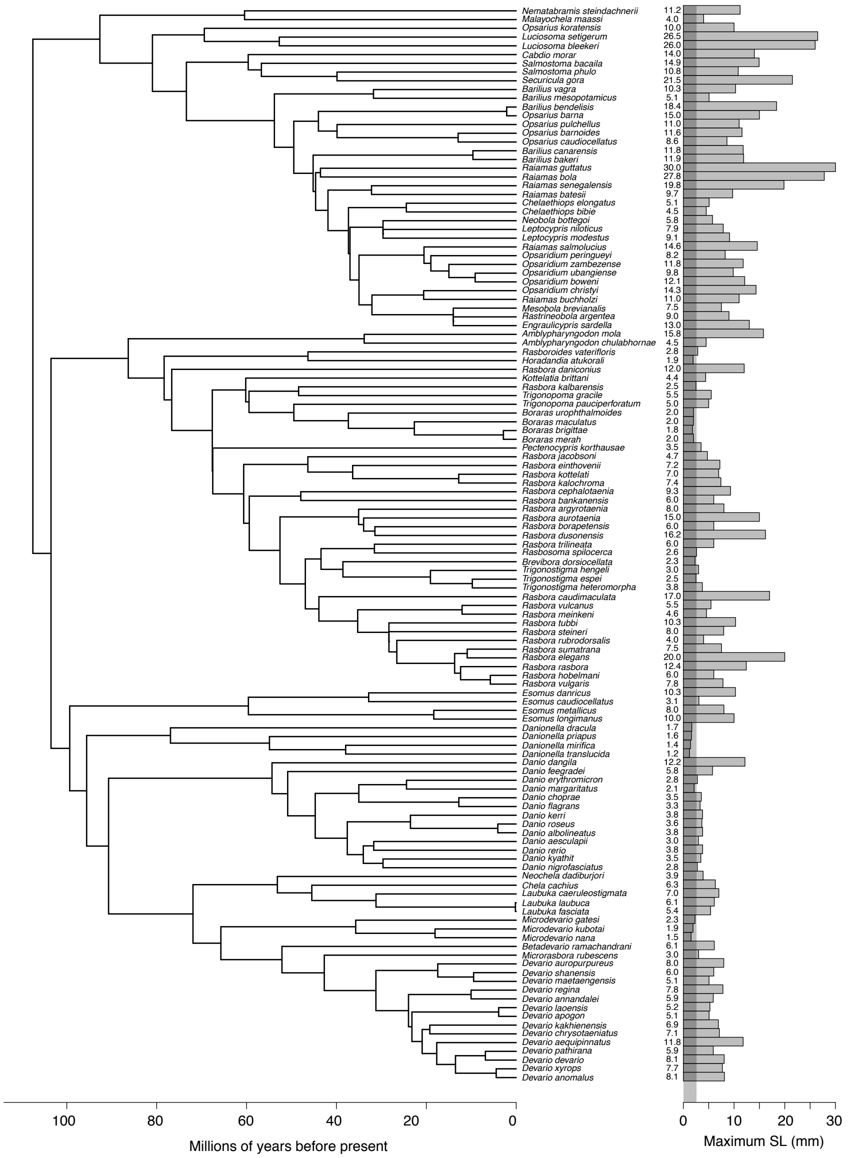


Figure 2. Time-calibrated phylogeny pruned to 123 species of Danionidae used in comparative phylogenetic analyses. Danionidae is divided into three tribes: Chedrini, Rasborini, and Danionini. Bar plots and adjacent numbers indicate maximum Standard Length (mm) for each species. Shaded area indicates the range of 0–26 mm SL used to define miniature species.

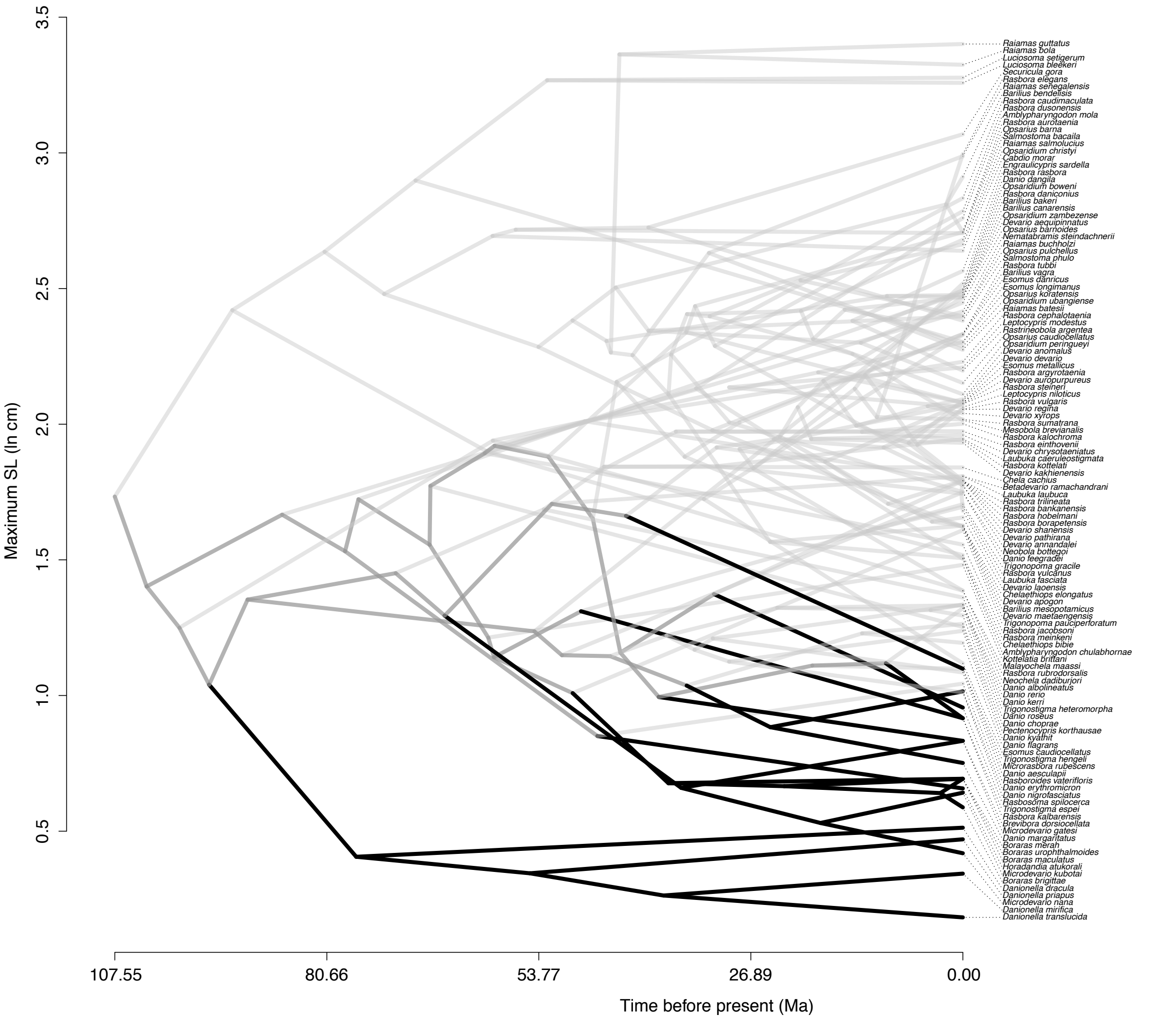


Figure 3. Traitgram showing ancestral body sizes as maximum body size (log cm SL) over time in the Danionidae. Black branches indicate branches leading from miniature taxa back to the root, blue branches indicate branches assigned to a miniature state (with the deepest node representing a non-miniature node that transitions to a miniature node or tip), grey branches indicate remaining taxa. Ancestral body sizes of miniature taxa are usually below the root state, showing a gradual decrease over time.

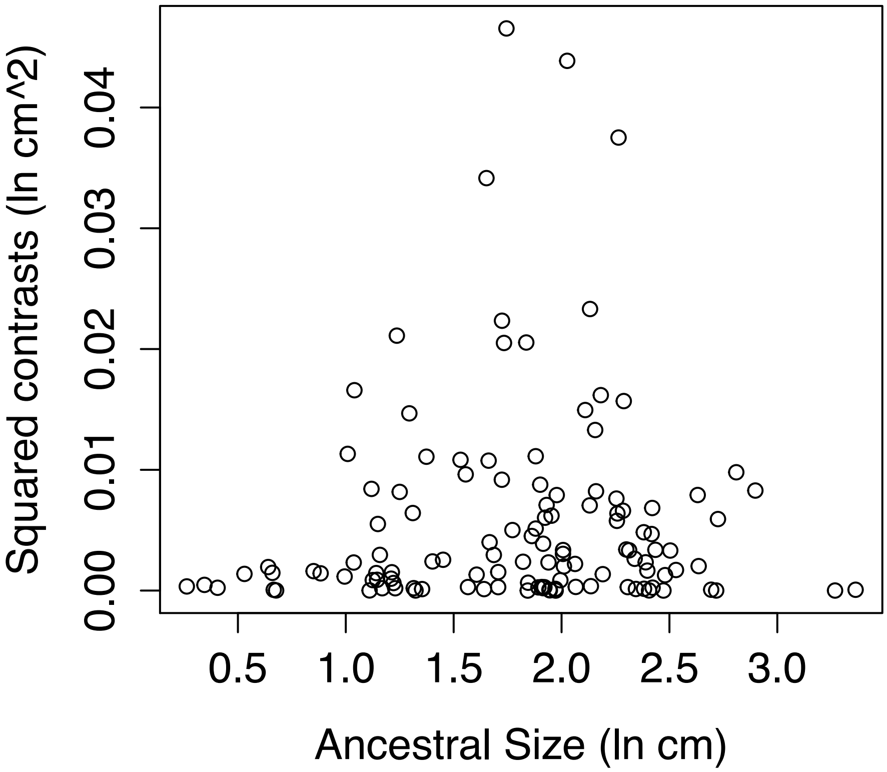


Figure 4. Rate-by-state plot displaying ancestral size (estimated in ln cm) versus rate (as estimated by squared independent contrasts). The correlation of rates of body size evolution to ancestral size are not significantly different from the null distribution (Spearman's rank correlation = .063, p = .485), however the rates of body size evolution appear to be reduced at the largest and smallest ancestral body sizes.

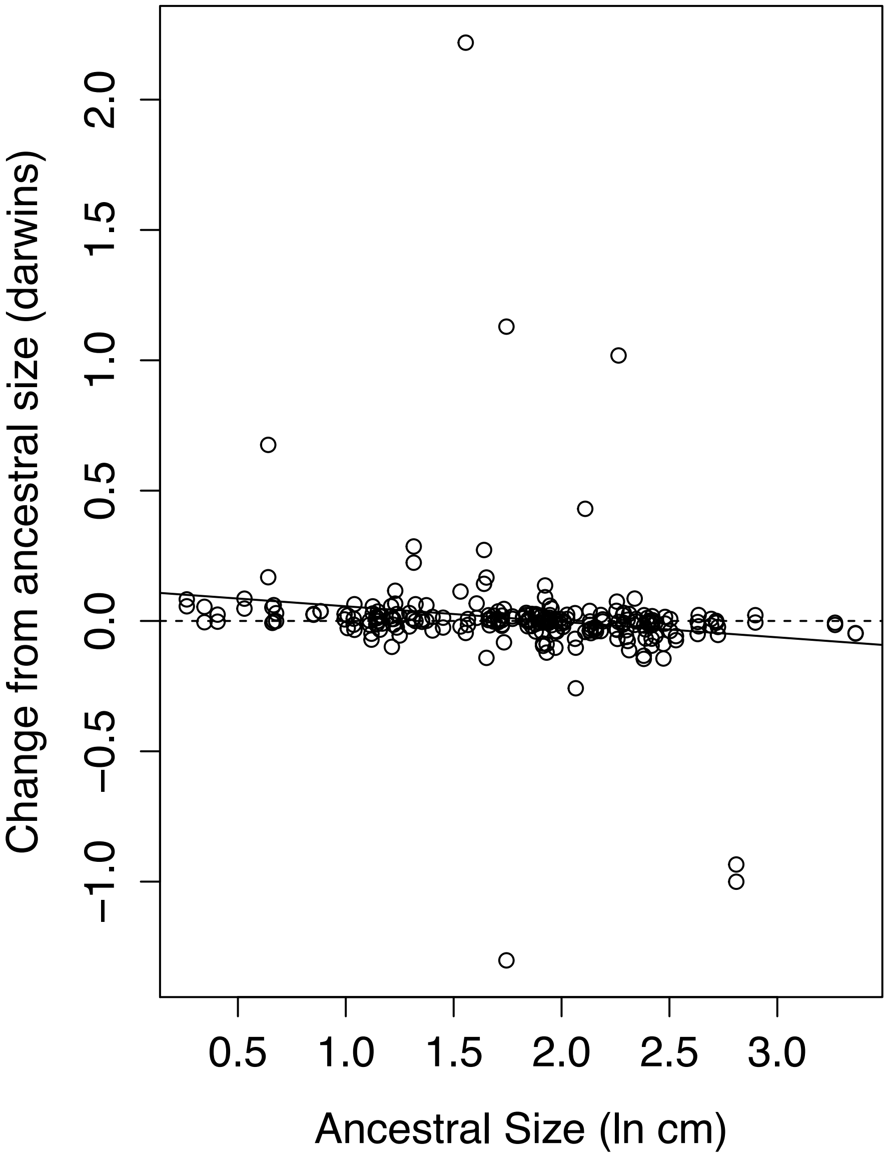


Figure 5. Ancestor-vs-change plot displaying ancestral size (estimated in ln cm) versus rate of change (as calculated from the difference between an ancestral size and descendent size over the branch length in time). A linear regression was used to fit a line with intercept = .119 ln cm/million years and slope = -0.061 ln cm/darwin. The observed slope was significantly different from the null distribution (p = .001).

Supplementary Material 1

GenBank accession numbers used for phylogenetic analysis.

		RAG1	Rh	Cyt b	COI
Danionidae	Amblypharyngodon chulabornae	HM224018	HM223898	HM224255	N/A
Danionidae	Amblypharyngodon mola	HM224019	HM223899	HM224256	HM224137
Danionidae	Aphyocypris chinensis	EU292692	FJ197066	AB218688	AB218688
Danionidae	Barilius cf. barila	HM224020	HM223900	HM224257	HM224138
Danionidae	Barilius mesopotamicus	N/A	N/A	HM48377	N/A
Danionidae	Barilius sp. "Bangladesh"	HM224021	N/A	HM224258	HM224139
Danionidae	Barilius vagra	HM224022	HM223901	HM224259	HM224140
Danionidae	Betadevario ramachandrani	N/A	GU327622	GU327623	N/A
Danionidae	Cabdio morar	EU711105	FJ531343	AP011335	AP011335
Danionidae	Chela cachius	EF452845	EF452914	EF452745	EF452891
Danionidae	Chela dadiburjori	EU292694	EF452915	EF452746	EF452892
Danionidae	Chelaethiops bibie	HM224023	HM223902	HM224260	HM224141
Danionidae	Chelaethiops bibie	HM224024	N/A	HM224261	HM224142
Danionidae	Chelaethiops elongatus	JX197014	JX197021	JX197006	JX196996
Danionidae	Danio aesculapii	N/A	EU241365	EU241430	N/A
Danionidae	Danio albolineatus	EU292696	EU409661	HM224262	HM224143
Danionidae	Danio cf. dangila	HM224025	HM223903	HM224263	HM224144
Danionidae	Danio cf. rerio "Assam"	N/A	EU241353	EU241420	N/A
Danionidae	Danio choprai	N/A	N/A	EF452740	EF452879
Danionidae	Danio choprai	HM224026	HM223904	HM224264	HM224145
Danionidae	Danio dangila	EU292697	EU409660	AP011235	AP011235
Danionidae	Danio erythromicron	HM224027	HM223905	EF452737	EF452867
Danionidae	Danio erythromicron	EU292698	GQ365222	AP011419	AP011419
Danionidae	Danio feegradei	N/A	N/A	EF452732	EF452861
Danionidae	Danio feegradei	HM224028	HM223906	HM224265	HM224146
Danionidae	Danio flagrans	N/A	EU241356	EU241421	N/A
Danionidae	Danio kerri	HM224029	HM223907	HM224266	HM224147
Danionidae	Danio kyathit	N/A	N/A	EF452733	EF452862
Danionidae	Danio kyathit "Spotted"	HM224030	HM223908	HM224267	HM224148
Danionidae	Danio margaritatus	EU292695	GQ365223	HM224268	HM224149
Danionidae	Danio nigrofasciatus	N/A	N/A	N/A	EF452863
Danionidae	Danio nigrofasciatus	EU292699	HM223909	HM224269	HM224150
Danionidae	Danio rerio	U71093	L11014	AC024175	AC024175
Danionidae	Danio roseus	N/A	N/A	EF452735	EF452865
Danionidae	Danio roseus	HM224031	HM223910	HM224270	HM224151
Danionidae	Danio sp. "Bangladesh"	HM224032	HM223911	HM224271	HM224152
Danionidae	Danio sp. "Hikari"	N/A	N/A	EF452731	EF452860
Danionidae	Danio sp. "Ozelot"	N/A	EU241364	EU241429	N/A
Danionidae	Danio sp. "Panther"	N/A	N/A	EF452734	EF452864
Danionidae	Danio sp. "Panther"	HM224033	HM223912	HM224272	HM224153
Danionidae	Danionella dracula	EF452841	HM223913	EF452741	EF452887
Danionidae	Danionella dracula	FJ753520	N/A	EF151099	FJ753484
Danionidae	Danionella mirifica	FJ753519	N/A	FJ753512	N/A
Danionidae	Danionella priapus	FJ753547	N/A	FJ753518	FJ753511

Danionidae	Danionella sp. "India"	EU292700	FJ531347	AP011424	AP011424
Danionidae	Danionella sp. "South Myanmar	FJ753543	N/A	FJ753514	FJ753507
Danionidae	Danionella translucida	FJ753544	N/A	FJ753515	FJ753508
Danionidae	Devario aequipinnatus	FJ410924	N/A	HM224273	HM224154
Danionidae	Devario annandalei	HM224034	HM223914	HM224274	HM224155
Danionidae	Devario anomalus	FJ410925	N/A	HM224282	HM224163
Danionidae	Devario apogon	N/A	EU241366	EU241431	N/A
Danionidae	Devario auropurpureus	EF452843	EF452912	EF452743	EF452889
Danionidae	Devario auropurpureus	EU292708	HM223915	HM224275	HM224156
Danionidae	Devario cf. malabaricus	HM224035	HM223916	HM224276	HM224157
Danionidae	Devario chrysotaeniatus	HM224036	HM223917	HM224277	HM224158
Danionidae	Devario devario	N/A	HM223918	EF452736	EF452866
Danionidae	Devario kakhienensis	N/A	EU241370	EU241435	N/A
Danionidae	Devario laoensis	HM224037	HM223919	HM224278	HM224159
Danionidae	Devario maetaengensis	N/A	EU241371	EU241436	N/A
Danionidae	Devario pathirana	N/A	EU241372	EU241437	N/A
Danionidae	Devario regina	EU292701	FJ531348	HM224279	HM224160
Danionidae	Devario regina	HM224038	HM223920	HM224280	HM224161
Danionidae	Devario shanensis	HM224039	HM223921	HM224281	HM224162
Danionidae	Devario sp.	HM224040	HM223922	HM224283	HM224164
Danionidae	Devario sp. "Laos"	HM224041	HM223923	HM224284	HM224165
Danionidae	Devario xyrops	N/A	EU241374	EU241439	N/A
Danionidae	Engraulicypris sardella	JX197015	JX197022	JX197007	JX196997a
Danionidae	Esomus caudiocellatus	N/A	EU241375	EU241440	N/A
Danionidae	Esomus cf. ahli	EF452842	HM223924	EF452742	EF452888
Danionidae	Esomus cf. danricus "DarkStripe"	HM224042	HM223925	HM224285	HM224166
Danionidae	Esomus cf. danricus "LightStripe"	HM224043	N/A	HM224286	HM224167
Danionidae	Esomus danricus	HM224044	HM223926	HM224287	HM224168
Danionidae	Esomus longimanus	FJ531248	FJ531349	HM224288	HM224169
Danionidae	Esomus metallicus	EU292702	FJ197067	AB239594	AB239594
Danionidae	Horadandia atukorali	EU292703	FJ531350	AP011400	AP011400
Danionidae	Laubuca caeruleostigmata	HM224046	HM223928	HM224289	HM224170
Danionidae	Laubuca fasciata	HM224047	HM223929	HM224290	HM224171
Danionidae	Laubuca laubuca	HM224048	HM223930	HM224291	HM224172
Danionidae	Laubuca sp.	HM224049	HM223931	HM224292	HM224173
Danionidae	Leptocypris niloticus "BurkinaFaso"	HM224051	N/A	HM224294	HM224175
Danionidae	Leptocypris niloticus "Ethiopia"	HM224050	HM223932	HM224293	HM224174
Danionidae	Leptocypris sp.	HM224052	HM223933	AP011428	AP011428
Danionidae	Luciosoma bleekeri	HM224053	HM223934	AP011399	AP011399
Danionidae	Luciosoma setigerum	EU292704	FJ531352	AP011423	AP011423
Danionidae	Luciosoma sp.	FJ753533	N/A	EF151104	FJ753497
Danionidae	Malayochela maassi	FJ753522	EU241379	EF151098	FJ753486
Danionidae	Mesobola brevianalis	HM224054	HM223935	HM224295	HM224176
Danionidae	Microdevario gatesi	N/A	EU241380	EU241445	N/A
Danionidae	Microdevario kubotai	EU292707	FJ531353	EF452738	EF452868
Danionidae	Microdevario kubotai	HM224055	N/A	N/A	HM224177
Danionidae	Microdevario nanus	EU292705	GQ365224	AP011402	AP011402
Danionidae	Microrasbora rubescens	EF452844	EF452913	EF452744	EF452890
Danionidae	Nematabramis steindachneri	FJ753532	N/A	EF151106	FJ753496

Danionidae	Neobola bottegoi	HM224056	HM223936	HM224296	HM224178
Danionidae	Opsaridium batesii	JX197017	JX197024	JX197010	JX197002a
Danionidae	Opsaridium boweni	JX197016	JX197023	JX197009	JX197000a
Danionidae	Opsaridium peringueyi	HM224072	HM223954	HM224311	HM224192
Danionidae	Opsaridium sp.	EF452846	HM223955	EF452747	EF452893
Danionidae	Opsaridium ubangiense	HM224073	HM223956	HM224312	HM224193
Danionidae	Opsaridium zambezense	N/A	N/A	N/A	HM224194
Danionidae	Opsarius bakeri	HM224076	HM223959	HM224315	HM224197
Danionidae	Opsarius barna	N/A	N/A	N/A	EU417797
Danionidae	Opsarius barnoides	HM224077	HM223960	HM224316	HM224198
Danionidae	Opsarius barnoides	HM224078	HM223961	HM224317	HM224199
Danionidae	Opsarius bendelisis	EU292693	FJ531346	AP011433	AP011433
Danionidae	Opsarius canarensis	HM224079	HM223962	HM224318	HM224200
Danionidae	Opsarius caudiocellatus	HM224080	HM223963	HM224319	HM224201
Danionidae	Opsarius cf. bakeri	HM224081	HM223964	HM224320	HM224202
Danionidae	Opsarius cf. shacra	HM224082	HM223965	HM224321	HM224203
Danionidae	Opsarius koratensis	HM224083	HM223966	N/A	HM224204
Danionidae	Opsarius koratensis	N/A	HM223967	HM224322	HM224205
Danionidae	Opsarius pulchellus	HM224084	HM223968	HM224323	HM224206
Danionidae	Opsarius pulchellus	HM224085	HM223969	HM224324	HM224207
Danionidae	Opsarius sp. "Myanmar"	HM224086	HM223970	HM224325	HM224208
Danionidae	Pectenocypris korthausae	HM224087	HM223972	HM224327	HM224210
Danionidae	Pectenocypris korthausae	HM224088	HM223973	HM224328	HM224211
Danionidae	Raiamas bola	HM224089	HM223974	HM224329	HM224212
Danionidae	Raiamas buchholzi	HM224090	HM223975	HM224330	HM224213
Danionidae	Raiamas christyi	JX197018	JX197025	JX197011	JX197003a
Danionidae	Raiamas guttatus	HM224091	HM223976	HM224331	HM224214
Danionidae	Raiamas guttatus	HM224092	HM223977	AP011222	AP011222
Danionidae	Raiamas salmolucius	JX197019	JX197026	JX197012	JX197004a
Danionidae	Raiamas senegalensis	HM224093	HM223978	HM224332	HM224215
Danionidae	Raiamas senegalensis	HM224094	HM223979	HM224333	HM224216
Danionidae	Raiamas sp. "Myanmar"	HM224095	HM223980	HM224334	HM224217
Danionidae	Rasbora argyrotaenia	EF452836	EF452907	HM224335	EF452880
Danionidae	Rasbora aurotaenia	HM224096	HM223981	HM224336	HM224219
Danionidae	Rasbora bankanensis	EU292709	FJ531357	HM224337	HM224220
Danionidae	Rasbora borapetensis	HM224097	HM223982	N/A	HM224221
Danionidae	Rasbora brigittae	N/A	EU241347	EU241414	N/A
Danionidae	Rasbora brittani	HM224098	HM223983	HM224338	EF452869
Danionidae	Rasbora caudimaculata	N/A	N/A	HM224339	EF452870
Danionidae	Rasbora cephalotaenia	N/A	N/A	HM224340	EF452881
Danionidae	Rasbora cephalotaenia	HM224099	HM223984	AP011430	AP011430
Danionidae	Rasbora cf. bankanensis	N/A	N/A	HM224341	EF452871
Danionidae	Rasbora cf. borapetensis	HM224100	HM223985	HM224342	HM224222
Danionidae	Rasbora cf. micros	EF452839	EF452910	HM224343	EF452885
Danionidae	Rasbora cf. micros	HM224118	HM224000	HM224361	HM224235
Danionidae	Rasbora cf. pauciperforata	N/A	EU241400	EU241465	N/A
Danionidae	Rasbora cf. paviana	HM224101	HM223986	HM224344	HM224223
Danionidae	Rasbora daniconius	HM224102	HM223987	N/A	AP011285
Danionidae	Rasbora daniconius	HM224103	HM223988	HM224345	EF452872

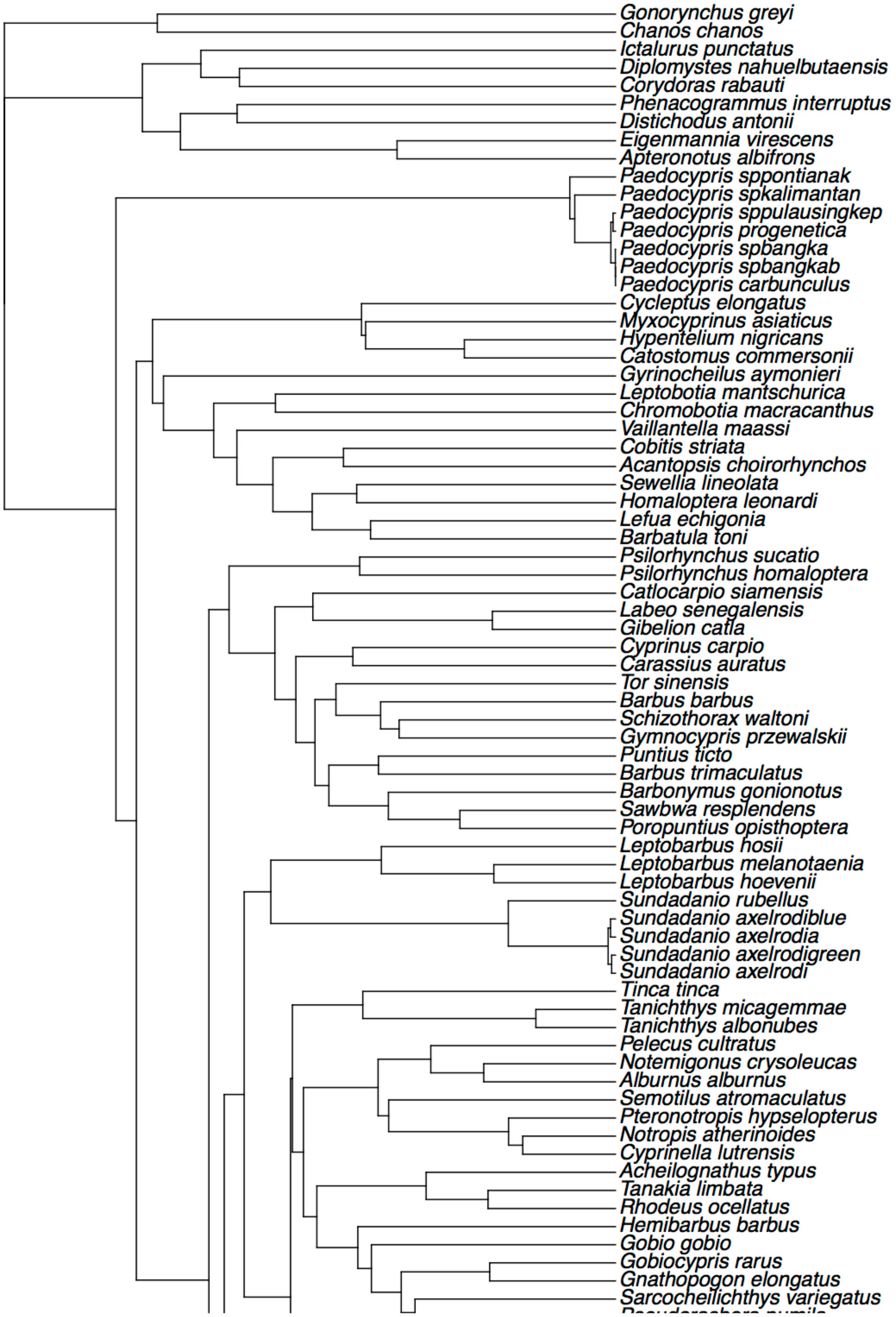
Danionidae	Rasbora dorsiocellata	N/A	N/A	HM224346	EF452873
Danionidae	Rasbora dorsiocellata	HM224104	HM223989	HM224347	HM224224
Danionidae	Rasbora dusonensis	HM224105	HM223990	HM224348	HM224225
Danionidae	Rasbora einthovenii	HM224106	HM223991	HM224349	HM224226
Danionidae	Rasbora elegans	HM224107	HM223992	HM224350	HM224227
Danionidae	Rasbora espei	N/A	N/A	HM224351	EF452877
Danionidae	Rasbora espei	HM224108	N/A	AP011449	AP011449
Danionidae	Rasbora gracilis	N/A	EU241402	EU241467	N/A
Danionidae	Rasbora hengeli	N/A	N/A	HM224352	EF452878
Danionidae	Rasbora hengeli	HM224109	HM223993	HM224353	HM224228
Danionidae	Rasbora heteromorpha	EU292712	FJ531360	AP011421	AP011421
Danionidae	Rasbora hobelmani	HM224110	HM223994	HM224354	HM224229
Danionidae	Rasbora jacobsoni	HM224111	HM223995	HM224355	HM224230
Danionidae	Rasbora kalbarensis	FJ753538	N/A	EF151116	FJ753502
Danionidae	Rasbora kalochroma	HM224112	HM223996	HM224356	HM224231
Danionidae	Rasbora kottelati	HM224113	HM223997	HM224357	HM224232
Danionidae	Rasbora maculata	EF452838	EF452909	HM224358	EF452884
Danionidae	Rasbora maculata	HM224114	N/A	AP011420	AP011420
Danionidae	Rasbora meinkeni	HM224115	HM223998	N/A	HM224233
Danionidae	Rasbora merah	HM224116	HM223999	HM224359	EF452859
Danionidae	Rasbora merah	HM224117	N/A	HM224360	HM224234
Danionidae	Rasbora pauciperforata	HM224119	HM224001	HM224362	HM224236
Danionidae	Rasbora rasbora	HM224120	HM224002	HM224363	HM224237
Danionidae	Rasbora rasbora	HM224121	HM224003	HM224364	HM224238
Danionidae	Rasbora rubrodorsalis	N/A	N/A	HM224365	EF452874
Danionidae	Rasbora rubrodorsalis	HM224122	HM224004	HM224366	HM224239
Danionidae	Rasbora sp. "Thailand"	N/A	N/A	N/A	HM224218
Danionidae	Rasbora spilocerca	HM224123	HM224005	HM224367	HM224240
Danionidae	Rasbora steineri	EU409631	EU409662	HM224368	HM224241
Danionidae	Rasbora sumatrana	EF452837	EF452908	HM224369	EF452882
Danionidae	Rasbora trilineata	HM224124	HM224006	HM224370	EF452883
Danionidae	Rasbora tubbi	HM224125	HM224007	HM224371	HM224242
Danionidae	Rasbora urophthalmoides	EF452840	EF452911	HM224372	EF452886
Danionidae	Rasbora vulcanus	FJ753539	N/A	EF151118	FJ753503
Danionidae	Rasbora vulcanus	N/A	N/A	N/A	EF452875
Danionidae	Rasbora vulgaris	HM224126	HM224008	HM224373	HM224243
Danionidae	Rasboroides vaterifloris	HM224127	HM224009	AP011432	AP011432
Danionidae	Rasboroides vaterifloris	N/A	N/A	HM224374	EF452876
Danionidae	Rastrineobola argentea	JX197020	JX197027	JX197013	JX197005a
Danionidae	Salmostoma bacaila "India"	HM224128	HM224010	AP011223	AP011223
Danionidae	Salmostoma bacaila "Nepal"	HM224130	HM224012	HM224376	HM224245
Danionidae	Salmostoma bacaila"Bangladesh"	HM224129	HM224011	HM224375	HM224244
Danionidae	Salmostoma cf. phulo"Longnose"	HM224131	N/A	HM224377	HM224246
Danionidae	Salmostoma cf. phulo"Shortnose"	HM224132	N/A	HM224378	HM224247
Danionidae	Salmostoma phulo	HM224133	HM224013	HM224379	HM224248
Danionidae	Salmostoma sp. "Myanmar"	HM224134	HM224014	HM224380	HM224249
Danionidae	Securicula gora	HM224135	HM224015	HM224381	HM224250
Danionidae	Securicula gora	N/A	HM224016	HM224382	HM224251
Outgroup	Acantopsis choirorhynchos	EU711139	FJ197039	AB242161	AB242161

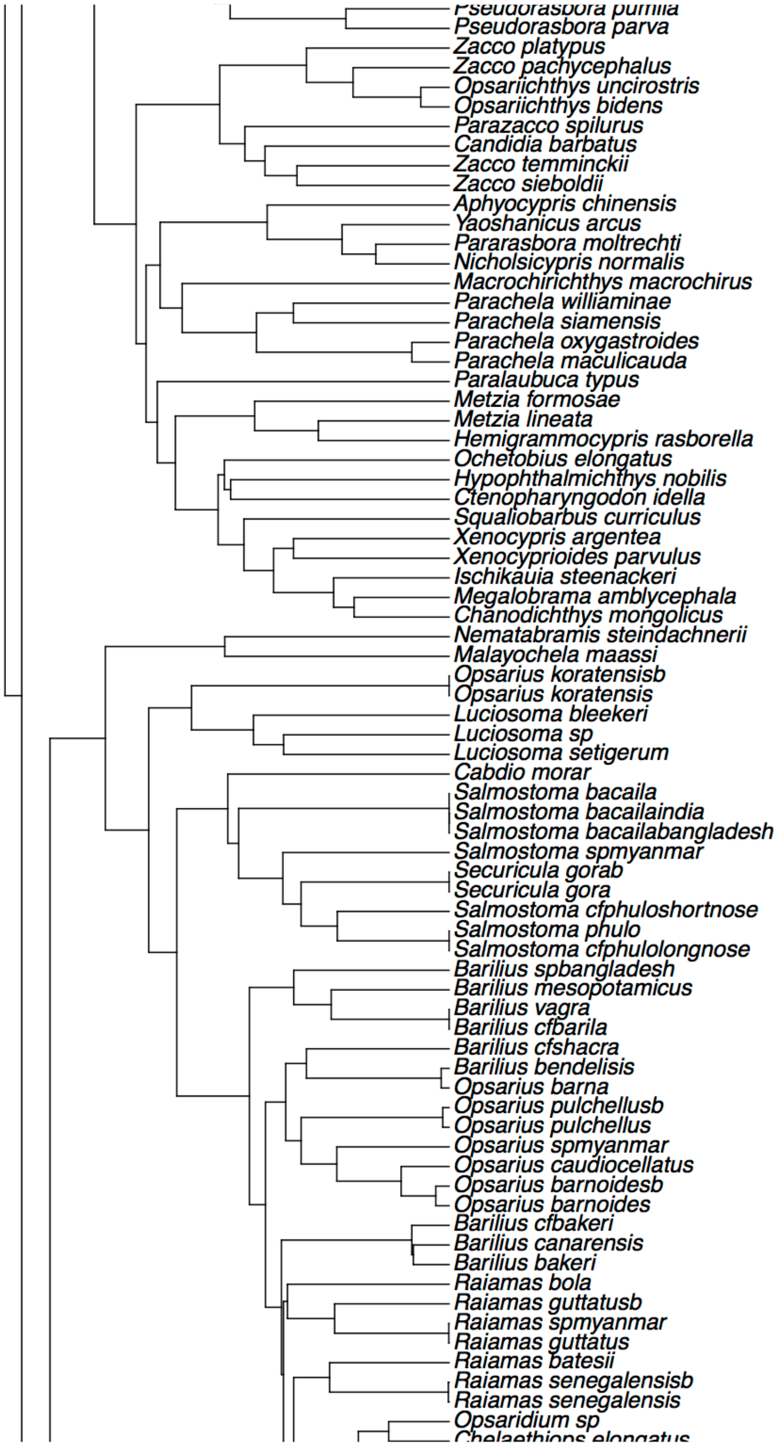
Outgroup	Acheilognathus typus	EU292688	FJ197042	AB239602	AB239602
Outgroup	Alburnus alburnus	EU711143	FJ197044	AB239593	AB239593
Outgroup	Barbatula toni	EU711133	FJ197030	AB242162	AB242162
Outgroup	Barbonymus gonionotus	EU711146	FJ531344	AB238966	AB238966
Outgroup	Barbus barbus	EU711147	FJ197049	AB238965	AB238965
Outgroup	Barbus trimaculatus	EU711148	FJ197050	AB239600	AB239600
Outgroup	Candidia barbatus	N/A	N/A	AY958200	N/A
Outgroup	Carassius auratus	DQ196520	L11863	AB006953	AB111951
Outgroup	Catostomus commersonii	EU409612	FJ197032	AB127394	AB127394
Outgroup	Chanodichthys mongolicus	EU711145	FJ197047	AP009060	AP009060
Outgroup	Chanos chanos	AY430207	FJ197072	AB054133	AB054133
Outgroup	Chromobotia macracantha	EU711137	FJ197037	AB242163	AB242163
Outgroup	Cobitis striata	EF458303	HM223938	AB054125	AB054125
Outgroup	Ctenopharyngodon idella	EF178284	HM223939	EU391390	EU391390
Outgroup	Cycleptus elongatus	EU409613	FJ197035	AB126082	AB126082
Outgroup	Cyprinella lutrensis	EU711158	FJ197061	AB070206	AB070206
Outgroup	Cyprinus carpio	AY787040	U02475	X61010	X61010
Outgroup	Gibelion catla	HM224057	HM223940	N/A	N/A
Outgroup	Gnathopogon elongatus	EU711153	FJ197055	AB218687	AB218687
Outgroup	Gobio gobio	EU292689	FJ197056	AB239596	AB239596
Outgroup	Gobiocypris rarus	N/A	N/A	AF309083	AY879113
Outgroup	Gonorynchus greyi	EU409606	EU409632	AB054134	AB054134
Outgroup	Gymnocypris przewalskii	EU711149	FJ197051	AB239595	AB239595
Outgroup	Gyrinocheilus aymonieri	EU292682	FJ197071	AB242164	AB242164
Outgroup	Hemibarbus barbus	EU711154	FJ197057	AB070241	AB070241
Outgroup	Hemigrammocypris rasborella	HM224045	HM223927	AP011422	AP011422
Outgroup	Homaloptera leonardi	EU711130	FJ197027	AB242165	AB242165
Outgroup	Hypentelium nigricans	EU711134	FJ197033	AB242169	AB242169
Outgroup	Hypophthalmichthys nobilis	HM224058	HM223941	EU343733	EU343733
Outgroup	Ictalurus punctatus	DQ492511	AF028016	AF482987	AF482987
Outgroup	Ischikauia steenackeri	EU292687	EU409648	AB239601	AB239601
Outgroup	Labeo senegalensis	EU711151	FJ197053	AB238968	AB238968
Outgroup	Lefua echigonia	EF458305	FJ197028	AB054126	AB054126
Outgroup	Leptobarbus hoevenii	FJ531249	FJ531351	AP011286	AP011286
Outgroup	Leptobarbus hosii	N/A	N/A	AY243350	N/A
Outgroup	Leptobarbus melanotaenia	N/A	N/A	N/A	JN646097
Outgroup	Leptobotia mantschurica	EU711138	FJ197038	AB242170	AB242170
Outgroup	Macrochirichthys macrochirus	EU409630	EU409659	AP011234	AP011234
Outgroup	Megalobrama amblycephala	EU409620	EU409647	EU434747	EU434747
Outgroup	Metzia formosae	HM224066	HM223949	HM224304	HM224186
Outgroup	Metzia lineata	HM224067	HM223950	HM224305	HM224187
Outgroup	Myxocyprinus asiaticus	EU711136	FJ197036	AB223007	AB223007
Outgroup	Nicholsicypris normalis	EU711123	HM223937	AP011396	AP011396
Outgroup	Nipponocypris sieboldii	EU292713	FJ197069	AB218898	AB218898
Outgroup	Nipponocypris temminckii	EF452849	EF452918	EF452750	EF452897
Outgroup	Notemigonus crysoleucas	EF452831	FJ197062	U01318	EF452854
Outgroup	Notropis atherinoides	HM224059	HM223942	HM224297	HM224179
Outgroup	Ochetobius elongatus	N/A	N/A	AF309506	N/A
Outgroup	Opsariichthys bidens	HM224074	HM223957	HM224313	HM224195

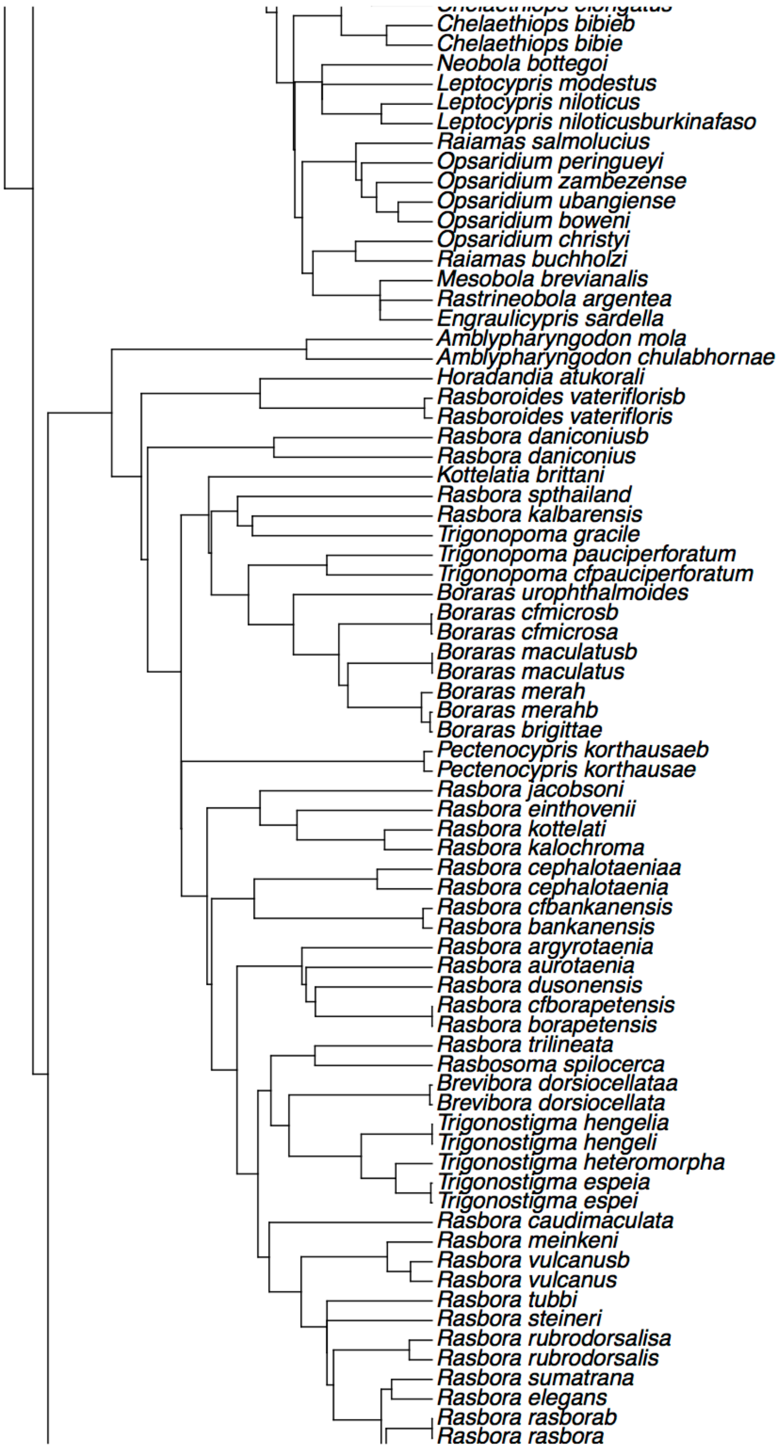
Outgroup	Opsariichthys pachycephalus	HM224075	HM223958	HM224314	HM224196
Outgroup	Opsariichthys uncirostris	EF452847	EF452916	EF452748	EF452894
Outgroup	Paedocypris carbunculus	GQ365218	GQ365226	HM224326	HM224209
Outgroup	Paedocypris sp. "Bangka"	N/A	N/A	EF151108	N/A
Outgroup	Paedocypris sp. "Bangka"	N/A	HM223971	AP011429	AP011429
Outgroup	Paedocypris sp. "Kalimantan"	N/A	N/A	EF151109	N/A
Outgroup	Paedocypris sp. "Pontianak"	N/A	N/A	EF151110	N/A
Outgroup	Paedocypris sp. "Pulau Singkep"	N/A	N/A	EF151111	N/A
Outgroup	Paedocypris sp. "Sumatra"	N/A	N/A	AP011287	AP011287
Outgroup	Parachela maculicauda	HM224060	HM223943	HM224298	HM224180
Outgroup	Parachela oxygastroides	HM224061	HM223944	HM224299	HM224181
Outgroup	Parachela siamensis	HM224062	HM223945	HM224300	HM224182
Outgroup	Parachela williaminaeP	HM224063	HM223946	HM224301	HM224183
Outgroup	Paralaubuca typus	EU409619	EU409646	AP011211	AP011211
Outgroup	Pararasbora moltrechti	N/A	N/A	FJ577897	N/A
Outgroup	Parazacco spilurus	N/A	N/A	AY958195	N/A
Outgroup	Pelecus cultratus	EU711144	FJ197045	AB239597	AB239597
Outgroup	Phenacogrammus interruptus	FJ197124	FJ197073	AB054129	AB054129
Outgroup	Pseudorasbora parva	HM224064	HM223947	HM224302	HM224184
Outgroup	Pseudorasbora pumila	EU711155	FJ197058	AB239599	AB239599
Outgroup	Psilorhynchus homaloptera	FJ531250	FJ531354	DQ026436	DQ026436
Outgroup	Psilorhynchus sucatio	FJ531251	FJ531355	AP011288	AP011288
Outgroup	Pteronotropis hypselopterus	HM224065	HM223948	HM224303	HM224185
Outgroup	Puntius ticto	EU711152	FJ197054	AB238969	AB238969
Outgroup	Rhodeus ocellatus	EU711142	FJ197043	AB070205	AB070205
Outgroup	Sarcocheilichthys variegatus	EU711157	FJ197060	AB054124	AB054124
Outgroup	Sawbwa resplendens	EU292686	N/A	HM224306	EF452895
Outgroup	Semotilus atromaculatus	EU409629	EU409658	HM224307	HM224188
Outgroup	Sewellia lineolata	HM224068	EU409635	AP011292	AP011292
Outgroup	Squaliobarbus curriculus	HM224069	HM223951	HM224308	HM224189
Outgroup	Sundadanio axelrodi	N/A	N/A	EF452739	HM224252
Outgroup	Sundadanio axelrodi	EU292711	GQ365228	HM224383	HM224253
Outgroup	Sundadanio axelrodi "blue"	N/A	JF915678	N/A	JF915678
Outgroup	Sundadanio axelrodi "green"	N/A	JF966222	N/A	JF915682
Outgroup	Sundadanio axelrodi "red"	N/A	JF966226	N/A	JF915683
Outgroup	Tanakia limbata	HM224070	HM223952	HM224309	HM224190
Outgroup	Tanichthys albonubes	FJ531253	FJ531359	AP011397	AP011397
Outgroup	Tanichthys micagemmae	HM224136	HM224017	HM224384	HM224254
Outgroup	Tinca tinca	EU711162	FJ197070	AB218686	AB218686
Outgroup	Vaillantella maassi	EU711132	FJ197031	AB242173	AB242173
Outgroup	Xenocyprioides parvulus	N/A	N/A	AF036207	N/A
Outgroup	Xenocypris argentea	HM224071	HM223953	HM224310	HM224191
Outgroup	Yaoshanicus arcus	FJ531254	FJ531361	AP011398	AP011398
Outgroup	Zacco platypus	EF452848	EF452917	EF452749	EF452896

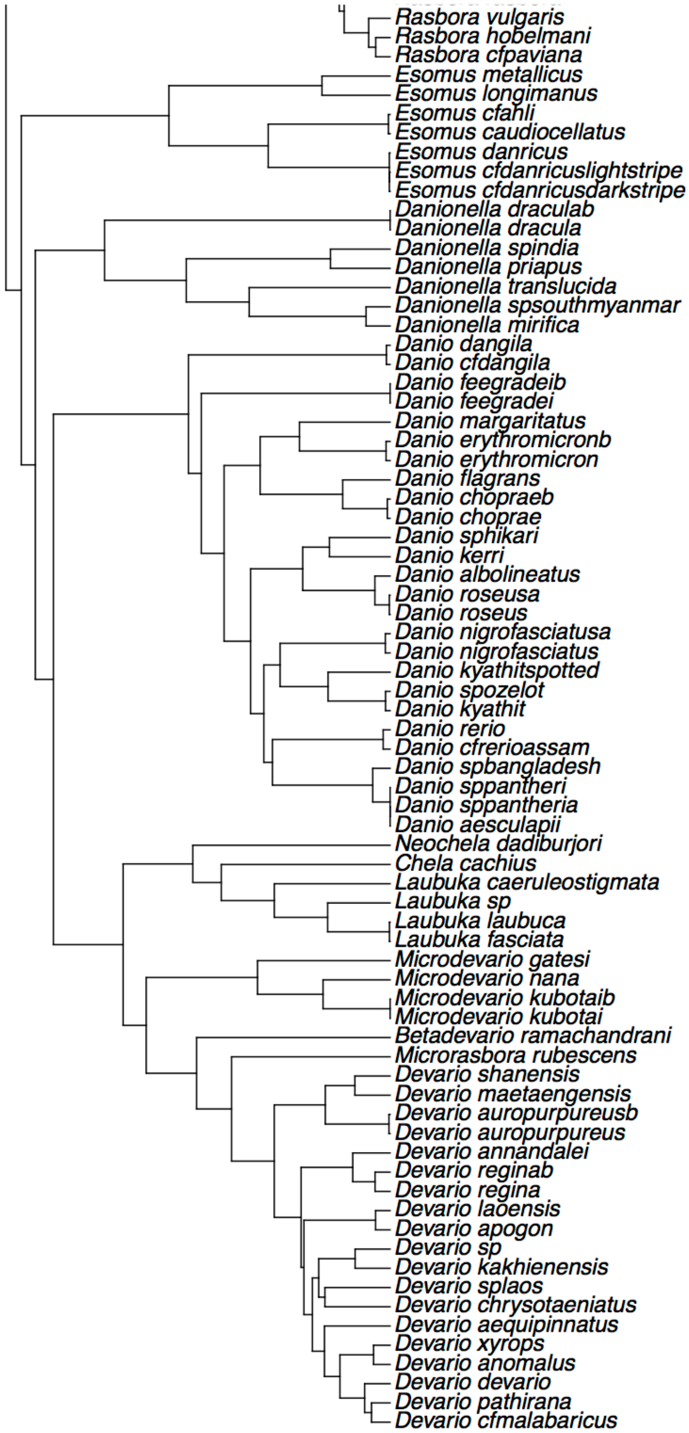
Supplementary Material 2

Time-calibrated phylogeny prior to pruning taxa for comparative phylogenetic analysis.




Pseudorasbora pumila
Pseudorasbora parva
Zacco platypus
Zacco pachycephalus
Opsariichthys uncirostris
Opsariichthys bidens
Parazacco spilurus
Candidia barbatus
Zacco temminckii
Zacco sieboldii
Aphyocypris chinensis
Yaoshanicus arcus
Pararasbora moltrechti
Nicholsicypris normalis
Macrochirichthys macrochirus
Parachela williaminae
Parachela siamensis
Parachela oxygastroides
Parachela maculicauda
Paralaubuca typus
Metzia formosae
Metzia lineata
Hemigrammocypris rasborella
Ochetobius elongatus
Hypophthalmichthys nobilis
Ctenopharyngodon idella
Squaliobarbus curriculus
Xenocypris argentea
Xenocyprioides parvulus
Ischikauia steenackeri
Megalobrama amblycephala
Chanodichthys mongolicus
Nematabramis steindachnerii
Malayochela maassi
Opsarius koratensisb
Opsarius koratensis
Luciosoma bleekeri
Luciosoma sp
Luciosoma setigerum
Cabdio morar
Salmostoma bacaila
Salmostoma bacailaindia
Salmostoma bacailabangladesh
Salmostoma spmyanmar
Securicula gorab
Securicula gora
Salmostoma cfphuloshortnose
Salmostoma phulo
Salmostoma cfphulolongnose
Barilius spbangladesh
Barilius mesopotamicus
Barilius vagra
Barilius cfbarila
Barilius cfshacra
Barilius bendelisis
Opsarius barna
Opsarius pulchellusb
Opsarius pulchellus
Opsarius spmyanmar
Opsarius caudiocellatus
Opsarius barnoidesb
Opsarius barnoides
Barilius cfbakeri
Barilius canarensis
Barilius bakeri
Raiamas bola
Raiamas guttatusb
Raiamas spmyanmar
Raiamas guttatus
Raiamas batesii
Raiamas senegalensisb
Raiamas senegalensis
Opsaridium sp
Chelaethiops elongatus


Chelaethiops elongatus
Chelaethiops bibieb
Chelaethiops bibie
Neobola bottegoi
Leptocypris modestus
Leptocypris niloticus
Leptocypris niloticusburkinafaso
Raiamas salmolucius
Opsaridium peringueyi
Opsaridium zambezense
Opsaridium ubangiense
Opsaridium boweni
Opsaridium christyi
Raiamas buchholzi
Mesobola brevianalis
Rastrineobola argentea
Engraulicypris sardella
Amblypharyngodon mola
Amblypharyngodon chulabhornae
Horadandia atukorali
Rasboroides vateriflorisb
Rasboroides vateriflorис
Rasbora daniconiusb
Rasbora daniconius
Kottelatia brittani
Rasbora spthailand
Rasbora kalbarensis
Trigonopoma gracile
Trigonopoma pauciperforatum
Trigonopoma cfpauciperforatum
Boraras urophthalmoides
Boraras cfmicrosb
Boraras cfmicrosa
Boraras maculatusb
Boraras maculatus
Boraras merah
Boraras merahb
Boraras brigittae
Pectenocypris korthausaeb
Pectenocypris korthausae
Rasbora jacobsoni
Rasbora einthovenii
Rasbora kottelati
Rasbora kalochroma
Rasbora cephalotaeniaa
Rasbora cephalotaenia
Rasbora cfbankanensis
Rasbora bankanensis
Rasbora argyrotaenia
Rasbora aurotaenia
Rasbora dusonensis
Rasbora cfborapetensis
Rasbora borapetensis
Rasbora trilineata
Rasbosoma spilocerca
Brevibora dorsiocellataa
Brevibora dorsiocellata
Trigonostigma hengelia
Trigonostigma hengeli
Trigonostigma heteromorpha
Trigonostigma espeia
Trigonostigma espei
Rasbora caudimaculata
Rasbora meinkeni
Rasbora vulcanusb
Rasbora vulcanus
Rasbora tubbi
Rasbora steineri
Rasbora rubrodorsalisa
Rasbora rubrodorsalis
Rasbora sumatrana
Rasbora elegans
Rasbora rasborab
Rasbora rasbora


Rasbora vulgaris
Rasbora hobelmani
Rasbora cfpaviana
Esomus metallicus
Esomus longimanus
Esomus cfahli
Esomus caudiocellatus
Esomus danricus
Esomus cfdanricuslightstripe
Esomus cfdanricusdarkstripe
Danionella draculab
Danionella dracula
Danionella spindia
Danionella priapus
Danionella translucida
Danionella spsouthmyanmar
Danionella mirifica
Danio dangila
Danio cfdangila
Danio feegradeib
Danio feegradei
Danio margaritatus
Danio erythromicronb
Danio erythromicron
Danio flagrans
Danio chopraeb
Danio choprae
Danio sphikari
Danio kerri
Danio albolineatus
Danio roseusa
Danio roseus
Danio nigrofasciatusa
Danio nigrofasciatus
Danio kyathitspotted
Danio spozelot
Danio kyathit
Danio rerio
Danio cfrerioassam
Danio spbangladesh
Danio sppantheri
Danio sppantheria
Danio aesculapii
Neochela dadiburjori
Chela cachius
Laubuka caeruleostigmata
Laubuka sp
Laubuka laubuca
Laubuka fasciata
Microdevario gatesi
Microdevario nana
Microdevario kubotaib
Microdevario kubotai
Betadevario ramachandrani
Microrasbora rubescens
Devario shanensis
Devario maetaengensis
Devario auropurpureusb
Devario auropurpureus
Devario annandalei
Devario reginab
Devario regina
Devario laoensis
Devario apogon
Devario sp
Devario kakhienensis
Devario splaos
Devario chrysotaeniatus
Devario aequipinnatus
Devario xyrops
Devario anomalus
Devario devario
Devario pathirana
Devario cfmalabaricus

CHAPTER 2

PHYLOGENOMICS OF PAEDOMORPHIC CYPRINIFORMES AND DANIONIDAE AS A PART OF RESOLVING CYPRINIFORMES RELATIONSHIPS USING AN ANCHORED ENRICHMENT APPROACH

BACKGROUND

My co-first author Carla C. Stout and I collaborated with Alan R. Lemmon and Emily M. Lemmon in reconstructing the evolutionary relationships of Cypriniformes using anchored phylogenomics. This section of my dissertation includes excerpted text from a co-first-authored manuscript we have submitted that represents dissertation chapters of both myself and Carla C. Stout, and additional text to highlight my research focus. My research questions were related to the phylogenetic relationships of the paedomorphic genera *Paedocypris*, *Sundadanio*, and *Danionella* relative to the remaining members of Danionidae among the Cypriniformes. The text includes an excerpted introduction with the addition of a review of the recent history of phylogenetic studies on paedomorphic Cypriniformes that are putatively danionids, excerpted methods from the paper, results containing additional text to focus on the Danionidae, and Discussion edited from the submitted manuscript. The remaining introduction, results, and discussion focused on other clades of Cypriniformes (including Cobitoidei, heavily addressed in Chapter 3) will be presented as part of Carla C. Stout's dissertation and in publication.

Abstract

As part of a larger study to infer relationships among the Cypriniformes, we infer the phylogenetic relationships among putative members of Danionidae. The relationships of the paedomorphic taxa *Paedocypris* and *Sundadanio* have been particularly problematic to infer. In this study we present the first phylogenomic analysis of Danionidae, using anchored hybrid enrichment for 172 taxa to represent the order Cypriniformes, including 34 tips representing putative members of the family Danionidae, and three outgroup taxa. This is the largest locus sampling for the order to date (219 loci, 315,288 bp, average locus length of 1011 bp). *Paedocypris* and *Sundadanio* are not members of Danionidae, and we recover *Paedocypris* as sister to Cyprinoidei and *Sundadanio* as an independent branch within Cyprinoidei sister to a clade formed by multiple cyprinoid families including Xenocyprididae, Leuciscidae, Acheilognathidae, Gobionidae, and Tanichthyidae. Danionidae is otherwise monophyletic with the exclusion of these two paedomorphic genera. *Esomus* is recovered as a separate branch of Danionidae sister to ((Danioninae + Chedrinae) + Rasborinae). The traditionally recognized subclades Danioninae, Chedrinae, and Rasborinae are each recovered as monophyletic with the exception of *Esomus* forming a distinct branch in the Danionidae. We reanalyze previously-collected morphological phylogenetic data and interpret these results relative to our phylogenomics results.

Introduction

Cypriniformes (minnows, carps, loaches, and suckers) is the largest group of freshwater fishes in the world. Diversity ranges from some of the smallest vertebrates in the world (*Paedocypris*, 7.9 mm in standard length) to members of *Tor* and *Catlocarpio* (almost 3 m SL) *(Mayden and Chen 2010)*. The number of valid species is currently estimated at around 4300 *(Eschmeyer et al. 2016)* with as many as 2500 still awaiting description *(Mayden et al. 2009)*. Species of Cypriniformes are distributed in freshwater habitats across Asia, Europe, Africa, and North America *(Nelson 2006)*. Example representatives include the zebrafish (*Danio rerio*), a model organism used in genomic and developmental biology, important aquaculture species like the common carp (*Cyprinus carpio*), major invasive species to North America such as *Hypophthalmichthys* (silver carp), and many popular aquarium species (rasboras and barbs).

For taxonomic clarity, this study follows the proposition by Mayden and Chen (2010) that elevates subfamilies within Cyprinidae to the family level based on repeated recovery of major clades. Superfamilies proposed by Mayden & Chen (2010) are elevated to the suborder level to be consistent with the recognition of suborders as the taxonomic level above family and below order in the most recent classification of bony fishes (Betancur-R et al. 2013). Because of the great diversity within Cypriniformes, most phylogenetic studies have focused on smaller groups within the order (Bufalino and Mayden 2010; e.g. Tang et al. 2010; 2011; 2013; Chang et al. 2014; Yang et al. 2015a). Approaches used to resolve relationships at these levels have typically included standard methods using PCR to amplify targeted mitochondrial and/or nuclear genes. These approaches have had varied success at elucidating relationships at these taxonomic levels, but deeper, all-inclusive studies have resulted in conflicting phylogenies, including some that are in direct conflict with each other and with morphological analyses (Mayden et al. 2007; 2009; Mayden and Chen 2010). The only nuclear genomic scale study to date consisted of 100 genes and was limited to only thirteen individuals, most of which belong to Xenocyprididae within Cyprinoidei (Tao et al. 2010). The large number of taxa in Cypriniformes has forced researches to either focus on a small subset of representatives with an increasing number of molecular loci, or focus on large taxonomic representation with relatively few loci.

Despite the importance of the zebrafish *Danio rerio* as a model organism in vertebrate biology (Howe et al. 2013), only recently have its closest relatives been resolved (Mayden et al. 2007). Danionidae is a diverse family of freshwater fishes, numbering approximately 300 species, distributed in Africa and Asia (Tang et al. 2010). Among the Cypriniformes, the relationships of Danionidae have been problematic to reconstruct until recently. Recent morphological and molecular studies generally are consistent on the recovery of three major clades of Danionidae: Danioninae, Rasborinae, and Chedrinae, and the exclusion of a clade of Far East ex-danionids now known to be more closely related to Xenocyprididae (Tang et al. 2010; Liao et al. 2011; Tang et al. 2013). On the other hand, phylogenetic resolution of some major relationships are poor. The relative relationships of Danioninae, Rasborinae, and Chedrinae are inconsistently recovered, and even the monophyly of these three groups forming a clade is repeatedly unsupported (Rüber et al. 2007; Fang et al. 2009; Mayden and Chen 2010; Tang et al. 2010; Liao et al. 2011). Also, the placement of the danionid genus *Esomus* is contradictory between multiple morphological and molecular analyses, including as sister to a clade formed by *Danio* and *Raiamas* (with only three members of Danionidae; Saitoh et al. 2011), as part of Danioninae (Mayden and Chen 2010; Tang et al. 2010), or as part of Chedrinae (Liao et al. 2011), but relationships are often weakly supported. Finally, the most challenging and controversial relationships of putative danionids has to do with the relationships of paedomorphic genera including *Paedocypris*, *Danionella*, and *Sundadanio* (as reviewed in Britz and Conway 2011).

The relationships of paedomorphic cypriniform fishes, in particular *Paedocypris*, have been extremely controversial and highly incongruent between morphological and multiple molecular studies. Soon after *Paedocypris* was described (Kottelat et al. 2006), the first phylogenetic hypothesis based on cytochrome b was presented. *Paedocypris* and *Sundadanio* were recovered as sister taxa with 100% posterior probability, and these two taxa were sister to the remaining Danionidae with 86% posterior probability (Rüber et al. 2007). Mayden et al. (2007) followed with a molecular phylogeny of Cypriniformes including *Sundadanio* and *Danionella*, with data comprising four mitochondrial genes and two nuclear genes. Both paedomorphic taxa were recovered with poorly-supported relationships in a polytomy amongst other members of Danioninae with parsimony, or *Danionella* with Danioninae and *Sundadanio* as an independent branch among Cyprinoidei under likelihood.

Britz & Conway (2009) then presented a detailed description of the osteology of *Paedocypris*, including inferences on the evolutionary relationships of *Paedocypris*, *Sundadanio*, and *Danionella*. Because paedomorphism is known to confound morphological phylogenetic analysis, Britz & Conway (2009) excluded characters that were absent in other miniature taxa and thus assumed to be convergent among small cypriniforms. They also excluded characters that were late-developing in zebrafish, to exclude characters that would be absent due to developmental truncation. Four absences were found that were shared amongst *Paedocypris*, *Sundadanio*, and *Danionella*, and five additional absences were found that were shared between *Paedocypris* and *Danionella*. Some other characters were described as progressive, where these characters are hypertrophied in the paedomorphic taxa (Britz and Conway 2009). Two of these progressive characters are apparently related to additional connections between different branchial arches in *Paedocypris* and *Danionella*. Four of these progressive characters are shared between *Paedocypris*, *Sundadanio*, and *Danionella*; these characters all involve hypertrophications of internal anatomy related to the Weberian appartus, an organ in fishes used for hearing (Britz and Conway 2009). In total, Britz & Conway (2009) found eight characters were shared among *Paedocypris*, *Sundadanio*, and *Danionella*, and an additional seven characters united *Paedocypris* and *Danionella*. Additional single-gene or few-gene analyses were published with cytochrome b, Rhodopsin, and RAG1 which had moderate support for a variety of relationships of *Sundadanio* with Danionidae, Cyprinidae, and other parts of Cyprinoidei, but fairly consistent placements for *Danionella* as part of Danioninae (Conway et al. 2008; Fang et al. 2009; Britz et al. 2009).

With six nuclear genes (RAG1, Rh, EGR1, EGR2B, EGR3, IRBP), Chen & Mayden (2009) reconstructed the relationships of a small number of species of Cypriniformes. Of the paedomorphic taxa, this study only included *Danionella*, which was recovered as sister to members of Danioninae, excluding *Esomus*. This was soon followed up by Mayden & Chen (2010), the first study since Rüber et al. (2007) to include the three paedomorphic genera. These authors announced that *Paedocypris* was sister to the remainder of Cypriniformes with 100% bootstrap support. *Sundadanio* was recovered sister to a cyprinoid genus *Leptobarbus* with moderate support (76% bootstrap support), and these two sister taxa were sister to a large clade of cyprinoid families including Xenocyprididae, Gobinoidae, Acheilognathidae, Tanichthyidae, Tincidae, and Leuciscidae (97% bootstrap support). *Danionella* was again recovered as part of the Danionidae, sister to the remaining Danioninae. Mayden & Chen (2010) argued that Britz & Conway's (2009) morphological study that supported the paedomorphic taxa as a clade did not include a phylogenetic analysis, as no other taxa were included to determine if shared characters were convergent or homoplasious, and thus criticizing the recovered grouping as due to similarity and not synapomorphy. They also mapped morphological characters by Fink and Fink (1981) for ostariophysan fishes onto the molecular phylogeny and reinterpret Britz & Conway's (2009) morphological characters in a phylogenetic context, arguing that the morphological characters are actually congruent with respect to the relationships recovered from molecular data. Lastly, given that cyprinoid clades were being elevated to family level, Mayden & Chen (2010) described family-rank names for *Paedocypris* (Paedocyprididae) and *Sundadanio* (Sundadanionidae). In the same issue of Molecular Phylogenetics & Evolution, Tang et al. (2010) published a molecular phylogeny focused on the Danionidae based on two mitochondrial (Cyt b, COI) and two nuclear genes (RAG1, Rh). These authors instead recover *Paedocypris* and *Sundadanio* as sister taxa within Danioninae. Despite the poor support for the sister relationship of *Paedocypris* and *Sundadanio*, the monophyly of Danioninae, and the monophyly of Danionidae, Tang et al. (2010) synonymized Paedocyprididae within 36 pages of its description. However, reanalysis of the data from Tang et al. (2010) does not yield the published topology (Chapter 1). Furthermore, Tang et al. (2011) included these sequences in part of their phylogenetic analysis of

Gobionidae; maximum parsimony and maximum likelihood analyses were congruent with Tang et al. (2010), while Bayesian analyses were congruent with Mayden & Chen (2010).

The high conflict for the placements of paedomorphic taxa in simultaneous publications from the Cypriniformes Tree of Life prompted Britz & Conway (2011) to criticize the Cypriniformes Tree of Life project with an article titled “The Cypriniformes Tree of Confusion.” Subsequently, Britz et al. (Britz et al. 2014) struck back at Mayden & Chen’s (2010) critique of their work by performing both a reanalysis of Mayden & Chen’s morphological data and an expanded dataset based on Britz & Conway’s (2009) earlier morphological data, as well as reanalysis of both Rüber et al.’s (2007) cytochrome b data and Mayden & Chen’s (2010) six-nuclear gene dataset. Reanalysis of Mayden & Chen’s (2010) morphological data without mapping on the molecular phylogeny supported a close relationship of paedomorphic taxa. Additionally, reanalyses of Britz & Conway’s (2009) morphological data combined with Conway’s (Conway 2011) morphological character matrix for Cypriniformes also supported a clade of the paedomorphic taxa, whether or not the putatively convergent characters were included or excluded (Britz et al. 2014). For re-analyses of molecular data, Britz et al. (2014) demonstrated that most of the loci were equivocal on the relationships of *Paedocypris* with a combination of tree topology tests, phylogenetic network analysis, splits support spectrum analysis, and per-site likelihood analysis; on the other hand, EGR3 strongly supported the placement of *Paedocypris* as sister to Cypriniformes, and they argued this gene was biased by systematic error.

Phylogenomics has emerged as a method to help resolve incongruent relationships due to conflict from sampling error or weak phylogenetic signal among individual genes (Philippe et al. 2005; Lemmon and Lemmon 2012). By increasing locus sampling, the signal-to-noise ratio can be increased and stochastic error can be reduced. New methods have been developed that have been specifically tailored for use in phylogenomics by enriching loci using bait-mediated sequence capture (Lemmon et al. 2012; Faircloth et al. 2012). Anchored hybrid enrichment is an attractive option for addressing the phylogenetic uncertainties still present within Cypriniformes, including of Danionidae and its putative members. This study provides the first phylogenomic analysis across the order and the largest study by locus-sampling for addressing the phylogenetic relationships of Danionidae to date.

METHODS

Taxon Selection and Tissue Preparation

The 172 taxa selected for this study (Additional file 1: Table S1) represent almost all major groups within the order. Species were chosen based on tissue availability and because of their incorporation in recent studies that will allow for direct comparisons (Saitoh et al. 2006; Mayden et al. 2007; Chen and Mayden 2009; Chen et al. 2009; Bufalino and Mayden 2010; Tang et al. 2011; 2013). Three outgroup taxa were chosen to represent the three other ostariophysan orders: Siluriformes, Gymnotiformes, and Characiformes.

Whole genomic DNA was prepared using the Omegabiotek E.Z.N.A. animal tissue extraction kit (product #D3396-02) and verified for quality and quantity using gel electrophoresis and nanodrop, respectively.

Locus Selection and Probe Design

Although the Anchored Hybrid Enrichment kit developed for vertebrates by Lemmon et al. (Lemmon et al. 2012; Lemmon and Lemmon 2012) contains a fish reference (*Danio*) and has been utilized in teleosts with moderate success (Eytan et al. 2015), we desired an enrichment tool more efficient and appropriate for phylogenomics in teleosts. Because of the complex nature of teleost genome evolution, which involved multiple whole-genome duplications and lineage-specific gene losses (Glasauer and Neuhauss 2014), it is impractical to identify a set of loci that are truly single-copy across all of Teleostei. Previous studies claiming to have identified single copy orthologs in teleosts (eg. Li 2007) likely only identified loci that were single-copy in the species they considered (an over-fitting problem). Evaluation of those loci in additional teleost lineages suggests that these loci are not universally single-copy (see below). Consequently, we aimed to target loci containing up to four gene copies in each of three diverse lineages of teleosts: zebrafish, platyfish, and cichlids.

Candidate target regions for Teleostei were derived by combining the 394 Vertebrate Anchor (v2) loci of Prum et al. (2015) and the 135 loci identified as Fugu-*Danio* single-copy orthologs by Li (2007). For the vertebrate anchor loci, teleost orthologs were obtained for *Danio rerio* (danRer7) using the human (hg19) coordinates and the USCS genome browser batch-coordinate (liftover) tool (Kent et al. 2002). For the Fugu-*Danio* orthologs, orthologous human (hg19) and chicken (galGal3) coordinates were obtained using the USCS liftover tool and the *Danio* coordinates identified by Li

(2007). Once the coordinates for *Danio*, *Homo*, and *Gallus* were obtained for all 529 candidate target regions, sequences corresponding to those regions [plus sufficient flanking region to obtain up to 3000 base pairs (bp) total] were extracted from the genomes and aligned by locus using MAFFT (Katoh and Standley 2013), v7.023b with "–genafpair" and "–maxiterate 1000" flags. The alignments were then used to generate a *Danio*-specific reference database containing spaced 20-mers. The *Danio* reference was then used to identify homologous regions in the genomes of zebrafish (Cypriniformes: Cyprinidae: *Danio rerio*; danRer7), platyfish (Cyprinodontiformes: Poeciliidae: Xiphophorus maculatus; Schartl et al. 2013), and cichlid (Loh et al. 2008).

As expected, we recovered multiple homologs for many of the candidate loci (only 64 loci were single copy in all three species). Consequently, only 277 loci had fewer than five homologs per species and were considered further. We aligned with MAFFT (Katoh and Standley 2013), v7.023b with "–genafpair" and "–maxiterate 1000" flags) all homolog sequences (up to 12 per locus) for each of the 277 candidates together with the homologous human probe region sequence from the Vertebrate Anchor (v2) design. Alignments were then manually inspected for misplaced and grossly misaligned sequences, which were removed. Finally, alignments were trimmed to include regions best suited for Anchored Hybrid Enrichment (conserved, low-gap, high taxon representation), taking care that the chosen region contained the human probe region. A total of 260 loci were retained.

Finally, in order to ensure efficient enrichment, we checked for high-copy regions (e.g. microsatellites and transposable elements) in each of the three teleost references as follows. First, a database was constructed for each species using all 15-mers found in the trimmed alignments for that species. We also added to the database all 15-mers that were 1bp removed from the observed 15-mers. The genome for the species was then exhaustively scanned for the presence of these 15-mers and matches were tallied at the alignment positions at which the 15mer was found. Alignment regions containing > 100,000 counts in any of the three species were masked to prevent probe tiling across these regions. Probes of 120bp were tiled uniformly at 5.5x tiling density.

Data Collection

Multilocus sequence data were collected at the Center for Anchored Phylogenomics at Florida State University (www.anchoredphylogeny.com) following Lemmon et al. (2012) with some adjustments. Each genomic DNA sample was sonicated to a fragment size of ~175-300 bp using a Covaris E220 Focused-ultrasonicator with Covaris microTUBES. Library preparation and indexing followed Meyer and Kircher (2010). Indexed libraries were pooled at equal quantities (12 pools of 16 samples each), and the library pools were enriched using a custom Agilent Custom SureSelect kit (Agilent Technologies), with probes designed as described above. The 12 enriched library pools were pooled with equal quantities for sequencing on 4 PE150 Illumina HiSeq2000 lanes with 8bp indexing. Sequencing was performed at Florida State University in the College of Medicine Translational Science Laboratory.

Data Analysis

Reads were quality filtered using Illumina's Casava software with the chastity filter set to high. In order to increase read length and accuracy overlapping reads were then merged following Rokyta, Lemmon, and Aronow (Rokyta et al. 2012). Non-overlapping read pairs were kept separate but still used in the assembly. All reads were then assembled into contigs following Prum et al. (2015) using mapping references derived from the zebrafish, platyfish, and cichlid sequences used for probe design. This assembler produces separate contigs for gene copies differing by more than 5% sequence divergence. To reduce errors caused by low-level indexing errors during sequencing, contigs were then filtered by removing those derived from fewer than 50 reads.

Sets of homologs were produced by grouping by target locus (across individuals) and the filtered consensus sequences. Orthology was then determined for each target locus as follows. First, a pairwise distance measure was computed for pairs of homologs, with distance being computed as the percentage of 20-mers observed in the two sequences that were found in both sequences. A neighbor-joining clustering algorithm was then used to cluster the consensus sequences in to orthologous sets, with at most one sequence per species in each orthologous set (see Prum et al. 2015 for details). In order to minimize the effects of missing data, clusters containing fewer than 130 (72%) of the species were removed from downstream processing.

Sequences in each orthologous set were aligned using MAFFT v7.023b (Katoh and Standley 2013) with --genafpair and --maxiterate 1000 flags. In order to remove poorly aligned regions raw alignments were then trimmed and masked following Prum et al. (2015), with the following adjustments: sites with > 50% similarity were identified as good, 20 bp regions containing < 14 good sites were masked, and sites with fewer than 30 unmasked bases were removed from the alignment.

For all phylogenetic analyses, sequences from the gymnotiform, siluriform, and characiform species were used as the outgroup. For the concatenated dataset, the alignment was partitioned by locus and the phylogeny estimated using RAxML using GTR+ Γ model with 500 bootstrap replicates. For the species tree analysis, a maximum likelihood phylogeny was estimated with 100 bootstrap replicates for each of the separate loci using RAxML with GTR+ Γ model assumed. We then used the RAxML bootstrap trees to estimate a species tree using STAR (Liu et al. 2009) with default parameters using STRAW (Shaw et al. 2013). ASTRAL-II (v4.10.2) was also used for species tree inference using the gene trees and their 100 bootstrap replicates (Mirarab and Warnow 2015). We performed 100 replicates of multi-locus bootstrapping.

To evaluate previous morphological hypotheses relative to our analyses, we re-examined the datasets in Conway (2011) and Britz et al. (2014) by running 1000 replicates of a heuristic search in PAUP* (Swofford 2002). We traced the characters in Mesquite v.3.04 (Maddison and Maddison 2015). We also performed Bayesian analyses on these morphological datasets under the Mk+Γ model in mrBayes 3.2 (Ronquist et al. 2012), which has been demonstrated to perform better than parsimony due to rate heterogeneity in character evolution (Wright and Hillis 2014). Estimating rate heterogeneity can be biased by sampling only variable or parsimony-informative characters, so we analyzed the data with correction for parsimony-informative characters for the Conway (2011) dataset and variable characters for the Britz et al. datasets (one character in these datasets was not parsimony-informative). For each dataset, we ran MCMC with two runs of four chains for 1,000,000 generations, sampling every 1,000. We assessed convergence using Tracer v1.6 (Rambaut and Drummond 2013).

RESULTS

A total of 315,288 base pairs (bp) spanning 219 loci were recovered for use in estimating the phylogenetic relationships. Average locus length was 1011bp with a range of 134-2119bp (Fig. 1). The total number of informative characters was 295,252 bp with only 3.48% missing data. Our results show promise for the ability of this method to provide robust support for relationships in Cypriniformes, with 97% of nodes recovered at 100% bootstrap support (Fig. 2-4). We find support for Mayden and Chen's (2010) recognition of Paedocyprididae (represented only by *Paedocypris*) and Sundadanionidae (represented only by *Sundadanio*), since neither was recovered within Danionidae. Paedocyprididae was recovered as sister to the remainder of Cyprinoidei. Sundadanionidae is recovered as sister to large clade of Cyprinoidei formed by Xenocyprididae, Gobionidae, Acheilognathidae, Tanichthyidae, and Leuciscidae (Fig. 2-4).

Danionidae, with the exclusion of *Paedocypris* and *Sundadanio*, is robustly supported as monophyletic, as well as its subclades Danioninae, Rasborinae, and Chedrinae. Danioninae and Chedrinae are recovered as sister clades, with Rasborinae sister to that clade. *Danionella* is recovered, as in previous results, as sister to the remaining Danioninae. Differing from previous results, we recover *Esomus* with robust support as sister to the remaining Danionidae, independent of the three other subclades.

Results from species tree analyses are congruent to the concatenation analysis with respect to the monophyly of Danionidae, the relationships of *Paedocypris* and *Sundadanio*, and the monophyly and relative relationships of *Esomus*, Rasborinae, Danioninae, and Chedrinae (Fig. 5, 6). The STAR analysis disagrees with concatenation on the placement of Danionidae in Cypriniformes (Fig. 5). In concatenation, Danionidae is recovered as sister to a large cyprinoid clade formed by Sundadanionidae, Xenocyprididae, Gobionidae, Acheilognathidae, Tanichthyidae, and Leuciscidae. In the STAR analysis, Danionidae is recovered as sister to Cyprinoidei except for *Paedocypris*. In the ASTRAL analysis, Danionidae is recovered in the same place as in the concatenation analysis, but with poor bootstrap support of only 45% (Fig. 6). Despite both STAR and ASTRAL being statistically consistent under the multispecies coalescent, algorithmic differences likely contribute to this conflict.

Discussion

We have presented the first order-wide, phylogenomic analysis of the Cypriniformes, and we demonstrate the utility of anchored enrichment at assessing the relationships of fishes from deep to more recent divergences. Our analyses have robustly supported the placement of *Paedocypris* as sister to all other cyprinoids, *Sundadanio* as sister to a major clade of cyprinoids, and confirmation of three major clades within Danionidae previously recovered. The anchored enrichment phylogenomic tree that we present provides the most robust phylogenetic analysis to date, supporting many of the previous hypotheses of relationships and providing new ideas that will require further scrutiny, such as the relationships among Cobitoidei (Chapter 3).

Differing from previous studies, we find the genus *Esomus* as a separate lineage sister to all remaining members of the Danionidae. The placement of *Esomus* has been contentious (Liao et al. 2011). *Esomus* has been placed as closely related to *Danionella* or *Sundadanio* within Danioninae with poor support (Conway et al. 2008; Fang et al. 2009;

Tang et al. 2010). Because of poorly supported nodes, molecular phylogenies are ambiguous on the placement of *Esomus* among the clades of Danionidae. Liao et al. (2011) remark that *Esomus* has a long branch in molecular phylogenetic analyses, and this may attract this branch towards other long branches such as *Danionella* and *Sundadanio*. Using morphological characters, Liao et al. (2011) recovered *Esomus* as sister to all other members of Chedrinae based on four characters, including two acquired states and two homoplasious states. In a subsequent paper, they admit this topology is never recovered in molecular analyses (Liao et al. 2012). Both of the acquired character states relate to the postcleithrum: first its presence, and secondly its orientation. In *Esomus*, the postcleithrum is absent, and the postcleithrum orientation was coded as missing, and thus may not be informative on its placement relative to the Chedrinae. Additionally, although postcleithrum absence within Danionidae is only found in the Chedrinae, postcleithrum absence is also found in disparate genera from multiple cyprinoid groups including leuciscids, cyprinids, and gobionids (Liao et al. 2011). Morphological homoplasy, long branch attraction, and short intervening branch lengths between danionid clades may have all contributed to the varying placement of *Esomus* between molecular and morphological studies.

We do not recover *Paedocypris* and *Sundadanio* within Danionidae. Our analyses are somewhat similar to Mayden and Chen (2010), who proposed the exclusion of these two genera from Danionidae. Our results are congruent with their placement of *Sundadanio*, but we recover *Paedocypris* as a lineage sister to the remainder of Cyprinoidei rather than sister to all of Cyprinformes. Although our finding is incongruent with the published topology of Mayden and Chen (2010), it is congruent with an unpublished mitogenome analysis mentioned by Mayden & Chen (2010), suggesting that the 13 mitochondrial genes may contain more phylogenetic signal than six nuclear genes for the recovery of this relationship. Our results robustly support the recognition of Paedocyprididae and Sundadanionidae based on the recovery of these genera as independent lineages among Cypriniformes, separate from the remaining members of Danionidae.

Britz et al. (2014) provide considerable morphological support for the paedomorphic taxa forming a monophyletic clade, even when using the dataset of Conway (2011) that did not include characters specific to paedomorphs. We reanalyzed the datasets of Britz et al. (2014), and found that even with their morphological dataset 3 (Conway 2011), that there were three character changes supporting all paedomorphs as monophyletic and nine character changes uniting *Paedocypris* and *Danionella*. Adding in characters specific to the paedomorphs (morphological datasets 4 and 5 from Britz et al. 2014) only increases the level of support. Under Bayesian analysis, the support for paedomorphic taxa forming a clade is weak in morphological dataset 3 (0.76 pp) but increases dramatically with addition of the paedomorphic-specific characters of datasets 4 and 5 (1.00 pp). We believe the weak support for the relationships of the various cyprinoids in the original dataset explains the disparity between the morphological and molecular hypotheses. In both the parsimony and Bayesian reanalyses of Britz et al.'s (2014) morphological dataset 3, the basal relationships of the cyprinoids are an almost complete comb. Without strong support for relationships within the Cyprinoidei, the dataset is insufficient for distinguishing synapomorphy from convergence among the paedomorphs, and adding characters specific to paedomorphs will only decrease the ability of the morphology to detect convergence. Conway's (2011) dataset already contains a high level of homoplasy before the addition of the paedomorphs, indicating that morphological evolution within Cypriniformes was rapid. The support in our dataset for three separate transitions to paedomorphism is strong, suggesting convergence in morphology, and we find at least five character changes in the Britz et al.'s morphological dataset 3 that support monophyly of the cyprinoids minus *Paedocypris* (21:1, 24:1, 34:1, 82:1, 101:0).

CONCLUSION

The Cypriniformes is among the most important clades of freshwater fishes and among the most studied with phylogenetic inference. This great deal of work makes it a key group in understanding the various pit-falls of phylogenetic studies, and it exemplifies the phylogenetic conflicts from the varying analyses of morphological, mitochondrial, and nuclear data. While some major clades of Cypriniformes have been long-supported, relationships within and among them have proven difficult to resolve across the entire order. Varying markers and morphological data have given different results and have been difficult to apply across such a large and diverse group. With the advent of phylogenomics, researchers can now acquire a substantial amount of highly informative, quality data for resolving dynamic relationships, and we demonstrate the efficacy of the approach using the complex relationships of cypriniforms.

The great diversity of Cypriniformes and the inclusion of perhaps the most important vertebrate model organism (Zebra Danio) also make Cypriniformes an ideal group for comparative analyses. Considerable insight into the functioning of genes within vertebrate organisms has been obtained from the analysis of the Zebra Danio including forced mutations that often result in unviable larvae. By comparing the genome of the Zebra Danio with close relatives, the role of mutations and gene expression can be determined. Comparative genomic studies within Cypriniformes have already benefited from the foundation and annotation of the Zebra Danio genome sequence to generate insights into the functional evolution of various adaptations including adaptation to harsh environments such as caves and high altitude streams (Meng et al. 2013; Yang et al. 2015b). With a robust phylogeny, we can get a much better understanding of the function of genes by treating relatives of the Zebra Danio as natural mutants screened by natural selection . As the Cypriniformes continues to become a more genome-enabled clade, with several new genomes published in the last few years (Xu et al. 2014; Burns et al. 2015; Yang et al. 2016), we expect our phylogeny to provide a useful framework for comparative genomics (Chapter 4).

REFERENCES

Betancur-R R., Broughton R.E., Wiley E.O., Carpenter K., Andrés López J., Li C., Holcroft N.I., Arcila D., Sanciangco M.D., Cureton J.C. II, Zhang F., Buser T., Campbell M.A., Ballestros J.A., Roa-Varon A., Willis S.C., Borden W.C., Rowley T., Reneau P.C., Hough D.J., Lu G., Grande T., Arratia G., Ortí G. 2013. The Tree of Life and a New Classification of Bony Fishes. PLOS Currents Tree of Life.:1–45.

Betancur-R R., Wiley E.O., Bailly N., Miya M., Lecointre G., Ortí G. Phylogenetic Classification of Bony Fishes --Version 3. Available from http://www.deepfin.org/Classification_v3.htm.

Britz R., Conway K.W. 2009. Osteology of *Paedocypris*, a miniature and highly developmentally truncated fish (Teleostei: Ostariophysi: Cyprinidae). J. Morphol. 270:389–412.

Britz R., Conway K.W. 2011. The Cypriniformes Tree of Confusion. In: de Carvalho M.R., Craig M.T., editors. Morphological and Molecular Approaches to the Phylogeny of Fishes: Integration or Conflict? Zootaxa. Magnolia Press. p. 73–78.

Britz R., Conway K.W., Rüber L. 2009. Spectacular morphological novelty in a miniature cyprinid fish, *Danionella dracula* n. sp. Proc. R. Soc. London Ser. B. 276:2179–2186.

Britz R., Conway K.W., Rüber L. 2014. Miniatures, morphology and molecules: *Paedocypris* and its phylogenetic position (Teleostei, Cypriniformes). Zool J Linn Soc. 172:556–615.

Bufalino A.P., Mayden R.L. 2010. Molecular phylogenetics of North American phoxinins (Actinopterygii: Cypriniformes: Leuciscidae) based on RAG1 and S7 nuclear DNA sequence data. Mol Phylogenet Evol. 55:274–283.

Burns F.R., Cogburn A.L., Ankley G.T., Villeneuve D.L., Waits E., Chang Y.-J., Llaca V., Deschamps S.D., Jackson R.E., Hoke R.A. 2015. Sequencing and de novo draft assemblies of a fathead minnow (*Pimephales promelas*) reference genome. Environ. Toxicol. Chem. 35:212–217.

Chang C.H., Li F., Shao K.T., Lin Y.-S., Morosawa T., Kim S., Koo H., Kim W., Lee J.-S., He S., Smith C., Reichard M., Miya M., Sado T., Uehara K., Lavoué S., Chen W.J., Mayden R.L. 2014. Phylogenetic relationships of Acheilognathidae (Cypriniformes: Cyprinoidea) as revealed from evidence of both nuclear and mitochondrial gene sequence variation: Evidence for necessary taxonomic revision in the family and the identification of cryptic species. Mol Phylogenet Evol. 81:182–194.

Chen W.J., Lheknim V., Mayden R.L. 2009. Molecular phylogeny of the Cobitoidea (Teleostei: Cypriniformes) revisited: position of enigmatic loach Ellopostomaresolved with six nuclear genes. J Fish Biol. 75:2197–2208.

Chen W.J., Mayden R.L. 2009. Molecular systematics of the Cyprinoidea (Teleostei: Cypriniformes), the world's largest clade of freshwater fishes: Further evidence from six nuclear genes. Mol Phylogenet Evol. 52:544–549.

Conway K.W. 2011. Osteology of the South Asian Genus Psilorhynchus McClelland, 1839 (Teleostei: Ostariophysi: Psilorhynchidae), with investigation of its phylogenetic relationships within the order Cypriniformes. Zool J Linn Soc.:no–no.

Conway K.W., Chen W.J., Mayden R.L. 2008. The "Celestial Pearl danio" is a miniature *Danio* (s.s) (Ostariophysi: Cyprinidae): evidence from morphology and molecules. Zootaxa. 1686:1–28.

Eschmeyer W.N., Fricke R., van der Laan R. 2016. Catalog of Fishes: Genera, Species, References.

Eytan R.I., Evans B.R., Dornburg A., Lemmon A.R., Lemmon E.M., Wainwright P.C., Near T.J. 2015. Are 100 enough? Inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. BMC Evol Biol.:1–20.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. Syst Biol. 61:717–726.

Fang F., Norén M., Liao T.Y., Källersjö M., Kullander S.O. 2009. Molecular phylogenetic interrelationships of the south Asian cyprinid genera *Danio*, *Devario* and *Microrasbora* (Teleostei, Cyprinidae, Danioninae). Zool Scripta. 38:237–256.

Fink S.V., Fink W.L. 1981. Interrelationships of the ostariophysan fishes (Teleostei). Journal of the Linnean Society of London, Zoology. 72:297–353.

Glasauer S.M.K., Neuhauss S.C.F. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol Genet Genomics. 289:1045–1060.

Howe K., Clark M.D., Torroja C.F., Torrance J., Berthelot C., Muffato M., Collins J.E., Humphray S., McLaren K., Matthews L., McLaren S., Sealy I., Caccamo M., Churcher C., Scott C., Barrett J.C., Koch R., Rauch G.-J., White S., Chow W., Kilian B., Quintais L.T., Guerra-Assunção J.A., Zhou Y., Gu Y., Yen J., Vogel J.-H., Eyre T., Redmond S., Banerjee R., Chi J., Fu B., Langley E., Maguire S.F., Laird G.K., Lloyd D., Kenyon E., Donaldson S., Sehra H., Almeida-King J., Loveland J., Trevanion S., Jones M., Quail M., Willey D., Hunt A., Burton J., Sims S., McLay K., Plumb B., Davis J., Clee C., Oliver K., Clark R., Riddle C., Eliott D., Threadgold G., Harden G., Ware D., Mortimer B., Kerry G., Heath P., Phillimore B., Tracey A., Corby N., Dunn M., Johnson C., Wood J., Clark S., Pelan S., Griffiths G., Smith M., Glithero R., Howden P., Barker N., Stevens C., Harley J., Holt K., Panagiotidis G., Lovell J., Beasley H., Henderson C., Gordon D., Auger K., Wright D., Collins J., Raisen C., Dyer L., Leung K., Robertson L., Ambridge K., Leongamornlert D., McGuire S., Gilderthorp R., Griffiths C., Manthravadi D., Nichol S., Barker G., Whitehead S., Kay M., Brown J., Murnane C., Gray E., Humphries M., Sycamore N., Barker D., Saunders D., Wallis J., Babbage A., Hammond S., Mashreghi-Mohammadi M., Barr L., Martin S., Wray P., Ellington A., Matthews N., Ellwood M., Woodmansey R., Clark G., Cooper J., Tromans A., Grafham D., Skuce C., Pandian R., Andrews R., Harrison E., Kimberley A., Garnett J., Fosker N., Hall R., Garner P., Kelly D., Bird C., Palmer S., Gehring I., Berger A., Dooley C.M., Ersan-Ürün Z., Eser C., Geiger H., Geisler M., Karotki L., Kirn A., Konantz J., Konantz M., Oberländer M., Rudolph-Geiger S., Teucke M., Osoegawa K., Zhu B., Rapp A., Widaa S., Langford C., Yang F., Carter N.P., Harrow J., Ning Z., Herrero J., Searle S.M.J., Enright A., Geisler R., Plasterk R.H.A., Lee C., Westerfield M., de Jong P.J., Zon L.I., Postlethwait J.H., Nüsslein-Volhard C., Hubbard T.J.P., Roest Crollius H., Rogers J., Stemple D.L. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 496:498–503.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. 2002. The Human Genome Browser at UCSC. Genome Res. 12:996–1006.

Kottelat M., Britz R., Tan H.H., Witte K.-E. 2006. *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. Proc. R. Soc. London Ser. B. 273:895–899.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Syst Biol. 61:727–744.

Lemmon E.M., Lemmon A.R. 2012. High-Throughput Genomic Data in Systematics and Phylogenetics. Annual review of Ecology, Evolution, and Systematics. 44:99–121.

Li C. 2007. A Genome-scale Approach to Phylogeny of Rayfinned Fish (Actinopterygii) and Molecular Systematics of Clupeiformes. :1–134.

Liao T.Y., Arroyave J., Stiassny M.L.J. 2012. Diagnosis of Asian *Raiamas* (Teleostei: Cyprinidae: Chedrina) with comments on chedrin relationships and previously proposed diagnostic characters for *Opsaridium* and *Raiamas*. Ichthyol Res. 59:328–341.

Liao T.Y., Kullander S.O., Fang F. 2011. Phylogenetic position of rasborin cyprinids and monophyly of major lineages among the Danioninae, based on morphological characters (Cypriniformes: Cyprinidae). J Zoolog Syst Evol Res. 49:224–232.

Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating Species Phylogenies Using Coalescence Times among Sequences. Syst Biol. 58:468–477.

Loh Y.-H.E., Katz L.S., Mims M.C., Kocher T.D., Yi S.V., Streelman J.T. 2008. Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. Genome Biology. 9:R113.

Maddison W.P., Maddison D.R. 2015. Mesquite: a modular system for evolutionary analysis. v3.04. Available from http://mesquiteproject.org.

Mayden R.L., Chen W.J. 2010. The world"s smallest vertebrate species of the genus *Paedocypris*: a new family of freshwater fishes and the sister group to the world"s most diverse clade of freshwater fishes (Teleostei: Cypriniformes). Mol Phylogenet Evol. 57:152–175.

Mayden R.L., Chen W.J., Bart H.L. Jr, Doosey M.H., Simons A.M., Tang K.L., Wood R.M., Agnew M.K., Yang L., Hirt M.V., Clements M.D., Saitoh K., Sado T., Miya M., Nishida M. 2009. Reconstructing the phylogenetic relationships of the earth's most diverse clade of freshwater fishes—order Cypriniformes (Actinopterygii: Ostariophysi): A case study using multiple nuclear loci and the mitochondrial genome. Mol Phylogenet Evol. 51:500–514.

Mayden R.L., Tang K.L., Conway K.W., Freyhof J., Chamberlain S., Haskins M., Schneider L., Sudkamp M., Wood R.M., Agnew M.K., Bufalino A.P., Sulaiman Z., Miya M., Saitoh K., He S. 2007. Phylogenetic relationships of *Danio* within the order Cypriniformes: a framework for comparative and evolutionary studies of a model species. J. Exp. Zool. 308B:642–654.

Meng F., Braasch I., Phillips J.B., Lin X., Titus T., Chungguang Z., Postlethwait J.H. 2013. Evolution of the eye transcriptome under constant darkness in *Sinocyclocheilus* cavefish. Mol Biol Evol. 30:1527–1543.

Meyer M., Kircher M. 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harbor Protocols. 2010:pdb.prot5448–pdb.prot5448.

Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 31:i44–i52.

Nelson J.S. 2006. Fishes of the World. New York: John Wiley & Sons.

Philippe H., Delsuc F., Brinkmann H., Lartillot N. 2005. Phylogenomics. Annual review of Ecology, Evolution, and Systematics. 36:541–562.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.

Rambaut A., Drummond A.J. 2013. Tracer v1.6. Available from http://tree.bio.ed.ac.uk/software/tracer/.

Rokyta D.R., Lemmon A.R., Margres M.J., Aronow K. 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (Crotalus adamanteus). BMC Genomics. 13:1.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol. 61:539–542.

Rüber L., Kottelat M., Tan H.H., Ng P.K.L., Britz R. 2007. Evolution of miniaturization and the phylogenetic position of *Paedocypris*, comprising the world's smallest vertebrate. BMC Evol Biol. 7:38.

Saitoh K., Sado T., Doosey M.H., Bart H.L. Jr, Inoue J.G., Nishida M., Mayden R.L., Miya M. 2011. Evidence from mitochondrial genomics supports the lower Mesozoic of South Asia as the time and place of basal divergence of cypriniform fishes (Actinopterygii: Ostariophysi). Zool J Linn Soc. 161:633–662.

Saitoh K., Sado T., Mayden R.L., Hanzawa N., Nakamura K., Nishida M., Miya M. 2006. Mitogenomic Evolution and Interrelationships of the Cypriniformes (Actinopterygii: Ostariophysi): The First Evidence Toward Resolution of Higher-Level Relationships of the World's Largest Freshwater Fish Clade Based on 59 Whole Mitogenome Sequences. J Mol Evol. 63:826–841.

Schartl M., Walter R.B., Shen Y., Garcia T., Catchen J., Amores A., Braasch I., Chalopin D., Volff J.-N., Lesch K.-P., Bisazza A., Minx P., Hillier L., Wilson R.K., Fuerstenberg S., Boore J., Searle S., Postlethwait J.H., Warren W.C. 2013. The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. Nature Genetics. 45:567–572.

Shaw T.I., Ruan Z., Glenn T.C., Liu L. 2013. STRAW: Species TRee Analysis Web server. Nucleic Acids Res. 41:W238–41.

Swofford D.L. 2002. Phylogenetic analysis using parsimony (PAUP), version 4.0 b10.

Tang K.L., Agnew M.K., Chen W.J., Hirt M.V., Raley M.E., Sado T., Schneider L.M., Yang L., Bart H.L. Jr, He S., Liu H.-Z., Miya M., Saitoh K., Simons A.M., Wood R.M., Mayden R.L. 2011. Phylogeny of the gudgeons (Teleostei: Cyprinidae: Gobioninae). Mol Phylogenet Evol. 61:103–124.

Tang K.L., Agnew M.K., Hirt M.V., Lumbantobing D.N., Raley M.E., Sado T., Teoh V.-H., Yang L., Bart H.L. Jr, Harris P.M., He S., Miya M., Saitoh K., Simons A.M., Wood R.M., Mayden R.L. 2013. Limits and phylogenetic relationships of East Asian fishes in the subfamily Oxygastrinae (Teleostei: Cypriniformes: Cyprinidae). Zootaxa. 3681:101–135.

Tang K.L., Agnew M.K., Hirt M.V., Sado T., Schneider L.M., Freyhof J., Sulaiman Z., Swartz E., Vidthayanon C., Miya M., Saitoh K., Simons A.M., Wood R.M., Mayden R.L. 2010. Systematics of the subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). Mol Phylogenet Evol. 57:189–214.

Tao W., Zou M., Wang X., Gan X., Mayden R.L., He S. 2010. Phylogenomic Analysis Resolves the Formerly Intractable Adaptive Diversification of the Endemic Clade of East Asian Cyprinidae (Cypriniformes). PLOS ONE. 5:e13508.

Wright A.M., Hillis D.M. 2014. Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. PLOS ONE. 9:e109210.

Xu P., Zhang X., Wang X., Li J.-T., Liu G., Kuang Y., Xu J., Zheng X., Ren L., Wang G., Zhang Y., Huo L., Zhao Z., Cao D., Lu C., Li C., Zhou Y., Liu Z., Fan Z., Shan G., Li X., Wu S., Song L., Hou G., Jiang Y., Jeney Z., Yu D., Wang L., Shao C., Song L., Sun J., Ji P., Wang J., Li Q., Xu L., Sun F., Feng J., Wang C., Wang S., Wang B., Li Y., Zhu Y., Xue W., Zhao L., Wang J.-T., Gu Y., Lv W., Wu K., Xiao J., Wu J., Zhang Z., Yu J., Sun X. 2014. Genome sequence and genetic diversity of the common carp, Cyprinus carpio. Nature Genetics. 46:1212–1219.

Yang J., Chen X., Bai J., Fang D., Qiu Y., Jiang W., Yuan H., Bian C., Lu J., He S., Pan X., Zhang Y., Wang X., You X., Wang Y., Sun Y., Mao D., Liu Y., Fan G., Zhang H., Chen X.Y., Zhang X., Zheng L., Wang J.-T., Le Cheng, Chen J., Ruan Z., Li J., Yu H., Peng C., Ma X., Xu J., He Y., Xu Z., Xu P., Wang J., Yang H., Wang J., Whitten T., Xu X., Shi Q. 2016. The *Sinocyclocheilus* cavefish genome provides insights into cave adaptation. BMC Biol.:1–13.

Yang L., Sado T., Hirt M.V., Pasco-Viel E., Arunachalam M., Li J., Wang X., Freyhof J., Saitoh K., Simons A.M., Miya M., He S., Mayden R.L. 2015a. Phylogeny and polyploidy: Resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). Mol Phylogenet Evol. 85:97–116.

Yang L., Wang Y., Zhang Z., He S. 2015b. Comprehensive Transcriptome Analysis Reveals Accelerated Genic Evolution in a Tibet Fish, Gymnodiptychus pachycheilus. Genome Biology and Evolution. 7:251–261.

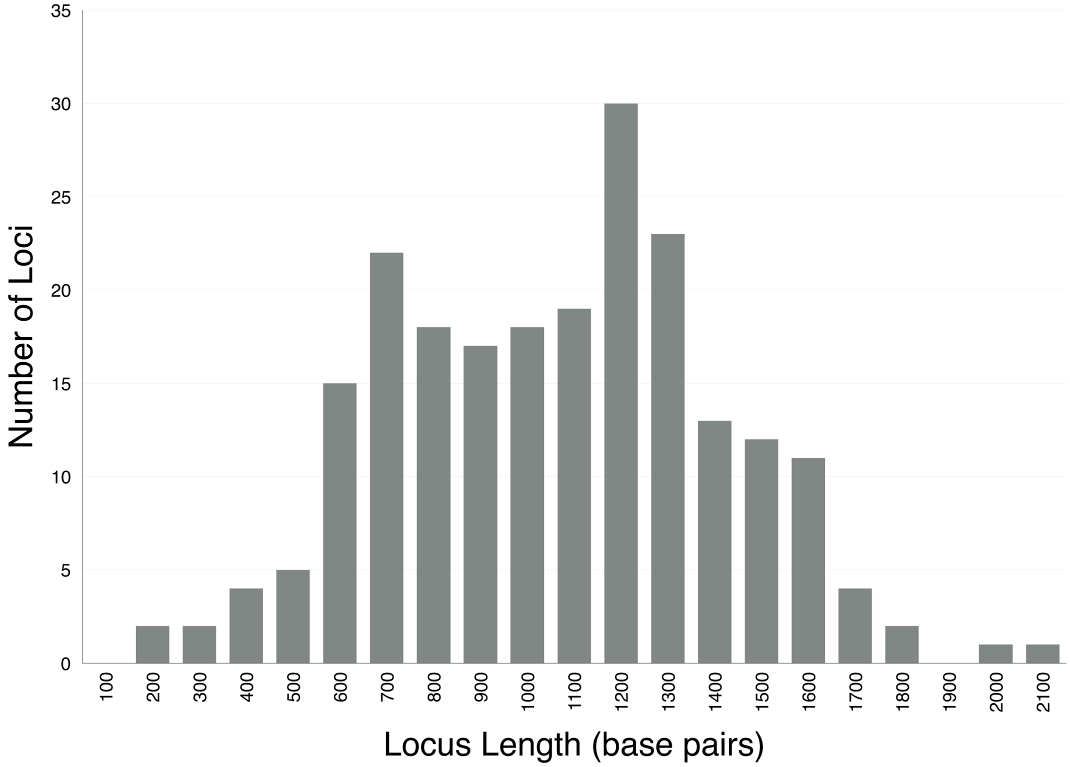


Figure 1. Histogram showing lengths of loci in base pairs.

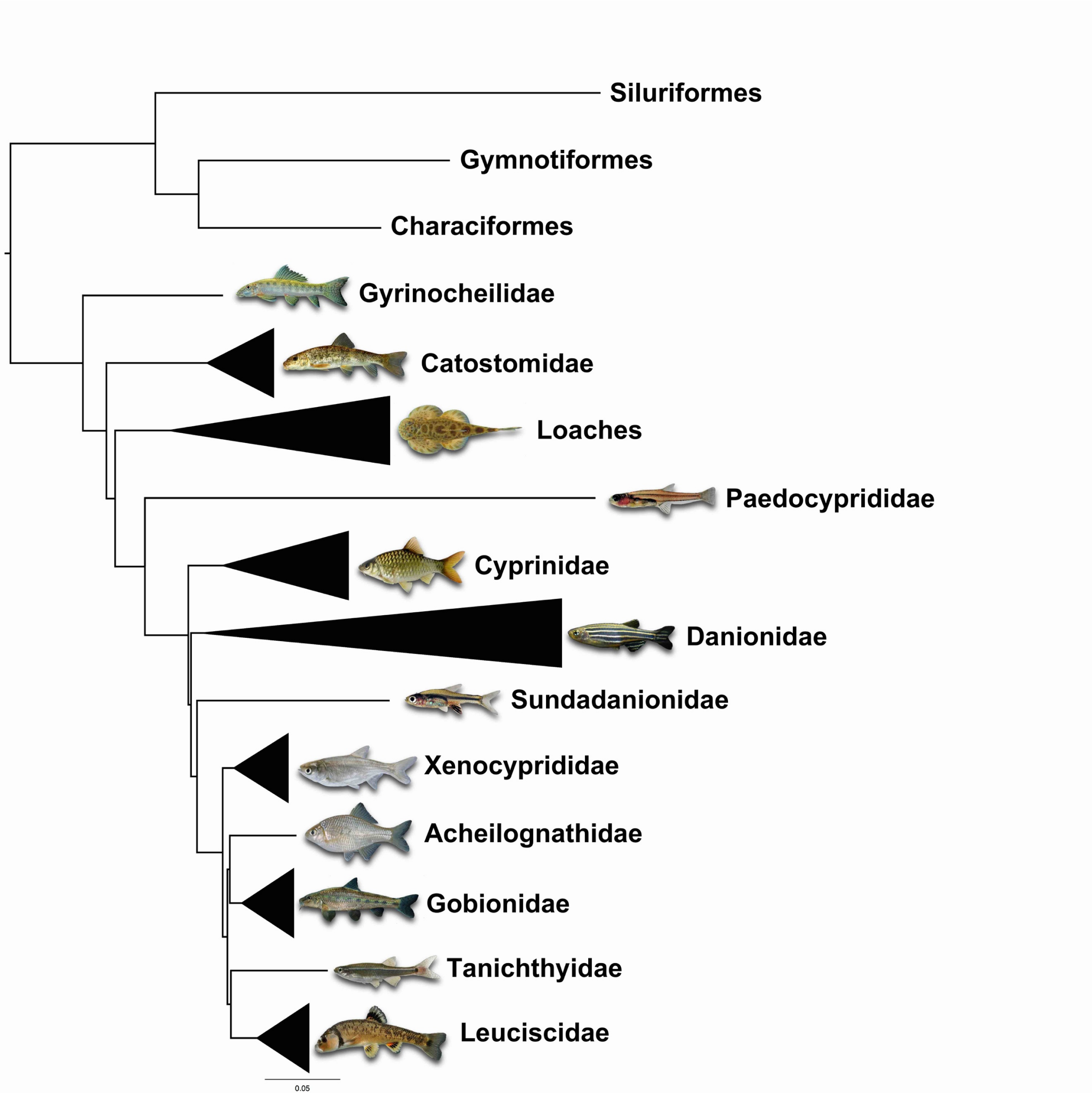


Figure 2. Maximum likelihood tree based on concatenation of all specimens collapsed into major clades. For all tree figures, all nodes shown are 100% bootstrap supported unless otherwise indicated, and the scale bar represents the number of nucleotide substitutions per site.

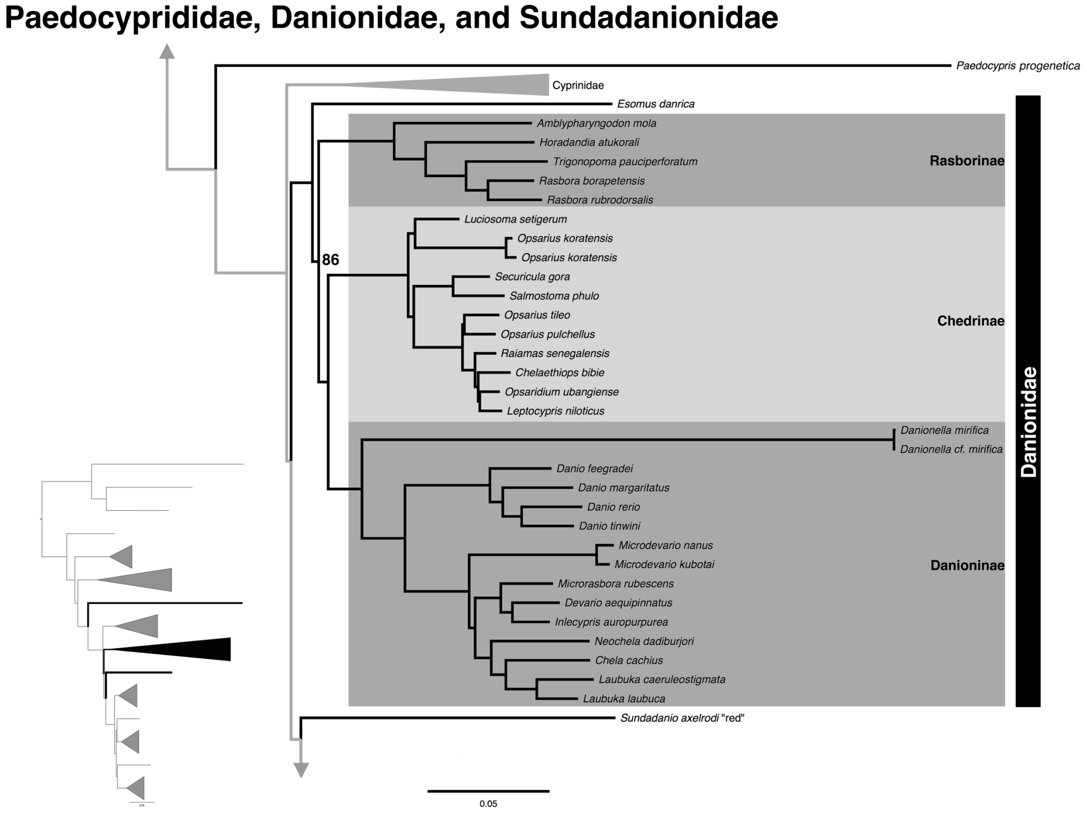


Figure 3. Expansion of Danionidae clade with labeled subfamilies. Also included are *Paedocypris* and *Sundadanio*, showing their placement outside of Danionidae.

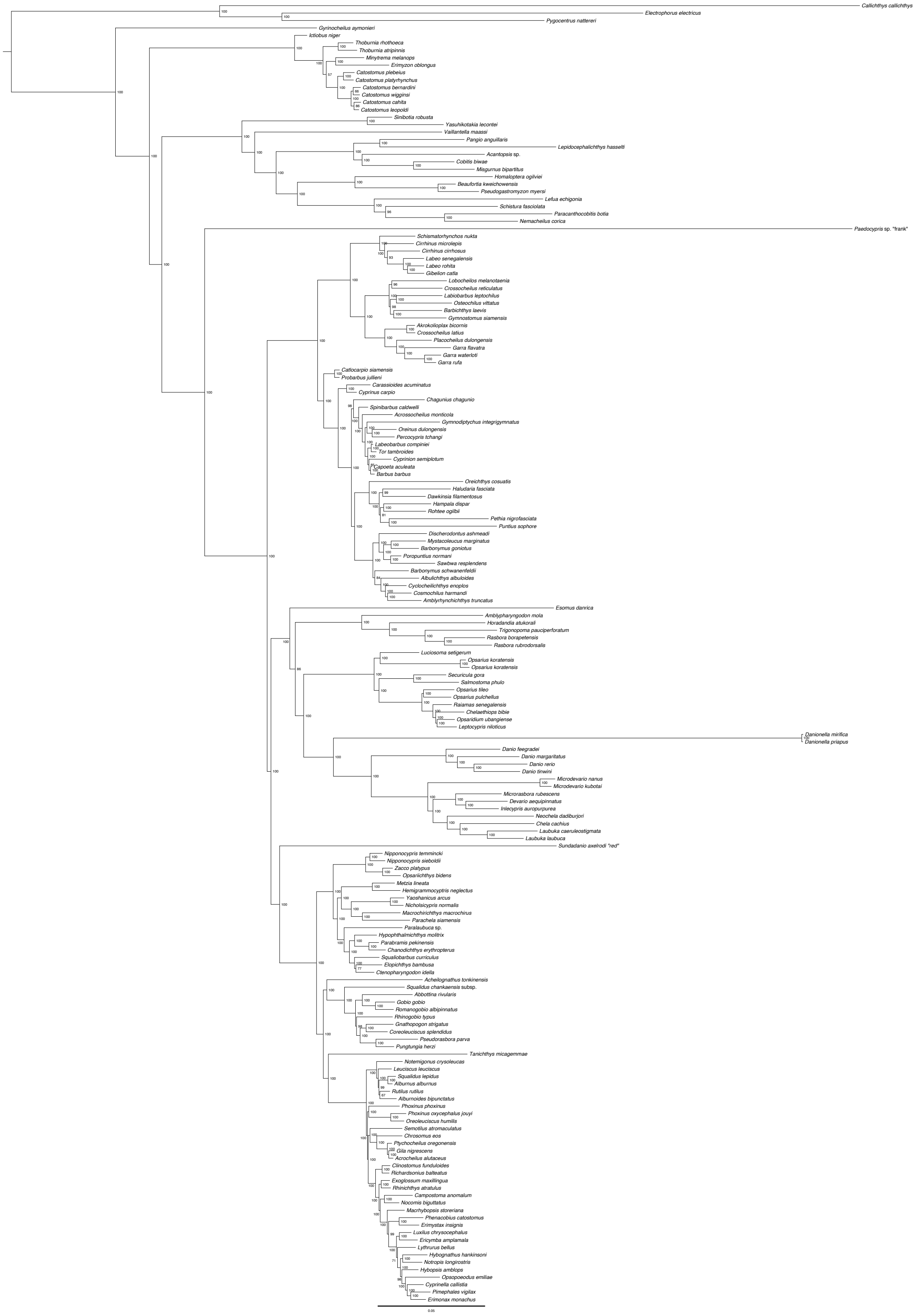


Figure 4. Complete maximum likelihood tree based on concatenation.

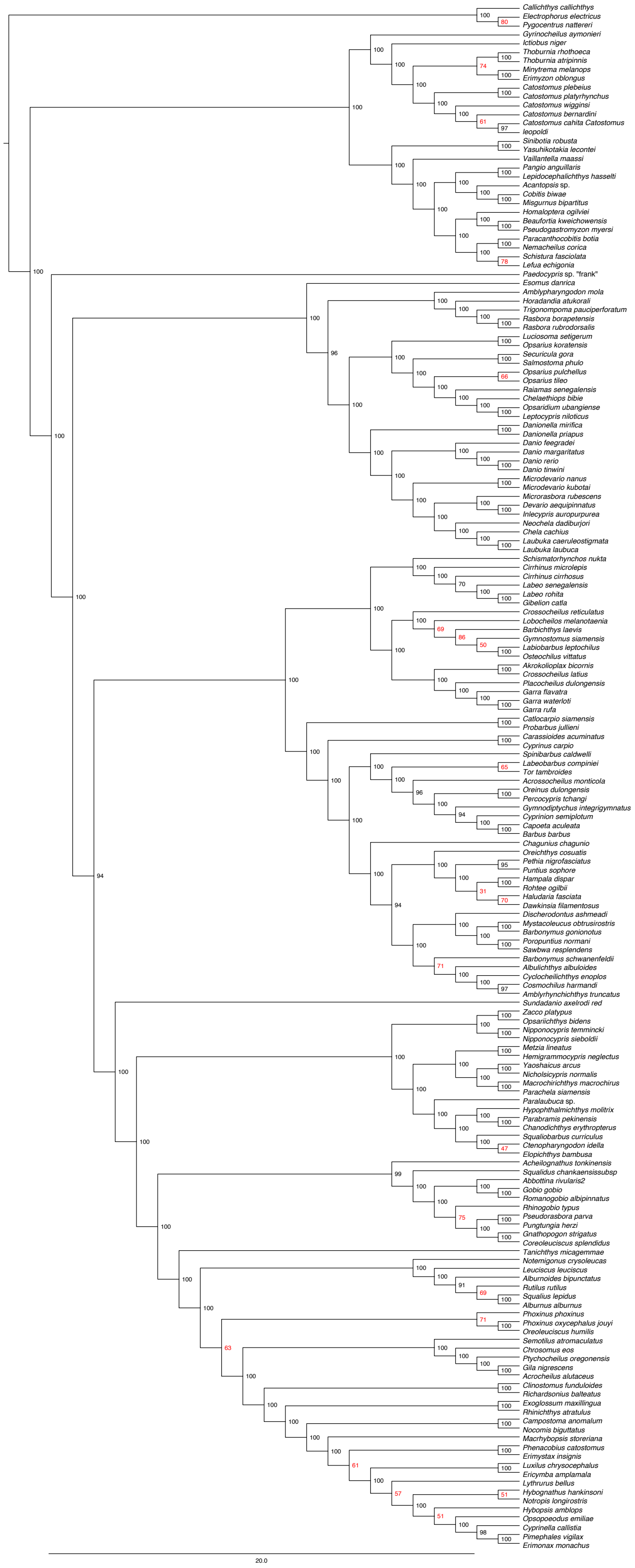


Figure 5. Collapsed cladogram of the STAR tree with Danionidae expanded.

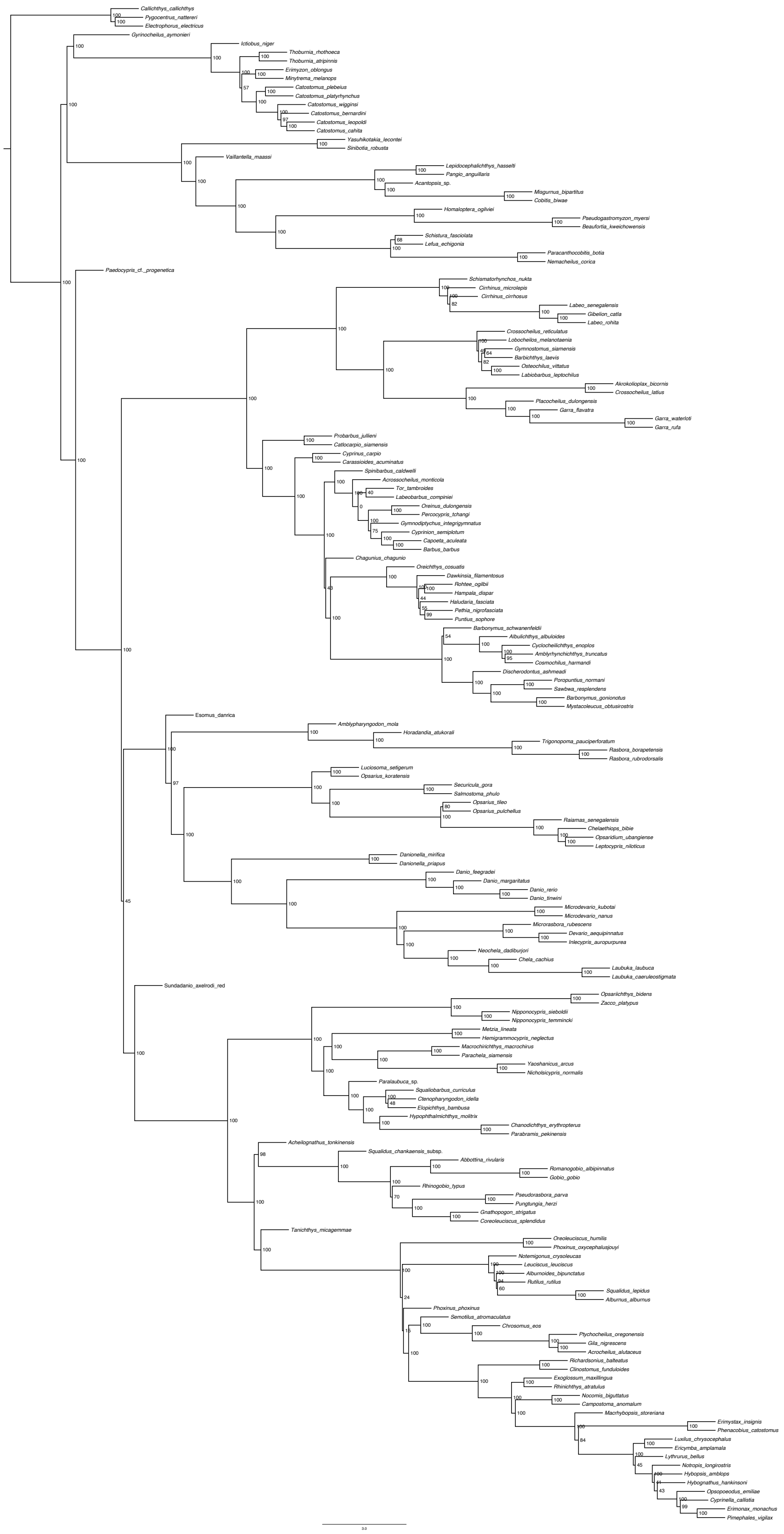


Figure 6. Collapsed cladogram of the ASTRAL tree with Danionidae expanded.

CHAPTER 3

PHYLOGENOMIC INTERROGATION OF SOURCES OF ERROR IN CONCATENATION AND COALESCENT ANALYSIS: A CASE STUDY ON CYPRINIFORMES

ABSTRACT

Phylogenomics has reduced issues with data limitation in estimating evolutionary relationships, however it has also highlighted the importance of data quality and model choice in phylogenetic analysis. Many sources of noise exist that can mislead phylogenetic inference, including base compositional heterogeneity, saturation, effects of long branches, and lack of phylogenetic signal. Furthermore, the multispecies coalescent model has arisen as a more-general model for phylogenomic inference relative to concatenation, but practical concerns with data quality remain. Using an anchored phylogenomics dataset for fishes of the order Cypriniformes, we study the effects of subsetting datasets to reduce biases that may confound phylogenetic inference on both concatenation and coalescent-based phylogenomic analysis. We demonstrate that both concatenation and coalescent-based analyses are sensitive to certain sources of phylogenetic noise, but this varies based on how the source of bias is quantified. Coalescent-based analyses appear to be more sensitive to data subsetting than concatenation on these data based on RF distances between analyses on different subsets. The relationships of the taxon *Paedocypris*, which has been particularly problematic to reconstruct, are insensitive to data subsetting, and is recovered as sister to Cyprinoidei in all analyses. The relationships of *Gyrinocheilus*, Catostomidae, and loaches conflict between concatenation and most coalescent-based analyses with respect to the placement

of the outgroup, though sensitivity to certain biases may confound coalescent-based analysis. These analyses demonstrate that previous results on major clades of Cypriniformes based on this dataset are insensitive to the biases that we studied herein, but long branches in the outgroup cause conflict between concatenation and coalescent-based analyses.

INTRODUCTION

Phylogenomics has burgeoned as a new hope in resolving the tree of life (Philippe et al. 2005). Empowered by new sequencing technologies, practitioners can now sequence hundreds to thousands of loci from across the genome to reconstruct evolutionary history, in particular to address difficult phylogenetic problems not solved by fewer loci. Utilizing many loci reduces the chances of unresolved phylogenetic relationships due to sampling error, where too little phylogenetic signal has been sampled to provide a clear picture of the relationships of a particular taxon (Philippe et al. 2005). On the other hand, large datasets do not reduce non-random, systematic error that represents phylogenetic noise, leading to overconfidence due to inflated estimates of support (Jeffroy et al. 2006; Philippe et al. 2011). With systematic error in phylogenomics potentially driving high precision for incorrect phylogenetic hypotheses, mutually exclusive topologies can have perfect support. To address this, exploration of sources of systematic error and their effect on phylogomic analysis can provide additional insight into the robustness of recovered relationships.

A variety of potential sources of error can bias phylogenomics. Paralogy and missing data are well-known sources of error that are important to minimize in phylogenomic studies (Lemmon and Lemmon 2012). However, even given datasets including only orthologous sequences and with low missing data, various other sources of systematic error exist. Various sequence biases (e.g. base compositional heterogeneity, saturation, GC bias) and evolutionary phenomena (e.g. evolutionary rate heterogeneity) can cause difficulty in reconstructing phylogenetic relationships given the limited ability of statistical models to accurately estimate evolution under real but complicated scenarios of molecular substitution (Struck 2014; Kück and Struck 2014). Pending the development of statistical models robust to these potential sources of noise, the selection of less misleading loci to phylogenetic inference can increase phylogenetic accuracy and reduce systematic error (e.g. Salichos and Rokas 2013; Kocot et al. 2013; Struck 2014; Doyle et al. 2015). In addition, the sensitivity of a phylogenomic analysis to various biases can be assessed by comparing analyses based on multiple subsets of the data with reduced bias, to determine the effect of those biases on phylogenomic analysis (Whelan et al. 2015; e.g. Borowiec et al. 2016). Subsetting should be a routine aspect of phylogenomic analysis to assess the support shown by various subsets of the data (Edwards et al. 2016). With many loci, there is the opportunity to interrogate the data and study the relative importance of a variety of sources of noise in phylogenomic analysis.

In addition, phylogenetic inferences are also highly dependent on model of evolution, and in particular much recent focus has concentrated on the multispecies coalescent model (Edwards 2009; Liu et al. 2015b; Springer and Gatesy 2016; Edwards et al. 2016). Coalescent methods reconstruct a species tree allowing for heterogeneous gene tree topologies to evolve within them (Edwards 2009), and the multispecies coalescent model simplifies to the model assumed by concatenation when gene tree topologies and branch lengths are the same (Liu et al. 2015b). Short internodes between rapid, successive branching events can cause an anomaly zone where the most common gene tree does not represent the species tree, and thus concatenation never recovers the true topology, leading to positively misleading results in which incorrect topologies are reconstructed with high support (Degnan and Rosenberg 2006; Kubatko and Degnan 2007; Degnan and Rosenberg 2009). Despite these theoretical concerns, concatenation has been useful for reconstructing much of the tree of life and is relatively rapid (Liu et al. 2015b). Method selection should be determined not just by theoretical concerns, but also practical concerns that arise from particular algorithms (Knowles et al. 2012). Fully-parameterized coalescent methods that reconstruct species trees from sequence alignments are too computationally intensive to perform on phylogenomic datasets, but coalescent methods that use gene trees as input data are relatively rapid, and are the only feasible option to analyze phylogenomic data under the multispecies coalescent (Mirarab et al. 2014a). These coalescent methods require two steps, the first step starting with inferring gene trees, and the second step that summarizes these relationships and gene tree statistics into a species tree (Mirarab et al. 2014a; Liu et al. 2015b).

Despite theoretical reasons why coalescent-based methods should be more accurate than concatenation, two-step coalescent methods may practically be more inaccurate than concatenation because of the low phylogenetic signal within genes, leading to inaccurate gene tree reconstruction that confounds analysis (Mirarab et al. 2014a). Coalescent-based analyses have been evaluated for their accuracy relative to a variety of sources of phylogenetic noise including long branch attraction, missing data, misrooting of gene trees, gene tree error, gene flow, and recombination (Liu et al. 2015a;

2015b). On the other hand, the effects of different sequence biases on phylogenomic analysis should be explored on a case-by-case basis (Kück and Struck 2014). Regardless, two-step coalescent methods are based on accurate gene trees, thus, accounting for potential biases that confound gene tree inference can improve both concatenation and coalescent methods. Furthermore, biases that can confound phylogenetic inference may bias one method more than the other. Also, although concatenation and coalescent-based models make different assumptions that can lead to mutually exclusive relationships, exclusion of biased sequences may allow these methods to converge towards a common tree topology.

We herein explore the sensitivity of phylogenomic analyses to numerous potential sources of biases and noise in a dataset focused on the relationships of fishes of the order Cypriniformes. Cypriniformes is the most diverse clade of freshwater fishes with over 4000 species (Nelson 2006; Mayden and Chen 2010; Eschmeyer et al. 2016). It includes a variety of fishes including algae eaters (Gyrinocheilidae), suckers (Catostomidae), and a monophyletic group formed by various families of fishes all colloquially called loaches, including Botiidae, Cobitidae, Vaillantellidae, Ellopostomatidae, Nemacheilidae, Balitoridae; (Mayden and Chen 2010; Kottelat 2012). Sometimes, the name name Cobitoidei is restricted to the loach clade (Kottelat 2012), while other authors choose to recognize Cobitoidei more broadly, so that it also includes Gyrinocheilidae and Catostomidae (Mayden and Chen 2010). The bulk of Cypriniformes has traditionally been classified within the family Cyprinidae. Mayden & Chen (2010) raised this grouping to superfamilial level, and Stout et al. (*submitted*) recognized this grouping at the suborder level (Cyprinoidei) to better fit with higher-level ray-finned fish taxonomy, where the taxonomic level below order and above family is generally the suborder (Betancur-R et al. 2013). The suborder Cyprinoidei includes a diversity of fishes including carps, minnows, barbs, gudgeons, bitterlings, and their kin, currently recognized to be grouped in a diversity of clades including Psilorhynchidae, Cyprinidae, Xenocyprididae, Paedocyprididae, Sundadanionidae, Acheilognathidae, Tanichthyidae, Tincidae, Gobionidae, Labeobarbidae, and Leuciscidae (Bufalino and Mayden 2010; Mayden and Chen 2010; Chang et al. 2014; Yang et al. 2015 Stout et al. submitted).

Cypriniform phylogeny has been controversial, particularly with respect to paedomorphic taxa such as *Paedocypris* and *Sundadanio* (Britz et al. 2014). These paedomorphic taxa have been recovered with different relationships between different studies. Rüber et al. (2007) and Tang et al. (2010; 2013) reported *Paedocypris* and *Sundadanio* group together as sister to other members of Danionidae, while Mayden & Chen (2010) reported *Paedocypris* was the sister group to Cypriniformes and *Sundadanio* was an independent branch in cyprinoid phylogeny. Tang et al. (2011) inconsistently recovered both topologies, with *Paedocypris* and *Sundadanio* within Danionidae in maximum likelihood and maximum parsimony, but *Paedocypris* sister to Cypriniformes and *Sundadanio* sister to *Leptobarbus* in Bayesian analysis. Britz & Conway (2009) and Britz et al. (2014), using morphological data, demonstrated numerous morphological characters that unite the paedomorphic taxa as a clade. Britz et al. (2014) implicated concerns with molecular data in reconstructing the relationships of paedomorphic taxa, and demonstrated that phylogenetic signal in molecular data was weak and unable to decisively place *Paedocypris*. Britz et al. (2014) also suggested that systematic error may affect a particular locus, EGR3, that supported an early-branching position for the

Paedocypris lineage. Long branches in the paedomorphic taxa *Paedocypris*, *Sundadanio*, and *Danionella* can potentially contribute to error and these taxa may be pulled towards the outgroup (Britz et al. 2014). With only a few loci available, it is possible the low signal and high conflict may be a combination of sampling error and systematic error that particular makes these relationships problematic to reconstruct.

To address data limitation and overcome sampling error, Stout et al. (*submitted*) utilized anchored phylogenomics to sequence hundreds of nuclear loci to infer the relationships across Cypriniformes. They recover the phylogenetic relationships of Cypriniformes with extremely high support with the majority of nodes at 100% bootstrap support. *Paedocypris* was recovered as sister to other cyprinoids, contrasting with previously published hypotheses (although mentioned as recovered in an unpublished mitogenomic study by Mayden and Chen 2010). Nevertheless, despite the large number of loci, Stout et al. (*submitted*) also recovered conflicting relationships for the Gyrinocheilidae, Catostomidae, and loaches between concatenation analysis versus coalescent-based species tree analysis. These groups were recovered as a clade in coalescent-based analysis, but Gyrinocheilidae is recovered as sister to Cypriniformes in concatenated analysis. Monophyly of Cobitoidei is supported by both morphological and molecular data (Mayden and Chen 2010; Conway 2011). On the other hand, Conway's (Conway 2011) morphological data for monophyly of Cobitoidei is not decisive, with most of the characters states shared variably by only two out of three of the major clades (Stout et al. *submitted*). Molecular data have previously been more limited and potentially confounded by sampling error. Strong conflict on a phylogenomic scale suggests some limitations with either data or methods in recovering cypriniform relationships. While the data were prepared to exclude paralogs and reduce missing data (Stout et al. *submitted*), other sources of systematic error may still confound phylogenetic analysis and artificially inflate bootstrap support. To address whether sources of systematic error have confounded the reconstruction of cypriniform relationships, removal of potentially biased loci can provide insight into the sensitivity of phylogenomic analyses to these biases. Data exclusion has been suggested as important in studying the relationships of *Paedocypris* with potentially biased molecular data (Britz et al. 2014).

We herein assess phylogenomic analyses on the Stout et al. (*submitted*) dataset for sensitivity to sources of error. For each locus, we quantified a variety of statistics representing various biases that may potentially mislead phylogenetic analysis. We studied the effects of these biases on phylogenomic analysis by analyzing subsets of loci that include less- or more-biased sets of loci, and analyzed both datasets under both concatenation and coalescent analyses. Data subsampling should be a routine method to evaluate robustness in phylogenomic analysis (Edwards et al. 2016). This study thus provides insight into the robustness of phylogenetic inferences relative to sources of bias, and a heuristic method to identify uncertain parts of phylogeny relative to sources of misleading signal. This study also provides clarity to the phylogenetic signal in the anchored phylogenomic dataset for cypriniform relationships and the impacts of biases in the data.

MATERIALS AND METHODS

Data

We reanalyze the data generated by Stout et al. (*submitted*), which was generated under a similar protocol as Prum et al. (2015). This multilocus nuclear dataset was targeted using anchored hybrid enrichment (Lemmon et al. 2012), which utilizes biotinylated-RNA baits designed to hybridize to DNA representing *a priori* selected loci. These selected loci were determined as having few copies based on genome alignments between zebrafish *Danio rerio*, platyfish *Xiphophorus maculatus*, and zebra mbuna cichlid *Metriaclima zebra*. Enriched genes were sequenced paired-end on the Illumina HiSeq platform. Assembly was performed using the known sequence for each locus as a seed sequence to map reads to, from which contigs were extended. For numerous loci, multiple copies were recovered. To exclude paralogs, sequences for each locus were clustered to identify clusters of gene copies, and clusters were selected to maximize taxon inclusion and minimize genetic differences within each locus. Following further data filtering to remove missing data, this resulted in 219 loci for 175 taxa. 172 taxa represent Cypriniformes, while three taxa represent the three other otophysan orders (*Callichthys callichthys*, Siluriformes; *Electrophorus electricus*, Gymnotiformes; *Pygocentrus nattereri*, Characiformes) which together form a monophyletic sister group to Cypriniformes (Characiphysi) (Nelson 2006). The data have an average locus alignment length of 1440 bp, a range in lengths of 192-3111 bp, and a total length of the concatenated alignment of 315,288 base pairs. The proportion of missing data was kept low to reduce the effects that shared missing data can have on phylogenetic reconstruction (Lemmon et al. 2009), and the overall matrix only has 3.476% missing data.

Outline of Phylogenetic Analyses

Analyses proceeded through several steps as outlined in Figure 1. First, we quantified a variety of potential sources of misleading signal in phylogenetic analysis (Table 1). Exploratory data analyses can be performed on alignments prior to phylogenetic tree inference (Morrison 2010). Such analyses allow for probing a dataset for characteristics such as conflict, phylogenetic signal, or sequence biases prior to moving onto further, definitive analyses, and are agnostic to model of evolution (Misof et al. 2014; Kück and Struck 2014). Second, we reconstructed a gene tree for each locus, and quantified additional statistics based on each gene tree; these types of statistics allow assessment of evolutionary characteristics of a locus such as evolutionary rate, or the level of information or conflict a dataset may have relative to a particular gene tree or model of evolution (Aberer et al. 2012; Salichos and Rokas 2013; Struck 2014; Doyle et al. 2015). We focused on four broad classes of sources of phylogenetic noise: base compositional heterogeneity, saturation, branch length effects, and phylogenetic signal, and quantified each of these with multiple bias metrics. From the numerous metrics for sources of phylogenetic noise, multilocus phylogenomic analyses were performed on subsets of data split between the least- and the most-biased halves, for which we quantified additional characteristics of bias and conflict within each of these combined datasets. Phylogenomic analyses were performed on either concatenated sets of loci or as coalescent analysis on sets of gene trees, and results of analyses were compared to determine the level of congruence overall and with respect to particular biases. Analyses were performed on the Auburn University CASIC High Performance Computing Cluster as well as the CIPRES web portal (Miller et al. 2010).

Quantifying Sources of Bias

We performed exploratory data analyses using two different programs, BaCoCa and MARE (Misof et al. 2014; Kück and Struck 2014). We utilized the BaCoCa pipeline to calculate various statistics describing alignment partitions that can introduce systematic biases into phylogenetic analysis. To quantify base compositional heterogeneity, we focused here on GC content, base compositional homogeneity as quantified by the X^2 statistic of a test of homogeneity (Foster 2004), relative composition frequency variability (RCFV; Zhong et al. 2011), proportion of gaps (i.e. missing data), and skews in A vs. T, C vs. G, C vs. T, and A vs. G (Perna and Kocher 1995; Zhong et al. 2011). For the four latter skew metrics, we were most interested in deviation from no skew as a potential source of bias, rather than the particular direction of skew towards any particular base. We assumed each skew metric represented orthogonal vectors and calculated overall skew as the square root of the sum of squares of all four skew metrics. We also used C value (congervence value) calculated by BaCoCa as a metric of saturation.

Second, we calculated tree-likeness for each locus using MARE (matrix reduction; Misof et al. 2014) as a measure of phylogenetic signal. MARE calculates the tree-likeness by assessing the support that each locus has for bifurcations when the dataset is randomly reduced to quartets. When bifurcations are not well-supported within many quartets, that particular gene may have lower phylogenetic signal (Misof et al. 2014). MARE was developed to use the BLOSUM62 amino acid substitution matrix to calculate distances between sequences; we extended this to utilize the DNAfull nucleotide substitution matrix (ftp://ftp.ncbi.nih.gov/blast/matrices/NUC.4.4, Accessed 29 Mar 2016) to calculate distances between nucleotide sequences. The DNAFull matrix scores matches as 5 and mismatches as -4, with ambiguity codes scored as intermediate scores.

We then calculated a variety of additional metrics based on analyses of gene trees. Unlike exploratory data analyses, these analyses follow a phylogenetic analysis under a particular model of evolution on each gene. As in the original phylogenomic analysis presented in Stout et al. (*submitted*), gene tree analyses were performed using the GTR+G model in RAxML. For each gene tree, we performed three methods as implemented in TreSpeX (Struck 2014). First, we calculated long-branch heterogeneity, using the standard deviation of LB scores as a metric. A higher standard deviation of LB scores indicates more variation in branch lengths. Secondly, we calculated the slope and the R^2 value of the linear regression of phylogenetic distance vs. uncorrected p-distance (i.e. a saturation plot) as two additional metrics of saturation; the greater the slope or the higher the R^2 value, the less saturated the data are. We also performed a test for deviation from clock-likeness, as more clock-like genes may be more accurate for phylogenetic inference, as outlined by Doyle et al. (2015). For each gene tree, the likelihoods of GTR+G and GTR+G under a strict clock are fit to the tree using PAUP* 4.0a147 (Swofford 2016), and the likelihood ratio of these models can be used as a metric of deviation from clock-likeness (i.e. deviation from a strict clock model), with larger values indicating a greater deviation (Doyle et al. 2015).

Next, we performed analysis with RogueNaRok as another assessment of phylogenetic signal. RogueNaRok identifies ‘rogue taxa’ that have inconsistent placement between bootstrap replicates, which may be attributed to low phylogenetic signal (Aberer et al. 2012). We recorded the number of rogue taxa for each locus as a metric of phylogenetic signal within each locus. We performed rogue taxon analysis using a maximum dropset size of three, because there is usually no need to exceed that number (Aberer et al. 2012). In addition, we also calculated the evolutionary rate of each locus by summing all branch lengths in each gene tree (Salichos and Rokas 2013) using ape (Paradis et al. 2004).

Finally, we assessed the phylogenetic signal with one other metric by quantifying splits support with spectral analysis using SAMS (Wägele and Mayer 2007). The placement of *Paedocypris* is not well-supported given the presence of other highly-supported splits not in the most likely topologies (Britz et al. 2014). SAMS quantifies the number of sites that support each split and ranks splits by their relative support. Comparison of the most highly-supported (i.e. highly-ranked) splits can provide an indication of how decisive a particular locus is. In data with higher information content, splits that agree with the maximum-likelihood gene tree will have a high number of sites supporting them; in other words, loci with high information content will have many highly-ranked splits that are also found in the gene tree (Wägele and Mayer 2007). On the other hand, in data with lower information content, there will be more conflict, with splits that are not found in the gene tree that are highly-ranked with many sites supporting them. Thus, this is a method that can allow for exploration of the internal consistency of a dataset. To quantify the information content for each locus, we focused on the first ten splits of each locus. Typically, only a few splits will be strongly supported by many sites, with a rapid decline to a background level of support. At this background level, the vast majority of splits in spectral analysis typically disagree with the gene tree, and they are interspersed by only a few weakly supported splits. From these top ten splits within each locus, we computed the average rank of splits found in the gene tree for each locus minus the average rank of conflicting splits, weighted by the number of sites supporting each split, and divided by the total number of sites supporting the top ten splits. This metric could be negative if there were more sites in the top ten splits that disagreed with the ML tree for each locus, or positive if more sites agreed with the ML tree.

Thus, for each locus, we had fourteen metrics describing various qualities that can impact phylogenetic analysis (Table 1). To visualize the variation for each metric, we plotted each bias against locus length, which is a good predictor of sampling error (Betancur-R et al. 2014). To visualize whether certain biases explained similar variation across loci, we performed principal components analysis (PCA) on the correlation matrix. Highly-correlated bias metrics are redundant and downstream analyses would not need to be repeated unnecessarily if a high overlap in variation explained is found. In the PCA, we also included gene length and average bootstrap support (the latter calculated using TreSpEx) because shorter genes generally contain less phylogenetic information (Rasmussen and Kellis 2012; Betancur-R et al. 2014) and may have some relationship to certain sources of bias. To normalize variables for PCA, we used log-transformation; for splits support, because this metric could be negative, we added 1 to transform all values to be positive. We determined if skew was decreased by log-transformation using the skewness command in moments (Komsta and Novomestky 2012), and variables were only log-transformed for the PCA if they reduced skewness (and thus increased normality). To quantify correlation in variation explained between biases, we calculated the Pearson correlation coefficients between normalized variables using the cor.table function implemented in picante (Kembel et al. 2010).

We quantified gene tree heterogeneity by computing Robinson-Foulds (RF) distances using phangorn (Robinson and Foulds 1981; Schliep 2011). RF distance between any two topologies indicates the number of splits that differ between the two unrooted topologies. For computing gene tree heterogeneity across loci, we computed the normalized RF distance because of the variable number of taxa within each locus, which normalizes the RF distance by the number of potential splits.

Lastly, as a coarse exploration of gene tree topologies and how they may be preferentially included in either the less-biased or more-biased subsets, we compared the level of bias and gene tree heterogeneity between groups of gene trees relative to the relationships of *Gyrinocheilus* and *Paedocypris*. First, we identified all gene gene trees that recovered *Gyrinocheilus* sister to Cypriniformes versus gene trees that recover Cobitoidei *sensu lato* as monophyletic. We also identified all gene trees that recovered *Paedocypris* as sister to Cypriniformes versus all gene trees that recovered *Paedocypris* sister to Cyprinoidei. For each comparison, we tested for a difference in mean level of bias and gene tree heterogeneity.

Phylogenetic Analysis

We split the loci into subsets for phylogenomic analysis based on ranking their level of bias. We ranked all loci for the bias metrics from least- to most-biased (Table 1). We split the loci into the top half and the bottom half of base pairs to produce two alignments of relatively equal length for each bias metric (Fig. 2). With 14 bias metrics, this resulted in 28 subsets of the data.

The least-biased and most-biased datasets for a particular bias may not actually differ much if the bias does not vary much across the loci. We used three methods to quantify the variation of data included in each matrix. First, to quantify the difference between the data subsets in the level of bias, we calculated Cohen's d. Cohen's d is typically used as a measure of effect size in power analysis, but is also useful here to quantify the difference in mean bias between dataset relative to the overall variation (i.e. standard deviation) in bias. Second, we calculated mean pairwise normalized RF distances across all loci within each dataset to quantify the heterogeneity of gene trees of loci in each dataset.

We then performed phylogenomic analyses on all data subsets and the complete dataset using both concatenation and coalescent analysis using RAxML v8.2.8 and ASTRAL-II v4.10.2, respectively (Stamatakis 2014; Mirarab and Warnow 2015). Loci were concatenated using AMAS (Borowiec 2016). For the coalescent-based analyses, we used as input the gene trees and their bootstrap replicates for each locus belonging to each dataset as inferred by RAxML under GTR+G for bootstrapping with 100 replicates. Both concatenation and coalescent-based phylogenomic analyses were bootstrapped with 100 replicates.

Summarizing Differences Among Phylogenomic Analyses

We summarized the differences in results across phylogenomic analyses in several ways. First, the RF distance between analyses of the least-biased and most-biased datasets was calculated as a metric of the sensitivity of the analysis to that bias. RF distance has been used as a measure of tree error when comparing a true tree and simulated (Knowles et al. 2012; Mirarab et al. 2014b). When a true tree is not known, however, RF distance is a measure of sensitivity. Lower tree distances should be recovered between different analyses if the topologies are more similar. We also standardized the sensitivity relative to effect size for comparison within least-biased and most-biased datasets for each bias, as datasets that are more different from each other might be expected to result in a larger difference between the topologies inferred.

We then inspected each phylogeny for whether it recovered several focal nodes of interest representing major relationships within Cypriniformes, and extracted their bootstrap support to assess confidence for certain topologies. To cluster trees by similarity, we also generated a UPGMA tree of the different phylogenies using phangorn, given pairwise RF distances across all trees as the distances. A tree of trees, or meta-tree, is a simple method of visualizing clusters of similar tree topologies (Nye 2008). Tips are represented by trees rather than taxa. Note that a meta-tree does not truly have an evolutionary interpretation and does not have polarity (so it is unrooted).

Taxon Removal Experiments

We further tested the sensitivity to long branches on phylogenetic reconstruction of Cobitoidei *sensu lato* for the concatenation analysis. The relationships of *Gyrinocheilus*, Catostomidae, and loaches were recovered with strongly conflicting topologies between concatenation and coalescent-based analyses in previous work (Stout et al. *submitted*). It is possible that this difference is caused by concatenation being more biased to long-branch attraction than coalescent-based methods (Liu et al. 2015a). To test for long-branch attraction with respect to particular branches, taxon removal experiments can be used. If two long branches attract when both branches are included, removing one of the long branches should change the position of the other. We performed four additional phylogenomic analyses on the entire concatenated dataset, but removing the following sets of taxa: 1) all three outgroup taxa, 2) *Gyrinocheilus*, 3) *Paedocypris*, and 4) *Gyrinocheilus* and *Paedocypris*.

RESULTS

Variation Across Loci

Estimation of the characteristics of a gene may be influenced by sampling error (Betancur-R et al. 2014), so we explored the relationship of bias metrics relative to alignment length (Fig. 3). Bias metrics varied in the strength of the relationship with alignment length, varying from no discernible relationship to clear patterns. Measures of base composition did not generally appear to have a relationship with alignment length; of these, only RCFV appears to decrease with higher alignment lengths. Deviation from clock-likeness had a clear pattern with alignment length. Although clock-like genes may be better for phylogenetic reconstruction (Doyle et al. 2015), this may be at the expense of sampling error if shorter genes are found to be more clock-like. Splits support from spectral analysis is highly variable with shorter loci and is relatively invariable once a certain alignment length is achieved among loci. This may have to do with relatively few sites to support splits in shorter loci. Number of rogue taxa decreases with alignment length also, again potentially an indication of reduced sampling error in longer loci.

Proportion of gaps is relatively invariable across loci, which is not surprising given missing data was minimized in these alignments (Stout et al. *submitted*).

A PCA displaying variation across loci in bias metrics demonstrates that there is little consistency across the metrics for explaining variation across loci (Fig. 4), given the variation in directions of the vectors for each bias metric. The difference in variation explained in bias metrics across loci is consistent with the lack of clear division between least-biased and most-biased loci across all bias metrics (black vs. white, Fig. 2). In addition, most of the bias metrics do not explain similar variation across loci as average bootstrap support within each locus and alignment length. Locus length and average bootstrap support explain similar variation across loci, consistent with previous work demonstrating a decrease in sampling error with locus length (Betancur-R et al. 2014).

Variation in bias across metrics for loci was corroborated by correlation coefficients (Table 2). Out of 91 correlation coefficients derived from pairwise comparisons between bias metrics, the only pairwise bias metric comparisons with a correlation coefficient greater in magnitude than 0.5 were the following: GC content and overall base skew (r = 0.557), RCFV and X^2 value of base heterogeneity (r = 0.667), proportion of gaps and tree-likeness (r = -0.503), RCFV and tree-likeness (-0.539), the saturation plot slope and saturation plot R^2 values (r = 0.707), proportion of gaps and saturation plot R^2 value (r = -0.519), and the evolutionary rate with X^2 value of base heterogeneity (r = 0.551), with C value (r = 0.539), with tree-likeness (r = -0.711), and with saturation plot slope (r = -0.646). A fairly strong relationship was recovered for average bootstrap support and alignment length (r = 0.756). Some of the bias metrics had a correlation coefficient with either alignment length or average bootstrap support greater than 0.5, including alignment length and RCFV value (r = -0.575), average bootstrap support and C value (r = -0.556), alignment length and average bootstrap support with deviation from clock-likeness (r = .754 and r = .602, respectively), and average bootstrap support with number of rogue taxa (r = -0.818). Almost none of the correlation coefficients were above 0.8, indicating none of the correlations were very strong, which is consistent with the directions of vectors in the PCA. The correlation between average bootstrap support and number of rogue taxa is easily explained because bootstrap support is explicitly used to identify rogue taxa.

All gene trees had different topologies, and gene tree heterogeneity was high, with a mean normalized RF distance of 0.475 (i.e. on average, datasets differed in 47.5% of their splits) and had a range from 0.282 to 0.733 (Fig. 5). Average bootstrap support for gene trees ranged from 41.02% to 81.85%, with a median average bootstrap support of 70.6%.

We then quantified bias and gene tree heterogeneity for sets of gene trees relative to the recovered placement of *Gyrinocheilus* and *Paedocypris* (Table 3). We found 95 genes recovered a relationship of *Gyrinocheilus* sister to Cypriniformes, while 64 genes recovered *Gyrinocheilus*, Catostomidae, and loaches as monophyletic. These sets of gene trees significantly differed in their alignment length, average bootstrap support, RCFV value, proportion of gaps, saturation plot slope, evolutionary rate, and gene tree heterogeneity. We found 52 genes supported *Paedocypris* as sister to Cypriniformes and 102 genes supported *Paedocypris* as sister to Cyprinoidei. These sets of gene trees significantly differed in their alignment length, average bootstrap support, RCFV value, deviation from clock-likeness, number of rogue taxa, saturation plot slope, saturation plot R^2, and evolutionary rate.

Differences Between Datasets

Because the various biases had different levels of variation (Fig. 2), splitting loci into least-biased and most-biased subsets resulted in some datasets that were relatively different for their bias, while other datasets were more similar, as quantified by effect size (Fig. 6). The decreasing effect size is apparent as a useful metric in quantifying the difference between datasets given the decreasing distance between mean values of bias between the least-biased and most-biased datasets.

With respect to gene tree heterogeneity within each dataset (Fig. 5), although a significant difference in mean normalized RF distances was found between all pairs of least-biased and most-biased datasets (t-test, maximum p-value ≤ .0000153) for 13/14 comparisons) except for overall base skew (p = .226). Visual inspection shows a high degree of overlap in ranges and thus a low effect size (Fig. 5).

Sensitivity of Phylogenomic Analyses

Identical topologies were never recovered between the least- and most-biased datasets, demonstrating some sensitivity to data subsetting and to the difference in level of bias between datasets (Fig. 7). When controlling for the relative difference between datasets in bias – by dividing the RF-distance between trees by effect size to get a measure of relative sensitivity – the variation in sensitivity is not greatly changed, although some biases have greater relative effect than their absolute effect (Fig. 7).

Overall, the largest differences in topologies were between the least- and most-biased datasets for percentage of gaps in the concatenation analyses, which is surprising given the already low level of missing data across loci (never exceeding 7%), but underscores its large effect in misleading phylogenetic analysis (Lemmon et al. 2009). In both concatenation and coalescent-based analyses, relatively large differences in topologies were found between the least- and most-biased datasets for RCFV, demonstrating the importance of base compositional heterogeneity and the effectiveness of this metric in quantifying this bias (Zhong et al. 2011).

Analyses were sensitive to long-branch heterogeneity (i.e. LB score standard deviation), and this effect is magnified when quantifying sensitivity relative to effect size. Thus, even relatively small differences in long-branch heterogeneity lead to relatively large topological differences. Analyses differing in the level of saturation by C value usually had a larger difference than analyses differing in saturation as quantified by slope of saturation plots, which always had a larger difference than analyses differing in saturation as quantified by R^2 of saturation plots. This ranking of saturation metrics suggests analyses are far more sensitive to C value than they are to other metrics of saturation, despite attempting to quantify the same phenomenon. Analyses also were fairly sensitive to datasets that differed in splits support. Dividing datasets by base compositional heterogeneity (as quantified by X^2) and overall base skew had relatively little influence on the RF distance.

The sensitivity of analyses to certain biases differed between concatenation and coalescent-based analyses. Concatenation analyses were more sensitive to deviations from clock-likeness and tree-likeness in genes than coalescent-based analysis.

Coalescent-based analyses were more sensitive to differences in the number of rogue taxa and evolutionary rate than concatenation.

When clustering phylogenomic trees by similarity into a meta-tree using pairwise RF-distances between trees (Fig. 8), two main clusters were recovered representing a division between topologies recovered in concatenated analyses and coalescent analyses, demonstrating consistent topological differences between these methods are not greatly lessened by removal of bias. As apparent from the branch lengths in the meta-tree, concatenated analyses are more similar to each other than coalescent analyses. On average, 13.5 more differing splits are found between coalescent-based analyses than between concatenation analyses ($p < .0001$). There is no distinct clustering of less-biased datasets across metrics exclusive of more-biased datasets, indicating less-biased datasets are not absolutely more similar than more-biased datasets. On the other hand, on average, less-biased datasets differ by 2.1 splits less than more-biased datasets ($p < .0001$), consistent with the expectation that the least-biased datasets are converging towards a well-supported topology.

Congruence and Conflict Among Phylogenomic Trees

The Cobitoidei differ in their phylogenetic relationships between concatenated and coalescent analyses (Fig. 8). In the concatenated phylogenies, Cobitoidei *sensu lato* is not monophyletic, with Gyrinocheilidae recovered as sister to the remaining Cypriniformes in analyses from all datasets regardless of level of bias. By contrast, the coalescent-based analyses on most datasets support monophyly of Cobitoidei *sensu lato*, except for four analyses on datasets with higher splits supports, slower genes, lowest saturation (measured by slope), and lower RCFV. On the other hand, in these coalescent-based analyses where Gyrinocheilidae is found sister to Cypriniformes, the sister relationship between Catostomidae and loaches are not well supported, and thus their relationships are ambiguous relative to *Gyrinocheilus* and Cyprinoidei.

All phylogenomic analyses recovered *Paedocypris* as the sister taxon to the Cyprinoidei. The lowest support this relationship ever reaches is 80% bootstrap support for the tree from the dataset with the fastest evolving genes analyzed using concatenation. *Sundadanio* is consistently recovered as the sister group to a clade formed by Tanichthyidae, Gobionidae, Xenocyprididae, and Leuciscidae across all analyses. *Danionella* is consistently recovered within the Danionidae.

Danionidae was fairly consistently recovered as the sister group to a large clade of cyprinoids including *Sundadanio*, Xenocyprididae, Gobionidae, Acheilognathidae, Tanichthyidae, and Leuciscidae, but this relationship is erratically supported across concatenation analyses and is poorly supported in most coalescent-based analyses. In seven of the ten analyses where this relationship has 100% bootstrap support, this relationship was found on analyses of datasets with more bias (e.g. lower splits support, more gaps, higher saturation). On occasion, Cyprinidae is found as sister to the large clade formed by *Sundadanio*, Xenocyprididae, Gobionidae, Acheilognathidae, Tanichthyidae, and Leuscidae with low to medium support. In one analysis, the concatenation analysis on the least-biased data for base skew, Danionidae and Cyprinidae are instead found as sister groups with weak support.

Analyses were fairly congruent in recovering the relative relationships between the major cyprinoid families Tanichthyidae, Gobionidae, Xenocyprididae, and

Leuciscidae (Fig. 8). These relationships have been robustly supported across numerous phylogenetic studies of Cypriniformes (Chen and Mayden 2009; Mayden and Chen 2010 Stout et al. submitted). *Tanichthys* is reconstructed as the sister group to Leuciscidae across all analyses. Gobionidae and Acheilognathidae are usually recovered as sister taxa, but, occasionally, Gobionidae is found as sister to Leuciscidae and *Tanichthys*, and in one analysis Acheilognathidae is found as sister to Leuciscidae and *Tanichthys*. These alternative relationships are only found in coalescent-based analyses on more-biased datasets.

Taxon Removal Experiments

We observed that the unrooted topology among all trees is identical for the relative relationships of *Gyrinocheilus*, Catostomidae, loaches, *Paedocypris*, and Cyprinoidei, where Catostomidae with *Gyrinocheilus* form one end of the split and loaches with Cyprinoidei form the other end of the split (Fig. 9). What differs between the unrooted topologies is the relative placement of the outgroup taxa. Strong conflict between the concatenation and coalescent-based analyses can potentially be explained by concatenation being potentially more prone to long-branch attraction than coalescent-based analysis.

When we removed the outgroup, the identical unrooted topology as the complete analysis is recovered (Fig. 9), demonstrating the placements of *Gyrinocheilus* and *Paedocypris* are not pulled towards the outgroup, consistent with the observation that all unrooted topologies for both concatenation and coalescent-based analyses are identical on the placements of these taxa.

When we removed only *Gyrinocheilus*, Catostomidae and loaches are recovered as sister taxa, and *Paedocypris* is recovered as sister to Cypriniformes. This perturbed topology is not recovered in any of the previous phylogenomic analyses based on complete taxon sampling, and demonstrates the outgroup and *Paedocypris* attract towards each other when *Gyrinocheilus* is not present. This is likely an effect of long-branch attraction. The recovery of Catostomidae and loaches as a clade is consistent with the coalescent-based analyses given that *Gyrinocheilus* is absent. When we removed only *Paedocypris*, we again recover the topology where *Gyrinocheilus* is sister to Cypriniformes, as in the full analysis.

Finally, when we removed *Paedocypris* and *Gyrinocheilus*, we recovered Catostomidae and loaches as sister taxa, and this clade sister to Cyprinoidei.

DISCUSSION

Sensitivity in Phylogenomics

This study reiterates the need for careful assessment of data to understand potential sources of systematic error and their effects in phylogenomic datasets (Jeffroy et al. 2006; Philippe et al. 2011). Data subsetting should be a routine step in evaluation of the robustness of phylogenomic data (Edwards et al. 2016). In non-parametric bootstrapping, consistent recovery of a particular topology indicates robustness to random signal in the data, but may be biased by non-random systematic error; thus, in the same spirit, consistent recovery of a topology across analyses based on different subsets of data, particularly subsets that differ in their level of systematic error, provides an exploration of the effect of various biases on recovered relationships. This data exploration is heuristic (Grant and Kluge 2003) because it allows for identifying unstable clades of interest deserving of further study that would not be identified by overconfident bootstrap values due to systematic error. Though these analyses are a biased sampling across the potential parameter space of bias variation, and therefore do not provide objective measures of support due to the subjective selection of biases assessed, they nevertheless provide insight into the sensitivity of the topology where non-parametric bootstrapping is biased towards invariant results. Testing the sensitivity of phylogenetic analysis to a variety of sources of noise can provide empirical evidence for the effect of a particular source of misleading signal that can guide locus exclusion in future phylogenetic studies.

Although data quality is important, many of these biases did not appear to have a large effect on the relationships among major clades of Cypriniformes. Significant biases discovered may not necessarily mislead phylogenetic analyses (Brown 2014; Doyle et al. 2015). The relative importance of excluding loci based on certain characteristics can vary across different datasets, and thus data should be assessed on a case-by-case basis (Kück and Struck 2014). While adjusting for biases can result in topological changes within concatenation vs. coalescent-based analyses, major analytical and theoretical differences exist between these methods that lead to a large difference between topologies recovered by each method. Coalescent-based analyses reconstruct phylogenies accommodating gene tree heterogeneity, thus how the models handle gene tree heterogeneity may have a larger effect on the difference in phylogenomic reconstruction than simply accounting for biased loci (Salichos and Rokas 2013; Liu et al. 2015b). Gene tree heterogeneity in our study was relatively similar across datasets (Fig. 5) even if levels of bias were relatively different (Fig. 6).

Data subsampling analyses have demonstrated that greater variation between topologies is found between concatenation analyses than coalescent-based analyses, with erratic and complete support for completely conflicting topologies between subsets in concatenation (Edwards 2009; Song et al. 2012; Edwards et al. 2016). Song et al. (2012) assessed sensitivity empirically by finding erratic bootstrap support for the topology of two focal relationships within mammals, while coalescent methods merely report low support. These nodes are particularly interesting because they are the most difficult to reconstruct in mammal phylogeny owing to incredibly short internodes. It is at these types of repeated, rapid branching events where the anomaly zone develops, and where coalescent methods demonstrably perform better than concatenation (Edwards 2009). Here, we find that trees based on concatenation are generally more similar than coalescent-based analyses measured by RF distance. This demonstrates that coalescent-based analyses may actually be more sensitive to data subsamples than concatenation analyses when studying the number of variable bipartitions across analyses. The higher precision in concatenation here is consistent with a low level of incomplete lineage sorting, where concatenation has higher accuracy relative to coalescent-based methods as measured by tree distances (Mirarab et al. 2014b). If we approach sensitivity by studying variation in bootstrap support across data subsets at certain focal nodes, bootstrap support consistency is similar to what Song et al. (2012) would predict. One example where we find erratic bootstrap support across data subsets in concatenation analyses is for the position of Danionidae relative to Sundadanionidae, Xenocyprididae, Gobionidae, Acheilognathidae, Tanichthyidae, and Leuciscidae. The coalescent-based analyses have fairly consistently low support, consistent with previous findings on the behavior of coalescent-based methods (Edwards 2009; Song et al. 2012). The rapid radiation of the Cyprinoidei into multiple clades may mean that it is a group where we do not here have enough data to resolve phylogeny using the two-step coalescent methods, which may require hundreds more loci (Song et al. 2012). In general, relationships among Xenocyprididae, Gobionidae, Acheilognathidae, Tanichthyidae, and Leuciscidae are less well-supported than in concatenation, however the recovered relationships are generally congruent with concatenation. On the other hand, concatenation and coalescent-based analyses differ in the relationships of *Gyrinocheilus* with other cypriniformes. *Gyrinocheilus* is consistently placed as sister to Cypriniformes across concatenation analyses. In coalescent-based analyses, however, the alternative relationships of *Gyrinocheilus* are both strongly supported, rather than weakly supported. This contrasts with the prediction that coalescent analyses across data subsets usually do not recover strong support for conflicting topologies, and also demonstrates an instance where data subsets demonstrated more erratic behavior in coalescent-based analysis relative to concatenation relative to reduction in some types of bias.

Paedocypris

Despite the vulnerability of *Paedocypris* to long-branch attraction (Britz et al. 2014), the placement of *Paedocypris* is insensitive to the biases as we treated them here. Presence of long branches alone does not actually confound concatenation, and is made problematic by short internodes connecting long branches in the presence of incomplete lineage sorting (Liu et al. 2015a). There is high consistency in the placement of the long-branched taxon *Danionella* among the Danionidae, not near the base nor near other long branches, even with a single mitochondrial gene (Rüber et al. 2007), which empirically demonstrates long branches do not always attract. We never recover the paedomorphic taxa as closely-related as suggested by morphological data of Britz & Conway (2009) and Britz et al. (2014). On the other hand, the morphological data available are not decisive on the relationships among the remaining cyprinoids; morphological phylogenetic analyses including paedomorphic taxa reconstruct most of the cyprinoids as a large polytomy, and are thus unable to distinguish the character states of paedomorphic cyprinoids as synapomorphy or homoplasy (Britz et al. 2014 Stout et al. submitted). Rüber et al. (2007) reconstructed a close relationship of paedomorphic taxa using cytochrome b, but these data have low phylogenetic signal and are fairly indecisive (Britz et al. 2014), and this is not supported by whole mitogenome analysis (Mayden and Chen 2010). Tang et al. (Tang et al. 2010; 2013) reconstructed a close relationship between paedomorphic taxa, but bootstrap supports were relatively low. In addition, re-analysis of these data does not yield the published topology, instead placing *Paedocypris* as sister to Cypriniformes and *Sundadanio* as closely related to *Leptobarbus* (Tan & Armbruster *in prep.*; Chapter 1).

Mayden & Chen (2010) recovered *Paedocypris* as the sister-group to Cypriniformes. The hypothesis of *Paedocypris* as being sister to Cypriniformes is consistent with a case of long-branch attraction (Britz et al. 2014). This result is not found with any of our phylogenomic analyses where we adjusted for sequence bias, however we did recover this relationship when *Gyrinocheilus* was removed. This highlights the importance of taxon sampling in correcting long-branch attraction (Wägele & Mayer 2007), even for phylogenomic analyses. However, Mayden & Chen (2010) recovered *Paedocypris* as sister to Cypriniformes with the inclusion of two *Gyrinocheilus* species. It is possible that the loci included in Mayden & Chen's (2010) analysis, particularly EGR3 as suggested by Britz et al. (2014), is particularly biased, even with important taxa represented to break long branches.

Mayden & Chen (2010) mention recovering *Paedocypris* as sister to Cyprinoidei in an unpublished mitogenome analysis. Given that fish mitochondrial genomes have thirteen protein-coding genes (Iwasaki et al. 2013), more than the six nuclear genes used in Mayden & Chen (2010), there may be more phylogenetic signal simply because of the longer alignment to reduce sampling error. This is interesting because mitochondrial DNA is well-known to evolve quite rapidly, and suggests that neither the high evolutionary rates nor saturation in those data lead to a relationship incongruent with nuclear phylogenomic data for major cypriniform relationships. The only gene in Mayden & Chen (2010) that decisively supported *Paedocypris* as sister to Cypriniformes is EGR3 (Britz et al. 2014), which suggests this gene may be particularly biased. Reanalysis of the Mayden & Chen (2010) matrix without EGR3 recovers *Paedocypris* as sister to cyprinoids, albeit with low bootstrap support (Britz et al. 2014). Thus, a signal of a relationship to cyprinoids was evident even prior to phylogenomic analysis by Stout et al. (*submitted*).

Cobitoidei

The placement of Gyrinocheilidae and the monophyly of the clade formed by Gyrinocheilidae, Catostomidae, and loaches has implications for the monophyly of Cobitoidei *sensu lato* and the evolution of Cypriniformes (Mayden and Chen 2010). The strongly differing placement of Gyrinocheilidae between the coalescent-based and concatenation analyses is cause for concern. These relationships may be difficult to reconstruct. It is possible that these branches are in the anomaly zone, because *Gyrinocheilus* as sister to Cypriniformes is the most prevalent gene tree and is not the recovered relationship of the species tree. In the anomaly zone, the most common gene tree will positively mislead concatenation analyses but not coalescent-based analyses (Edwards 2009).

It is also possible there is an effect of long-branch attraction. Different from the expectation that *Gyrinocheilus* may be pulled towards the outgroup by its long branch, we found that the unrooted topology representing the relative placements of *Gyrinocheilus*, Catostomidae, loaches, *Paedocypris*, and Cyprinoidei are consistent across all of our phylogenomic analyses, regardless if taxa or biases are removed. We found it was actually the outgroup that varied in placement between phylogenomic analyses, and it is influenced by long-branch attraction, as demonstrated by an attraction towards *Paedocypris* when *Gyrinocheilus* was removed. Also, when both *Paedocypris* and *Gyrinocheilus* are removed, the outgroup placement results in Cyprinoidei sister to a clade formed by Catostomidae and loaches, suggesting Cyprinoidei sister to Cobitoidei.

To better address long-branch attraction, taxon sampling is more important than locus sampling, as increased locus sampling only serves to continue sampling a biased lineage (Wägele and Mayer 2007). It is best to include other species that may help break the long branch. Unfortunately, both Gyrinocheilidae and Paedocyprididae consist of only one genus and three closely-related species (Roberts and Kottelat 1993; Kottelat et al. 2006; Britz and Kottelat 2008). Both are clearly distinct lineages from other Cypriniformes, so no extant taxa along these branches exists. Long-branch attraction can also be addressed by sampling an outgroup with a more recent common ancestor, to shorten the long branch of the outgroup (Wägele and Mayer 2007). The other otophysan taxa form a monophyletic group, Characiphysi, that is the closest sister group to Cypriniformes (Nelson 2006; Betancur-R et al. 2013), and thus increased sampling of additional otophysans will continue sampling lineages with the same most recent common ancestor with Cypriniformes. Given the extremely long branches in the outgroup, phylogenomic analysis of the relationships of Cobitoidei may be improved by breaking long branches in the outgroup. Alternatively, using a more distant outgroup may be valid if a more distant outgroup has shorter branch lengths than a closer outgroup with long branch lengths, as the long branches can cause long-branch attraction effects (Takezaki and Nishihara 2016).

Coalescent methods have been demonstrated to be more robust to effects of long-branch attraction (Liu et al. 2015a), so this may indicate that monophyletic Cobitoidei *sensu lato* may be the less-biased topology. Monophyly of Cobitoidei would be consistent with other analyses on morphological data and molecular data (Mayden and Chen 2010; Conway 2011). Interestingly, a few occasions of bias reduction in gene trees led to a result where *Gyrinocheilus* was sister to Cypriniformes, and thus congruent with the concatenation analyses. One of these datasets was on the slowest half of genes, while another was on the least saturated genes (by slope of saturation plot). Conserved genes with low saturation are effective for reconstructing ancient divergences (Betancur-R et al. 2014). In previous work on land plants, fast-evolving sites were found to more strongly bias concatenation than coalescent-based methods for reconstructing the relationships of plants near the base of different clades (Xi et al. 2013; 2014; Edwards et al. 2016); however, in our study, concatenation was not sensitive to difference in evolutionary rate for the position *Gyrinocheilus*, but coalescent-based analyses were. ASTRAL may be sensitive to saturation (Edwards et al. 2016), which also suggests that removing more saturated genes improves the coalescent analysis. Another of the coalescent-based analyses that supported *Gyrinocheilus* as sister to Cypriniformes was on the dataset optimized for splits support, a measure of data decisiveness, suggesting that the placement of *Gyrinocheilus* is supported in these analyses by more decisive gene trees. Accurate gene trees are necessary for accurate species tree reconstructions, and reducing the effect of low phylogenetic signal within genes can improve species tree analysis (Mirarab et al. 2014a). Weak genes can confound phylogenomic inference in coalescent-based analysis (Liu et al. 2015b). By contrast, the removal of two long branches, *Paedocypris* and *Gyrinocheilus*, results in a topology in concatenated analysis similar to that seen in coalescent-based analysis, with the remaining cobitoid clades (Catostomidae and loaches) forming a monophyletic group sister to Cyprinoidei. This suggests that taxon sampling with respect to long branches is also extremely important in affecting the placement of the outgroup taxa included here.

In summary, the taxon removal experiments suggest that concatenation is biased by long-branch attraction of the outgroup towards *Gyrinocheilus*, however the removal of biases suggests that monophyly of Cobitoidei *sensu lato* across most coalescent-based analysis is biased by low phylogenetic signal, saturation, and quickly-evolving genes in gene trees. Caution should be placed on too heavily interpreting the coalescent-based analyses because the data subsets are relatively small. Two-step coalescent-based analyses may actually need hundreds more genes to accurately reconstruct relationships than used in our subsets, although usually lack of strong signal presents itself as low support rather than high but contradictory support across subsets (Song et al. 2012; Mirarab et al. 2014a), such as in the relationship of Danionidae to other cyprinoid clades. When presence of ILS is low, as in when tree topologies are more similar between concatenation analyses than they are between coalescent-based analyses (Mirarab et al. 2014b), and when there are relatively few genes (i.e. <1000 genes), concatenation is more likely to reconstruct the correct tree even in the presence of long branches than coalescent-based analyses (Liu et al. 2015a). Even given a phylogenomic scale dataset, data limitation may be a problem for this comparison. Future work should explore the effect of increased taxon sampling in the outgroup as well as increased locus sampling overall in evaluating the relationships of Cobitoidei *sensu lato*.

CONCLUSION

Use of multiple loci for phylogenomic reconstruction has the benefit of reducing sampling error, but the other benefit of the release from data limitation is the ability to select subsets of loci that are less biased. This allows identifying the most unbiased evolutionary relationships a dataset may support, and study the sensitivity of phylogenetic analysis to sources of misleading signal. Phylogenomics is not simply the application of the same principles as decades of molecular phylogenetics, but has transformed the field by introducing new principles due to its broader scope and the variability in how genes evolve across the genome. Evaluating the data thus provides insight into how rapidly growing data may help elucidate or obfuscate the branches of the Tree of Life.

ACKNOWLEDGEMENTS

We thank ED Burress, MM Hayes, and NV Whelan for involvement in discussions that improved this manuscript. We would like to thank Bradley Morgan for assistance with maintenance of the Auburn University CASIC high-performance computing cluster. We thank Auburn University Cellular & Molecular Biosciences fellowship for their generous support of MT which allowed this work to be completed. We thank the National Science Foundation for their generous support of the All Cypriniformes Species Inventory (DEB-1023403). Thanks to you, the reader, if you made it through this entire dense chapter, and I owe you a beverage.

REFERENCES

Aberer A.J., Krompass D., Stamatakis A. 2012. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. Syst Biol. 62:162–166.

Betancur-R R., Broughton R.E., Wiley E.O., Carpenter K., Andrés López J., Li C., Holcroft N.I., Arcila D., Sanciangco M.D., Cureton J.C. II, Zhang F., Buser T., Campbell M.A., Ballestros J.A., Roa-Varon A., Willis S.C., Borden W.C., Rowley T., Reneau P.C., Hough D.J., Lu G., Grande T., Arratia G., Ortí G. 2013. The Tree of Life and a New Classification of Bony Fishes. PLOS Currents Tree of Life.:1–45.

Betancur-R R., Naylor G.J.P., Ortí G. 2014. Conserved Genes, Sampling Error, and Phylogenomic Inference. Syst Biol. 63:257–262.

Betancur-R R., Wiley E.O., Bailly N., Miya M., Lecointre G., Ortí G. Phylogenetic Classification of Bony Fishes --Version 3. Available from http://www.deepfin.org/Classification_v3.htm.

Borowiec M.L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ. 4:e1660–10.

Borowiec M.L., Lee E.K., Chiu J.C., Plachetzki D.C. 2016. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics.:1–15.

Britz R., Conway K.W. 2009. Osteology of *Paedocypris*, a miniature and highly developmentally truncated fish (Teleostei: Ostariophysi: Cyprinidae). J. Morphol. 270:389–412.

Britz R., Conway K.W., Rüber L. 2014. Miniatures, morphology and molecules: *Paedocypris* and its phylogenetic position (Teleostei, Cypriniformes). Zool J Linn Soc. 172:556–615.

Britz R., Kottelat M. 2008. Paedocypris carbunculus, a new species of miniature fish from Borneo (Teleostei: Cypriniformes: Cyprinidae). Raffles B Zool. 56:415–422.

Brown J.M. 2014. Detection of Implausible Phylogenetic Inferences Using Posterior Predictive Assessment of Model Fit. Syst Biol. 63:334–348.

Bufalino A.P., Mayden R.L. 2010. Molecular phylogenetics of North American phoxinins (Actinopterygii: Cypriniformes: Leuciscidae) based on RAG1 and S7 nuclear DNA sequence data. Mol Phylogenet Evol. 55:274–283.

Chang C.H., Li F., Shao K.T., Lin Y.-S., Morosawa T., Kim S., Koo H., Kim W., Lee J.-S., He S., Smith C., Reichard M., Miya M., Sado T., Uehara K., Lavoué S., Chen W.J., Mayden R.L. 2014. Phylogenetic relationships of Acheilognathidae (Cypriniformes: Cyprinoidea) as revealed from evidence of both nuclear and mitochondrial gene sequence variation: Evidence for necessary taxonomic revision in the family and the identification of cryptic species. Mol Phylogenet Evol. 81:182–194.

Chen W.J., Mayden R.L. 2009. Molecular systematics of the Cyprinoidea (Teleostei: Cypriniformes), the world's largest clade of freshwater fishes: Further evidence from six nuclear genes. Mol Phylogenet Evol. 52:544–549.

Conway K.W. 2011. Osteology of the South Asian Genus Psilorhynchus McClelland, 1839 (Teleostei: Ostariophysi: Psilorhynchidae), with investigation of its phylogenetic relationships within the order Cypriniformes. Zool J Linn Soc.:no–no.

Degnan J.H., Rosenberg N.A. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. PLoS Genet. 2:e68.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution. 24:332–340.

Doyle V.P., Young R.E., Naylor G.J.P., Brown J.M. 2015. Can We Identify Genes with Increased Phylogenetic Reliability? Syst Biol. 64:824–837.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S., Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol Phylogenet Evol. 94:447–462.

Eschmeyer W.N., Fricke R., van der Laan R. 2016. Catalog of Fishes: Genera, Species, References.

Foster P.G. 2004. Modeling compositional heterogeneity. Syst Biol. 53:485–495.

Grant T., Kluge A.G. 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. Cladistics. 19:379–418.

Iwasaki W., Fukunaga T., Isagozawa R., Yamada K., Maeda Y., Satoh T.P., Sado T., Mabuchi K., Takeshima H., Miya M., Nishida M. 2013. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol Biol Evol. 30:2531–2540.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends in Genetics. 22:225–231.

Kembel S.W., Cowan P.D., Helmus M.R., Cornwell W.K., Morlon H., Ackerly D.D., Blomberg S.P., Webb C.O. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 26:1463–1464.

Knowles L.L., Lanier H.C., Klimov P.B., He Q. 2012. Full modeling versus summarizing gene-tree uncertainty: Method choice and species-tree accuracy. Mol Phylogenet Evol. 65:501–509.

Kocot K.M., Citarella M.R., Moroz L.L., Halanych K.M. 2013. phyloTreepruner: A phylogenetic Tree-Based Approach for selection of Orthologous sequences for phylogenomics. Evolutionary Bioinformatics Online. 2013:429–435.

Komsta L., Novomestky F. 2012. moments: Moments, cumulants, skewness, kurtosis and related tests.

Kottelat M. 2012. Conspectus Cobitidum: an Inventory of the Loaches of the World (Teleostei: Cypriniformes: Cobitoidei). Raffles B Zool. 26:1–199.

Kottelat M., Britz R., Tan H.H., Witte K.-E. 2006. *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. Proc. R. Soc. London Ser. B. 273:895–899.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. Syst Biol. 56:17–24.

Kück P., Struck T.H. 2014. BaCoCa--a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Mol Phylogenet Evol. 70:94–98.

Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference. Syst Biol. 58:130–145.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Syst Biol. 61:727–744.

Lemmon E.M., Lemmon A.R. 2012. High-Throughput Genomic Data in Systematics and Phylogenetics. Annual review of Ecology, Evolution, and Systematics. 44:99–121.

Liu L., Xi Z., Davis C.C. 2015a. Coalescent Methods Are Robust to the Simultaneous Effects of Long Branches and Incomplete Lineage Sorting. Mol Biol Evol. 32:791–805.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015b. Estimating phylogenetic trees from genome-scale data. Annals of the New York Academy of Sciences. 1360:36–53.

Mayden R.L., Chen W.J. 2010. The world"s smallest vertebrate species of the genus *Paedocypris*: a new family of freshwater fishes and the sister group to the world"s most diverse clade of freshwater fishes (Teleostei: Cypriniformes). Mol Phylogenet Evol. 57:152–175.

Miller M.A., Pfeiffer W., Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Computer Environment Workshop (GCE).:1–8.

Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science. 346:1250463–1250463.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014b. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 30:i541–i548.

Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 31:i44–i52.

Misof B., Meusemann K., Reumont von B.M. 2014. *A priori* assessment of data quality in molecular phylogenetics. Algorithms Mol ….

Morrison D.A. 2010. Using data-display networks for exploratory data analysis in phylogenetic studies. Mol Biol Evol. 27:1044–1057.

Nelson J.S. 2006. Fishes of the World. New York: John Wiley & Sons.

Nye T. 2008. Trees of Trees: An Approach to Comparing Multiple Alternative Phylogenies. Syst Biol. 57:785–794.

Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 20:289–290.

Perna N.T., Kocher T.D. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J Mol Evol. 41:353–358.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biology. 9:e1000602.

Philippe H., Delsuc F., Brinkmann H., Lartillot N. 2005. Phylogenomics. Annual review of Ecology, Evolution, and Systematics. 36:541–562.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.

Rasmussen M.D., Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res. 22:755–765.

Roberts T.R., Kottelat M. 1993. Revision of the Southeast Asian freshwater fish family Gyrinocheilidae. Ichthyol Explor Freshwaters. 4:375–383.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math Biosci. 53:131–147.

Rüber L., Kottelat M., Tan H.H., Ng P.K.L., Britz R. 2007. Evolution of miniaturization and the phylogenetic position of *Paedocypris*, comprising the world's smallest vertebrate. BMC Evol Biol. 7:38.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327–331.

Schliep K.P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics. 27:592–593.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci USA. 109:14942–14947.

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol Phylogenet Evol. 94:1–33.

Stamatakis A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics. 30:1312–1313.

Struck T.H. 2014. TreSpEx-Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. Evolutionary Bioinformatics Online. 10:51–67.

Swofford D.L. 2016. Phylogenetic Analysis Under Parsimony (PAUP*), version 4.0a147. Available from http://paup.csit.fsu.edu/.

Takezaki N., Nishihara H. 2016. Resolving the Phylogenetic Position of Coelacanth: The Closest Relative Is Not Always the Most Appropriate Outgroup. Genome Biology and Evolution. 8:1208–1221.

Tang K.L., Agnew M.K., Chen W.J., Hirt M.V., Raley M.E., Sado T., Schneider L.M., Yang L., Bart H.L. Jr, He S., Liu H.-Z., Miya M., Saitoh K., Simons A.M., Wood R.M., Mayden R.L. 2011. Phylogeny of the gudgeons (Teleostei: Cyprinidae: Gobioninae). Mol Phylogenet Evol. 61:103–124.

Tang K.L., Agnew M.K., Hirt M.V., Lumbantobing D.N., Raley M.E., Sado T., Teoh V.-H., Yang L., Bart H.L. Jr, Harris P.M., He S., Miya M., Saitoh K., Simons A.M., Wood R.M., Mayden R.L. 2013. Limits and phylogenetic relationships of East Asian fishes in the subfamily Oxygastrinae (Teleostei: Cypriniformes: Cyprinidae). Zootaxa. 3681:101–135.

Tang K.L., Agnew M.K., Hirt M.V., Sado T., Schneider L.M., Freyhof J., Sulaiman Z., Swartz E., Vidthayanon C., Miya M., Saitoh K., Simons A.M., Wood R.M., Mayden R.L. 2010. Systematics of the subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). Mol Phylogenet Evol. 57:189–214.

Wägele J., Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol Biol. 7:147.

Whelan N.V., Kocot K.M., Moroz L.L., Halanych K.M. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci USA. 112:5773–5778.

Xi Z., Liu L., Rest J.S., Davis C.C. 2014. Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. Syst Biol. 63:919–932.

Xi Z., Rest J.S., Davis C.C. 2013. Phylogenomics and Coalescent Analyses Resolve Extant Seed Plant Relationships. PLOS ONE. 8:e80870–11.

Yang L., Sado T., Hirt M.V., Pasco-Viel E., Arunachalam M., Li J., Wang X., Freyhof J., Saitoh K., Simons A.M., Miya M., He S., Mayden R.L. 2015. Phylogeny and polyploidy: Resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). Mol Phylogenet Evol. 85:97–116.

Zhong M., Hansen B., Nesnidal M., Golombek A., Halanych K.M., Struck T.H. 2011. Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. BMC Evol Biol. 11:369.

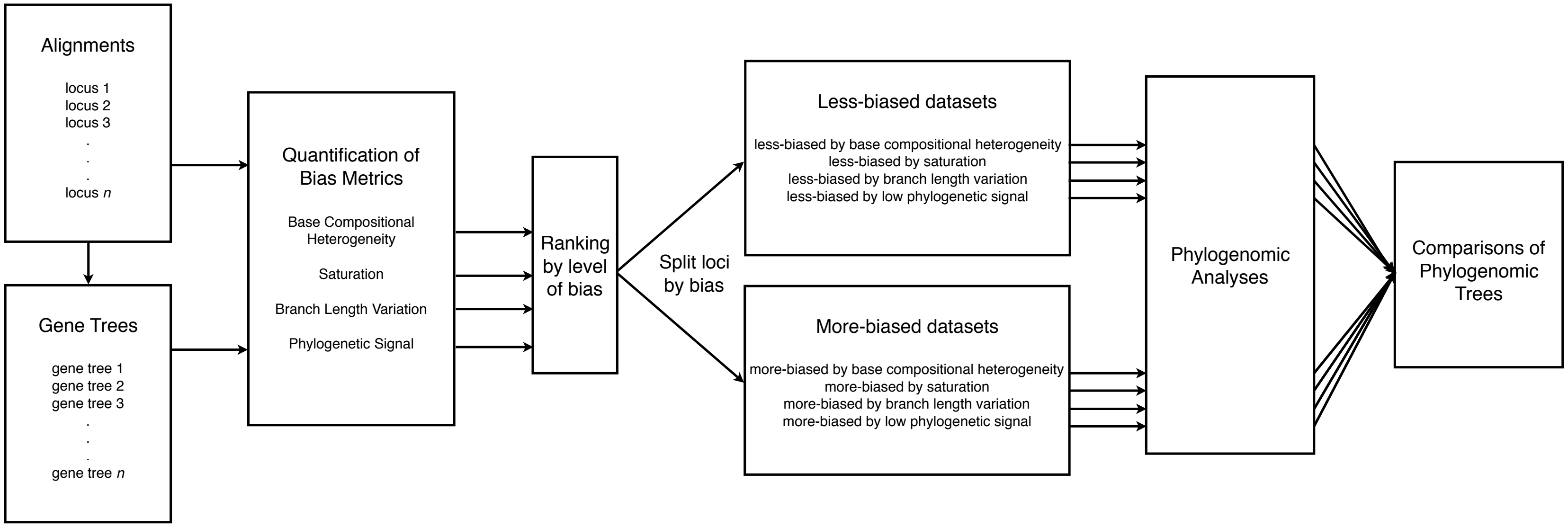

Alignments
locus 1
locus 2
locus 3
.
.
.
locus *n*
Gene Trees
gene tree 1
gene tree 2
gene tree 3
.
.
.
gene tree *n*
Quantification of Bias Metrics
Base Compositional Heterogeneity
Saturation
Branch Length Variation
Phylogenetic Signal
Ranking by level of bias
Split loci by bias
Less-biased datasets
less-biased by base compositional heterogeneity
less-biased by saturation
less-biased by branch length variation
less-biased by low phylogenetic signal
More-biased datasets
more-biased by base compositional heterogeneity
more-biased by saturation
more-biased by branch length variation
more-biased by low phylogenetic signal
Phylogenomic Analyses
Comparisons of Phylogenomic Trees

Figure 1. A conceptual map of the pipeline of analyses. First, various bias metrics that quantify potential sources of phylogenetic noise were calculated for each locus from gene alignments and their gene trees. Next, loci were ranked by their level of bias and were split into least-biased and most-biased datasets for phylogenomic analyses with both concatenation and coalescent-based analyses. Finally, differences between topologies were quantified and summarized.

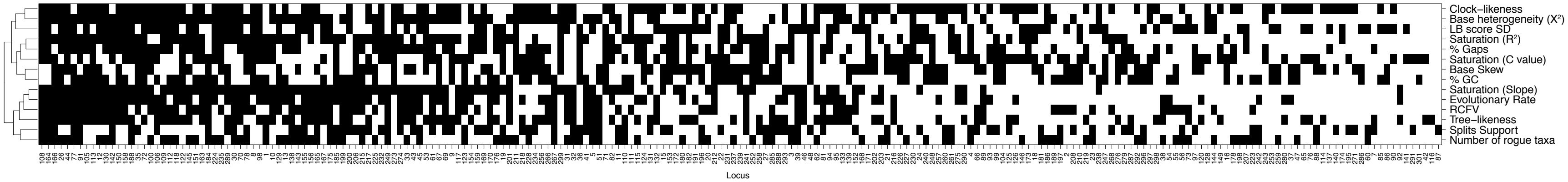


Figure 2. Distribution of loci in the least-biased (black) or the most-biased (white) data by each bias metric, as they were split into data matrices for phylogenomic analysis. Loci are ordered on the x-axis by the number of least-biased datasets each locus was included in. Bias metrics are clustered along the y-axis by the degree of overlap in the loci in the least-biased datasets. There was not a clear separation between least-biased and most-biased loci across different bias metrics, thus many loci are not necessarily consistently the least-biased or most-biased across all bias metrics. Refer to Table 1 and text for definitions of bias metrics.

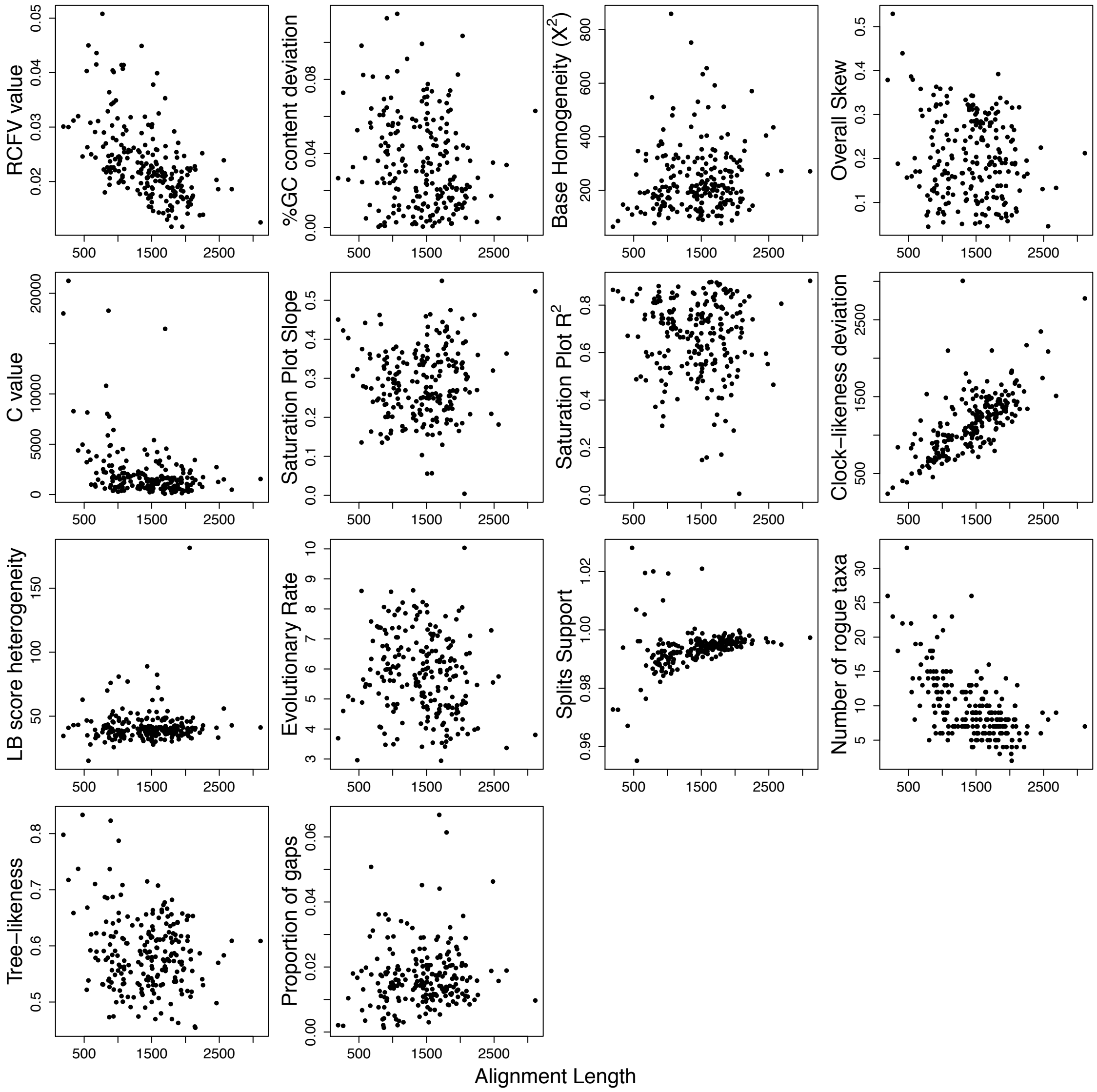


Figure 3. Variation in bias metric values as a function of locus length. Refer to Table 1 for definitions of bias metrics.

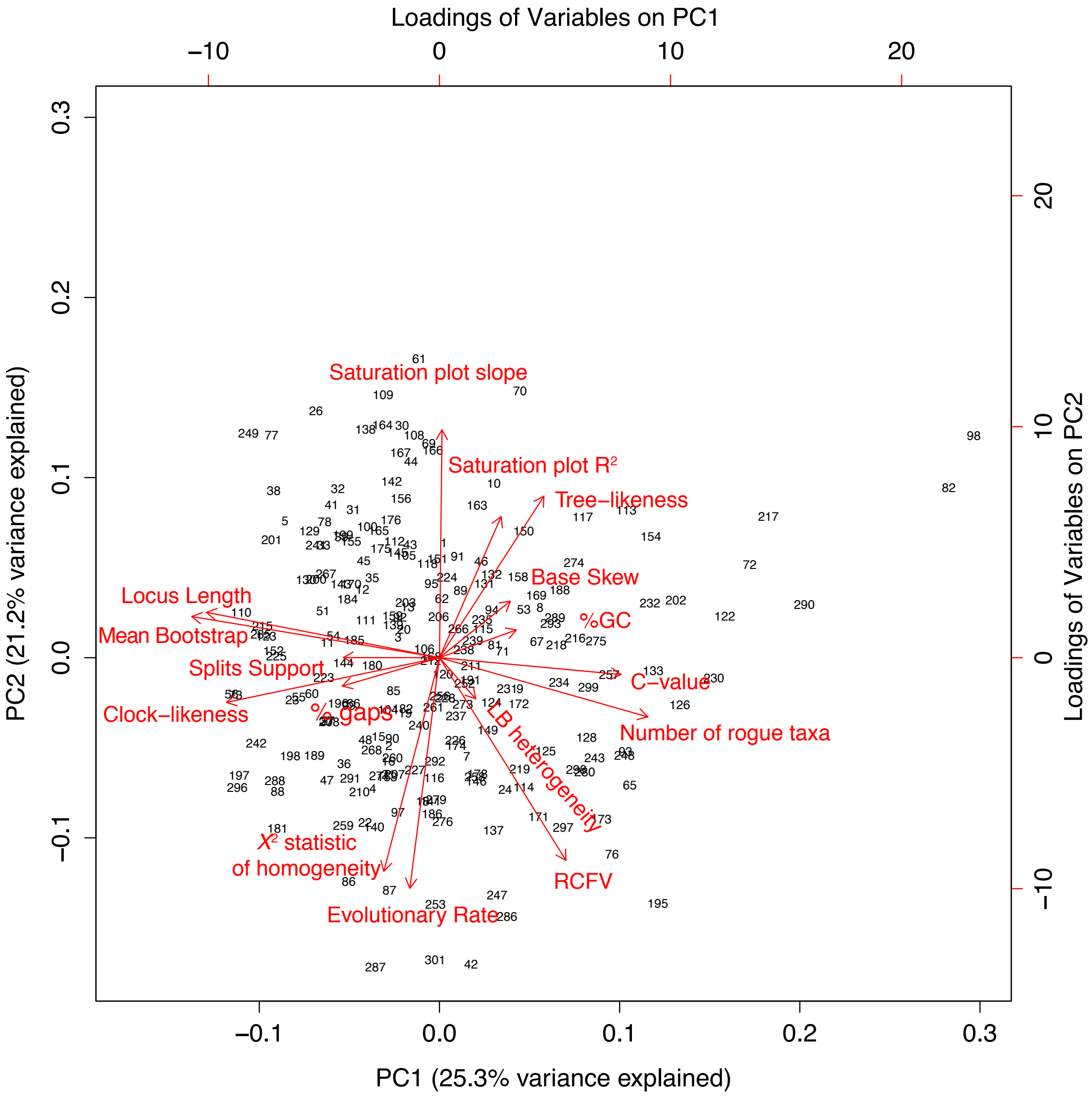


Figure 4. First two principal components showing major axes of variation in bias metric values across loci. Only some bias metrics contribute qualitatively smaller proportions of variation to the first two PC axes based on most bias metrics having longer vectors. Based on the variable directions of the vectors, the different bias metrics quantify a variety of different qualities of the loci.

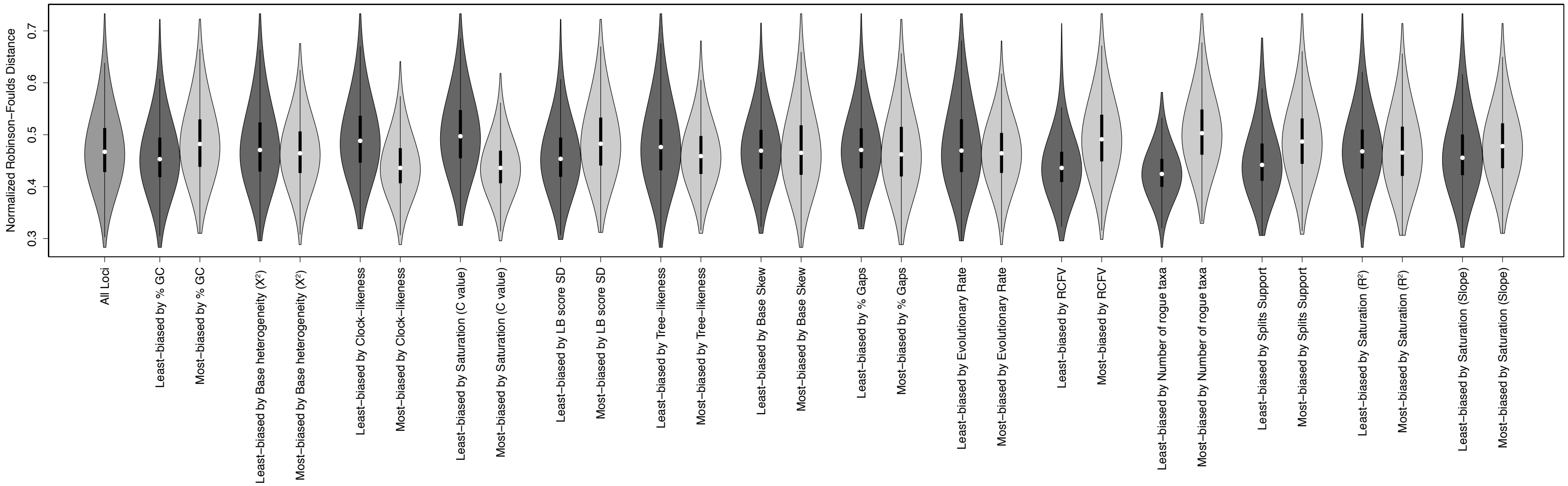


Figure 5. Gene tree heterogeneity as quantified by the distribution of Robinson-Foulds distances across gene trees of all loci within a dataset, for all loci as well as the least-biased (dark grey) and most-biased (light grey) dataset split for each bias metric. In general, little qualitative difference in gene tree heterogeneity is visible across datasets.

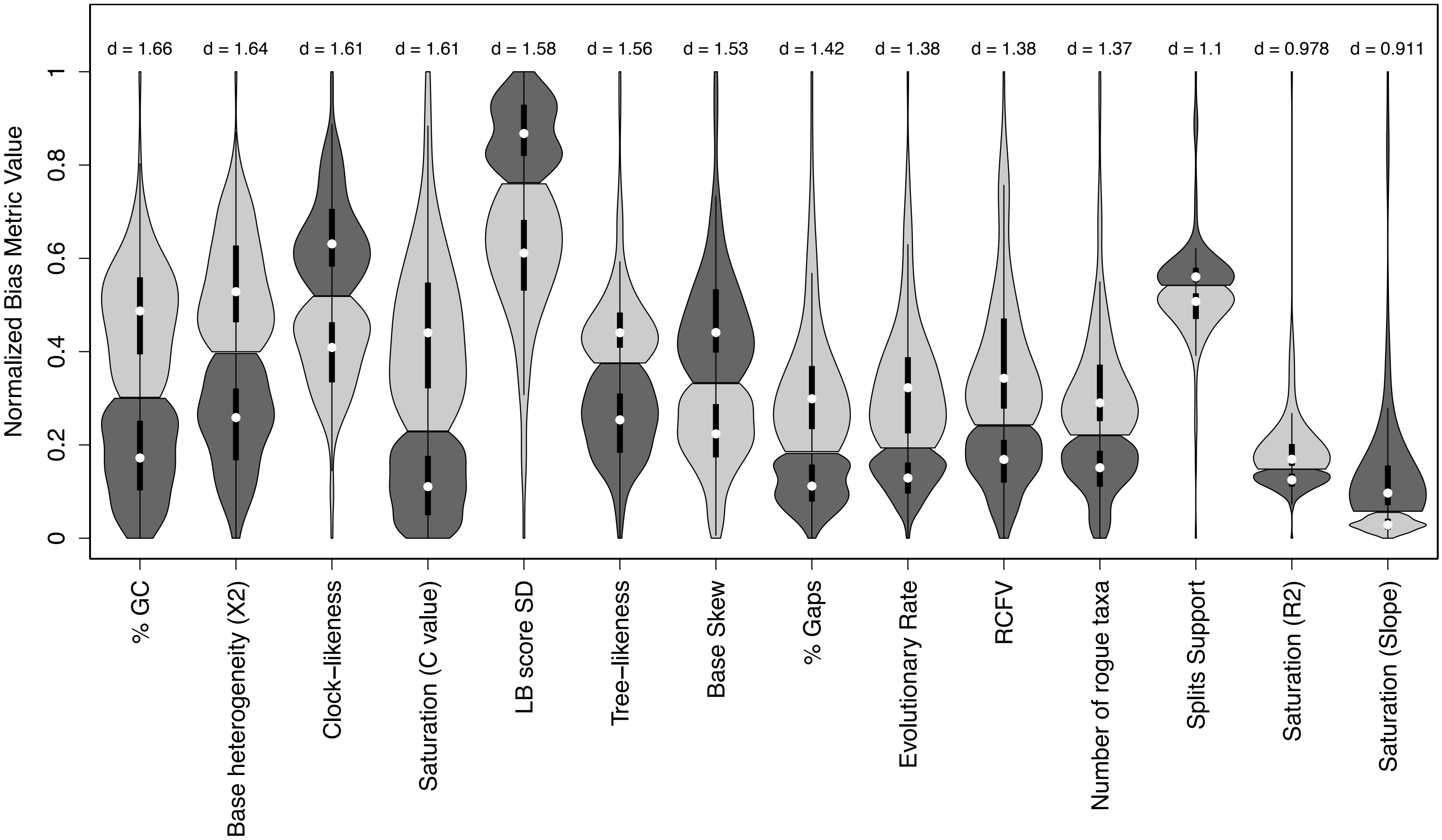


Figure 6. Variation within the least-biased (dark grey) and most-biased (light grey) dataset bins split for each bias metric. Overall variation of each bias metric was standardized to range from 0 to 1, and bias metrics are ranked by decreasing effect size (Cohen’s d) as a measure of how different the two dataset bins are for that particular bias.

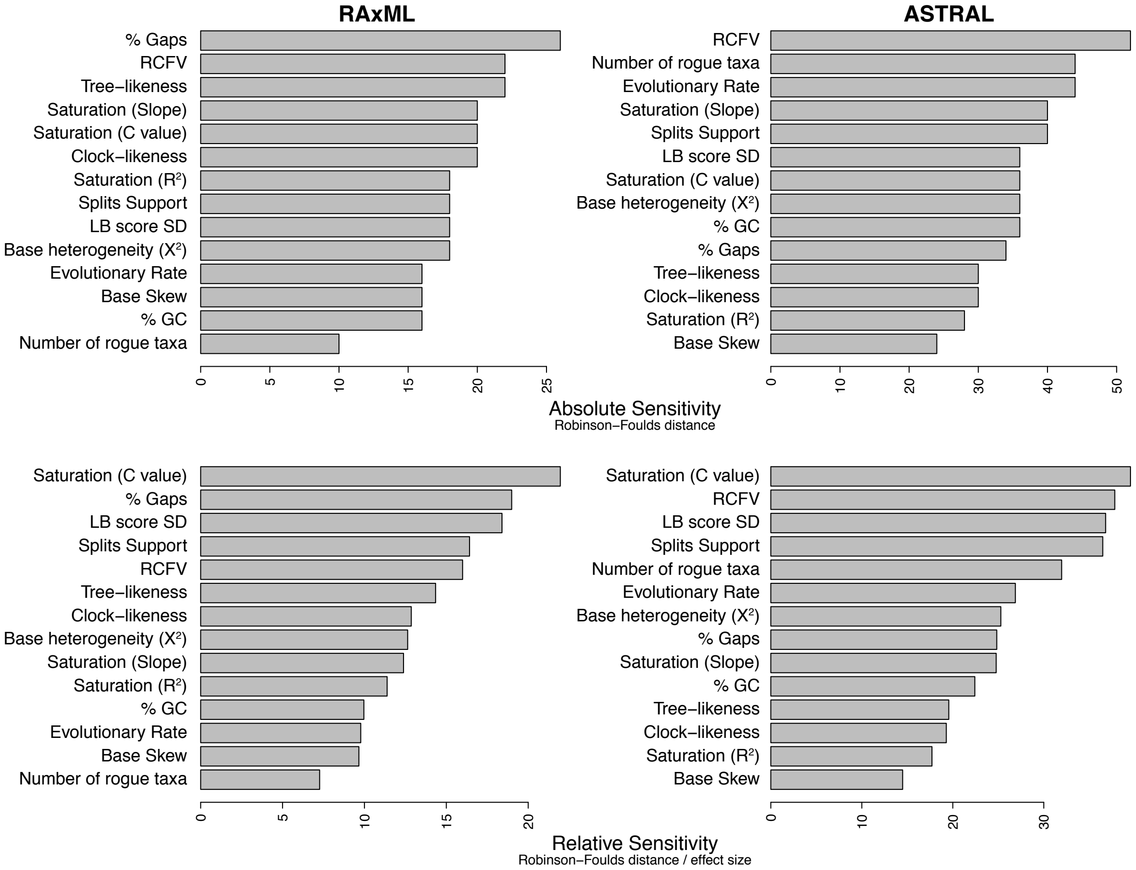


Figure 7. Sensitivity of phylogenomic analyses to each bias metric. Phylogenomic analyses result in different topologies between least-biased and most-biased datasets for each bias metric in (a) concatenation and (b) coalescent-based analyses. When sensitivity is standardized relative to difference in bias between datasets (c,d), certain biases have a greater relative effect despite a relatively low difference in bias between datasets.

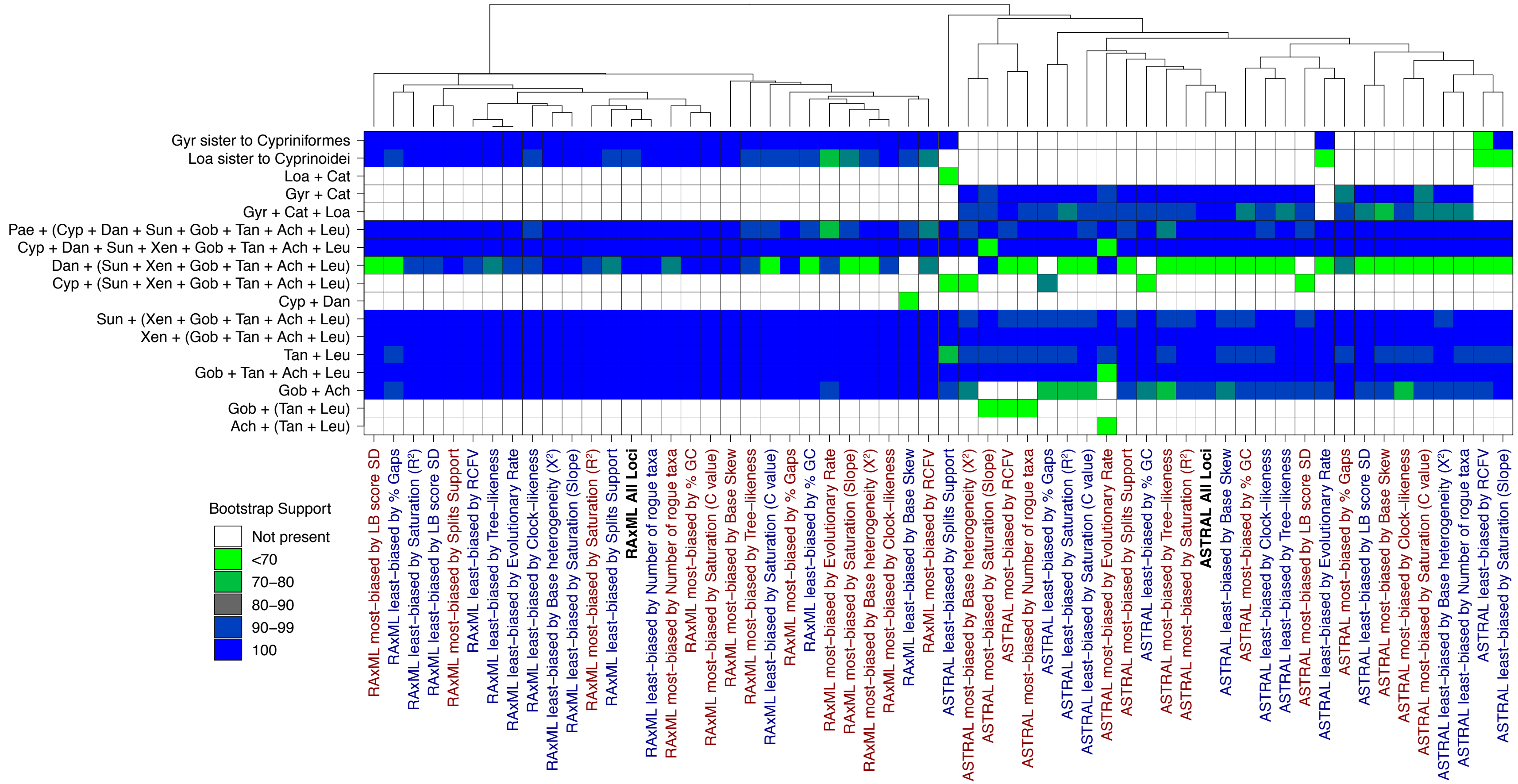


Figure 8. Consistency of selected clades recovered across phylogenomic analyses. Black boxes indicates 100% bootstrap support for that node, grey indicates 70-100% bootstrap support, and white indicates the split is not found in that topology. Labels to the left of the

graph indicate selected relationships tested. Parentheses indicate a focus on the node where the clade outside of parentheses is sister to the clade formed by clades within parentheses. Above the plot is a meta-tree visualizing clustering of phylogenomic trees using UPGMA and distances measured by Robinson-Foulds metric. Trees resulting from all loci are indicated in bold, trees resulting from the least-biased datasets are indicated in blue, and trees resulting from the most-biased datasets are indicated in red. Clades are abbreviated as the following: Gyr = Gyrinocheilidae, Cat = Catostomidae, Loa = loaches, Pae = *Paedocypris*, Cyp = Cyprinidae, Dan = Danionidae, Sun = *Sundadanio*, Xen = Xenocyprididae, Tan = Tanichthyidae, Leu = Leuciscidae, Gob = Gobionidae, and Ach = Acheilognathidae.

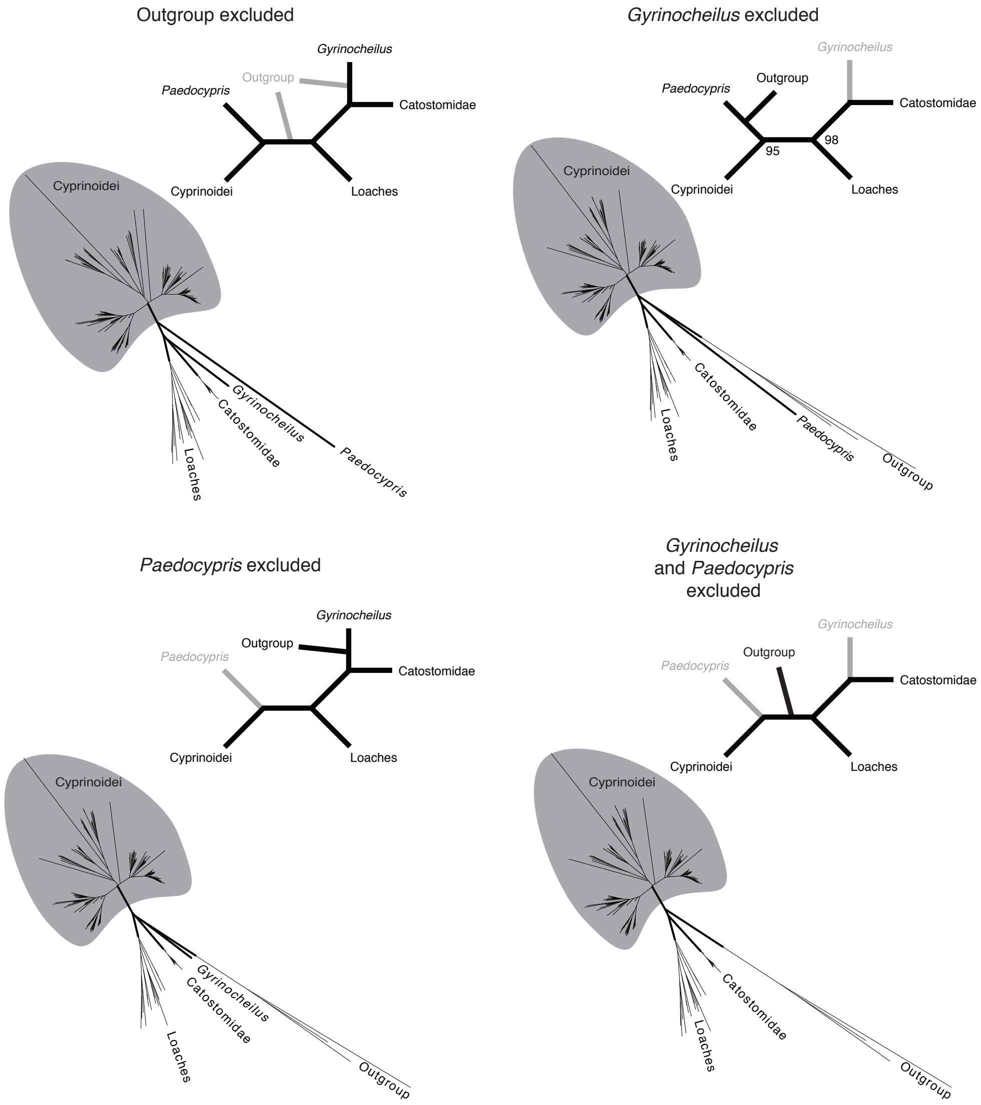


Figure 9. Unrooted trees inferred by excluding taxa on all loci using concatenation. Simplified topologies to the upper right of each topology have excluded clade in grey, with nodes labeled if <100% bootstrap support. For the outgroup exclusion (upper left), alternative grey branches indicate both placements of outgroup based on concatenation and coalescent-based analyses. Unrooted topology of the relative relationships of Gyrinocheilidae, Catostomidae, Cobitoidei *sensu stricto* (loaches), *Paedocypris* and the remaining Cyprinoidei are congruent across taxon removal experiments (as they were in bias sensitivity analyses), with the outgroup shifting when *Gyrinocheilus* is removed.

Table 1. Summary of various sources of phylogenetic error and bias metrics quantified for each locus, order of sorted values from least-biased to most-biased loci, and software used to calculate each one.

Class	Bias Metric	Metric Description	Order for ranking	Software	References
Base Compositional Heterogeneity	Relative Compositional Frequency Variability (RCFV)	RCFV value indicates the level of deviation from the mean base frequency.	Increasing	BaCoCa	Kück & Struck 2014
	Deviation in GC content	Absolute value of difference in proportion of GC (or AT) from all-equal base frequencies	Increasing	BaCoCa	Kück & Struck 2014
	X^2 test statistic of a test of homogeneity	X^2 test of homogeneity is used to test if base frequencies are homogeneous. Thus the X^2 test statistic provides a relative measure of heterogeneity.	Increasing	BaCoCa	Kück & Struck 2014
	Overall Skew	Square root of sum of squares of A/T, G/C, A/G, and C/T skews. Each skew metric provides a different measure of the bias between two bases; the overall skew provides provides a metric of skew in all four directions.	Increasing	BaCoCa	Kück & Struck 2014
Saturation	Convergence-value (C value)	Ratio of the standard deviation of transition-transversion ratio to the standard deviation of uncorrected genetic *p* distance. Smaller values indicate convergence in transition-transversion ratios and thus higher saturation.	Decreasing	BaCoCa	Kück & Struck 2014
	Saturation plot slope	Slope of linear regression of evolutionary distance vs. uncorrected genetic *p* distance	Decreasing	TreSpEx	Struck 2014
	Saturation plot R^2	R^2 of linear regression of evolutionary distance vs. uncorrected genetic *p* distance	Decreasing	TreSpEx	Struck 2014
Branch length variation	Clock-likeness	Log-likelihood ratio of the likelihood of a strict-clock model vs. the GTR+G model for each locus. Larger values indicate a larger deviation from clock-likeness.	Increasing	PAUP*	Swofford; Doyle et al. 2015
	Long branch heterogeneity	LB scores quantify the pairwise differences in branch lengths within a phylogeny. Greater variation (quantified by standard deviation) indicates branches have greater heterogeneity in long branches	Increasing	TreSpEx	Struck 2014
	Evolutionary rate	Sum of all branch lengths in the gene tree	Increasing	ape	Paradis et al. 2004; Salichos & Rokas 2013

Phylogenetic signal or Information Content	Splits support	Difference in average rank of top ten splits that agree vs. splits that disagree with the gene tree, weighted by the number of sites supporting each split, and divided by the sum of the number sites supporting all top ten splits	Decreasing	SAMS	Wägele & Mayer 2007
	Number of rogue taxa	Rogue taxa have unstable placement between bootstrap replicates, thus a locus with many rogue taxa is uncertain for many taxa	Increasing	RogueNarok	Aberer et al. 2012
	Tree-likeness	Proportion of resolved vs. unresolved randomly-selected quartets. A higher proportion of resolved quartests indicates more phylogenetic information in the data.	Decreasing	MARE	Misof et al. 2014
	Proportion of gaps	Proportion of undetermined characters within the alignment	Increasing	BaCoCa	Kück & Struck 2014

Table 2. Correlation between alignment length, average bootstrap support, and normalized bias metrics across loci. Pearson correlation coefficient above diagonal, p-value below diagonal.

	Locus length	Average bootstrap support	Deviation in %GC content	Overall Base Skew	X^2 test of homogeneity statistic	RCFV	C value	% gaps	Tree-likeness	Deviation from clock-likeness	Number of rogue taxa	LB score heterogeneity	Saturation plot slope	Saturation plot R^2	Splits support	Evolutionary Rate
Locus length	-	7.56E-01	-1.88E-01	-0.160647	1.49E-01	-5.75E-01	-3.81E-01	2.05E-01	-1.88E-01	7.54E-01	-5.83E-01	7.17E-02	7.00E-02	-8.06E-02	0.289492	-0.148752
Average bootstrap support	9.85E-42	-	-2.22E-01	-0.119664	9.46E-02	-4.79E-01	-5.56E-01	1.82E-01	-3.36E-01	6.02E-01	-8.18E-01	-2.92E-01	1.63E-01	-4.84E-02	0.215877	0.058412
Deviation in %GC	5.16E-03	9.42E-04	-	0.556519	-1.48E-01	5.27E-02	4.80E-02	4.41E-02	1.46E-01	-1.44E-01	1.63E-01	5.15E-03	-3.40E-03	-8.92E-02	0.015514	-0.116708
Overall Base Skew	1.73E-02	7.72E-02	3.30E-19	-	-2.00E-01	2.20E-02	7.31E-02	6.80E-03	2.50E-01	-1.55E-01	1.76E-02	-6.01E-02	1.02E-01	-1.03E-02	-0.167943	-0.106365
X^2 test of homogeneity statistic	2.73E-02	1.63E-01	2.85E-02	0.0029849	-	6.66E-01	4.60E-02	4.22E-02	-5.39E-01	4.22E-01	8.74E-03	2.98E-02	-4.78E-01	-1.84E-01	0.023325	0.55094
RCFV	1.05E-20	6.08E-14	4.38E-01	0.7456564	1.89E-29	-	3.10E-01	-9.96E-02	-2.55E-01	-2.08E-01	4.20E-01	-6.33E-02	-4.35E-01	-1.40E-01	-0.219992	0.538705
C value	5.58E-09	3.93E-19	4.79E-01	0.2817201	4.98E-01	2.85E-06	-	-5.03E-01	1.00E-01	-3.75E-01	4.53E-01	1.53E-01	-1.91E-02	3.72E-01	-0.190897	-0.021206
% gaps	2.34E-03	6.91E-03	5.16E-01	0.9203178	5.35E-01	1.42E-01	1.85E-15	-	1.26E-01	2.60E-01	-1.03E-01	1.11E-01	-2.31E-01	-5.19E-01	0.105733	-0.042224
Tree-likeness	5.25E-03	3.56E-07	3.05E-02	0.0001845	6.54E-18	1.40E-04	1.39E-01	6.21E-02	-	-3.22E-01	2.19E-01	3.23E-01	3.10E-01	4.64E-02	-0.083428	-0.711206
Deviation from clock-likeness	1.67E-41	6.13E-23	3.29E-02	0.0215128	7.18E-11	2.00E-03	1.00E-08	9.67E-05	1.16E-06	-	-4.26E-01	5.23E-02	-9.85E-02	-1.83E-01	0.246648	0.116214
Number of rogue taxa	2.47E-21	4.96E-54	1.56E-02	0.7957678	8.98E-01	8.87E-11	1.68E-12	1.29E-01	1.08E-03	4.46E-11	-	2.94E-01	-1.92E-01	1.74E-02	-0.130682	-0.015854
LB score heterogeneity	2.91E-01	1.13E-05	9.40E-01	0.3760274	6.61E-01	3.51E-01	2.31E-02	1.01E-01	1.00E-06	4.41E-01	9.86E-06	-	-4.37E-01	-3.03E-01	0.160532	-0.09765
Saturation plot slope	3.03E-01	1.58E-02	9.60E-01	0.1313491	7.02E-14	1.67E-11	7.79E-01	5.57E-04	3.02E-06	1.46E-01	4.30E-03	1.33E-11	-	7.07E-01	-0.069106	-0.646476
Saturation plot R^2	2.35E-01	4.76E-01	1.88E-01	0.8789957	6.37E-03	3.87E-02	1.31E-08	1.74E-16	4.94E-01	6.53E-03	7.98E-01	5.09E-06	1.65E-34	-	-0.056763	-0.355797
Splits support	1.34E-05	1.31E-03	8.19E-01	0.0128171	7.31E-01	1.05E-03	4.58E-03	1.19E-01	2.19E-01	2.28E-04	5.35E-02	1.74E-02	3.09E-01	4.03E-01	-	0.023549
Evolutionary Rate	2.77E-02	3.90E-01	8.49E-02	0.1165296	8.75E-19	7.00E-18	7.55E-01	5.34E-01	4.64E-35	8.62E-02	8.16E-01	1.50E-01	2.63E-27	6.19E-08	0.72893	-

Table 3. Comparison of loci characteristics based on what relationships they recover for *Gyrinocheilus* and *Paedocypris*, including locus length, average gene tree bootstrap support, level of bias, and gene tree heterogeneity. P-value derived from a two-tailed t-test, marked with asterisk if p < .05.

		Mean: *Gyrinocheilus* sister to Cypriniformes (n = 95)	Mean: Monophyletic Cobitoidei *sensu lato* (n = 64)	p-value	*Paedocypris* sister to Cypriniformes (n = 52)	*Paedocypris* sister to remaining Cyprinoidei (n = 102)	p-value
Sampling Error	Locus Length	1560.189	1306.563	0.001340*	1341.135	1628.892	0.000439*
	Mean Bootstrap Support	71.39699	67.98032	0.002255*	68.25379	72.09099	0.002289*
Base Compositional Heterogeneity	% GC	0.033445	0.029102	0.263511	0.028112	0.031372	0.414701
	Overall Skew	0.205895	0.178222	0.057523	0.184191	0.210898	0.118141
	Base heterogeneity (X^2)	232.9573	236.7752	0.842827	247.5383	225.8014	0.252971
	RCFV	0.022157	0.024419	0.032632*	0.02421	0.021281	0.006292*
Saturation	Saturation (C Value)	1474.551	2213.704	0.055548	2326.523	1733.688	0.256937
	Saturation (Slope)	0.305673	0.276866	0.029073*	0.27375	0.310615	0.011287
	Saturation (R^2)	0.678314	0.668144	0.679228	0.666272	0.680586	0.614736
Branch Length Variation	Clock-likeness	1243.898	1122.422	0.087862	1042.164	1264.961	0.000951*
	LB score heterogeneity	39.54958	40.48854	0.487793	43.48548	40.3311	0.293231
	Evolutionary Rate	5.705226	6.106417	0.046336*	6.236591	5.510276	0.001196*
Phylogenetic Signal	Tree-likeness	0.583351	0.575599	0.428021	0.571347	0.592363	0.054124
	Number of rogue taxa	8.852632	10.17188	0.052780	10.44231	8.431373	0.017983*
	Splits Support	0.994148	0.993016	0.242082	0.992694	0.994262	0.197308
	% Gaps	0.017653	0.015028	0.043002*	0.016273	0.016381	0.950415
Gene Tree Heterogeneity		0.474855	0.472476	0.004130*	0.469864	0.468648	0.068079

CHAPTER 4

FUNCTIONAL GENOMIC EVOLUTION OF PAEDOMORPHIC CYPRINIFORMES

ABSTRACT

Large phenotypic changes throughout evolution may have their foundation in widespread functional genomic changes. The order Cypriniformes includes multiple, independent evolutionary transitions to an extreme miniature, paedomorphic phenotype, represented by the genera *Paedocypris*, *Sundadanio*, and *Danionella*. To study functional genomic changes in the evolution of paedomorphism in these taxa, we study whether rates of nonsynonymous mutation vs. synonymous mutation differ in paedomorphic lineages relative to other teleosts for 8,687 genes. We discovered 2,686 genes that had a relatively greater level of purifying selection in paedomorphic taxa relative to other teleosts, and 258 genes that had a relatively greater level of positive selection in paedomorphic taxa relative to other teleosts. Genes related to bone development did not have a significantly different level in paedomorphic taxa relative to other genes on average, but genes related to growth did. In addition, a few particular genes for these classes evolving under positive selection in paedomorphic taxa may have functional consequences for their phenotype. Functional categories significantly enriched (i.e. over-represented) among genes evolving under greater purifying selection in paedomorphic taxa included a variety of genes related to transmembrane protein function and oxidation pathways. A variety of anatomical structures are significantly enriched among genes evolving under greater purifying selection in paedomorphic taxa, including yolk layers, cardiovascular anatomy, neural anatomy, and muscle segments. Positively-selected genes in paedomorphic taxa included

genes functional in transcription and cell cycle control, and also included genes related to anatomical development of the axis, eye, pronephric mesoderm, and optic tectum, though no significant enrichment was found for any functional or anatomical category.

INTRODUCTION

Understanding the basis of phenotypic diversity across taxa has been a dominant theme of evolution since its conception (Darwin 1882; Raff 2000). Although it has been decades since genes have been demonstrated to encode the blueprint necessary for development and thus underlying organism phenotype (Raff 2000), the interplay between genomic evolution and phenotypic evolution is still poorly understood for the vast majority of taxa and phenotypic characters. Developmental biology experiments in model organisms have been a powerful tool for understanding the genetic basis of phenotype (Haffter et al. 1996; Raff 2000), but how these same genes evolve to influence phenotype in nature remains unknown (Edmunds et al. 2015). Comparisons between taxa on an evolutionary scale allow for addressing how findings in model organisms are evolutionarily relevant. They also allow for the discovery of potential new genes and evolutionary patterns involved in the evolution of phenotypes. Genetic screens in experimental research target a subset of potential genotypes and phenotypes (Schier et al. 1997), and the evolutionary relevance of these mutations are unknown. Biodiversity provides a natural experiment where extant species are natural mutants that have survived the screen of natural selection, thus providing another avenue to discover genotype-phenotype relationships (Mayden and Chen 2010).

The order Cypriniformes includes some of the smallest known vertebrates, including paedomorphic fishes such as *Paedocypris*, *Sundadanio*, and *Danionella* (Roberts 1986; Kottelat et al. 2006; Conway et al. 2011). These genera share truncation in their development, which leads to the loss of numerous bones that normally develop late in ontogeny (Britz and Conway 2009; Britz et al. 2009). Phylogenomic analysis robustly demonstrates that these three genera are not closely related (Stout et al. *submitted*; Tan & Armbruster *in prep*.; Chapters 2, 3), and thus their extreme morphology evolved convergently. Miniaturization has been found to release developmental constraints in a variety of taxa (Weitzman and Vari 1988), and likely plays a role in numerous phenotypic novelties found in taxa such as *Paedocypris* and *Danionella* (Britz and Conway 2009; Britz and Conway 2016). All three paedomorphic genera are represented by long branch lengths in phylogenetic analyses (Mayden and Chen 2010), demonstrating that these taxa have undergone rapid molecular evolution. The zebrafish, due to its inclusion within Cypriniformes, provides a comparative resource for developmental and genomic biology (Meyer et al. 1993; Howe et al. 2013), making this order is a potential clade to study the evolution of functional and developmental genes on a genomic scale. The relationship of functional genomic evolution to phenotypic evolution in cypriniforms besides zebrafish has only begun to be studied (Meng et al. 2013; Xu et al. 2014; e.g. Wang et al. 2015).

To study the functional genomic evolution of paedomorphic cypriniforms, we sequenced and assembled transcriptomes for the paedomorphic taxa *Paedocypris*, *Danionella*, and *Sundadanio*. We compared the level of selection on thousands of single-copy protein-coding genes genes between paedomorphic taxa to their orthologs across teleosts to discover genes of importance in the evolution of paedomorphic taxa. Of these genes, we focused on the level of selection belonging to two major groups based on their involvement in either bone development or growth. Finally, we determined if any functional classes of genes were more likely to be differentially-selected relative to all genes tested.

RESULTS

We generated transcriptomic data for the paedomorphic cypriniform species *Paedocypris*, *Sundadanio*, and *Danionella* and two non-paedomorphic cypriniformes *Tanichthys* and *Leptobarbus* (Table 1). Transcriptomes were also assembled from publically available transcriptome raw reads for several members of Cypriniformes (Table 1). Transcriptome assemblies varied considerably in total assembly size and number of isoforms clusters assembled by Trinity (Grabherr et al. 2011). Transcriptome assemblies also varied in the number of non-redundant sequences after redundant sequences were excluded using CD-HIT-EST (Li and Godzik 2006).

Using biomaRt to query Ensembl (Durinck et al. 2005; Durinck et al. 2009), we determined a reference set of 12,457 single-copy core ortholog groups across teleost fishes. This reference set was derived by obtaining peptide sequences of one-to-one orthologs (as annotated by Ensembl) for all pairwise combinations of the following nine teleost genomes: zebrafish (*Danio rerio*), Mexican cave tetra (*Astyanax mexicanus*), Atlantic cod (*Gadus morhua*), three-spined stickleback (*Gasterosteus aculeatus*), Japanese ricefish (*Oryzias latipes*), Fugu (*Takifugu rubripes*), green spotted puffer (*Tetraodon nigrividiris*), Nile tilapia (*Oreochromis niloticus*), and platyfish (*Xiphophorus *maculatus*). We were able to recover between 8,370 to 11,640 of these 12,700 single-copy orthologs among the Cypriniformes using HaMStR (Ebersberger et al. 2009 Table 1). After several filtering steps to exclude genes not present in all three paedomorphic taxa, genes with extremely short alignments, and genes with poor taxon sampling, a set of 8,687 genes were retained for further testing for positive selection.

The branch test of positive selection allows testing for whether the ratio of nonsynonymous to synonymous codon substitutions (ω) differs between foreground lineages vs. background lineages (Yang 1998 Fig. 1), with relatively larger values of ω indicating more nonsynonymous changes vs. synonymous changes than expected, and thus increased positive selection. Based on tests of positive selection where the three branches leading to paedomorphic taxa were classified as foreground branches, we determined 2,944 genes as significantly differentially selected (SDS) genes in the paedomorphic taxa at adjusted p < .05 (i.e. after adjusting the p-value for multiple testing; Fig. 2; Wright 1992). Thus, approximately one-third of the genes tested had significantly different rates of positive selection in paedomorphic taxa. These genes could be divided into two groups. There were 2,686 genes which had a lower ω in paedomorphic taxa relative to the other teleosts, and thus had fewer nonsynonymous substitutions than expected (i.e. greater purifying selection) in paedomorphic taxa. There were 258 genes that had a higher ω in paedomorphic taxa relative to other teleosts, which demonstrates more nonsynonymous substitutions than expected (i.e. relatively greater positive selection).

Genes tested for differences in positive selection were distributed throughout the zebrafish genome (Fig. 2), with two exceptions. First, no genes were tested on the zebrafish mitochondrial genome. Second, a large portion of the zebrafish chromosome 4 arm did not have any genes tested. This corresponds to a genomic region that is known to have relatively few protein-coding genes, highly repetitive DNA, and gene expansions due to a high rate of gene duplication relative to other teleost genomes (Howe et al. 2013). Not only were genes tested distributed across chromosomes, SDS genes were also distributed across chromosomes. Using DAVID to test whether certain chromosome locations were enriched (i.e. over-represented) among SDS genes (Huang et al. 2008), we found that zebrafish chromosome 1 was enriched for SDS genes with greater positive selection in paedomorphic taxa, while zebrafish chromosomes 5, 8, 13, 18, 19, and 20 were enriched for SDS genes with greater purifying selection in paedomorphic taxa.

Because the paedomorphic taxa have many bones that are lost or do not ossify, it is of specific interest whether or not these genes are under differential selection. From the full dataset of 8,687 genes tested overall, 84 of these genes were related to bone formation, morphogenesis, development, remodeling, or mineralization. These 84 genes did not have a significantly different level of selection relative to the remaining genes, (Mann-Whitney U test, p = 0.506), demonstrating that these bone development genes did not have greater or lower selection than expected on average. On an individual basis, 20 genes were significantly differentially selected between paedomorphic taxa and non-paedomorphic taxa (Table 2). Only two of these genes had significantly greater positive selection in paedomorphic taxa, scube3 (ENSDARG00000011490) and hhip (ENSDARG00000060397). The hhip gene is a pleotropic hox-interacting protein that can have large phenotypic consequences for zebrafish mutants, as described by (Koudijs et al.

2005), including greater proliferation of cells in the eyes, ears, and fins, but also whole-body dwarfism.

Because of the extremely small body size of paedomorphic fishes, we were also interested in genes related to growth. In our set of single-copy genes tested, 106 of the genes contained the term ‘growth’ in their Gene Ontology (GO) annotation (Harris et al. 2004), which includes genes related to growth ranging from whole organism level to cellular level. On average, this set of 106 genes was found to differ in level of selection from the remaining genes (Mann-Whitney U test, $p = 0.033$). Of these 106 genes, 41 of these genes were SDS genes. Only a single one of these SDS genes had a greater level of positive selection in paedomorphic taxa relative to non-paedomorphic teleosts, bmp3 (bone morphogenetic protein 3, ENSDARG00000060526), (adjusted $p = 0.0173$), a gene known to be relevant for development of the shape of craniofacial structure, and knockdown experiments for this gene in zebrafish can even result in the absence of craniofacial bones (Schoenebeck et al. 2012). The remaining 40 genes encompassed a variety of genes including several growth factors and growth factor receptors (Table 3).

Tests of functional enrichment allow for determining whether a functional category of genes is over-represented relative to the functions across all genes tested, to discover what classes of genes may be important. SDS genes undergoing greater purifying selection in paedomorphic taxa were significantly enriched ($p < 0.05$, false discovery rate cut-off of 0.05) for functional categories including two major clusters: a main cluster of protein superfamily functions including G-coupled proteins, cell adhesion, membrane adhesion, signal transduction, and transport across membranes and GO functions including proteins intrinsic to the cell membrane, integral to the membrane, and proteins that are part of the plasma membrane; and a second, small cluster including transferase and kinase protein superfamily functions (Fig. 3).

Genotype-phenotype relationships discovered in mutant zebrafish on the ZFIN database (Sprague et al. 2003; Bradford et al. 2011) allow for testing for if the development of particular anatomical structures is more commonly affected by differentially-selected genes (Fig. 4). A variety of ZFIN anatomy annotations were recovered forming several disconnected networks, including networks formed by epidermis, periderm, and embryonic enveloping layer (EVL); liver and gut; and heart tube and heart. Several other anatomical structures that did not share genes with other structures included yolk syncytial layer (YSL), cardiovascular system, vein, musculature system, peripheral olfactory organ, and pronephric duct (Fig. 4). This demonstrates that a diversity of anatomical structures are affected by greater purifying selection in paedomorphic taxa, and that this is not due to pleiotropic genes that affect the development of many anatomical structures.

No functional categories or anatomical structures were enriched among genes under greater positive selection in paedomorphic taxa. This is likely due to the relatively small number of genes overall, with only 258 genes with significantly higher positive selection in paedomorphic taxa relative to non-paedomorphic taxa. Of these, only 188 were recognized by DAVID, and only 74 had ZFIN anatomical annotations. However, differentially positively selected gene functions are still of interest whether or not they are more highly biased towards any particular relative to the background. Among the functional categories recovered among genes under purifying selection, clusters were recovered from DAVID that were related to transcription, including transcription factor binding, regulation, and nucleotide excision repair; phosphorylation by serine-threonine kinases; cell cycle regulation by ubiquitin-mediated proteolysis; and complex 1 LYR protein. Positive selection was found on several motifs, including C2H2-type Zinc finger, tetratricopeptide repeats, WD40 repeats, and RNA recognition RNP-1 motifs (Fig. 5). Among the anatomical structures annotated for genes under positive selection, genes were related to the eye, optic tectum, pronephric mesoderm, and axis (Fig. 6).

Discussion

Widespread functional genomic evolutionary changes may be related to the evolution of extreme phenotypes. We discovered that a significantly different level of selection is associated with approximately one-third of the genes we tested. Most of this was associated with increased purifying selection in paedomorphic taxa relative to non-paedomorphic teleosts. In other words, there are fewer nonsynonymous mutations relative to synonymous mutations than expected in paedomorphic taxa when compared to other teleosts. The finding of a widespread increase in the level of purifying selection across a third of the genes tested contradicts the hypothesis that large genomic changes driven by positive selection led to the morphological convergences in paedomorphic taxa. On the other hand, this is an intuitive result given the rapid rate of molecular evolution in paedomorphic taxa. Because paedomorphic taxa have rapid molecular evolution, as indicated by long branch lengths in previous phylogenetic analyses (Mayden and Chen 2010), if paedomorphic taxa had the same level of selection as in other non-teleosts, they would also be expected to have many more nonsynonymous mutations, in proportion to the number of synonymous mutations. Thus, maintenance of the function of a particular gene under natural selection in the face of rapid mutations would result in a decrease in the rate of nonsynonymous mutations relative to synonymous mutations. These genes thus demonstrate genes that may have important basic function in teleosts that are conserved in zebrafish. The functions affected include cellular processes that do not necessarily have macroscopic, phenotypic effects that could be predicted based on prior morphological studies. In particular, these genes were significantly enriched for functions related to membrane transport and signaling. Membrane proteins involved with signaling are an important player in development, where cell-to-cell communication is a major player in morphogenesis (Kimmel et al. 2001). Genes under purifying selection were enriched for functions in a variety of anatomical organs. Extending the hypothesis that rapid molecular evolution in paedomorphic taxa leads to fewer relative nonsynonymous mutations in genes that are more constrained by selection, these genes should also be related to some of the more important organ systems whose functions are more constrained by selection. Thus, the finding of purifying selection for genes related to the development of a variety of organ systems that are inherent to organismal function – such as yolk layers, epidermis, heart, muscle, liver, and gut – is logical. These are all also structures that develop early and are not lost in paedomorphic taxa, as opposed to various anatomical structures such as bones.

The anatomical structures that could be affected by genes under positive selection include the pronephric mesoderm. The pronephros is the anterior part of the kidney, and it generally regresses early during development in gnathostomes after formation of tubules in the the mesonephric region (Kardong 2014). However, a paedomorphic goby, *Schindleria,* was found to have the pronephros as the functional kidney, and it was similar in form to that of hagfishes, which makes it highly derived within teleosts (Schindler 1932; Johnson and Brothers 1993). It is unknown whether the pronephros is also functional in the paedomorphic cyprinids nor what its form is, but if it is functional, the positive selection on genes related to the pronephros could be explained by the pronephros needing to maintain urinary function in a life stage where it is normally not active.

The karyotype and genome evolution of *Paedocypris* is particularly interesting among cypriniforms, as it represents one of the smallest vertebrate genomes with the fewest number of chromosomes among cypriniforms. *Paedocypris* has a genome size not far above 300 megabases, and two species studied have 30 and 34 chromosomes (Liu et al. 2012). The paedomorphic phenotype of *Paedocypris* might lead to speculation that there are large numbers of genes missing that are no longer functional given the lack of the adult morphology. While gene loss in *Paedocypris* cannot be directly assessed with transcriptome data, the distribution of the orthologs across the zebrafish genome provides some information. Orthologs of the genes we recovered in *Paedocypris* are distributed across the zebrafish genome, with the exception of the zebrafish-specific segment of chromosome 4. This arm of chromosome 4 in zebrafish contains many duplicated genes and does not share homology with any previously sequenced teleost genome, *Cyprinus carpio*, or *Hypophthalmichthys* (Howe et al. 2013; Xu et al. 2014; Zhu et al. 2015), indicating this chromosome arm arose after the divergence of zebrafish from other teleosts, from Cyprinidae, and from Xenocyprididae. Genome size variation across life is not related much to gene number; rather, genome size is well known to correlate with an increase in the proportion of noncoding DNA in the genome, represented primarily by transposable elements within the genome (Gregory 2005). The *Tetraodon nigroviridis* genome, the smallest vertebrate genomes sequenced so far, confirms that the loss of transposable elements is a major factor in genome size reduction in fishes (Jaillon et al. 2004). It is likely that the genome miniaturization in *Paedocypris* is also driven by substantial decreases in transposable elements. Additionally, although their function is not enriched relative to the background (potentially given the lack of gene ontology annotations for some of these proteins in zebrafish) the anaphase promoting complex protein subunits 5, 7, and 16 are some of the proteins with greatest level of difference in positive selection in paedomorphic taxa relative to other teleosts; the anaphase promoting complex is important in humans in maintaining genomic stability and therey preventing cancer (Wäsch et al. 2010).

The absence of secretory calcium-binding phosphoproteins SCPP1 and SCPP5 has been implicated in the evolution of bony structures in the elephant shark, which lacks endochondral bone (Venkatesh et al. 2014), and the channel catfish, which lacks scales (Liu et al. 2016). The paedomorphic fishes have a number of endochondral bones that remain unossified, and *Paedocypris* and *Danionella* lack scales (Britz and Conway 2009). Two of the scpp genes, sparc (osteonectin) and spp1 genes, were both found to have significantly greater purifying selection in paedomorphic taxa relative to non-paedomorphic taxa. The remaining genes were not tested. This was potentially due to our high requirement for orthologs to be present in at least seven of the sequenced teleost genomes, so we filtered out all of the scpp genes in the preparation of our reference ortholog set. Of the SCPP genes found in zebrafish, only SCPP1 and SCPP5 have orthologs between zebrafish and other teleosts, and neither have over five species represented. Liu et al. (2016) also noted the lack of numerous scpp copies in various sequenced teleost genomes. A lower threshold requirement for number of orthologs in non-cypriniform teleosts could increase the number of loci primarily found in zebrafish and other cypriniforms, at the expense of lower power in testing positive selection due to lower taxon sampling.

Given a significantly higher level of positive selection on bmp3 in paedomorphic taxa than non-paedomorphic teleosts, a role for bmp3 in craniofacial structuring of paedomorphic fishes is possible. Mutant zebrafish for bmp3 have extremely reduced craniofacial development of bones and cartilage, including the absence of numerous cartilaginous elements in the embryo (Schoenebeck et al. 2012). Evolution of expression in another bmp gene, bmp4, is an important regulator in the craniofacial evolution in cichlids (Parsons and Albertson 2009). This is consistent with the perturbed craniofacial morphology, with the absence of numerous bones in paedomorphic fishes (Britz and Conway 2009). In particular, there are numerous craniofacial modifications in *Danionella* species, which is exemplified by extreme craniofacial modifications such as the absence of a kinethmoid (a synapomorphy for Cypriniformes), the presence of tooth-like odontoid processes on the jaws in *Danionella dracula* (lack of oral teeth is a synapomorphy of Cypriniformes), and the upper jaw bones being represented by a single element of unknown homology to the premaxilla and maxilla of other teleosts (Britz and Conway 2016).

Although evolution in functional gene sequences are an important factor shaping the phenotype of organisms, development of phenotype is a complex interplay between gene function and gene expression. Although differences in gene expression can be assessed using RNA-seq (Wang et al. 2009), tissues were not standardized across all species used in this study, making expression tests meaningless. In addition, we lack replication. Thus, we did not address the evolution of gene expression in paedomorphic taxa here. Causal mutations that can affect gene expression, such as mutations in *cis*-regulatory regions (Hoekstra and Coyne 2007), cannot be discovered from transcriptomes as they are not transcribed. Mutations in noncoding RNAs, important in regulation of gene expression (Prasanth and Spector 2007), also cannot be assessed given the use of polyA-tail enrichment used to enrich for mRNAs; even if they were sequenced, tests of selection that depend on codon substitution models are meaningless in this respect, and thus different approaches must be used. Furthermore, even if evolution on the gene sequence level has a role in the evolution of paedomorphism, if these genes are not expressed in the adult fish then they will not be captured in the transcriptome. Nonetheless, we were able to test a variety of genes representing roughly a third of the known genes in the zebrafish genome, even though it has the largest number of vertebrate genes in a sequenced vertebrate genome so far (Collins et al. 2012; Howe et al. 2013). As approaches towards genome sequencing and assembly continue to improve (Bradnam et al. 2013), genomic comparisons will provide a more comprehensive picture of the functional genomic evolution of paedomorphic taxa, particularly in regards to genome miniaturization in *Paedocypris* (Liu et al. 2012).

CONCLUSION

We discovered that the rapid molecular evolution in paedomorphic taxa has led to a decrease in the level of selection on functional genes across the genome, demonstrating an important characteristic of evolutionary rate in influencing the inferred level of selection, and highlighting the genes that are relatively conserved in paedomorphic taxa. In addition, we discovered a much smaller complement of genes had greater levels of selection in paedomorphic taxa relative to other teleosts, some with potential relationship to their extreme phenotype, including genes related with bone development and morphogenesis. The comparison of independent lineages that have converged on a similar phenotype allows for a replicated natural experiment for the study of evolutionary patterns. This demonstrates the opportunity for paedomorphic fishes to provide insights into both the most conserved genes among fishes as well as genes underlying their extreme, convergent phenotype.

MATERIALS AND METHODS

Taxon sampling

The Cypriniformes is a diverse order with a number of genomic and transcriptomic resources available. We generated five additional cypriniform transcriptomes, including three paedomorphic taxa *Paedocypris* cf. *progenetica*, *Danionella* cf. *translucida*, and *Sundadanio* sp., as well as two non-miniature taxa, *Tanichthys micagemmae* and *Leptobarbus hoevenii*. Included taxa and sequence information is provided in Table 1. Cypriniformes includes a number of taxa that are sometimes or always tetraploid (Leggatt and Iwama 2003; Yang et al. 2015). Because we are interested in functional genetic evolutionary rates relative to paedomorphism, rather than the effect of polyploidization and subsequent diploidization on rates (Li et al. 2015), we excluded these taxa so that we could focus on only evolution under diploidy.

Specimen collection, RNA isolation and sequencing

All protocols followed IACUC 2014-2451. Specimens were obtained from the ornamental pet trade. Live specimens were anesthetized using MS-222. Whole specimens of *Paedocypris*, *Sundadanio*, *Danionella*, and *Tanichthys* were preserved in RNA Later, while the head was dissected from a *Leptobarbus* specimen for preservation. Samples were stored at –80ºC prior to extraction. RNA was extracted from whole specimens (*Paedocypris*, *Sundadanio*, *Danionella*) or the entire head (*Tanichthys*, *Leptobarbus*). Tissues were homogenized, and whole RNA using TRIzol extraction and cleaning with Omega Bio-Tek RNA extraction kit. Whole RNA extract was submitted for sequencing at HudsonAlpha Genome Sequencing Lab (Huntsville, AL), where polyA+ cDNA library preparation with directional module, barcoding of libraries, paired-end sequencing on the Illumina HiSeq 2000 platform, and demultiplexing of barcoded reads were performed.

De novo *transcriptome assembly*

Analyses were performed on the Auburn University CASIC HPC cluster. In addition to the newly generated transcriptome data, we downloaded raw reads of additional cypriniform taxa from NCBI SRA using the sra-toolkit (Table 1). Quality control checks were performed on raw reads using FastQC (Andrews 2010). Raw reads were pre-processed using Trimmomatic (Bolger et al. 2014) to trim adapters, trim the first 13 bases to eliminate random hexamer binding (HEADCROP:13), trim low quality sequence of phred score less than 20 (SLIDINGWINDOW:4:20) at fragment ends, and filter short sequences (MINLENGTH:50). Remaining reads were normalized and assembled *de novo* using Trinity version 2014-07-17 (Grabherr et al. 2011). We used TransDecoder to extract coding regions from transcripts (Haas and Papanicolaou). Redundant transcripts were removed using CD-HIT-EST v4.6.1 with the following settings: -c .99 -w 10 -r 0 (Li and Godzik 2006).

Orthology inference and sequence alignment

We used HaMStR to infer orthologs, which uses profile Hidden Markov Models from a predefined core ortholog set to extend orthology groups to target taxa (Ebersberger et al. 2009). Using biomaRt to query Ensembl (Durinck et al. 2005; Durinck et al. 2009), we generated a reference set of core ortholog groups by obtaining peptide sequences of one-to-one orthologs from Ensembl for all pairwise combinations of the following nine fish genomes: zebrafish (*Danio rerio*), Mexican cave tetra (*Astyanax mexicanus*), Atlantic cod (*Gadus morhua*), three-spined stickleback (*Gasterosteus aculeatus*), Japanese ricefish (*Oryzias latipes*), Fugu (*Takifugu rubripes*), green spotted puffer (*Tetraodon nigrividiris*), Nile tilapia (*Oreochromis niloticus*), and platyfish (*Xiphophorus maculatus*). The peptide sequence for the longest transcript for each ortholog was selected to represent each ortholog for each species. Because peptide sequences are not always available for orthologs, ortholog groups for which peptides were not found for zebrafish and for which less than seven of the nine species had sequence were excluded. This resulted in 12,458 ortholog groups, each with a unique zebrafish ortholog (with an Ensembl gene identifier) that we used as the annotation of the ortholog group for downstream functional analyses. Each ortholog group was split into individual fasta files for alignment using the emboss command seqret (Rice et al. 2000).

To generate profile Hidden Markov Models for HaMStR, reference sequences were first aligned using MAFFT v7.221 with the following settings: --auto --maxiterate 1000 --localpair (Katoh and Standley 2013). Then, pHMMs are generated from ortholog alignments internally using the hmmbuild function of HMMER 3.1b2 (Eddy 1998). HaMStR 13.2.6 was then used to extend orthologs from the reference set for each query transcriptome, which uses a combination of pHMM search using hmmsearch, orthology prediction of hmmsearch hits using blast against a reference-taxon (in this case, zebrafish) (Altschul et al. 1990), and post-processing using GeneWise to determine coding sequence and reading frame of query transcripts based on alignment to proteins of the reference taxon (Birney 2004). We used the restrictive setting in HaMStR to select a single best-hit putative ortholog from each query dataset. Query datasets include all of the above assembled cypriniform transcriptomes, as well as the CDS sequence for the *Ctenopharyngodon idella* genome (female gene models v1) (Wang et al. 2015). We were able to successfully expand all 12,458 orthologous clusters, although not all were present in all taxa.

For reference sequences from genomes from Ensembl, we used exonerate v2.2.0 to trim cDNA sequences to match protein sequences, using the protein2dna model and a score threshold of 30 (Slater and Birney 2005). These reference sequences were used as reliable sequences for the alignment using MACSE release 1.01b, a codon-aware multiple sequence alignment method (Ranwez et al. 2011). Although the GeneWise step in HaMStR should determine coding sequence based on protein alignment, we chose to treat sequences we assembled as ‘less reliable’ sequences for the MACSE alignment, and thus these sequences had a lower frame shift cost (10 vs. 30) and stop codon cost (60 vs. 100) relative to the Ensembl sequences.

Further bioinformatic processing was performed with a modified version of the bioinformatics pipelines employed by Kocot et al. (2011) and Garrison et al. (2016). Alignments from MACSE were trimmed with ALISCORE and ALICUT to remove ambiguously aligned regions, but in the current analysis, ALISCORE was performed on the amino acid alignments produced by MACSE to exclude codons from the nucleotide alignment. Any sequences of 20 or fewer base pairs surrounded by ten or more gaps on either side were also removed to further reduce potential misaligned sequence. Sequences that overlapped others by less than 20 nucleotides were removed. Redundant sequences were excluded using uniqhaplo.pl (available at http:://raven.iab.alaska/edu/ntakebay/). Codons consisting of all gaps, or with all gaps except in less than 4 taxa, were trimmed from the alignment. Trimmed alignments below 75 bp were excluded, and ortholog clusters were excluded if they did not have at least 7 taxa and did not include all three paedomorphic taxa, *Paedocypris*, *Sundadanio*, and *Danionella*.

Tests of Differential Selection and Functional Enrichment

We performed tests of differences in level of selection in paedomorphic taxa using the PAML 4.8 program codeml (Yang 2007). The codeml software fits models of codon substitution to test for positive selection, using the ratio of nonsynonymous to synonymous codon substitutions (dN/dS = ω). In addition, foreground branches can be specified to have a different ω than background branches as an alterantive hypothesis to the null hypothesis that all taxa share the same level of positive selection (a single value for ω). Finally, significance of the alternative hypothesis relative to the null can be assessed using a likelihood ratio test. For the alternative hypothesis, we assigned all three paedomorphic taxa as foreground branches, which assigns a different level of positive selection (ω) to these branches compared to the rest of the tree (background branches). We used a multifurcating topology based on the consensus of relationships from phylogenomic data as a starting tree for maximum likelihood model fitting (Fig. 1). Relationships outside of Cypriniformes are based on a consensus among phylogenomic studies based on fully sequenced genomes (Austin et al. 2015; Takezaki and Nishihara 2016). Relationships within Cypriniformes are based on a consensus of the phylogeny of Mayden & Chen (2010), anchored phylogenomic data (Stout et al. *submitted*; Chapters 2, 3), and preliminary phylogenomic analyses based on the transcriptome data (Supplementary Material). Because not all genes were present in all taxa, the phylogeny was pruned for each gene that did not have complete taxon sampling using ape in R (Paradis et al. 2004).

We tested for significance of the alternative hypothesis relative to the null hypothesis using likelihood ratio tests, with degrees of freedom equal to 1 (the difference in number of parameters between the alternative and null models). We adjusted the p-values for multiple testing based on the Benjamin-Hotchberg procedure (Wright 1992; Benjamini and Hochberg 1995) as implemented in R using the function p.adjust. Adjusted p-values have an intuitive interpretation; a test that is significant at an adjusted p-value of 0.05 is also significant at a false-discovery rate cut-off of 0.05. We used $p < 0.05$ as our cut-off for significance.

We focused on two broad categories of genes where a difference in level of selection could be related to paedomorphic phenotype. First, a list of genes related to bone development were derived from Venkatesh et al. (2014), who provided a list of human protein RefSeq accession numbers. We supplemented their list with genes containing the word 'bone' in their gene ontology (GO) terms, which added an additional 17 genes for which we determined single-copy orthologs. Secondly, a list of genes related to growth was derived by determining gene ontology functional annotations containing the word 'growth.' We determined if this level of selection on foreground taxa in each group of genes was significantly differentially selected relative to the remaining genes tested using a Mann-Whitney U test (i.e. Wilcoxon rank-sum test).

We compiled a list of the significantly differentially selected (SDS) genes and a list of all genes tested, then performed tests of enrichment using DAVID 6.8 (Dennis et al. 2003; Huang et al. 2007; Huang et al. 2008). We used the zebrafish Ensembl identifiers for each ortholog group. DAVID performs tests of enrichment to determine if certain, for example, functional categories are significantly over-represented among a focal gene group (e.g. SDS genes) relative to the background list (e.g. all single-copy orthologs tested for differential selection). We first tested if certain zebrafish chromosomes were enriched for SDS genes, to test if selection was different across chromosomes. Second, we tested for enrichment of functional categories. Functional categories are simultaneously queried and tested for enrichment from annotations originating from a variety of databases including SwissProt keywords (Bairoch and Boeckmann 1991), UniProt features (The UniProt Consortium 2007), Gene Ontology terms (Harris et al. 2004), KEGG pathways (Kanehisa et al. 2004), and protein domains from InterPro (Apweiler et al. 2001), Protein Information Resource (Barker et al. 2000), and SMART (a simple modular architecture research tool; Schultz et al. 1998). We also determined over-represented anatomical parts affected by SDS genes using DAVID, which is based on knowledge of genotype-phenotype interactions in zebrafish originating from ZFIN (Sprague et al. 2003; Bradford et al. 2011).

ACKNOWLEDGEMENTS

We would like to thank Tonia S. Schwartz and Nicole L. Garrison for information and help on analyses for analyses used in this work. We would like to acknowledge Wet Spot Tropical Fish and Frank M. Greco for assistance in procuring specimens. We would like to thank Bradley Morgan with his work related to maintaining the Auburn University CASIC HPC cluster, including maintenance and installation of software. This work was supported by Auburn University (Cellular and Molecular Biosciences Peaks of Excellence Research Fellowship to MT); and by National Science Foundation (DEB-0315963, DEB-0107751, DEB-1022720) to JWA.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403–410.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, et al. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29:37–40.

Austin CM, Tan MH, Croft LJ, Hammer MP, Gan HM. 2015. Whole Genome Sequencing of the Asian Arowana (*Scleropages formosus*) Provides Insights into the

Evolution of Ray-Finned Fishes. Genome Biology and Evolution 7:2885–2895.

Bairoch A, Boeckmann B. 1991. The SWISS-PROT protein sequence data bank. Nucleic Acids Res.

Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao C, Yeh L-SL, Ledley RS, Janda JF, et al. 2000. The Protein Information Resource (PIR). Nucleic Acids Res 28:41–44.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57:289–300.

Birney E. 2004. GeneWise and Genomewise. Genome Res 14:988–995.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Bradford Y, Conlin T, Dunn N, Fashena D, Frazer K, Howe DG, Knight J, Mani P, Martin R, Moxon SAT, et al. 2011. ZFIN: enhancements and updates to the Zebrafish Model Organism Database. Nucleic Acids Res 39:D822–D829.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. GigaScience 2:10.

Britz R, Conway KW, Rüber L. 2009. Spectacular morphological novelty in a miniature cyprinid fish, *Danionella dracula* n. sp. Proc. R. Soc. London Ser. B 276:2179–2186.

Britz R, Conway KW. 2009. Osteology of *Paedocypris*, a miniature and highly developmentally truncated fish (Teleostei: Ostariophysi: Cyprinidae). J. Morphol. 270:389–412.

Britz R, Conway KW. 2016. *Danionella dracula*, an escape from the cypriniform *Bauplan* via developmental truncation? J. Morphol. 277:147–166.

Collins JE, White S, Searle SMJ, Stemple DL. 2012. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. Genome Res 22:2067–2078.

Conway KW, Kottelat M, Tan HH. 2011. Review of the Southeast Asian miniature cyprinid genus *Sundadanio* (Ostariophysi: Cyprinidae) with descriptions of seven new species from Indonesia. Ichthyol Explor Freshwaters 22:251–288.

Darwin C. 1882. On the Origin of Species by Means of Natural Selection.

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane H, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4:R60–11.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21:3439–3440.

Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 4:1184–1191.

Ebersberger I, Strauss S, Haeseler von A. 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol 9:157.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

Edmunds RC, Su B, Balhoff JP, Eames BF, Dahdul WM, Lapp H, Lundberg JG, Vision TJ, Dunham RA, Mabee PM, et al. 2015. Phenoscape: Identifying Candidate Genes for Evolutionary Phenotypes. Mol Biol Evol 33:13–24.

Garrison NL, Rodriguez J, Agnarsson I, Coddington JA, Griswold CE, Hamilton CA, Hedin M, Kocot KM, Ledford JM, Bond JE. 2016. Spider phylogenomics: untangling the Spider Tree of Life. PeerJ 4:e1719.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652.

Gregory TR. 2005. Synergy between sequence and size in large-scale genomics. Nature Reviews Genetics 6:699–708.

Haas B, Papanicolaou A. TransDecoder (Find Coding Regions Within Transcripts). Available from: http://transdecoder.github.io

Haffter P, Granato M, Brand M, Mullins MC, Nüsslein-Volhard C, Hammerschmidt M, Kane DA, Odenthal J, van Eeden FJM, Jiang YJ, et al. 1996. The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. Development 123:1–36.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: Evo devo and the genetics of adaptation. Evolution 61:995–1016.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503.

Huang DW, Sherman BT, Lempicki RA. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57.

Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, et al. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res 35:W169–W175.

Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946–957.

Johnson GD, Brothers EB. 1993. *Schindleria*: A Paedomorphic Goby (Teleostei: Aobioidei). BMS.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277–D280.

Kardong KV. 2014. Vertebrates: Comparative Anatomy, Function, Evolution. 7 ed. McGraw-Hill Education

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780.

Kimmel CB, Miller CT, Moens CB. 2001. Specification and Morphogenesis of the Zebrafish Larval Head Skeleton. Developmental Biology 233:239–257.

Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, et al. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477:452–456.

Kottelat M, Britz R, Tan HH, Witte K-E. 2006. *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. Proc. R. Soc. London Ser. B 273:895–899.

Koudijs MJ, Broeder den MJ, Keijser A, Wienholds E, Houwing S, van Rooijen EMHC, Geisler R, van Eeden FJM. 2005. The Zebrafish Mutants dre, uki, and lep Encode Negative Regulators of the Hedgehog Signaling Pathway. PLoS Genet 1:e19–12.

Leggatt RA, Iwama GK. 2003. Occurrence of polyploidy in the fishes. Reviews in Fish Biology and Fisheries 13:237–246.

Li J-T, Hou G-Y, Kong X-F, Li C-Y, Zeng J-M, Li H-D, Xiao G-B, Li X-M, Sun X-W. 2015. The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (Cyprinus carpio). Sci. Rep. 5:8199.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Liu S, Tan HH, Tan SL, Yunhan H. 2012. Chromosome Evolution and Genome Miniaturization in Minifish. PLOS ONE 7:e37305.

Liu Z, Liu S, Yao J, Bao L, Zhang J, Li Y, Jiang C, Sun L, Wang R, Zhang Y, et al. 2016. The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nat Commun 7:1–13.

Mayden RL, Chen WJ. 2010. The world"s smallest vertebrate species of the genus *Paedocypris*: a new family of freshwater fishes and the sister group to the world"s most diverse clade of freshwater fishes (Teleostei: Cypriniformes). Mol Phylogenet Evol 57:152–175.

Meng F, Braasch I, Phillips JB, Lin X, Titus T, Chungguang Z, Postlethwait JH. 2013. Evolution of the eye transcriptome under constant darkness in *Sinocyclocheilus* cavefish. Mol Biol Evol 30:1527–1543.

Meyer A, Biermann CH, Ortí G. 1993. The phylogenetic position of the zebrafish (Danio rerio), a model system in developmental biology: an invitation to the comparative method. Proc. R. Soc. London Ser. B 252:231–236.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.

Parsons KJ, Albertson RC. 2009. Roles for Bmp4 and CaM1 in Shaping the Jaw: Evo-Devo and Beyond. Annu. Rev. Genet. 43:369–388.

Prasanth KV, Spector DL. 2007. Eukaryotic regulatory RNAs: an answer to the "genome complexity" conundrum. Genes & Development 21:11–42.

Raff RA. 2000. Evo-devo: the evolution of a new discipline. Nature Reviews Genetics 1:74–79.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons.Murphy WJ, editor. PLOS ONE 6:e22594.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277.

Roberts TR. 1986. *Danionella translucida*, a new genus and species of cyprinid fish from Burma, one of the smallest living vertebrates. Env Biol Fish 16:231–241.

Schier AF, Driever W, Solnica-Krezel L, Neuhauss SCF, Malicki J, Stemple DL, Stainier DYR, Zwartkruis F, Abdelilah S, Rangini Z, et al. 1997. A genetic screen for mutations affecting embryogenesis in zebrafish. Development 123:37–46.

Schindler O. 1932. Sexually mature larval Hemirhamphidae from the Hawaiian Islands. Bernice P. Bishop Museum Bulletin 197:1–28.

Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, Rimbault M, Decker B, Kidd JM, Sood R, et al. 2012. Variation of BMP3 Contributes to Dog Breed Skull Diversity.Leeb T, editor. PLoS Genet 8:e1002849–11.

Schultz J, Milpetz F, Bork P, Ponting CP. 1998. SMART, a simple modular architecture research tool: Identification of signaling domains. Proc Natl Acad Sci USA 95:5857–5864.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics.

Sprague J, Clements D, Conlin T, Edwards P, Frazer K, Schaper K, Segerdell E, Song P, Sprunger B, Westerfield M. 2003. The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Res 31:241–243.

Takezaki N, Nishihara H. 2016. Resolving the Phylogenetic Position of Coelacanth: The Closest Relative Is Not Always the Most Appropriate Outgroup. Genome Biology and Evolution 8:1208–1221.

The UniProt Consortium. 2007. The Universal Protein Resource (UniProt). Nucleic Acids Res 36:D190–D195.

Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, Ohta Y, Flajnik MF, Sutoh Y, Kasahara M, et al. 2014. Elephant shark genome provides unique insights into gnathostome evolution. Nature 505:174–179.

Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, Lu H, Huang R, Xia X, Feng Q, et al. 2015. The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation. Nature Genetics 47:625–631.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10:57–63.

Wäsch R, Robbins JA, Cross FR. 2010. The emerging role of APC/CCdh1 in controlling differentiation, genomic stability and tumor suppression. Oncogene 29:1–10.

Weitzman SH, Vari RP. 1988. Miniaturization in South American freshwater fishes; and overview and discussion. Proc Biol Soc Wash 101:444–465.

Wright SP. 1992. Adjusted p-values for simultaneous inference. Biometrics 48:1005.

Xu P, Zhang X, Wang X, Li J-T, Liu G, Kuang Y, Xu J, Zheng X, Ren L, Wang G, et al. 2014. Genome sequence and genetic diversity of the common carp, Cyprinus carpio. Nature Genetics 46:1212–1219.

Yang L, Sado T, Hirt MV, Pasco-Viel E, Arunachalam M, Li J, Wang X, Freyhof J, Saitoh K, Simons AM, et al. 2015. Phylogeny and polyploidy: Resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). Mol Phylogenet Evol

85:97–116.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24:1586–1591.

Zhu C, Tong J, Yu X, Guo W. 2015. Comparative mapping for bighead carp (Aristichthys nobilis) against model and non-model fishes provides insights into the genomic evolution of cyprinids. Mol Genet Genomics 290:1313–1326.

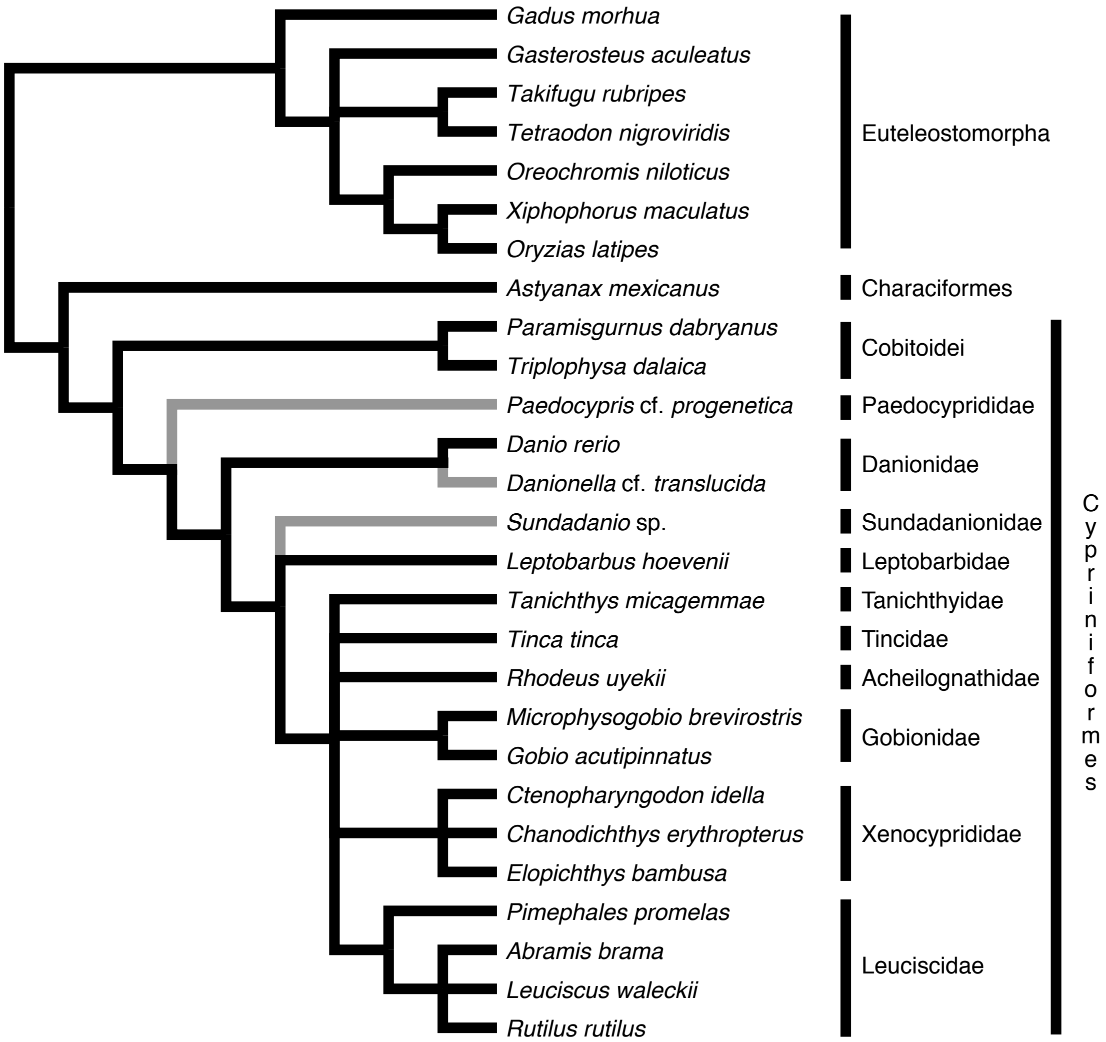


Figure 1. Cladogram of teleosts with a focus on Cypriniformes used as a starting tree for tests of positive selection. Grey branches lead to paedomorphic taxa and were set as foreground branches in testing for differential positive selection. The topology is conservatively multifurcating, allowing for some uncertainty in topological relationships.

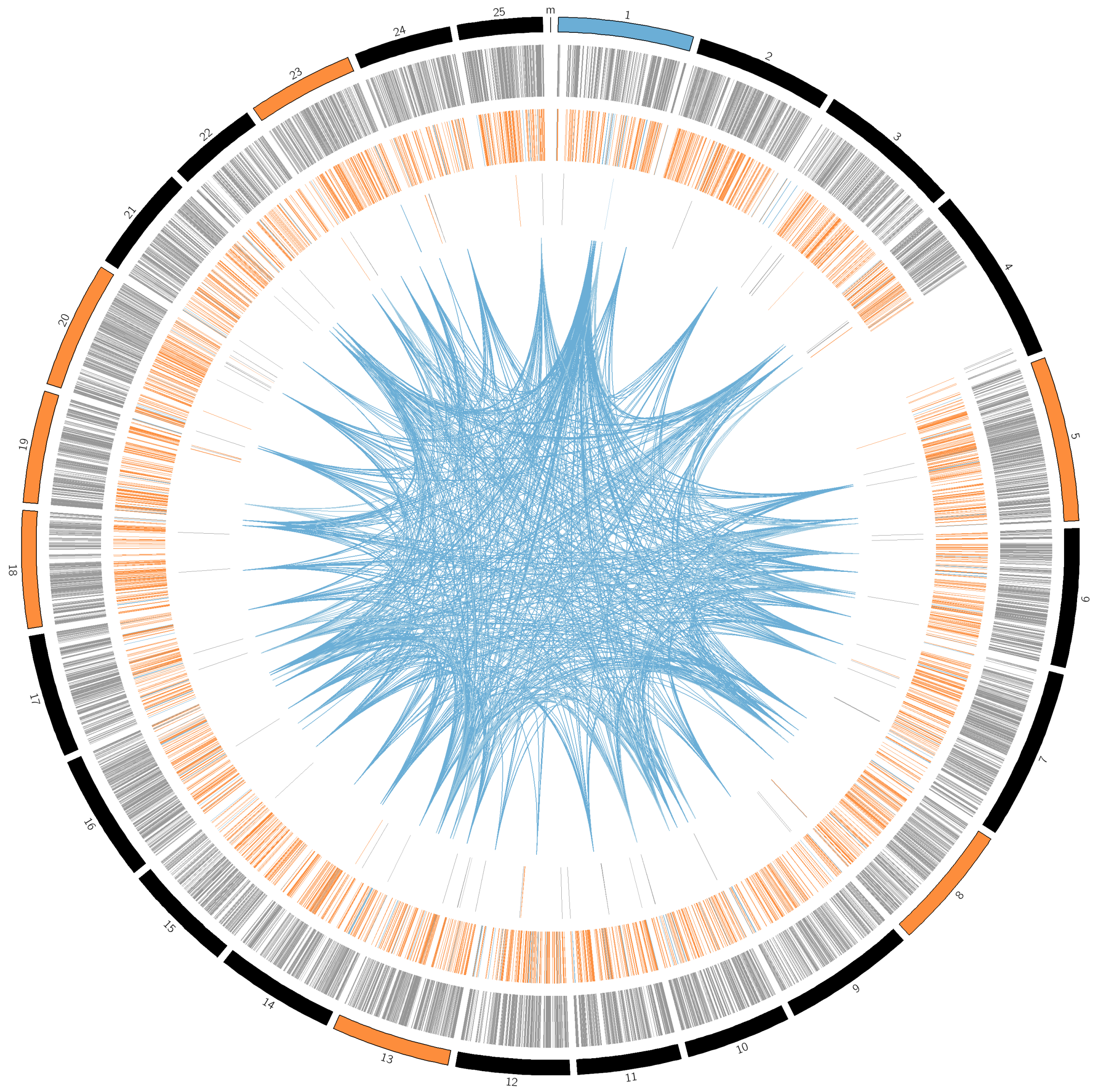


Figure 2. Caption on following page.

Figure 2. Circos plot displaying genomic locations relative to the zebrafish genome for results of tests for positive selection and functional enrichment. From the outside in: 1) zebrafish chromosomes, starting with the mitochondrial genome, followed by chromosomes 1-25; chromosomes highlighted if they have a significant enrichment for genes undergoing greater purifying selection (orange) or positive selection (blue) in paedomorphic taxa relative to non-paedomorphic teleosts; 2) level of selection inferred with a single rate across all taxa, with darker grey indicating higher level of selection (null model of selection); 3) relative level of selection on genes that were significantly different between the paedomorphic taxa and non-paedomorphic teleosts, either undergoing greater purifying selection (red) or greater positive selection (blue) (alternative model of selection); 4) relative level of selection on genes between paedomorphic taxa and non-paedomorphic teleosts of bone genes, either undergoing greater purifying selection (red), greater positive selection (blue), or no differential selection (grey). Blue links between genes display shared enriched gene function annotations and zebrafish anatomical annotations for genes undergoing greater positive selection, corresponding to annotation networks of Figures 5 and 6.

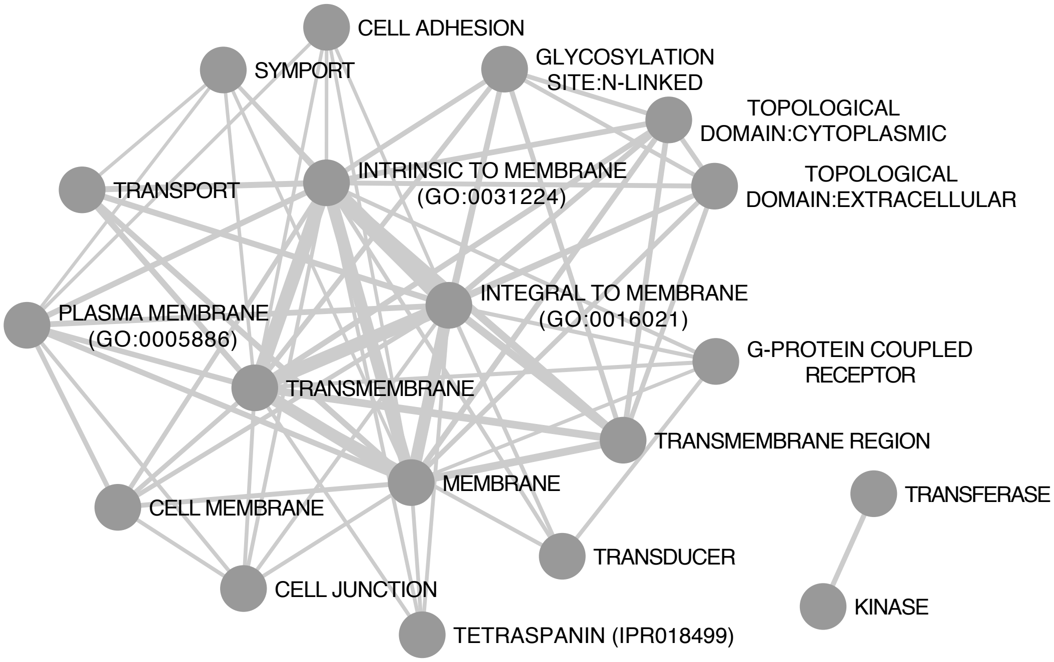

CELL ADHESION
SYMPORT
GLYCOSYLATION SITE:N-LINKED
TOPOLOGICAL DOMAIN:CYTOPLASMIC
TRANSPORT
INTRINSIC TO MEMBRANE (GO:0031224)
TOPOLOGICAL DOMAIN:EXTRACELLULAR
INTEGRAL TO MEMBRANE (GO:0016021)
PLASMA MEMBRANE (GO:0005886)
G-PROTEIN COUPLED RECEPTOR
TRANSMEMBRANE
TRANSMEMBRANE REGION
MEMBRANE
TRANSFERASE
CELL MEMBRANE
TRANSDUCER
CELL JUNCTION
KINASE
TETRASPANIN (IPR018499)

Figure 3. Functional enrichment for genes under greater purifying selection in paedomorphic cypriniforms relative to other teleosts, p < .05, FDR < .05. Node sizes are relative to the number of genes in each functional category and edge widths are relative to the number of shared genes between categories.

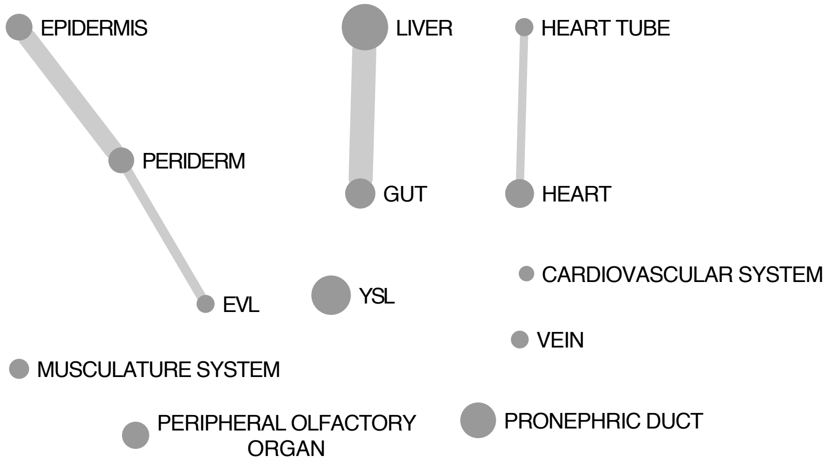


Figure 4. Functional anatomical enrichment for genes under greater purifying selection in paedomorphic cypriniforms relative to other teleosts, p < .05, FDR < .05. Node sizes are relative to the number of genes in each functional category and edge widths are relative to the number of shared genes between categories.

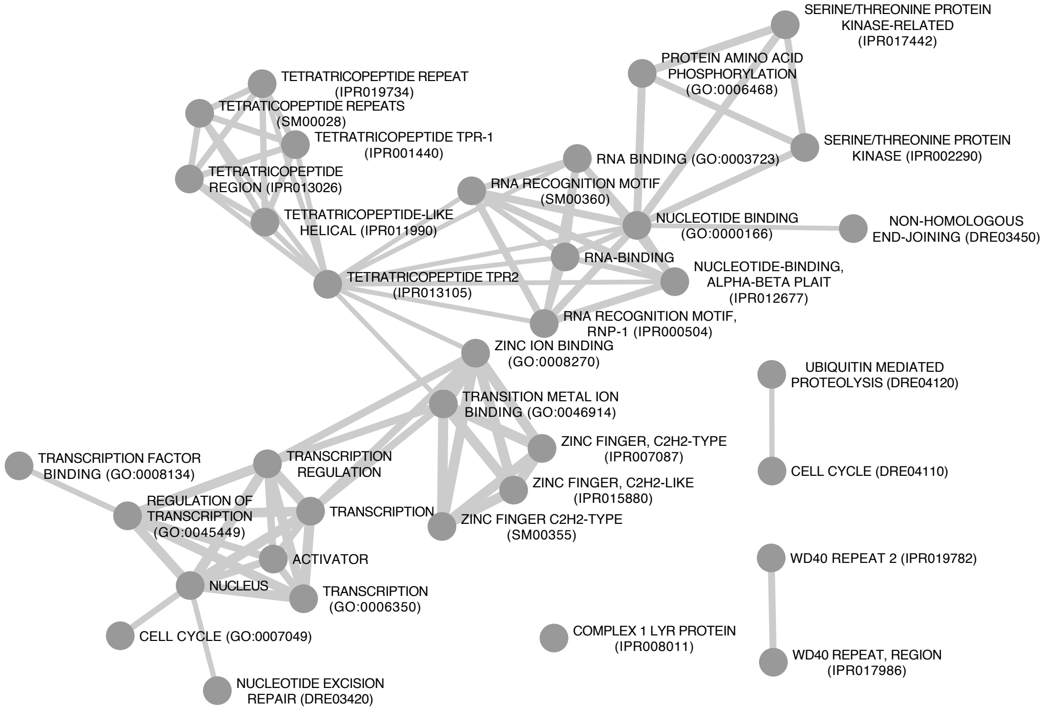

SERINE/THREONINE PROTEIN KINASE-RELATED (IPR017442)
PROTEIN AMINO ACID PHOSPHORYLATION (GO:0006468)
TETRATRICOPEPTIDE REPEAT (IPR019734)
TETRATICOPEPTIDE REPEATS (SM00028)
TETRATRICOPEPTIDE TPR-1 (IPR001440)
SERINE/THREONINE PROTEIN KINASE (IPR002290)
RNA BINDING (GO:0003723)
TETRATRICOPEPTIDE REGION (IPR013026)
RNA RECOGNITION MOTIF (SM00360)
TETRATRICOPEPTIDE-LIKE HELICAL (IPR011990)
NUCLEOTIDE BINDING (GO:0000166)
NON-HOMOLOGOUS END-JOINING (DRE03450)
RNA-BINDING
NUCLEOTIDE-BINDING, ALPHA-BETA PLAIT (IPR012677)
TETRATRICOPEPTIDE TPR2 (IPR013105)
RNA RECOGNITION MOTIF, RNP-1 (IPR000504)
ZINC ION BINDING (GO:0008270)
UBIQUITIN MEDIATED PROTEOLYSIS (DRE04120)
TRANSITION METAL ION BINDING (GO:0046914)
ZINC FINGER, C2H2-TYPE (IPR007087)
TRANSCRIPTION FACTOR BINDING (GO:0008134)
TRANSCRIPTION REGULATION
CELL CYCLE (DRE04110)
ZINC FINGER, C2H2-LIKE (IPR015880)
REGULATION OF TRANSCRIPTION (GO:0045449)
TRANSCRIPTION
ZINC FINGER C2H2-TYPE (SM00355)
ACTIVATOR
WD40 REPEAT 2 (IPR019782)
NUCLEUS
TRANSCRIPTION (GO:0006350)
CELL CYCLE (GO:0007049)
COMPLEX 1 LYR PROTEIN (IPR008011)
WD40 REPEAT, REGION (IPR017986)
NUCLEOTIDE EXCISION REPAIR (DRE03420)

Figure 5. Functional categories from DAVID for all genes under greater positive selection in paedomorphic cypriniforms relative to other teleosts. Note there are no significantly enriched categories at p < .05 and FDR < .05. Node sizes are relative to the number of genes in each functional category and edge widths are relative to the number of shared genes between categories.

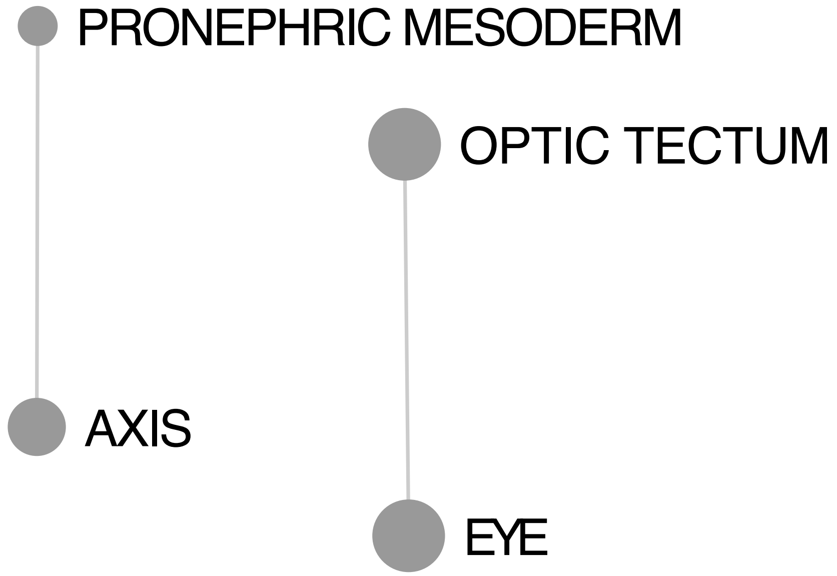


Figure 6. Functional anatomical annotations for all genes under greater positive selection in paedomorphic cypriniforms relative to other teleosts. Note there are no significantly enriched categories at p < .05 and FDR < .05. Node sizes are relative to the number of genes in each functional category.

Table 1. Transcriptome statistics. Total clusters indicates number of unique transcript clusters assembled by Trinity. Non-redundant sequences corresponds to the number of sequences after redundant sequences were excluded by CD-HIT-EST. Single-copy orthologs identified corresponds to the number out of the 12,457 single-copy orthologs identified in each transcriptome using HaMStR. Orthologs analyzed corresponds to the number out of the 8,687 genes used in tests of positive selection for each species. *Ctenopharyngodon idella* gene models were downloaded from official genome website (http://www.ncgr.ac.cn/grasscarp/).

Species	Accessions*	Total assembly size (bp)	Total clusters	Non-redundant sequences	Single-copy orthologs identified	Orthologs analyzed
This study						
Danionella cf. *translucida*	N/A	155,728,530	131,541	42,695	11,588	8,687
Leptobarbus hoevenii	N/A	189,023,600	184,361	49,526	11,527	8,336
Paedocypris cf. *progenetica*	N/A	117,593,428	100,906	41,198	10,812	8,687
Sundadanio sp.	N/A	94,645,990	80,119	33,601	10,974	8,687
Tanichthys micagemmae	N/A	182,413,513	170,201	47,527	11,600	8,389
Previously sequenced						
Abramis brama	SRR1752897	151,938,099	155,688	44,311	10,527	7,274
Chanodichthys erythropterus	SRR2179924, SRR2179946, SRR2182152, SRR2182178, SRR3336604	228,426,137	193,700	65,575	11,640	8,054
Ctenopharyngodon idella	–	–	–	30,266	11,634	8,241
Elopichthys bambusa	SRR886276	136,767,001	147,250	44,311	10,579	7,407
Gobio acutipinnatus	SRR1660441	221,663,564	183,936	56,082	11,260	7,962
Leuciscus waleckii	SRR949612	65,937,706	97,108	24,844	8,483	6,030
Microphysogobio brevirostris	SRR1185341	80,156,283	117,293	38,500	9,894	6,981
Paramisgurnus dabryanus	SRR1652368, SRR1652322, SRR1652342	130,248,762	125,306	45,876	10,437	7,485
Pimephales promelas	SRR1582202	144,494,376	105,993	40,331	10,812	7,996
Rhodeus uyekii	SRR2043485, SRR2043486	276,266,590	241,951	50,163	10,686	7,916
Rutilus rutilus	SRR1776878	276,266,590	241,951	60,667	11,149	7,843
Tinca tinca	SRR1622030	83,818,531	94,000	29,678	8,370	6,011
Triplophysa dalaica	SRR1698202	198,029,199	127,780	45,924	10,682	7,812

Table 2. Results of statististical tests for positive selection for genes related to bone formation.

Ensembl Gene ID	Gene Name	Log Likelihood of Null Model	Parameter # in Null Model	ω under null model	Log Likelihood of Alternative Model	Parameter # in Alternativ e Model	Background ω under alternative model	Foreground ω under alternative model	p-value	Adjusted p-value
ENSDARG00000005789	enpp1	-31300.9183	44	0.18004	-31291.63596	45	0.1906	0.13265	1.6423E-05	1.8797E-04
ENSDARG00000008107	src	-12607.88356	42	0.03192	-12601.42126	43	0.03387	0.00862	3.2429E-04	2.1055E-03
ENSDARG00000011490	scube3	-5872.718631	40	0.05179	-5868.966493	41	0.04804	0.09555	6.1553E-03	2.2331E-02
ENSDARG00000012066	dcn	-9879.445092	46	0.08723	-9875.939726	47	0.09304	0.06074	8.1023E-03	2.7613E-02
ENSDARG00000015686	bmp6	-9229.36766	40	0.09175	-9224.345333	41	0.09876	0.05946	1.5279E-03	7.3903E-03
ENSDARG00000016086	smurf1	-11969.1731	42	0.01921	-11962.4117	43	0.02142	0.00836	2.3568E-04	1.5995E-03
ENSDARG00000019353	sparc	-6336.412979	46	0.08354	-6332.919294	47	0.09022	0.0545	8.2087E-03	2.7866E-02
ENSDARG00000019646	twist3	-3999.416018	43	0.02779	-3995.369524	44	0.03188	0.01011	4.4437E-03	1.7210E-02
ENSDARG00000020007	col1a2	-6404.407925	34	0.23358	-6395.370655	35	0.26797	0.10921	2.1242E-05	2.3022E-04
ENSDARG00000026811	extl3	-18890.99093	44	0.0268	-18884.19425	45	0.02987	0.0182	2.2699E-04	1.5502E-03
ENSDARG00000027552	mapk1	-6851.442387	44	0.02491	-6847.594307	45	0.02814	0.01408	5.5338E-03	2.0500E-02
ENSDARG00000028071	bmp1a	-18402.75242	41	0.03994	-18397.4703	42	0.04294	0.02851	1.1530E-03	5.9055E-03
ENSDARG00000030215	matn1	-5296.290694	38	0.04878	-5291.139564	39	0.05724	0.02263	1.3287E-03	6.6229E-03
ENSDARG00000031894	lef1	-6240.872452	41	0.05906	-6237.276049	42	0.06444	0.03527	7.3196E-03	2.5506E-02
ENSDARG00000039577	ptk2bb	-32011.9134	43	0.08413	-31998.25657	44	0.08981	0.05071	1.7299E-07	4.3182E-06
ENSDARG00000045071	chad	-9120.876385	37	0.1072	-9109.299844	38	0.12039	0.05858	1.4960E-06	2.5940E-05
ENSDARG00000045802	hapln3	-10459.92246	41	0.12227	-10453.2472	42	0.13244	0.07989	2.5835E-04	1.7331E-03
ENSDARG00000056152	fam3c	-8285.416538	46	0.16531	-8277.17938	47	0.1806	0.08743	4.9314E-05	4.5525E-04
ENSDARG00000060397	hhip	-16433.99267	44	0.09598	-16430.44727	45	0.0906	0.11923	7.7481E-03	2.6625E-02
ENSDARG00000069463	alox12	-17206.555	41	0.13479	-17200.79987	42	0.14334	0.10131	6.9213E-04	3.9195E-03

Table 3. Results of statististical tests for positive selection for genes that contain the term growth in their GO description.

Ensembl Gene ID	Gene Name	Log Likelihood of Null Model	Parameter # in Null Model	ω under null model	Log Likelihood of Alternative Model	Parameter # in Alternativ e Model	Background ω under alternative model	Foreground ω under alternative model	p-value	Adjusted p-value
ENSDARG00000008107	src	-12607.88356	42	0.03192	-12601.42126	43	0.03387	0.00862	3.2429E-04	2.1055E-03
ENSDARG00000008388	mmp14b	-14481.54342	44	0.09729	-14470.69751	45	0.10474	0.05387	3.2015E-06	4.8537E-05
ENSDARG00000009021	chrna1	-7920.171669	36	0.04417	-7916.672448	37	0.04795	0.02798	8.1581E-03	2.7748E-02
ENSDARG00000010207	smad3b	-12019.43741	46	0.02779	-12016.12714	47	0.02941	0.01717	1.0081E-02	3.2762E-02
ENSDARG00000011496	ppm1bb	-9702.513032	46	0.06498	-9695.174148	47	0.07099	0.0346	1.2754E-04	9.7274E-04
ENSDARG00000014907	htra1b	-14859.29674	46	0.08788	-14852.93253	47	0.0942	0.05274	3.6014E-04	2.2923E-03
ENSDARG00000015472	gpc4	-15194.57905	46	0.09837	-15189.89236	47	0.10405	0.0733	2.2016E-03	9.8584E-03
ENSDARG00000015686	bmp6	-9229.36766	40	0.09175	-9224.345333	41	0.09876	0.05946	1.5279E-03	7.3903E-03
ENSDARG00000016086	smurf1	-11969.1731	42	0.01921	-11962.4117	43	0.02142	0.00836	2.3568E-04	1.5995E-03
ENSDARG00000016623	si:ch211-195b13.1	-9320.217553	43	0.05975	-9316.091982	44	0.0639	0.0372	4.0726E-03	1.6100E-02
ENSDARG00000017367	rhbdf1b	-12275.57214	35	0.10294	-12271.83848	36	0.11022	0.07752	6.2829E-03	2.2647E-02
ENSDARG00000019367	tgfb3	-9996.547955	45	0.05269	-9972.938137	46	0.06205	0.019	6.3461E-12	6.6420E-10
ENSDARG00000020072	thbs4b	-25838.93208	42	0.03786	-25822.02559	43	0.04286	0.02092	6.0673E-09	2.4629E-07
ENSDARG00000026811	extl3	-18890.99093	44	0.0268	-18884.19425	45	0.02987	0.0182	2.2699E-04	1.5502E-03
ENSDARG00000027087	tgfb2	-7936.793739	39	0.04847	-7930.478549	40	0.05466	0.03084	3.7953E-04	2.3995E-03
ENSDARG00000027290	nrp1b	-22102.43648	41	0.11561	-22088.81506	42	0.12766	0.08016	1.7944E-07	4.4158E-06
ENSDARG00000034434	igf1rb	-10717.18344	34	0.10756	-10702.64517	35	0.12231	0.05999	6.9574E-08	2.0214E-06
ENSDARG00000034541	tgfbr2	-16084.69079	44	0.07855	-16073.8649	45	0.08565	0.04682	3.2691E-06	4.9357E-05
ENSDARG00000034700	vegfab	-5244.829843	44	0.18361	-5228.912953	45	0.21693	0.06846	1.6795E-08	5.9549E-07
ENSDARG00000035056	fgf13a	-4786.867753	41	0.06015	-4773.747889	42	0.07198	0.01015	3.0155E-07	6.8397E-06
ENSDARG00000035563	znf703	-9996.071993	43	0.08989	-9990.102748	44	0.09851	0.05883	5.4986E-04	3.2694E-03
ENSDARG00000035899	lingo1b	-13283.63731	34	0.07434	-13277.73077	35	0.07927	0.04563	5.8815E-04	3.4353E-03
ENSDARG00000036541	rhbdf1a	-21997.60542	46	0.05004	-21993.21205	47	0.05342	0.03684	3.0343E-03	1.2740E-02
ENSDARG00000037238	smad5	-8660.635398	46	0.01756	-8657.231966	47	0.01884	0.00712	9.0808E-03	3.0178E-02
ENSDARG00000039577	ptk2bb	-32011.9134	43	0.08413	-31998.25657	44	0.08981	0.05071	1.7299E-07	4.3182E-06

ENSDARG00000041449	spred1	-10637.76886	45	0.07141	-10632.84349	46	0.07387	0.01741	1.6976E-03	7.9865E-03
ENSDARG00000045557	socs2	-4471.905506	44	0.08774	-4468.39305	45	0.09563	0.04958	8.0383E-03	2.7449E-02
ENSDARG00000051746	cecr1a	-17015.73123	46	0.18326	-17006.82005	47	0.19613	0.12531	2.4252E-05	2.5475E-04
ENSDARG00000052279	osgn1	-11713.59801	41	0.08275	-11709.90121	42	0.08895	0.06258	6.5456E-03	2.3381E-02
ENSDARG00000053939	tgfa	-2628.160357	40	0.06458	-2625.002009	41	0.07409	0.02264	1.1961E-02	3.7751E-02
ENSDARG00000055136	si:dkey-101k6.5	-9390.269276	37	0.11534	-9381.00301	38	0.1296	0.06681	1.6703E-05	1.8992E-04
ENSDARG00000060526	bmp3	-8464.325676	39	0.06094	-8460.290103	40	0.05578	0.0851	4.4976E-03	1.7372E-02
ENSDARG00000061213	rabep2	-11054.67046	44	0.16007	-11045.07419	45	0.17696	0.09844	1.1817E-05	1.4554E-04
ENSDARG00000070617	vhl	-5809.464437	43	0.18382	-5806.347354	44	0.19822	0.1277	1.2531E-02	3.9171E-02
ENSDARG00000070914	dusp6	-8696.801564	46	0.01856	-8692.554016	47	0.0209	0.00976	3.5610E-03	1.4489E-02
ENSDARG00000075593	trim71	-14824.5989	37	0.03805	-14816.17929	38	0.042	0.02326	4.0684E-05	3.8880E-04
ENSDARG00000079306	rlim	-9377.956636	41	0.0729	-9371.491898	42	0.07982	0.04433	3.2345E-04	2.1030E-03
ENSDARG00000079862	kl	-22335.96311	36	0.07461	-22331.10783	37	0.07885	0.0583	1.8321E-03	8.5143E-03
ENSDARG00000086778	pdgfba	-4901.20253	41	0.15563	-4896.490351	42	0.17471	0.09406	2.1412E-03	9.6527E-03
ENSDARG00000103403	sar1b	-4431.874189	46	0.03867	-4425.278443	47	0.04434	0.01229	2.8122E-04	1.8592E-03
ENSDARG00000104039	errfi1	-10043.37082	43	0.17069	-10040.4944	44	0.18115	0.13317	1.6462E-02	4.8908E-02

Supplementary Material

Preliminary Transcriptome Phylogeny for Cypriniformes

A preliminary maximum-likelihood phylogeny was inferred based on transcriptome data. From the 8,867 single copy orthologs (i.e. after loci were excluded due to low taxon sampling or exclusion of any of the paedomorphic taxa; Chapter 4), we concatenated the loci using FASconCAT-G (Kück and Longo 2014). Next, to optimize the dataset for phylogenetic information, we reduced the matrix using MARE (Misof et al. 2014), modified to calculate phylogenetic information based on nucleotide sequence data (Chapter 3). This resulted in 5,674 remaining loci, and no taxa were excluded. The alignment (6,110,544 base pairs) was partitioned by gene, with each partition having the GTR+G model, and the maximum-likelihood tree (Supp. Fig. 1) was inferred using ExaML version 3.0.17 (Kozlov et al. 2015).

A major difference between the anchored phylogenomic tree (Chapters 2,3) and the inferred transcriptome phylogeny is the placement of *Paedocypris* as sister to Cypriniformes. Although this has been hypothesized before (Mayden and Chen 2010), we hypothesize this recovered relationship is an artefact due to the absence of *Gyrinocheilus* (Gyrinocheilidae), which has not yet been sampled for transcriptome data. In anchored phylogenomic analyses, we recover *Paedocypris* as sister to Cyprinoidei; however, when *Gyrinocheilus* is excluded, the placement of *Paedocypris* shifts to sister to the remaining members of Cypriniformes due to long branch attraction towards the outgroup (Chapter 3). Thus, the placement of *Paedocypris* cannot be confidently assessed without increasing taxon sampling for transcriptome data, in particular *Gyrinocheilus*, and potentially other outgroup taxa.

The transcriptome phylogeny is consistent with the anchored phylogenomics tree in recovering an early-branching position for Danionidae, and a sister relationship between *Sundadanio* (Sundadanionidae) and a large clade of cyprinoids formed by Tanichthyidae, Leuciscidae, Xenocyprididae, Gobionidae, and Acheilognathidae. Also, the recovery of Xenocyprididae as sister to a clade formed by Tanichthyidae, Tincidae, Gobionidae, Acheilognathidae and Leuciscidae from the transcriptome-based analysis is consistent with in Mayden & Chen (2010) and anchored phylogenomics analyses (Chapters 2, 3). Acheilognathidae and Gobionidae are supported as closely related in both anchored phylogenomic data and transcriptomic data, although this was not recovered by Mayden & Chen (2010). Some of the other differences in the relationships of cyprinoids can likely be attributed to short branch lengths between many of the cyprinoid families, resulting in little phylogenetic signal to reconstruct the relative relationships among cyprinoid families, even for phylogenomic data.

The relationships of *Leptobarbus* (Leptobarbidae) and *Tinca* (Tincidae) recovered from transcriptome data cannot be compared with the anchored phylogenomics tree because they were not sampled. *Leptobarbus* was recovered as sister to *Sundadanio* by Mayden & Chen (2010), but *Leptobarbus* was recovered as sister to the clade formed by Xenocyprididae, Tanichthyidae, Tincidae, Gobionidae, Acheilognathidae, and Leuciscidae in the transcriptome phylogeny. *Tinca* was recovered with poorly-resolved relationship with *Tanichthys* and Leuciscidae by Mayden & Chen (2010), but with 100% bootstrap support as sister to Xenocyprididae by Tao et al. (2013). In the transcriptome phylogeny, *Tinca* is recoverd as sister to Acheilognathidae.

Except for recovery of a close relationship of Acheilognathidae and Gobionidae, the relative relationships of Tanichthyidae, Acheilognathidae, Gobionidae, and Leuciscidae all differ between Mayden & Chen (2010), the anchored phylogenomics tree, and the transcriptome phylogeny. In the anchored phylogenomics tree, *Tanichthys* is recovered as sister to Leuciscidae, however in the transcriptome phylogeny, *Tanichthys* is recovered as sister to a clade formed by Gobionidae, Acheilognathidae, Tincidae, and Leuciscidae.

REFERENCES

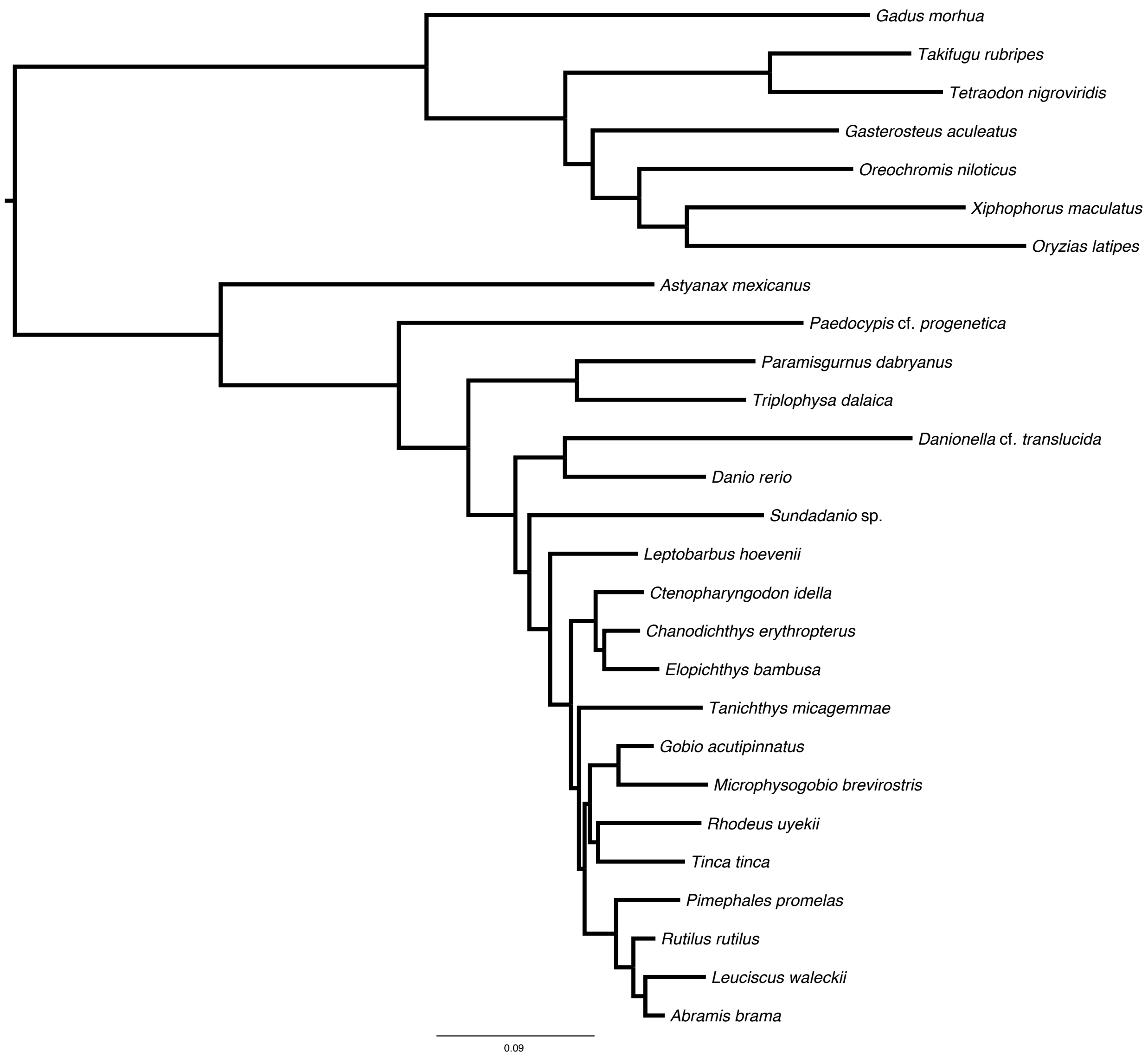
Kozlov AM, Aberer AJ, Stamatakis A. 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31:2577–2579.

Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool 11:81.

Mayden RL, Chen WJ. 2010. The world"s smallest vertebrate species of the genus *Paedocypris*: a new family of freshwater fishes and the sister group to the world"s most diverse clade of freshwater fishes (Teleostei: Cypriniformes). Mol Phylogenet Evol 57:152–175.

Misof B, Meusemann K, Reumont von BM, Kück P, Prohaska SJ, Stadler PF. 2014. A priori assessment of data quality in molecular phylogenetics. Algorithms for Molecular Biology 9:1.

Tao W, Mayden RL, He S. 2013. Remarkable phylogenetic resolution of the most complex clade of Cyprinidae (Teleostei: Cypriniformes): A proof of concept of homology assessment and partitioning sequence data integrated with mixed model Bayesian analyses. Mol Phylogenet Evol 66:603–616.



Supplemental Figure 1. Maximum likelihood tree inferred from 5,674 loci, partitioned by gene, using ExaML. Note that this preliminary analysis is not bootstrapped.