GENOME EVOLUTION AND GENE EXPRESSION DIVERGENCE IN THE GENUS *DANIO*

by

BRAEDAN MARSHALL MCCLUSKEY

A DISSERTATION

Presented to the Department of Biology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2016

DISSERTATION APPROVAL PAGE

Student: Braedan Marshall McCluskey

Title: Genome Evolution and Gene Expression Divergence in Genus *Danio*

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Biology by:

William Cresko	Chairperson
John H. Postlethwait	Advisor
Monte Westerfield	Core Member
John Conery	Core Member
Nelson Ting	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2016

© 2016 Braedan Marshall McCluskey
This work is licensed under a Creative Commons
Attribution (United States) License.

DISSERTATION ABSTRACT

Braedan Marshall McCluskey

Doctor of Philosophy

Department of Biology

June 2016

Title: Genome Evolution and Gene Expression Divergence in Genus *Danio*

Genus *Danio* includes zebrafish (*Danio rerio*) and several other phenotypically diverse species. To understand the history of these species and how they acquired the genetic differences underlying their diverse phenotypes, I performed two phylogenomic studies using Restriction-Site Associated DNA Sequencing and DNA hybridization-based exome enrichment. The results of these studies highlight important methodological considerations applicable to future experiments across taxa. Furthermore, these studies provide detailed understanding of the relationships within *Danio* including extensive introgression between lineages. The extent of introgression varies across the genome with regions of high recombination at the ends of chromosomes having the most evidence for introgression. Together, this work gives vital insight into the history of a model organism and the evolutionary processes that give rise to phenotypic diversity.

This dissertation includes published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Braedan Marshall McCluskey

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Augustana College, Sioux Falls, SD (Now Augustana University)

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2016, University of Oregon
Bachelor of Arts, Biology, 2009, Augustana College, Sioux Falls, SD

AREAS OF SPECIAL INTEREST:

Evolution
Gene Expression
Phylogenomics
Zebrafish

PROFESSIONAL EXPERIENCE:

Graduate Research Fellow, University of Oregon, 2010-2016

Graduate Teaching Fellow, University of Oregon, 2009-2010

Undergraduate Research Fellow, Augustana College, Sioux Falls, SD, 2008

Undergraduate Research Fellow, VA Hospital, Sioux Falls, SD, 2007

Teaching Assistant, Augustana College, Sioux Falls, SD, 2006-2009

GRANTS, AWARDS, AND HONORS:

NRSA F32 Postdoctoral Fellowship for study in the lab of Dr. David Parichy at the University of Washington, “Morphogenesis of Pigmentation Patterns in Zebrafish and its Relatives: Testing the Genetic and Developmental Basis of Diverse Adult Phenotypes across Evolution.”

Genetics Society of America Travel Award, 2016

Genetics Training Grant Fellow, University of Oregon, 2010-2013

PUBLICATIONS:

McCluskey BM & Postlethwait JH. 2015. Phylogeny of Zebrafish, a "Model Species," within *Danio*, a "Model Genus". Molecular Biology and Evolution 32:635-652.

Braasch I, Peterson SM, Desvignes T, **McCluskey BM**, Batzel P, Postlethwait JH. 2015. A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 324:316-341.

Wilson CA, High SK, **McCluskey BM**, Amores A, Yan YL, Titus TA, Anderson JL, Batzel P, Carvan MJ, Schartl M, Postlethwait, JH. 2014. Wild Sex in Zebrafish: Loss of the Natural Sex Determinant in Domesticated Strains. Genetics 198:1291-1308.

ACKNOWLEDGMENTS

I wish to express sincere appreciation to Professor John H. Postlethwait for his mentorship and assistance in preparing this manuscript. Additionally, I owe thanks to Peter Batzel and Jason Sydes for their informatics support, Trevor Enright, Amanda Rapp, and Bill Trevarrow for support in the form of fish husbandry, Tom Titus, Angel Amores, and Catherine Wilson for assistance with molecular techniques, and all members of the Postlethwait lab for their continued intellectual and moral support. Thank you to the members of the Institute of Ecology and Evolution for openness both personally and intelectually and for critical discussions of evolutionary processes. I also wish to acknowledge the contributions of Drs. Fang Fang and Sven Kullander, for their invaluable contributions to our understanding of *Danio* biodiversity, without which these studies would not have been possible. These investigations were supported in part by the University of Oregon Genetics Training Grant (5-T32-GMO7413) and NIH grants awarded to Dr. John H. Postlethwait (R01OD011116).

I dedicate this work to my parents, Tabb and Joan McCluskey,

who made little boys ask questions.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION……………………………………………………………….1

The Genome as a Forest, not a Tree………………………………………..2

Brief Outline of this Dissertation……………………………………………...5

II. PHYLOGENY OF ZEBRAFISH, A “MODEL SPECIES,” WITHIN *DANIO*, A “MODEL GENUS” ………………………………………………………………..6

Abstract……………………………………………………………………………6

Introduction…………..……………………………………………………………7

Results……………………………………………………………………………15

The Number of Total RAD-tags Varies Widely Across Species Due to Expansion of Repeats and Gene Families…………………………………...15

Genomic Distribution of RAD-tag Loci……………………………………...23

Maximum Likelihood Phylogenetic Inference Based on Concatenated Datasets……………………………………………………………………26

Bayesian Inference and Maximum Parsimony Analyses of Concatenated Datasets……………………………………………………………………30

Introgression Testing with pyRAD…………………………………………...32

Multi-Locus Phylogeny………………………………………………………34

Phylogeography of the Danio *Genus*………………………………………...37

Discussion………………………………………………………………………..40

The Phylogeny of Danio……………………………………………………...40

Potential Biases in RAD-seq Phylogenomics………………………………..42

Methodological Considerations for Application of RAD-seq to Phylogenomics………………………………………………………………43

Conclusions…………………………………………………………………..46

Materials and Methods…………….……………………………………………..47

Animals………………………………………………………………………47

Genomic DNA Extraction and Sequencing…………………………………..48

Locus Generation and Orthology Inference…………………………………49

Phylogenetic inference……………………………………………………………51

Analyses of genomic features……………………………………………………………52

Biogeographic analyses…………………………………………………………...52

Bridge……………………………………………………………………………53

III. A HYBRID HISTORY OF ZEBRAFISH REVEALED BY PHYLOGENOMICS AND GENOME STRUCTURE…………………………………………………………..54

Abstract………………………………………………………………………..54

Introduction…………….…………………………………………………………..55

Results……………………………………………………………………………57

Genealogical Heterogeneity within Danio…………………………………..57

Genomic Regions Affected by Introgression……………………………………64

Discussion……………………………………………………………………..69

A Hybrid History of the Zebrafish Genome……………………………………69

Implications for Phylogenomic Studies……………………………………...75

Conclusions………………………………………………………………………..78

Materials and Methods………………….…………………………………………..79

Library preparation and Exome Alignment Creation………………………..79

Maximum Likelihood Phylogenetic analyses………………………………...80

Genomic Structure analyses……………………………………………………81

Bridge……………………………………………………………………………82

IV. CONCLUSIONS……………………………………………………………………83

V. APPENDIX…….…………………….…………………………………………86

Supplemental Figures………………….…………………………………………86

Supplemental Tables….………………….…………………………………………88

LIST OF FIGURES

Figure Page

1.1. DNA and Population Topologies………………………………………………..3

2.1. Recent *Danio* phylogenies disagree on the sister group to zebrafish and other relationships…………………………………………...10

2.2. Analysis of RAD-tag alignments to the zebrafish genome and description of constructed datasets……………………………………………16

2.3. Enrichment of SbfI sites in the zebrafish genome………………………..............25

2.4. Maximum likelihood phylogeny of the *Danio* genus based on RAD-tag sequences……………………………………………………...28

2.5. Bipartition frequencies of phylogenies inferred from Individual RAD-tag loci..…………………………………………………………37

2.6. Phylogeography of *Danio* species……………………………………..............39

3.1. Total Evidence and Mitochondrial Phylogenies……………………………………59

3.2. Genealogical history discordance and effects of chromosome structure…………………………………………………………………..61

3.3. Relationships of taxon trios vary based on chromosome position…………………..67

3.4. A model for the recent evolutionary history of zebrafish…………………………...70

S.2.1. Phylogeny of the *Danio* genus based on RAD-tag sequences and Bayesian inference…………………………………………….86

S.2.2. Cladogram of the *Danio* genus based on RAD-tag sequences and maximum parsimony…………………………………………87

LIST OF TABLES

Table Page

1. Data Set Characteristics……………………………………………………………….24

2. Partitioned Patterson's *D*-Statistic Test………………………………………………….34

S.2.1. Reads and RAD-tags per Sample………………………….................................88

CHAPTER I

INTRODUCTION

Just as any two individuals can trace their ancestry to a single individual within an ancestral population, any two orthologous stretches of DNA can trace their way back to a single stretch of DNA within an ancestral genome. Genetic differences between individuals (and ultimately between species) are the result of random mutations that occurred since this DNA replication event. Some of these genetic differences between genomes cause phenotypic differences between individuals. To understand the evolution of genetic differences underlying the phenotypic differences that distinguish two species, we need to understand the evolutionary history of the lineages that carried the genomes as they diverged and the types of genetic differences the genomes acquired.

Recently, phylogenomic studies are enhancing our understanding of the histories of species and genomes (Baack and Rieseberg 2007; Harrison and Larson 2014; Soucy, et al. 2015). Studies in diverse taxa find evidence that a single tree-like model is insufficient to explain the histories of many groups of species (Linder and Rieseberg 2004; Cui, et al. 2013a; Eaton and Ree 2013; Jones, et al. 2013; Martin, et al. 2013a, b; Hipp, et al. 2014; Fontaine, et al. 2015; McCluskey and Postlethwait 2015; Li, et al. 2016; Pease, et al. 2016). The "Tree of Life" is more of a "Web of Life" where the genetic material within organisms has more than one evolutionary history. Horizontal gene flow in microorganisms is the predominant example of this phenomenon, but we are now beginning to appreciate how processes including population structure, speciation with gene flow, introgression, and hybridization have shaped the history of multicellular, diploid organisms. One aspect that has become clear as genome-level data become increasingly available is that more sequence data is not sufficient to answer many question in evolutionary biology (Roure, et al. 2007; Philippe, et al. 2011). What we need, and what is gradually happening, are fundamental changes in the way we look at the histories of species that explicitly include these non-tree-like processes (Larget, et al. 2010; Pickrell and Pritchard 2012; Faircloth, et al. 2013; Ting and Sterner 2013; Hipp, et al. 2014; Martin, et al. 2015).

The Genome as a Forest, not a Tree

The genetic similarities and differences between DNA sequences provide evidence for the history of those sequences. By comparing DNA sequences at a single orthologous location in the genome (a locus), we can infer the order in which mutations occurred and how sequences at that locus are related by common descent (Felsenstein 1981). Under most circumstances, a tree-like structure known as a bifurcating topology, which contains nodes corresponding to DNA replication events (Figure 1.1a) can represent these relationships. If processes such as recombination or gene conversion occurred within a locus, the locus will include more than one history and the topology explaining the history of that locus as a whole will be reticulate, not strictly bifurcating (Figure 1.1b). Although many types of characters can be used for molecular phylogenetic analyses, one of the most commonly used is single nucleotide polymorphisms (SNPs). In most circumstances, SNPs can often be assumed to have one of two states: ancestral or derived, which are often represented simply as “A” and “B” for simplicity (Durand, et al. 2011; Eaton and Ree 2013). Using SNPs on short timescales, this two-state assumption is likely to be valid because it is unlikely that independent mutations at a locus will cause changes to identical character states, a condition known as homoplasy (Felsenstein 1978). Derived mutations, such as SNPs, that accumulate within a DNA sequence are passed on following replication and provide evidence for the common ancestry of sequences containing those mutations (Figure 1.1c).

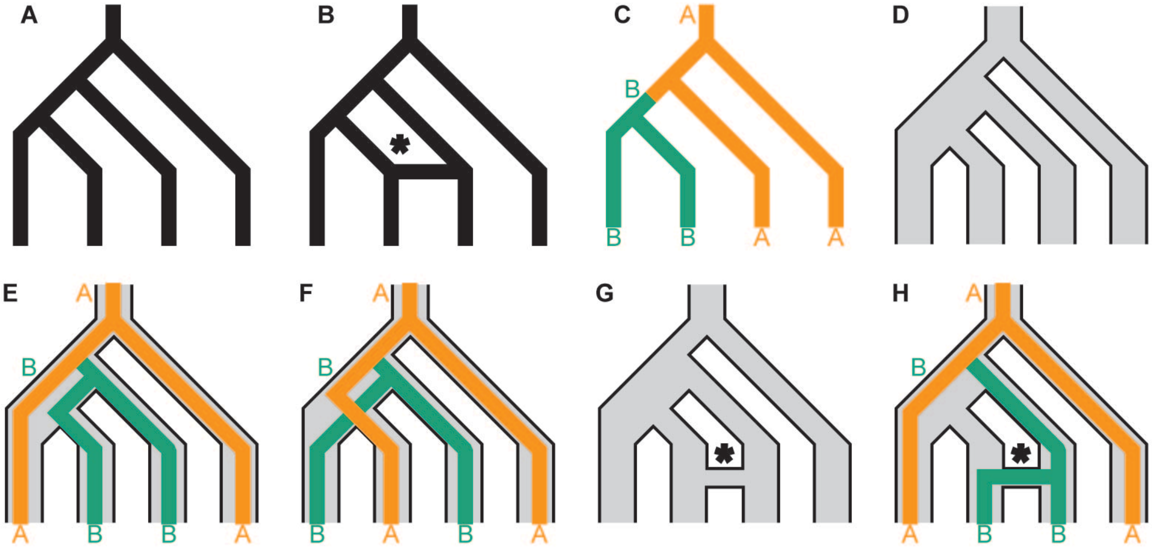


Figure 1. DNA and Population Topologies. (a) A strictly bifurcating topology representing the history of a DNA locus. (b) A reticulate topology. Asterisk denotes reticulation event. (c) Derived character states, such as SNPs, are inherited by all descendants. The BBAA pattern is consistent with the population topology. (d) In addition to the histories of genetic loci, topologies can also describe the histories of populations of organisms. Populations have many DNA sequences, and the history of some sequences may not match the population topology for any of several reasons. (e) An ABBA pattern can arise when a genetic locus experiences ILS and has a different history than the populations that carried it. (f) A BABA pattern is as likely as an ABBA pattern under ILS. (g) Population topologies can be reticulate. Asterisk denotes secondary contact of two partially reproductively isolated populations. (h) Introgression can occur when population topologies are reticulate. Introgressed genetic loci have a topology that may differ from the majority of the genome.

Every nucleotide in every locus in every genome has a history that can be represented by a single topology. In sexually reproducing diploid taxa, however, the genome is not inherited as a single locus and local topologies can vary across the genome

(Martin, et al. 2013a; Fontaine, et al. 2015; Li, et al. 2016; Pease, et al. 2016). Just as the histories of DNA sequences have topologies, the histories of populations carrying those sequences have topologies, which we can infer from the DNA sequences the populations carry. Ancestral populations of sexually reproducing diploid organisms that experienced instantaneous and complete reproductive isolation (i.e. instant speciation) will show a single topology that describes the history of those populations (Figure 1.1d). In population topologies, splits represent reproductive isolation just as splits represent DNA replication events in genetic topologies.

Even with instantaneous speciation, loci in the genome may disagree with the population topology under certain circumstances (Holder, et al. 2001). The histories of loci should often agree with the topology of the populations in which they exist (compare Figure 1.1c and Figure 1.1d). If speciation events are closely spaced in time, then ancestral genetic variation that existed prior to the first speciation event can sort out according to a topology that differs from the population topology. This process is known as incomplete lineage sorting of alleles (ILS). ILS can result in either of two topologies that differ from the population topology (Figures 1.1e&f). These two ILS topologies should occur with equal likelihood. Because ILS topologies differ from the population topology, the history of a genome with both topologies will be reticulate.

Reticulate evolution can also occur via introgression (gene flow between populations following reproductive isolation). If populations were reproductively isolated, but came into secondary contact, their population history is reticulate (Figure 1.1g). Loci can pass from one population to another via this reticulation, a process known as introgression (Figure 1h) (Harrison and Larson 2014). If only one of two populations receives introgressed alleles from a more distant relative, the introgressed alleles will have a topology that matches only one of the ILS topologies (i.e. always ABBA as in Figure 1.1e; never BABA as in Figure 1.1f) (Durand, et al. 2011). By extending these few logical principles, we can use mutations present in DNA sequences from related species to determine the history of the populations that carried those DNA sequences as they accumulated mutations.

Brief Outline of this Dissertation

Herein, I use the zebrafish, *Danio rerio*, and other members of genus *Danio* to study genome evolution. In Chapter II, I use novel methodologies to infer the relationships between zebrafish and several of its relatives. This work was coauthored by my advisor, Dr. John H. Postlethwait, and published in Volume 32 of *Molecular Biology and Evolution* in February of 2015. In Chapter III, I further investigate these relationships and the role that chromosome structure and population history had on genome sequences between diverging *Danio* lineages. In writing this chapter, I received detailed feedback from my advisor, Dr. John H. Postlethwait. Finally, in Chapter IV, I summarize and synthesize the conclusions of these studies and how they impact our understanding of genome evolution in the genus *Danio*.

CHAPTER II

PHYLOGENY OF ZEBRAFISH, A "MODEL SPECIES," WITHIN *DANIO*, A "MODEL GENUS"

This work was coauthored by my advisor, Dr. John H. Postlethwait, and published in Volume 32 of *Molecular Biology and Evolution* in February of 2015. I performed all experiments and analyses as well as the majority of the writing. Dr. Postlethwait contributed substantially to the experimental design, implementation of novel methodologies, discussion of results, writing, and editing.

Braedan M. McCluskey[1], and John H. Postlethwait[1]

[1] Institute of Neuroscience, University of Oregon, Eugene, OR 97403-1254

Abstract

Zebrafish (*Danio rerio*) is an important model for vertebrate development, genomics, physiology, behavior, toxicology, and disease. Additionally, work on numerous *Danio* species is elucidating evolutionary mechanisms for morphological development. Yet, the relationships of zebrafish and its closest relatives remain unclear possibly due to incomplete lineage sorting, speciation with gene flow, and interspecies hybridization. To clarify these relationships, we first constructed phylogenomic datasets from 30,801 RAD-tag loci (483,026 variable positions) with clear orthology to a single location in the sequenced zebrafish genome. We then inferred a well-supported species tree for *Danio* and tested for gene flow during the diversification of the genus. An approach independent of the sequenced zebrafish genome verified all inferred relationships. Although identification of the sister taxon to zebrafish has been contentious, multiple RAD-tag datasets and several analytical methods provided strong evidence for *Danio aesculapii* as the most closely related extant zebrafish relative studied to date. Data also displayed patterns consistent with gene flow during speciation and post-speciation introgression in the lineage leading to zebrafish. The incorporation of biogeographic data with phylogenomic analyses put these relationships in a phylogeographic context and supplied additional support for *D. aesculapii* as the sister species to *D. rerio*. The clear resolution of this study establishes a framework for investigating the evolutionary biology of *Danio* and the heterogeneity of genome evolution in the recent history of a model organism within an emerging model genus for genetics, development, and evolution.

Introduction

The zebrafish, *Danio rerio* (Hamilton, 1822), is an important model for understanding vertebrate developmental mechanisms (Kinkel and Prince 2009), genome evolution (Postlethwait, et al. 2004), physiology (Lohr and Hammerschmidt 2011), behavior (Norton and Bally-Cuif 2010), toxicology (Peterson and MacRae 2012), and disease (Lieschke and Currie 2007; Santoriello and Zon 2012). Furthermore, along with mouse and human, zebrafish has the best genome assembly and gene annotation among vertebrates (Howe et al., 2013).

In addition to zebrafish, the genus *Danio* (sensu Fang, 2003) contains several other species (hereafter referred to as danios) that differ from zebrafish in size, pigment patterns, skeletal morphologies, growth control, and behaviors (Fang 2003; Rosenthal and Ryan 2005; Froelich, Fowler, et al. 2013). Phenotypic differences in species closely related to zebrafish expedite the investigation of the molecular biology of evolution through comparative studies among related species coupled with functional tests in zebrafish. Several developmental studies using various danios showed that seemingly homologous features can develop by different cellular, molecular, and genetic mechanisms: for example, striped pigment patterns derive from different primary cell populations in different species (Quigley, et al. 2004a); certain molecular pathways are crucial for pattern formation in some species but not others (Quigley, et al. 2005; McMenamin, et al. 2014b); and mutations in the same gene can reduce stripe formation in one species but increase stripe formation in another species (Mills, et al. 2007). Other recent work studied the development of keratinized breeding tubercles (a synapomorphy of *Danio*) and used gain-of-function and loss-of-function mutants in zebrafish to recapitulate the range of morphological variation seen in these structures in other danios (Rodriguez 2013). Another study used functional tests in zebrafish embryos to determine when regulatory modules arose in the lineage leading to zebrafish (Camp, et al. 2012).

A feature of genus *Danio* that facilitates evolutionary analysis is that zebrafish can form hybrids with its congeners and even more distantly related relatives (Parichy and Johnson 2001; Wong, et al. 2011). Particularly informative studies identified genes involved in the evolution of species-specific pigment patterns by the strategy of mating zebrafish pigmentation pattern mutants to other danios to test for complementation of phenotypes (Parichy and Johnson 2001). Similar strategies using more distantly related species elucidated the mechanisms that led to the evolution of different patterns of muscle growth (Froelich, Fowler, et al. 2013; Froelich, Galt, et al. 2013). A clear understanding of the historical relationships among species used in such experiments is necessary to interpret the results in an evolutionary context. Only with a well-supported phylogeny can we confidently infer ancestral states, distinguish synapomorphic traits from homoplasic traits, and determine the order of events in evolution.

Recent phylogenetic studies involving *Danio* addressed the placement of the genus in relation to other groups within the incredibly diverse order Cypriniformes (Mayden, et al. 2007; Fang, et al. 2009; Tang, et al. 2010). These studies used numerous taxa to accommodate the diversity of species within Cypriniformes, but used sequences from relatively few loci. As such, these datasets were well suited to resolving relationships at the genus level and above, but relationships below the genus level, particularly between closely related species, were often unresolved. These studies agree on the placement of several groups of species within *Danio* (Figure 2.1), although not all species within those groups are represented in every study. First, the large danios—*D. feegradei* (yoma danio) and *D. dangila* (moustached danio), which is the type species of the genus—are consistently recovered basal to all other danios. Second, three species—*D. choprae* (glowlight danio)*, D. erythromicron* (emerald dwarf danio)*,* and *D. margaritatus* (celestial pearl danio)—are recovered as a monophyletic clade hereafter referred to as the *D. choprae* species group. Third, four taxa—*D. albolineatus* (pearl danio)*, D. roseus* (rose danio)*, D. kerri* (Kerr’s danio)*,* and *D.* sp *“hikari”*—form a clade hereafter referred to as the *D. albolineatus* species subgroup. Fourth, a clade within *Danio* that excludes the large, basal danios and the *D. choprae* species group has high support in all three studies; we refer to this clade as the "*D. rerio* species group" because it includes *D. rerio* and all species recovered as members of its sister group in one or more of these three studies—*D. aesculapii* (panther danio), *D. kyathit* (orange-finned danio), *D. nigrofasciatus* (dwarf danio) and the *D. albolineatus* species subgroup. The *D. rerio* species group likely includes three other recently described species—*D. jaintianensis* (Sen 2007), *D. quagga* (Kullander, et al. 2009), and *D. tinwini* (Kullander and Fang 2009b)—that were not included in previous molecular phylogenetic studies.

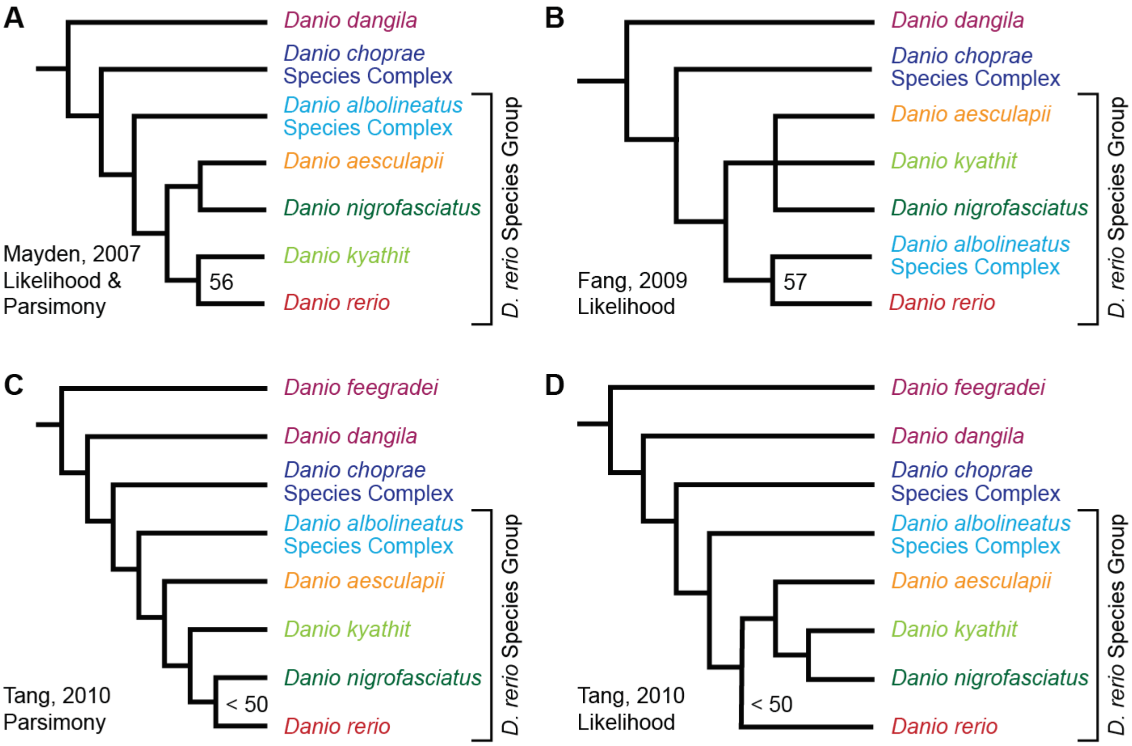


Figure 2.1. Recent *Danio* phylogenies disagree on the sister group to zebrafish and other relationships. (*A*) Topology from Mayden et al. (2007) recovered by parsimony and maximum likelihood from two nuclear and four mitochondrial genes that involved 6,921 positions. (*B*) Topology from Fang et al. (2009) recovered by maximum likelihood from Rhodopsin and Cytb sequences that involved 1,542 positions. (*C*) Topology from Tang et al. (2010) recovered by parsimony based on two nuclear and two mitochondrial genes that involved 4,117 positions. (*D*) Topology from Tang et al. (2010) recovered by maximum likelihood based on the same data as in c. To simplify comparisons across studies, all topologies are presented as phylograms and show support only for the sister group to *D. rerio*. In each phylogeny, tips labeled as *D. albolineatus* and *D. choprae* species groups include two or more taxa included in these groups.

Despite the emergence of the *Danio* genus as a model system for understanding the molecular genetics of functional evolution, relationships within the *D. rerio* species group remain contested, particularly the identity of the sister group to *D. rerio* (Figure 2.1). The aforementioned recent phylogenetic studies of *Danio* and related taxa (Mayden, et al. 2007; Fang, et al. 2009; Tang, et al. 2010) inferred four different species or clades to be the sister group to zebrafish, with each relationship having only limited support. As previous phylogenetic studies were unable to clearly resolve the relationships of species closely related to zebrafish, we performed a phylogenomic study of genus *Danio* focused on the *D. rerio* species group. Phylogenomic datasets, which consist of sequences from hundreds or thousands of loci from throughout the genome, can offer several advantages over phylogenetic datasets, which use sequences from only a handful of loci. Phylogenomic approaches are particularly important in clades with short internal branches (Pollard, et al. 2006), clades with a history of hybridization (Cui, et al. 2013b), and clades with introgressed genomic regions (Martin, et al. 2013a) because under these conditions different regions of the genome have different histories.

The low cost of short-read sequencing makes representational sequencing that samples a large subset of the genome an attractive option for obtaining a genome-wide analysis of informative characters for phylogenomics. A particularly promising method is RAD-seq (restriction-associated DNA sequencing), in which short DNA sequences adjacent to restriction enzyme cutting sites (i.e. RAD-tags) provide phylogenetic signal through polymorphisms within the DNA tag (Baird, et al. 2008). The nature of RAD-seq, however, presents two main challenges for phylogenomics. First, RAD-tag loci are shorter than some sequences used for phylogenomic inferences, such as contigs generated from transcriptome data (Cui, et al. 2013b) and ultraconserved elements (Faircloth, et al. 2013). The relatively short length of individual RAD-tag loci makes it a challenge to correctly identify orthologous sequences and to distinguish among alleles, orthologs, and paralogs, especially ohnologs (Wolfe 2001). Moreover, because they are not restricted to coding genes (as with transcriptome-based methods) or predefined loci (as with ultraconserved elements), RAD-tags include sequences from all regions of the genome with the appropriate restriction enzyme cut site including repetitive elements, transposons, and low-complexity regions. To address this challenge, previous phylogenomic studies using RAD-seq employed clustering methods (Emerson, et al. 2010; Hohenlohe, et al. 2011; Eaton and Ree 2013; Jones, et al. 2013; Hipp, et al. 2014), BLASTx searches (Wang, et al. 2013), or alignments to EST-based transcriptome sequences (Andrew, et al. 2013) to infer orthology. Here, we introduce the use of a reference genome and annotations of repetitive elements to define orthology across species.

The second hurdle to using RAD-seq for phylogenomics is that the restriction enzyme cut site must be conserved across taxa to sequence the adjacent genomic DNA and obtain orthologous sequences. The turnover of restriction enzyme recognition sites through evolutionary time results in large amounts of “missing” data as the distance between taxa increases. The long, eight base pair recognition sequences we generally employ exacerbate this problem because, although they have the advantage that they cut only about 25,000 times per genome, they suffer the disadvantage of incurring mutations faster than enzymes with four or six base pair recognition sequences. These issues have been addressed at length for population genomic studies within species (Baird, et al. 2008; Emerson, et al. 2010; Amores, et al. 2011; Catchen, et al. 2011; Hohenlohe, et al. 2011), but theoretical and modeling work in addition to that performed on sequenced genomes (Rubin, et al. 2012; Cariou, et al. 2013), would improve the utility of RAD-seq data across taxa separated by longer timescales.

In recent years, RAD-seq has emerged as a common tool for population genomic studies within species, but exploration of its utility for phylogenomics is only beginning. RAD-seq has been used for phylogenomics in instances of recent radiations: mosquitos diverged <22,000 years (Emerson, et al. 2010), cichlids diverged <15,000 years (Keller, et al. 2013; Wagner, et al. 2013), pupfishes diverged <10,000 years (Martin and Feinstein 2014), and a clade of flowering plants (Eaton and Ree 2013). The utility of RAD-seq for answering phylogenetic questions on longer timescales was verified *in silico* for Drosophila (crown age 60 million years) (Rubin, et al. 2012; Cariou, et al. 2013), but empirical studies outside of recent radiations are rare. A RAD-seq phylogenomic approach investigating relationships among Xiphophorus fishes (crown age 2.44 million years) (Jones, et al. 2013) recovered a topology nearly identical to the topology obtained in an independent phylogenomic study based on transcriptomic data (Cui, et al. 2013b).

Other groups (Cruaud, et al. 2014) used RAD-seq to infer relationships of ground beetles with a crown age of 17 million years and American oaks with a crown age of 23-33 million years (Hipp, et al. 2014). Several aforementioned studies (Eaton and Ree 2013; Jones, et al. 2013; Keller, et al. 2013; Hipp, et al. 2014) as well as a recent study in sunflowers (Andrew, et al. 2013) also used RAD-seq to test for hybridization and gene flow in the diversification of their respective study species.

Here, we used RAD-tag sequences flanking cut sites for the restriction enzyme SbfI to resolve relationships in the genus *Danio.* We investigated twelve danios and seven outgroup taxa and analyzed data either by aligning reads to the *D. rerio* reference genome or by clustering RAD-tags into *de novo* loci based on sequence similarity independent of the reference genome. Using concatenated datasets, which were orders of magnitude larger than previous datasets and originated from thousands of loci across the genome, both approaches recovered the same topology with *Danio aesculapii* as the sister taxon to *Danio rerio* using both maximum likelihood and Bayesian inference. The topology inferred using maximum parsimony differed only in the placement of the two basal danios. A third analysis using a multilocus approach and Patterson's D-statistic revealed complicated historical relationships consistent with rapid speciation and introgression during the diversification of genus *Danio*, particularly in the lineage leading to zebrafish. We found that the biogeographical distribution of danio species compared to recovered historical relationships among *D. rerio* and its closest congeners is consistent with species distributions across geographically distinct hydrological basins. Our findings suggest that previous phylogenetic studies obtained limited support for the relationships of zebrafish and its closest relatives due to a lack of phylogenetic signal on the internal branches near the base of the *D. rerio* species group. Furthermore, our findings provide evidence that the discordance among previous studies could be due to different gene trees underlying the limited number of loci investigated in each study. The new, more detailed understanding of the history of *Danio* given by this RAD-seq study provides a better framework for understanding the molecular biology and evolution of a preeminent vertebrate "model species" and an emerging "model genus."

Results

The Number of Total RAD-tags Varies Widely Across Species Due to Expansion of Repeats and Gene Families

Using the restriction enzyme SbfI, we digested genomic DNA and prepared RAD-tag libraries from 41 individuals representing twelve species within *Danio* and seven outgroup species (Baird, et al. 2008). Sequencing the libraries provided 1.0 million to 3.8 million quality filtered reads per sample (supplementary table S.2.1, Supplementary Material online). This sequencing depth is equivalent to 38x to 152x average coverage of the 58,720 RAD-tags we identified on the 25 chromosomes of the zebrafish reference genome (Zv9 version 72). We then used Stacks (Catchen, et al. 2011) to create RAD-tag loci from highly similar sequences with sufficient sequencing depth. The total number of RAD-tags per individual ranged from 26,132 to 58,750 (Figure 2.2a). This variation was significantly associated with species ($F(18,22) = 17.81$, $p < 0.001$) and independent of sequencing depth ($t(39) = 3.70$, $p = 0.062$, $R^2 = 0.087$), suggesting that the observed variation is not due to under-sequencing, but rather to a fluctuation in the total number of RAD-tags across species.

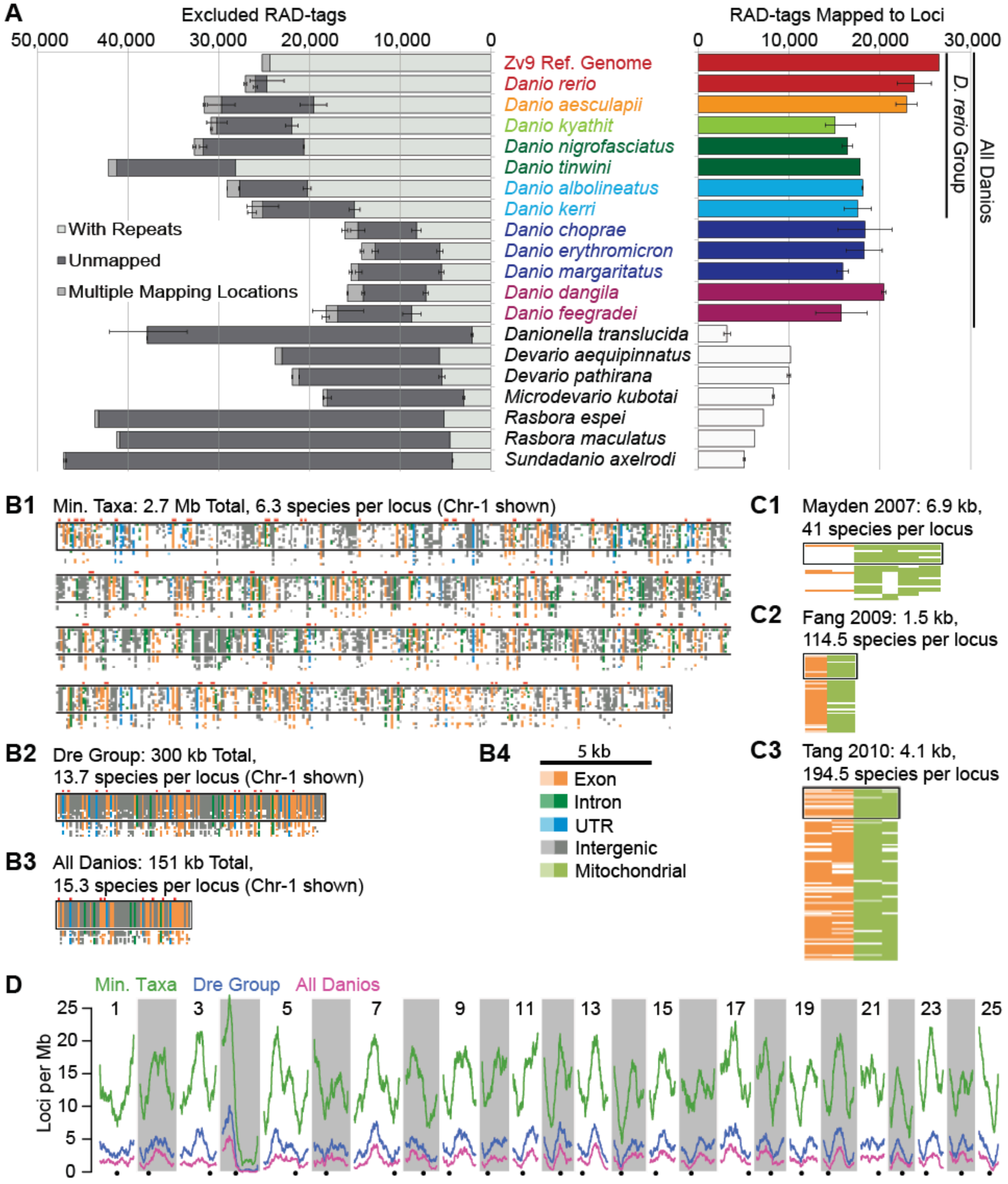


Figure 2.2. Analysis of RAD-tag alignments to the zebrafish genome and description of constructed datasets. (*A*) RAD-tags across taxa. Top bars represent *in silico* RAD-tags *from* the Zv9 reference genome aligned back *to* the genome in the same manner as the RAD-tags from the biological samples. Species are ordered according to their relatedness to *D. rerio* (middle column) as recovered by the maximum likelihood phylogeny of Tang et al. (2010) such that species near the top of the graph are more closely related to *D. rerio*. Error bars represent the range of RAD-tags obtained for each species except for *D. rerio*, where error bars represent the standard error of the mean for the seven zebrafish individuals. The left bar graph (gray tones) shows the number of RAD-tags that were excluded from phylogenetic analyses due to questionable orthology across taxa, including tags that were repetitive (pale gray), were unmapped (dark gray), or that mapped to multiple locations (medium gray). The right bar graph (color) shows the number of RAD-tag loci (i.e. RAD-tags with clear orthology to a single, non-repetitive, locus in the zebrafish genome). Use of a reference genome identifies a large number of RAD-tag loci with well-supported orthology in species closely related to the reference genome. (*B*) Comparison of datasets used for phylogenetic inference in this study. The three graphics represent all RAD-tag loci from (*B1*) chromosome 1 (chr-1) for the Min. Taxa dataset, (*B2*) chr-1 of the Dre Group dataset, and (*B3*) chr-1 of the All Danios dataset. Each narrow row of blocks represents one of the 19 species included in this study (listed in the order shown in Figure 2A). Each narrow column of blocks represents a RAD-tag locus (1,318 in Min. Taxa, 133 in Dre Group, and 66 in All Danios). Note that the Min. Taxa dataset is broken into four segments. Taxa within *Danio* are surrounded by a black rectangle. Loci at splice acceptors are denoted by red blocks over the loci. The horizontal scale and the color code for each block denoting the genomic feature of the locus (*B4*) applies to B and C. The darkness of each block denotes whether that locus was sequenced in all samples (dark), some samples (midtone), or no samples (white) for each taxon. (*C*) Datasets used for phylogenetic inference in previous studies. The datasets of (*C1*) Mayden et al. (2007), (*C2*) Fang et al. (2009), and (*C3*) Tang et al. (2010) are shown with taxa restricted to the genera included in the current study. (*D*) Density of RAD-tag loci from the Min. Taxa, Dre Group, and All Danios datasets mapped across the Zebrafish Genome. Alternating light and dark vertical bars represent chromosomes with centromeres denoted as black dots below each column. Colored lines represent the average number of loci in a 10 Mb (megabase) sliding window for Min. Taxa dataset (green), Dre Group dataset (blue), and All Danios dataset (fuschia). See text for explanation of the anomalously low number of tags on the right arm of chr-4.

Unlike previous phylogenomic RAD-seq studies, we could use the high quality reference genome and annotated repeat information for one of our species to provide protection against incorrect orthology assumptions. Using these resources, we performed stringent quality filtering. We removed RAD-tags if they matched annotated repeats, failed to align to the zebrafish genome with high support, or aligned to more than one location with equal support. Using these filters, we limited the RAD-tags in our phylogenomic datasets to those that aligned with high support to a single, non-repetitive location in the zebrafish genome. We therefore refer to these sets of orthologous RAD-tags mapped to the same genomic location as "RAD-tag loci," or simply "loci" to distinguish them from RAD-tags excluded from our datasets due to unclear orthology. To evaluate how well these filtering steps worked, we also applied them *in silico* to the 58,720 RAD-tags flanking SbfI cut sites on the 25 chromosomes in the zebrafish reference genome (Zv9). We include these *in silico* results for comparison to the actual results obtained for our seven zebrafish samples.

The percentage of total RAD-tags removed from further analyses because of similarity to zebrafish repetitive elements varied greatly across species (Figure 2.2A, light gray bars) from 9.0% in *Danionella translucida* to 55.9% in *Danio tinwini*. Compared to species more distantly related to zebrafish, members of the *D. rerio* species group had on average more than twice as many RAD-tags with high similarity to annotated repetitive elements. DNA transposons, which make up 39% of the entire zebrafish genome (Howe, et al. 2013), comprised 26.2% of the RAD-tags generated *in silico* from the zebrafish genome and 26.9% of the RAD-tags from the seven sequenced zebrafish specimens, the highest of any species in the study. DNA transposons made up

14.2% to 26.0% of the RAD-tags for the other species in the *D. rerio* species group, 5.6% to 8.8% in danios outside of the *D. rerio* species group and 2.2% to 6.8% in outgroups. This trend suggests that the 39% of the zebrafish genome made up of DNA transposons is the result of a recent expansion of DNA transposons in the lineage leading to zebrafish. Satellite repeats, which constitute 0.9% of the total zebrafish reference genome (Howe, et al. 2013), made up 6.3% of our *in silico* RAD-tag dataset based on the reference sequence and 7.2% of the RAD-tags from our zebrafish samples. Other species in the *D. rerio* species group had similar levels of satellite repeats (4.1% to 13.8%) with the exception of *D. tinwini*, in which satellite repeats accounted for a staggering 32.4% of RAD-tags, suggesting a significant recent expansion of these elements in *D. tinwini*. Satellite repeats were less frequent in the more basally diverging danios (0.3% to 2.7% of RAD-tags) and in outgroups (0.3% to 0.9% of RAD-tags). This difference in DNA transposon and satellite content explains much of the variation in the number of total RAD-tags across species and suggests that the high repeat content of the zebrafish genome (52.2%, the highest reported repeat content of any sequenced vertebrate) (Howe, et al. 2013) is a relatively recent occurrence. For comparison, the closest relative to zebrafish with a sequenced genome is currently the common carp, *Cyprinus carpio*, the genome of which is estimated to contain just 11.7% to 28.0% repeats (Henkel, et al. 2012). We note, however, that the nature of the present RAD-seq experiment limits the repeat families that we can assay to those with SbfI sites. Further investigations into the historical origin of repeats in the zebrafish genome will require other forms of data, such as whole genome sequences from related species.

After removing RAD-tags related to repetitive elements, we found that 0.8% of the total RAD-tags from the *D. rerio* samples mapped to genomic scaffolds not incorporated into the 25 chromosomes and 1.6% of RAD-tags failed to align anywhere in the zebrafish reference genome (Figure 2.2a, dark gray). Some of these unmappable RAD-tags may represent sequencing artifacts, but many unmapped RAD-tags appeared independently in several different zebrafish samples suggesting that they belong to elements absent from the TU-strain-based zebrafish reference genome. RAD-tags from species other than zebrafish that failed to map to the *D. rerio* genome (15.8% to 29.1% per danio sample, 49.4% to 76.8% in the outgroups) likely represent sequences from three sources: regions orthologous to unassembled zebrafish genomic scaffolds, loci orthologous to locations on the 25 zebrafish chromosomes but diverged beyond recognition, and sequences with no orthologous locus in the zebrafish genome.

Across species, a small proportion of RAD-tags mapped to multiple locations on chromosomes in the zebrafish reference genome with equal support (Figure 2.2A, medium gray). Among RAD-tags with multiple mapping locations were 884 RAD-tags (1.7% of the total) generated *in silico* from chromosomes in the zebrafish reference genome. As we knew the genomic locations of these markers, we could identify the genetic elements from which they came. The long (right) arm of chromosome 4 (chr-4R), which has been noted for its elevated GC-content and highly duplicated gene families (Howe, et al. 2013), contained 232 multiple-mapping RAD-tags with an average of 11.1 mapping locations in chr-4R alone. Many multiply mapped RAD-tags occurred within members of duplicated gene families in this region including a subfamily of NOD-like receptors, which have been reported as unique to teleosts (Laing, et al. 2008). Other genomic elements that generated RAD-tags with multiple best mapping positions included centromeres, lncRNAs, and some tandemly duplicated genes (e.g., *ints12* at Chr-1: 25,414,014-25,419,172 and 1:25321186-25326343). In other sampled taxa, many of the 0.6% to 4.7% of RAD-tags that mapped to multiple locations appeared to come from similar elements, but may also represent sequences from other paralogous gene families, unannotated repeats, or constrained regions of ohnologs tracing back to genome duplications at the base of the teleost or the vertebrate radiations.

Across all 41 samples, 648,326 RAD-tags aligned to a single location on one of the 25 chromosomes in the zebrafish reference genome. These mapping locations corresponded to 89,593 loci with a RAD-tag alignment from at least one sample (table 1) with 30,801 loci having RAD-tags from four or more samples. Most loci that occurred in fewer than four samples were species-specific and therefore of little use for phylogenetic inference. We expected to recover 26,476 RAD-tag loci in our zebrafish samples based on our *in silico* analysis of the zebrafish reference genome. The mean number of loci identified across our *D. rerio* samples (23,770 loci) was lower than this genomic estimate, but substantially higher than the average of all species in the *D. rerio* species group (18,339 loci), the average across all danio species (17,734 loci), and the average for outgroup taxa (7,148 loci). This decrease in the number of RAD-tag loci with increasing phylogenetic distance from the reference genome mirrors the increase in unmappable RAD-tags mentioned previously. Interestingly, in contrast to the high variability in repetitive RAD-tags across species, the total number of non-repetitive RAD-tags (mapped RAD-tag loci and unmappable RAD-tags) was relatively constant across the phylogenetic distance investigated in this study with the mean across *Danio* (25,301

RAD-tags) being only slightly higher than the mean across outgroup species (24,953 RAD-tags).

Based on alignment locations, we inferred RAD-tags from different samples to represent orthologous loci if they mapped to the same unique location in the zebrafish genome. We constructed three datasets based on the degree of conservation across species (table 1 and Figure 2.2b). Compared to the datasets used by previous phylogenetic studies (Figure 2.2c), the resulting phylogenomic datasets were orders of magnitude larger in terms of total alignment positions and had similar or greater taxon sampling within *Danio*. The 'minimum taxa' dataset ("Min. Taxa") contains all 30,801 loci with RAD-tags from four or more samples including 5,370 loci with RAD-tags exclusive to zebrafish samples. The '*Danio rerio* species group' dataset ("Dre Group") consists of the 3,406 loci present in all 20 individuals sampled from seven species in the *Danio rerio* species group. The 'all danios' dataset ("All Danios") is restricted to the 1,720 loci that are conserved in all 28 sequenced samples of twelve species in genus *Danio*.

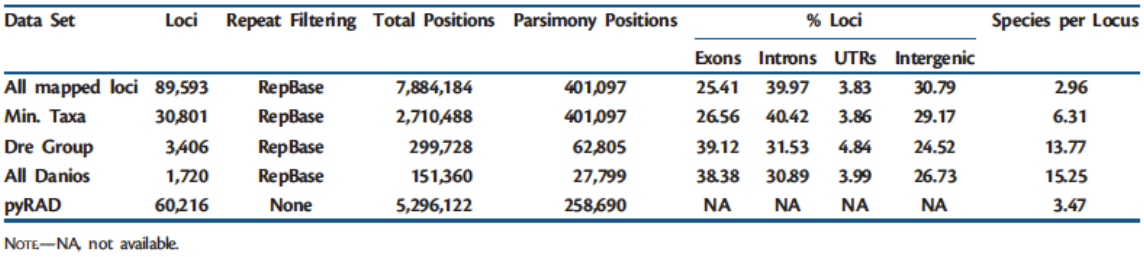
Most RAD-based phylogenomic studies have not had the luxury of a sequenced genome for any of the investigated taxa and have therefore inferred orthology based solely on sequence similarity. To complement our genome-assisted analyses and to better compare analyses across studies, we therefore also employed a genome-independent approach to infer RAD-tag orthology based only on sequence similarity. We generated the resulting dataset ("pyRAD Dataset" in table 1) using the pyRAD package (Eaton and Ree 2013), which was designed explicitly for phylogenomic analysis of RAD-seq data. This dataset contained 60,216 loci with RAD-tags from four or more samples. By comparing results from our genome-assisted approach and a genome-independent approach, we were able to infer the benefit or detriment of using available genomic resources.

Genomic Distribution of RAD-tag Loci

Having identified orthologous loci, we used the annotation of the *D. rerio* reference genome to determine the distribution and degree of conservation of RAD-tag loci in our datasets. In all three genome-assisted datasets, the left arm of chr-4 (chr-4L) had the greatest density of RAD-tag loci per megabase in the genome, while the highly-repetitive, heterochromatic right arm of chr-4 (chr-4R), the putative sex chromosome of *D. rerio* (Anderson, et al. 2012), had the fewest loci per megabase (Figure 2.2d). As mentioned previously, chr-4R contains a number of highly duplicated gene families, which limits our ability to make well-supported orthology inferences for RAD-tags within those loci. Thus, the paucity of chr-4R loci included in our datasets is partly due to excluding loci mapping to these gene families and does not reflect perfectly the distribution of RAD-seq loci across the zebrafish genome. In addition to chr-4R, other regions of the genome, particularly regions surrounding the centromeres of several chromosomes, also have fewer loci than the genome-wide average.

To find the proportion of SbfI-based RAD-tags occurring in exons, introns, UTRs, and intergenic regions, we compared the sequences of RAD-seq loci to sequences of genes annotated in Ensembl Zv9 (table 1). Results showed that RAD-tags in our datasets were more than four-fold enriched within and around coding sequences. Annotated transcripts from Zv9 (Ensembl version 75) contained 278 RAD-tag alignments per megabase compared to the average of 66 RAD-tag alignments per megabase in non-repetitive regions of the genome. Because SbfI recognizes a GC-rich octamer (CCTGCAGG; Figure 2.3a), the enrichment of loci around coding sequences is likely due in part to the high GC content of zebrafish transcripts (47.6% GC) relative to the AT-rich genome-wide average (38.6% GC). This enrichment for GC content continued outside of the SbfI sites in the RAD-tags themselves with observed GC content ranging from 47.6% to 49.3% across taxa. Conserved regions within coding sequences also affect the enrichment of SbfI sites around coding sequence. For instance, SbfI sites are enriched in the NOD-like receptor gene family, which includes 135 annotated paralogs on chr-4R and 190 annotated paralogs elsewhere in the genome. In particular, genes in subfamily C of the NOD-like receptors have repetitive exons with a highly conserved leucine-rich region that often contains an SbfI site (e.g., see Figure 2.3b).

Table 1. Data Set Characteristics



Data Set	Loci	Repeat Filtering	Total Positions	Parsimony Positions	% Loci				Species per Locus
					Exons	Introns	UTRs	Intergenic	
All mapped loci	89,593	RepBase	7,884,184	401,097	25.41	39.97	3.83	30.79	2.96
Min. Taxa	30,801	RepBase	2,710,488	401,097	26.56	40.42	3.86	29.17	6.31
Dre Group	3,406	RepBase	299,728	62,805	39.12	31.53	4.84	24.52	13.77
All Danios	1,720	RepBase	151,360	27,799	38.38	30.89	3.99	26.73	15.25
pyRAD	60,216	None	5,296,122	258,690	NA	NA	NA	NA	3.47

NOTE—NA, not available.

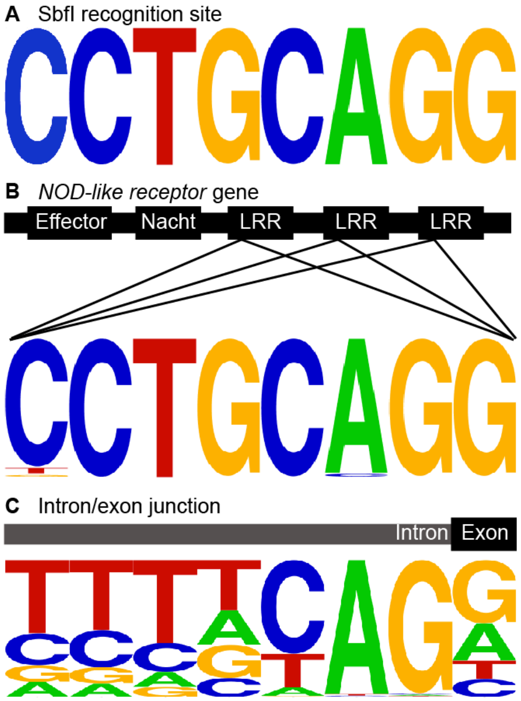


Figure 2.3. Enrichment of SbfI sites in the zebrafish genome. (*A*) The SbfI restriction cut site. *(B)* SbfI sites are enriched at splice acceptors due to the zebrafish consensus splice acceptor sequence. Base heights represent frequency of each base at each position. (*C*) SbfI sites often occur in NOD-like receptor genes (subfamily C) due to the consensus sequence of Leucine-rich repeat (LRR) domains.

We noticed that a striking number of RAD-tag loci were located at splice acceptor sites. *In silico* analysis of the zebrafish reference genome showed that SbfI sites are enriched about 1,600-fold at splice acceptor sites (738 observed vs. <0.5 expected assuming genome-wide nucleotide levels). We surmise that this is due to the similarity of the SbfI recognition site to the zebrafish splice acceptor motif (Figure 2.3c). Because each SbfI cut site has two associated RAD-tags, a cut site at a splice junction has one RAD-tag extending into the exon and its directly adjacent sister RAD-tag extending in the opposite direction into the intron. Realizing this situation allowed us to ask how well our approach recovered RAD-tag loci in exons compared to introns. Of the nearly 90,000 loci across all samples that had orthology to the zebrafish genome, 1,404 loci began at splice acceptors. Most of these loci (66.4% of 1,404 loci) occurred as pairs with both the exonic and intronic loci recovered in our dataset. When our dataset contained only the exonic or only the intronic locus, the exonic RAD-tag was obtained more often (25.9% of loci) than the intronic RAD-tag (7.7% of loci). We presume that this asymmetry arises because the higher rate of sequence conservation in exons than introns provided more frequent alignment of the exon partner of the RAD-tag pair to the zebrafish reference genome. In support of this hypothesis, the number of RAD-tags at splice acceptors with a confidently mapped intronic locus but no exonic locus generally occurred for duplicated genes. For example, of the 16 paralogs of *ms4a17a* in a 500 kb section of chr-4, only *ms4a17a.11* had a confidently mapped intronic RAD-tag at a splice acceptor. Recovery of the intronic rather than the exonic RAD-tag likely occurred due to our exclusion of exonic reads that mapped to multiple paralogs, while the corresponding intronic reads retained sufficient paralog-specific sequence to align to a unique spot in the reference genome.

Maximum Likelihood Phylogenetic Inference Based on Concatenated Datasets

Having identified orthologous loci, we constructed maximum likelihood (ML) trees from all three genome-assisted datasets and the genome-independent pyRAD dataset. All four datasets supported the same topology with most nodes having full bootstrap support (Figure 2.4). For nodes with less support, datasets containing more RAD-tags (Min. Taxa and pyRAD) had higher support than did datasets with fewer sites. All currently recognized species and genera were recovered as monophyletic clades across datasets and bootstraps. All four datasets support the same relationships for the outgroup species used in this study, demonstrating the potential for using RAD-seq on longer timescales. Consistent with previous studies, analysis placed the *Danio* species with the largest body size, *Danio dangila* and *Danio feegradei,* basally within the genus across all ML analyses. These two species formed a clade to the exclusion of all other danios in all bootstrap replicates of the Min. Taxa and pyRAD datasets as well as in the majority of bootstrap replicates of the other two datasets. The *D. choprae* species group was recovered basal to the *D. rerio* species group with full support across all datasets. Within the *D. rerio* species group, the *D. albolineatus* species subgroup had full support and was the most basally diverging. Because all datasets fully supported the monophyly of the remaining species in the *D. rerio* species group to the exclusion of the *D. albolineatus* species group, we conclude that the *D. albolineatus* species group is unlikely to constitute the sister group to *D. rerio* in contrast to previous findings (Fang, et al. 2009). RAD-tag analysis recovered the formerly unplaced *D. tinwini* as a close relative of *D. nigrofasciatus,* suggesting that reduced adult body size is a synapomorphy of these two species and evolved independently of reduced body size in *D. margaritatus* and *D. erythromicron* within the *D. choprae* species group.

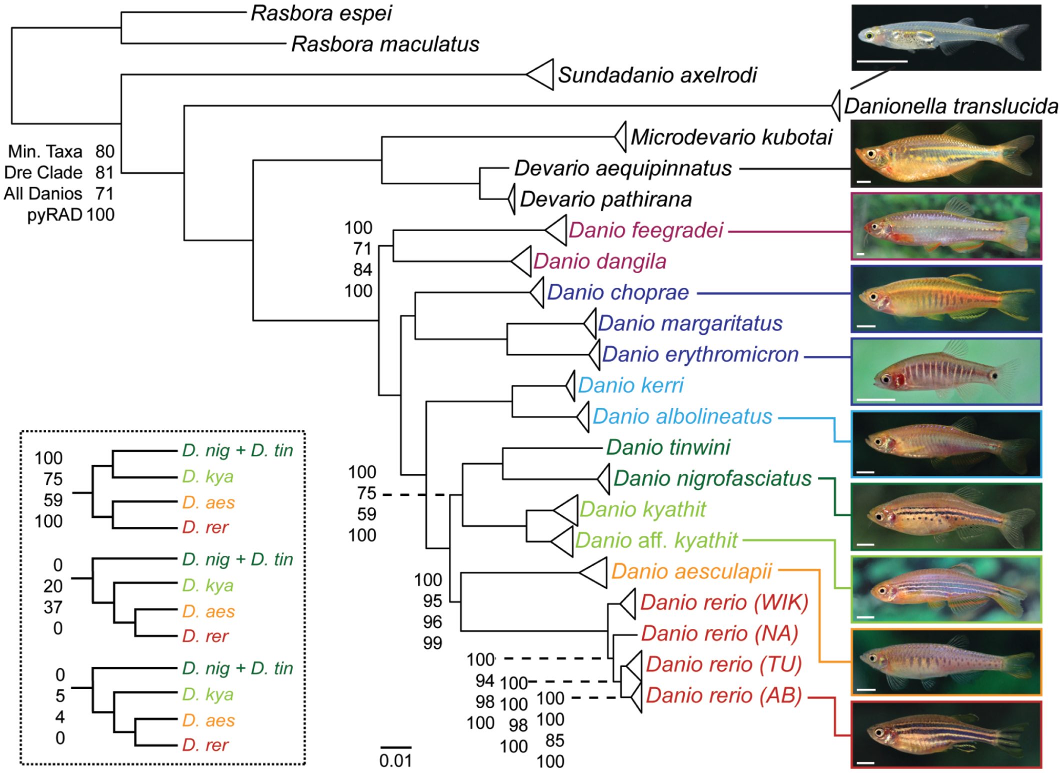


Figure 2.4. Maximum likelihood phylogeny of the *Danio* genus based on RAD-tag sequences. Unlabeled nodes have 100% bootstrap support across all datasets. Labeled nodes give the bootstrap support for the node in (from top to bottom) the Min. Taxa dataset, the Dre Group dataset, the All Danios dataset, and the pyRAD dataset. Branch lengths are based on the maximum likelihood analysis of Min. Taxa dataset, the largest genome-based dataset. The inset shows the support across datasets for the relationships within a subclade of the *D. rerio* species group. White scale bars in photos are 1 cm in length.

All four datasets recovered *D. aesculapii* as the sister species of *D. rerio*. This finding was surprising because, to our knowledge, this relationship had not been recovered in previous phylogenetic studies that included these species (Mayden, et al. 2007; Fang, et al. 2009; Tang, et al. 2010). Moreover, the adult pigmentation pattern of the two species is markedly different (see Figure 2.4).

Within *D. rerio*, we sequenced two lab strains (AB and Tübingen) and two wild strains (Nadia and WIK). Other groups have investigated the relatedness and population structure of zebrafish strains (Guryev, et al. 2006; Whiteley, et al. 2011), but the strains used in each study overlapped relatively little across studies. The relationships we recovered do, however, confirm the previous finding that AB and Tubingen are more closely related to each other than either is to WIK (Guryev, et al. 2006).

The remaining member of the *D. rerio* species group used in this study, *D. kyathit*, was originally described as having morphs with different pigment patterns—spotted or continuous stripes—with the holotype being conspicuously spotted (Fang 1998). The present study included a pair of spotted *Danio kyathit* individuals and a pair of striped individuals. We refer to the striped *D. kyathit* specimens as *Danio* aff. *kyathit* to denote the difference in pigmentation pattern relative to the holotype while recognizing that the type series of the species included one striped individual. Branch lengths separating striped *D.* aff. *kyathit* from spotted *D. kyathit* were greater than branch lengths separating any other conspecific samples, including the four *D. rerio* strains, but they are shorter than branch lengths separating any other two species in this study. The intermediate branch lengths separating *Danio kyathit* morphs offers no solid answer as to whether spotted and striped *D. kyathit* pigment morphs are different species. The relatedness and species status of the spotted and striped *D. kyathit* morphs and taxa from other studies—*Danio* aff. *kyathit* (Quigley, et al. 2004a)*, Danio kyathit* “spotted” (Tang, et al. 2010), and *Danio* sp. ‘Ozelot’ (Fang, et al. 2009; Tang, et al. 2010)—remain unclear and warrant further investigation by the examination of morphologies, vigor, and fertility of hybrids.

An additional issue that makes *D. kyathit* problematic is its variable phylogenetic location within the *D. rerio* species group in two of our four datasets. Analyses of the Min. Taxa and the pyRAD datasets placed *D. kyathit* as sister to the (*D. nigrofasciatus, D. tinwini*) clade with full support. The Dre Group and All Danios datasets also recovered this topology in the majority of bootstrap replicates, but without full support. Variant minority placements of *D. kyathit* in the Dre Group and All Danios datasets did not corroborate previous phylogenetic studies (Figure 2.4, inset). In our bootstrap replicates where *D. kyathit* did not fall with (*D. nigrofasciatus, D. tinwini*), it usually fell basal to the (*D. rerio, D. aesculapii*) clade (86% of the replicates) and infrequently as sister to *D. rerio* (14% of the replicates) but never as sister to *D. aesculapii*. The asymmetrical placement of *D. kyathit* with respect to *D. rerio* and *D. aesculapii* is striking and is investigated further in later subsections.

Bayesian Inference and Maximum Parsimony Analyses of Concatenated Datasets

Bayesian analysis of the Dre Group and the All Danios datasets supported the same topology as ML, with equivalent support for most terminal nodes (Supplementary Figure S.2.1, Supplementary Material online). Notably, Bayesian analyses of the Dre Group dataset and the All Danios dataset placed *D. kyathit* as sister to the (*D. nigrofasciatus, D. tinwini*) clade with posterior probabilities of 0.95 and 0.94 respectively. This result seems to contrast to the relatively weak support for this relationship obtained from these two datasets using ML (75% bootstrap support in the Dre Group dataset; 59% in the All Danios dataset). Close inspection of topologies sampled from the posterior distribution in these analyses showed that the reduced support for several deeper nodes, including even the base of genus *Danio*, is due to the erratic behavior of one of the outgroups, *Danionella translucida*. The only member of its genus included in the present study, this species was separated from all others by a long branch, consistent with previous studies, which inferred high mutation rates for *Danionella* species (Mayden, et al. 2007; Ruber, et al. 2007; Fang, et al. 2009; Tang, et al. 2010). The two *Danionella translucida* samples also had the highest proportion of unmappable RAD-tags and the fewest mapped loci of any taxon across all four datasets. Thus, the erratic nature of *Danionella translucida* in the current study is likely due to its failure to clear the two hurdles that RAD-seq faces for application to phylogenomics: sufficient sequence similarity to infer orthology across taxa and conservation of the restriction enzyme cut sites. The behavior of *Danionella translucida* reiterates the importance of thorough taxon sampling and the impact of long branches when using RAD-seq for phylogenomics. Bayesian analyses of the larger datasets (Min. Taxa dataset and the pyRAD Dataset) were not completed because the chains failed to converge after two weeks of running on mpi nodes and were estimated by MrBayes to take several months to complete.

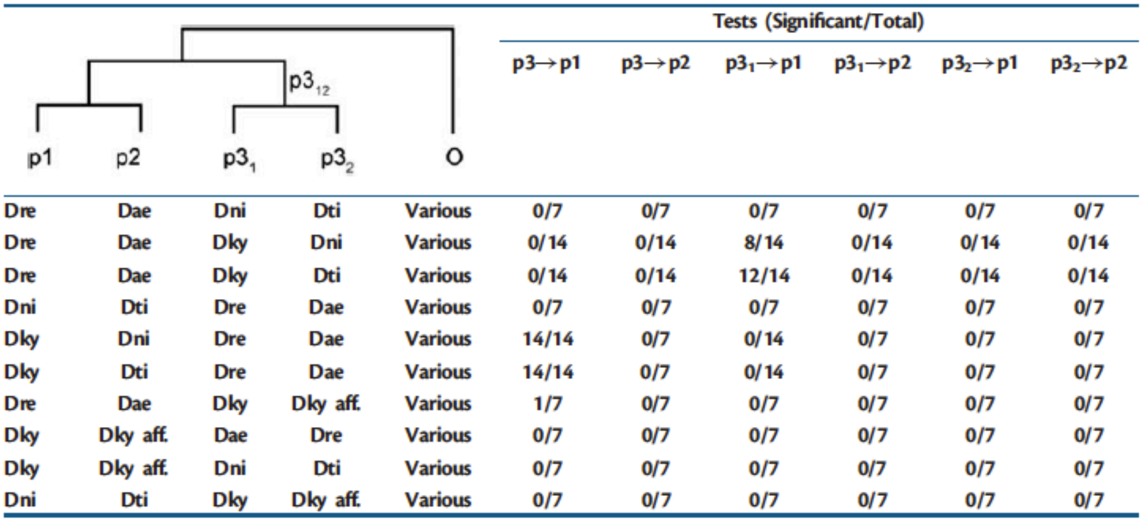
Topologies obtained through maximum parsimony (MP) analyses of the three genome-assisted datasets were nearly identical to the topology obtained by ML across all three datasets (supplementary Figure S.2.2, Supplementary Material online). Under ML, *D. feegradei* and *D. dangila* formed a monophyletic clade; in contrast, using MP, *D. feegradei* consistently fell basal to *D. dangila*. This minor difference between the findings of the two methods is not surprising given that the topology inferred using parsimony was recovered in some of the likelihood-based bootstrap analyses and because of the relatively long branches leading to *D. feegradei* and *D. dangila*, which can be particularly susceptible to long branch attraction, especially when using parsimony (Felsenstein 1978).

The few nodes lacking full bootstrap support using MP were the same nodes that lacked full support using ML. For instance, among the four sampled strains within *D. rerio*, MP and ML recovered WIK as the most basally diverging strain in most, but not all, bootstrap replicates of the Dre Group dataset and the All Danios dataset. The other two datasets, which each contained more than 250,000 parsimony-informative characters, unambiguously supported WIK as the most basally diverging *D. rerio* strain. Similarly, *D. kyathit* formed a clade with (*D. nigrofasciatus, D. tinwini*) in all four datasets, but this relationship lacked full support in the Dre group dataset (80% bootstrap support with MP; 75% with ML) and the All Danios dataset (67% bootstrap support with MP; 59% with ML). Consistent with our findings using ML, the variant topologies supported a (*D. kyathit,* (*D. rerio*, *D. aesculapii*)) clade.

Introgression Testing with pyRAD

A possible explanation for variability and asymmetry in the placement of *Danio kyathit* in our analyses of concatenated datasets is the process of introgression within the *D. rerio* species group. To test for past gene flow between the various taxa within the *Danio rerio* species group, we used a partitioned Patterson's D-statistic on the topology recovered with ML, MP, and Bayesian inference. This approach, which tests for an imbalance of character state frequencies on a hypothesized five-taxon topology with an outgroup and two pairs of sister taxa, has been used with RAD-tag sequences to measure unidirectional gene flow from outcrossing flowers into closely related incrossing species (Eaton and Ree 2013) and to test for past introgression in American oaks (Hipp, et al. 2014). To avoid biasing our analysis by testing for signs of gene flow on only a subset of topologies supported by our prior analyses, we tested all five-taxon topologies consistent with the topology recovered by maximum likelihood and Bayesian analysis. Because a partitioned Patterson's D-statistic can only be determined with four ingroup taxa, we sequentially excluded two of the six members of the *D. rerio* species group with variable placement in ML bootstrap replicates (*D. rerio*, *D. aesculapii*, *D. kyathit*, *D.* aff. *kyathit*, *D. nigrofasciatus*, and *D. tinwini*) while retaining the other four taxa as the ingroup taxa. The outgroup alternated between all more basally diverging danios (table 2).

Table 2. Partitioned Patterson's *D*-Statistic Tests



					Tests (Significant/Total)					
p1	p2	$p3_1$	$p3_2$	O	p3→p1	p3→p2	$p3_1$→p1	$p3_1$→p2	$p3_2$→p1	$p3_2$→p2
Dre	Dae	Dni	Dti	Various	0/7	0/7	0/7	0/7	0/7	0/7
Dre	Dae	Dky	Dni	Various	0/14	0/14	8/14	0/14	0/14	0/14
Dre	Dae	Dky	Dti	Various	0/14	0/14	12/14	0/14	0/14	0/14
Dni	Dti	Dre	Dae	Various	0/7	0/7	0/7	0/7	0/7	0/7
Dky	Dni	Dre	Dae	Various	14/14	0/7	0/14	0/7	0/7	0/7
Dky	Dti	Dre	Dae	Various	14/14	0/7	0/14	0/7	0/7	0/7
Dre	Dae	Dky	Dky aff.	Various	1/7	0/7	0/7	0/7	0/7	0/7
Dky	Dky aff.	Dae	Dre	Various	0/7	0/7	0/7	0/7	0/7	0/7
Dky	Dky aff.	Dni	Dti	Various	0/7	0/7	0/7	0/7	0/7	0/7
Dni	Dti	Dky	Dky aff.	Various	0/7	0/7	0/7	0/7	0/7	0/7

The observed partitioned Patterson's D-statistics provide strong support for two instances of past introgression in the recent history of the *Danio rerio* species group. Tests for introgression into *D. rerio* or *D. aesculapii* (rows 1-5 in table 2) provided evidence for introgression of alleles present in *D. kyathit* into *D. rerio* after its divergence from *D. aesculapii*. Tests for introgression of alleles originating in *D. rerio, D. aesculapii,* or their common ancestor (rows 6-10 in table 2) revealed strong evidence for introgression of alleles predating the divergence of the two species into *D. kyathit*. Analysis provided no evidence, however, for introgression between *D. kyathit* and *D. rerio* or *D. aesculapii* since the divergence of *D. kyathit* and *D.* aff. *kyathit* (rows 11-14 in table 2). Taken together, these tests support a model in which alleles present in a common ancestor of *D. rerio* and *D. aesculapii* were introduced into the common ancestor of *D. kyathit* and *D.* aff. *kyathit* after they diverged from *D. nigrofasciatus* and *D. tinwini.* Subsequently, after it diverged from *D. aesculapii,* the *D. rerio* lineage acquired alleles from an ancestor of *D. kyathit* and *D.* aff. *kyathit*. These instances of inferred introgression are consistent with the asymmetric results we observed previously in which *D. kyathit* occasionally appeared as sister to *D. rerio* or (*D. rerio*, *D. aesculapii*) but never as sister to *D. aesculapii* in the ML and MP analyses of the Dre Group dataset and the All Danios dataset.

Multi-Locus Phylogeny

Evidence for introgression from the partitioned Patterson's D-statistics demonstrated that a single bifurcating tree is unlikely to best describe the diversification of the *Danio* genus. Thus, while analyses of concatenated datasets can approximate underlying phylogenies, a more detailed methodology is required to surpass methods that assume a single underlying bifurcating topology. We therefore supplemented the concatenation-based phylogenomic approaches described above with BUCKy, a method based on posterior probabilities of topologies from multiple loci (Larget, et al. 2010). Given topologies sampled by Bayesian inference from the posterior distribution of topologies across all loci, BUCKy determines concordance factors (CFs) and confidence intervals (CI) for bipartitions in a topology based on the proportion of topologies that include each bipartition. We restricted our analysis to the *D. rerio* species group to reduce the effects of missing data while still maintaining sufficient loci for analysis and addressing the relationships of the species that interest us the most. All of our previous analyses concluded that the *D. albolineatus* species subgroup diverged basal to all other species in the *D. rerio* species group, so we therefore used *D. albolineatus* and the closely-related *D. kerri* to root the tree.

The BUCKy analysis recovered topologies across loci that are largely consistent with the phylogeny obtained from analyzing the concatenated datasets. Rather than supporting a single species tree, however, results of this multi-locus analysis supported a complex history for *D. rerio* and its closest relatives (Figure 2.5). Most individual RAD-tag loci had enough phylogenetic information to separate out closely related taxa as shown by the high CFs for the bipartition basal to the two members of the *D. albolineatus* species group (CF=0.978, 95% CI=(0.970,0.985)) and the bipartition basal to the two *D. kyathit* color morphs (CF=0.956, 95% CI=(0.943,0.967)). The bipartitions corresponding to the short internal nodes of the ML topology, however, had far less support than the long branches basal to the two *D. kyathit* color morphs and basal to the *D. albolineatus* species group, as would be expected under an instance of rapid speciation. When analyzing individual loci, *D. kyathit* fell as sister to the (*D. rerio, D. aesculapii*) clade (CF=0.246, 95% CI=(0.225,0.268)) slightly more often than the (*D. nigrofasciatus, D.*

tinwini) clade (CF=0.233, 95% CI=(0.211,0.258)) consistent with introgression from the ancestor of *D. rerio* and *D. aesculapii* into *D. kyathit* as inferred by the partitioned Patterson's D-statistics. Because BUCKy determines CFs for all bifurcations in the topologies sampled from the posterior distribution, we were also able to measure support for relationships that were not in concordance with the topology that was inferred using ML, MP, and Bayesian inference. Of particular note, *D. kyathit* was found to be sister to *D. rerio* (CF=0.207, 95% CI=(0.184,0.229)) significantly more often than *D. kyathit* was found to be sister to *D. aesculapii* (CF=0.144, 95% CI=(0.125,0.166)). The introgression of alleles from *D. kyathit* into *D. rerio* explains this asymmetry, as mentioned previously, and recapitulates results from our maximum likelihood analyses in which *D. kyathit* grouped with *D. rerio* in a small percentage of bootstrap replicates (5% in Dre Group dataset and 4% in All Danios dataset), but never with *D. aesculapii*. Due in part to the short length of RAD-tag loci and in part to recent divergence times of species within the *D. rerio* species group, many RAD-tag loci for these closely related taxa lacked strong phylogenetic signal and often supported equally two or more topologies. It remains unclear how often these topologies are the result of synapomorphic changes that support sister group relationships and how many are due to homoplastic changes, incomplete lineage sorting, or interspecies hybridization.

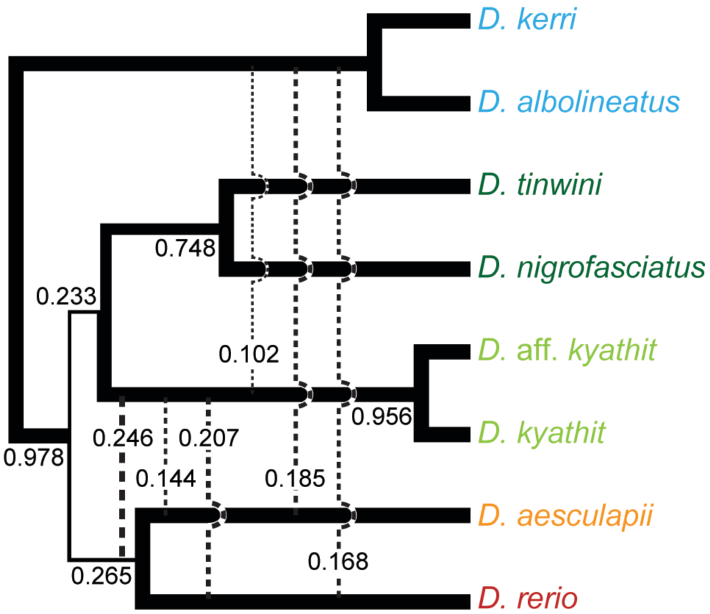


Figure 2.5. Bipartition frequencies of phylogenies inferred from Individual RAD-tag loci. Values represent Bayesian concordance factors for bipartitions. Solid lines show terminal branches and splits in the topology recovered by analysis of the concatenated datasets. Dashed lines that connect lineages show splits with concordance factors (CFs) larger than 0.10 but not concordant with the topology recovered by analysis of the concatenated datasets. Bipartition widths are proportional to their CFs.

Phylogeography of the Danio *Genus*

A more satisfying appreciation of *Danio* history can incorporate the group's phylogeny into its biogeography. *Danio* species occur across Southeast Asia, from northern India to Malaysia, with the majority of species being endemic to one or two major hydrological basins (Fang, 2009 and our Figure 2.6). *D. rerio* and *D. albolineatus*, however, have considerably wider ranges, distributed across a larger area than all other members of the genus combined, and yet the distributions of *D. rerio* and *D. albolineatus* do not overlap. *D. rerio* occurs in several hydrological basins across the Indian plateau as well as in the Ganges/Brahmaputra basin at elevations ranging from sea level to well over 1,000 meters (Engeszer, et al. 2007). *D. albolineatus* covers a similarly large range across several major hydrological basins from central Myanmar to southern Thailand. On the border of India and Myanmar, the Arakan Mountains separate the Ganges/Brahmaputra basin from the Irrawaddy basin and delimit the native ranges of a number of *Danio* species; no *Danio* species described east of the Arakan Mountains has been collected on the west of the Arakan Mountains and vice versa. Two apparent exceptions appear in the literature, but species designations for the relevant samples have since been amended (Fang 2000; Kullander and Fang 2009a). Hora (1937) designated several individuals collected from the west coast of Myanmar as *Danio choprae*, but these individuals were later assigned to *Danio aesculapii* in the first formal description of the latter species (Kullander and Fang 2009a). Similarly, specimens collected from the northwestern Irrawaddy basin were initially deemed to be *Danio rerio* (Chen, et al. 1988), but were later identified as *Danio albolineatus* (Fang 2000).

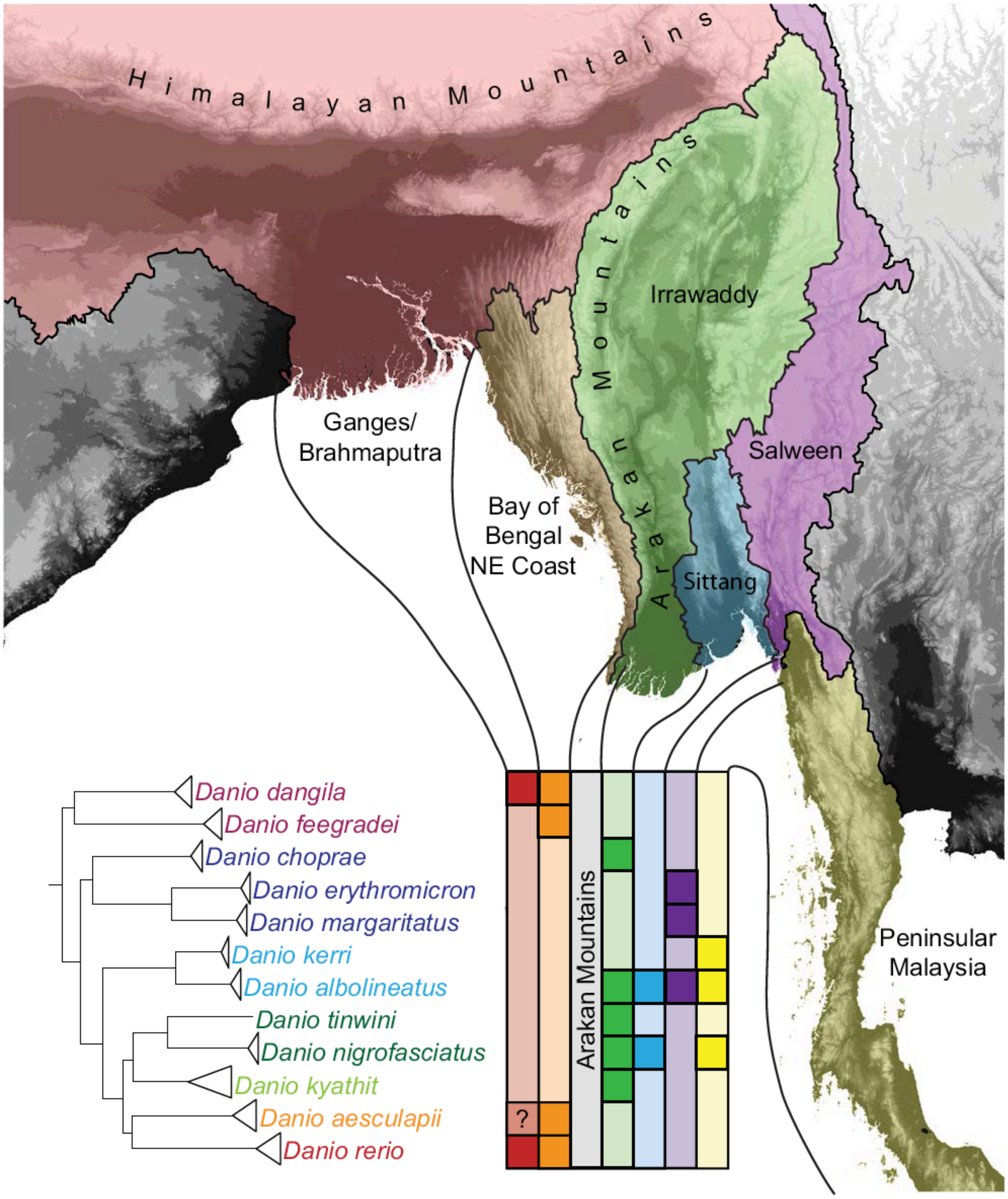


Figure 2.6. Phylogeography of *Danio* species. The Arakan Mountains of Myanmar separate *D. rerio*, *D. aesculapii*, and the large, basal *Danio* species from all other danios in this study. *D. rerio* occurs in several basins to the West of the area shown. *D. albolineatus* occurs in several basins to the East of the area shown.

Discussion

The Phylogeny of Danio

RAD-tag analysis of twelve species of *Danio* and seven outgroups recovered well-supported species relationships and suggested an explanation for the disparate results of previous phylogenetic studies of this group. Relationships among major clades within *Danio* (the basal danios, the *D. choprae* species group, and the *D. rerio* species group) are consistent across previous studies and the present work. Recovered associations within the *D. rerio* species group, however, vary across studies and involve nodes with low support both in previous studies and in our most restrictive dataset. Variable topologies and weak support for relationships within the *D. rerio* species group can be explained by two phenomena apparent in our RAD-tag data: 1) different trees for different loci helps explain variable topologies across studies, and 2) short branches at the base of the *D. rerio* species group explain weak support for relationships within the group.

The three recent molecular phylogenetic studies of *Danio* (Mayden, et al. 2007; Fang, et al. 2009; Tang, et al. 2010) varied not only in the topologies they inferred, but also in the numbers and types of loci investigated. Mayden (2007) used sequences from six genes but only the four mitochondrial genes, which are maternally inherited and segregate as a single locus, were sampled in more than one *Danio*. Fang (2009) used one mitochondrial gene and one nuclear gene. Tang (2010) used two mitochondrial genes and two nuclear genes. The RAD-tag data presented in this study are solely of nuclear origin. Thus, the results of Mayden (2007) reflect historical relationships of *Danio* mitochondria and the results of Fang (2009) and Tang (2010) are based on both mitochondrial and nuclear relationships; in contrast the RAD-tags results presented here reflect only the history of *Danio* nuclear genomes. Given the evidence for introgression and multiple gene trees uncovered in our data, it is not surprising that inferred topologies within *Danio* using small datasets are variable across studies because the gene trees from which the species trees were inferred may well have been different for each study.

The topology we recover is consistent with the distribution of species in various hydrological basins of Southeast Asia (Figure 2.6). All danios sampled in this study can be found in the same major hydrological basin as their closest known relative. Even *Danio choprae*, which appears to be an exception to this pattern, has a close relative native to the same basin—*Danio flagrans* the most recently described danio (Kullander 2012), which has yet to become widely available and was not included in our study. The range of *D. rerio* encompasses most of India and Bangladesh, extending as far east as the upper Indus basin and as far west as the Brahmaputra basin south of the Himalaya. The large, basal *Danio* species used in this study (*D. feegradei* and *D. dangila*) also occur in the eastern part of this range as well as in adjacent hydrological basins along the western coast of Myanmar in which *D. aesculapii* is also found (Kullander and Fang 2009a). Interestingly, an uncharacterized specimen referred to as *Danio* sp. "Bangladesh" appears closely related to *D. aesculapii*, but was collected in Bangladesh, where *D. rerio* also occurs (Tang, et al. 2010). Whether this individual represents a subspecies of *D. aesculapii*, a new species altogether, a hybrid between *D. rerio* and *D. aescualapii*, or an introduction remains unknown, but warrants further investigation. Two other *Danio* species—*Danio jaintianensis* (Sen 2007) and *Danio meghalayensis* (Sen and Day 1985)—have been described from the eastern extent of the *D. rerio* range, but little is known about how they are related to other species and whether or not they can form fertile hybrids with any of their congeners.

Potential Biases in RAD-seq Phylogenomics

Various groups have warned of potential biases that can occur when using phylogenomic datasets and RAD-seq data in particular for building phylogenies and inferring introgression. Recently, Roure (2013) showed that larger concatenated datasets with substantial amounts of missing data are more susceptible to phylogenetic artifacts than are smaller more complete datasets, particularly when using an inadequate model of sequence evolution (Roure, et al. 2013). To address this potential issue, we used ModelTest to identify the best model of sequence evolution for our RAD-tag data and used datasets with varying degrees of missing data (Min. Taxa, Dre Group, All Danios, and pyRAD datasets). We recovered the same topology for all of these datasets, although analyses based on the larger datasets provided more support for a few nodes (most notably at the base of the *D. rerio* species group). This increase in support with increasing dataset size would be expected if the topology for the *D. rerio* species group falls in the “Anomaly Zone” (Rosenberg and Tao 2008; Rosenberg 2013), another potential bias that can plague phylogenomic analyses using concatenated datasets. The results of our analysis of individual loci, however, would not be expected if the *Danio* species tree we recover was, in fact, anomalous. Instead, partitioned Patterson’s D-statistic tests for gene flow between taxa suggest that the lack of resolution within the *D. rerio* species group by concatenation-based methods is due to instances of introgression, which violate the assumption of a single underlying tree. The paper originally describing the partitioned Patterson's D-statistic test, however, suggested caution when using D-statistics for RAD-seq data due to the limited scope of the simulation studies upon which the test was based (Eaton and Ree 2013). Rather than using RAD-seq data, a recent phylogenomic study showing ancient introgression among fish species relied on contigs generated from transcriptome data (Cui, et al. 2013b). Although loci used for RAD-seq analysis tend to be considerably shorter than contigs assembled from transcriptome sequences and most RAD-tag loci do not fall in coding sequence, they have higher polymorphism rates than exons in transcripts and thus provide substantial rates of phylogenetically informative characters. Moreover, because D-statistic tests are based on the frequency of character states, the length of individual loci does not matter as long as orthology is accurately assigned. Because the inferred instances of introgression were highly significant and were recovered using multiple outgroup taxa, they are unlikely to be artifacts of incorrect orthology assignment. Thus, we conclude that our results are unlikely to be affected by two known sources of bias affecting phylogenomic datasets.

Methodological Considerations for Application of RAD-seq to Phylogenomics

RAD-seq has been successfully used to answer biogeographic and phylogenomic questions for short timescales (Emerson, et al. 2010; Eaton and Ree 2013; Jones, et al. 2013; Keller, et al. 2013; Wagner, et al. 2013; Martin and Feinstein 2014), but its utility on longer timescales has been questioned (Rubin, et al. 2012; Cariou, et al. 2013; Jones, et al. 2013) and remains largely unexplored. While a study has yet to explicitly estimate the root age of *Danio*, investigations estimating the root ages of various Cyprinid clades from relaxed molecular clock methods for the *cytB* gene placed the origin of genus *Danio* in the mid Miocene, at least 13 million years ago, and the last common ancestor of all taxa included in the present study at more than 31 million years ago (Ruber, et al. 2007). If *Danio* is indeed a 13 million years old genus, the present study represents, to the best of our knowledge, the longest timescale to which RAD-seq has been empirically applied in vertebrates. Despite the antiquity of this divergence time, we were still able to infer orthologous RAD-tags across both ingroup and outgroup taxa via both our genome-assisted method and genome-independent, *de novo* method. Application of RAD-seq on this timescale empirically demonstrates its utility for answering phylogenomic questions across larger phylogenetic distances than those to which it has previously been applied.

The number of RAD-tags that lacked discernable orthologs in other species increased in outgroup samples, as expected. With RAD-seq, as with all phylogenetic methods, dense taxon sampling is likely to improve results. As seen with our *Danionella translucida* samples, the lack of a close relative combined with a high mutation rate can result in unusable loci due to mutations in the restriction enzyme site on the branch leading to the divergent taxon, to the creation of new restriction enzyme sites in the divergent taxon, and to the inability to determine the orthology of loci in the divergent taxon to loci in other species due to a high number of mutations within individual loci.

Our study provides several key insights for designing and analyzing future phylogenetic investigations based on RAD-seq. First, we show that our analysis using a reference genome yielded results that are highly consistent with a similar genome-independent approach. Our validation of the reference genome independent approach is an important result given that the majority of organisms lack a close relative with a sequenced genome. Second, we analyzed the genomic positions of RAD-seq loci aligned to a reference genome and showed that mapping locations are not a random sample of the genome. Rather, when using SbfI*,* RAD-seq loci are enriched in and around exons, particularly at splice acceptor sites, in some repeat elements, and in certain conserved protein domains (e.g. leucine-rich repeat domains). A large proportion of the data, in some samples up to half of the mapped loci, are orthologous to repetitive regions of the zebrafish genome. These loci may warrant exclusion from phylogenomic investigations due to the uncertainty associated with separating alleles from paralogs within a sample and identifying orthologs across samples. We note, however, that the pyRAD analysis, which was agnostic as to the identity of annotated repeats, performed comparable to our largest genome-assisted analysis. This finding suggests that the substantial number of repetitive loci filtered out of the genome-assisted datasets due to similarity to repetitive elements did not ultimately impact the efficacy of the genome-independent pyRAD approach. The extent to which this finding will apply to other taxa, however, warrants further investigation in future RAD-seq studies of taxa closely related to other model species. Third, in our study, concatenated datasets with fewer, more conserved loci (e.g. the All Danios dataset) provided less support for certain species relationships than did concatenated datasets with more loci that were less conserved (e.g. Min. Taxa and pyRAD datasets). This increase in support with increased data is only advantageous if the concatenated datasets converge on the true species tree. When a single species tree is not the best descriptor of the group's history, using more data may result in looking past important historical relationships such as gene flow between species unless additional analyses are performed that do not assume a single species tree.

Conclusions

We provide the first well-supported phylogeny of the cyprinid genus *Danio* based on more than 30,000 nuclear RAD-tag loci from four strains of *Danio rerio*, eleven other danios, and seven closely related outgroup species. Our analyses of concatenated datasets all provide strong support for *D. aesculapii* as the sister species to *D. rerio*, a relationship that has not been recovered in previous phylogenetic studies. Tests for introgression and analysis of phylogenies based on individual loci revealed striking asymmetries that are inconsistent with a single bifurcating species tree, suggesting that the diversification of *Danio* involved instances of rapid speciation and introgression. The two inferred instances of introgression within the *D. rerio* species group both involve *D. rerio*, demonstrating the necessity of understanding the history of the genus as a whole to understand more fully this important model organism. Given the evidence we see for multiple topologies explaining the recent evolution of *Danio rerio* and its closest relatives, we infer that the seemingly incongruent results of previous phylogenetic studies, while likely to be correct gene trees for the loci included in each of those studies, are not accurate representations of the genome-wide species tree for *Danio.* To better understand the recent evolutionary history of zebrafish, future work will necessarily include: further phylogenomic inferences involving increased taxon sampling to characterize the mosaic history of its genome; whole genome sequencing of several other danios to identify recent structural rearrangements; hybrid studies and examination of natural isolates to explore the extent to which these species can and do interbreed; as well as cellular and developmental studies to understand the evolution of pigmentation, size, and other phenotypic differences among these species.

To varying degrees, all species in genus *Danio* share with zebrafish the biological characteristics that allowed *Danio rerio* to become a preeminent model organism (e.g. small size, high fecundity, externally developing transparent embryos, and ease of laboratory culture) so many designed for zebrafish can also be used to study its closest relatives. Because they can hybridize with zebrafish, the natural genetic variation in evolutionarily important traits that these species possess can be seen as an extension to the genetic resources available for zebrafish including induced null activity alleles in over 45% of its 26 thousand genes, more than any other vertebrate (Kettleborough, Busch-Nentwich et al. 2013, and http://www.sanger.ac.uk/Projects/D_rerio/zmp/). These features, along with the phylogenetic context we provide in the current work, promises to make *Danio* the premier vertebrate 'model genus'.

Materials and Methods

Animals

DNAs were collected from caudal fin clips of 41 individuals representing twelve *Danio* species and seven outgroup species. Taxa from the University of Oregon Fish Facility included: *Danio rerio* (AB), S#23336; *Danio rerio* (Tu), S#23891; *Danio rerio* (Nad), S#22583; WIK strain, S#23069; *Danio nigrofasciatus,* S#23139; and *Danio albolineatus,* S#23658. Taxa from Eugene Research Aquatics, LLC included: *Danio aesculapii*, S#ERA.Daes.1; *Danio dangila*, S#ERA.Ddan.1; and *Danio feegradei*, S#ERA.Dfee.1. Taxa from the aquarium trade included: *Danio kyathit*, WS24.Dkyp and WS25.Dkyp; Danio aff. kyathit, WS22.Dkyt and WS23.Dkyt; Danio tinwini, WS02.Dtin;

Danio kerri, WS08.Dker and WS20.Dker; Danio choprae, WS12.Dcho and NT01.Dcho; Danio margaritatus, WS05.Dmar and S#23036; Danio erythromicron, WS14.Dery and WS15.Dery; Danionella translucida, WS120729.DLAtr; Devario aequipinnatus, NT05.DEVaeq; Devario pathirana, WS120416.DEVpat; Microdevario kubotai, WS120416.MICkub; Sundadanio axelrodi, S#Q12824.SUNaxe; Rasbora espei, WS07.RASesp; and Rasbora maculata, WS10.RASmac. The University of Oregon Animal Care and Use Committee approved all protocols associated with this work. We compared mitochondrial sequences from our animals to COI and CytB sequences published in previous phylogenetic studies (Mayden, et al. 2007; Tang, et al. 2010) to confirm species identification.

Genomic DNA Extraction and Sequencing

We used the restriction enzyme SbfI-HF (New England Biolabs) to digest genomic DNA and ligated barcoded Illumina sequencing adapters to the four-base overhangs left by the enzyme. We sequenced RAD-tags flanking these sites for all samples on an Illumina HiSeq 2000 with single-end, 100 base pair reads. Sequences passed through several filtering steps prior to use for phylogenetic inference. To sort by sample barcode and exclude sequences without an SbfI site, we used *process_radtags.pl* from the *Stacks* software package (Catchen, et al. 2011) with the following parameters: -b barcodes.txt -e sbfI -E phred33 –D. Sequences containing Illumina adapters were excluded from analysis. Remaining sequences were run through *condetri* v 2.2 (Smeds and Kunstner 2011) to exclude reads with quality < 20 at any site. After these quality filtering steps, we obtained 1.0 million to 3.8 million reads per sample (supplementary table S.2.1, Supplementary Material online).

From caudal fin clips collected from euthanized or anesthetized adults of each species, we extracted genomic DNA and prepared sequencing libraries as described (Amores, et al. 2011), except that adapters had six-nucleotide barcode sequences and were optimized for sequencing on the Illumina HiSeq 2000. All barcodes differed by at least two nucleotides to prevent attribution of sequence to the wrong sample due to sequencing error in the barcode. Samples were sequenced in one lane of an Illumina HiSeq 2000 using single end 100-nucleotide reads. Three samples had low coverage (fewer than one million sequences) and were resequenced.

Locus Generation and Orthology Inference

For the Min. Taxa, Dre Group, and All Danios datasets, we used the *ustacks* program in the *Stacks* package (Catchen, et al. 2011) to merge RAD-tag alleles into loci within individuals allowing up to two mismatches (-M 2) between alleles. We enforced a minimum stack depth of three reads (-m 3) to account for possible read misattribution from other samples and rare sequencing artifacts. Repetitive and over-merged stacks were accounted for with the parameters (-r -d). For polymorphic loci, the consensus sequence was extracted from the output of *Stacks*. Following locus generation with *ustacks*, we removed sequences from repetitive regions of the Zebrafish genome with *RepeatMasker* v 3.3.0 (Smit, et al. 1996-2010) (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) using the zebrafish repeat database (-species danio). For *in* *silico* RAD-seq analysis of the zebrafish reference genome (Zv9 version 72), we extracted sequences flanking SbfI sites and ran them through the *ustacks* and *RepeatMasker* steps. In *ustacks*, the minimum stack depth requirement was removed to accommodate the 1x coverage of the *in silico* sequences.

For the genome-independent approach, quality-filtered sequences were run through *pyRAD* v 1.5.1 (Eaton and Ree 2013) with a minimum sample cutoff of four samples, a clustering threshold of 0.90, and a maximum of three heterozygous samples to exclude merging potentially paralogous loci. The 60,216 loci meeting these requirements were concatenated to form the pyRAD dataset.

In addition to using a completely *de novo* approach, we used the available and well-annotated *D. rerio* reference genome (Zv9 version 72) to define orthology. Quality-filtered RAD-tag loci from each sample were aligned against the zebrafish reference genome using GSNAP (Wu and Watanabe 2005). Several sets of parameters varying minimum percent identity to reference, indel penalty, and mismatch trimming were tested to accommodate the genetic distance of some species from the zebrafish reference genome (data not shown). Ultimately, parameters were chosen that maximized the number of reads across samples that mapped best to a single site in the genome with high support (mapping quality >30); these parameters were: -m 0.5 --indel-penalty=1 --trim-mismatch-score=0 --trim-indel-score=0 --max-middle-insertions=20 --max-middle-deletions=20 --max-end-insertions=20 --max-end-deletions=20. Aligned RAD-tag loci were inferred to be orthologous based on genomic location and were placed into a database using a custom python script (*gRad_parser_dynamic.py*, Supplementary Material online).

For each genomic locus with sequences from four or more samples, sequences were aligned against each other using *Muscle* v 3.8.31 (Edgar 2004) with default settings. The resulting alignments were trimmed of their SbfI restriction sites and concatenated to give the Min. Taxa, Dre Group, and All Danios datasets based on the number of samples and species possessing each locus. The Min. Taxa dataset contained all 30,801 loci present in at least four samples; the Dre Group dataset contained the 3,406 loci present in all members of the *D. rerio* species group; the All Danios dataset contained 1,720 loci present in all danio samples.

Phylogenetic inference

To determine the most appropriate model of sequence evolution for the datasets in this study, we employed *ModelTest* v 2.1.4 (Guindon and Gascuel 2003; Darriba, et al. 2012). Based on these results, maximum likelihood phylogenies were inferred under a GTR+I+**Γ** model in *RAxML* v 7.3.0 (Stamatakis 2006). Maximum parsimony analyses were also run with *RAxML* v 7.3.0. For Bayesian inference, we used *MrBayes* v 3.2.1 (Huelsenbeck, et al. 2001; Ronquist and Huelsenbeck 2003; Ronquist, et al. 2012) to analyze the Dre Group dataset and All Danios dataset. The Min. Taxa dataset and the pyRAD dataset were estimated to take several months to complete even using MPI nodes on the University of Oregon's super computer (http://aciss-computing.uoregon.edu). We allowed one million generations for burn-in, then sampled every thousand generations for ten million generations. For the multilocus analysis, we used *MrBayes* v 3.2.1 with the same parameters to sample posterior probability distributions for individual loci with sequence for all species in the *D. rerio* species group. These distributions were combined and analyzed in *BUCKy* v 1.4.2 (Larget, et al. 2010) with α = 1 and default settings. For the analysis of Patterson's D-statistic tests, we sampled and analyzed character states in the pyRAD dataset using the partitioned D-statistic test integrated into the *pyRAD* v 1.5.1 package (Eaton and Ree 2013).

Analyses of genomic features

We downloaded genomic feature files in *bed* format from Ensembl and UCSC using the BioMart tool and UCSC Table Browser respectively. We created our own *bed* files for genomic alignments using the *bamtobed* tool from the *bedtools* package (Quinlan and Hall 2010). To determine overlap between these various sets of genomic features, we used the *intersect* tool from the *bedtools* package (Quinlan and Hall 2010).

For analyses of splice acceptor sequences and Nod-like receptors, we downloaded the appropriate reference sequences from UCSC genome browser (splice acceptors) or Ensembl (Nod-like receptors). Nucleotide sequences for Nod-like receptors sequences were aligned with Muscle v 3.8.31 and the region corresponding to an SbfI site in most sequences was extracted. Base frequency graphics were generated using WebLogo (Crooks, et al. 2004).

Biogeographic analyses

We retrieved hydrological basin data from Aquastat, the Food and Agricultural Organization of the United Nations's global water information system (fao.org/nr/water/aquastat/gis/index.stm) and danio locality information from the Global

Biodiversity Information Facility (gbif.org) and imported them into *ArcGIS* (esri.com/software/arcgis) for visualization and comparison. Some additions and corrections were made to the locality information based on recent publications and redescriptions. Namely, two localities of *D. choprae* from the Western Ghats were corrected to *D. aesculapii* according to (Kullander and Fang 2009a).

Bridge

In contrast to phylogenetic studies based on few genetic loci, phylogenomic studies sample many loci across the genome to infer species relationships. Phylogenomic studies are, therefore, assumed to offer a better representation of the history of the genomes examined. In the next chapter, I describe a second phylogenomic study of genus *Danio,* wherein I use genome structure in addition to sequence information to infer a species history at odds with the history presented in Chapter II. Taken together, Chapters II and III provide a cautionary example of drawing conclusions of species histories agnostic to possible effects of genome structure.

CHAPTER III

A HYBRID HISTORY FOR ZEBRAFISH REVEALED BY PHYLOGENOMICS AND GENOME STRUCTURE

This work was coauthored by my advisor, Dr. John H. Postlethwait. I performed all experiments and analyses as well as the majority of the writing. Dr. Postlethwait contributed substantially to the experimental design, use of novel methodologies, interpretation of results, and editing.

Braedan M. McCluskey[1], and John H. Postlethwait[1]

[1] Institute of Neuroscience, University of Oregon, Eugene, OR 97403-1254

Abstract

Reticulate evolution due to introgression is increasingly recognized as an important process shaping the evolution of genomes in a variety of taxa. Here, we investigate the extent, causes, and implications of reticulate evolution in the genus *Danio*, which includes zebrafish, a major model organism. Using exome sequencing of ten *Danio* species and one outgroup, we find that historical relationships across the genus are clear with the exception of the relationships between zebrafish and its three closest relatives. By incorporating genome structure into phylogenetic analyses, we show that the genealogical discordance in this group depends on chromosome location. The centers of chromosomes consistently place *D. rerio* with *D. kyahit*, while distal portions of chromosomes consistently place *D. rerio* with *D. aesculapii.* Approximately half of the

genome supports each relationship. These observations are consistent with zebrafish being an ancient hybrid species, although other demographic scenarios could contribute to the observed pattern. Our knowledge of *Danio* natural history, genetics, and recent evolutionary theory suggest that variation in the rate of recombination across the genome may help explain the patterns we see repeated across chromosomes. The resolution provided by this study is a cautionary tale for phylogenomic studies in general, and has important implications for comparative studies across the genus *Danio* and our understanding of the recent evolutionary history of a major model organism.

Introduction

Phylogenomic analyses of a multitude of organisms have revealed that the genomes of many groups of taxa have heterogeneous histories. These groups include systems with ecology and mating systems predisposing them to gene flow (Eaton and Ree 2013; Hipp, et al. 2014; Fontaine, et al. 2015; Li, et al. 2016), groups that radiated rapidly (Meyer, et al. 2015; Prum, et al. 2015; Pease, et al. 2016), and key model organisms including drosophila (Clark, et al. 2007; Garrigan, et al. 2012) and primates (Scally, et al. 2012; Ting and Sterner 2013; Prufer, et al. 2014; Sankararaman, et al. 2014; Gordon, et al. 2016). Heterogeneous genome histories result from various forms of reticulate evolution that can occur when lineages speciate rapidly or when reproductively isolated lineages exchange genes. Processes proposed to generate reticulated genomic histories vary across studies due to varying biology of each system and patterns apparent in the genome. Incomplete lineage sorting of alleles (ILS) is a random process that occurs when populations bifurcate in quick succession and alleles segregating in the ancestral population are inherited in a pattern that does not reflect the history of population bifurcation (Holder, et al. 2001). All loci inherited via ILS will have a most recent common ancestor in the ancestral population before populations bifurcated. In contrast to ILS, introgression, the exchange of genetic material between diverged populations with incomplete reproductive isolation, is nonrandom and leaves a signature in the genome (Durand, et al. 2011; Eaton and Ree 2013; Supple, et al. 2013; Martin, et al. 2015; Li, et al. 2016). Regions of the genome that are introgressed will have a common ancestor after the initial divergence of the populations and share more derived alleles in introgressed regions than expected by the order of population bifurcation. In many systems with reticulate evolutionary histories, genome structure, namely chromosomal location, gene density and recombination rate, can greatly affect the likelihood of incomplete lineage sorting between closely spaced speciation events and the permeability of partially diverged genomes to introgression (Mallet 2005; Turner, et al. 2005; Ellegren, et al. 2012; Scally, et al. 2012; Burri, et al. 2015; Fontaine, et al. 2015). These effects can result in distinct patterns across the genome, which are not apparent without the incorporation of genomic structure.

Despite its status as a major model organism, several questions remain unanswered regarding the evolutionary history of zebrafish, *Danio rerio*, and its congeners. Previous phylogenetic studies differed on the inferred relationships within this group, in particular, the relationship of zebrafish and its several closest relatives (Meyer, et al. 1993; Fang 2003; Quigley, et al. 2004a; Mayden, et al. 2007; Mayden, et al. 2008; Fang, et al. 2009; Tang, et al. 2010; McCluskey and Postlethwait 2015). A recent phylogenomic study suggested that introgression involving the zebrafish lineage could explain the discordance across studies (McCluskey and Postlethwait 2015). As more groups use the genus *Danio* as a model for evolution (Parichy and Johnson 2001; Quigley, et al. 2005; Rosenthal and Ryan 2005; Mills, et al. 2007; Parichy 2007; Wong, et al. 2011; Camp, et al. 2012; Froelich, Galt, et al. 2013; Mahalwar, et al. 2014; McMenamin, et al. 2014b; Patterson, et al. 2014), it is increasingly important that we understand these relationships as interpretations of experimental results depend upon historical relationships among species and how those relationships vary across the genome. Here, we investigate how the genomes of zebrafish, *Danio rerio*, and its closest relatives were shaped by genome structure, incomplete lineage sorting, and introgression during their evolutionary history.

Results

Genealogical Heterogeneity within Danio

To investigate the evolutionary history of the *Danio* genus, we used a molecular hybridization-based exome-enrichment protocol with RNA baits complementary to annotated zebrafish genes. We sequenced exomes from ten danio species varying in body size, pigment pattern, and geographic distribution, as well as one species from the closely related *Devario* genus. We aligned exome-enriched Illumina reads to the zebrafish genome and examined only reads aligning to annotated APPRIS gene models (Rodriguez, et al. 2013), which represent the longest isoform of protein-coding genes and constitute 4.2% of the genome. For further analyses, we focused primarily on coding regions in these nuclear APPRIS gene models that were sequenced in all our species to at least 5x coverage. Concatenating these nuclear positions gives a 15.45 megabase alignment with no missing data and an average coverage of 23x per species. This alignment corresponds to approximately 1.1% of the nuclear genome, 26.5% of the total annotated APPRIS gene sequence in GRCz10 v82 (Kersey, et al. 2016), and more than 100 times more data than the comparable dataset used in the previous phylogenomic study with RAD-seq (McCluskey and Postlethwait 2015).

As a first approximation of the evolutionary history of these species, we built a Total Evidence Phylogeny (TEP) (Eernisse and Kluge 1993) using maximum likelihood for this 15.45 megabase alignment. This approach frequently used in phylogenomics (Cui, et al. 2013a; Fontaine, et al. 2015; Martin, et al. 2015) and implicitly assumes a single bifurcating model of genome evolution, an assumption that will be violated if these species experienced considerable gene flow. The inferred TEP topology (Figure 3.1a) agreed with recent phylogenetic (Mayden, et al. 2007; Fang, et al. 2009; Tang, et al. 2010) and phylogenomic (McCluskey and Postlethwait 2015) studies of genus *Danio* for the relationships of major groups within *Danio* ("TEP Topology" in Fig 3.2c). The relationships inferred for *D. rerio* and its closest relatives varied across previous studies, possibly due to differences in taxon sampled, limited loci used for analyses, and/or hypothesized gene flow (McCluskey and Postlethwait 2015). The relationships recovered in the TEP differ from these previous studies. The TEP supports *D. rerio* as the sister species of *D. aesculapii,* with *D. kyathit* basal to these two species. This topology is nearly identical to our previously published phylogenomic topology based on RAD-seq data (McCluskey and Postlethwait 2015), which grouped *D. kyathit* with *D. nigrofasciatus* ("RAD Topology" in Fig 3.2c). As with many large phylogenomic datasets, all nodes in the TEP were highly supported by bootstrap resampling and approximate likelihood ratio tests (Figure 3.1a). A maximum likelihood phylogeny inferred for positions from the mitochondrial genome recovered a topology that differed from the TEP, but had only limited support and was not significantly better than alternatives (Figure 3.1b). A poorly supported mitochondrial tree suggests that zebrafish and its close relatives speciated rapidly. The relationships within the TEP from nuclear sequence, however, are well supported and are likely to be true if these species do not have reticulate evolutionary histories.

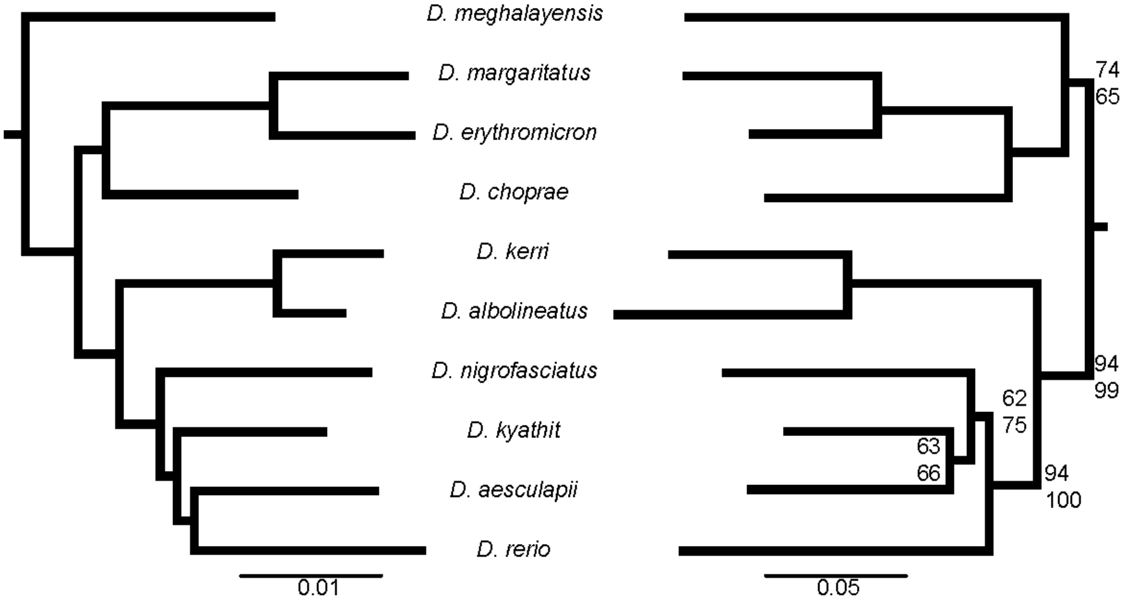


Figure 3.1. Total Evidence and Mitochondrial Phylogenies. (a) Total evidence phylogeny inferred using maximum likelihood analyses of 15.45 megabases of aligned nuclear coding sequence. All nodes have > 99% support from bootstrap and approximate likelihood ratio tests. (b) Mitochondiral phylogeny inferred using maximum likelihood analyses of five kilobases of aligned mitochondrial sequences. Bootstrap values are above nodes. Approximate likelihood ratio test results are below. Nodes with values of 100 for both measurements are not shown.

Widespread Genealogical Incongruence Mediated by Genome Structure

Given disagreements between our RAD-seq dataset and our exome dataset, it became necessary to map the extent of genealogical heterogeneity across the genome.

Recent studies have inferred phylogenies for adjacent genomic windows of equal size (Fontaine, et al. 2015; Pease, et al. 2016). In data sets such as exomes or transcriptomes, this results in drastically different amounts of sequence between windows (Pease, et al. 2016). To account for this phenomenon, we inferred maximum likelihood phylogenies for non-overlapping 10,000 nucleotide windows in the concatenated dataset rather than for windows standardized by a set genomic window. Each of the 1,532 non-overlapping 10-kilobase long windows contained coding sequence from one or more genes from the same region of a single zebrafish chromosome. Thus, these windows are all the same size in "transcript space," but represent different amounts of "genome space" due to variation in gene length and spacing across the genome. Using these "window trees" explicitly allow for variable histories across the genome. Comparing the window tree topologies, we saw that most nodes across window trees were consistent with the TEP (Fig 3.2a). Differences from the TEP were almost entirely restricted to the placement of *D. choprae* or variations in the relationships between *D. rerio* and its three closest relatives: *D. aesculapii, D. kyathit,* and *D. nigrofasciatus*. These variations could be due to ILS as these species rapidly diversified or due to introgression at a later point in the speciation process.

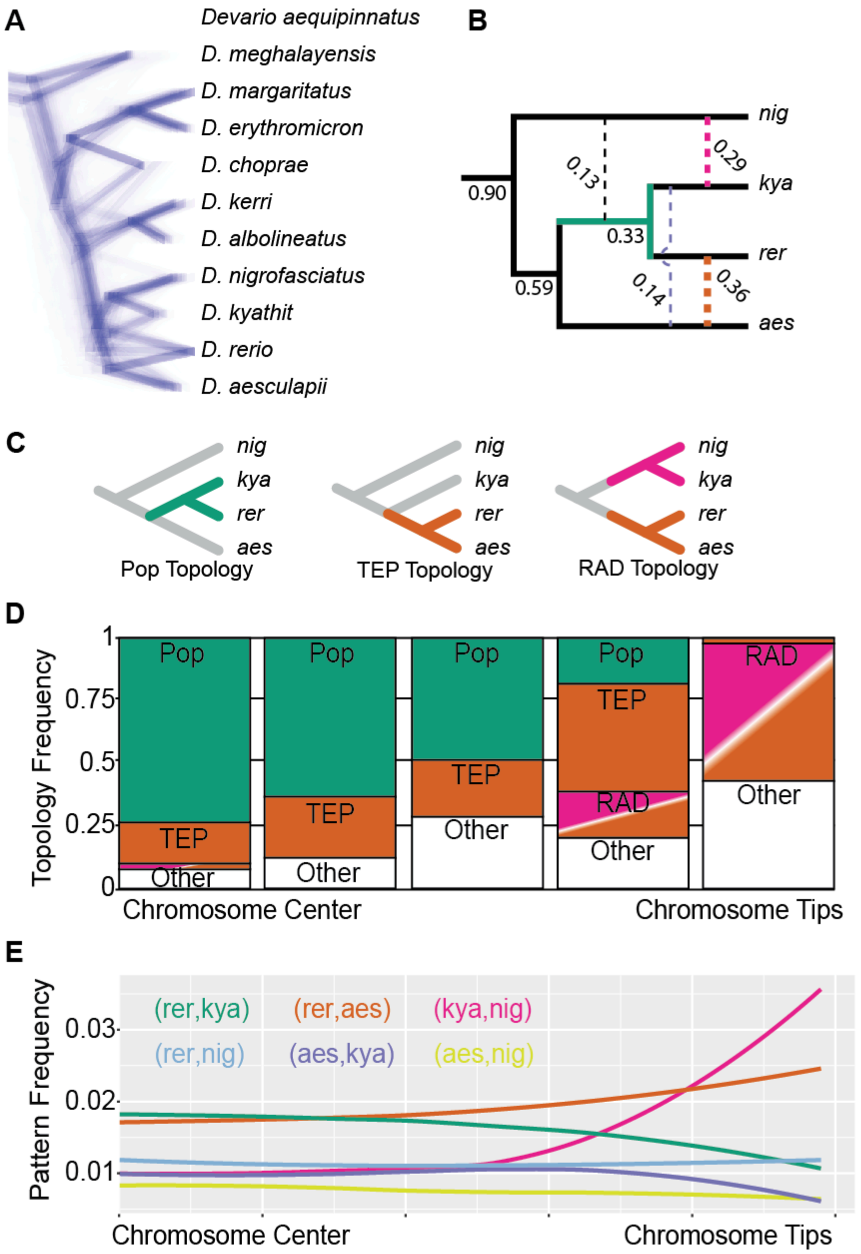


Figure 3.2. Genealogical history discordance and effects of chromosome structure. (a) Window trees constructed for 10,000 aligned coding positions from 1,532 adjacent exonic regions display genealogical heterogeneity as shown by a DensiTree plot of window trees. (b) Concordance factors for relationships between zebrafish and its closest relatives support three major pair-wise relationships. (c) The three most frequent topologies inferred for window trees are recovered at different frequencies at different positions along the chromosome as shown by frequency distributions (top) and DensiTree plots (bottom). (d) The frequency of shared, derived SNPs between two species follows a pattern similar to the topologies those relationships support. Frequencies are calculated using only variable sites and therefore robust to the effects of regions with low diversity.

We next performed a concordance analysis (Larget, et al. 2010) of the 1,532 window trees to infer the population tree that is most likely to produce the collection of window trees. Unlike the maximum likelihood methods that recovered the TEP and RAD topologies, this concordance analysis does not assume a strictly bifurcating species process, and is therefore a more accurate model of histories that include high levels of incomplete lineage sorting and/or introgression of alleles between species (Larget, et al. 2010). The population tree, which represents the history of a population most likely to contain the observed set of gene trees, inferred from concordance analyses did not match the TEP topology or the previously inferred RAD topology. Rather, the population topology placed *D. rerio* sister to *D. kyathit* with *D. aesculapii* sister to these two species ("Pop Topology" in Fig 3.2c). The short internal branch that separates these species on the population tree was estimated at 0.134 coalescent units, consistent with rapid subsequent lineage splitting events.

Recent studies have shown that genome structure can have considerable impact on the relatedness of taxa across the genome, particularly in recently formed or incipient species (Hohenlohe, et al. 2010; Ellegren, et al. 2012; Hohenlohe, et al. 2012; Burri, et al. 2015; Fontaine, et al. 2015; Li, et al. 2016; Pease, et al. 2016). To directly test if genome structure contributed to the heterogeneous genealogical histories between zebrafish and its three closest relatives in our window tree data set, we partitioned the alignments for each chromosome into 10 bins based on their relative location along the chromosome (i.e. the 10% of sequence closest to the left and right ends of Chr1 each constitute a partition). We then constructed maximum likelihood trees for all 250 bins (Fig 3.2d). Results showed that chromosome location had a drastic impact on whether the Pop, TEP, or RAD topologies was likely to be recovered. The Pop topology was most frequently recovered in the centers of chromosomes, while the TEP and RAD topologies were recovered most often at the ends of chromosomes. The Pop topology constituted more than 40% of the total topologies, which is more than the TEP and RAD topologies combined. Given the high frequency of the Pop topology in the topologies of these large data partitions and the results of the concordance analyses, which supported the Pop topology as the overall population tree, we assume the Pop topology represents the lineage branching order or "species tree" in further analyses.

To determine if the heterogeneous genealogical histories we observed are consistent with incomplete lineage sorting or introgression, we performed genome-wide D-statistic tests (Durand, et al. 2011), also known as ABBA/BABA tests, for the relationships supported by high concordance factors in the window tree analyses (Figure 3.2b). Both genome-wide D-statistic tests performed were highly significant and had magnitudes among the highest reported in vertebrates (Cui, et al. 2013a; Sankararaman, et al. 2014; Li, et al. 2016). These results support high levels of gene flow between the lineages leading to *D. rerio* and *D. aesculapii* ($D=0.223$, $p < 0.0001$) as well as between the lineages leading to *D. kyathit* and *D. nigrofasciatus* ($D=0.215$, $p < 0.0001$). Consistent with an effect of genome structure on introgression, the frequency of derived SNPs shared between taxa due to gene flow (i.e. ABBA patterns in the D-statistic test), are enriched at the ends of chromosomes (orange and magenta lines in Fig 3.2e). The enrichment of introgression patterns at the ends of chromosomes explains why the TEP and RAD topologies, which are supported by one or more of these introgression patterns, are found primarily at the ends of chromosomes. Taken together, the exceptionally high genome-wide D-statistics and the enrichment for introgression patterns at the ends of chromosomes suggests that the maximum likelihood trees corresponding to 1) the TEP in this study and 2) the RAD tree in (McCluskey & Postlethwait, 2015) were recovered due to high levels of introgression between lineages. While these topologies may support the history of the genome if it were inherited as a single locus, they likely do not reflect the lineage branching order of these species. Overall, these observations from window trees and SNP patterns demonstrate a heterogeneous history of the *Danio* genome that is affected by chromosome structure.

Genomic Regions Affected by Introgression

To visualize which regions of the genome have histories consistent with introgression and which supported the population branching order, we used the frequency of pairwise non-trivial splits (i.e. ABBA/BABA patterns) to generate chromoplots (Fontaine, et al. 2015) in addition to plotting the frequency of each pattern. Chromoplots include three species (and an outgroup taxon) and show which of three relationships between the three species has the most support in a genomic window. Line plots of the same data show how much support all three relationships have in a window. In these analyses, BBAA patterns support the species tree, ABBA patterns support the hypothesized introgression event, and BABA patterns support the other possible relationship, which is expected to show up at the same frequency as the ABBA pattern in the absence of introgression. Across a variety of taxa, known or hypothesized chromosomal inversions with a paraphyletic distribution (relative to the species tree) cause an accumulation of ABBA or BABA patterns in discrete genomic regions. Regions that experienced gene flow are often spread across chromosomes rather than localized to a few discrete regions, as is the case for structural variants (Hohenlohe, et al. 2010; Hohenlohe, et al. 2012; Fontaine, et al. 2015; Li, et al. 2016; Pease, et al. 2016).

To identify which regions of the genome experienced introgression, we performed three chromoplot comparisons. To determine the extent of genealogical heterogeneity across the genome not caused by introgression, we first performed chromoplot analysis for *D. rerio, D. aesculapii*, and the more distantly related *D. nigrofasciatus* (Fig 3.2a&b). The relationship supported by the Pop tree was recovered across most of the genome with the exception of some regions primarily near the ends of chromosomes including several contiguous windows on the left arm of Chromosome 20 (white box in Fig 3.2a). To determine which regions of the genome supported introgression between *D. kyathit* and *D. nigrofasciatus*, we next investigated relationships between *D. rerio, D. kyathit*, and *D. nigrofasciatus* (Fig 3.2c&d). The genomic history of these species showed a striking pattern with high support for the Pop tree in the centers of chromosomes and high support for the introgression pattern at the ends of chromosomes. Chromosome 4 proved to not follow this pattern. With the exception of several megabases at the distal end, the right half of Chromosome 4 is non-recombining, repeat rich, and full of duplicated gene families (Anderson, et al. 2012; Howe, et al. 2013; Wilson, et al. 2014). Despite spanning tens of megabases, because it does not recombine, this region is effectively a single locus near the end of the chromosome. Support for the BABA topology was largely restricted to the left arm of chromosome 20 including the region supporting the BABA topology in the previous comparison. Finally, to determine specifically which regions of the genome placed *D. rerio* with *D. kyathit* as in the Pop topology and which regions supported introgression between *D. rerio* and *D. aesculapii*, we performed chromoplot analyses for those species (Fig 3.2e&f). Just over half of all windows (51.2%) placed *D. rerio* with *D. kyathit* as predicted by the species tree, while 46.4% of windows supported the introgression hypothesis predicted by the TEP and RAD topologies. The introgression windows were enriched at the ends of chromosomes, albeit not as much as in the other introgression scenario. Interestingly, the left arm of chromosome 20 again had strong support for the BABA pattern in the same region as the other two comparisons (white box in Fig 3.2e). Across these comparisons, we see that topologies supporting introgression are located primarily at the ends of chromosomes, which correspond to regions of high recombination (Fig 3.2g), which may suggest a role for recombination in generating these patterns.

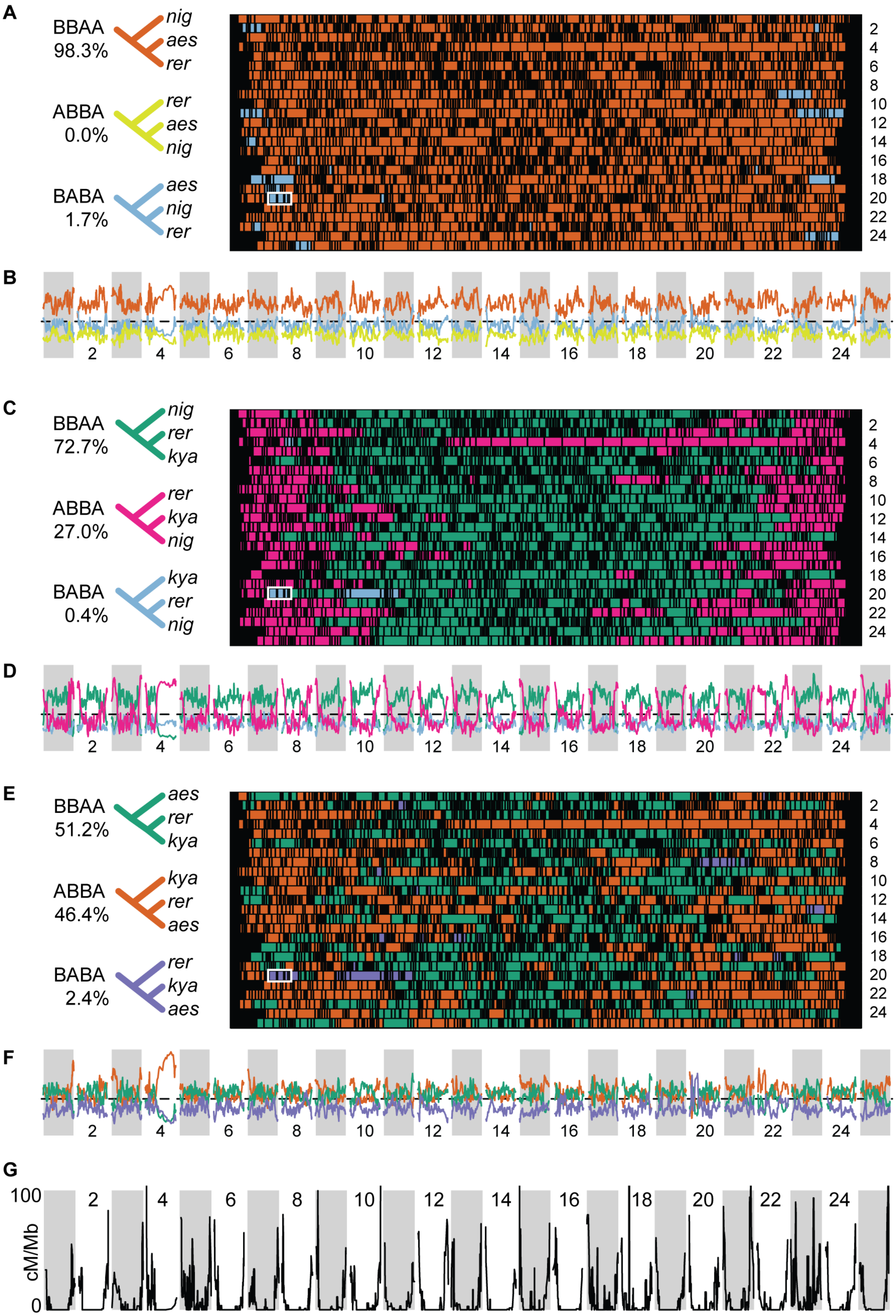

A
BBAA 98.3%
nig aes rer
ABBA 0.0%
rer aes nig
BABA 1.7%
aes nig rer
2 4 6 8 10 12 14 16 18 20 22 24
B
2 4 6 8 10 12 14 16 18 20 22 24
C
BBAA 72.7%
nig rer kya
ABBA 27.0%
rer kya nig
BABA 0.4%
kya rer nig
2 4 6 8 10 12 14 16 18 20 22 24
D
2 4 6 8 10 12 14 16 18 20 22 24
E
BBAA 51.2%
aes rer kya
ABBA 46.4%
kya rer aes
BABA 2.4%
rer kya aes
2 4 6 8 10 12 14 16 18 20 22 24
F
2 4 6 8 10 12 14 16 18 20 22 24
G
100
0
cM/Mb
2 4 6 8 10 12 14 16 18 20 22 24

Figure 3.3. Relationships of taxon trios vary based on chromosome position. (a, c, and e) Chromoplots are color-coded based on the pair of species that shares the most derived SNPs (BBAA, ABBA, and BABA sites) in a genomic window. The 25 chromosomes in zebrafish are stacked on top of each other. (b, d, and f) Line plots show the relative frequencies of BBAA, ABBA, and BABA patterns along chromosomes. Black line represents the null expectation of one third. Alternating gray and white boxes denote chromosomes arranged next to each other. The color of the highest line in a line plot corresponds to the color of that genomic window in the corresponding chromoplot. Window size for chromoplots and line plots is 5000 aligned base pairs jumping by 500. (a) Chromoplot and (b) line plot for relationships that are not expected to vary due to hypothesized introgression events. (c) Chromoplot and (d) line plot for relationships expected to vary due to hypothesized introgression between *D. kyathit* and *D. nigrofasciatus* (magenta). (e) Chromoplot and (f) line plot for the relationship of zebrafish and its two closest relatives. Green lines support the species tree, while orange lines support hypothesized introgression between *D. rerio* and *D. aesculapii.* Note the predicted inversion on Chromosome 20 (white box) that supports a minority topology in all three comparisons. (g) Recombination rate in male zebrafish is heavily biased to the ends of chromosomes. Data from (Anderson, et al. 2012).

In all three comparisons, four contiguous windows had low support for the BBAA and ABBA patterns, but high support for the BABA pattern, which occurs only at low frequency in other parts of the genome. Given the frequency of BABA windows across the three chromoplots, the probability of four windows co-occuring by chance across all three comparisons is exceptionally low ($p=4.30 \times 10^{-20}$). The patterns observed across comparisons support a discrete genomic region with a unique history different from the histories of the rest of the genome as has been seen for non-recombining chromosome regions in other systems (Hohenlohe, et al. 2010; Hohenlohe, et al. 2012; Fontaine, et al. 2015). The patterns observed suggest this region is a chromosomal inversion that was inherited as a single locus either via ILS as these species diversified or via an introgression event not detected in this study. If either of these scenarios is correct, a discrete segment of the *Danio* genome has a fourth history in addition to the Pop, TEP, and RAD topologies supported in other regions of the genome.

Discussion

A Hybrid History of the Zebrafish Genome

The relationships of zebrafish and its closest relatives display patterns of variation across the genome, in contrast to the relationships of most basal *Danio* species, which are largely unaffected by genome structure. Gene flow between these basal *Danio* species may warrant further investigation. For zebrafish and its closer congeners, the genome supports one of three topologies, which occur at different frequencies at genomic location relative to the centers and ends of chromosomes. We believe different genomic regions have different histories representing three demographic events. First, the shared ancestry of *D. rerio* with *D. kyathit* causes the centers of chromosomes to support the proposed species tree (Figure 3.4a) due to derived SNPs accumulated early in the divergence process that are inherited by both *D. rerio* and *D. kyathit*. Second, introgression between the *D. rerio* and *D. aesculapii* lineages causes support for the TEP topology in genomic regions biased toward the ends of chromosomes (Figure 3.4b). Third, introgression between the *D. rerio* and *D. aesculapii* lineages as well as between the *D. kyathit* and *D. nigrofasciatus* lineages results in support for the RAD topology almost exclusively at the ends of chromosomes (Figure 3.4c). The mutations accumulated prior to or during these demographic events result in a hybrid history across chromosomes (Figure 3.4d).

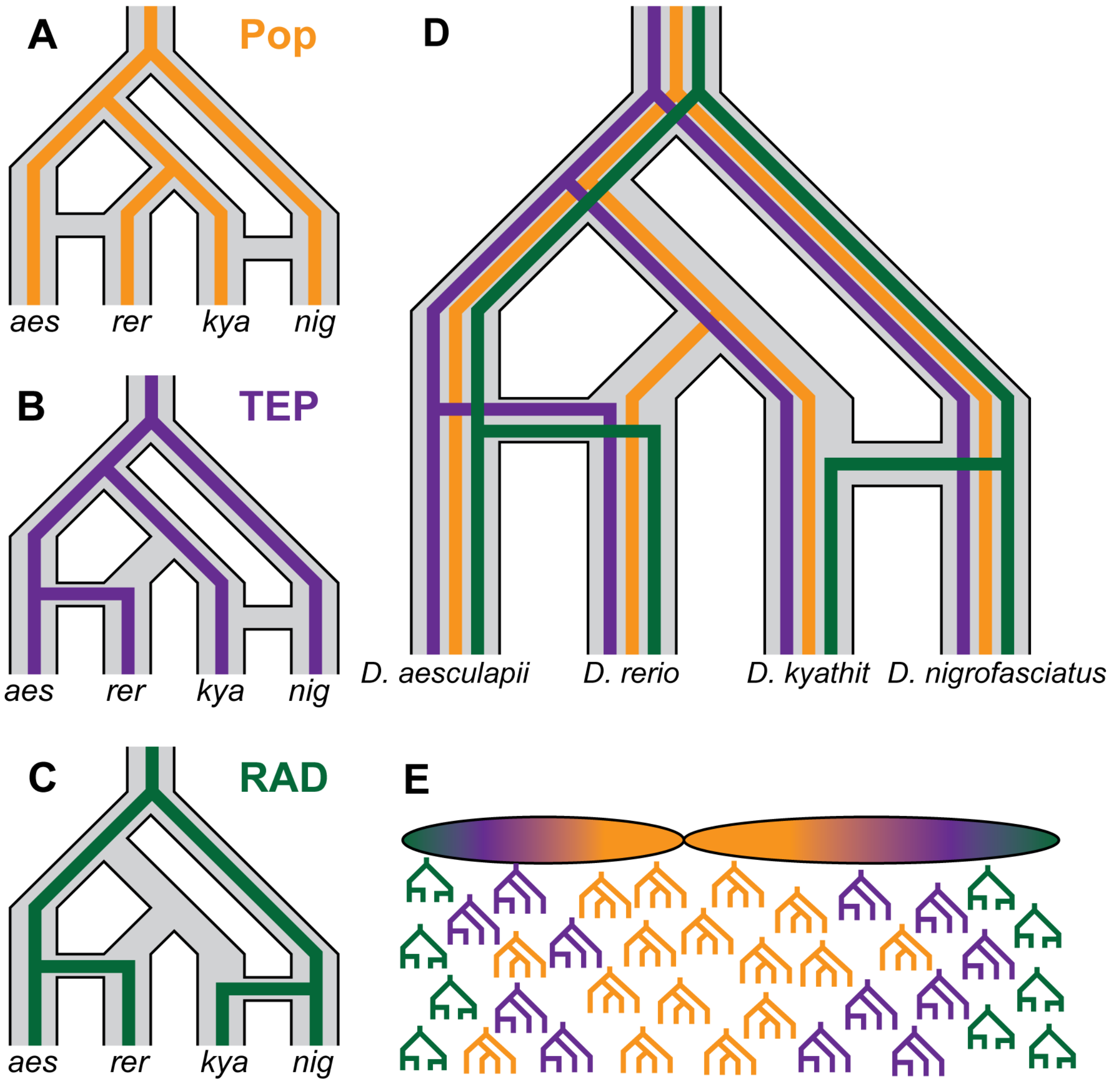


Figure 3.4. A model for the recent evolutionary history of zebrafish. The genomes of zebrafish and its close relatives support three distinct histories in different parts of the genome resulting from reticulate evolution due to introgression (shown in only one direction for simplicity). These histories include (a) the Population topology, or 'species tree', which reflects the initial bifurcating pattern of ancestral populations and occurs predominantly at the centers of chromosomes; (b) the TEP topology, which supports introgression between *D. rerio* and *D. aesculapii* and is biased toward the ends of chromosomes; and (c) the RAD topology, which supports two instances of introgression, the first between *D. rerio* and *D. aesculapii* and the second between *D. kyathit* and *D. nigrofasciatus*, and is found almost exclusively at the ends of chromosomes. (d) All of these histories occurred within a bifurcating population tree. (e) A representative chromosome shows a non-random distribution of topologies similar to what we observe in the *Danio* genome, with the Pop topology dominating near the center, the RAD topology dominating at the ends, and the TEP topology occurring between them.

What is the mechanism that gives rise to introgressed regions congregating at the ends of chromosomes? Incomplete lineage sorting and introgression can both lead to heterogeneous genome histories. Distinguishing between these two processes can be difficult (Holder, et al. 2001; Martin, et al. 2013b; Sankararaman, et al. 2014; Martin, et al. 2015; Li, et al. 2016; Pease, et al. 2016). Consistent with ILS, about 1% of variable nucleotide positions across the *Danio* genome have patterns across species that support relationships that disagree with the species tree and also disagree with both introgression events (Figure 3.2e). Two lines of evidence argue strongly against ILS driving the consistent patterns that distinguish chromosome centers from chromosome ends despite evidence for ILS across the genome, First, as evidenced by significant D-statistics, data show considerable bias for some relationships, which are expected to be random under an ILS scenario. Second, two sources of mutations (mutations occurring after the initial population divergence and mutations present in the ancestral population) generate character states concordant with the species tree (BBAA patterns), but only one source of mutations (mutations present in the ancestral population) generate ILS character states (ABBA and BABA). Thus, ILS patterns should not occur more frequently than the species tree pattern. At multiple chromosome ends in multiple comparisons, however, ABBA and BABA patterns occur more often than BBAA patterns, arguing against ILS being the only source of reticulate evolution in *Danio*.

Unlike an ILS scenario, introgression can potentially explain the spatial distribution of different genomic histories if we incorporate our knowledge of *Danio* biogeography, zebrafish genetics, and evolutionary theory. For introgression to occur, lineages must co-occur and be able to hybridize. Unlike the hypothesized introgression scenarios suggested in (McCluskey and Postlethwait 2015), which involved species from different hydrological drainages (*D. rerio* and *D. kyathit*), the introgression events presented herein involve species pairs with distributions that currently overlap. Populations derived from the *D. rerio* and *D. aesculapii* lineages occur in drainages west of the Arakan mountains and overlap in Bangladesh (Kullander and Fang 2009a; Tang, et al. 2010; Whiteley, et al. 2011). *D. kyathit* and *D. nigrofasciatus* occur in the Irrawaddy drainage in Myanmar to the east of the Arakan mountains (Fang 1998; Kullander and Fang 2009b).

Several *Danio* species hybridize in captivity (Collins, et al. 2012) and a variety of hybrids have been produced *in vitro* (Parichy and Johnson 2001; Quigley and Parichy 2002; Quigley, et al. 2005; McMenamin, et al. 2014a). While many hybrids between different *Danio* species appear to be sterile, some species pairs generate hybrids with low fertility when crossed to other hybrids or a parental species (Parichy and Johnson 2001). These reproductive incompatibilities may not have been present between different ancestral lineages in the diversification of *Danio*. Additionally, hybrid clutches are extremely male-biased (personal observation) fitting Haldane's rule for a likely ZZ/ZW system (Anderson, et al. 2012; Wilson, et al. 2014). These observations suggest a scenario wherein introgression occurred through a male hybrid with recombination between chromosomes during meiosis in the hybrid occurring similar to their distribution in male zebrafish, (but not in chromosomal inversions with different orientations in the two genomes in the hybrid). We know from decades of zebrafish genetics that recombination events are enriched at the ends of chromosomes (Postlethwait, et al. 1994; Johnson, et al. 1996; Knapik, et al. 1998; Gates, et al. 1999; Geisler, et al. 1999; Shimoda, et al. 1999; Woods, et al. 2000; Hukriede, et al. 2001; Woods, et al. 2005; Bradley, et al. 2011; Patowary, et al. 2013); in males, recombination is almost exclusive to the ends of chromosomes (Anderson, et al. 2012; Howe, et al. 2013; Wilson, et al. 2014) matching the distribution of introgressed genomic regions we observe (Figure 3.2g).

Population genetic theory has long noted the effects of recombination rate on nucleotide variation within populations (Felsenstein 1974; Hill and Robertson 2007). The effects of recombination rate on longer timescales (i.e. between species), however, remain subject to debate (Turner, et al. 2005; Noor and Bennett 2009; Turner and Hahn 2010). Orthologous regions of the genomes in two populations will diverge in sequence due to mutations accumulated in each population. Gene flow between populations breaks down this divergence through recombination, bringing together genetic variants from both populations. Recombination can be completely abolished by geographical separation of the populations or by changes in chromosome structure (e.g. inversions). Recombination can also be decreased if hybrid fitness is reduced through Bateson-Dobzhansky-Müller incompatibilities (Orr 1995; Presgraves 2010), or adaptation to different selective pressures in the two populations (Cutter 2012). If selection acts on a genomic region, genomic divergence is expected to spread to linked genomic regions via genetic hitchhiking (Smith and Haigh 2007). These regions affected by "linked selection" will be larger in genomic regions with reduced recombination, such as the centers of *Danio* chromosomes. Moreover, selection is expected to act more frequently in gene dense regions, such as the centers of *Danio* chromosomes, with more targets for selection. Taken together, when genomes from two *Danio* lineages came into contact via hybridization, selection likely affected genomic regions differently based on chromosomal location. Genes near the centers of chromosomes would be inherited along with hundreds of other genes that had experienced different selection pressures in the two lineages. Genes near chromosome ends, however, would be initially inherited in smaller linkage blocks, which would be broken down further in subsequent generations. These different selection pressures may explain the structured genomic histories we see across chromosomes. This interpretation, however, is tentative as similar patterns of heterogeneous genomic histories can also been attributed to diverse other biological processes. In previous studies, nearly equivalent levels of evidence for two different relationships have been interpreted, for instance, as 1) due to speciation with gene flow (Martin, et al. 2013b) in cases with little structure across the genome, 2) due to hybrid speciation (formation of a lineage reproductively isolated from its two parent lineages) when genome structure was not accounted for (Jones, et al. 2013; Kang, et al. 2013), and 3) due primarily to ILS during several instances of rapid speciation (Scally, et al. 2012; Meyer, et al. 2015). A recent study across many tomato species investigated ILS and introgression using chromoplots and window trees similar to our analyses (Pease, et al. 2016). Interestingly, in this study, a species trio including a taxon (*hua-1360*) previously shown to be admixed (Labate, et al. 2014), displayed a pattern across chromosomes similar to what we see in the present study, with high support for one relationship primarily in the center of chromosomes and high support for another relationship biased to the ends of chromosomes. While not initially interpreted as due to differences in gene flow mediated by chromosome structure, this similar result corroborates our interpretation that the pattern we describe is due to hybridization and may be applicable across broad taxonomic scales.

Implications for Phylogenomic Studies

The structured relationships we see across the *Danio* genome have implications for the interpretation of past and future phylogenetic studies. For example, the multiple histories of the *Danio* genome may explain the discordance between previous phylogenetic studies involving the *Danio* genus (Mayden, et al. 2007; Fang, et al. 2009; Tang, et al. 2010; McCluskey and Postlethwait 2015). Several molecular phylogenetic studies were based on loci from nuclear and mitochondrial DNA. Moreover, the nuclear loci used in these and other phylogenetic studies occur on different parts of the chromosome. In particular, Rhodopsin, a locus used in several of these *Danio* phylogenetic studies, occurs near the end of chromosome 8 in *D. rerio* and is therefore more likely to be subject to introgression than nuclear loci residing near the centers of chromosomes. When histories vary across the genome and studies use just a few loci for inference, inferred relationships will be subject to stochastic sampling bias based on the location of loci used for inference. When large phylogenomic datasets are used, variation from stochastic sampling will be negligible. However, without incorporating genome structure, strong systematic biases in these data sets can lead to different inferred relationships than when genome structure is incorporated into analyses. In our present study, for example, the inferred species tree (Pop topology) based on the most frequent topology obtained from hundreds of window trees differed from the TEP topology, which assumed a single history of the genome. Both of these topologies differed from the RAD topology inferred in our previous phylogenomic study (McCluskey and Postlethwait 2015). A likely cause of this discordance is that the RAD loci used in our previous study occurred disproportionately towards the ends of chromosomes, which are enriched for

SNPs supporting the RAD topology. Drawing inferences under the supposition of different species trees, such as trees inferred from different phylogenomic datasets, can have considerable implications on inferred evolutionary origins of species-specific biology in the taxa examined (Thomas and Hahn 2015). As a cautionary example of this principle, the maximum likelihood topology inferred in the RAD study was the maximum likelihood topology for only 5.4% of genomic windows in this study, which showed an extreme bias toward the ends of chromosomes. Drawing inferences about the source and direction of introgression events based on this RAD topology would have led to very different results than obtained in the present dataset.

This cautionary example is just one instance of how genome-agnostic approaches can lead to biologically implausible inferences. Vertebrate genomes do not evolve as a single locus. Mitochondria, autosomes, and regions near sex-associated loci all have different modes of inheritance. Furthermore, recombination rates in zebrafish and other species vary drastically across the genome and between sexes. These factors, in addition to random processes like incomplete lineage sorting of alleles and deterministic processes like purifying and positive selection, all affect how various lineages inherit different parts of the genome (Hill and Robertson 2007; Smith and Haigh 2007; Scally, et al. 2012). Many current phylogenomic methods for understanding how species are related make the implicit assumption that a genome has a single "species tree" inferred from treating the entire genome as a single locus that is expected to represent the one true lineage divergence pattern. Variation from this "Total Evidence Phylogeny" is interpreted as due to incomplete lineage sorting or introgression. In many scenarios, these assumptions may be valid. In cases such as *Danio*, however, a genome-agnostic view can be dominated by high levels of phylogenetic information in introgressed regions of the genome, which do not represent the historical lineage divergence order.

Conclusions

Analysis of our exome sequence data has illuminated some of the complex recent evolutionary history of zebrafish and raises new questions about the evolution of this model genus. With current taxon sampling and the topology of the inferred species tree (Pop topology in Figure 3.2b), we cannot determine the direction of introgression or the proportion of the genome that experienced introgression using conventional methods (Durand, et al. 2011; Martin, et al. 2015; Pease, et al. 2016), which require comparison of two pairs of sister species (as in the RAD topology in Figure 3.2b). This problem leaves key questions about the history of zebrafish unresolved. On one end of the spectrum, introgression was unidirectional with the zebrafish lineage donated genes to other lineages, but not receiving any introgressed alleles. In this extreme scenario, all of the zebrafish genome could have been theoretically passed through time in a single population. On the other end of the spectrum, zebrafish represents an ancient hybrid species with about half of its genome closely related to *D. kyathit* and half of its genome introgressed from the lineage leading to *D. aesculapii*. Future analyses incorporating more taxa, population level sampling across species, and whole genomic analyses of multiple individuals from many taxa may help to resolve these questions as has been done in other vertebrate models of evolution (Schumer, et al. 2013). Additionally, due to the extreme genealogical diversity in this group, interpreting the evolution of phenotypes across species should be done with caution, as phenotypes may be caused by genes with histories differing from the species tree, an increasingly recognized phenomenon recently termed “hemiplasy” (Hahn and Nakhleh 2016). Hemiplasy explains many instances of pigmentation patterns that have paraphyletic distributions across species (Cui, et al. 2013b; Martin, et al. 2013a; Stankowski and Streisfeld 2015), suggesting some phenotypes may be more prone to the process. The results we find suggest that genes at the ends of *Danio* chromosomes will also be predisposed to hemiplasy due to increased levels of gene flow. These biases may prove important for understanding the evolution of diverse pigmentation patterns in danios and other fishes (Innes 1956; Geisel 1960; Quigley and Parichy 2002; Quigley, et al. 2005; Ekker, et al. 2008; McMenamin, et al. 2014a; Patterson, et al. 2014) and phenotypic evolution in general.

In addition to clarifying the history of zebrafish, we show here that under complex demographic scenarios, genome agnostic approaches initially based on results treating the genome as a single locus (e.g. “Total Evidence Phylogenies” from maximum likelihood analyses using concatenated datasets) can lead to biologically implausible results if the phylogenetic signal in the data set is dominated by patterns resulting from introgression. For instance, using as our species tree the previously published RAD topology, which placed *D. kyathit* closer to the *D. nigrofasciatus* than to *D. rerio*, we would have inferred that the centers of all 25 chromosomes in *D. kyathit* were introgressed from the *D. rerio* lineage, while introgression at the ends of chromosomes was limited. Combining genome structure with the data from past phylogenomic studies will likely lead to more insight into the natural history of species investigated. When possible, genome structure should be incorporated into phylogenomic studies, especially in instances where historical introgression is likely.

Results of this study leave many details of the evolutionary history of zebrafish unanswered, in particular the identity of the closest extant relative of *Danio rerio*. The answer depends on how we define “closest relative”. For species investigated here, the majority of the genome in terms of megabases supports *D. rerio* and *D. kyathit* as sister species. However, the *D. rerio* lineage likely exchanged genes most recently via introgression with the *D. aesculapii* lineage. Confusing the matter even more, to our knowledge, the only species demonstrated to be reproductively compatible (i.e. producing fertile hybrids) with *D. rerio* is *D. nigrofasciatus* (Parichy and Johnson 2001). Alternatively, a species not included in the study or an undescribed species could be the “closest relative” of zebrafish. Regardless of how we ask the question, the answer is clearly not black and white.

Materials and Methods

Library preparation and Exome Alignment Creation

DNAs were collected from caudal fin clips or larvae of ten *Danio* species and one outgroup species (Figure 3.1a). The University of Oregon Animal Care and Use Committee approved all protocols associated with this work. We extracted and purified genomic DNA using a Blood and Tissue Kit (Qiagen). We constructed libraries and performed exome enrichment with SureSelectXT2 RNA Baits (Agilent) according to manufacturer’s instructions. The exome-enriched libraries were quantified using a Qubit fluorimeter and sequenced on the Illumina HiSeq 2500 (SE100 bp reads) and the Illumina NextSeq (SE75 bp reads). We quality filtered Illumina reads with Trimmomatic using a per base quality score minimum set to 20 (1% error rate) and minimum read length of 30 nucleotides. To address the sequence divergence between zebrafish and other species, we used bbmap (Bushnell), a global alignment algorithm with permissive parameters, but requiring single best alignments to the zebrafish genome (GRCz10 v 82). BBMap parameters were: ambiguous=best minidentity=0.70 maxindel=100 idtag=t k=12. From these alignments, we called variants using vcftools (Danecek, et al. 2011). We generated a custom python script to parse these variant call files. The script takes as input exon locations (in this case, APPRIS transcripts as a .bed file), one VCF file per species, a minimum depth, and a minimum genotyping quality score. The script parses through the VCF file for all positions in the .bed file. Positions near any indels are marked as "N" (ambiguous nucleotides) in order to avoid alignment problems. The script also converts to "N" all VCF positions below the minimum depth or genotyping quality cut offs. For all other sites, the script assigns a consensus base from the .vcf genotype field. Once all per-species consensus sequences are called, the script parses sequences in all species, and converts all codons with an "N" at any codon position in any species to "N" in for that codon in all species. The script outputs FASTA or Phylip files based on chromosomal location as specified in the .bed file.

Maximum Likelihood Phylogenetic analyses

To infer phylogenies for different genomic windows, we concatenated alignments from the same chromosome and divided each chromosome's alignment into 1) 10 kb non-overlapping windows starting at the ends of each chromosome used to generate window trees in Figure 3.2a, and 2) alignments representing 10% of the total sequence for a chromosome relative to the chromosome ends and center for inferring the phylogenies represented in Figure 3.2d. For all windows, the entire concatenated alignment, and the mitochondrial alignment, we inferred maximum likelihood phylogenies and approximate likelihood ratio tests under a GTR+I+**Γ** model in *RAxML* v 8.2.3 (Stamatakis 2006) on the University of Oregon's super computer (http://aciss-computing.uoregon.edu). We determined topology frequencies and visualized phylogenies using DensiTree (Bouckaert 2010). We performed concordance analyses in *BUCKy* v 1.4.2 (Larget, et al. 2010) with α = 1 and default settings.

Genomic Structure analyses

To investigate relationships along the genome at a finer scale, we extracted genotypes from the same .vcf file used with the custom python script mentioned previously. We coded genotype patterns to splits using custom Unix and R scripts (available upon request). To avoid confounding factors of regions of exceptionally low diversity (Martin, et al. 2015), split frequencies were calculated for variable positions only. Due to the discontiguous nature of our exome data across the genome, we performed all analyses across chromosomes using windows of variable size relative to the genome, but standardized to include the same number of bases in our alignments. This approach, together with excluding invariant positions, ameliorates the effects of stochastic sampling that can cause false positives in certain types of genome scans (Martin, et al. 2015). Chromoplots and line plots were plotted in R using custom scripts (available upon request) based on the relative frequency of derived character states in our standardized genomic windows. D-statistics were calculated as previously described (Durand, et al. 2011). The D-statistic reported for *D. kyathit* and *D. nigrofasciatus* is the average of comparisons using *D. rerio* and *D. aesculapii* as the lineage not receiving gene flow. The probability of observing the patterns in the putative inversion on chromosome 20 was calculated from the joint probability of the observed patterns in each comparison corrected for the total number of genomic windows. Recombination rate data from (Anderson, et al. 2012), but were lifted over to GRCz10 from the previous zebrafish genome assembly (Zv9).

Bridge

Understanding species relationships is central to understanding the evolution of phenotypic characters (Mayden, et al. 2009). Because genes and populations can have different histories, the genetic changes underlying phenotypes may not mirror the overall history of species. Studies across taxa are demonstrating the importance of speciation with gene flow and introgression for understanding the evolution of phenotypes (Pollard, et al. 2006; Jones, et al. 2013; Martin, et al. 2013b; Supple, et al. 2013; Fontaine, et al. 2015; Soucy, et al. 2015; Pease, et al. 2016). Few studies of introgression, however, incorporate the effects of genome structure and recombination rate variation. Using a genome-scale approach in *Danio*, we demonstrate that genome structure played a large role in mediating introgression in the history of zebrafish, a major model species. Future studies will be able to apply these principles to understanding incongruence between demographic and genomic histories across taxa.

CHAPTER IV

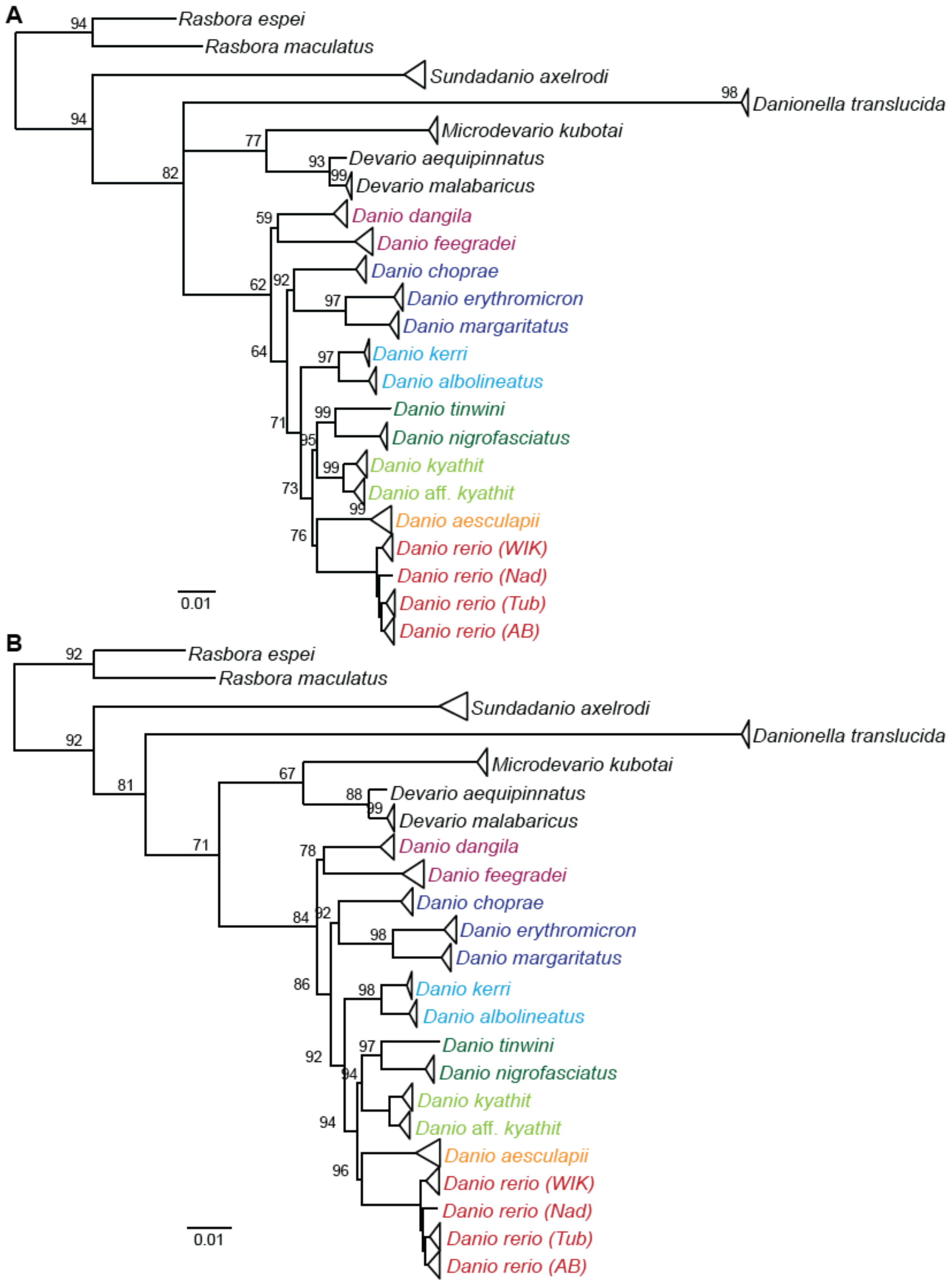
CONCLUSIONS

Mutations, the product of random imperfections in DNA replication, are the ultimate source of the genetic diversity that drives phenotypic evolution (Waddington 1962). To understand how phenotypic differences between species evolved, we must understand the histories of species and how mutations affect their phenotypes. We can begin to understand the histories of species by using mutations in DNA from different species to infer how they are related by shared ancestry. As we sequence more DNA from more species, we obtain a higher resolution view of the speciation and divergence processes that shaped species. To understand how these processes shaped the history of zebrafish, *Danio rerio* and its congeners, I performed two phylogenomic studies. These studies were necessary because several previous phylogenetic studies using a few loci arrived at different inferences as to the history of these species (Quigley, et al. 2004b; Mayden, et al. 2007; Fang, et al. 2009; Tang, et al. 2010). My first phylogenomic study (McCluskey and Postlethwait 2015) used Restriction-Site Associated DNA Sequencing (Baird, et al. 2008), a technique applicable to any group of species regardless of available genomic resources, to sample thousands of short sequences from across the genome. My analyses supported recent ancestry for two geographically overlapping species pairs: *D. rerio* and *D. aesculapii* in India and Bangladesh, and *D. kyathit* and *D. nigrofasciatus* in Myanmar. These analyses also revealed evidence for another relationship, which we interpreted as representing putative gene flow between *D. rerio* and *D. kyathit* despite their current geographic separation.

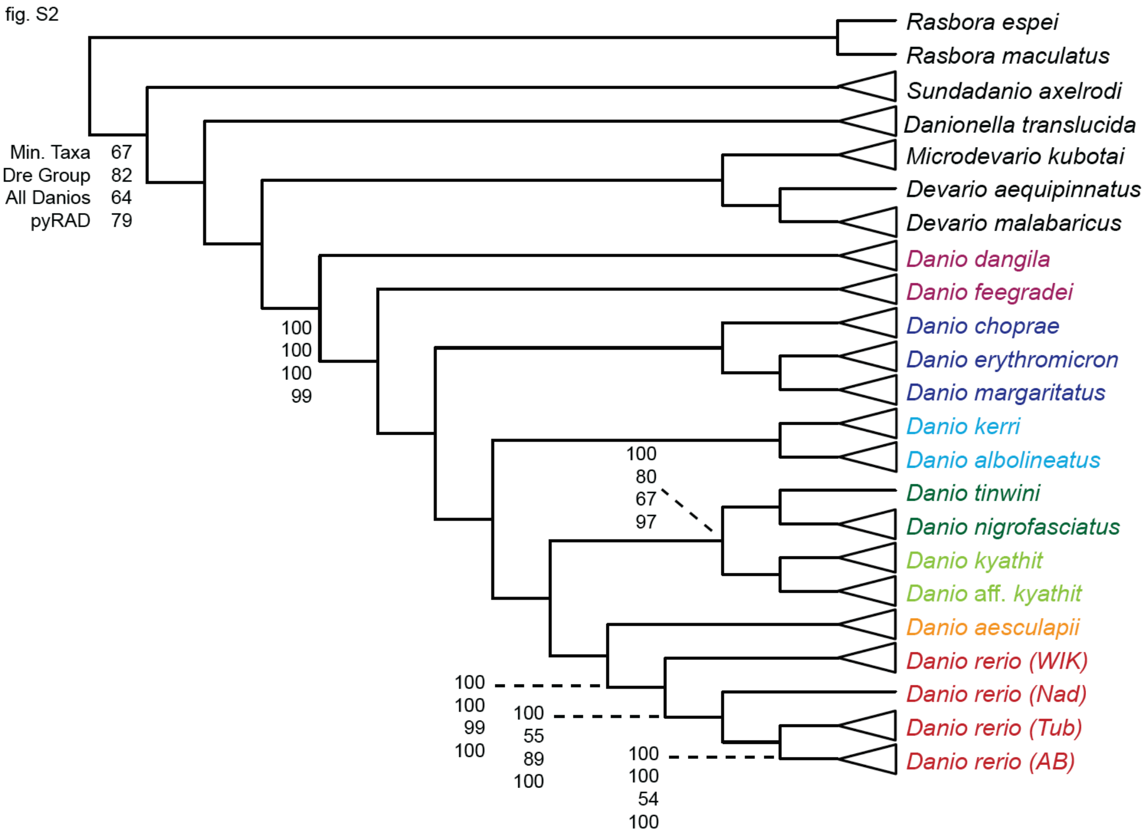
My second phylogenomic study used resources available for zebrafish to investigate the effects of chromosome structure on the relationships of these species across the genome. By coupling exome enrichment, genomic annotations, and knowledge of zebrafish genetics, we arrived at a complex demographic scenario that disagreed with the interpretations of the RAD-seq study (McCluskey and Postlethwait 2015) regarding the history of *D. rerio*, *D. kyathit, D. aesculapii,* and *D. nigrofasciatus*. This scenario supports three key demographic events shaping the recent history of zebrafish and its relatives at different parts of the genome, possibly due to linked selection in regions of low recombination. The centers of chromosomes have evidence for *D. rerio* and *D. kyathit* being closest relatives. We believe this relationship reflects the lineage splitting order of these four species. Support for *D. rerio* and *D. aesculapii* being closest relatives occurs primarily toward the ends of chromosomes, but not exclusively. We believe this relationship reflects gene flow between these lineages after the *D. rerio* lineage split with *D. aesculapii*. Finally, support for *D. kyathit* and *D. nigrofasciatus* being closest relatives occurs almost exclusively at the ends of chromosomes. We believe this relationship also represents gene flow, which occurred after *D. rerio* and *D. kyathit* shared a common ancestor. This detailed history helps us understand the history of these species and also the history of specific parts of their genomes. Distinguishing which genes have which history is essential for interpreting when mutations affecting those genes arose as the species diversified.

Ultimately, the “Tree of Life” in general, and individual phylogenies in particular, are simply models that help us comprehend and describe the complex, multifaceted process that is evolution. It is important to understand the limitations of these models, while also recognizing their utility for describing data, conveying concepts, and ultimately testing hypotheses regarding the biology of a system. Studies of genome evolution in the context of evolutionary relationships will provide vital insight into the evolutionary process, but will only tell part of the story. No matter how detailed our description of the history of species or the effects of that history on genome evolution, actually testing specific hypotheses in future studies is the only way to truly discern how random mutations ultimately give rise to phenotypic diversity.

APPENDIX



Supplementary Figure S.2.1. Phylogeny of the *Danio* genus based on RAD-tag sequences and Bayesian inference. (*A*) 50% majority-rule consensus tree based on the Dre Group dataset, with clade credibility values reported. Unlabeled nodes represent clade credibility values of 100. (*B*) 50% majority-rule consensus tree based on the All Danios dataset.



Supplementary Figure S.2.2. Cladogram of the *Danio* genus based on RAD-tag sequences and maximum parsimony. Unlabeled nodes have 100% bootstrap support across all datasets. Labeled nodes give the bootstrap support for the node in (from top to bottom) the Min. Taxa dataset, the Dre Group dataset, the All Danios dataset, and the pyRAD dataset.

Supplemental Table S.2.1. Reads and RAD-tags per Sample

Sample	Quality Filtered Illumina Reads	# ustacks	Unique Alignments (Loci)	With Repeats (Filtered out)	Multiple Alignments (Filtered out)	Unmappable (Filtered out)
Zv9	N/A	51,730	26,476	24,370	884	0
Drer_AB_f	1,247,229	52,717	24,549	25,607	1,114	1,447
Drer_AB_m	3,818,954	40,238	18,526	20,206	805	701
Drer_Nad_f	1,379,572	56,505	26,269	27,457	1,362	1,417
Drer_Tub_f	1,298,032	52,155	24,591	25,016	1,103	1,445
Drer_Tub_m	994,860	47,516	23,016	22,452	1,050	998
Drer_WIK_f	1,408,571	54,849	25,422	26,933	1,325	1,169
Drer_WIK_m	1,491,580	52,050	24,018	25,317	1,208	1,507
Daes_f	1,695,109	50,274	21,754	18,106	1,778	8,636
Daes_m	2,004,864	58,750	24,060	21,068	1,966	11,656
Dkya_f	1,530,556	51,438	14,065	21,948	581	7,617
Dkya_m	1,398,456	50,187	13,958	21,489	598	7,316
Dkya_aff_f	1,614,209	53,302	17,321	22,920	794	9,534
Dkya_aff_m	1,182,616	50,158	14,933	21,534	645	8,539
Dnig_f	1,275,029	47,582	16,987	20,615	1,170	10,691
Dnig_m	1,066,263	47,575	15,865	20,681	775	11,505
Dtin_q	3,426,584	50,293	17,819	28,138	933	13,124
Dalb_f	2,274,943	51,118	18,150	19,839	1,371	7,446
Dalb_m	3,540,105	53,334	18,045	20,715	1,336	7,513
Dker_f	1,414,838	40,669	19,049	15,616	1,699	11,871
Dker_m	1,358,909	33,632	16,058	14,433	725	8,380
Dcho_f	1,435,296	32,822	15,358	7,659	1,130	5,667
Dcho_m	1,505,205	36,393	21,322	8,744	1,788	7,272
Dery_f	1,537,193	30,056	20,248	5,939	1,659	6,791
Dery_m	1,189,215	27,520	16,253	5,257	1,256	7,544
Dmar_f	1,459,906	27,061	15,259	5,189	686	8,741
Dmar_m	1,511,014	30,421	16,548	5,735	1,058	9,476
Ddan_f	1,216,524	33,972	20,171	6,890	1,670	6,783
Ddan_m	1,242,728	36,295	20,679	7,447	1,744	7,067
Dfee_f	1,087,858	35,844	12,921	7,699	930	5,348
Dfee_m	1,242,573	41,598	18,584	9,722	1,713	10,982
DEVaeq_f	1,036,139	33,968	10,153	5,683	759	17,373
DEVpat_f	1,118,879	31,405	9,803	5,093	750	15,759
DEVpat_m	1,208,614	32,379	10,151	5,726	774	15,728
MICkub_f	1,171,149	27,356	8,344	3,092	467	15,453
MICkub_m	1,087,620	26,132	8,170	2,902	454	14,606
DLAtra_f	1,544,491	25,870	3,519	2,204	156	19,991
DLAtra_m	1,194,180	20,644	2,818	1,977	136	15,713
SUNaxe_f	1,366,967	30,987	5,195	4,202	226	21,364
SUNaxe_m	1,778,313	30,763	4,986	4,263	216	21,298
RASesp_q	1,350,104	31,795	7,173	5,136	403	19,083
RASmac_q	1,382,004	29,246	6,216	4,522	287	18,221

References

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. 2011. Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication. Genetics 188:799-U779.

Anderson JL, Mari AR, Braasch I, Amores A, Hohenlohe P, Batzel P, Postlethwait JH. 2012. Multiple Sex-Associated Regions and a Putative Sex Chromosome in Zebrafish Revealed by RAD Mapping and Population Genomics. Plos One 7.

Andrew RL, Kane NC, Baute GJ, Grassa CJ, Rieseberg LH. 2013. Recent nonhybrid origin of sunflower ecotypes in a novel habitat. Molecular Ecology 22:799-813.

Baack EJ, Rieseberg LH. 2007. A genomic view of introgression and hybrid speciation. Current Opinion in Genetics & Development 17:513-518.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. Plos One 3.

Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics 26:1372-1373.

Bradley KM, Breyer JP, Melville DB, Broman KW, Knapik EW, Smith JR. 2011. An SNP-Based Linkage Map for Zebrafish Reveals Sex Determination Loci. G3-Genes Genomes Genetics 1:3-9.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bures S, Garamszegi LZ, et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Research 25:1656-1665.

Bushnell B. BBMap.

Camp JG, Jazwa AL, Trent CM, Rawls JF. 2012. Intronic Cis-Regulatory Modules Mediate Tissue-Specific and Microbial Control of angptl4/fiaf Transcription. Plos Genetics 8.

Cariou M, Duret L, Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecology and Evolution 3:846-852.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3-Genes Genomes Genetics 1:171-182.

Chen YR, Cui GH, Shao JJ. 1988. Three Cyprinid Fishes new to Chinese Fauna. Zoological Research 9:439-440.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203-218.

Collins RA, Armstrong KF, Meier R, Yi YG, Brown SDJ, Cruickshank RH, Keeling S, Johnston C. 2012. Barcoding and Border Biosecurity: Identifying Cyprinid Fishes in the Aquarium Trade. Plos One 7.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. Genome Research 14:1188-1190.

Cruaud A, Gautier M, Galan M, Foucaud J, Saune L, Genson G, Dubois E, Nidelet S, Deuve T, Rasplus JY. 2014. Empirical Assessment of RAD Sequencing for Interspecific Phylogeny. Molecular Biology and Evolution 31:1272-1274.

Cui R, Schumer M, Kruesi K, Walter R, Andolfatto P, Rosenthal GG. 2013a. Phylogenomics reveals extensive reticulate evolution in Xiphophorus fishes. Evolution 67:2166-2179.

Cui RF, Schumer M, Kruesi K, Walter R, Andolfatto P, Rosenthal GG. 2013b. Phylogenomics Reveals Extensive Reticulate Evolution in Xiphophorus Fishes. Evolution 67:2166-2179.

Cutter AD. 2012. The polymorphic prelude to Bateson-Dobzhansky-Muller incompatibilities. Trends in Ecology & Evolution 27:209-218.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9:772-772.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. Molecular Biology and Evolution 28:2239-2252.

Eaton DAR, Ree RH. 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). Systematic Biology 62:689-706.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Eernisse DJ, Kluge AG. 1993. Taxonomic Congruence Versus Total Evidence, and Amniote Phylogeny Inferred from Fossils, Molecules, and Morphology. Molecular Biology and Evolution 10:1170-1195.

Ekker SC, Parichy DM, Cheng KC. 2008. Research implications of pigment biology in zebrafish. Zebrafish 5:233-235.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491:756-760.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proceedings of the National Academy of Sciences of the United States of America 107:16196-16200.

Engeszer RE, Patterson LB, Rao AA, Parichy DM. 2007. Zebrafish in the Wild: A Review of Natural History and New Notes from the Field. Zebrafish 4:21-U126.

Faircloth BC, Sorenson L, Santini F, Alfaro ME. 2013. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). Plos One 8.

Fang F. 1998. *Danio kyathit*, a new species of cyprinid fish from Myitkyina, northern Myanmar. Ichthyological Exploration of Freshwaters 8:273-280.

Fang F. 2003. Phylogenetic analysis of the Asian cyprinid genus Danio (Teleostei, Cyprinidae). Copeia:714-728.

Fang F. 2000. A Review of Chinese Danio Species (Teleostei:Cyprinidae). Acta Zootaxonomica Sinica 25:213-227.

Fang F, Noren M, Liao TY, Kallersjo M, Kullander SO. 2009. Molecular phylogenetic interrelationships of the south Asian cyprinid genera Danio, Devario and Microrasbora (Teleostei, Cyprinidae, Danioninae). Zoologica Scripta 38:237-256.

Felsenstein J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Systematic Zoology 27:401-410.

Felsenstein J. 1974. Evolutionary Advantage of Recombination. Genetics 78:737-756.

Felsenstein J. 1981. Evolutionary Trees from DNA-Sequences - a Maximum-Likelihood Approach. Journal of Molecular Evolution 17:368-376.

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, et al. 2015. Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347:1258524.

Froelich JM, Fowler ZG, Galt NJ, Smith Jr. DL, Biga PR. 2013. Sarcopenia and piscines: the case for indeterminate-growing fish as unique genetic model organisms in aging and longevity research. Frontiers in Genetics 4:159.

Froelich JM, Galt NJ, Charging MJ, Meyer BM, Biga PR. 2013. In vitro indeterminate teleost myogenesis appears to be dependent on Pax3. In Vitro Cellular & Developmental Biology - Animal 49:371-385.

Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC. 2012. Genome sequencing reveals complex speciation in the Drosophila simulans clade. Genome Research 22:1499-1511.

Gates MA, Kim L, Egan ES, Cardozo T, Sirotkin HI, Dougan ST, Lashkari D, Abagyan R, Schier AF, Talbot WS. 1999. A genetic linkage map for zebrafish: Comparative analysis and localization of genes and expressed sequences. Genome Research 9:334-347.

Geisel TS. 1960. One fish, two fish, red fish, blue fish. New York,: Beginner Books; distributed by Random House.

Geisler R, Rauch GJ, Baier H, van Bebber F, Bross L, Dekens MPS, Finger K, Fricke C, Gates MA, Geiger H, et al. 1999. A radiation hybrid map of the zebrafish genome. Nature Genetics 23:86-89.

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome. Science 352:52-+.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52:696-704.

Guryev V, Koudijs MJ, Berezikov E, Johnson SL, Plasterk RHA, van Eeden FJM, Cuppen E. 2006. Genetic variation in the zebrafish. Genome Research 16:491-497.

Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution 70:7-17.

Harrison RG, Larson EL. 2014. Hybridization, Introgression, and the Nature of Species Boundaries. Journal of Heredity 105:795-809.

Henkel CV, Dirks RP, Jansen HJ, Forlenza M, Wiegertjes GF, Howe K, van den Thillan GEEJM, Spaink HP. 2012. Comparison of the Exomes of Common Carp (Cyprinus carpio) and Zebrafish (Danio rerio). Zebrafish 9:59-67.

Hill WG, Robertson A. 2007. The effect of linkage on limits to artificial selection (Reprinted). Genetics Research 89:311-336.

Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, Manos PS. 2014. A Framework Phylogeny of the American Oak Clade Based on Sequenced RAD Data. Plos One 9:e93975.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Molecular Ecology Resources 11:117-122.

Hohenlohe PA, Bassham S, Currey M, Cresko WA. 2012. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. Philosophical Transactions of the Royal Society B-Biological Sciences 367:395-408.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. Plos Genetics 6.

Holder MT, Anderson JA, Holloway AK. 2001. Difficulties in detecting hybridization. Systematic Biology 50:978-982.

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.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Evolution - Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-2314.

Hukriede N, Fisher D, Epstein J, Joly L, Tellis P, Zhou Y, Barbazuk B, Cox K, Fenton-Noriega L, Hersey C, et al. 2001. The LN54 radiation hybrid map of zebrafish expressed sequences. Genome Research 11:2127-2132.

Innes WT. 1956. Exotic aquarium fishes; a work of general reference. In. Norristown etc. Pa.,: Aquarium Pub. Co. etc. p. v.

Johnson SL, Gates MA, Johnson M, Talbot WS, Horne S, Baik K, Rude S, Wong JR, Postlethwait JH. 1996. Centromere-linkage analysis and consolidation of the zebrafish genetic map. Genetics 142:1277-1288.

Jones JC, Fan SH, Franchini P, Schartl M, Meyer A. 2013. The evolutionary history of Xiphophorus fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. Molecular Ecology 22:2986-3001.

Kang JH, Schartl M, Walter RB, Meyer A. 2013. Comprehensive phylogenetic analysis of all species of swordtails and platies (Pisces: Genus Xiphophorus) uncovers a hybrid origin of a swordtail fish, Xiphophorus monticolus, and demonstrates that the sexually selected sword originated in the ancestral lineage of the genus, but was lost again secondarily. Bmc Evolutionary Biology 13.

Keller I, Wagner CE, Greuter L, Mwaiko S, Selz OM, Sivasundar A, Wittwer S, Seehausen O. 2013. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. Molecular Ecology 22:2848-2863.

Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C, et al. 2016. Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Research 44:D574-D580.

Kinkel MD, Prince VE. 2009. On the diabetic menu: Zebrafish as a model for pancreas development and function. Bioessays 31:139-152.

Knapik EW, Goodman A, Ekker M, Chevrette M, Delgado J, Neuhauss S, Shimoda N, Driever W, Fishman MC, Jacob HJ. 1998. A microsatellite genetic linkage map for zebrafish (Danio rerio). Nature Genetics 18:338-343.

Kullander SO. 2012. Description of Danio flagrans, and redescription of D. choprae, two closely related species from the Ayeyarwaddy River drainage in northern Myanmar (Teleostei: Cyprinidae). Ichthyological Exploration of Freshwaters 23:245-262.

Kullander SO, Fang F. 2009a. Danio aesculapii, a new species of danio from south-western Myanmar (Teleostei: Cyprinidae). Zootaxa:41-48.

Kullander SO, Fang F. 2009b. Danio tinwini, a new species of spotted danio from northern Myanmar (Teleostei: Cyprinidae). Ichthyological Exploration of Freshwaters 20:223-228.

Kullander SO, Liao TY, Fang F. 2009. Danio quagga, a new species of striped danio from western Myanmar (Teleostei: Cyprinidae). Ichthyological Exploration of Freshwaters 20:193-199.

Labate JA, Robertson LD, Strickler SR, Mueller LA. 2014. Genetic structure of the four wild tomato species in the Solanum peruvianum s.l. species complex. Genome 57:169-180.

Laing KJ, Purcell MK, Winton JR, Hansen JD. 2008. A genomic view of the NOD-like receptor family in teleost fish: identification of a novel NLR subfamily in zebrafish. Bmc Evolutionary Biology 8.

Larget BR, Kotha SK, Dewey CN, Ane C. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26:2910-2911.

Li G, Davis BW, Eizirik E, Murphy WJ. 2016. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). Genome Research 26:1-11.

Lieschke GJ, Currie PD. 2007. Animal models of human disease: zebrafish swim into view. Nature Reviews Genetics 8:353-367.

Linder CR, Rieseberg LH. 2004. Reconstructing patterns of reticulate evolution UN plants. American Journal of Botany 91:1700-1708.

Lohr H, Hammerschmidt M. 2011. Zebrafish in Endocrine Systems: Recent Advances and Implications for Human Disease. Annual Review of Physiology, Vol 73 73:183-211.

Mahalwar P, Walderich B, Singh AP, Nusslein-Volhard C. 2014. Local reorganization of xanthophores fine-tunes and colors the striped pattern of zebrafish. Science 345:1362-1364.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends in Ecology & Evolution 20:229-237.

Martin CH, Feinstein LC. 2014. Novel trophic niches drive variable progress towards ecological speciation within an adaptive radiation of pupfishes. Molecular Ecology 23:1846-1862.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013a. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Research 23:1817-1828.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013b. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Research 23:1817-1828.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the Use of ABBA-BABA Statistics to Locate Introgressed Loci. Molecular Biology and Evolution 32:244-257.

Mayden RL, Chen WJ, Bart HL, Doosey MH, Simons AM, Tang KL, Wood RM, Agnew MK, Yang L, Hirt MV, et al. 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. Molecular Phylogenetics and Evolution 51:500-514.

Mayden RL, Tang KL, Conway KW, Freyhof J, Chamberlain S, Haskins M, Schneider L, Sudkamp M, Wood RM, Agnew M, et al. 2007. Phylogenetic relationships of Danio within the order cypriniformes: A framework for comparative and evolutionary studies of a model species. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 308B:642-654.

Mayden RL, Tang KL, Wood RM, Chen WJ, Agnew MK, Conway KW, Yang L, Simons AM, Bart HL, Harris PM, et al. 2008. Inferring the Tree of Life of the order Cypriniformes, the earth's most diverse clade of freshwater fishes: Implications of varied taxon and character sampling. Journal of Systematics and Evolution 46:424-438.

McCluskey BM, Postlethwait JH. 2015. Phylogeny of Zebrafish, a "Model Species," within Danio, a "Model Genus". Molecular Biology and Evolution 32:635-652.

McMenamin SK, Bain EJ, McCann AE, Patterson LB, Eom DS, Waller ZP, Hamill JC, Kuhlman JA, Eisen JS, Parichy DM. 2014a. Thyroid hormone-dependent adult pigment cell lineage and pattern in zebrafish. Science 345:1358-1361.

McMenamin SK, Bain EJ, McCann AE, Patterson LB, Eom DS, Waller ZP, Hamill JC, Kuhlman JA, Eisen JS, Parichy DM. 2014b. Thyroid hormone-dependent adult pigment cell lineage and pattern in zebrafish. Science 345:1358-1361.

Meyer A, Biermann CH, Orti G. 1993. The phylogenetic position of the zebrafish (Danio rerio), a model system in developmental biology: an invitation to the comparative method. Proc Biol Sci 252:231-236.

Meyer BS, Matschiner M, Salzburger W. 2015. A tribal level phylogeny of Lake Tanganyika cichlid fishes based on a genomic multi-marker approach. Molecular Phylogenetics and Evolution 83:56-71.

Mills MG, Nuckels RJ, Parichy DM. 2007. Deconstructing evolution of adult phenotypes: genetic analyses of kit reveal homology and evolutionary novelty during adult pigment pattern development of Danio fishes. Development 134:1081-1090.

Noor MAF, Bennett SM. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity 103:439-444.

Norton W, Bally-Cuif L. 2010. Adult zebrafish as a model organism for behavioural genetics. Bmc Neuroscience 11.

Orr HA. 1995. The Population-Genetics of Speciation - the Evolution of Hybrid Incompatibilities. Genetics 139:1805-1813.

Parichy DM. 2007. Homology and the evolution of novelty during Danio adult pigment pattern development. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 308B:578-590.

Parichy DM, Johnson SL. 2001. Zebrafish hybrids suggest genetic mechanisms for pigment pattern diversification in Danio. Development Genes and Evolution 211:319-328.

Patowary A, Purkanti R, Singh M, Chauhan R, Singh AR, Swarnkar M, Singh N, Pandey V, Torroj C, Clark MD, et al. 2013. A Sequence-Based Variation Map of Zebrafish. Zebrafish 10:15-20.

Patterson LB, Bain EJ, Parichy DM. 2014. Pigment cell interactions and differential xanthophore recruitment underlying zebrafish stripe reiteration and Danio pattern evolution. Nat Commun 5:5299.

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. Plos Biology 14:e1002379.

Peterson RT, MacRae CA. 2012. Systematic Approaches to Toxicology in the Zebrafish. Annual Review of Pharmacology and Toxicology, Vol 52 52:433-453.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Worheide G, Baurain D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. Plos Biology 9.

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. Plos Genetics 8.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in Drosophila: Evidence for incomplete lineage sorting. Plos Genetics 2:1634-1647.

Postlethwait J, Amores A, Cresko W, Singer A, Yan YL. 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends in Genetics 20:481-490.

Postlethwait JH, Johnson SL, Midson CN, Talbot WS, Gates M, Ballinger EW, Africa D, Andrews R, Carl T, Eisen JS, et al. 1994. A Genetic-Linkage Map for the Zebrafish. Science 264:699-703.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. Nature Reviews Genetics 11:175-180.

Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43-+.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569-U247.

Quigley AK, Turner JM, Nuckels RJ, Manuel JL, Budi EH, MacDonald EL, Parichy DM. 2004a. Pigment pattern evolution by differential deployment of neural crest and post-embryonic melanophore lineages in Danio fishes. Development 131:6053-6069.

Quigley IK, Manuel JL, Roberts RA, Nuckels RJ, Herrington ER, MacDonald EL, Parichy DM. 2005. Evolutionary diversification of pigment pattern in Danio fishes: differential fms dependence and stripe loss in D. albolineatus. Development 132:89-104.

Quigley IK, Parichy DM. 2002. Pigment pattern formation in zebrafish: a model for developmental genetics and the evolution of form. Microsc Res Tech 58:442-455.

Quigley IK, Turner JM, Nuckels RJ, Manuel JL, Budi EH, MacDonald EL, Parichy DM. 2004b. Pigment pattern evolution by differential deployment of neural crest and post-embryonic melanophore lineages in Danio fishes. Development 131:6053-6069.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Rodriguez A. 2013. The Zebrafish as a Model for the Evolution and Development of Breding Tubercles in Fishes. [University of Colorado - Boulder.

Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, Valencia A, Tress ML. 2013. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Research 41:D110-D117.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Systematic Biology 61:539-542.

Rosenberg NA. 2013. Discordance of Species Trees with Their Most Likely Gene Trees: A Unifying Principle. Molecular Biology and Evolution 30:2709-2713.

Rosenberg NA, Tao R. 2008. Discordance of species trees with their most likely gene trees: The case of five taxa. Systematic Biology 57:131-140.

Rosenthal GG, Ryan MJ. 2005. Assortative preferences for stripes in danios. Animal Behaviour 70:1063-1066.

Roure B, Baurain D, Philippe H. 2013. Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets. Molecular Biology and Evolution 30:197-214.

Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. Bmc Evolutionary Biology 7.

Ruber L, Kottelat M, Tan HH, Ng PKL, Britz R. 2007. Evolution of miniaturization and the phylogenetic position of Paedocypris, comprising the world's smallest vertebrate. Bmc Evolutionary Biology 7.

Rubin BER, Ree RH, Moreau CS. 2012. Inferring Phylogenies from RAD Sequence Data. Plos One 7.

Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507:354-+.

Santoriello C, Zon LI. 2012. Hooked! Modeling human disease in zebrafish. Journal of Clinical Investigation 122:2337-2343.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483:169-175.

Schumer M, Cui RF, Boussau B, Walter R, Rosenthal G, Andolfatto P. 2013. An Evaluation of the Hybrid Speciation Hypothesis for Xiphophorus Clemenciae Based on Whole Genome Sequences. Evolution 67:1155-1168.

Sen N. 2007. Description of a new species of *Brachydanio* Weber and de Beaufort, 1916 (Pisces : Cypriniformes : Cyprinidae) from Meghala, North East India with a note on comparative studies of other known species. Rec. zool. Surv. India 107:27-31.

Shimoda N, Knapik EW, Ziniti J, Sim C, Yamada E, Kaplan S, Jackson D, de Sauvage F, Jacob H, Fishman MC. 1999. Zebrafish genetic map with 2000 microsatellite markers. Genomics 58:219-232.

Smeds L, Kunstner A. 2011. CONDETRI - A Content Dependent Read Trimmer for Illumina Data. Plos One 6.

RepeatMasker Open-3.0 [Internet]. 1996-2010 [cited 2013. Available from: http://www.repeatmasker.org

Smith JM, Haigh J. 2007. The hitch-hiking effect of a favourable gene. Genetics Research 89:391-403.

Soucy SM, Huang JL, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nature Reviews Genetics 16:472-482.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Stankowski S, Streisfeld MA. 2015. Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. Proc Biol Sci 282.

Supple MA, Hines HM, Dasmahapatra KK, Lewis JJ, Nielsen DM, Lavoie C, Ray DA, Salazar C, McMillan WO, Counterman BA. 2013. Genomic architecture of adaptive color pattern divergence and convergence in Heliconius butterflies. Genome Research 23:1248-1257.

Tang KL, Agnew MK, Hirt MV, Sado T, Schneider LM, Freyhof J, Sulaiman Z, Swartz E, Vidthayanon C, Miya M, et al. 2010. Systematics of the subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). Molecular Phylogenetics and Evolution 57:189-214.

Thomas GWC, Hahn MW. 2015. Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals. Molecular Biology and Evolution 32:1232-1236.

Ting N, Sterner KN. 2013. Primate molecular phylogenetics in a genomic era. Molecular Phylogenetics and Evolution 66:565-568.

Turner TL, Hahn MW. 2010. Genomic islands of speciation or genomic islands and speciation? Molecular Ecology 19:848-850.

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in Anopheles gambiae. Plos Biology 3:1572-1578.

Waddington CH. 1962. The nature of life. New York,: Atheneum.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Molecular Ecology 22:787-798.

Wang XQ, Zhao L, Eaton DAR, Li DZ, Guo ZH. 2013. Identification of SNP markers for inferring phylogeny in temperate bamboos (Poaceae: Bambusoideae) using RAD sequencing. Molecular Ecology Resources 13:938-945.

Whiteley AR, Bhat A, Martins EP, Mayden RL, Arunachalam M, Uusi-Heikkila S, Ahmed ATA, Shrestha J, Clark M, Stemple D, et al. 2011. Population genomics of wild and laboratory zebrafish (Danio rerio). Molecular Ecology 20:4259-4276.

Wilson CA, High SK, McCluskey BM, Amores A, Yan YL, Titus TA, Anderson JL, Batzel P, Carvan MJ, Schartl M, et al. 2014. Wild Sex in Zebrafish: Loss of the Natural Sex Determinant in Domesticated Strains. Genetics 198:1291-+.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nature Reviews Genetics 2:333-341.

Wong TT, Saito T, Crodian J, Collodi P. 2011. Zebrafish Germline Chimeras Produced by Transplantation of Ovarian Germ Cells into Sterile Host Larvae. Biology of Reproduction 84:1190-1197.

Woods IG, Kelly PD, Chu F, Ngo-Hazelett P, Yan YL, Huang H, Postlethwait JH, Talbot WS. 2000. A comparative map of the zebrafish genome. Genome Research 10:1903-1914.

Woods IG, Wilson C, Friedlander B, Chang P, Reyes DK, Nix R, Kelly PD, Chu F, Postlethwait JH, Talbot WS. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Research 15:1307-1314.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859-1875.