CASTING A WIDER NET: USING RAY-FINNED FISH GENOMES TO GAIN NOVEL INSIGHTS INTO VERTEBRATE MOLECULAR EVOLUTION

by

Rittika Mallik

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2024

Approved by:

_____
Dr. Alex Dornburg

_____
Dr. Cynthia Gibas

_____
Dr. Anthony Fodor

_____
Dr. Jun Tao Guo

_____
Dr. Adam Reitzel

ABSTRACT

RITTIKA MALLIK. Casting A Wider Net: Using Ray-Finned Fish Genomes To Gain Novel Insights Into Vertebrate Molecular Evolution.
(Under the direction of DR. ALEX DORNBURG)

The past decade has provided unprecedented insights into the molecular evolutionary pathways that have given rise to the present day diversity of vertebrates. Comparative genomic studies have repeatedly revealed that many key ecological traits, novel functional phenotypes, and even disease states are governed by genomic regions characterized by frequent mutations, duplications, or deletion events. However, the evolutionary origins and early diversification history of many of these regions remain poorly understood. My work focuses on providing a resolution to this history, focusing on the evolution of the vertebrate mobilome and a clustered gene family of innate immune receptors with putative links to the origin of the adaptive immune response. To accomplish this, I sequenced the genomes of *Polypterus bichir* and *Lepisosteus osseus*, two taxa that fill critical genomic sampling gaps for early diverging vertebrate lineages. Integrating these genomes into a comparative dataset of over 100 genomes that span all major ray-finned fish lineages, I investigated the effect of teleost genome duplication (TGD) on the diversification of the ray-finned fish mobilome. My findings reveal no substantial shift in mobilome composition following the TGD event, in line with a growing body of evidence that this historical ploidy event has not left a signature of a burst of molecular diversification and innovation across half of living vertebrates. I next expanded my taxonomic coverage to include all major vertebrate lineages to investigate the evolutionary origin of signal regulatory proteins (SIRPs) and their ligand CD47. In mammals, SIRPs are essential for regulating macrophage function and have become important targets for cancer therapy. These receptors also contain variable and joining exons and are hypothesized to have arisen in tetrapods out of a complex of innate immune receptor gene families that also gave rise to recom-

bining T-cell receptors and antibody encoding Immunoglobulin domains. My work demonstrates this is not the case. Instead, SIRPs have evolutionary origins coincident with the origin of the adaptive immune response. In contrast, we find no evidence for an ancient origin of the CD47 ligand, which interacts with SIRPs. Instead, CD47 appears to have arisen at the beginning of amniote evolution, suggesting a decoupling of the evolutionary origins of this ligand and receptor pair. These findings provide a new perspective on the origins and diversification of innate immune receptor gene families and their relationship to the emergence of the adaptive immune system.

ACKNOWLEDGEMENTS

I extend my sincere gratitude to my advisor, Dr. Alex Dornburg, for his invaluable guidance and support throughout my doctoral journey. His mentorship has been a source of inspiration, fostering intellectual growth and creativity. I enjoyed discussions about research, anime, and languages and I have learned a lot from his expertise, insights, and vision. Dr. Dornburg's willingness to grant me the independence and creativity to explore and develop novel ideas has been instrumental in my journey. I am deeply grateful for his open-mindedness in embracing my opinions and ideas, even when they diverged from his own, demonstrating the true essence of mentorship.

I am grateful to my committee members, Dr. Cynthia Gibas, Dr. Anthony Fodor, Dr. Jun Tao Guo, and Dr. Adam Reitzel, for their insightful comments and suggestions on my dissertation. Their mentorship and availability for discussions have been invaluable in shaping my research path. I also extend my heartfelt thanks to Dr. Jeffrey A. Yoder for his significant contributions to my research and assistance in writing. A huge thanks especially to Dr. Gibas for believing in me and supporting me in my lab transition.

I also extend my gratitude to my colleagues Katerina Zapfe, Gerard Nasser, Arielle Zapata and Juan Bolanos for their friendship, collaboration, and support in my research. Their insights and assistance have been invaluable. I am equally thankful to my peers Brandon Turner, Josh Sikder, Daisy Fry Brumit, and Gabe O'Reilly for our wide-ranging conversations and the camaraderie we've shared. Working with you all and forming lasting friendships has been the best part of my past five years. Spending time with you all and enjoying fun outings was the highlight of my week.

To all my friends, Madhumita Paul, Antardipan Pal, and Dev Mashruwala thank you for keeping me sane during this rollercoaster ride. You were available when I needed you the most, and I am so happy to have friends like you. I would also like to thank my childhood friends, Aishwarya Kumar and Upasona Paul, for always being

there for me, no matter how far we are. Last but not least, I would like to dedicate this dissertation to my boyfriend, Sayantan Datta, for encouraging and supporting my endeavors over the years. He has been my pillar of strength and support throughout all the highs and lows over the years. As I look forward to the next chapter of my life, I am filled with gratitude for the incredible people who have contributed to my growth and shaped me into who I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Genome stability

The duplication of genes and genetic elements is a hallmark of genome evolution. However, this hallmark also raises the question of what general rules underlie the fate of such duplications. Answering this question is of particular importance as genome stability is a critical aspect of cellular health and survival. It encompasses various genetic alterations, ranging from point mutations to chromosome rearrangements. The breakdown of mechanisms that protect the genome and the ensuing instability are key factors in the aging process and are associated with diseases such as cancer [3]. Consequently, a comprehensive network of interconnected pathways is dedicated to preserving genome integrity in response to the continuous challenges that induce DNA damage, including epigenetic mechanisms [73], DNA modifications [204], histone variants and modifications [38], chromatin structure [81], and non-coding RNAs [38]. all perform various functions to ensure the maintenance of genome stability, protect the genome from invasion by transposable elements and contribute to various DNA repair pathways.

The etiology of genome instability is multifaceted, encompassing both trans-acting and cis-acting elements, each playing distinct yet interrelated roles in the preservation of genomic integrity [169].; and S-phase checkpoint factors, which are essential for the orchestration of replication, repair processes, chromosome segregation, and cell-cycle progression. On the other hand, cis-acting elements represent chromosomal regions particularly prone to instability. These hotspots include fragile sites-specific DNA sequences predisposed to gaps, constrictions, and breaks due to compromised

replication progression - and highly transcribed DNA sequences, which are linked to transcription-dependent recombination and rearrangements, further underscoring the complexity of maintaining genome stability [3]. These mechanisms are highly important for maintaining organismal health, as illustrated by their role in human cancer.

1.2 Genome instability and cancer

Genomic instability is a characteristic of most cancer cells and is associated with an increased tendency of genome alteration during cell division [227]. This instability is a hallmark of various pathological disorders, including cancer and conditions related to premature aging, emphasizing the importance of genomic integrity for human health. Genomic instability often results from damage to genes controlling cell division and tumor suppression and is closely monitored by surveillance mechanisms such as the DNA damage checkpoint, DNA repair machinery, and mitotic checkpoint [3]. Defects in these regulatory mechanisms can lead to genomic instability, predisposing cells to malignant transformation.

1.2.1 Genome stability and relationship to clustered gene families

It is clear that genome instability can lead to harmful effects at the organismal level. However, it is now clear that genomic instability is a double-edged sword, and also plays a crucial role in generating genetic diversity and evolution [146]. It provides the variation necessary for natural selection, facilitating adaptive processes that shape species over time. This complex interplay between the detrimental and beneficial outcomes of genomic instability underscores its significant role in both health and evolutionary biology. In particular, the breakdown of genome stability has given rise to complex suites of gene families that shape many of the very features we require for survival. For example, light sensing is a fundamental trait [228], with proteins

involved in human vision being studied for over a century using various model organisms. Recent large comparative studies have placed gene families in the human genome within an evolutionary context, revealing associations between gene family evolution and lineage ecology, as well as a remarkable diversity of genotypes in genes such as genes like crystallins and opsins related to light sensing phenotypes across the Tree of Life.

Opsins are a prime example of gene diversification, with over 1000 opsins identified across various species, including humans, flies, mice, and zebrafish, since the sequencing of the first opsin gene nearly 40 years ago in bovids [155, 140]. This diversity has shed light on color vision across the Tree of Life, demonstrating functional convergences, correlations between opsin repertoire and ecological factors, gene losses or loss of function associated with sensory trade-offs, amino acid convergence among distantly related taxa, and convergences in the genetic basis of human retinal diseases [17, 46] Surprisingly, opsin genes are expressed not only in the eye but also in the skin and peripheral tissues [195], carrying out functions beyond light reception. Comparative genomics of opsin gene families has implications for wound healing [46], hair growth, optogenetics, and metabolic physiology, with future investigations poised to reveal the full diversity of these gene families. The example from opsins underscores how understanding the functional relationships within gene clusters can provide insights with high translational relevance. For my dissertation, I am particularly interested in understanding the evolution of clustered gene families and how these contribute to the diversification of novel functional forms.

1.3 Clustered gene families

Clustered gene families represent groups of genes closely arranged within specific genomic regions, often residing on the same chromosome. These gene clusters exhibit

shared nucleotide sequences or functions and are subject to collective regulation [167]. Clustered gene families are important for understanding the function and evolution of genomes. They provide evidence of candidate homologous regions and are thought to be valuable for characterizing the general structure of genetic variation among human populations [167]. Relative to surrounding regions, clustered gene families often display disproportionately high levels of sequence and gene copy diversity. As such, clustered gene families may be hotspots for additional gene birth events, higher rates of gene death (loss), nucleotide substitution, sharing of regulatory sequences among multiple genes in tandem [63]. However, it has been difficult to characterize the functional diversity and sequence of gene families across species. Both variations in the number of genes within each gene family and sequence divergence among duplicate genes can be seen in organisms; these variations can even be seen at the population level (i.e., copy number variation) or among closely related species. The complex evolutionary history of gene duplication over time that has produced two or more paralogs on the same chromosome (i.e., cis) that are either grouped or at different chromosomal positions can be seen in this molecular diversity.

The physical proximity of these genes implies a common evolutionary history and shared regulatory elements, influencing their coordinated expression. These gene families encompass diverse members, including those encoding proteins with related functions or participating in shared biological pathways [224]. This non-random organization is thought to enhance gene regulation efficiency and streamline cellular processes. A comprehensive understanding of the composition and arrangement of clustered gene families is pivotal for unraveling the intricacies of gene regulation, functional genomics, and evolutionary biology. The study of these clusters yields valuable insights into the molecular mechanisms underpinning cellular processes, contributing significantly to deciphering the complexity of genetic interactions within the genome.

Much of our understanding of clustered gene families involved in the immune response comes from the human genome. The currently assembly GRCh38 is comprised of approximately 3,100 megabases, encoding up to 20,000 protein-coding genes, over 25,000 non-protein-coding genes, and around 15,000 pseudogenes [63]. A significant portion of the protein-coding genes have paralogs, indicating a high frequency of gene duplication. It is now clear that the study of human gene families benefits from comparisons with a diverse range of organisms across the Tree of Life due to the deep evolutionary history of gene duplication and divergence. Phylogenetic approaches are crucial in understanding the evolutionary processes that shape the origins and functions of human gene families. Gene family evolution is generally expected to mirror the evolutionary relationships among species, with divergences occurring due to selective pressures, mutations, and genetic drift over millions of years. However, gene families vary in their evolutionary dynamics: some are highly labile, generating new sequences and functional diversity, while others are constrained, preserving essential genetic information across the Tree of Life. Although these dynamics-lability and constraint-appear as opposites, they can coexist within gene families. Some highly conserved gene families may still contribute to sequence diversity and innovation, underscoring the complex interplay of evolutionary forces in shaping gene family evolution.

Gene families that are clustered could be hotspots for more gene birth events, increased rates of gene death (loss), nucleotide substitution, exon swapping, interlocus gene conversion, and the sharing of regulatory sequences among several genes simultaneously. Examples of clustered gene families are very common in immunogenetics and include the Signal regulatory proteins (SIRPs), the nucleotide oligomerization domain (NOD) and the NACHT leucine-rich repeat and PYD containing (NALP) gene

families, the cysteine rich scavenger receptor gene family, the IL-17 cytokine genes , and the SpTransformer (SpTrf) genes [14]. A key aspect of my dissertation and research interest involves understanding the evolutionary history of clustered gene families such as these to provide the comparative framework for future functional studies, with an emphasis on understudied genes involved in the innate immune response. For my current work, I focus on one such family: SIRPs.

1.3.1 Signal Regulatory Proteins

The Signal Regulatory Protein (SIRP) family encompasses a group of transmembrane glycoproteins predominantly expressed on immune cells and neurons [12]. The principal members of this family include SIRPα (CD172a), SIRPβ, and SIRPγ. SIRPα is primarily found on hematopoietic progenitors, myeloid cells, dendritic cells, natural killer (NK) cells, and neurons [138, 208]. In the context of this discussion, the focus is on SIRPγ, which is a T-cell-restricted surface molecule. SIRPγ, along with SIRPα, interacts with CD47, a transmembrane protein that is broadly expressed across various cell types [138].

The interaction between CD47 on red blood cells and SIRPα on macrophages is particularly noteworthy, as it prevents the phagocytosis of red blood cells by macrophages, thereby regulating their lifespan. This interaction is crucial for self-recognition and the prevention of autoimmune responses. However, cancer cells can exploit this mechanism by overexpressing CD47 to evade immune surveillance and phagocytosis. The disruption of the SIRPα-CD47 interaction has become a target for cancer immunotherapy, aiming to enhance the clearance of tumor cells by the immune system [47].

In Chapter 3 of my dissertation, I am undertaking a comprehensive analysis to elucidate the evolutionary origins of Signal Regulatory Proteins (SIRPs) across ver-

tebrates. This investigation aims to determine the temporal depth of SIRP emergence in the evolutionary timeline. Concurrently, I am tracing the evolutionary history of CD47, a crucial ligand for SIRPs, to assess the degree of evolutionary concordance between these two critical components of the immune system. This comparative analysis will provide insights into the co-evolutionary dynamics and the functional interdependence of SIRPs and CD47 throughout vertebrate evolution. This work was made possible by sequencing additional genomes for major early diverging vertebrates that are currently lacking genomic resources in my first and second chapters. In addition, this work empowered me to also assess patterns of rapid evolution over deep time scales in other class of sequences prone to rapid birth and death events: transposable elements.

1.4 Transposable elements

Transposable elements (TEs) are DNA sequences that can change their position within a genome, sometimes creating or reversing mutations and altering the cell's genomic size. They are also known as "jumping genes" due to their ability to move from one location to another within the genome [91]. TEs are found in virtually all organisms, from bacteria to humans, and constitute a significant portion of the genome in many species. In humans, transposable elements make up nearly half of the genome.The activity of transposable elements can have various impacts on the genome and the organism. On one hand, they can be sources of genetic diversity, as their movement can lead to the creation of new genes or regulatory elements, potentially providing evolutionary advantages [95].

On the other hand, their insertion into functional genes or regulatory regions can disrupt gene function, leading to mutations that may be harmful or even lethal. TEs can also influence genome structure by promoting recombination, leading to deletions, duplications, and chromosomal rearrangements. Transposable elements are

classified into two main types based on their mechanism of transposition: Class I elements, or retrotransposons, which move via an RNA intermediate and employ a "copy-and-paste" mechanism; and Class II elements, or DNA transposons, which move directly as DNA using a "cut-and-paste" mechanism [75]. The human genome contains a significant proportion of repetitive sequences derived from transposable elements, including retrotransposons that replicate through a copy-and-paste mechanism and DNA transposons that use a cut-and-paste method. However, in humans, only retrotransposons, specifically long interspersed element-1 (LINE-1 or L1), remain active as mobile DNAs. LINE-1 retrotransposition is ongoing in humans, contributing to genomic structural variations in populations and alterations in cancer genomes. Endogenous retroviruses are also present in the genome, serving as promoter and protein-coding sequences [39]. The activity of TEs is usually tightly regulated in organisms, as uncontrolled transposition can lead to genomic instability. However, their ability to mobilize and mutate has made them valuable tools in genetic research and biotechnology for gene tagging, mutagenesis, and gene therapy.

Given the importance of TEs for empirical research, there has been a call for investigations of TE diversity over deeper evolutionary timescales to provide a comparative basis for such studies. However, for ray-finned fishes, which collectively comprise over half of all living vertebrates, such studies have not been attempted. Using results from my first chapter, my second chapter provides the first comprehensive perspective on the evolution of the ray-finned fish mobilome. This work explored the extent to which the teleost genome duplication (TGD) influenced the diversification trajectory of the ray-finned fish mobilome. We combined a new high-coverage genome of Polypterus bichir with data from over 100 publicly available actinopterygian genomes to evaluate the macroevolutionary consequences of genome duplication events on the evolution of transposable elements (TEs). Contrary to expectations, our findings do

not support a significant shift in mobilome composition following the TGD event. Instead, the diversity of the actinopterygian mobilome seems to have been shaped by a history of lineage-specific changes in composition that are not associated with commonly cited drivers of diversification, such as body size, water column usage, or latitude. These results offer a fresh perspective on the early diversification of the actinopterygian mobilome and suggest that historical ploidy events may not necessarily lead to bursts of TE diversification and innovation. Overall, these findings underscore the clade-specific heterogeneity within ray-finned fishes and suggest that factors other than extinction events may have shaped mobilome divergences between ray-finned fish lineages and other vertebrates.

CHAPTER 2: A CHROMOSOME-LEVEL GENOME ASSEMBLY OF LONGNOSE GAR, *LEPISOSTEUS OSSEUS*

Published in G3: Genes | Genomes | Genetics 2023

2.1 Abstract

Holosteans (gars and bowfins) represent the sister lineage to teleost fishes, the latter being a clade that comprises over half of all living vertebrates and includes important models for comparative genomics and human health. A major distinction between the evolutionary history of teleosts and holosteans is that all teleosts experienced a genome duplication event in their early evolutionary history. As the teleost genome duplication occurred after teleosts diverged from holosteans, holosteans have been heralded as a means to bridge teleost models to other vertebrate genomes. However, only three species of holosteans have been genome sequenced to date and sequencing of more species is needed to fill sequence sampling gaps and provide a broader comparative basis for understanding holostean genome evolution. Here we report the first high quality reference genome assembly and annotation of the longnose gar (*Lepisosteus osseus*). Our final assembly consists of 22,709 scaffolds with a total length of 945 bp with contig N_{50} of 116.61 kb. Using BRAKER2, we annotated a total of 30,068 genes. Analysis of the repetitive regions of the genome reveals the genome to contain 29.12% transposable elements, and the longnose gar to be the only other known vertebrate outside of the spotted gar and bowfin to contain CR1, L2, Rex1, and Babar. These results highlight the potential utility of holostean genomes for understanding the evolution of vertebrate repetitive elements, and provide a critical reference for comparative genomic studies utilizing ray-finned fish models.

2.2 Significance

Over half of all living vertebrates are teleost fishes, including numerous experimental models such as zebrafish (*Danio rerio*) and medaka (*Oryzias latipes*). However, translating research in teleost models to other organisms such as humans is often challenged by the fact that teleosts experienced a genome duplication event in their early evolutionary history. Recent genome sequencing of three holosteans, the sister

lineage to teleosts, has revealed these taxa to be critical for linking homologs between teleosts and other vertebrates. Sequencing of holostean genomes remains limited, thereby impeding further comparative genomic studies. Here we fill this sampling gap through the genomic sequencing of the longnose gar (*Lepisosteus osseus*). This annotated reference genome will provide a useful resource for a range of comparative genomic applications that span fields as diverse as immunogenetics, developmental biology, and the understanding of regulatory sequence evolution.

2.3 Introduction

Teleost fishes represent over half of all living vertebrates and have successfully radiated in nearly all of the planet's aquatic habitats [142, 82]. Teleosts are of vital ecological importance, form the basis of several multi-billion dollar industries [113, 196, 82], and act as important model species (e.g. zebrafish and medaka) that are of high utility for human health research [71, 163]. The rapid accumulation of hundreds of genome sequences spanning the teleost Tree of Life has empowered unprecedented insights into the genomic basis for their evolutionary success [53], and provided key insights into teleost molecular biology with translational relevance to human health![97]. However, the development of both a deeper understanding of teleost genome evolution and the connection between teleost and human genomes has been challenged by the teleost-specific genome duplication (TGD) event that occurred during the early evolution of teleosts. This duplication event has complicated investigations of genomic novelty, homology, and synteny [93, 28] In contrast, the few living species of holosteans, non-teleost fishes (bowfin and gar) dubbed "living fossils" by Darwin [54] diverged from teleosts prior to the TGD [202, 64]. Holostean genomes have been demonstrated to be critical for understanding gene synteny and homology of complex genomic regions between teleosts and other vertebrates [18, 28, 202]. Being the closest living relatives of teleosts, holosteans provide particularly informative context for understanding whether genomic novelties identified in teleosts are in fact unique to teleosts

and for linking teleost and other vertebrate genomes [68, 66, 202, 28].

Extant holosteans include seven species of Lepisosteidae (gar; [209]) and two species of Amiidae (bowfin; [33, 225]). Analyses of the spotted gar *(Lepisosteus oculatus)* genome demonstrated the potential of holostean genomes for comparative studies, providing critical insights into the evolution of vertebrate immunity, development, and the function of regulatory sequences [28]. Recently, the alligator gar (*Atractosteus spatula*) genome was incorporated into an analysis of how vertebrates made the transition from water to land [18], and the genome of the distantly related eyetail bowfin (*Amia ocellicauda*, previously *Amia calva*, see [33]), provided understanding into other aspects of early vertebrate diversification including the evolution of scales, loci associated with the vertebrate adaptive immune response, and numerous other traits [202]. These studies have been imperative for our understanding of vertebrate evolution and molecular biology. However, they also underscore the potential insights that genomic sequencing of the remaining holostean genomes would provide. In particular, sequencing additional holostean species, with more focused investigations of within-clade sequence evolution, would facilitate a better understanding of highly fragmented regions in currently available holostean genome assemblies. In this study, we present a high-quality assembly and annotation for the longnose gar (*Lepisosteus osseus*). This fourth holostean genome fills a critical sampling gap among holosteans, providing a valuable resource for genomic investigations of early vertebrate evolution as well as the necessary context for bridging research between model teleosts and the human genome.

2.4 Results and Discussion

Assembly and coverage of universal orthologs

Here we report a high-quality assembly of the longnose gar (Figure 2.1(A) genome (Supplementary Table 2.1 and NCBI Bioproject PRJNA811181). Dovetail Genomics (Scotts Valley, CA) performed DNA extraction from a longnose gar blood sample, library preparation, sequencing, and genome assembly. Genomic DNA was extracted using a Qiagen Blood and Cell Culture DNA Midi Kit (Germantown, MD), yielding DNA with an average fragment length of 95 kbp that was used in the construction of Chicago HiC sequencing libraries. The 10X supernova assembly resulted in 27,738 scaffolds forming a total final genome size of 1,014,182,714 bp, with 2.4% of the genome (24,076,280 bp) comprised of the ambiguous base 'N' and a GC content of 40.1%. During Dovetail Hi-Rise assembly the input assembly was further incorporated into 22,745 longer scaffolds. The total length of the resulting Dovetail Hi-Rise assembly was 1014.98 Mbp, with a contig N_{50} of 116.6 kbp. The N_{50} of the assembly was 52.996 Mbp scaffolds with a L_{50} of 8 scaffolds. This is similar to the spotted gar genome that is 945 Mbp long, with a contig N_{50} size of 68.3 kbp, and a scaffold N50N_{50} size of 6.9 Mbp [28] and the eyetail bowfin reference genome which is 527 Mbp, with a scaffold N_{50} of 41.2 Mbp, an L_{50} of 9 scaffolds, and contig N_{50} of 21.1 kbp ([202]) (Figure: 2.1B).

The L_{90} based on this assembly is 26 which is close to the known karyotype of 28 for longnose gar [166] and the known karyotype of 28 for the Tropical Gar (*Atractosteus tropicus*, [9]).

However, this contrasts with the spotted gar, where a chromosomal spread suggested 29 linkage groups [28]. Our results support this change in karyotype between the spotted gar and longnose gar (Figure 2.1C), revealing a scaffold in the spotted gar genome with no strong syntenic relationship to any in the longnose gar.

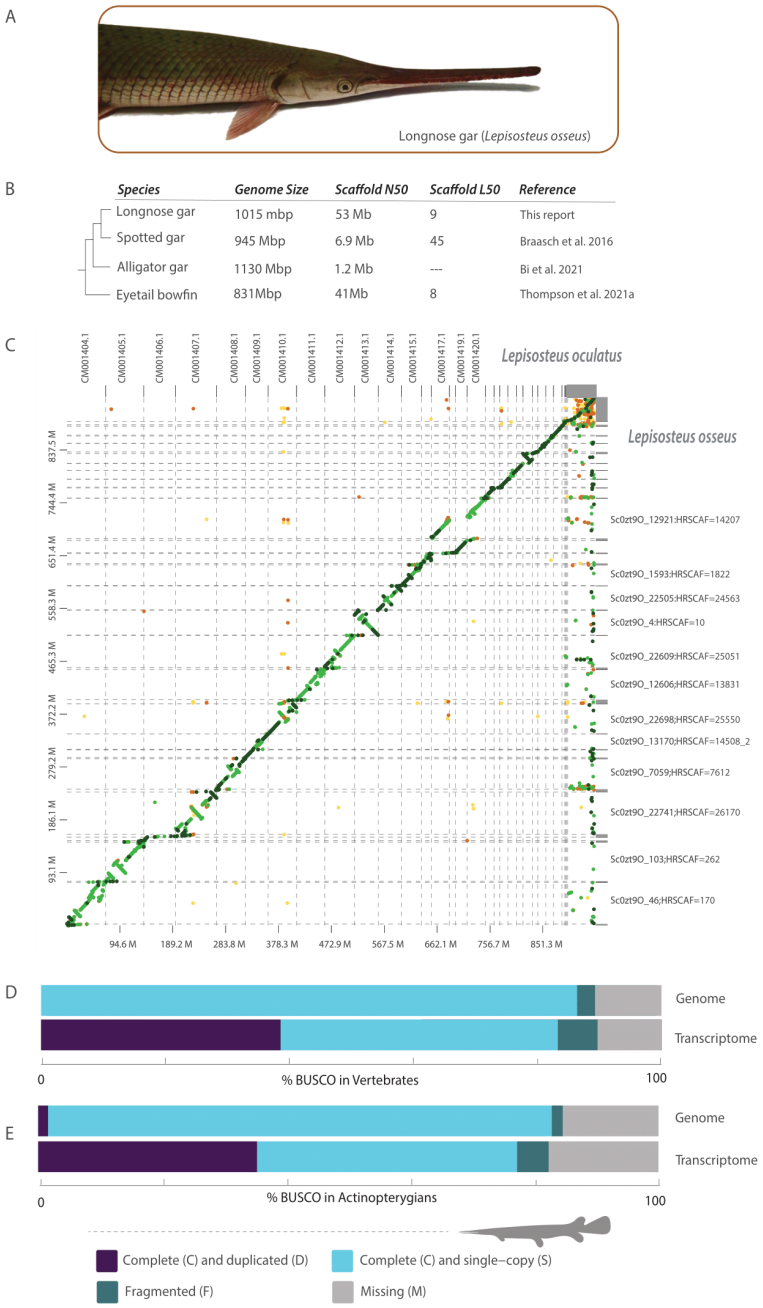


Figure 2.1: The annotated longnose gar genome.(A) Photo of a longnose gar wild-caught from North Carolina. (B) Comparison of holostean genome metrics. (C) Synteny between the longnose gar and the spotted gar shows 26.26% of the sequences have >75% match (dark green), 53.60% sequences have a match between 50% to 70% (light green), 5.77% sequences are between 25% to 50% (orange) and 1.42% fall below 25% (yellow). 12.95% sequences have no match. (D) BUSCO scores of the longnose gar genome and transcriptome compared to vertebrates (E) and actinopterygians. Photo credit: AD.

Analysis of the remainder of the spotted gar genome reveals a high degree of synteny with the longnose gar genome, with several possible inversions. These contrasts in karyotypes between longnose and spotted gars illuminated by our analyses have implications for gar conservation and management. When they co-occur in the wild, both longnose gars and spotted gars can hybridize with alligator gars (*Atractosteus spatula*) [22], a species that is experiencing declines in population across its range [191, 150, 23]. Given the differences in chromosomes between longnose and spotted gar species, this raises questions concerning hybrid offspring fitness and fertility that are of fundamental importance for species management. Our sequencing of the longnose gar genome could provide a useful tool for such efforts by providing a framework for marker development for hybrid identification as well as for the detection of historic introgression events.

Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis (Manni et al. 2021) comparing the longnose gar genome against the Actinopterygii dataset recovered a total of 3082 out of 3640 loci. Of these, 3017 (82.8%) are complete (2957 (81.2%) complete and single copy, 60 (1.6%) complete and duplicated), 65 (1.8%) fragmented, and the remaining 558 (15.4%) missing (**Figure 2.1D** and **Supplemental Table 2.2**). These numbers change slightly when compared to the Vertebrata BUSCO dataset. Out of 3354 total BUSCO groups, we recovered 2898 (86.4%) complete sequences [(2867 (85.5%) complete and single copy, 30 (0.9%) complete and duplicated], 97 (2.9%) fragmented, and 359 (10.7%) missing (**Figure 2.1E** and **Supplemental Table 2.2**). These results were similar for transcriptomes, with a higher number of duplicated genes reflecting likely splice variants. The higher number of sequences recovered when using the vertebrate versus actinopterygian databases mirrors a similar near twofold difference in missing data in the bowfin genome [202]. This may reflect a teleost bias for Actinopteryigan BUSCOs, or stem from a difference in

the number of target loci. More work is needed to assess if the development of a BUSCO dataset for early diverging, non-teleost, actinopterygians is warranted.

New insights into the transposable elements of Holostean genomes

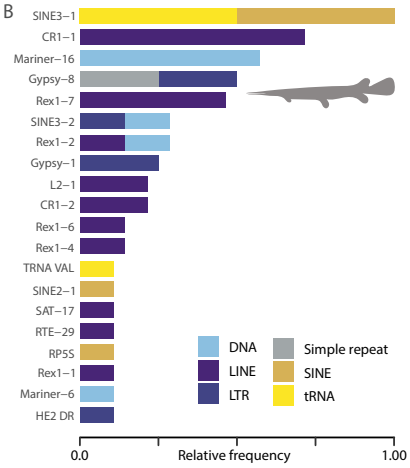
Our RepeatModeler analysis [76] reveals 29.12% of the longnose gar genome is composed of transposable elements (TEs; Figure 2.2A). Retroelements account for 14.02% of the transposons, 2.06% of which are short interspersed nuclear elements (SINEs) and 6.53% of which are long interspersed nuclear elements (LINEs). Among LINEs, L2/CR1/Rex elements represent 4.31% of the total diversity, while L1/CIN4 elements represent only 0.35% (Figure 2.2A). DNA transposons cover 3.72% of the genome, with Tc1-IS630-Pogo elements reflecting 2.65% of the total diversity. In general SINES and LINES are some of the most frequent TE types (Figure 2.2B). Our results are on par with the 20% TE content found in the spotted gar, and contrasts with the levels of near 50% TE content in humans or zebrafish (Danio rerio) (Figure 2.2C) [28]. These comparisons are based on values reported in the literature. However, given the magnitude of the contrasts between some species and overlap with others, a formal reanalysis of mobile elements across these representative organisms would likely yield similar contrasts. Similar to spotted gar, we find a high diversity of eukaryote TEs in the longnose gar genome after conducting a BLAST search against the Repbase [102] database. 235 sequences of repeats matched to Repbase: 46 DNA transposons, 71 LINEs, 25 SINEs, 28 Long Terminal Repeats (LTRs), 20 tRNA and 37 sequences classified as Unknown. Our finding of CR1 parallels a similar discovery in the spotted gar, which was used as evidence to suggest that the absence of CR1 is a teleost-specific loss, and not a general condition of early ray-finned fish (Figure 2.2A) [28]. Additionally, our finding of CR1, L2, Rex1, and Babar reveals longnose gar to be the third known vertebrate with all four CR1-like families. As the only other known vertebrates with all four CR1-like families are the spotted gar and bowfin, this

finding highlights the potential utility of holosteans for understanding the evolution of early vertebrate TEs as well as the need for additional studies of TEs in earlier diverging ray-finned fishes to provide additional evolutionary context for changes between holosteans and teleosts.

A

	Number of sequences	Length Occupied(bp)	Percentage of sequence
Retroelements	450725	130442851	14.02 %
SINEs:	95652	19192539	2.06 %
Penelope	16978	4542618	0.49 %
LINEs :	213318	60798079	6.53 %
L2/CR1/Rex	147677	40096152	4.31 %
R2/R4/NeSL	737	258046	0.03 %
RTE/Bov -B	29740	9195814	0.99 %
L1/CIN4	9577	2907845	0.31 %
LTR elements:	141755	50452233	5.42 %
Ty1/Copia	614	131016	0.01 %
Gypsy/DIRS1	13022	9321648	1.00 %
Retroviral	18520	16657741	1.79 %
DNA transposons	129138	34607661	3.72 %
hobo - Activator	23979	6005361	0.65 %
Tc1 -IS630 -Pogo	90891	24674839	2.65 %
PiggyBac	459	104448	0.01 %
Tourist/Harbinger	1464	429370	0.05 %
Rolling-circles	2089	383539	0.04 %
Unclassified:	472697	96609907	10.38
Satellites:	1533	572446	0.06 %
Simple repeats:	131239	6012273	0.65 %

B



C

Species	DNA	LTR	LINE	SINE	Unclassified	Total TE	
Eyetail bowfin	7.09	2.27	3.93	0.33	6.17	22.12	Holostei
Longnose gar	3.72	5.42	6.53	2.06	10.38	28.11	
Spotted gar	3.54	2.55	5.76	2.73	5.54	20.12	
Zebrafish	38.50	4.37	4.02	2.22	2.10	51.21	
Atlantic cod	6.39	4.88	3.30	0.58	2.81	17.96	Teleostei
Medaka	3.10	0.70	2.40	0.80	9.20	16.20	
Coelacanth	1.65	0.86	4.47	1.87	13.60	22.45	Actinistia
Western clawed frog	25.00	2.35	5.90	0.40	0.50	34.15	
Western painted turtle	0.93	0.77	5.52	1.75	0.14	9.11	
Green anole	8.70	2.90	12.20	5.11	0.75	29.66	
Chinese alligator	1.99	10.29	29.13	0.80	2.16	44.37	
Chicken	0.80	1.30	6.40	0.10	-	8.60	
Domestic mouse	0.88	9.87	19.20	8.22	0.38	38.55	
Modern human	3.03	8.55	20.99	13.64	0.15	46.36	Tetrapoda

Figure 2.2: Overview of repetitive sequences within the longnose gar genome. (A) The distribution of repeat types identified from the longnose gar genome using RepeatModeler [76]. (B) The relative frequency of the 20 most frequent TEs that matched to Repbase. Colors correspond to TE types. (C) Comparison of TE class (DNA, LTR, LINE and SINES) percentages in the longnose gar with other vertebrates. Values for other vertebrates were obtained from literature: spotted gar [28]; eyetail bowfin [202]; zebrafish [94]; Atlantic cod [193]; medaka [105]; coelacanth [7]; Western clawed frog [90]; western painted turtle [185]; green anole [5]; Chinese alligator [216]; Chicken [98]; domestic mouse [136]; and modern human [114]. Photo credits: AD.

Transcriptome sequencing and gene ontology

Transcriptome sequences were assembled from RNA-Seq derived from immune tissues (spleen, kidney and intestine) of the same longnose gar individual from which the genome originated that were pooled prior to sequencing. The transcriptome sequences were de novo assembled using Trinity [84, 86] and mapped to the genome using HISAT2. Processed reads were mapped to the genome and assessed using samtools v.1.18, with 30901143 out of 44987853 reads (68.69%) mapping to the genome. BUSCO [124] analysis was employed on the assembled transcriptomic sequences to quantify the completeness of the transcriptome (**Figure 2.1** and **Supplemental Table 2.3**). Comparing the longnose gar transcriptome to the Actinopterygii and Vertebrata databases yielded 2809 (77.2%) and 2792 (83.3%) complete sequences, respectively. The vertebrate database yielded a higher number of complete and single-copy (44.7%) and complete and duplicated (38.6%) sequences than the comparison to the actinopterygian database (41.9% and 35.3% respectively).

The transcriptome was used to annotate the genome (Zenodo DOI: 10.5281/zenodo.7435126), predicting 30,068 proteins. This is similar to the 25,645 proteins predicted in the spotted gar genome by MAKER [28]). Gene ontology analysis of transcripts from pooled spleen, kidney and intestine RNA reveals the top molecular functions to include cytokine activity, signaling receptor regulator activity, signaling receptor activator activity and also include proteasomal complexes in the cellular components. These are all reflective of the immunological roles of these tissues **(see Supplemental Figures S2.1-S2.3)**.

2.5 Conclusion

Our sequencing of the longnose gar genome fills a sampling gap in the genome sequences available for early diverging ray-finned fishes, thereby providing a critical resource for comparative genomic investigations. This genome has immediate utility in investigations concerning the capacity for gar species to hybridize that have implications for management as well as our understanding of hybridization in general. As the evolutionary divergence between longnose gars and alligator gars is estimated to be over 100 million years old [33], leveraging this new genomic resource could provide key insights into how hybridization remains possible over such an extreme evolutionary timescale. Additionally, this genome provides the framework for numerous comparative genomic investigations that could expand our understanding of transposable element evolution, the evolution of vertebrate gene families, developmental biology, or even aiding in the linking of translational research in fish models such as zebrafish or medaka to the human genome.

2.6 Methods and Materials

Sample acquisition

All research involving live animals was performed in accordance with relevant institutional and national guidelines and regulations, and was approved by the North Carolina State University Institutional Animal Care and Use Committee. The longnose gar specimen was wild-caught on the Haw River, North Carolina (35.626174, -79.057769) by the NC Wildlife Resources Commission using standard boat electroshocking methods and housed at the NC State University College of Veterinary Medicine in a 300 gallon tank with recirculating water at 18-23°C. The individual was anesthetized using MS-222 and 2.5 mL of blood was collected into 0.5 mL of 87 mM EDTA for genomic sequencing. The fish was euthanized, and supplemental tissue

samples (spleen, kidney and intestine) were collected for transcriptome sequencing.

Chicago and Dovetail Hi-C library prep and sequencing

A Chicago library was prepared by Dovetail Genomics using ∼500 ng genomic DNA and methods described in [165]. In brief, chromatin was reconstituted in vitro and crosslinked with formaldehyde. Chromatin was digested (*Dpn*II) and the resulting 5' overhangs were filled with biotinylated nucleotides. Blunt ends were ligated and DNA was purified. Purified DNA was obtained from protein by reversing the crosslinks and subsequently treated to remove the biotin that was not initially internal to the ligated fragments. The Hi-C library was then created using the methods as described in [117] shearing the DNA to ∼350 bp mean fragment size. The sequencing libraries were generated using NEBNext Ultra enzymes and Illumina compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeqX and yielded 163 million paired end reads (2 × 150 bp) that provided 7514.79X physical coverage of the genome (10-10,000 kbp).

Scaffolding the assembly with Hi-Rise

Dovetail staff used HiRise [165] to scaffold genome assemblies. The de novo assembly, Chicago library reads and the Dovetail HiC library reads were used as inputs for HiRise. The shotgun and Chicago library sequences were first aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu), that masked base pairs that followed a restriction enzyme junction. The Chicago data was aligned and scaffolded following aligning and scaffolding of the dovetail HiC library. Once all the sequences were aligned and scaffolded, shotgun sequences were used to close the gaps between contigs.

Contamination removal and species verification

Contaminated and adaptor sequences were identified with feedback from NCBI and removed using custom scripts. The species' identity was confirmed using tBLASTn searches with the universal barcode for fish species Cytochrome c Oxidase I (COI) as a query [99]. Custom scripts have been archived on Zenodo (DOI: 10.5281/zenodo.7435126)

RNA sequencing

RNA was extracted (Qiagen RNeasy kit) from the spleen, kidney and intestine of the same individual longnose gar as the genome sequence and quantity and integrity of the isolated RNA were assessed using a NanoDrop 1000 (Thermo Fisher) and Agilent Bioanalyzer respectively. In brief, mRNA was enriched using oligo(dT) beads, rRNA was removed using a Ribo-Zero kit (Epicentre, Madison, WI) and mRNA was randomly fragmented. Each RNA extraction was equalized for a final concentration of 180 ng/μL. Library preparation and sequencing was performed by Novogen Corporation (Sacramento, CA). Next-gen sequencing (2 $\times$ 150 bp paired end reads) was performed on a NovaSeq 6000 instrument (Illumina). Adapter sequences and poor quality reads were filtered with Trimmomatic v34 [25]. The transcriptome was de novo assembled with Trinity v2.11.0 [84]. Completeness of the transcriptomes was assessed using a BUSCO analysis [124]. Raw reads and computationally assembled transcriptome sequences were deposited onto NCBI under the accession numbers SRR19528583 and GKEG00000000, respectively.

Transcriptome sequences were further analyzed to assign gene ontology. The assembled RNA-seq from Trinity was translated using Transdecoder to identify the candidate coding regions in the transcript sequences. The longest open reading frames (orfs) output was used for BLASTx and BLASTp analysis against the uniprot

database (Nov 2021 release) to get the top target sequences for every transcript. Hmmscan v.3.3.2 was used to search for protein sequences in the Pfam-A (Nov 2021 release) library. Signalp v.5.0b [201] and TMHMM v.2.0c [110] were used to identify the signal peptides and the transmembrane proteins. Trinotate v.3.2.2 used sqlite database along with the Trinity assembled transcriptome and the longest orfs from Transdecoder to create a gene transmap [36]. The transcripts were finally annotated using Trinotate and then further analyzed to obtain the GO annotations. The GO terms were visualized using the enrichplot and ggupset packages in R. All code has been made available on Zenodo: (DOI: 10.5281/zenodo.7435126).

Annotation and genome quality assessments

BRAKER2 [34] was used to annotate the longnose gar genome, which uses GeneMark-ET [119] to predict the preliminary genes and generate a genemark.gtf output that was used for training with Augustus (Stanke et al. 2006). The genome was filtered to remove any duplicates and adapters. The transcriptome sequences were aligned using HISAT2 [107] to get an aligned sorted bam file. RepeatModeler [76] identified the repeats in the genome to prevent mis-annotation of the repeats as protein coding genes. The consensus file containing repeats was used as input for RepeatMasker [200] to soft-mask the repeats for BRAKER. The masked genome and the aligned transcriptomes were loaded into BRAKER to obtain the annotated proteins.

2.7 Data Availability

The longnose gar genome sequence is available through NCBI Bioproject PRJNA811181. Raw transcriptome reads and computationally assembled transcriptome sequences are available through NCBI under the accession numbers SRR19528583 and GKEG00000000, respectively. All code used for analyses and the BRAKER2 genome annotation are available on Zenodo (DOI: 10.5281/zenodo.7435126)

2.8 Acknowledgements

We thank Ingo Braasch and Andrew Thompson (Michigan State University) for very helpful discussions on genome analyses strategies, Kent Passingham (NC State University) for assistance with blood collection, and the NCBI staff for assistance with identifying adapter and other contaminating sequences. We thank A. Bogan for facilitating the photography of the *Anolis* in Figure 2.

2.9 Author Contribution

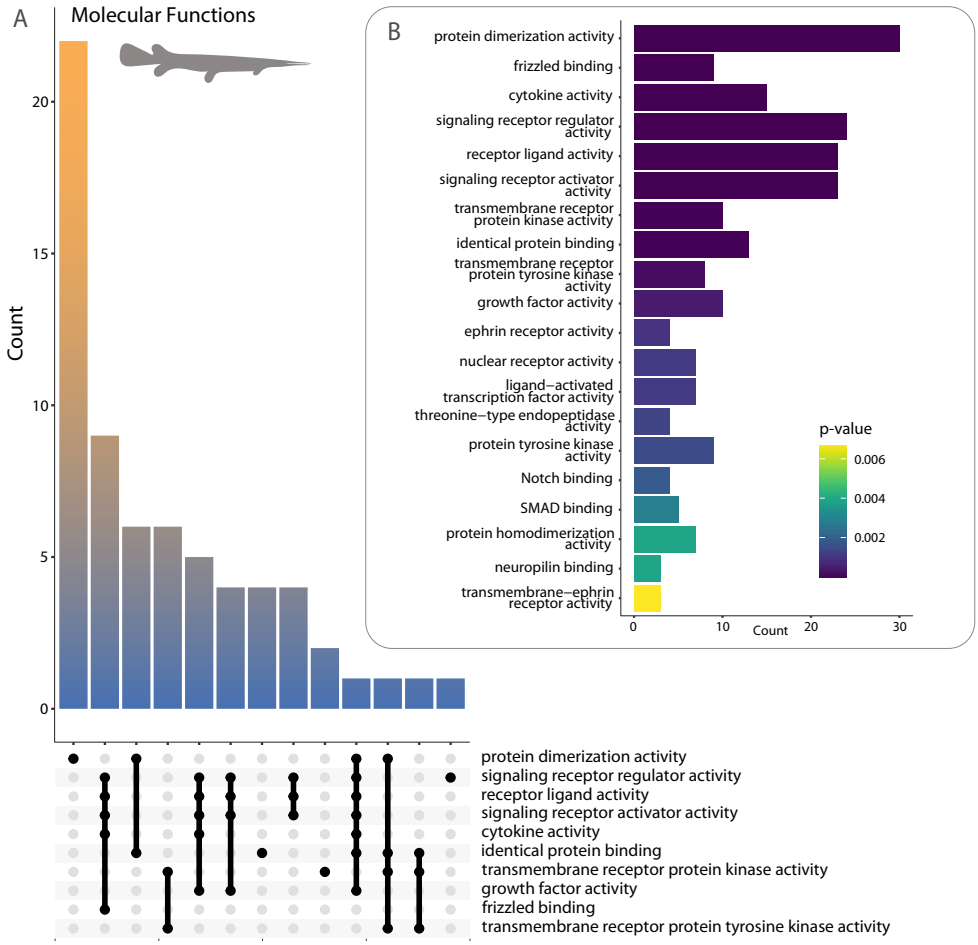
JAY and AD conceived of the project; MF, DJW, JAY and AD collected *Lepisosteus osseus*; RM executed genomic analyses including analysis of repetitive elements and genome annotation; DJW and RM assembled transcriptome sequences; RM, KBC and AD analyzed transcriptome sequences; RM conducted transcriptome annotation; RM and AD visualized data; RM, KBC, JAY and AD wrote the first draft of the manuscript; all other authors contributed to the subsequent writing; JAY and AD supervised the project. All authors read and approved the manuscript.

2.10 Funding

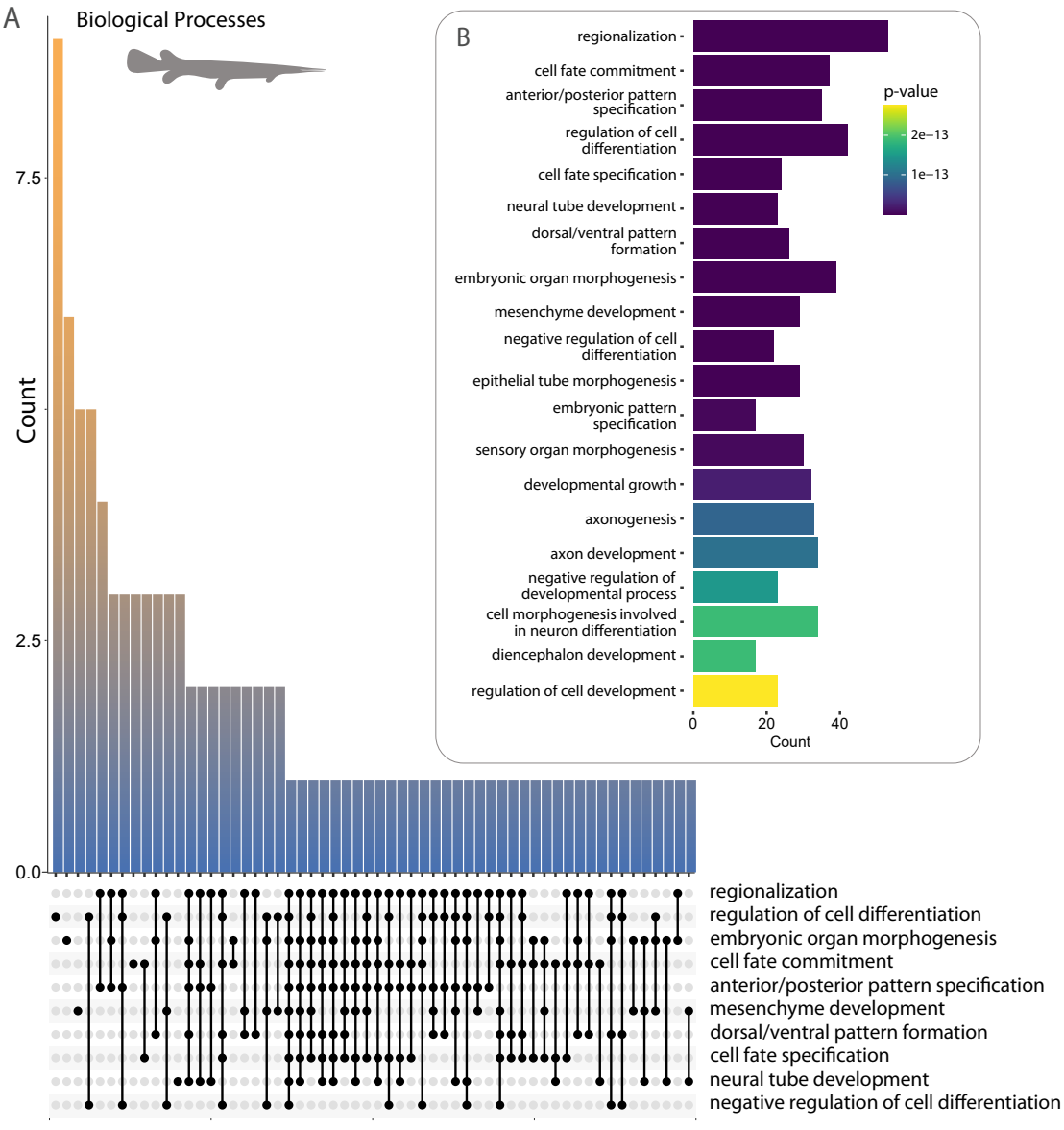
This work was supported by the National Science Foundation (IOS1755330 to JAY and IOS1755242 to AD), the National Evolutionary Synthesis Center, NSF EF0905606 (DJW), and funding from the Triangle Center for Evolutionary Medicine (AD and JAY).

2.11 Supplemental Tables and Figures

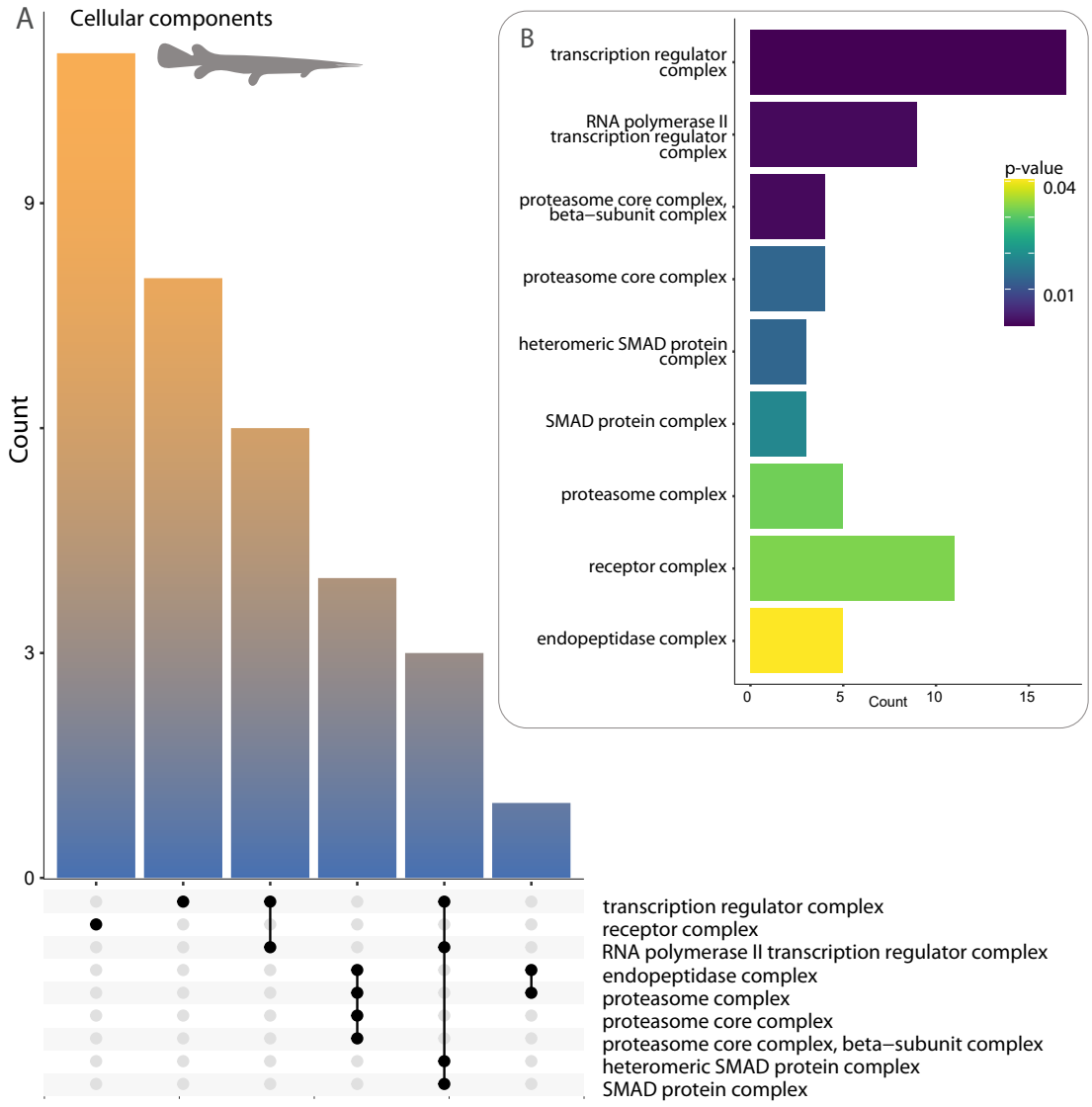
2.11.1 Supplemental Figures



Supplemental Figure S2.1: Summary of Gene Ontology *Molecular Functions* analysis from the longnose gar transcriptome. Predicted proteins from the transcriptome were used as inputs to assess (A) functions and their intersections and (B) most common terms.



Supplemental Figure S2.2: Summary of Gene Ontology *Biological Processes* analysis from the longnose gar transcriptome. Predicted proteins from the transcriptome were used as inputs to assess (A) functions and their intersections and (B) most common terms.



Supplemental Figure S2.3: Summary of Gene Ontology *Cellular Components* analysis from the longnose gar transcriptome. Predicted proteins from the transcriptome were used as inputs to assess (A) functions and their intersections and (B) most common terms.

2.11.2 Supplemental Tables

Supplemental Table 2.1: Genome assembly statistics of longnose gar

Feature	Value
GC content	40.1%
Number of Scaffolds	22,745
Number of scaffolds >1 kbp	22,709
Contig N50	116.61 kb
Scaffold N50	52.996Mb
Scaffold L50	8
L90	26 scaffolds
N90	5.560 Mb
Longest scaffold	74,198,471 bp
Number of gaps	27,358
Percent of genome in gaps	2.45%

Supplemental Table 2.2: BUSCO analysis of the longnose gar genome assembly

Feature	Actinopterygii	Vertebrata
Complete BUSCOs (C)	3017 (82.8%)	2898 (86.4%)
Complete and single-copy BUSCOs (S)	2957 (81.2%)	2867 (85.5%)
Complete and duplicated BUSCOs (D)	60 (1.6%)	30 (0.9%)
Fragmented BUSCOs (F)	65 (1.8%)	97 (2.9%)
Missing BUSCOs (M)	558 (15.4%)	359 (10.7%)
Total BUSCO groups searched (n)	3640	3354

Supplemental Table 2.3: BUSCO analysis of the longnose gar transcriptome assembly

Feature	Actinopterygii	Vertebrata
Complete BUSCOs (C)	2809 (77.2%)	2792 (83.3%)
Complete and single-copy BUSCOs (S)	1524 (41.9%)	1492 (44.7%)
Complete and duplicated BUSCOs (D)	1285 (35.3%)	1294 (38.6%)
Fragmented BUSCOs (F)	187 (5.1%)	214 (6.4%)
Missing BUSCOs (M)	644 (17.7%)	348 (10.3%)
Total BUSCO groups searched (n)	3640	3354

CHAPTER 3: INVESTIGATING THE IMPACT OF WHOLE GENOME DUPLICATION ON TRANSPOSABLE ELEMENT EVOLUTION IN RAY-FINNED FISHES

Submitted to Genome Biology and Evolution 2024

3.1 Abstract

Transposable elements (TEs) can make up more than 50% of any given vertebrate's genome, with substantial variability in TE composition among lineages. TE variation is often linked to changes in gene regulation, genome size, and speciation. However, the role that genome duplication events have played in generating abrupt shifts in the composition of the mobilome over macroevolutionary timescales remains unclear. We investigated the degree to which the teleost genome duplication (TGD) shaped the diversification trajectory of the ray-finned fish mobilome. We integrate a new high coverage genome of *Polypterus bichir* with data from over 100 publicly available actinopterygian genomes to assess the macroevolutionary implications of genome duplication events on TE evolution. Our results provide no evidence for a substantial shift in mobilome composition following the TGD event. Instead, the diversity of the actinopterygian mobilome appears to have been shaped by a history of lineage specific shifts in composition that are not correlated with commonly evoked drivers of diversification such as body size, water column usage, or latitude. Collectively, these results provide a new perspective on the early diversification of the actinopterygian mobilome and suggest that historic ploidy events may not necessarily catalyze bursts of TE diversification and innovation.

3.2 Significance

We investigate the role of the teleost genome duplication on transposable element (TE) diversification in ray-finned fishes by integrating an analysis of the mobilome from a newly sequenced genome from Polypterus bichir with analyses from over 100 ray-finned fish genomes. We reveal that ray-finned fish TE diversity depicts a signature of lineage-specific shifts rather than a major burst of novelty near the origin of teleosts, suggesting a complex, nuanced history of TE evolution. Our results challenge the impact of ploidy events on long-term TE evolution and set a new direction

for future research into the genomic and evolutionary mechanisms influencing TE diversity across half of all living vertebrates.

3.3 Introduction

Transposable elements (TEs) can account for over 50% of a vertebrate's total genome content [192]. Aptly dubbed "jumping genes", TEs possess the astonishing ability to rearrange and reposition themselves within a given genome [129], through the use of two primary strategies for transposing. Class I retrotransposons replicate through a "copy and paste mechanism" involving an RNA intermediate. This RNA molecule is reverse-transcribed back into a DNA copy, which is then seamlessly integrated into the genome, leaving the original template element untouched [92]. In contrast, class II DNA transposons use a "cut-and-paste" technique to excise themselves from their current location and relocate to an alternate genomic location [137]. Comparative studies of vertebrate TEs have revealed substantial heterogeneity in the composition of class I and class II TEs between major clades of vertebrates [49], These studies have implicated changes in TE composition with altered gene regulation [205, 170], evolutionary changes in genome size [24, 141, 223], evolutionary novelties [181], changes in life history [147], and speciation [183, 175] to name but a few. While the past decade has yielded tremendous strides towards developing an understanding of the general hallmarks of TE evolution, the evolutionary fate of TEs following whole genome duplication events is just beginning to emerge [159, 176, 158].

Genome duplication events have the potential to amplify genome complexity, with emerging evidence highlighting a possible role for duplication events enabling phenotypic evolution through duplicated genes [135]. Studies within individual or closely related species that vary in their level of ploidy have revealed substantial modifications that can facilitate profound epigenetic repatterning within the domains of TEs. However, the impact of such changes on the mode of TE diversification remains unclear.

For example, surplus gene copies can compensate for potential losses or modifications in expression caused by TE insertions, thereby facilitating extensive genomic modification through the actions of transposable elements. This redundancy hypothesis [128] suggests a substantial shift in TE content and composition following a genome duplication event that continues as lineages diversify [159]. An alternate hypothesis argues that genome duplication events correspond to a transitory phase for a species that is characterized by diminished population size [120]. This bottleneck hypothesis argues that as the efficacy of selection decreases, the likelihood of moderately deleterious TE insertions within nascent polyploid genomes becoming fixed is increased. Consequently, this hypothesis would expect a pulse of TE diversification coincident with a genome duplication event [159]. Recently, a hypothesis framed around the concept of "Hopeful Monster" suggested that the balance between the deleterious and beneficial aspects of TE proliferation can lead to occasional beneficial mutations that can promote speciation and contribute to the emergence of new traits [207]. In contrast to the previous two hypotheses, this hypothesis predicts that genome duplications would result in only a limited number of TE changes, not a substantial pulse of diversification. Given that these hypotheses represent corner cases on a continuum of possibilities and alternate hypotheses, empirical comparative genomic studies at different evolutionary timescales are key to understanding the implications of genome duplication events on the evolution of TEs.

Ray-finned fishes (Actinopterygii) offer an exemplar group for the macroevolutionary study of TEs and genome duplication event. Comprising half of all living vertebrate species, ray-finned fishes have successfully radiated across virtually all aquatic habitats including swamps [4], abyssal ocean depths [55, 131], polar and high altitude regions [160, 62, 57, 143] and caves [221]. The evolutionary success of ray-finned fishes is unusual relative to other vertebrate groups. Over 99% of the over 35,000 living

species of actinopterygians are teleosts, a clade that experienced a genome duplication event in its early evolutionary history [83, 35]. In contrast, the three remaining extant ray-finned fish lineages – Holostei (gars and bowfin, 8 species), Acipenseriformes (sturgeon and paddlefish, 29 species), and Polypteridae (bichirs and ropefishes, 14 species) – did not undergo the teleost specific genome duplication event. Recent studies of holostean genomes have highlighted differences in the composition of TEs between holosteans and model teleosts such as zebrafish (*Danio rerio*) and medaka *(Oryzias latipes*) [28, 202, 122], raising the possibility of an evolutionary shift following a genome duplication event in teleosts. Comparative genomic studies are needed to rule out the possibility that the mobilomes of holosteans, and not teleosts, are unusual relative to all other ray-finned fishes. To date, no comparative studies of the ray-finned fish mobilome have included a broad representative sampling of teleosts, holosteans, acipenseriforms, and polypterids. With the growing number of genomes of ray-finned fishes deposited in public data repositories, such studies are now possible.

Over twenty years ago, the *Takifugu rubripes* genome represented the first ray-finned fish genome sequenced [8]. Decreased costs of sequencing have since led to a surge of efforts to sequence additional actinopterygian genomes [72, 168] that now present a wealth of resources for phylogenetically comprehensive comparative studies. In addition to the availability of hundreds of teleost genomes, there are genomes of a few for non-teleost ray-finned fishes. For example, the sterlet (*Acipenser ruthenus*) genome revealed an independent historic genome duplication in this chondrostein lineage that provides an additional ancient duplication for understanding evolutionary patterns of TE diversification in actinopterygians [51]. Similarly, the sequencing of the Senegal bichir (*Polypterus senegalus*) leveraged the anatomy of polypterids to provide critical insights into how vertebrates achieved the transition from water to land [18]. In addition, the recently sequenced genome of a second polypterid, *Erpeto-*

ichthys calabaricus [Reedfish; Assembly (ErpCal1.1; NCBI Annotation Release 100)], has been deployed alongside the *P. senegalus* genome to provide further insights into other aspects of early vertebrate diversification that include the evolution of keratins [108], olfactory receptors [235], and numerous other traits [89, 130]. However, based on BUSCO scores, the currently sequenced genome of *P. senegalus*, remains incomplete. As the only *Polypterus* genome sequenced, the lack of an additional *Polypterus* genome challenges interpretation of results concerning the distribution of TEs in this lineage. Given the phylogenetic position of polypterids, additional genome sequencing efforts are of extreme value for contextualizing the diversification of ray-finned fish TEs.

Here we present a high-quality chromosome-level assembly and annotation for an additional *Polypterus* species, *Polypterus bichir*. We integrate analyses of TEs within this genome with data on TE content for all other major lineages of actinopterygians to investigate the impact of genome duplication on the early evolution of the teleost mobilome. Using a comparative phylogenetic framework, we test for associations between genome size and the composition of the mobilome, variations in mobilome composition between teleosts and non-teleost ray-finned fish lineages, as well as possible correlations between TE content and aspects of actinopterygian biodiversity such as habitat, body size, or water column occupation. We additionally reconstruct the ancestral mobilome of actinopterygians through the TGD event and assess the phylogenetic signal of Class I and Class II TEs. These results provide a new perspective on the early diversification of the actinopterygian mobilome and the impact of the TGD on its evolution.

3.4 Results and Discussion

3.4.1 *Bichir genomes: an example of evolutionary conservation or recent divergence?*

We present a high-quality assembly of the *Polypterus bichir* genome (NCBI Bioproject PRJNA811142). The results of a BLAST search of mitochondrial COI from our specimen against polypterid barcode sequences on NCBI verified the identification as *Polypterus bichir lapradei*, a currently not recognized subspecies based on morphology that has been suggested to represent a genetically distinct lineage [144]. The 10X supernova assembly from Dovetail Genomics (Scotts Valley, CA) resulted in 130,773 scaffolds forming a total final genome size of 3,962,089,718bp. 13.9% of the genome (550,533,170bp) is composed of the ambiguous base "N" and a GC content of 39.21%. During Dovetail Hi-Rise assembly, the input assembly was further incorporated into 70,587 longer scaffolds. The total length of the resulting Dovetail Hi-Rise assembly was 3905.43 Mbp, with a contig N50 of 37.39 kbp. The N50 of the assembly was 202.693 Mbp scaffolds with a L50 of seven scaffolds. This is similar to the *E. calabaricus* genome which is 3.6Gb long, with a contig N50 size of 6.8 Mb, and a scaffold N50 size of 217.7 Mbp, and the *P. senegalus* reference genome, which is 3.7 Gbp, with a scaffold N50 of 189.69 Mbp, and contig N50 of 4528.14 kbp (**Supplemental Table S1**). However, a comparison of the BUSCO analysis on the *P. senegalus* reference genome to the newly sequenced *P. bichir* genome reveal a striking difference between the two assemblies (**Supplemental Table S2 and Supplemental Figure S4.2**). We find nearly triple the number of actinopterygian orthologs in the *P. bichir* genome relative to the *P. senegalus* genome (**Supplemental Figure S4.2**). This suggests that the *P. bichir* sequence may fill additional gaps in our understanding of the genomic evolution of polypteriforms outside of the mobilome elements discussed in this study.

Contrasting patterns of synteny between our new sequenced genome and those of *P. senegalus* and *E. calabaricus* using D-genies supports a high level of synteny between these species (Supplemental Figure S3.2). Further, the number of *P. bichir* super-scaffolds (18) reflect the described karyotypes of most polypterids: *E. calabaricus* (n=18), *Polypterus palmas* (n=18), *Polypterus delhezi* (n=18), *P. senegalus* (n=18), and *Polypterus ornatipinnis* (n=18)[10, 111, 133, 212]. The only exception to this chromosome count known in polypterids is in *P. weeksi* (n=19) [212], the sister lineage to *Polypterus ornatipinnis*. While closely related vertebrate taxa can often diverge substantially in their TE content over time, we find that this is not the case in bichirs. Instead, the conservation of karyotype and synteny across these species is also reflected in the relative abundances of TEs across these species. For *P. bichir*, RepeatModeler quantified 53.40% of the genome to be composed of transposable elements. This is similar to *P. senegalus* with (54.66%) and *Erpetoichthys calabaricus* (60.35%) and overall posits a surprising level of genomic conservation between species of polypterids. This conservation could be a consequence of multiple factors including slow rates of molecular evolution such as those observed in gars and Bowfin [225], the possible geologically recent Early Miocene crown age of Polypteridae estimated using molecular clocks [144], or a combination of the two to name but one set of possibilities.. Indeed, Kimura distances of Polypterus bichir TEs S3.28, reveal that the bichir genome is dominated by recent copies of mostly DNA and LINE transposons (K 5). Regardless of the mechanism, our sequencing of the *P. bichir* genome reveals polypterids as a candidate lineage for future studies of the mechanisms that promote genomic stability within a clade.

3.4.2 *The evolution of the ray-finned fish mobilome*

Placing our characterization of the polypterid mobilome into the context of other major ray-finned fish lineages reveals a complex history of mobilome evolution over

the past 400 million years (**Figure 3.1A**). The reconstructed phylogenetic history of TE evolution in actinopterygians is not consistent with a sudden TE proliferation coincident with the TGD event. Instead, our analyses reveal that compositional TE patterns at this scale largely reflect shifts within actinopterygian subclades. For example, cichlids have similar compositional patterns relative to pufferfish, the latter of which exhibits an increase in the number of LINE elements (**Figure 3.1A**). Independently, Lampriformes (*Regalecus & Lampris*) also exhibit an expansion of their LINE elements to a relative level similar to that observed in pufferfishes (Figure 3.1A). The ubiquity of such patterns of clade-specific heterogeneity are strongly supported by tests of phylogenetic signal using both Pagel's lambda (λ) and Blomberg's K (K) (**Supplemental Table S3**). In all cases, K values are significant (DNA | *K=0.478, p=6e-04*; LTR | *K=0.405, p=1e-04*; LINE | *K=0.945, p=1e-04*; and SINE | *K=0.529, p=3e-04*). Likewise, quantifications of lambda values for DNA (λ*=0.98, p=1.89e-13*), LTR (λ*=0.819, p=6.64e-11*), LINE (λ*=0.975, p=1.51e-19*), and SINE (λ *1.0, p=4.37e-15*) are very close to 1, indicating strong phylogenetic signal, supporting a pattern of similar relative TE abundances between closely related taxa and increasing disparity between subclades.

Contrasting relative TE content (**Figure 3.1B**) against total genome size reveals numerous expansions and contractions of both genome size and relative TE abundances across the phylogenetic diversity of ray-finned fishes. Importantly, there is no signal of a shift in the rate of TE evolution with the origin of teleosts. Instead, the reconstructed history indicates that expansions and contractions of both relative TE abundance and genome size appear to heterogeneously occur across various teleost and non-teleost lineages.

Phylogenetic regression analyses indicate that shifts in genome size are weakly

correlated with changes in the relative abundances of TEs (**Figure 3.1C; Supplemental Table S3**). Correspondingly, we find strong evidence that genome size in ray-finned fishes also exhibits a pattern of strong phylogenetic signal (*K=0.659, p=6e-04;* $\lambda = 0.717$*, p=4.45e-15*), paralleling a major trend in genome size evolution that has been found across the Tree of Life. Lineages as disparate as angiosperms [15, 6], *Drosophila* [184], marine dinoflagellates [118], bacteria and Archaea [126] exhibit strong phylogenetic signal in genome size evolution. Given that the relative abundance of TE content in a genome is weakly correlated with genome size in ray-finned fishes [116] (**Figure 3.1D**; and numerous other lineages [49]), it is possible that the evolution of TEs exhibits a signature of strong phylogenetic signal across the Tree of Life [61].

Multiple factors promote genome size evolution ([153]), including hypothesized correlations between the molecular evolution of the mobilome and overall increases in genome size. For example, transposition rates generally exceed excision rates ([106]), enabling TEs to contribute to the enlargement of genomes and exert an influence on genome size evolution. However, whether the TGD event resulted in a shift in the mode of evolution between the teleosts and non teleosts mobilome remains unclear. Using a phylogenetic analysis of variance (ANOVA), we find no support for a significant difference in TE content between teleosts and non-teleosts ($p = 0.06$; **Figure 3.1D**). Additionally, phylogenetic regression analyses strongly support a correlation between the relative content of all mobilome elements, indicating that the increase or decrease in one will also result in an increase in the other (**Figure 3.2**). These results suggest that the overall relative composition of the ray-finned fish mobilome has been shaped more by recent lineage specific shifts in genome evolution than historical legacy stemming from the TGD event.

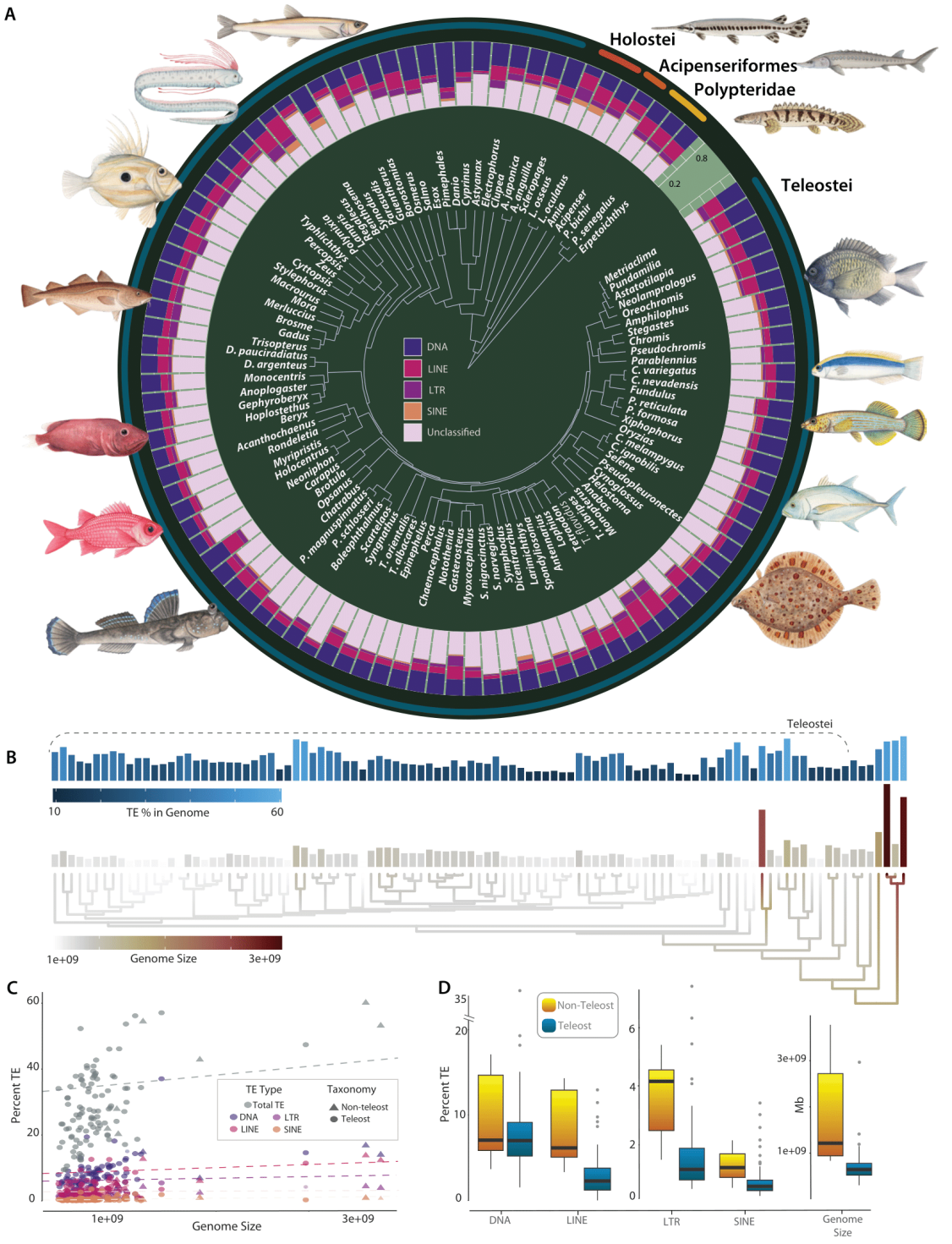


Figure 3.1: Major patterns of TE evolution across the evolutionary history of ray-finned fishes. (A) The relative abundance of DNA, LTR, LINE, SINE transposons, and unknown transposons relative to actinopteryigiian phylogeny are shown. (B) Total TE% and absolute genome size in the context of evolutionary history are compared. (C) Results of a phylogenetic regression to assess the relationship between genome size and TE abundance are provided. (D) TE% and genome size between teleosts and non-teleost actinopterygians are compared. Colors in A correspond to elements labeled in the panel. Colors in B are shaded relative to high (light blue or red) and low values (dark blue, or beige) in the upper and lower panels respectively. Colors in C correspond to the labels in A for individual elements. Panels in D correspond to the quantiles of each category with dark horizontal lines indicating the mean value

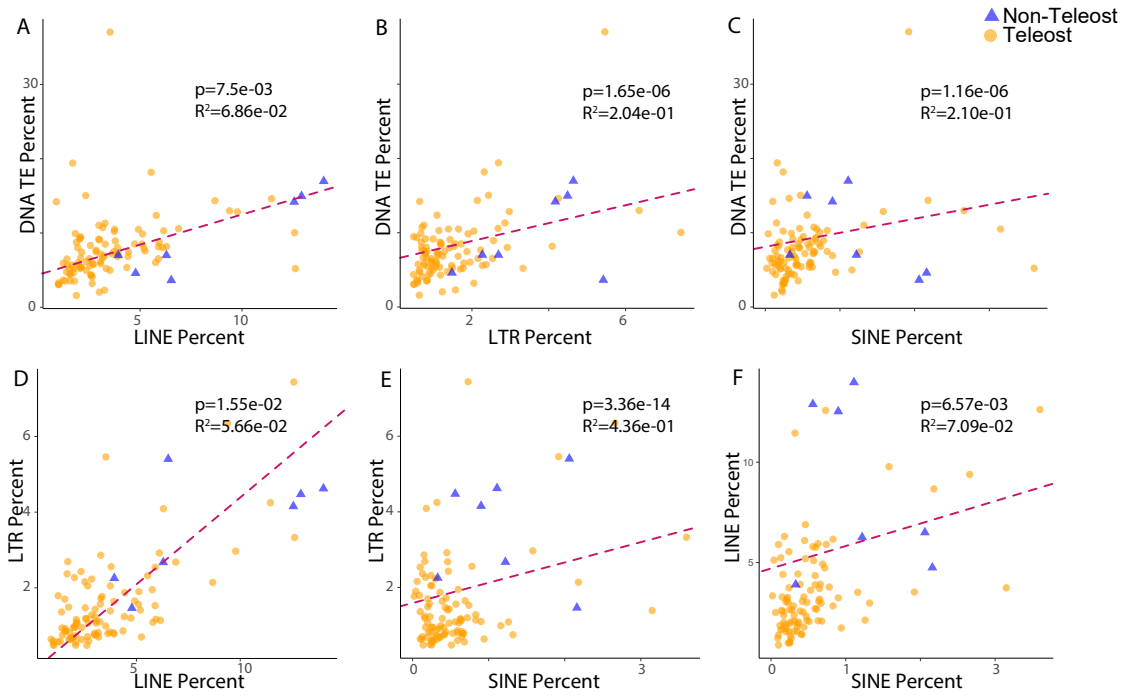


Figure 3.2: Correlation between relative class I and class II mobilome content. Panels (A), (B), and (C) display the correlations between DNA transposons and LINE, LTR, and SINE transposons, respectively. Panels (D) and (E) illustrate the correlations between LTR transposons and LINE and SINE transposons, respectively. Panel (F) presents the correlation between LINE and SINE transposons

3.4.3 *Concerning the TGD and the appearance of novel mobilome elements*

It is possible that the TGD catalyzed the evolution of major groups of novel TEs. However, when placing the twenty-five main superfamilies of the teleost mobilome [49] into the context of the phylogenetic diversity of ray-finned fishes, we reveal that this effect was muted. We estimated the ancestral mobilome of actinopterygians utilizing a model-averaged stochastic character mapping approach [174]. Out of 25 superfamilies, the presence of 24 superfamilies observed in teleosts is mirrored within non-teleosts (**Figure 3.3**). The only exception to this is the DNA transposon superfamily Chaepev ([104]). It is unlikely that this superfamily arose as a consequence of the TGD, as it is present in several arthropods as well as a range of vertebrate species including anoles [148], and lampreys [235]. Chapaev sequences have also been

documented from White Sturgeon (*Acipenser transmontanus*) [235], suggesting that additional species of non-teleost ray-finned fishes possess this superfamily.

Genome duplication events are hypothesized to form a substrate for genomic innovation [52]. In contrast to this expectation, our results reveal numerous instances of numerous likely TE losses as well as independent gains. For example, we find repeated losses of TE superfamilies such as LINE Dong, DNA TcMariner, and DNA Crypton in teleosts (Figure 3.3). LTR-Pao is absent in Longnose Gar and Bowfin, suggesting a loss in holosteans. LTR-DIRS are absent in all three polypterids (*P. bichir, P. senegalus, and E. calabaricus*). In addition, we identify the TE superfamily Jockey, a LINE element, exclusively in *Lepisosteus osseus* and *P. senegalus*. Jockey has been previously confirmed in coelacanth [48] and lampreys, and this marks the first reported instance of its presence in actinopterygians. It is possible that this element has been lost multiple times independently, however Jockey is known to have high rates of horizontal transfer in other lineages [172, 199], raising the possibility that these were independent gains.

Studies of vertebrate genome evolution often assume vertical transmission as the dominant mode of evolution. However, recent work has highlighted the impact of horizontal transfer events in the evolution of the vertebrate mobilome [162, 80, 236]. Our ancestral mobilome reconstructions of the DNA Kolobok superfamily are in line with phylogenetic patterns expected by a model of horizontal transfer (Figure 3.3). In our sampling, the presence of Kolobok is limited to five distantly related teleost lineages, as well as the Longnose Gar (*Lepisosteus osseus*) among non-teleost ray-finned fishes. It is possible that the Kolobok element has been repeatedly transferred throughout the evolutionary history of Actinopterygii. Such transfer events are considered more likely by recent observations of horizontal transfer at more recent time scales. For

example, there is evidence of horizontal gene transfer of DNA transposons like Merlin, TcMariner, and PiggyBAC in salmonids [56]. Likewise, there has been evidence of horizontal transfer of Chapaev transposons in White Sturgeon, Pacific Bluefin Tuna (*Thunnus orientalis*) and Blue Catfish (*Ictalurus furcatus*) [235]. As the number of high quality genomes for species of ray finned fishes continue to accumulate, future studies of the extent of horizontal transfer events across the ray-finned fish mobilome offer an exciting research prospect.

3.4.4 *Lineage-specific expansions of the teleost mobilome*

Transposable elements (TEs) likely play a beneficial role by enhancing an organism's ability to respond to dynamic environmental conditions [234]. Numerous studies have linked TE activity to an organism's responsiveness to environmental factors [11, 77, 95, 42]. This association between TEs and environment suggests that there may be a correlation between relative TE abundance and specific aspects of an organism's ecological niche or life history. We tested this expectation using three commonly tested abiotic/biotic factors associated with evolutionary diversification: latitude, body size, and depth. However, we find no such association for several factors often invoked to explain diversification patterns at this evolutionary scale. Under the Brownian motion model, our phylogenetics least squares regression (PGLS) analysis between TE content and maximum body size revealed no evidence for a strong correlation between these traits (**Figure 3.4A** and **Supplemental Tables S4, S5, S6 and S7**; Likewise, we find no statistically supported correlation between TE content and depth occupancy in marine fishes (**Figure 3.4B**). There is also no general pattern of a correlation between TE content and the average latitude of a species geographic distribution (**Figure 3.4C**). The only exception to this lack of correlation between latitude and TE content occurs in SINEs (**Supplemental Table S6**) under the Brownian motion model ($p=0.0197$). This correlation is likely driven by

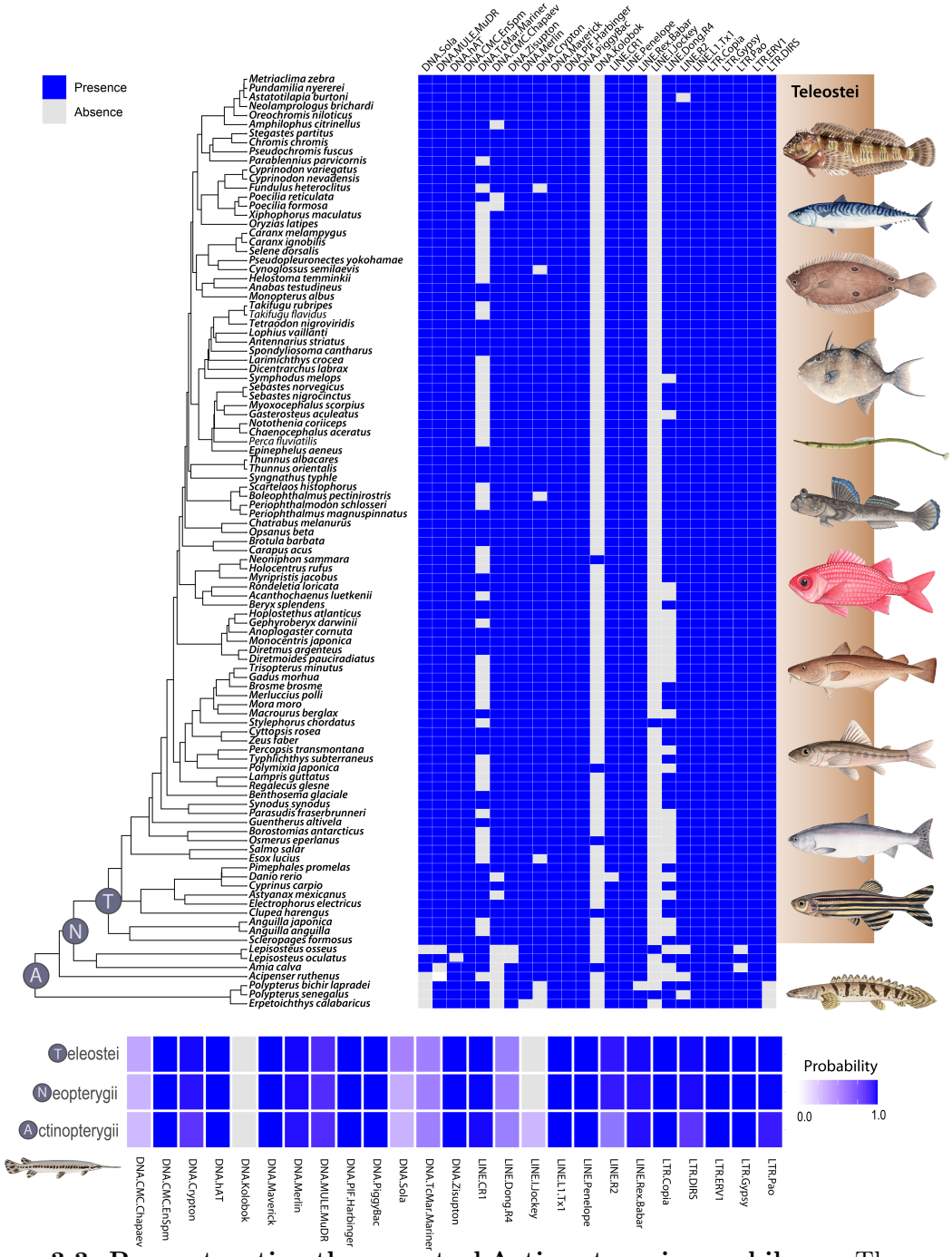


Figure 3.3: Reconstructing the ancestral Actinopterygian mobilome. The presence and absence of transposable element (TE) superfamilies (listed above) are represented by blue and grey squares, respectively, for various species on the left (top panel). The brown-shaded area represents teleosts, while the unshaded region below represents non-teleosts with evolutionary relationships between taxa depicted by the phylogenetic tree on the far-left. This input data was used to estimate the ancestral mobilome (the probability of each TE superfamily being present or absent) for the most recent common ancestor of all Actinopterygii (A), Neopterygii (N), or Teleostei (T) using SIMMAP (bottom panel) With a resulting heatmap plotted using ggplot [222].

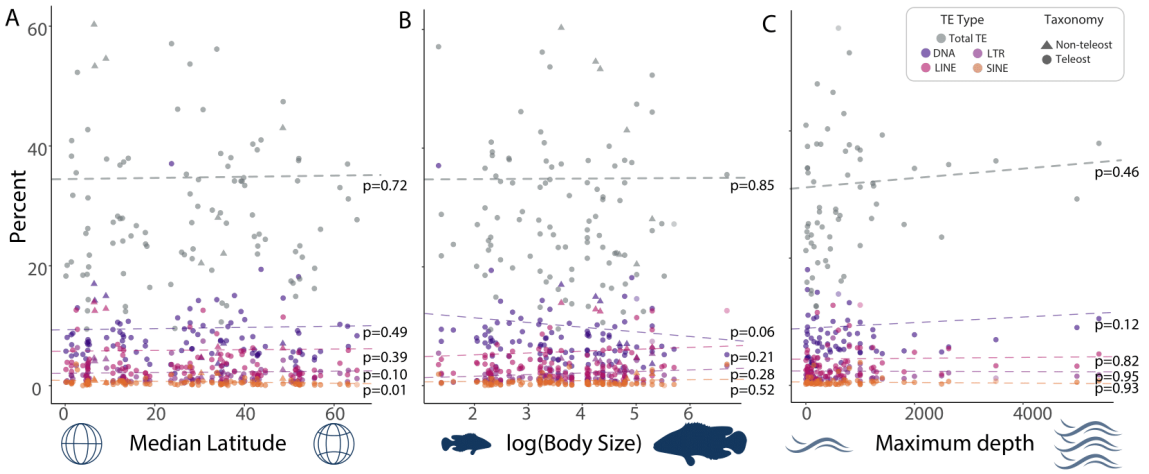


Figure 3.4: Correlation of TEs with biotic and abiotic factors. (A) This panel displays the results of phylogenetic generalized least-squares (PGLS) regressions between TE percentages and median latitude (B) This panel portrays the PGLS regression results between TE percentage and body size (log-transformed). Lastly, (C) depicts PGLS results between TE abundance and the maximum habitat depth of the species. The colors and shapes of each data point on the plot are defined in the key. .

a reduction in SINE elements in Antarctic notothenioids, which experienced an unusual bout of genomic evolution prior to their diversification in the Southern Ocean [53]. Regardless, depth, body size, or latitude appear not to be predictors of mobilome evolution, even when differentiating between marine, freshwater, or estuarine fishes at this evolutionary scale. In all cases, phylogenetic regression results were largely similar between a brownian and Orstein-Uhlenbeck model of trait evolution (**Supplemental Table S6 & S7**).

The TGD certainly could have presented opportunities for a burst of genomic novelty. However, the next 300 million years of ray-finned fish evolution would certainly be expected to shape the genomes of different lineages responding to different biotic and abiotic conditions. As such, it is possible that the signal of the genome duplication has either eroded, or that. For example, TEs could replace themselves in a way analogous to multigene families, in which recently diverging paralogs often replace older genes. In such a scenario, rapid rates of mobilome evolution manifest as

high between-clade heterogeneity at the scale of all ray-finned fishes. Clade-specific changes in the mobilome are readily apparent when considering a phylogenetic principal components analysis (PCA) of mobilome abundances (**Figure 3.5**). For example, PC1 corresponds to sharp divergence between *Esox*, salmonids, and three bichirs, from acanthomorphs **(Figure 3.5** and **Supplementary Figures S3.4, S3.5, S3.6**). This alignment of taxa with PC1 values coincides with the prevalence of high percentage of DNA transposons that comprise the transposable elements within fish genomes that influence genome size variation among teleosts such as zebrafish (**Supplemental Figure S3.3**), medaka, stickleback, and *Tetraodon* [101]. It is certainly not possible to discount a possible role for life history shifts shaping the mobilome among closely related species or that differences between Class 1 and Class II replications serve as an opportunity for substantial expansion of the mobilome.

3.5 Conclusion

We find no evidence for a discernible temporal signature of TE diversification coincident with the TGD, raising the possibility that this genome duplication event did not exert a long-lasting influence on transposable element diversification. The absence of a discernible temporal signature of diversification coincident with the TGD raises important questions about the long-term effect of ploidy on TE diversification. First, our results suggest that the ray-finned fish mobilome appears to maintain a consistent profile, with substantial similarity between the mobilomes of teleosts and non-teleost ray-finned fishes. However, mobilome content between taxa varied substantially, from around 55% in Zebrafish to about 6% in pufferfish genomes. These shifts cannot be entirely explained by differences in genome size as genome compaction in smooth pufferfish like *T. rubripes* and *Tetraodon nigroviridis* is not associated with an overall reduction of retrotransposon diversity [215]. Similarly, Medaka and Zebrafish have been found to have very similar L1 retrotransposon diversity, despite a large dispar-

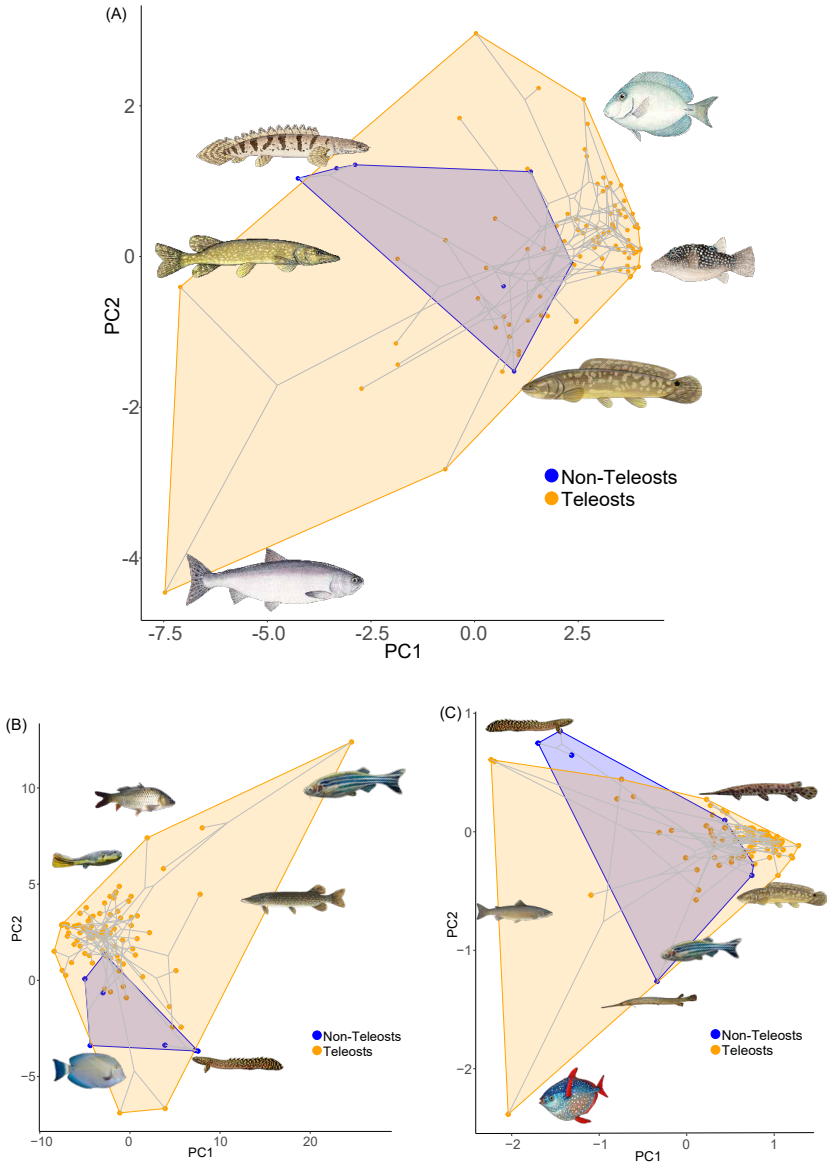


Figure 3.5: Considering Transposable Element (TE) diversity in the context of evolutionary history. (A) Results of a Phylogenetic PCA on the abundances of LTRs, DNA transposons, LINEs, and SINEs that project the phylogeny and points onto a 2D plot of PC1 and PC2. (B) depicts the resulting space based on the residual variation following linear regression between major TE classes and genome size. (C) Depicts the results of a PCA on the residual variation of the abundances of all 25 superfamilies accounting for differences in genome size, consisting of 13 DNA transposons, 7 LINEs, and 5 LTR subfamilies. Blue indicates non-teleost actinopterygians, orange indicates teleosts. All the images of the fishes are from Wikimedia commons.

ity in genome size [109]. Instead, these results strongly imply that the history of the ray-finned fish mobilome is one of lineage-specific evolution. In this way, the diversification of TEs may mirror the general dynamics of gene-birth found in multi-gene families in which there is consistent turnover of paralogs through time [161]. If true, then the impact of ploidy events may be short-lived at best, and shifts in TEs may be more aligned with the concepts of hopeful monsters over deeper evolutionary scales.

Our analyses provide the first global overview of the actinopterygian mobilome, supporting the observation that early diverging aquatic vertebrates (e.g., coelacanth, ray finned fishes, cartilaginous fish, and lamprey) contain a significantly broader spectrum of mobilome diversity in the genome relative to birds or mammals, ranging from 22 to 27 TE superfamilies. Among these, certain autonomous elements like ERVs, LINE1 retrotransposons, TcMariner, or hAT DNA transposons, as well as non-autonomous ones like V-SINE [164], are widely distributed across all the vertebrates. This suggests their likely presence in ancestral jawed vertebrate genomes. In contrast, we see a lower distribution of endoretroviruses (ERVs) in actinopterygians than mammals and birds, with the amount ranging from 0.033% in Fugu to 0.76% in zebrafish [43]. Our findings of clade-specific heterogeneity within ray-finned fishes, suggests that future studies between more closely related clades to be particularly fruitful as genomic resources become available. In particular, such studies could reveal superfamilies that have been lost or are on the path to extinction in specific lineages, similar to studies revealing such findings for L2 and Helitrons in birds [49] and gypsy retrotransposons in birds and mammals [215]. It is clear that we are only beginning to unmask the complexity of transposon dynamics and decipher the intricate processes that contribute to TE diversification over deep evolutionary timescales.

3.6 Materials and Methods

3.6.1 *Sample acquisition and sequencing*

All research involving live animals was performed in accordance with relevant institutional and national guidelines and regulations, and was approved by the North Carolina State University Institutional Animal Care and Use Committee (IACUC). We acquired an individual *Polypterus bichir* specimen (Yoder lab ID 0051) through the pet trade that was anesthetized using MS-222 for the extraction of blood (10 ml). The fish was then euthanized for dissection of tissue samples and the voucher specimen was deposited in the North Carolina Museum of Natural Sciences Ichthyology Collection (NCSM 111902). Blood was shipped to Dovetail Genomics, LCC (Scotts Valley, CA) for genomic DNA extraction, library preparation, sequencing, and assembly. Samples were extracted by Dovetail staff using Qiagen and Cell Culture Midi Kit (Qiagen, Gmbh), yielding DNA with an average fragment length of 95 kbp that was used in the construction of HiC sequencing libraries.

3.6.2 *Chicago Library preparation and sequencing*

A Chicago library was prepared using the methods as described in [165]. About 500 ng of High Molecular Weight genomic DNA, with mean fragment length of 95 kbp. Chromatin was reconstituted in vitro by incorporating the DNA with purified histones and chromatin assembly factors and then fixed by formaldehyde. DpnII was used to digest the chromatin, followed by filling in the 5' overhangs with biotinylated nucleotides, and then ligating the free blunt ends. After the ligations step, DNA was purified from protein by reversing the crosslinks. Biotinylated free ends were removed from the purified DNA by treating it. The DNA was sheared to ∼350 bp fragment size and Illumina-compatible adapters along with NERNext Ultra enzymes were used to generate sequencing libraries. Streptavidin beads were used ahead of PCR enrich-

ment of each library to isolate the biotin-containing fragments. The libraries were sequenced on an Illumina HiSeqX providing 30.58x physical coverage of the genome (1-100 kbp).

3.6.3 *Dovetail Hi-C library preparation and sequencing*

The preparation of a dovetail Hi-C library was executed as described [117]. For each library, the chromatin was fixed in place in the nucleus by crosslinking with formaldehyde, and the extracted fixed chromatin was then digested with the restriction enzyme DpnII that produces 5' overhangs. These 5' overhangs were filled in with biotinylated nucleotides, and then the resulting blunt ends were ligated. Crosslinks were then reversed to obtain the purified DNA, and which was treated to remove excess biotin. A Hi-C library was then created by shearing the DNA to ∼350 bp mean fragment size. Sequencing libraries were generated using NEBNext Ultra enzymes and Illumina compatible adapters. Streptavidin beads were used prior to PCR enrichment of each library to isolate the biotin containing fragments. Libraries were sequenced using an Illumina HiSeqX which yielded 191 million paired end reads (2 × 150 bp) and provided 8,427.66 x physical coverage of the genome (10-10,000 kbp).

3.6.4 *Scaffolding the assembly with Hi-Rise*

Sequencing reads were assembled with Hi-Rise, a software pipeline designed to scaffold genome assemblies using proximity ligation data [165]. It uses the *de novo* assembly, shotgun reads, Chicago library reads and Dovetail HiC library reads as input and conducts an iterative analysis. The shotgun and Chicago library sequences are first aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu), with modifications such as masking out the base pairs that followed a restriction enzyme junction. Hi-Rise then modeled the separations

of Chicago read pairs mapped within draft scaffolds, using a likelihood model for genomic distance between read pairs. The likelihood model produced by HiRise was used to identify and break putative misjoins and also score and make prospective joins. Dovetail HiC library sequences were analyzed using the same methods after aligning and scaffolding the Chicago data. Once all the sequences were aligned and scaffolded, shotgun sequences were used to close the gaps between contigs.

3.6.5 *Contamination removal and species verification*

Dovetail staff have noted that when pooled with other samples on Illumina sequencing platforms, 10X Chromium Genome solution libraries are susceptible to a small degree of index hopping that can result in minor incorrect assignment of sequenced reads during demultiplexing. This low level of index misassignment, typically results in sequence contaminants impacting, but limited to, small scaffolds (typically less than 10 kb) in the final assembly. To mitigate this, dovetail staff leveraged any uncharacteristic number of reads per barcode associated with impacted scaffolds to identify and reliably isolate them from the final assembly. This was accomplished by aligning the 10X reads to the supernova assembly, recording the barcode count for the aligned reads, and recording the number of reads that aligned to each scaffold and were tagged with the same barcode. The median number of reads per barcode for each scaffold were then calculated and scaffolds with a distinct anomalously low ratio were removed.

Currently *Polypterus bichir* is described as a single species with the subspecies *Polypterus bichir bichir* and *Polypterus bichir lapradei* no longer considered valid [134]. This sinking of subspecies is a consequence of the high degree of morphological similarity [30, 29]. However, molecular investigations have suggested the possibility that *P. bichir bichir* and *P. bichir lapradei* may be genetically distinct [197]

and these have been treated as independent lineages [144]. As no genetic species delimitation analyses have been conducted between these putative genetic lineages, we extracted the mitochondrial barcode COI from our assembly and used a BLAST search against polypterid sequences on NCBI to verify identification that included barcodes from *Polypterus bichir bichir* and *Polypterus bichir lapradei.* This ensured that we accounted for possible future taxonomic revisions while remaining consistent with current taxonomy.

3.6.6 *RNA sequencing and assembly*

RNA was extracted (Qiagen RNeasy kit) from the spleen, kidney, gill, heart, eyes and intestine of the same individual *P. bichir* as the genome sequence. Quantity and integrity of the isolated RNA were assessed using a NanoDrop 1000 (Thermo Fisher) and Agilent Bioanalyzer respectively. The process of mRNA enrichment was done using oligo(dT) beads, and rRNA was removed using a Ribo-Zero kit (Epicentre, Madison, WI). Each RNA extraction was equalized for a final concentration of 180 ng/μL. Library preparation and sequencing was performed by Novogen Corporation (Sacramento, CA). Next-gen sequencing (2×150 bp paired end reads) was performed on a NovaSeq 6000 instrument (Illumina). Adapter sequences and poor quality reads were filtered with Trimmomatic v34 [26]. The transcriptome was de novo assembled with Trinity v2.11.0 [84] followed by BUSCO analysis to assess completeness of the transcriptomes [124]. Raw reads and computationally assembled transcriptome sequences were deposited onto NCBI under the accession numbers SRR19537224 - SRR19537230 and GKOV01000000 respectively.

3.6.7 *Gene ontology assessment*

We conducted a series of analyses for gene ontology assignment. First, the RNA-seq data assembled using Trinity was translated with Transdecoder, enabling the identification of potential coding regions in the transcript sequences. The longest open

reading frames (ORFs) were extracted and subjected to BLASTx and BLASTp analyses against the Uniprot database (November 2021 release), yielding the top target sequences for each transcript. Subsequently, Hmmscan v.3.3.2 was employed to search for protein sequences in the Pfam-A database (November 2021 release). Signalp v.5.0b and TMHMM v.2.0c were used to detect signal peptides and transmembrane proteins, respectively. Trinotate v.3.2.2, in combination with the Trinity assembled transcriptome and the longest ORFs from Transdecoder, generated a gene transmap. This process facilitated the annotation of transcripts using Trinotate, followed by further analysis to obtain the Gene Ontology (GO) annotations. For visualization of the GO terms, the 'enrichplot' and 'ggupset' packages in R were employed. These steps collectively provided a comprehensive understanding of the functional attributes of the transcriptome data, to understand the underlying biological processes and pathways associated with the studied organisms (Supplemental Figures S3.7-S3.27).

3.6.8 *Annotation and genome quality assessments*

The *P. bichir* genome annotation was done using the commonly used BRAKER2 [34]. To accomplish this, we first used RepeatModeler (Version 5.8.8) [76] to model the repeats in the genome sequence. We then used RepeatMasker [1] to mask the repeats found with RepeatModeler and remove them from the genome. The masked genome and the transcriptome aligned using HISAT2 (version 2.1.0) [107] were then used to annotate the genome using the Genemark-ET option in BRAKER. The genemark.gtf file was subsequently used by Augustus [79] to model the proteins.

The completeness of the protein sequences was assessed using BUSCO v. (5.5.0) [189, 182] with both the Actinopterygii (actinopterygii_odb10, busco.ezlab.org) and vertebrate (vertebrata_odb10, busco.ezlab.org) databases. As polypterids are several hundred million years divergent from all other actinopterygians [65, 144], the

use of the second vertebrate-wide database allowed us to verify similar levels of assembly completeness and mitigate against a potentially teleost-biased ray-finned fish database that may not capture loci in a deeply divergent taxon. We additionally conducted an analysis of synteny between our assembly of the *P. bichir* genome and the previously sequenced *P. senegalus* and *E. calabaricus* genomes using D-Genies [40]. To estimate the relative ages of transposable elements within the genome, we generated a Kimura distance plot using Repeatmasker with the -a parameter to get the alignment file (.align). This was analyzed using 'calcDivergenceFromAlign.pl' to get the divergence summary divsum file (.divsum), which was further analyzed using 'createRepeatLandscape.pl' to create the comprehensive repeat landscape.

3.6.9 *Comparative analyses of the Actinopterygian mobilome*

To assemble a dataset of TE content across major ray-finned fish lineages, we integrated the results of the repeat analysis of the *P. bichir* genome with other publicly available analyses of transposable elements in ray-finned fishes. For non-teleosts, this captured TE content from two additional polypteriform genomes for *E. calabaricus* [89] and *P. senegalus* [78], Acipenser ruthenus [69] as a representative of Chondrostei, as well as *Lepisosteus oculatus* [28], *L. osseus* [122], and Amia calva [202] as representative holosteans. These data were integrated with data from 98 teleost genomes previously analyzed for TE content [171], that capture the majority of major teleost lineages including Elopomorpha, Osteoglossomorpha, Otocephala, and a large number of acanthomorph and non-acanthomorph euteleosts. We generated a consensus sequence file using RepeatMasker to obtain detailed information TEs that were extracted from the resulting .out and .tbl files. This yielded a dataset of TE content for 105 ray-finned fish genomes. To place this data into a comparative phylogenetic framework, we first obtained a time calibrated phylogeny of all taxa from TimeTree v5 [112]. As this tree lacked resolution for acanthomorph lineages, we modified

branch lengths to reflect the age estimates from a recent analysis of acanthomorph divergence times based on ultraconserved elements [82] that is consistent with other published divergence time estimates of this superradiation [142, 96]. We additionally modified the topology to reflect the proposed sister relationship between Osteoglossomorpha and Elopomorpha based on genomic and transcriptomic sequence analyses [96, 220, 213, 87]

We used the ggtree and ggtreeExtra packages [232] in R 4.3.1 to visualize the distribution of TE abundances (LTR, LINE, SINE, DNA) across the evolutionary history of actinopterygians. We additionally used ggtree in conjunction with phytools [173] to visualize genome size and TE content variation between species alongside a likelihood based ancestral state estimation of changes in genome size across the phylogeny conducted in phytools. To assess if changes in genome size were correlated with the changes in the overall abundance of TEs, or changes in the abundances of LTRs, LINEs, SINEs, or DNA elements, we conducted a series of phylogenetic linear regressions using the phylolm package in R. Model fits were assessed by quantification of Akaike Information Criterion (AIC) scores (**Supplemental Table S4**) Regressions were conducted under a Brownian model of trait evolution (**Supplemental Table S6**). To assess whether results were robust to the underlying model of character evolution, analyses were repeated using an Ornstein-Uhlenbeck (OU) model of character evolution (**Supplemental Table S7**). A similar set of phylogenetic regressions was next conducted to assess if increases in the abundance of elements (e.g., LINEs, SINEs, etc) were positively correlated with each other, or if negative correlations exist, allowing us to assess whether elements have antagonistic evolutionary dynamics (Supplemental Table S8). Next, we performed a set of regressions assessing whether changes in the abundances of elements were correlated with changes in maximum body size, latitude, or maximum depth of occurrence. Body size and depth

data for each species were taken from fishbase using the rfishbase package in R [21]). Latitudinal data was calculated using occurrence data from the Global Biodiversity Information Facility (gbif) using rgbif v3.7.8. We further tested for differences in TE content between teleosts and non-teleosts using a phylogenetic ANOVA in phytools phylANOVA with multiple comparison correction using the Benjamini-Hochberg procedure [16].

To reconstruct the ancestral mobilome of early ray-finned fishes, we focused on the 25 elements most commonly studied in analyses of ray-finned fish TEs. Presence/absence data was scored for each species and each element and used as input data for a model averaged stochastic character mapping approach [27] implemented in phytools. This approach expands the standard phytools implementation to ancestral state estimation by allowing possible models of character change (e.g., “Equal Rates", “All Rates Different") in phytools to contribute to the reconstruction in relation to their Akaike weight. We assessed the degree to which variation in TE content could be explained by evolutionary history through quantification of the phylogenetic signal of overall TE abundance as well as the abundances of each TE type. This was accomplished using the phylosig function in phytools to calculate both Pagel's λ [156] and Blomberg's K [20] Values of λ are distributed between zero and 1, with a zero value representing the absence of phylogenetic signal and a value of 1 corresponding to the expectations of Brownian motion on a phylogeny. To assess statistical significance, we compared our empirical λ values to the null expectation that $\lambda = 0$ for each trait via a likelihood ratio test. Blomberg's K compliments estimates of λ , with values less than 1 indicating less phylogenetic signal than would be expected given a model of Brownian motion, and value greater than 1 indicating a higher than expected coupling between the distribution of traits and the underlying phylogeny. To assess statistical significance of K we compared our empirical K values to null

distributions of expected K values based on 10,000 permutations of each trait on the phylogeny.

To visualize the major axes of variation of the ray-finned fish mobilome, we conducted a phylogenetic principal component analysis (pPCA) using the phyl.pca function in phytools. The resulting PC axes were then used with the phylomorphospace function in phytools to project the phylogeny into the resulting PCA space using ggplot2. Additionally, we conducted a pPCA on the abundances of the subtypes of all TEs based on the amounts derived from the .out files. As preliminary analyses indicated a correlation between elements and overall genome size, pPCAs were repeated on the residuals resulting from a regression of genome size vs TE content, mirroring similar approaches to accounting for traits that covary with another trait (e.g., limb proportions and body size, etc).

3.7 Data Availability

The *P. bichir* genome sequence is available through NCBI Bioproject PRJNA811142. Raw transcriptome reads are available through NCBI under the accession numbers SRR19537224, SRR19537225, SRR19537226, SRR19537227, SRR19537228, SRR19537229, SRR19537230. Computationally assembled transcriptome sequences are available on NCBI under the accession number GKOV000000000. All the files and code used for TE analysis and visualization are available on Zenodo (DOI: 10.5281/zenodo.10398557).

3.8 Acknowledgements

We thank Kent Passingham (NC State University) for assistance with blood collection and the NCBI staff for help with sequence contamination identification. We thank Katerina Zapfe for her invaluable input on figure refinement, and Brandon

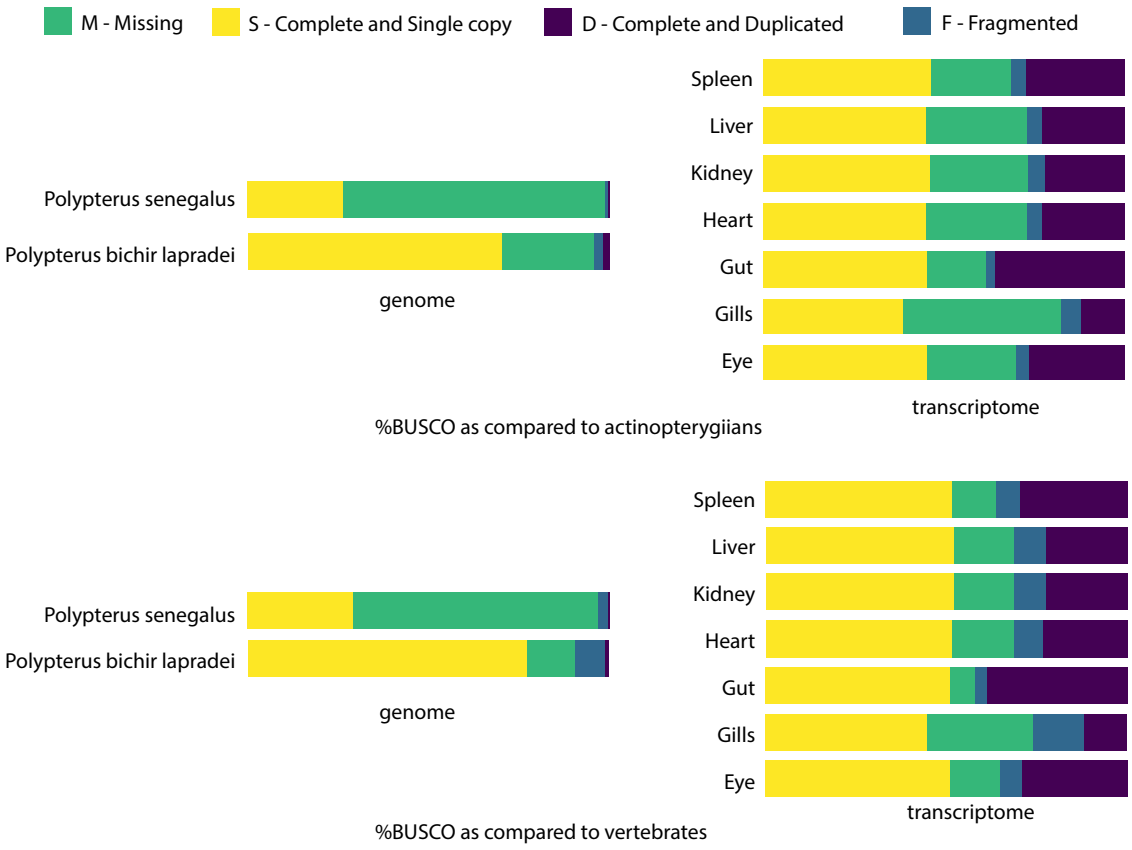
Turner for helpful insights on transposable elements.

3.9 Funding

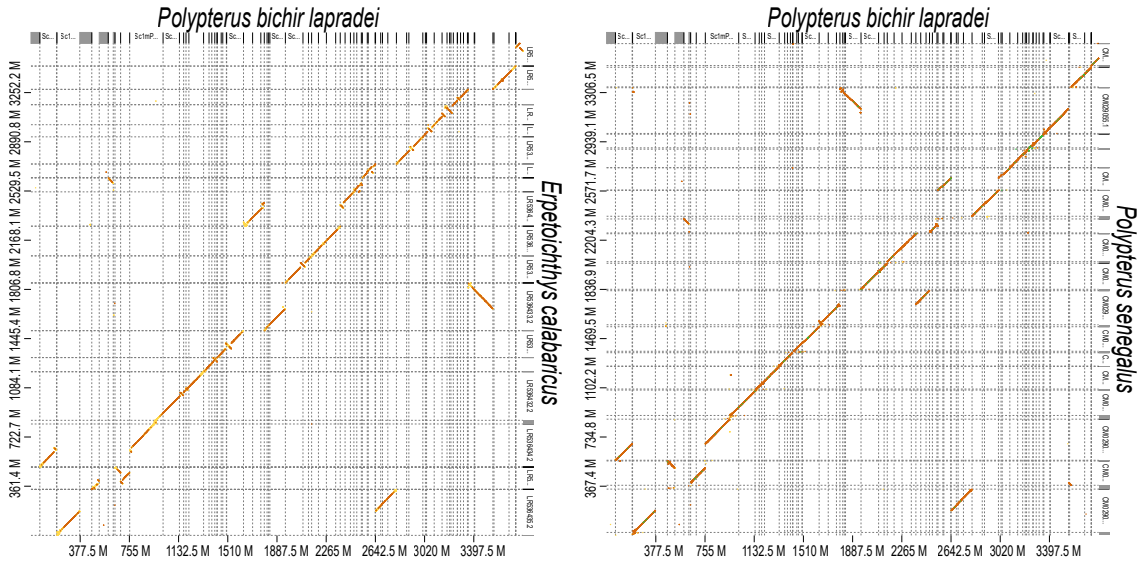
This work was supported by the National Science Foundation (IOS1755242 to AD and IOS1755330 to JAY).

3.10 Supplemental Tables and Figures

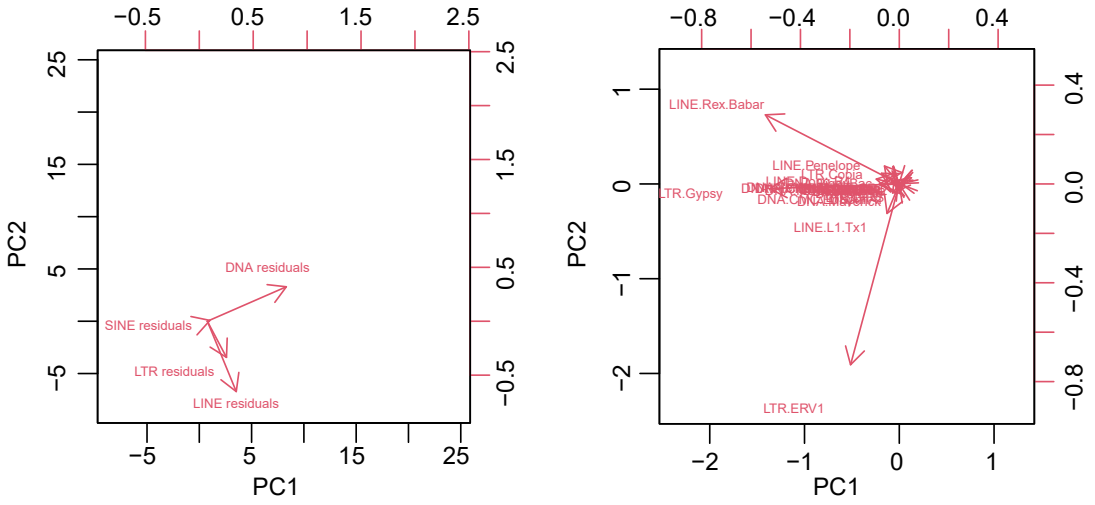
3.10.1 Supplemental Figures



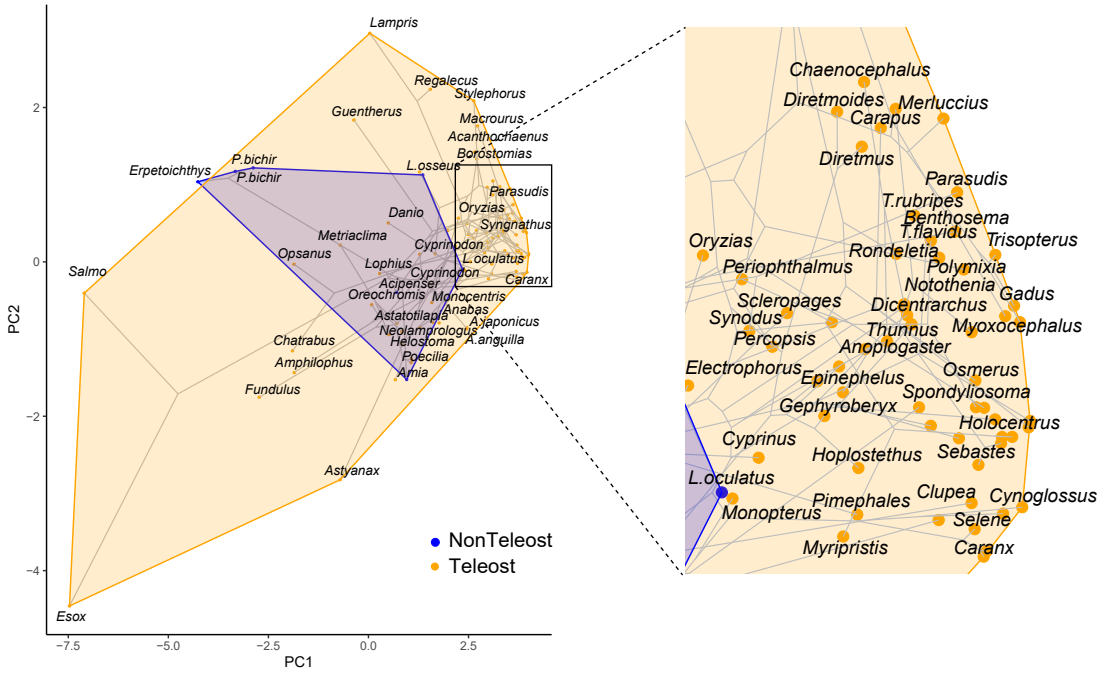
Supplemental Figure S3.1: BUSCO analysis comparing the Polypterus bichir and Polypterus senegalus reference genomes and our transcripts for Polypterus bichir from various tissues. BUSCO scores for the Polypterus bichir reference genome were higher than those of the Polypterus senegalus reference , indicating the higher completeness and quality of the Polypterus bichir genome assembly. The upper panel depicts the loci captured when using the actinopterygian BUSCO reference set, the lower panel indicates the loci captured when utilizing the vertebrate reference set.



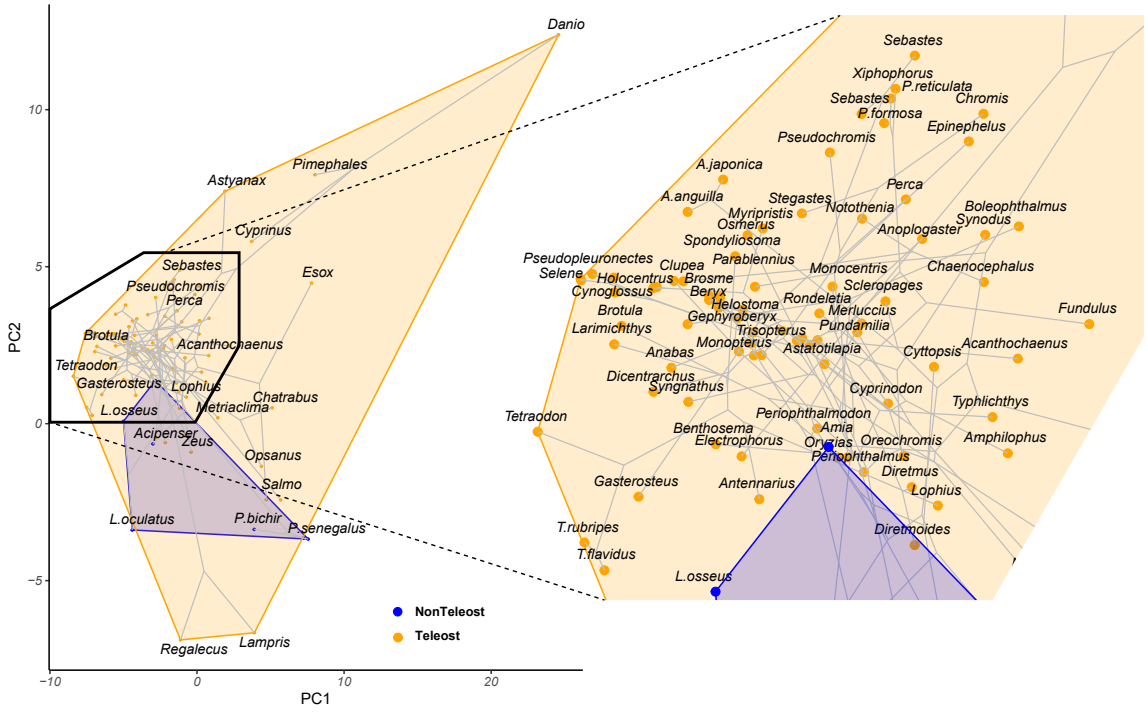
Supplemental Figure S3.2: D-genies plot showing synteny between the Polypterus bichir and Polypterus senegalus reference genomes (left) and between the Polypterus bichir and Erpetoichthys calabaricus reference genomes (right).



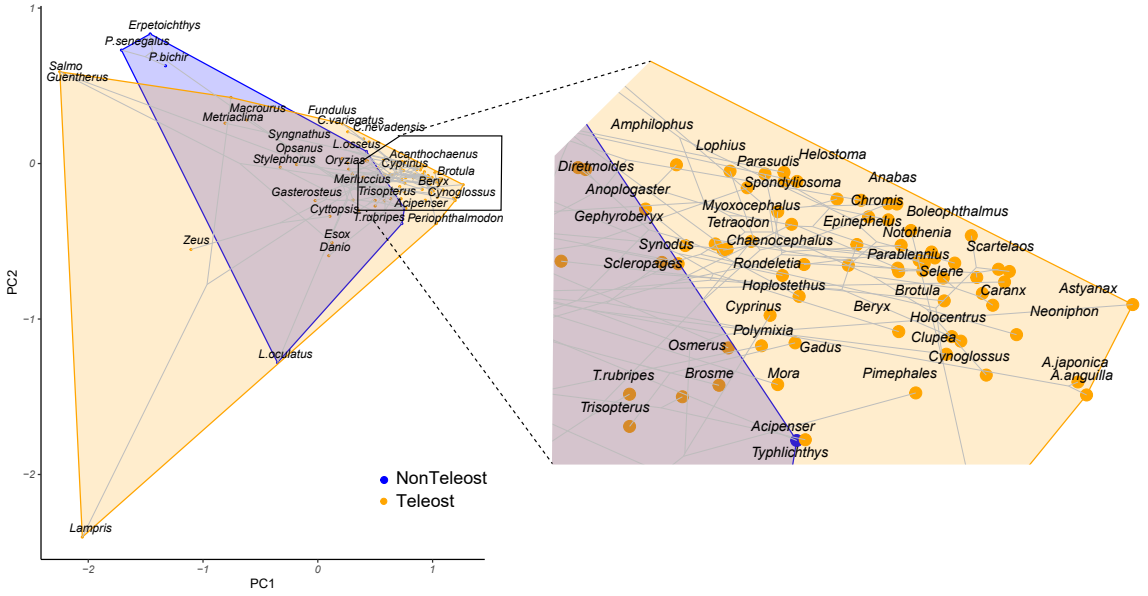
Supplemental Figure S3.3: Biplots depicting the loading of the variables in the phylogenetic PCA spaces depicted in Figure 5



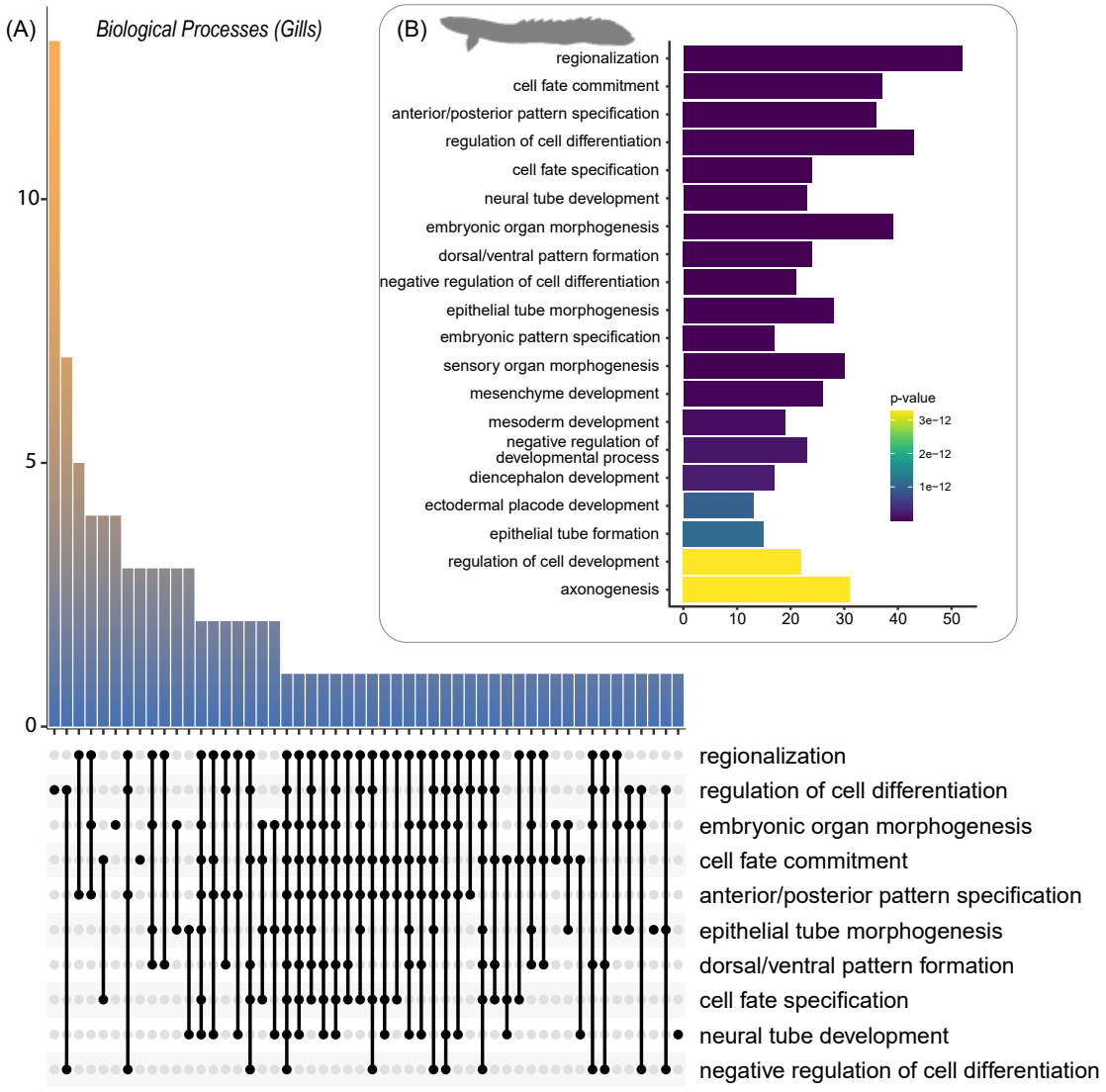
Supplemental Figure S3.4: Labeled plots of results of a phylogenetic PCA on the abundances of LTRs, DNA transposons, LINEs, and SINEs that project the phylogeny and points onto a 2D plot of PC1 and PC2. Blue indicates non-teleost actinopterygians, orange indicates teleosts.



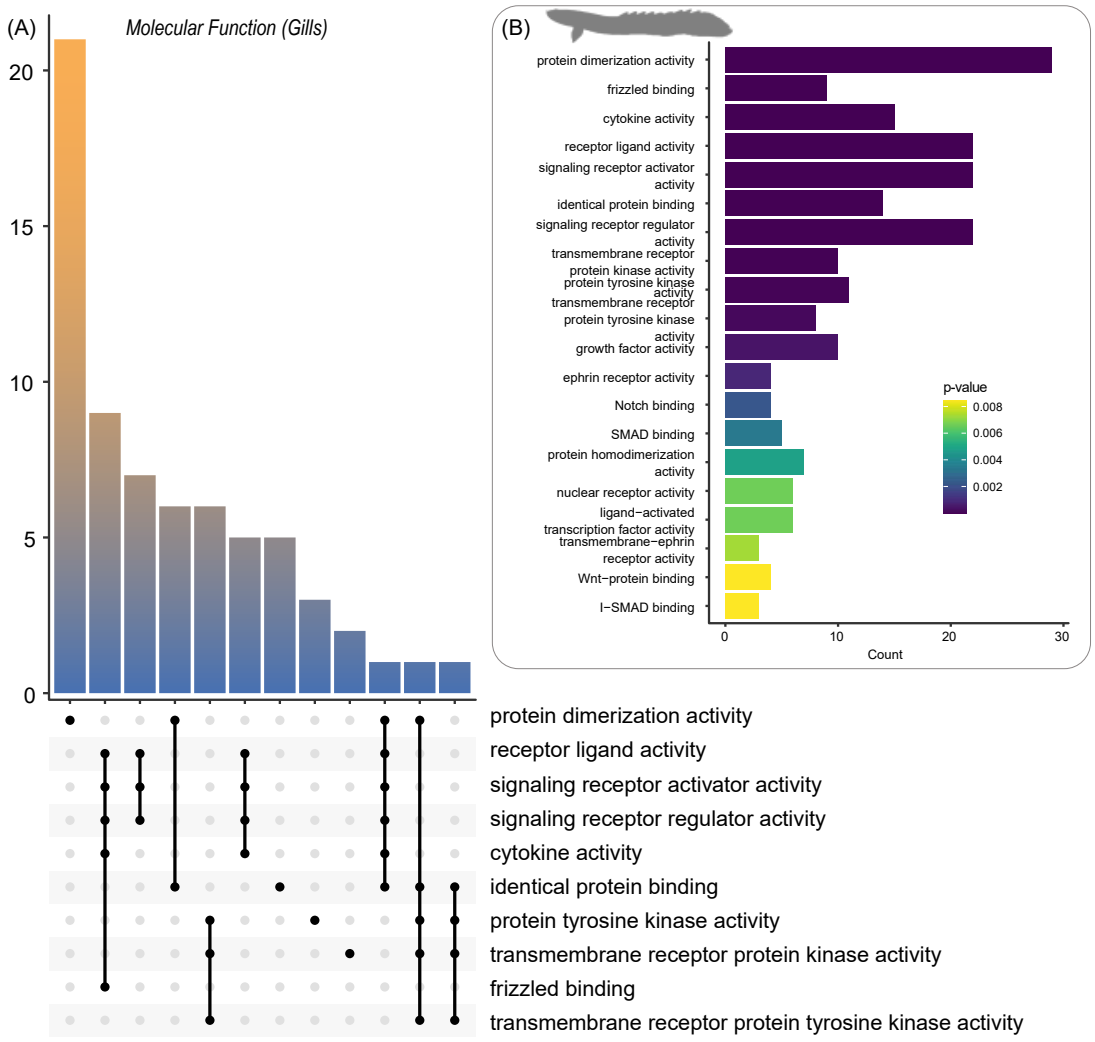
Supplemental Figure S3.5: Labeled plots of phylomorphospace depicting the resulting space based on the residual variation following linear regression between major TE classes and genome size. Blue indicates non-teleost actinopterygians, orange indicates teleosts



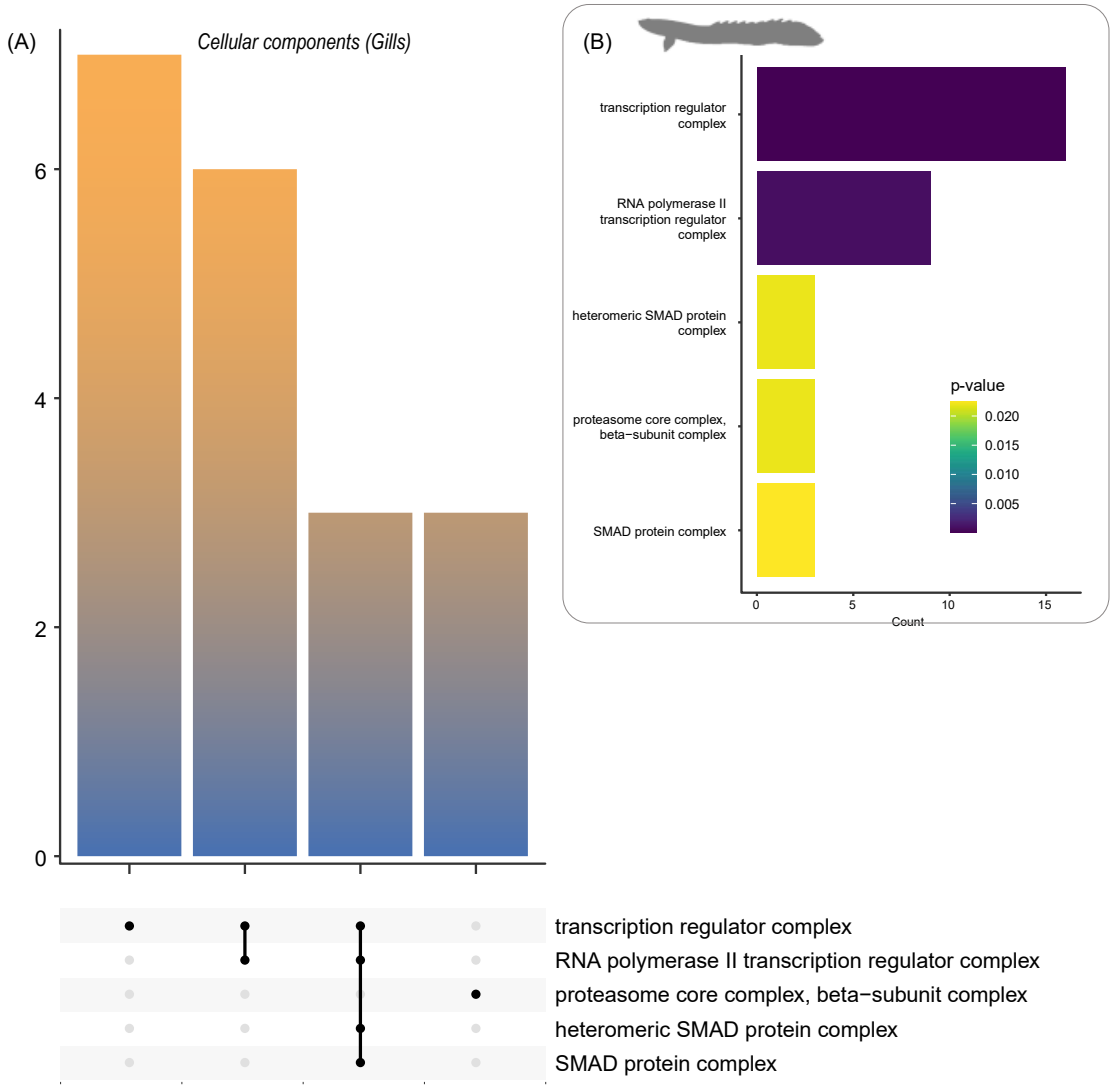
Supplemental Figure S3.6: : Labeled plots depicting the results of a PCA on the residual variation of the abundances of all 25 superfamilies accounting for differences in genome size, consisting of 13 DNA transposons, 7 LINEs, and 5 LTR subfamilies. Blue indicates non-teleost actinopterygians, orange indicates teleosts.



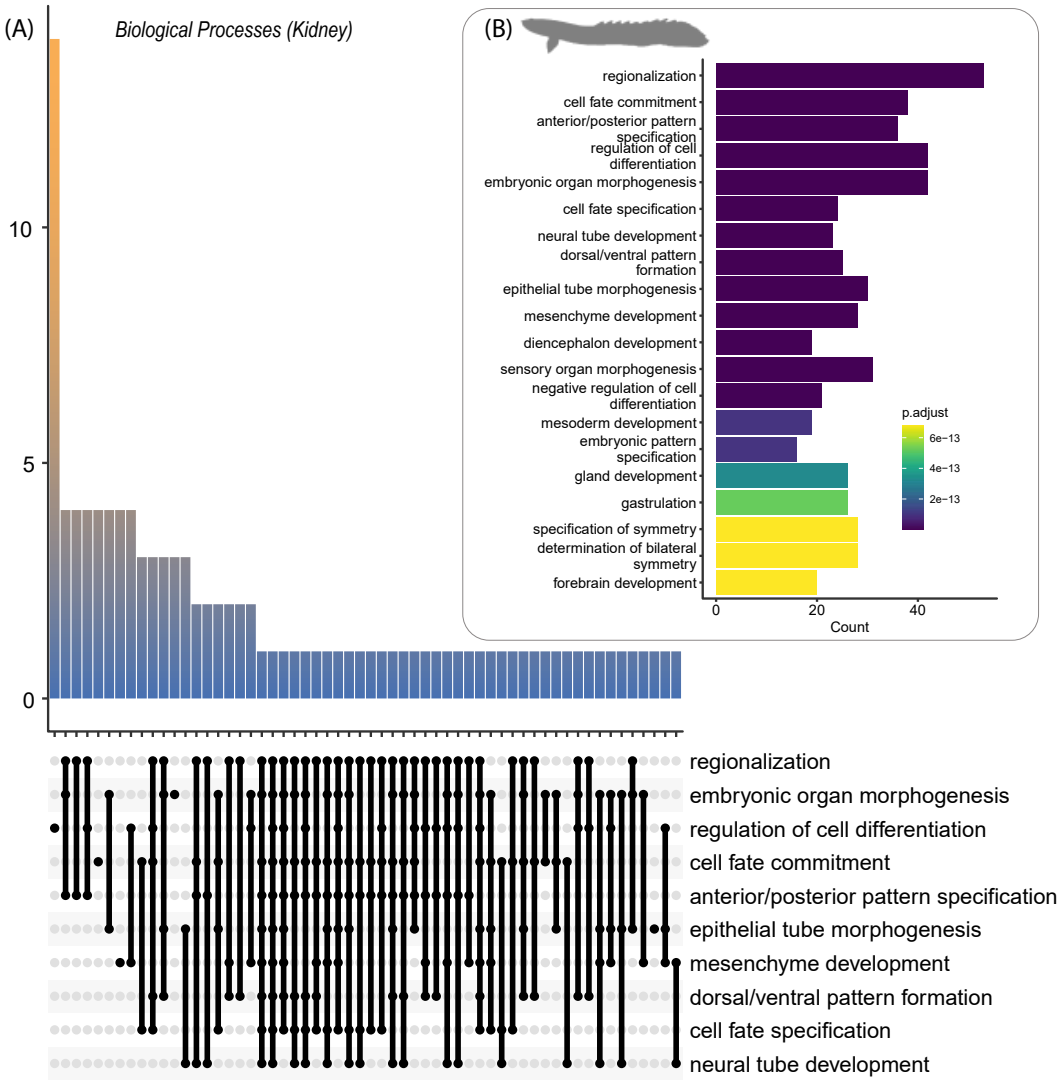
Supplemental Figure S3.7: Summary of Gene Ontology *Biological Processes* analysis from the *Polypterus bichir* gill transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



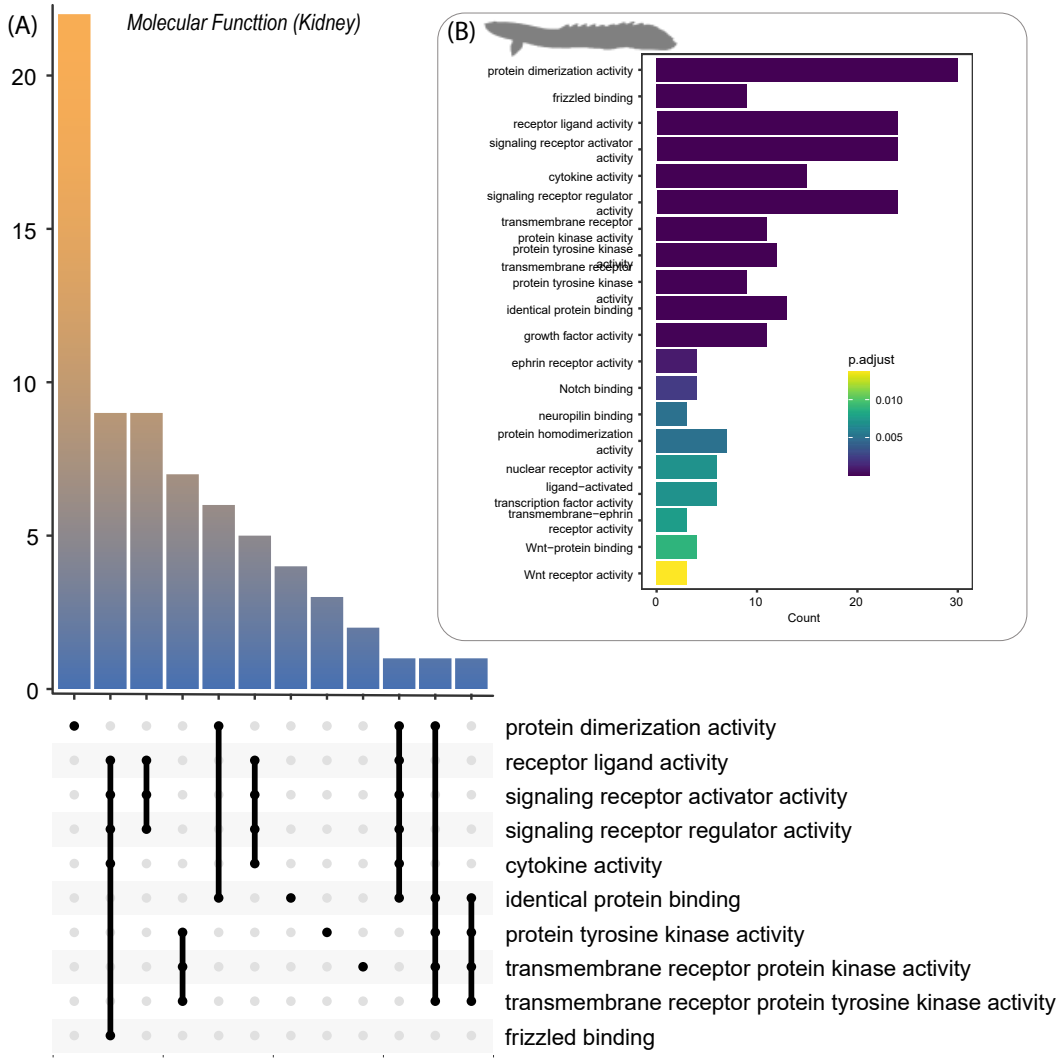
Supplemental Figure S3.8: Summary of Gene Ontology *Molecular Function* analysis from the *Polypterus bichir* gill transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



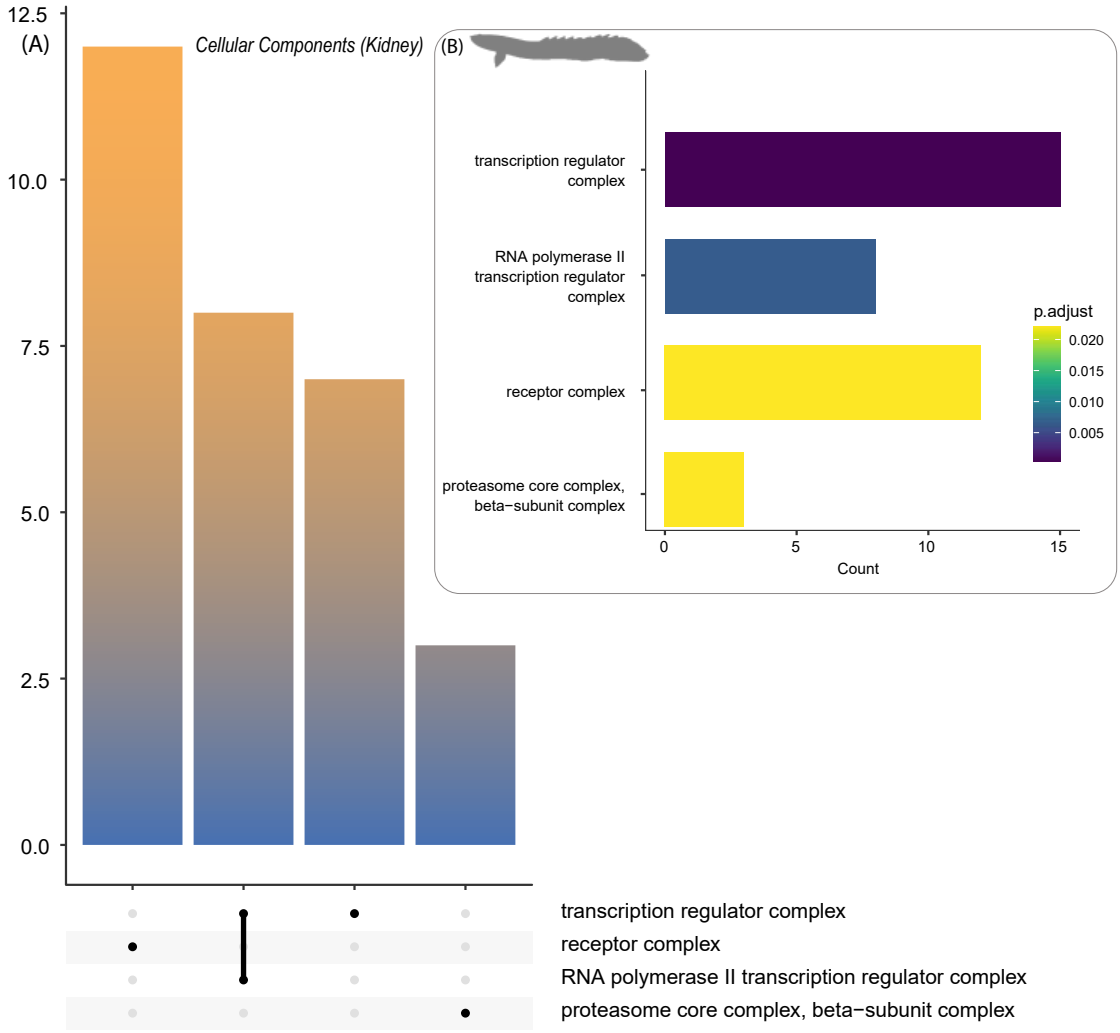
Supplemental Figure S3.9: Summary of Gene Ontology *Cellular Components* analysis from the *Polypterus bichir* gill transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



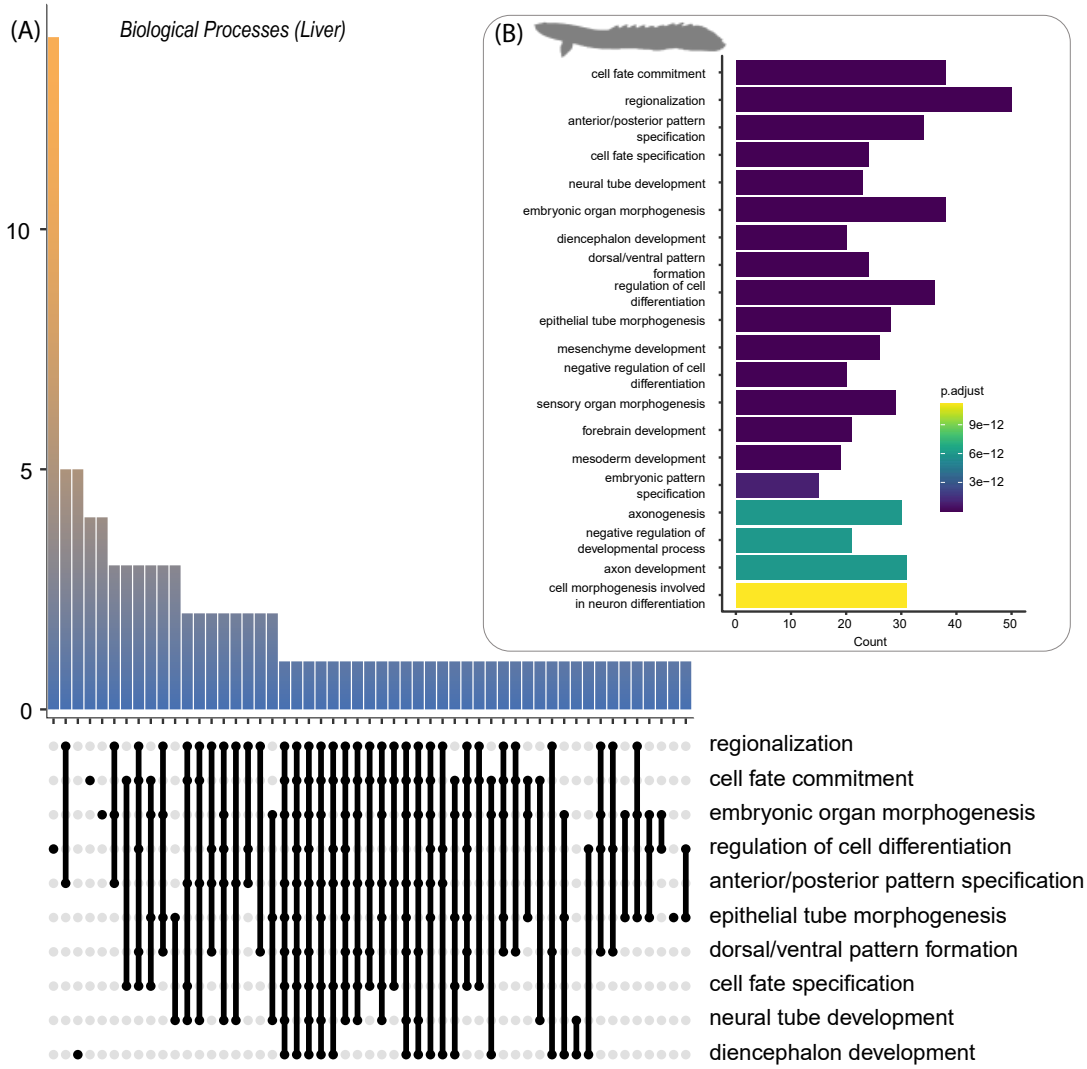
Supplemental Figure S3.10: Summary of Gene Ontology *Biological Processes* analysis from the *Polypterus bichir* kidney transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



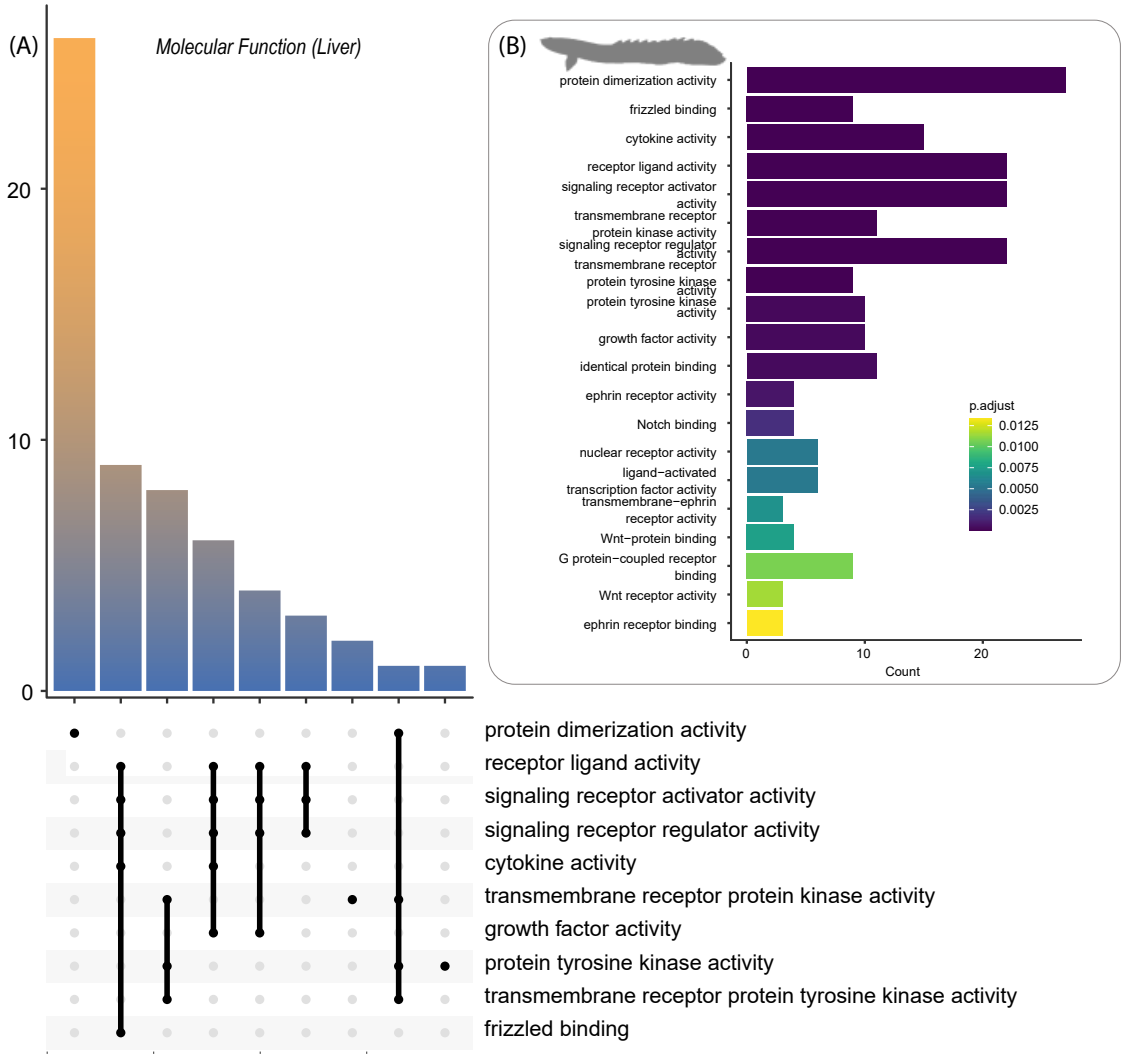
Supplemental Figure S3.11: Summary of Gene Ontology *Molecular Function* analysis from the *Polypterus bichir* kidney transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



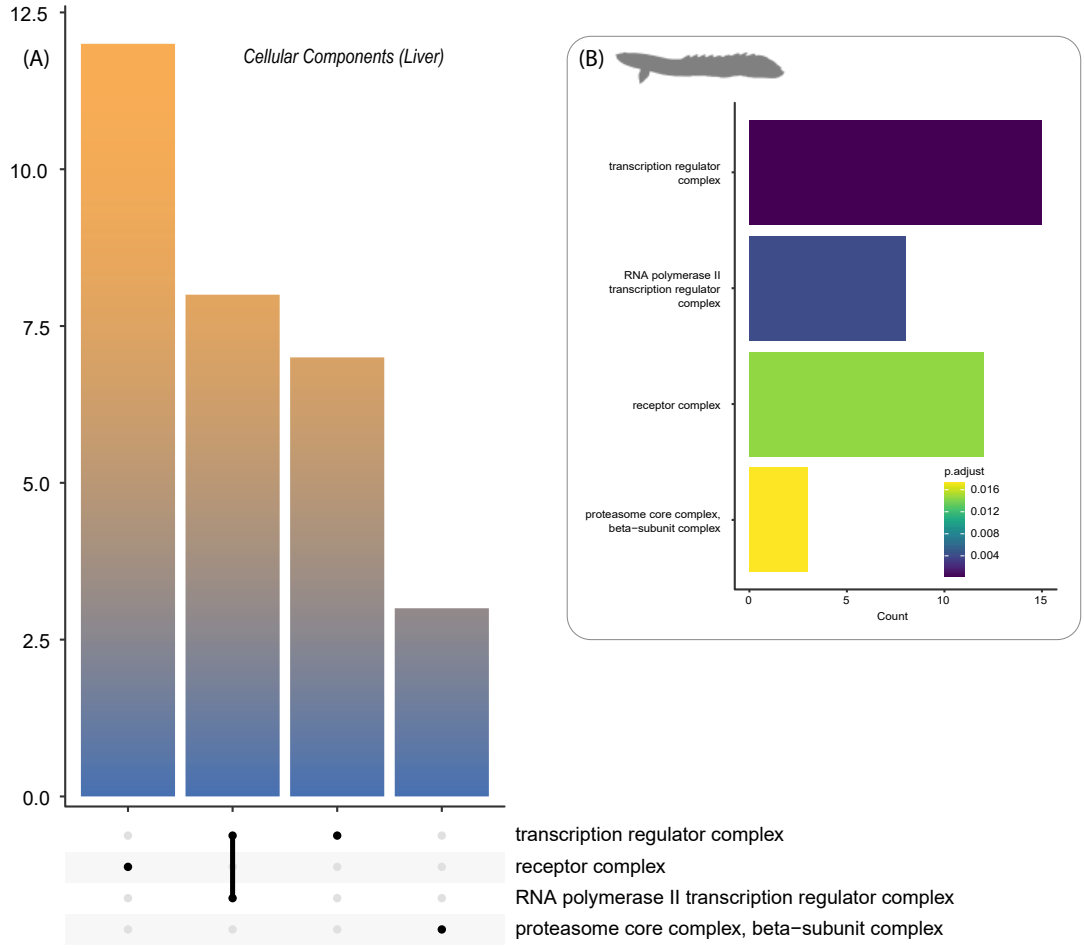
Supplemental Figure S3.12: Summary of Gene Ontology *Cellular Components* analysis from the *Polypterus bichir* kidney transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



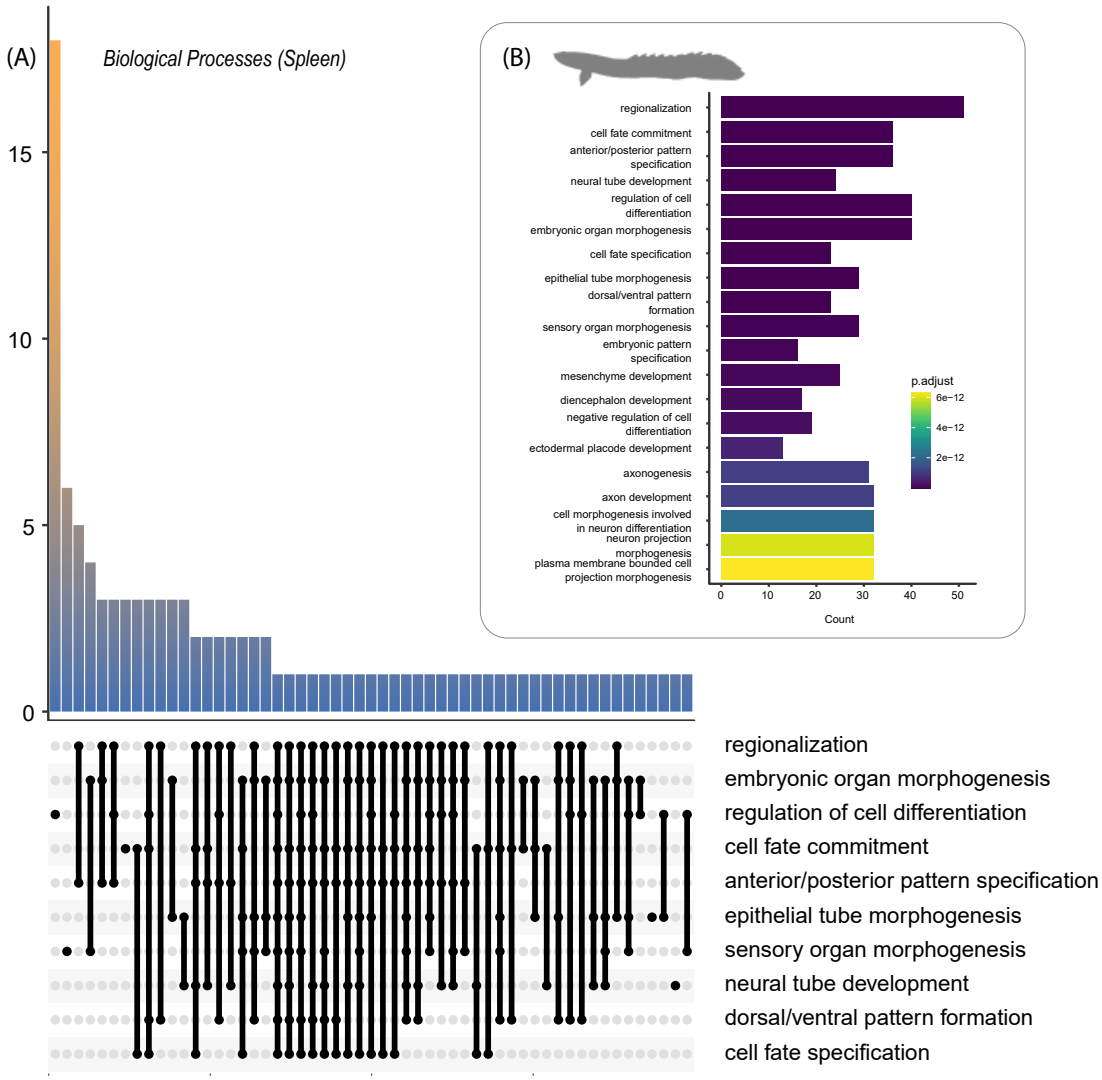
Supplemental Figure S3.13: Summary of Gene Ontology *Biological Processes* analysis from the *Polypterus bichir* liver transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



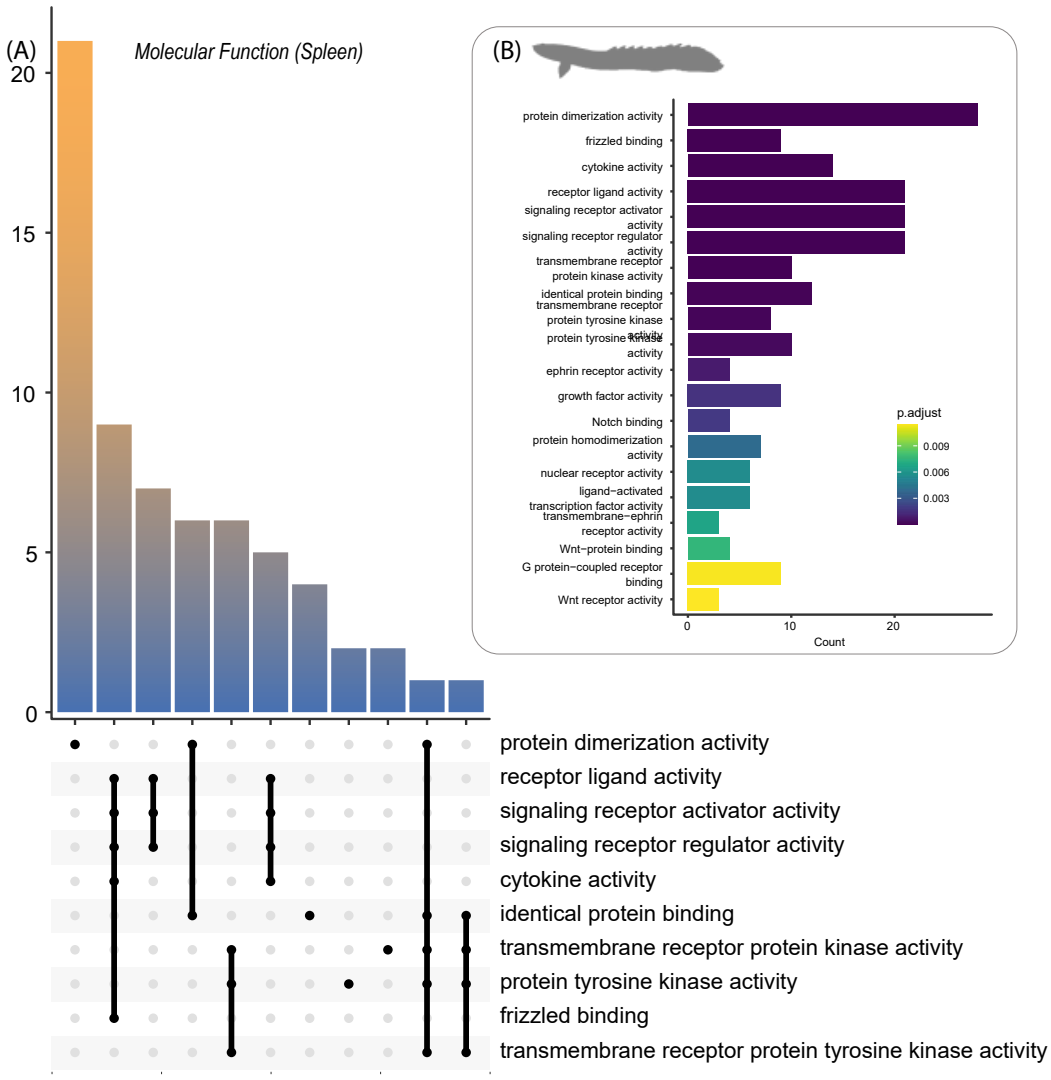
Supplemental Figure S3.14: Summary of Gene Ontology *Molecular Function* analysis from the *Polypterus bichir* liver transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



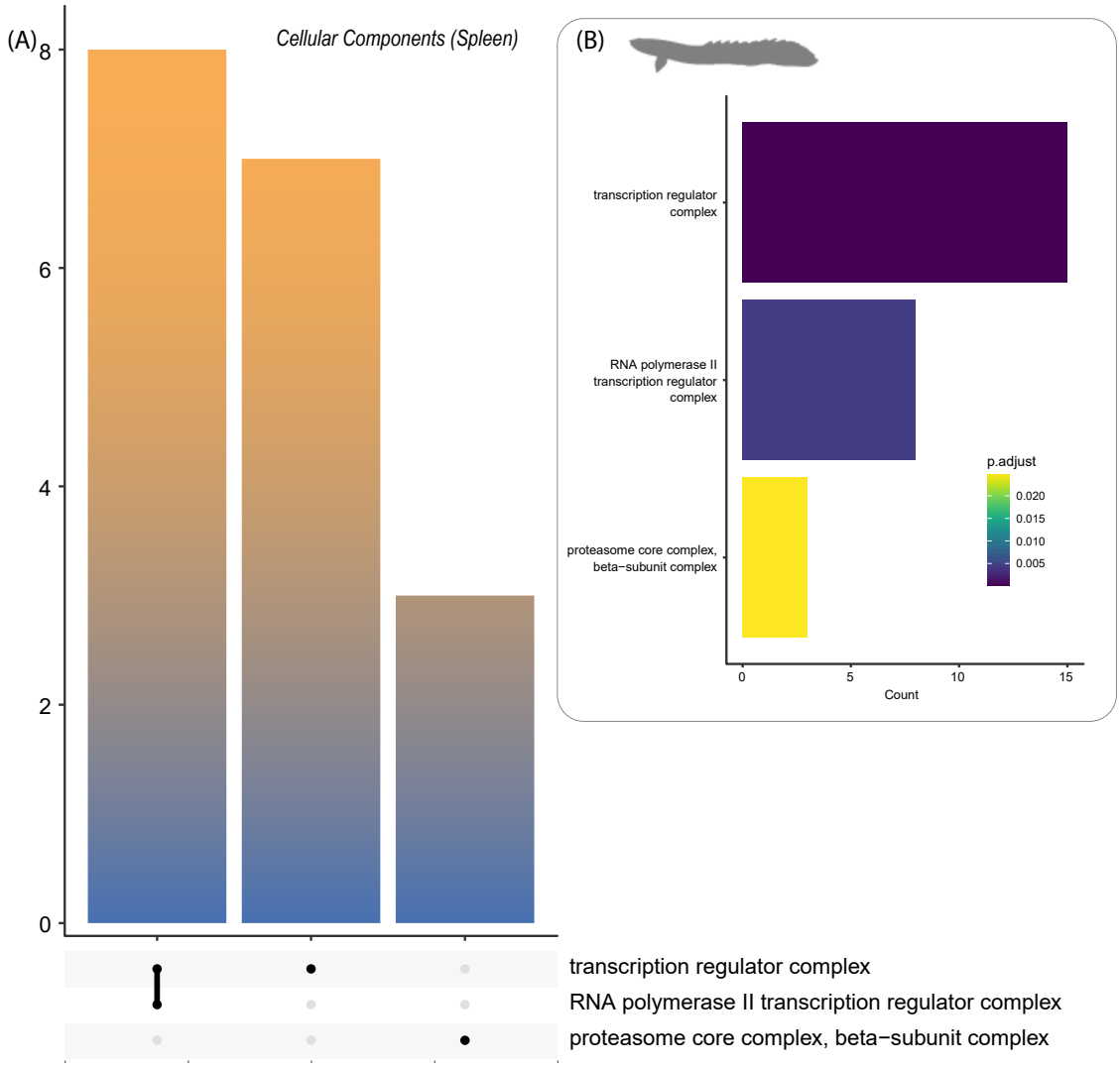
Supplemental Figure S3.15: Summary of Gene Ontology *Cellular Components* analysis from the *Polypterus bichir* liver transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



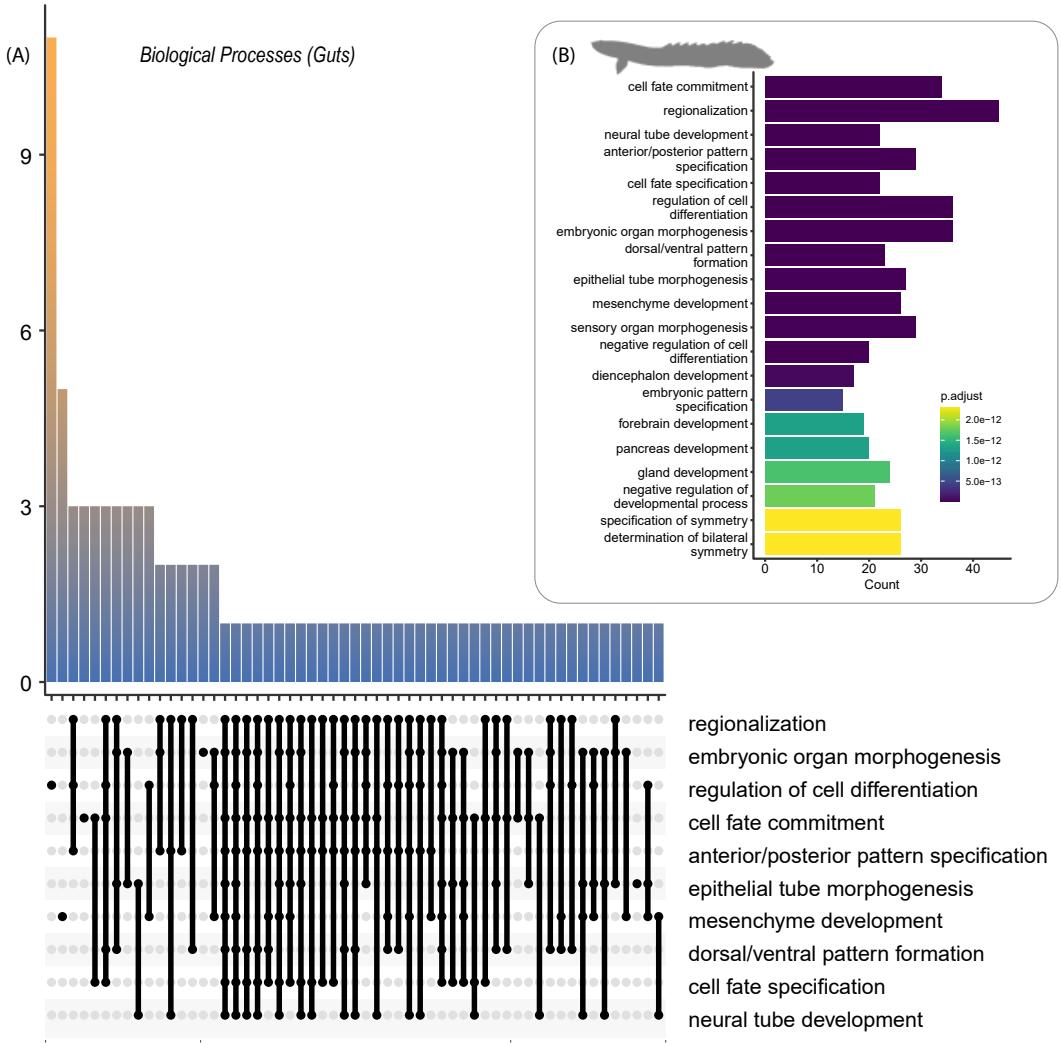
Supplemental Figure S3.16: Summary of Gene Ontology *Biological Processes* analysis from the *Polypterus bichir* spleen transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



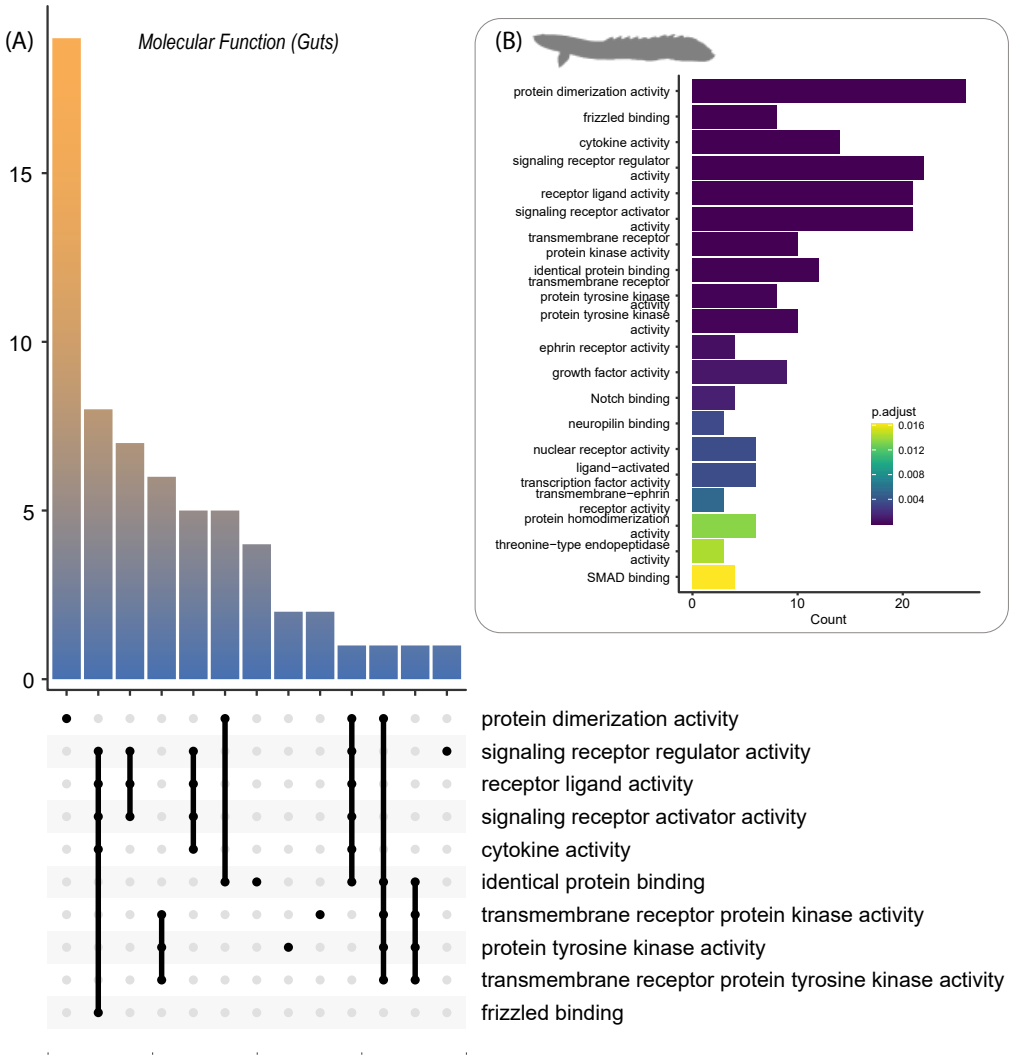
Supplemental Figure S3.17: Summary of Gene Ontology *Molecular Function* analysis from the *Polypterus bichir* spleen transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



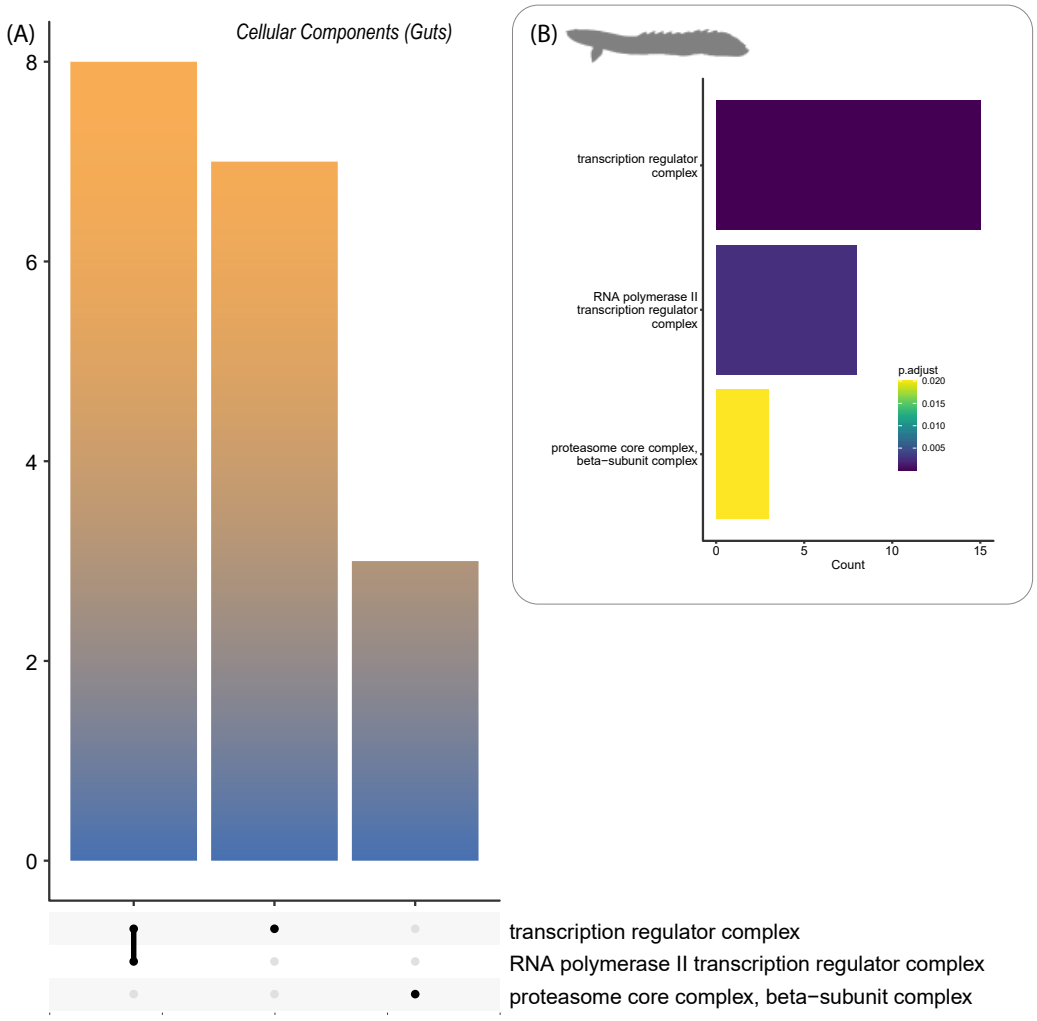
Supplemental Figure S3.18: Summary of Gene Ontology *Cellular Components* analysis from the *Polypterus bichir* spleen transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



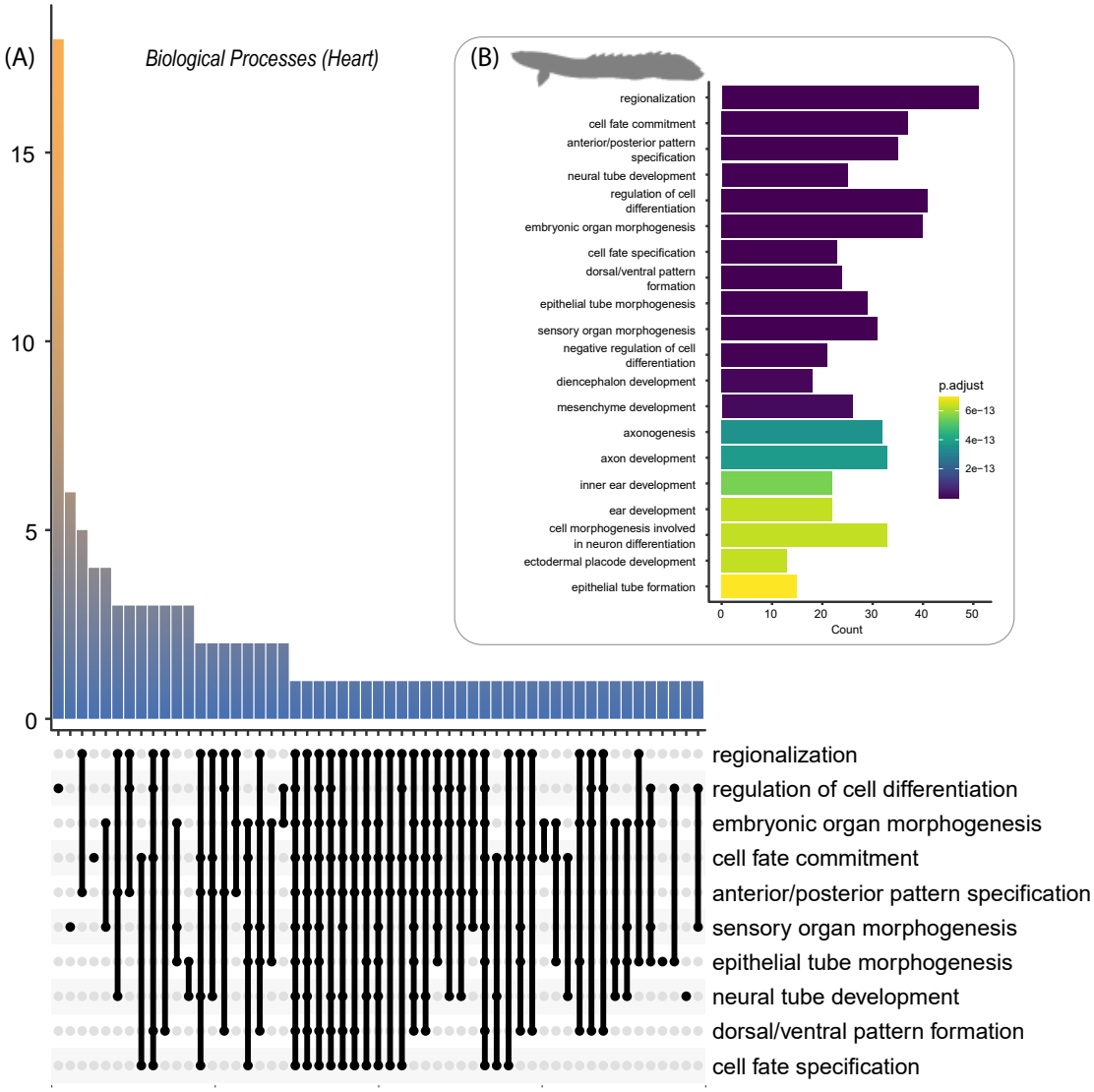
Supplemental Figure S3.19: Summary of Gene Ontology textitBiological Processes analysis from the *Polypterus bichir* spleen transcriptome. Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



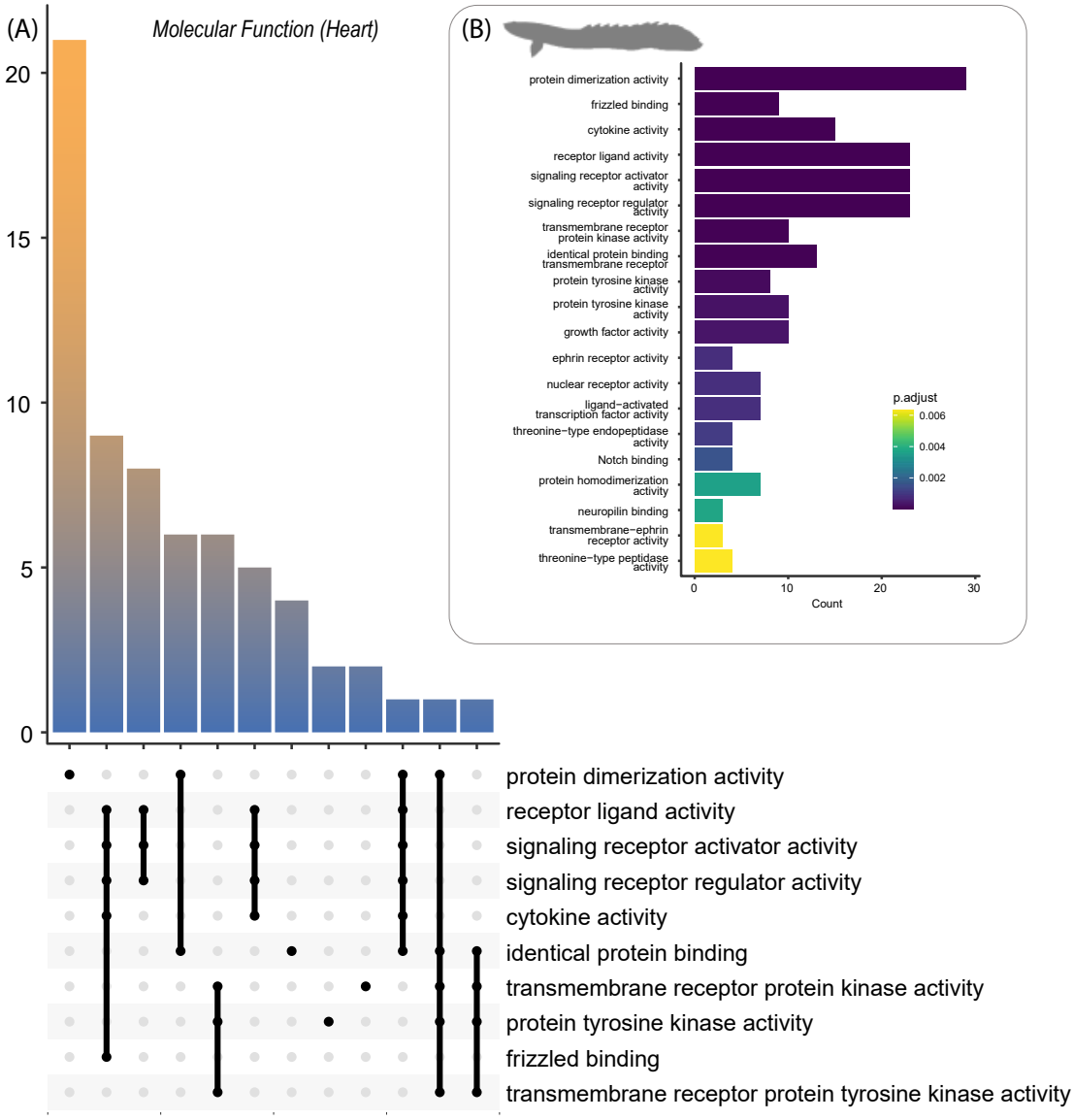
Supplemental Figure S3.20: Summary of Gene Ontology *Molecular Function* analysis from the *Polypterus bichir* spleen transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



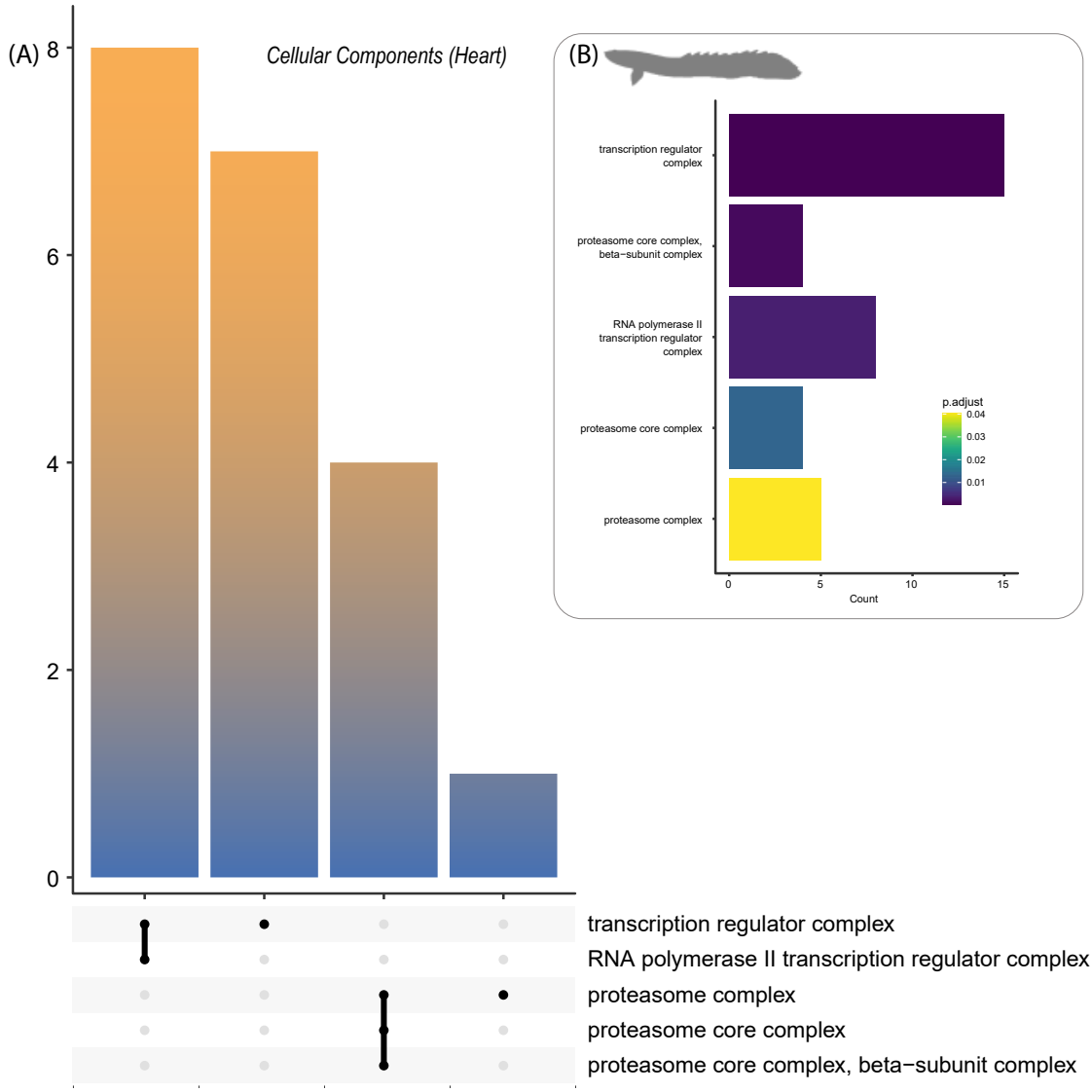
Supplemental Figure S3.21: Summary of Gene Ontology *Cellular Components* analysis from the *Polypterus bichir* guts transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



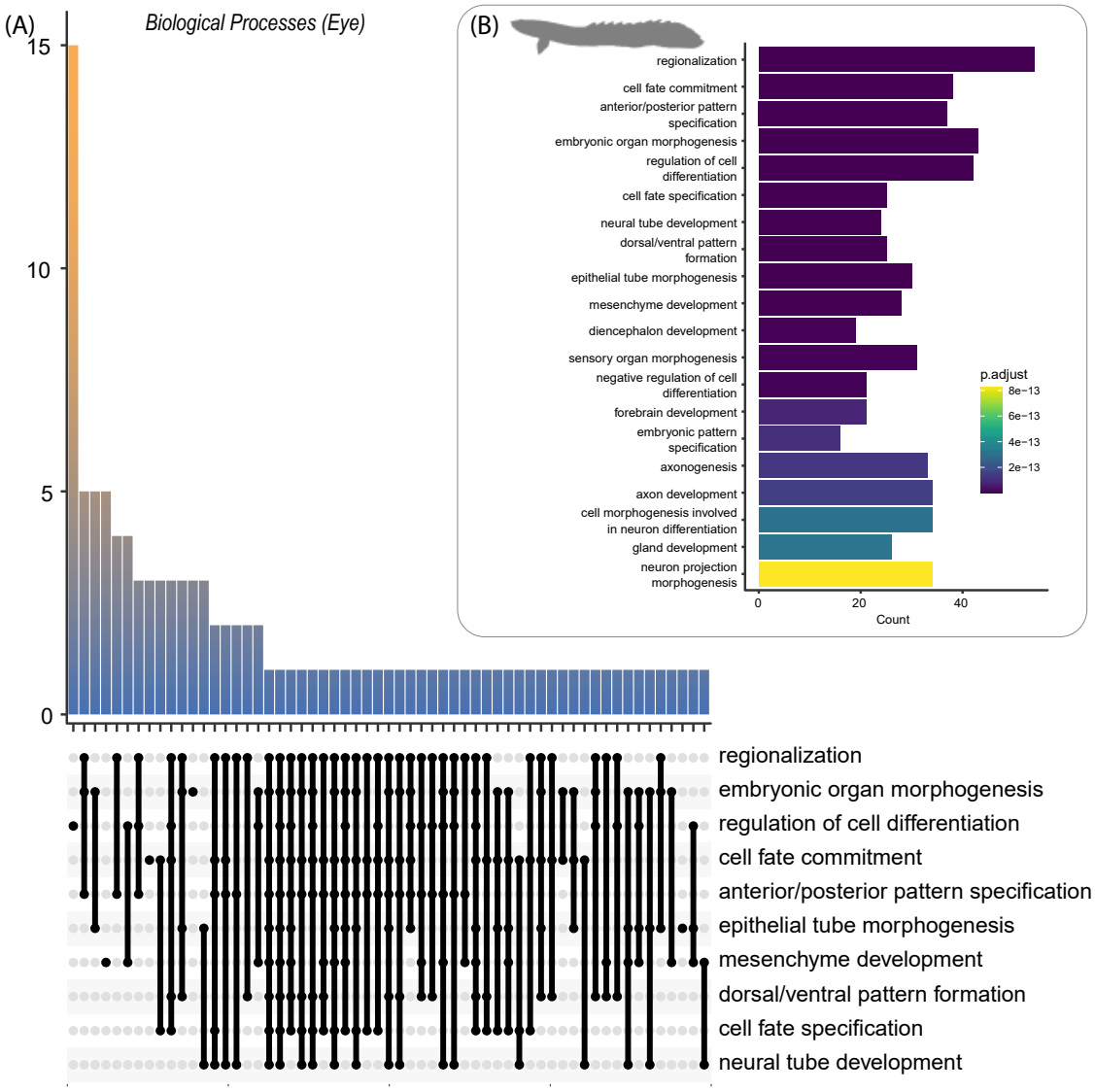
Supplemental Figure S3.22: Summary of Gene Ontology *Biological Processes* analysis from the *Polypterus bichir* heart transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



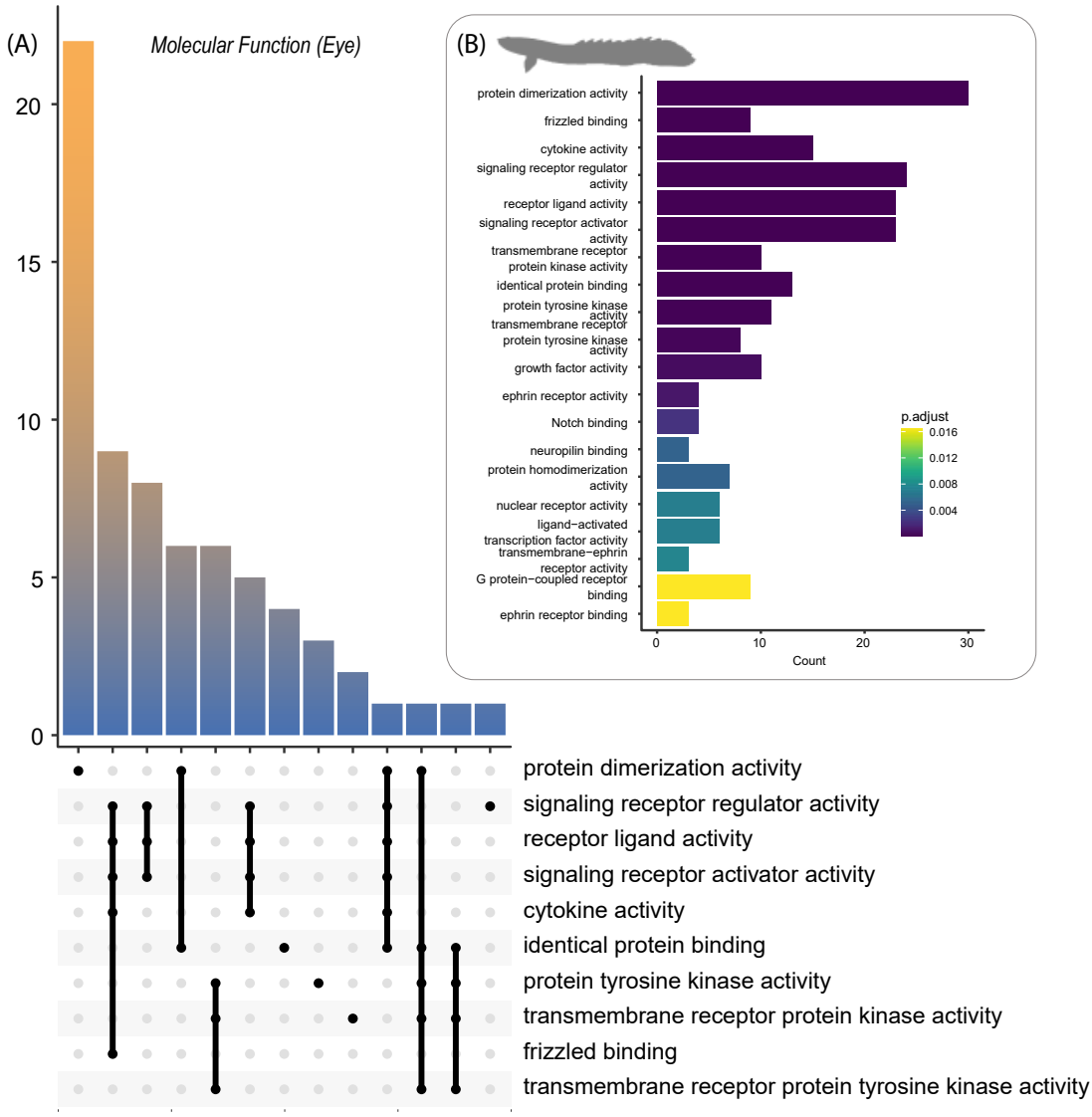
Supplemental Figure S3.23: Summary of Gene Ontology *Molecular Function* analysis from the *Polypterus bichir* heart transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



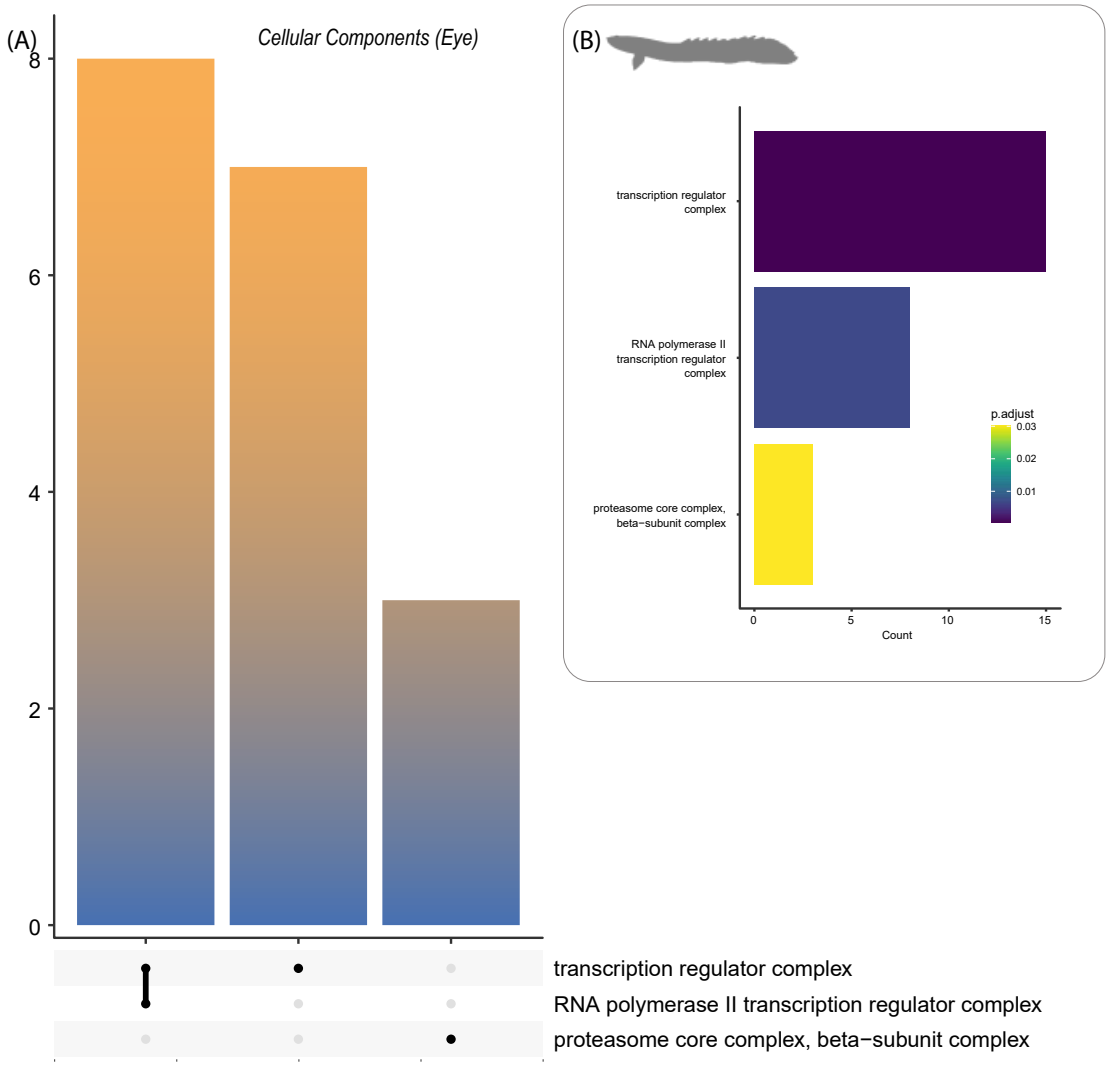
Supplemental Figure S3.24: Summary of Gene Ontology *Cellular Components* analysis from the *Polypterus bichir* heart transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



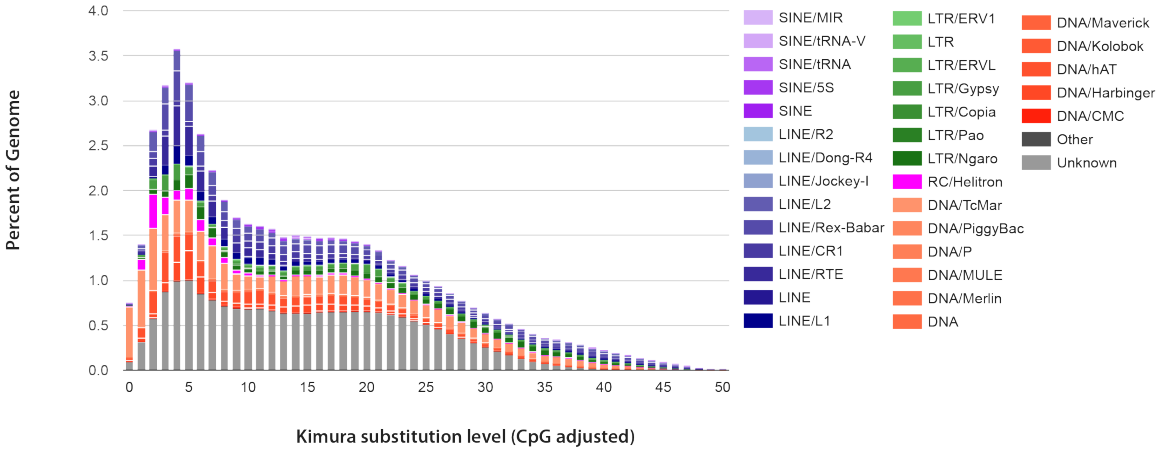
Supplemental Figure S3.25: Summary of Gene Ontology *Biological Processes* analysis from the *Polypterus bichir* eye transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



Supplemental Figure S3.26: Summary of Gene Ontology *Molecular Function* analysis from the *Polypterus bichir* eye transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



Supplemental Figure S3.27: Summary of Gene Ontology *Cellular Components* analysis from the *Polypterus bichir* eye transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.



Supplemental Figure S3.28: Summary of Gene Ontology *Cellular Components* analysis from the *Polypterus bichir* eye transcriptome.
Predicted proteins from the transcriptome were used as inputs to assess (A) processes and their intersections and (B) most common terms.

3.10.2 Supplemental Tables

Supplemental Table S1: Comparative genome assembly metrics of *Polypterus bichir* and related species

Species	Genome Size	Scaffold N50	Scaffold L50	Reference
Polypterus bichir	3.9Gb	202Mb	7	(This study)
Erpetoichthys calabaricus	3.8Gb	199Mb	7	[149]
Polypterus senegalus	3.6Gb	189Mb	8	[18]
Amia Calva	831Mb	41Mb	9	[67]
Lepisosteus oculatus	945Mb	6.9Mb	45	[28]
Lepisosteus osseus	930Mb	53Mb	8	[122]

Supplemental Table S2: BUSCO scores of *Polypterus bichir*

	Actinopterygii	Vertebrata
Complete BUSCOs (C)	2630 (72.2%)	2638 (78.77%)
Complete and single-copy BUSCOs (S)	2563 (70.4%)	2596 (77.4%)
Complete and duplicated BUSCOs (D)	67 (1.8%)	42 (1.3%)
Fragmented BUSCOs (F)	90 (2.5%)	272 (8.1%)
Missing BUSCOs (M)	920 (25.3%)	444 (13.2%)
Total BUSCO groups searched (n)	3640	3354

Supplemental Table S3: Phylogenetic Signal and Variance Analysis using Pagel's Lambda, Blomberg's K, and ANOVA

	Pagel's lambda		**Blomberg's K**		**ANOVA**	
	λ	P-value	K	P-value	F	P-value
DNA	0.98	**1.89e-13**	0.48	**6e-04**	1.34	0.78
LTR	0.82	**6.64e-11**	0.41	**1e-04**	19	0.23
LINE	0.98	**1.51e-19**	0.94	**1e-04**	27	0.14
SINE	1.01	**4.37e-15**	0.53	**3e-04**	19	0.23
Genome Size	0.72	**4.45e-15**	0.66	**6e-04**	**47.7**	**0.04**

Bolded values indicate P-values below 0.05

Supplemental Table S4: Evaluation of model fit for different TE types, genome size, and habitat characteristics under Brownian motion and Ornstein-Uhlenbeck (OU) models

	Brownian Motion			**OU**		
Traits	Likelihood	AIC	AICw	Likelihood	AIC	AICw
DNA	-276.21	556.43	0.12	**-273.25**	**552.49**	**0.877**
LTR	-181.09	366.18	<0.001	**-165.23**	**336.47**	**>0.999**
LINE	**-219.31**	**442.61**	**0.57**	-218.61	443.22	0.43
SINE	**-69.81**	**143.63**	**0.73**	-69.80	145.60	0.27
Total TE	**-360.08**	**724.17**	**0.61**	-359.53	725.07	0.39
Size	-601.31	1206.63	0.05	**-597.29**	**1200.58**	**0.95**
Depth	-528.20	1060.40	0.02	**-523.05**	**1052.10**	**0.98**
Mean Lat.	-481.17	966.34	<0.001	**-466.67**	**939.35**	**>0.999**
Median Lat.	-498.38	1000.76	<0.001	**-479.89**	**965.79**	**>0.999**
Lower Lat.	-505.02	1014.04	<0.001	**-490.84**	**987.68**	**>0.999**
Upper Lat.	-494.66	993.32	<0.001	**-467.44**	**940.89**	**>0.999**
Genome Size	-2224.41	4452.83	<0.001	**-2208.42**	**4422.85**	**>0.999**

Bolded rows indicate best-fit model based on the highest AIC weight (AICw).

Supplemental Table S5: AIC scores of phylogenetic linear models evaluating each type of transposable element (TE)

TE (y) Name	**TE (x) Name**	**TE(x)**	**TE(x) + Habitat**	**TE(x) + Taxonomy**	**TE(x) + Taxonomy + Habitat**
DNA	LTR	534.53	537.62	536.22	539.19
DNA	LINE	534.30	537.20	535.54	538.46
DNA	SINE	551.10	552.68	552.85	560.85
LTR	LINE	309.25	307.19	311.22	309.19
LTR	SINE	368.15	366.66	369.91	366.36
LINE	SINE	440.59	443.64	441.93	445.01

Supplemental Table S6: Phylogenetic linear models depicting the correlation between TE abundance and genome size, latitude, body size, or depth under a Brownian motion model of trait evolution.

Trait	TE	Intercept (a)	Coefficient (b)	SE (b)	T value	P value
Genome size	All	32.72	2.65E-03	1.35E-03	1.97	0.05
	DNA	8.35	9.59E-10	6.00E-10	1.60	0.11
	LINE	5.96	4.80E-04	3.50E-04	1.39	0.17
	LTR	2.67	1.56E-10	2.41E-10	0.65	0.52
	SINE	**0.62**	**1.89E-10**	**8.00E-05**	**2.37**	**0.02**
Latitude	All	36.44	1.41E-02	3.92E-02	0.36	0.07
	DNA	9.50	1.20E-02	1.73E-02	0.69	0.49
	LINE	6.46	8.60E-03	9.96E-03	0.86	0.39
	LTR	2.60	1.12E-02	6.81E-03	1.64	0.10
	SINE	**1.06**	**-6.06E-03**	**2.36E-03**	**-2.57**	**0.01**
Body Size	All	36.15	1.55E-01	8.23E-01	0.19	0.85
	DNA	12.91	1.29E+01	3.58E-01	-1.95	0.05
	LINE	5.56	2.60E-01	2.08E-01	1.25	0.22
	LTR	2.24	1.54E-01	1.44E-01	1.07	0.29
	SINE	0.77	3.16E-02	4.91E-02	0.64	0.52
Depth	All	30.63	6.62E-04	8.91E-04	0.74	0.46
	DNA	9.06	4.44E-04	2.88E-04	1.54	0.13
	LINE	4.04	5.10E-05	2.24E-04	0.22	0.82
	LTR	2.10	-6.60E-06	1.09E-04	-0.06	0.95
	SINE	0.67	-5.50E-06	6.50E-05	-0.08	0.93

Bolded rows indicate P-values below 0.05.

Supplemental Table S7: Results of phylogenetic linear models evaluating the correlation between TE abundance and genome size, latitude, body size, or depth under an OU model.

Trait	**TE**	**Intercept (a)**	**Coefficient (b)**	**SE (b)**	**T value**	**P value**
Genome size	All	20.22	1.15e-08	1.79e-09	6.42	**4.55e-09**
	DNA	4.76	4.15e-09	7.44e-10	5.58	**2.05e-07**
	LINE	1.45	3.21e-09	4.24e-10	7.58	**1.75e-11**
	LTR	0.09	1.01e-09	2.23e-10	4.55	**1.52e-05**
	SINE	0.04	2.51e-10	1.14e-10	2.21	**2.95e-02**
Latitude	All	1.52e-05	-5.62e-03	0.06	-0.09	0.93
	DNA	7.89	1.24e-03	0.02	0.05	0.96
	LINE	4.66	-2.73e-02	0.01	-1.79	0.07
	LTR	1.67	1.60e-04	0.01	0.02	0.98
	SINE	0.77	-6.45e-03	0.01	-1.93	0.06
Body Size	All	29.69	-0.19	1.03	-0.19	0.85
	DNA	11.18	-0.89	0.40	-2.20	**0.03**
	LINE	2.04	0.51	0.25	2.01	**0.05**
	LTR	0.59	0.29	0.11	2.56	**0.01**
	SINE	0.32	0.07	0.06	1.29	0.19
Depth	All	27.89	1.813e-03	1.13e-03	1.61	0.11
	DNA	7.24	1.50e-04	3.72e-04	0.40	0.69
	LINE	3.45	2.26e-04	3.14e-04	0.72	0.47
	LTR	1.58	4.3e-05	1.52e-04	0.28	0.78
	SINE	0.57	2.38e-05	8.11e-05	0.29	0.77

Bolded rows indicate P-values below 0.05.

Supplemental Table S8: Evidence for correlated evolution between TEs from phylogenetic linear models

TE (x)	TE (y)	Intercept (a)	Coeff. (b)	SE (b)	T Value	P Value	R^2	R^2 Adj.
LINE	DNA	8.07	1.94	0.71	2.73	**7.53e-3**	0.07	0.06
LTR	DNA	6.41	1.23	0.24	5.09	**1.65e-06**	0.20	0.20
SINE	DNA	4.51	0.80	0.15	5.18	**1.16e-06**	0.21	0.20
LINE	LTR	1.64	0.54	0.22	2.46	**0.01549**	0.06	0.05
SINE	LTR	-0.14	0.46	0.05	8.83	**3.36e-14**	0.44	0.43
SINE	LINE	4.70	1.12	0.40	2.78	**0.006573**	0.07	0.06

Bolded values indicate P-values below 0.05

CHAPTER 4: ON THE EVOLUTIONARY ORIGINS OF SIGNAL REGULATORY PROTEINS AND CD47

Manuscript in preparation

To be submitted to Immunology

4.1 Abstract

Immunoglobulin (Ig) genes that encode antibodies and T-cell receptors (TCRs) are essential components of the vertebrate adaptive immune response that arose out of ancestral vertebrate innate immune gene families. However, the evolutionary history among TCRs, Ig proteins, and certain extant innate immune receptor gene families that share motifs of a variable (V) domain and joining (J) segments remain unclear. Here we focus on the evolutionary history of the signal-regulatory protein (SIRP) multi-gene family which contain extracellular Ig-like domains with VJ exons and encode transmembrane glycoproteins that play a crucial role in cellular communication and the innate immune response [74]. Integrating genomic data from all major vertebrate lineages with phylogenetic and syntenic analyses, we demonstrate that SIRPs are far more widespread and highly variable in copy number across vertebrates than previously hypothesized. Our phylogenetic analyses reveal SIRPs to have evolutionary origins that span all major jawed vertebrate lineages, suggesting a far more ancient origin than previously hypothesized. Further, we find no evidence for an ancient origin of the CD47 ligand that interacts with SIRPs. Instead, CD47 appears to have arisen at the dawn of amniotes, suggesting that the evolutionary origins of this ligand and receptor are decoupled. Collectively, our findings provide a new perspective on our understanding of the origins and diversification of innate immune receptor gene families related to the emergence of the adaptive immune system.

4.2 Introduction

The ability to mount an adaptive immune response using immunoglobulin (Ig) genes that encode antibodies and T-cell receptors (TCRs) is a hallmark of jawed vertebrates. Over the past 50 years, the search for the evolutionary origin of TCRs and antibody encoding Ig domains has yielded the view that they were derived

from the ancestral innate immune system in early jawed vertebrates (gnathostomes). With continued genome sequencing efforts, analyses have begun to reveal which extant innate immune share recent common ancestry with Ig-type adaptive immune receptors, providing new insights into the genomic substrate that gave rise to these key components of the adaptive immune response. For example, TCRs and Ig proteins share a general structure that includes a single antigen-recognizing variable type (V) domain, the carboxy terminus of which undergoes somatic recombination with either a single joining (J) segment, or a diversity (D) segment and a J segment. The innate immune receptors believed to share the closest evolutionary relationships with these recombining receptors also possess VJ-joined exons. This raises a question: did early gnathostomes possess a diversity of VJ-containing innate immune gene families, or did these arise after the origin of TCRs and antibody encoding Ig genes? Currently, our understanding of the distribution of VJ-containing innate immune receptors across jawed vertebrates remains limited. For example, NKp30, CD8, and CD49 are present in all major gnathostome lineages, while Novel Immune Type Receptors (NITRs) are hypothesized to be unique to ray-finned fishes. Investigations that fill the knowledge gap for other VJ-containing innate immune receptor gene families are vital for understanding the molecular diversification of the gnathostome innate immune system and the pathways that gave rise to TCRs and antibody encoding Ig genes.

Among known VJ-containing gene families, the evolution of signal-regulatory proteins (SIRPs) remains particularly poorly known. SIRPs are transmembrane glycoproteins that play a crucial role in cellular communication [2] and the innate immune response [74]. In humans, SIRPs possess three extracellular Ig-like domains with VJ exons. In addition, the cytoplasmic domain of SIRP contains two immunoreceptor tyrosine-based inhibition motifs that can interact with Src homology

2 domain-containing phosphatases. The human SIRP gene family includes SIRPα (also known as SHPS-1, BIT, MFR, CD172a, or p84), SIRPβ1, SIRPβ2, SIRPγ, and the SIRPβ3p pseudogene [208]. However, comparative investigations of SIRP family members have been restricted primarily to mammals, where they appear to be widespread and diverse [31]. SIRPs have been identified on myeloid and other cells in humans, mice, rats, and cattle[208], where multiple SIRPβ-like molecules with activating potential possibly interact with the adaptor molecule DAP12 to activate and induce phagocytosis in myeloid cells[59, 41]. Their diversity, even among related species like mice and rats, suggests additional roles in host defense, possibly including pathogen recognition [19]. Comparisons between humans and rodents have additionally revealed two SIRP gene clusters in the latter with unique structures that suggest an ancestral SIRP with a more complex domain arrangement [208]. Outside of mammals, three SIRP family members with conserved synteny to mammals have been identified in chickens, leading to speculation as to whether SIRPs may be widespread in amniotes or perhaps a more ancient gene cluster that arose proximate to the origin of V(D)J recombining receptors [68].

The diversity of SIRPs in mammals has immediate translational relevance as SIRP dysregulation and interactions with other molecules have been implicated in a variety of diseases including autoimmune disorders [186], diabetes [190, 186], and cancer [198]. Comparative investigations of SIRPs would likely be fruitful for understanding the molecular biology of these cases. However, a phylogenetic perspective is needed to disentangle homology from convergence. As gene functions and regulation evolve, accurate assessments of homology are necessary references for translating research between model organisms and humans. In the case of SIRPs, knowledge of the evolutionary history of its interacting partner CD47 would be particularly valuable as this ligand is critical for the ability of SIRPs to play a regulatory role.

CD47, also known as integrin-associated protein, is a transmembrane protein characterized by a single immunoglobulin-like (Ig-like) domain and five membrane-spanning regions [100]. Serving as a cellular ligand for SIRPα, bidirectional signaling from the interaction between SIRPα and CD47 causes several cell-to-cell responses, including T-cell activation, cell-cell fusion stimulation, and phagocytosis inhibition [188]. As the CD47/SIRPα interaction is a rare negative regulator of phagocytosis in humans [152], this interaction has begun receiving considerable attention. For example, increased CD47 on the cell surface has been shown to shield hematopoietic cells from phagocytosis as they migrate from the bone marrow [188], and changes in expression have been associated with cancer [70, 188], stress [178], and differences in aging [217]. Correspondingly, blocking CD47 interaction with SIRPα can enhance cancer cell clearance by macrophages [226]. Although the translational relevance of the SIRP/CD47 interaction is well established, investigations into the evolutionary history of this ligand remain lacking.

In this study, we use comparative genomic approaches to establish the evolutionary origins of SIRPs and CD47 using genomes that span all major groups of vertebrates. We first use identify SIRP orthologs and paralogs, and use phylogenetic approaches to support a far earlier origin of SIRPs than previously hypothesized. Our results dramatically expand the known distribution of SIRPs, demonstrate their pervasiveness across many vertebrate clades in which they are currently undocumented, and assess the degree to which gene regions surrounding these clusters are syntenic between distantly related vertebrate groups. We then test whether the number of Ig domains within SIRP genes represents an evolutionary conserved or labile trait, while also using ancestral state reconstructions to reveal the history of functional gains and losses predicted from sequence data. Finally, we use phylogenetic methods to

determine the evolutionary origin of CD47, revealing that the origin of the SIRP gene family and this ligand may be evolutionarily decoupled. Collectively, these findings represent the most comprehensive comparative analyses conducted on this gene family and its interacting partner, filling a critical knowledge gap for translational research and our understanding of the evolutionary history of VJ-containing innate immune receptors.

4.3 Methods

4.3.1 *Dataset Assembly*

All protein sequences used for this study were retrieved from the National Center for Biotechnology Information (NCBI) using the NCBI Datasets tool. To ensure a comprehensive representation of the evolutionary diversity of the Signal Regulatory Protein Superfamily (SIRPS), we targeted 107 species that contain, representative taxa from all major vertebrate lineages, including mammals (29), squamates (16), archosaurs (3), coelacanths, amphibians (5), ray-finned fishes (27), cartilaginous fishes (5), and agnathans (19). This strategic taxon selection aimed to provide a broad phylogenetic perspective on the distribution and evolution of SIRPS across the early history of vertebrates.

4.3.2 *Identification of SIRP genes*

The curated protein sequences dataset was transformed into a searchable database using DIAMOND, an efficient and high-performance tool for aligning protein sequences [37]. To establish a reference framework for the identification of Signal Regulatory Protein (SIRP) members, confirmed sequences of SIRPs from *Homo sapiens*, *Gallus gallus*, and *Bos taurus* were obtained from the National Center for Biotechnology Information (NCBI). These reference sequences were then employed as queries in a BLASTp search against the DIAMOND-generated database [37], with an E-value

threshold of $1e^{-10}$. Custom Python scripts were used to generate a more focused dataset by eliminating redundant isoforms from search results. The sequences retrieved from this search were subsequently aligned using FAMSA [58], a fast and accurate multiple sequence alignment tool designed to handle large datasets efficiently. Finally, aligned sequences were used as input for IQ-TREE2 [132], which employed maximum likelihood methods to construct a phylogenetic tree that represents the evolutionary relationships among the identified SIRPs sequences.

4.3.3 *Nomenclature for SIRPs*

Preliminary analyses revealed pervasive misannotation of Signal Regulatory Protein (SIRPS) members by annotation software. In this work, we label any discovered SIRP genes consecutively within each species (e.g., *Naja naja sirp1*, *Naja naja sirp2*). Due to the high numbers of lineage-specific expansions within the SIRP gene cluster, one-to-one orthologs are generally not identifiable between species. As such, *Naja naja sirp1* would not necessarily reflect a true ortholog of a gene labeled *sirp1* in another species. Pseudogenes were identified based on the appearance of internal stop codons and designated with a "*p*".

4.3.4 *Genomic Identification of Immunoglobulin Domains*

We utilized the NCBI's Conserved Domain Database (CDD) [125, 218] to identify immunoglobulin (Ig) domains within putative SIRP sequences. The CDD is a comprehensive protein annotation resource that comprises a collection of well-annotated multiple sequence alignment models for both ancient domains and full-length proteins. These models are available as position-specific score matrices (PSSMs), enabling the rapid identification of conserved domains in protein sequences through Reverse Position-Specific BLAST (RPS-BLAST). For each sequence, we used CDD to provide the positions of the Ig domains. These domains were subsequently extracted from the sequences using custom Python scripts and named sequentially (e.g., ig1,

ig2).

4.3.5 *Phylogenetic Analysis of Ig Sequences*

Extracted Ig domains were aligned using FAMSA [58] and visualized in Aliview [115] to ensure sequences did not contain any erroneous flanking regions. Sequences with incomplete domains were also excluded from the dataset. Using the resulting dataset, we conducted two analyses. We first used the dataset of all individual Ig domains as input for IQ-TREE2 [132] and estimated the maximum likelihood tree topology conditioned on the best-fit model identified using Bayesian Information Criterion [145] in ModelFinder [103]. We additionally generated a concatenated alignment of the Ig domains for each species and repeated our phylogenetic analyses to assess how high variability in flanking regions impacted the topological inference when the entire sequence was used (above). In all cases, topological support was assessed using 5000 bootstrap replicates.

The paralogs were enumerated using a Python script, and species were identified for each query. Additionally, we quantified the number of immunoglobulins (Ig) domains per species by analyzing the hit data file obtained from the NCBI Conserved Domain Database (CDD). The list of species was utilized to obtain a time-calibrated phylogenetic tree from Timetree.org [112], which was also employed to evaluate the phylogenetic signal of immunoglobulin domains. The phylogenetic signal was assessed using the phylosig function in the phytools package, calculating both Pagel's lambda (λ) [156] and Blomberg et al.'s K [20]. Lambda values range from zero to one, where zero indicates no phylogenetic signal and one signifies complete phylogenetic correlation. The empirical λ values were statistically compared to the null hypothesis of $\lambda = 0$ for each trait using a likelihood ratio test to determine significance. Blomberg's K serves as a complementary measure to λ, with values below one suggesting less phylogenetic signal than expected under a Brownian motion model, and values above one indicating a stronger association between trait distribution and the phylogeny

than expected.

4.3.6 *Identification and Analysis of CD47 sequences*

CD47 sequences for *Homo sapiens*, *Bos taurus*, and *Gallus gallus* were downloaded from the National Center for Biotechnology Information (NCBI) database. This approach was consistent with the pipeline previously utilized for SIRP sequences, ensuring comparability between the datasets. The retrieved CD47 sequences were subjected to BLAST (Basic Local Alignment Search Tool) analysis against a curated list of vertebrates to identify homologous sequences. Subsequently, the sequences were aligned using MAFFT (Multiple Alignment using Fast Fourier Transform), a widely recognized tool for sequence alignment. Following alignment, phylogenetic tree construction was performed using IQ-TREE 2, a powerful and efficient phylogenetic software. The accuracy and reliability of the dataset were ensured by removing duplicates of sequences using custom Python scripts, a critical step for preventing redundancy and ensuring the integrity of the subsequent analyses.

4.3.7 *Syntenic Analyses of the SIRP gene cluster and CD47*

To assess synteny, we first used Genomicus v100.1 to explore the degree of synteny in the 15 genes flanking either side of the human SIRP gene cluster on chromosome 20 in the following representative vertebrate lineages: (mammals, squamates, archosaurs, coelacanths, amphibians, ray-finned fishes, cartilaginous fishes, and agnathans.) As Genomicus is based on the assumption of correct annotation, we additionally conducted reciprocal DIAMOND searches of protein sequences from the annotated human chromosome 20 against the following taxa. Reciprocal DIAMOND searches were restricted to scaffolds or chromosomes containing SIRPs. A maximum of five hits was returned based on an e-value cutoff of e^{-10}, and only the first gene encountered in the annotation file for each genome was used to ensure that genes with multiple isoforms were only represented by

one species. The identification of syntenic regions across the selected organisms was accomplished using collinear analyses in MCScanX [219] the circlize v.0.4.16 package in R [85]. The syntenic analysis was repeated with the CD47 sequences as well using 5 genomes, *Homo sapiens*, *Gallus gallus*, *Chelonia mydas*, and *Naja naja*.

4.4 Results and Discussion

4.4.1 *SIRPs have Ancient Evolutionary Origins*

Diamond searches revealed a large pool of candidate sequences from genomes on NCBI that were potentially homologous to sequences of known SIRPs used as queries. We identified a total of 214 SIRP sequences from the genomes of 107 representative vertebrates, including the newly sequenced *Polypterus bichir* [122,]. Potential homologs were similar between queries, so all reported findings below are based on the human query unless otherwise noted. We found putative SIRPs as nearly ubiquitous across amniote lineages spanning mammals, archosaurs, squamates, and turtles. Likewise, we found candidate SIRPs outside of amniotes, including in early diverging ray-finned fishes, coelacanths, and chondrichthyes. Phylogenetic analyses of SIRP Ig domains from putative sequences revealed strong support [Bootstrap Support (BSS) = 94] for the existence of multiple SIRP clades (Fig. 4.1), suggesting that these clades hold evolutionary origins that exceed the currently known distribution in amniotes [68].

In addition to their ancient origin, our results reveal that ancient paralog events form the foundation of modern SIRP diversity in vertebrates. The majority of these sequences can be classified into eight distinct clades. The clade containing human SIRPβ1, SIRPα, SIRPγ, and SIRPβ3 spans *Dasypus*, *Myotis* (BSS = 81). The clade with human SIRPβ1 exhibits a high degree of similarity to SIRPs spanning macaques to camels, and it is part of a larger clade that includes human SIRPα and SIRPγ, which are unique to primates. Similarly, the clade containing human

Figure 4.1: The figure shows a representative SIRP phylogenetic tree using Human SIRPα as query. The dark grey inner circle shows the phylogenetic tree of the entire dataset. The light blue ring shows the presence of SIRPs, while the dark blue ring shows the presence of CD47.

SIRPd and SIRPb2 groups with sequences from camels, guinea pigs, and pumas. Collectively, these results suggest that the paralogs in the SIRP gene cluster are specific to placental mammals. This corresponds to the observation of other lineage specific paralog clusters such as those in snakes such as *Naja*, *Pantherophis*, and *Thamnophis* or early diverging ray-finned fishes such as *Polypterus* and *Erpetoichthys*.

Given aspects of their similarity to novel immune type receptors (NITRs) and phylogenetic distribution in Holosteans and Teleosts, [68] hypothesized that SIRPs may be the sister lineage to NITRs. In this proposed scenario, SIRPs were restricted

to sarcopterygians, and NITRs restricted to actinopterygians. However, our results show that this is not the case. Instead, SIRPs are found to span all jawed vertebrates, including several early diverging ray-finned fishes 4.1. In contrast, efforts to find NITRs in bichirs, sturgeon, paddlefish, or cartilaginous fishes have failed [177, 68, 230]. Integrating what is known about NITRs with our findings raises an intriguing alternate hypothesis concerning the relationship between these receptors: SIRPs may have given rise to NITRs prior to the common ancestor of holosteans and teleosts. Additional bichir, paddlefish, and sturgeon genomic and transcriptomic resources are needed to test such a hypothesis, as the extremely limited existing resources do not preclude the possibility that NITRs are not being found due to assembly/sequencing errors. Alternatively, an alternate gene family may have an analogous function in these lineages. Such a scenario would require NITRs to have been lost in bichirs, chondrosteans, and sarcopterygians independently. Regardless, our survey of SIRPs suggests that this gene family was present along with CD8Beta, CD79b, NKp30, and PRAPs during the genesis of the V(D)J recombining adaptive immune response.

4.4.2 *The evolutionary origin of CD47s is decoupled from the origin of SIRPs*

Our searches reveal a candidate pool of sequences that were potential homologs to CD47 across a wide range of vertebrates, as shown by the dark blue circle in 4.1. However, this pool did not span the same level of taxonomic breadth as the SIRP pool, and was restricted to amniotes. Phylogenetic analysis of putative CD47 orthologs provide no support for early diverging clades of paralogous origin (Fig. 4.2). Instead, CD47 is largely single copy in most amniotes with notable exceptions in *Bos taurus* and *Lacerta agilis*. These results suggest that CD47 interactions arose long after the evolutionary origin of SIRPs and is consistent with the hypothesis of that the CD47/SIRP interaction is conserved in amniotes [214].

In humans and other mammals, the CD47-SIRPα interaction functions as an inhibitory function, as it prevents signaling from phagocytosis receptors [154, 211]. Similar to SIRPα, SIRPγ serves as a counter-receptor to CD47, albeit with a significantly reduced binding affinity. This interaction is further characterized by its unidirectional signaling, primarily due to the absence of the intracellular cytoplasmic tail in SIRPγ. Consequently, the signaling cascade initiated by the CD47/SIRPγ interaction is confined to CD47, highlighting a distinct mechanism of action compared to the bidirectional signaling observed in the CD47/SIRPα axis [180, 194]. However, the absence of CD47 outside of amniotes suggests the high possibility of alternate patterns of SIRP function and expression. In particular, an unusual aspect of SIRPs is their expression in myeloid cells [12, 2]. This expression pattern contrasts sharply with Ig and TCR genes that are primarily expressed in lymphocytes [157]. This shift in expression patterns has been used to argue for a functional shift between SIRPs and other genes with VJ exons. When an amniote origin of CD47 is considered in the context of our finding of SIRPs in earlier diverging jawed vertebrates, this raises the possibility that myeloid expression represents an amniote evolutionary novelty that arose in concert with the SIRP/CD47 interaction. Further tests of expression patterns outside of amniotes are critically needed.

Testing for a possible shift in SIRP expression patterns may also help reconcile the presumed functional divergence of these receptors from other VJ-exon bearing genes that were present at the genesis of V(D)J recombination. However, such tests also have implications for clinical research. The CD47-SIRP axis represents a crucial mechanism of immune modulation, with significant implications for the development of therapeutic strategies targeting this pathway in cancer and other diseases [32, 198] gents with the potential to disrupt the CD47/SIRPα axis can mitigate immune evasion and enhancing immune responses against cancer cells

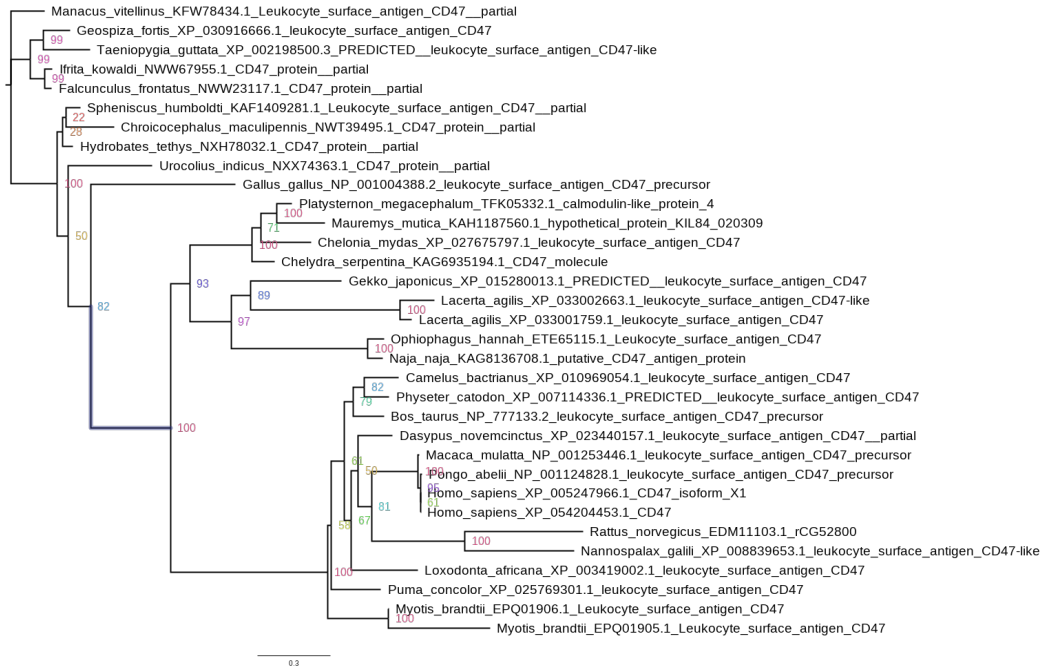


Figure 4.2: CD47 phylogenetic tree using Homo sapiens CD47 as a query

[88, 121]. In addition to activating macrophage-mediated tumor killing, the interruption of the CD47-SIRPα axis also has the potential to enhance antibody-dependent cellular cytotoxicity through the inhibition of SIRPα expressed on the surface of natural killer cells [139]. Moreover, targeting the CD47-SIRPα axis through CD47 or SIRPα blockade may enhance the function of macrophages as antigen-presenting cells [60]. [32, 198, 203], such work represents an exciting research frontier.

4.4.3 *The number of SIRP genes is highly variable between vertebrates*

Our comparative analysis of the SIRP gene cluster across representative vertebrate taxa revealed high variance in the numbers of paralogs within each lineage (Figure 4.3 A). The average paralog count across the studied species was 2.67 using humans as query. The values with other queries are quite similar (2.71 Bos taurus query; 3.2 chicken query). We compute a Pagel's Lambda value of 0.47 suggesting some degree of phylogenetic signal to the distribution of paralogs ($p = 0.01$). However, phylogenetic

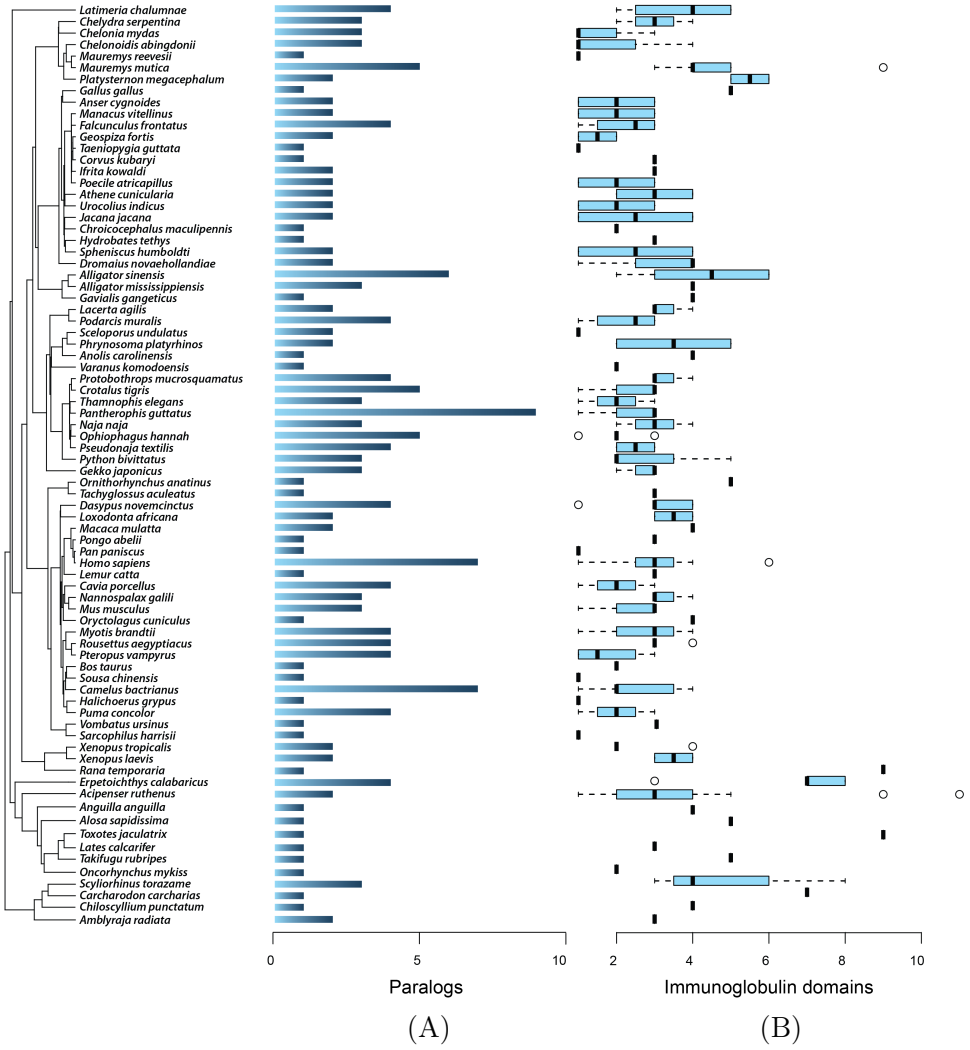


Figure 4.3: Phylogenetic tree with paralog counts and Ig domain variance using Human SIRPα as query

signal was not supported by Blomberg's K (0.16, p = 0.67). This mismatch between lambda and K was similar when using using chicken (Lambda=0.4460971 | p=0.009; K = 0.26, p = 0.12) as a query. In the case of cattle as a query, the additional paralogs found do support phylogenetic signal under both metrics (Lambda=0.60 | p=0.0009; K = 0.33, p = 0.04). In general, Selachimorphs (sharks) exhibited the lowest degree of paralogs (mean = 2.75). In contrast, Homo sapiens (humans) demonstrated a higher than average number of paralogs, along with other species such as Erpetoichthys calabaricus (reedfish), Alligator sinensis (Chinese alligator), and Mauremys mutica

(yellow pond turtle). Polypterus bichir exhibited the highest number of paralogs among the species analyzed, though some of this variation is likely attributed to annotation error. Manual annotation coupled with additional genomic resources for other Polypterus species is needed to assess the scope of paralogs in this lineage.

The diversity of paralogs between vertebrate lineages is similar to overall patterns of gene content variation in innate immune genes [44, 202, 45] and clustered gene families in general [179, 63]. However, the functional consequences of gene birth and death events in SIRPs remain largely unknown. At the population level, gene content variation between individuals can allow the innate immune response to defend against a broader range of pathogens at the species level [63]. Such variation can scale up between species to include both core and conserved functions, as well as species-specific responses between lineages [187]. It is also possible that the variation in copy number may reflect alternate ecological strategies for persisting in pathogen rich environments between lineages. For example, species exposed to a high diversity of pathogens might evolve a greater number of receptor copies to cope with this challenge, positing a possible link between ecological niche and immune function. For example, work across migratory and African birds has provided direct evidence that loss in immunogenetic diversity is linked to the ecological release of leaving the pathogen rich continent for Europe [151]. The degree to which lineages possess unique functions in SIRPs and whether those are linked to any aspect of vertebrate ecology remains unknown.

4.4.4 *Variation in Ig domains suggests shifts in domain architecture and function*

Our analysis of the number of Ig domains revealed that the average number of immunoglobulin (Ig) domains per species in the phylogenetic tree ranged from 4 to 6. The calculated Blomberg's K [20] value was 0.236, $p = 0.204$, indicating a weak phylogenetic signal, suggesting that the distribution of Ig domains across species is not

strongly influenced by their shared evolutionary history. Conversely, the Lambda [156] value was 0.675, $p = 3.29e-06$, suggesting a moderate degree of phylogenetic dependence. These metrics are commonly used to assess how traits are distributed across a phylogeny [206, 229, 123, 50]. However, a difference between lambda and K is not unexpected given the difference in the metrics. The calculation of lambda is based on scaling of branch lengths given a model of trait evolution to determine whether trait distributions are influenced by phylogeny [156]. In contrast, Blomberg's K is a measure of the observed variance in traits among species to the expected variance under a Brownian motion model of evolution, normalized by the phylogenetic contrasts' variance [20]. This means that K focuses on variance partitioning versus the fit to a scaled model of phylogeny. The significant λ with a nonsignificant K may reflect the presence of phylogenetic signal that affects the scaling of the tree but does not manifest as a significant deviation from the stringent Brownian expectation of trait variance among species. Given that our sampling is limited to only a few representative taxa per clade, additional taxon sampling may also be needed to better determine the trait variance between major vertebrate lineages. Moreover, as different sequences were returned as a function of query, we also observed fluctuation in the lambda and K values highlighting the potential influence of the choice of reference species on the perceived phylogenetic signal in comparative genomic analyses.

Tests of phylogenetic signal do not overwhelmingly support a strong difference in the average number of Ig domains per gene between species. However, the observed variation does indicate a very likely difference in Ig domain architecture and function within and between lineages. For NITRs, variation in the number of Ig domains has been associated with the presence of inhibitory (cytoplasmic ITIM), activating (charged residue in the transmembrane domain), and secreted forms as well as functionally ambiguous protein structures [66]. Future comparative transcriptomic studies are vital to determining the degree to which such forms exists

between SIRP paralogs and species. Given the functional diversity linked to these structural variations in NITRs, it's plausible that SIRPs exhibit a comparable degree of functional and structural heterogeneity. Integrating comparative transcriptomic studies into a phylogenetic framework can reveal whether such heterogeneity reflects lineage-specific immune adaptations thereby revealing patterns of convergent evolution, functional novelty, lineage specificity, and overall functional diversification.

4.4.5 *Syntenic analyses support an early vertebrate origin for SIRPs*

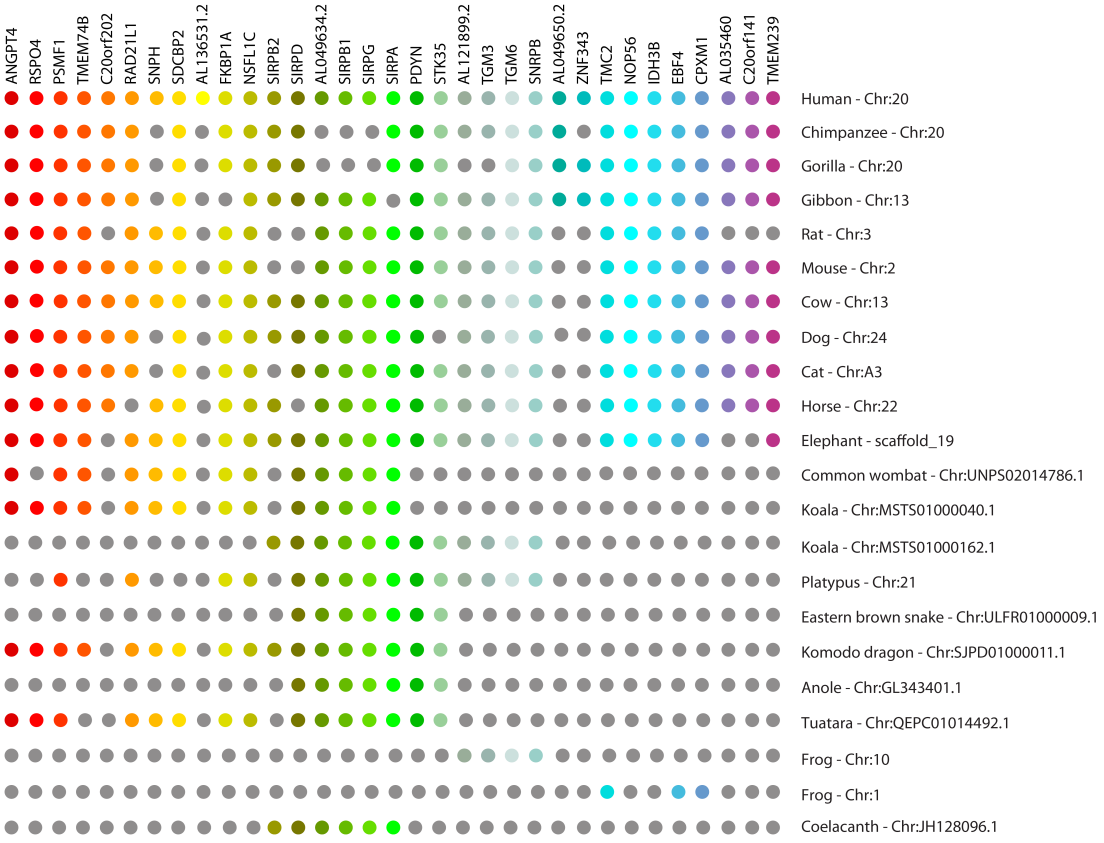


Figure 4.4: Visualization of conserved syntenic blocks using Genomicus. Each gene's presence across the species' chromosomes is represented by a uniquely colored dot, aligned vertically to correlate with its homolog. Gene names are indicated at the top of each column and species names are provided at the end of each row along with scaffold/chromosome number.

Using Genomicus, we found evidence for synteny to human SIRPα (Figure 4.4), across an array of species spanning Sarcopterygii. SIRPG, SIRPD, SIRPB1 and

SIRPB2. SIRPG, SIRPB1, and SIRPD are mostly conserved across all examined mammalian genomes. However, SIRPB2 is absent in Wombats, Platypus, Anoles, Eastern Brown Snakes, and Tuatara. The genes to the right of the SIRP gene cluster in humans that span TCM2 to TMEM239 are restricted entirely to placental mammals. In contrast, genes to the left of the SIRP cluster are widespread among all amniotes. ALO496432 was the only gene present in all major sarcopterygian lineages, being absent only in gorillas and chimpanzees.

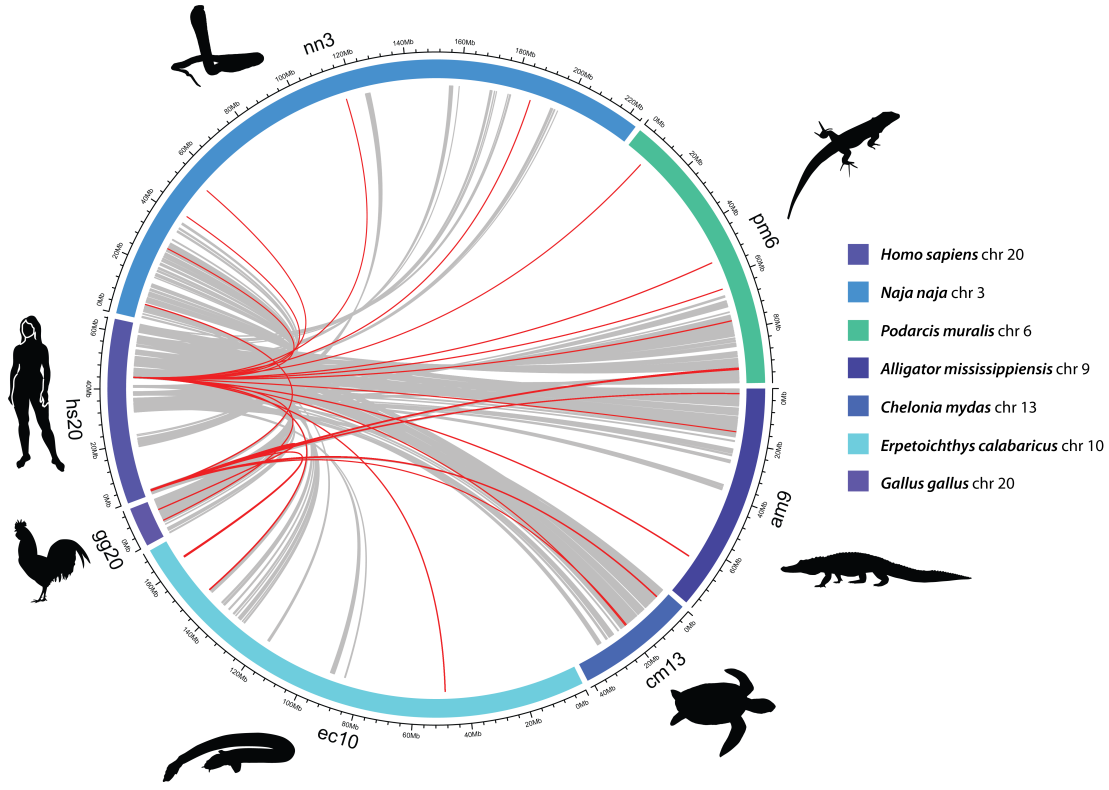


Figure 4.5: Circos plot synteny with SIRPs. The plot shows the comprehensive analysis of synteny among distinct chromosomes, each from a different species. The SIRPs are marked using red lines, and the grey lines denote general synteny.

Zooming out to include a pairwise comparison of all genes located on chromosomes or scaffolds with putative SIRP clusters reveals a high degree of synteny between lineages that span humans, sea turtles, and ray-finned fish (Figure 4.5). This high degree of syntenic relationships between chromosomes suggests shared genomic architecture and potentially reflects conserved sequences or regions of critical

biological function. Critically, the regions with the highest degree of synteny tend to be clustered near the SIRP gene family in all cases. Such a degree of conservation may indicate genes that coexpress or otherwise interact with SIRPs similar to syntenic regions for other genes [233]. Additionally, the conservation at evolutionary divergences that span over 400 million years [143, 231] suggests the presence of core motifs that are shared outside of amniotes. For example, we reveal a high concentration of syntenic relationships between regions containing the human SIRP cluster on chromosome 20 and *Chelonia mydas* chromosome 13 and *Naja naja* chromosome 3. In the case of the reedfish chromosome 10, there are notable syntenic connections, though far fewer in number relative to *Chelonia mydas* and *Naja naja*.

4.4.6 *Syntenic analyses support that the origin of SIRPs is decoupled from the origin of CD47*

The CD47 synteny plot reveals a dense network of syntenic relationships, particularly extensive conserved genomic architecture near the CD47 gene (Figure 4.6). The linearity and absence of crossing over in these connections suggest that these regions have maintained their relative orientation and order since they diverged from a common ancestor. However, interspersed gaps suggest gene losses, loss of evolutionary signal (noise), potential genomic rearrangements, or the presence of species-specific genomic sequences that have diverged significantly. Currently, the syntenic regions around CD47 in other vertebrates remain entirely unexplored. Similarly, the genes that interact with CD47 in other vertebrate lineages are unknown, challenging resolution to the question of whether syntenic regions have coevolved with CD47 to ensure function. Future studies are needed to assess if the genes that govern regulatory mechanisms for the interaction between CD47 and SIRPα for phagocytosis by macrophages are conserved and syntenic. Regardless of outcome, our findings are consistent with an amniote origin of CD47, thereby

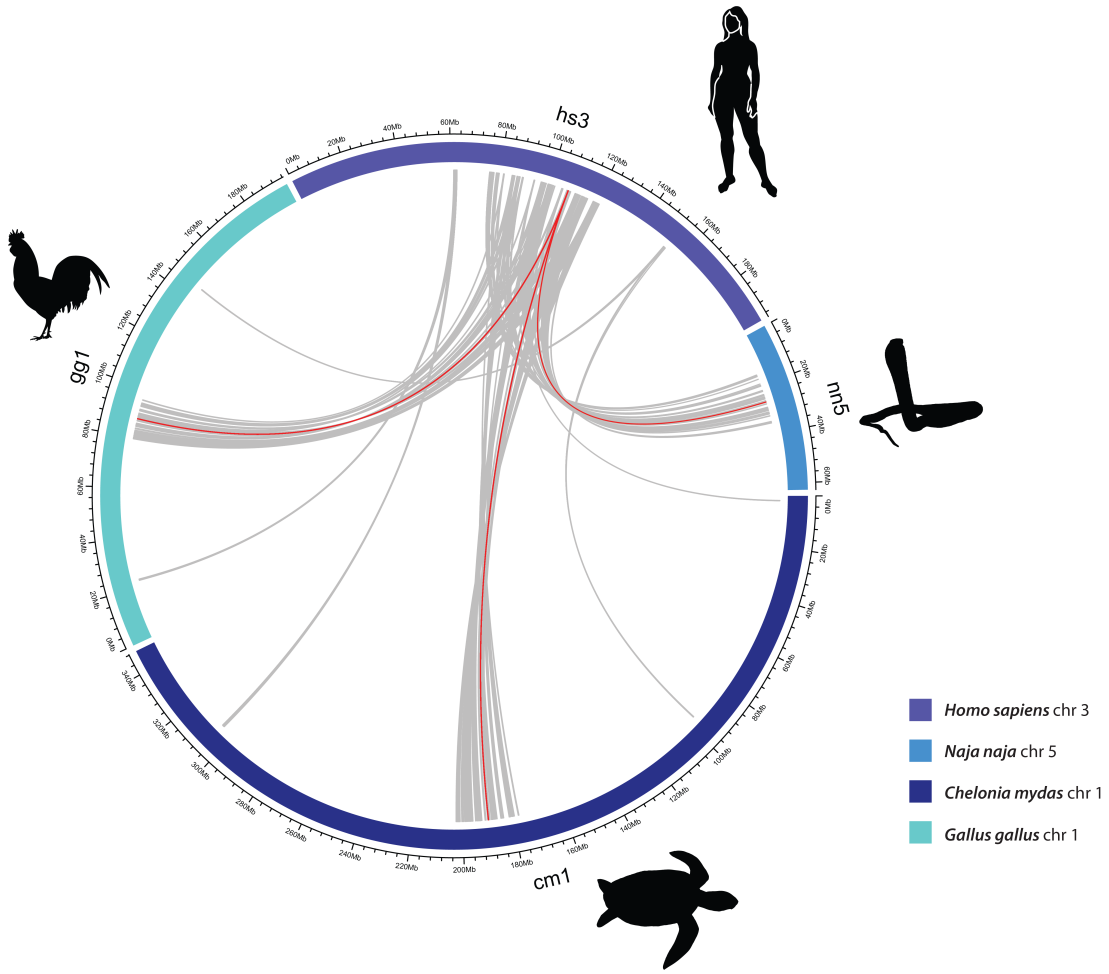


Figure 4.6: Circos plot synteny with CD47. The plot shows the comprehensive analysis of synteny among distinct chromosomes, each from a different species. The CD47 are marked using red lines, and the grey lines denote general synteny.

providing guidance for the choice of models for future functional studies.

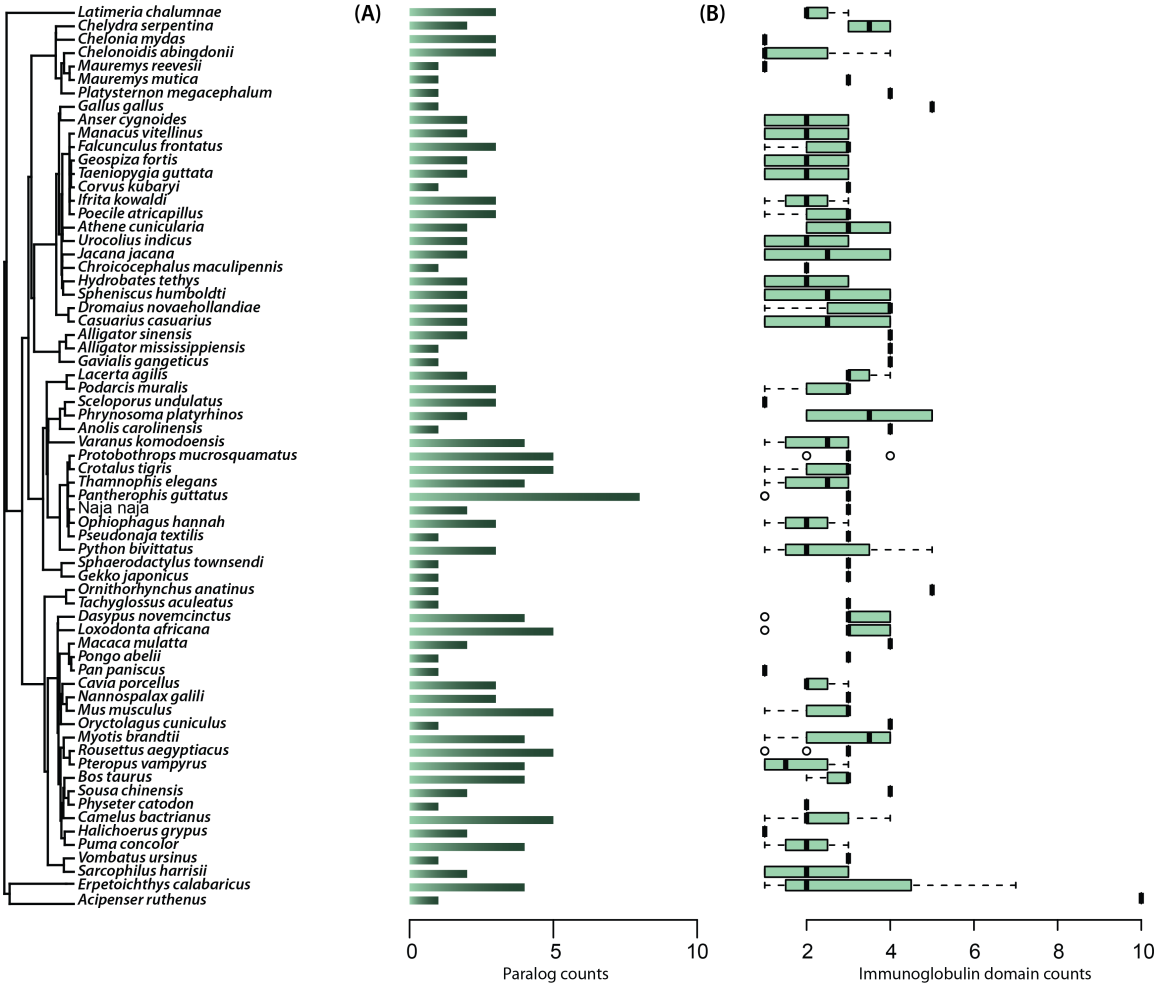
4.4.7 *Summary: A new perspective on the evolutionary history of SIRPs*

Our investigation into the evolutionary history and distribution of signal regulatory protein (SIRP) family members and their interaction partner CD47 across vertebrate lineages provides a new perspective on the origins and diversification of this innate immune receptor gene family and how it may be related to the emergence of the adaptive immune system. By revealing a phylogenetic distribution of SIRPs that spans all major jawed vertebrate lineages, we challenge prior assumptions regarding

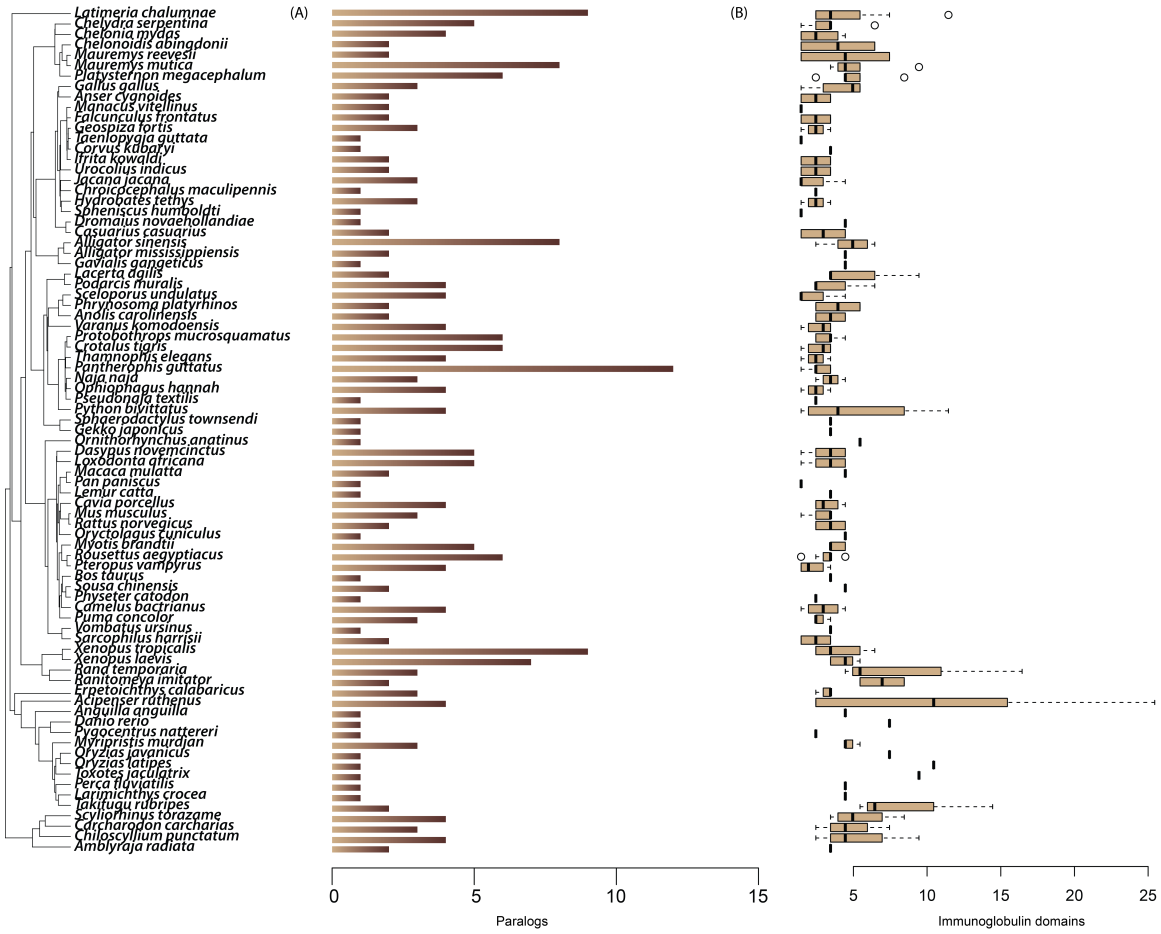
the place of this gene family in the history of V(D)J recombination. Additionally, the restricted presence of CD47 in amniotes suggests that the functionality and interactions of these proteins represent an evolutionary novelty. Given the focus of this interaction as a possible target for cancer suppression [127, 13, 60, 210], whether this novelty is associated with shifts in expression or other functions between amniotes and non-amniotes represents an exciting research prospect with potentially high relevance to cancer biology.

4.5 Supplementary Tables and Figures

4.5.1 Supplementary Figures



Supplemental Figure S4.1: Phylogenetic tree with paralog counts and Ig domain variance using *Bos taurus* SIRPα as query.



Supplemental Figure S4.2: Phylogenetic tree with paralog counts and Ig domain variance using chicken SIRPα as query.

CHAPTER 5: CONCLUSIONS

This dissertation has contributed to the field of vertebrate comparative genomics by addressing critical gaps in the genomic sampling of early diverging fishes. Through our sequencing efforts, notably of *Lepisosteus osseus*, this research has increased our holostean genomic sampling, enhancing our understanding of this pivotal group. The sequencing of *Polypterus bichir* represents an additional effort to bridge the existing sampling gap in Polypterids, enriching the genomic database necessary for comparative studies within this ancient lineage. The integration of the sequenced genomes into a comprehensive comparative dataset, comprising over 100 genomes representative of all principal lineages of ray-finned fishes, enabled a detailed analysis of the impact of the teleost genome duplication (TGD) on the diversification of the ray-finned fish mobilome.

Contrary to the expectation that such a significant genomic event would catalyze a substantial diversification in the mobilome, the findings of this study indicate no marked alteration in the composition of mobile elements post-TGD. This observation supports the emerging consensus within the scientific community that the TGD has not precipitated a pronounced phase of molecular diversification and innovation across a major fraction of extant vertebrates. While genome duplication events have long been heralded as drivers of genomic evolution, the findings from this dissertation suggest a more nuanced role, pointing instead to the importance of lineage-specific adaptations in ray-finned fishes. At the present, limited genomic sampling in these lineages precludes our understanding; however, with initiatives like the 10000 Fish Genomes Project by the Earth BioGenome Project, we are on the cusp of revealing the nuanced roles of transposable elements in evolution. As more genomes are sequenced,

the new frontier will be in interpreting the TE landscape across diverse fish lineages, advancing our grasp of their unique evolutionary trajectories.

Furthermore, this study has unveiled significant insights into the evolution of the immune system by linking an additional gene family with the rise of adaptive immunity. My work suggests that SIRPs may have emerged concurrently with the advent of the adaptive immune system. This finding underscores the ancient and integral role of SIRPs in the immune response.

The decoupling of the emergence of CD47 and SIRP gene families across vertebrate evolution offers another area for future exploration. Specifically, CD47 appears evolutionarily limited to amniotes, whereas SIRPs have far more ancient origins. This divergence raises pivotal questions about the functional roles of SIRPs in these early-diverging groups. Given the established role of SIRPs in modulating immune responses through interaction with CD47 in mammals, their presence in species lacking CD47 suggests potential alternative functions or interactions with different ligands. Investigating these aspects could significantly advance our understanding of immune system evolution by uncovering potentially diverse immunological strategies employed by various vertebrate lineages. Understanding these mechanisms not only contributes to basic biological knowledge but may also inform therapeutic strategies exploiting SIRP-mediated pathways.

REFERENCES

[1] R Hubley & P Green A. F. A. Smit. RepeatMasker. `https://www.repeatmasker.org/webrepeatmaskerhelp.html`. Accessed: 2022-12-14.

[2] S Adams, L J van der Laan, E Vernon-Wilson, C Renardel de Lavalette, E A Döpp, C D Dijkstra, D L Simmons, and T K van den Berg. Signal-regulatory protein is selectively expressed by myeloid and neuronal cells. *J. Immunol.*, 161(4):1853–1859, August 1998.

[3] Andrés Aguilera and Belén Gómez-González. Genome instability: a mechanistic view of its causes and consequences. *Nat. Rev. Genet.*, 9(3):204–217, March 2008.

[4] James S Albert, Victor A Tagliacollo, and Fernando Dagosta. Diversification of neotropical freshwater fishes. *Annu. Rev. Ecol. Evol. Syst.*, 51(1):27–53, November 2020.

[5] Jessica Alföldi, Federica Di Palma, Manfred Grabherr, Christina Williams, Lesheng Kong, Evan Mauceli, Pamela Russell, Craig B Lowe, Richard E Glor, Jacob D Jaffe, David A Ray, Stephane Boissinot, Andrew M Shedlock, Christopher Botka, Todd A Castoe, John K Colbourne, Matthew K Fujita, Ricardo Godinez Moreno, Boudewijn F ten Hallers, David Haussler, Andreas Heger, David Heiman, Daniel E Janes, Jeremy Johnson, Pieter J de Jong, Maxim Y Koriabine, Marcia Lara, Peter A Novick, Chris L Organ, Sally E Peach, Steven Poe, David D Pollock, Kevin de Queiroz, Thomas Sanger, Steve Searle, Jeremy D Smith, Zachary Smith, Ross Swofford, Jason Turner-Maier, Juli Wade, Sarah Young, Amonida Zadissa, Scott V Edwards, Travis C Glenn, Christopher J Schneider, Jonathan B Losos, Eric S Lander, Matthew Breen, Chris P Ponting, and Kerstin Lindblad-Toh. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, 477(7366):587–591, August 2011.

[6] Conchita Alonso, Ricardo Pérez, Pilar Bazaga, and Carlos M Herrera. Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms. *Front. Genet.*, 6:4, January 2015.

[7] Chris T Amemiya, Jessica Alföldi, Alison P Lee, Shaohua Fan, Hervé Philippe, Iain MacCallum, Ingo Braasch, Tereza Manousaki, Igor Schneider, Nicolas Rohner, Chris Organ, Domitille Chalopin, Jeramiah J Smith, Mark Robinson, Rosemary A Dorrington, Marco Gerdol, Bronwen Aken, Maria Assunta Biscotti, Marco Barucca, Denis Baurain, Aaron M Berlin, Gregory L Blatch, Francesco Buonocore, Thorsten Burmester, Michael S Campbell, Adriana Canapa, John P Cannon, Alan Christoffels, Gianluca De Moro, Adrienne L Edkins, Lin Fan, Anna Maria Fausto, Nathalie Feiner, Mariko Forconi, Junaid Gamieldien, Sante

Gnerre, Andreas Gnirke, Jared V Goldstone, Wilfried Haerty, Mark E Hahn, Uljana Hesse, Steve Hoffmann, Jeremy Johnson, Sibel I Karchner, Shigehiro Kuraku, Marcia Lara, Joshua Z Levin, Gary W Litman, Evan Mauceli, Tsutomu Miyake, M Gail Mueller, David R Nelson, Anne Nitsche, Ettore Olmo, Tatsuya Ota, Alberto Pallavicini, Sumir Panji, Barbara Picone, Chris P Ponting, Sonja J Prohaska, Dariusz Przybylski, Nil Ratan Saha, Vydianathan Ravi, Filipe J Ribeiro, Tatjana Sauka-Spengler, Giuseppe Scapigliati, Stephen M J Searle, Ted Sharpe, Oleg Simakov, Peter F Stadler, John J Stegeman, Kenta Sumiyama, Diana Tabbaa, Hakim Tafer, Jason Turner-Maier, Peter van Heusden, Simon White, Louise Williams, Mark Yandell, Henner Brinkmann, Jean-Nicolas Volff, Clifford J Tabin, Neil Shubin, Manfred Schartl, David B Jaffe, Byrappa Venkatesh, Federica Di Palma, Eric S Lander, Axel Meyer, and Kerstin Lindblad-Toh. The african coelacanth genome provides insights into tetrapod evolution. *Nature*, 496(7445):311–316, April 2013.

[8] Samuel Aparicio, Jarrod Chapman, Elia Stupka, Nik Putnam, Jer-Ming Chia, Paramvir Dehal, Alan Christoffels, Sam Rash, Shawn Hoon, Arian Smit, Maarten D Sollewijn Gelpke, Jared Roach, Tania Oh, Isaac Y Ho, Marie Wong, Chris Detter, Frans Verhoef, Paul Predki, Alice Tay, Susan Lucas, Paul Richardson, Sarah F Smith, Melody S Clark, Yvonne J K Edwards, Norman Doggett, Andrey Zharkikh, Sean V Tavtigian, Dmitry Pruss, Mary Barnstead, Cheryl Evans, Holly Baden, Justin Powell, Gustavo Glusman, Lee Rowen, Leroy Hood, Y H Tan, Greg Elgar, Trevor Hawkins, Byrappa Venkatesh, Daniel Rokhsar, and Sydney Brenner. Whole-genome shotgun assembly and analysis of the genome of fugu rubripes. *Science*, 297(5585):1301–1310, August 2002.

[9] Lenin Arias-Rodríguez, Salomón Páramo-Delgadillo, Wilfrido M Contreras-Sánchez, and Carlos A Álvarez-González. Cariotipo del pejelagarto tropical atractosteus tropicus (lepisosteiformes: Lepisosteidae) y variación cromosómica en sus larvas y adultos. *Revista de Biología Tropical*, 57(3):529–539, 2009.

[10] Konrad Bachmann, Olive B Goin, and Coleman J Goin. The nuclear DNA of polypterus palmas. *Copeia*, 1972(2):363–365, 1972.

[11] Pierre Baduel, Basile Leduque, Amandine Ignace, Isabelle Gy, José Gil, Jr, Olivier Loudet, Vincent Colot, and Leandro Quadrana. Genetic and environmental modulation of transposition shapes the evolutionary potential of arabidopsis thaliana. *Genome Biol.*, 22(1):138, May 2021.

[12] A Neil Barclay and Marion H Brown. The SIRP family of receptors and immune regulation. *Nat. Rev. Immunol.*, 6(6):457–464, June 2006.

[13] A Neil Barclay and Timo K Van den Berg. The interaction between signal regulatory protein alpha (SIRPα) and CD47: structure, function, and therapeutic target. *Annu. Rev. Immunol.*, 32:25–50, 2014.

[14] Megan A Barela Hudgell and L Courtney Smith. Sequence diversity, locus structure, and evolutionary history of the SpTransformer genes in the sea urchin genome. *Front. Immunol.*, 12:744783, November 2021.

[15] Jeremy M Beaulieu, Ilia J Leitch, Sunil Patel, Arjun Pendharkar, and Charles A Knight. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytol.*, 179(4):975–986, June 2008.

[16] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 57(1):289–300, January 1995.

[17] Michael H Berry, Amy Holt, Autoosa Salari, Julia Veit, Meike Visel, Joshua Levitz, Krisha Aghi, Benjamin M Gaub, Benjamin Sivyer, John G Flannery, and Ehud Y Isacoff. Restoration of high-sensitivity and adapting vision with a cone opsin. *Nat. Commun.*, 10(1):1221, March 2019.

[18] Xupeng Bi, Kun Wang, Liandong Yang, Hailin Pan, Haifeng Jiang, Qiwei Wei, Miaoquan Fang, Hao Yu, Chenglong Zhu, Yiran Cai, Yuming He, Xiaoni Gan, Honghui Zeng, Daqi Yu, Youan Zhu, Huifeng Jiang, Qiang Qiu, Huanming Yang, Yong E Zhang, Wen Wang, Min Zhu, Shunping He, and Guojie Zhang. Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes. *Cell*, 184(5):1377–1391.e14, March 2021.

[19] Zachary B Bjornson-Hooper, Gabriela K Fragiadakis, Matthew H Spitzer, Han Chen, Deepthi Madhireddy, Kevin Hu, Kelly Lundsten, David R McIlwain, and Garry P Nolan. A comprehensive atlas of immunological differences between humans, mice, and Non-Human primates. *Front. Immunol.*, 13:867015, March 2022.

[20] Simon P Blomberg, Theodore Garland, Jr, and Anthony R Ives. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4):717–745, April 2003.

[21] C Boettiger, D T Lang, and P C Wainwright. rfishbase: exploring, manipulating and visualizing FishBase data from R. *J. Fish Biol.*, 81(6):2030–2039, November 2012.

[22] Sandra Bohn, Brian R Kreiser, Daniel J Daugherty, and Kristopher A Bodine. Natural hybridization of lepisosteids: Implications for managing the alligator gar. *N. Am. J. Fish. Manage.*, 37(2):405–413, March 2017.

[23] Sandra E Bohn, Brian R Kreiser, Damon Williford, Joel Anderson, and William Karel. To all the gar I loved before: range-wide population genetic structure in alligator gar. *Conserv. Genet.*, 24(4):501–521, August 2023.

[24] Astrid Böhne, Frédéric Brunet, Delphine Galiana-Arnoux, Christina Schultheis, and Jean-Nicolas Volff. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res.*, 16(1):203–215, 2008.

[25] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[26] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[27] Jonathan P Bollback. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics*, 7:88, February 2006.

[28] Ingo Braasch, Andrew R Gehrke, Jeramiah J Smith, Kazuhiko Kawasaki, Tereza Manousaki, Jeremy Pasquier, Angel Amores, Thomas Desvignes, Peter Batzel, Julian Catchen, Aaron M Berlin, Michael S Campbell, Daniel Barrell, Kyle J Martin, John F Mulley, Vydianathan Ravi, Alison P Lee, Tetsuya Nakamura, Domitille Chalopin, Shaohua Fan, Dustin Wcisel, Cristian Cañestro, Jason Sydes, Felix E G Beaudry, Yi Sun, Jana Hertel, Michael J Beam, Mario Fasold, Mikio Ishiyama, Jeremy Johnson, Steffi Kehr, Marcia Lara, John H Letaw, Gary W Litman, Ronda T Litman, Masato Mikami, Tatsuya Ota, Nil Ratan Saha, Louise Williams, Peter F Stadler, Han Wang, John S Taylor, Quenton Fontenot, Allyse Ferrara, Stephen M J Searle, Bronwen Aken, Mark Yandell, Igor Schneider, Jeffrey A Yoder, Jean-Nicolas Volff, Axel Meyer, Chris T Amemiya, Byrappa Venkatesh, Peter W H Holland, Yann Guiguen, Julien Bobe, Neil H Shubin, Federica Di Palma, Jessica Alföldi, Kerstin Lindblad-Toh, and John H Postlethwait. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.*, 48(4):427–437, April 2016.

[29] Ralf Britz and G David Johnson. On the homology of the posteriormost gill arch in polypterids (cladistia, actinopterygii). *Zool. J. Linn. Soc.*, 138(4):495–503, September 2008.

[30] Ralf Britz and G David Johnson. Occipito-vertebral fusion in actinopterygians: conjecture, myth and reality. part 1: non-teleosts. *Origin and Phylogenetic Interrelationships of Teleosts Honoring Gloria Arratia*, 2010.

[31] G P Brooke, K R Parsons, and C J Howard. Cloning of two members of the SIRP alpha family of protein tyrosine phosphatase binding proteins in cattle that are expressed on monocytes and a subpopulation of dendritic cells and which mediate binding to CD4 T cells. *Eur. J. Immunol.*, 28(1):1–11, January 1998.

[32] Gary Brooke, Joanna D Holbrook, Marion H Brown, and A Neil Barclay. Human lymphocytes interact directly with CD47 through a novel member of the signal regulatory protein (SIRP) family. *J. Immunol.*, 173(4):2562–2570, August 2004.

[33] Chase D Brownstein, Daemin Kim, Oliver D Orr, Gabriela M Hogue, Bryn H Tracy, M Worth Pugh, Randal Singer, Chelsea Myles-McBurney, Jon Michael

Mollish, Jeffrey W Simmons, Solomon R David, Gregory Watkins-Colwell, Eva A Hoffman, and Thomas J Near. Hidden species diversity in an iconic living fossil vertebrate. *Biol. Lett.*, 18(11):20220395, November 2022.

[34] Tomáš Brůna, Katharina J Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform*, 3(1):lqaa108, March 2021.

[35] Frédéric G Brunet, Hugues Roest Crollius, Mathilde Paris, Jean-Marc Aury, Patricia Gibert, Olivier Jaillon, Vincent Laudet, and Marc Robinson-Rechavi. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.*, 23(9):1808–1816, September 2006.

[36] Donald M Bryant, Kimberly Johnson, Tia DiTommaso, Timothy Tickle, Matthew Brian Couger, Duygu Payzin-Dogru, Tae J Lee, Nicholas D Leigh, Tzu-Hsing Kuo, Francis G Davis, Joel Bateman, Sevara Bryant, Anna R Guzikowski, Stephanie L Tsai, Steven Coyne, William W Ye, Robert M Freeman, Jr, Leonid Peshkin, Clifford J Tabin, Aviv Regev, Brian J Haas, and Jessica L Whited. A Tissue-Mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Rep.*, 18(3):762–776, January 2017.

[37] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12(1):59–60, November 2014.

[38] Irina V Bure, Marina V Nemtsova, and Ekaterina B Kuznetsova. Histone modifications and Non-Coding RNAs: Mutual epigenetic regulation and role in pathogenesis. *Int. J. Mol. Sci.*, 23(10), May 2022.

[39] Kathleen H Burns. Transposable elements in cancer. *Nat. Rev. Cancer*, 17(7):415–424, July 2017.

[40] Floréal Cabanettes and Christophe Klopp. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, 6:e4958, June 2018.

[41] Charles Cant. Association of signal-regulatory proteins β with KARAP/DAP12. *Eur. J. Immunol.*, January 2000.

[42] P Capy, G Gasperi, C Biémont, and C Bazin. Stress and transposable elements: co-evolution or useful parasites? *Heredity*, 85 (Pt 2):101–106, August 2000.

[43] Federica Carducci, Marco Barucca, Adriana Canapa, Elisa Carotti, and Maria Assunta Biscotti. Mobile elements in Ray-Finned fish genomes. *Life*, 10(10), September 2020.

[44] Kara B Carlson, Cameron Nguyen, Dustin J Wcisel, Jeffrey A Yoder, and Alex Dornburg. Ancient fish lineages illuminate toll-like receptor diversification in early vertebrate evolution. *Immunogenetics*, 75(5):465–478, October 2023.

[45] Kara B Carlson, Dustin J Wcisel, Hayley D Ackerman, Jessica Romanet, Emily F Christiansen, Jennifer N Niemuth, Christina Williams, Matthew Breen, Michael K Stoskopf, Alex Dornburg, and Jeffrey A Yoder. Transcriptome annotation reveals minimal immunogenetic diversity among wyoming toads, anaxyrus baxteri. *Conserv. Genet.*, April 2022.

[46] Irene Castellano-Pellicena, Natallia E Uzunbajakava, Charles Mignon, Bianca Raafs, Vladimir A Botchkarev, and M Julie Thornton. Does blue light restore human epidermal barrier function via activation of opsin during cutaneous wound healing? *Lasers Surg. Med.*, 51(4):370–382, April 2019.

[47] Rodrigo Catalán, Mario Orozco-Morales, Norma Y Hernández-Pedro, Alberto Guijosa, Ana L Colín-González, Federico Ávila-Moreno, and Oscar Arrieta. CD47-SIRPα axis as a biomarker and therapeutic target in cancer: Current perspectives and future challenges in nonsmall cell lung cancer. *J Immunol Res*, 2020:9435030, September 2020.

[48] Domitille Chalopin, Shaohua Fan, Oleg Simakov, Axel Meyer, Manfred Schartl, and Jean-Nicolas Volff. Evolutionary active transposable elements in the genome of the coelacanth. *J. Exp. Zool. B Mol. Dev. Evol.*, 322(6):322–333, September 2014.

[49] Domitille Chalopin, Magali Naville, Floriane Plard, Delphine Galiana, and Jean-Nicolas Volff. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol. Evol.*, 7(2):567–580, January 2015.

[50] Zhuo Chen and John J Wiens. The origins of acoustic communication in vertebrates. *Nat. Commun.*, 11(1):369, January 2020.

[51] Peilin Cheng, Yu Huang, Hao Du, Chuangju Li, Yunyun Lv, Rui Ruan, Huan Ye, Chao Bian, Xinxin You, Junmin Xu, Xufang Liang, Qiong Shi, and Qiwei Wei. Draft genome and complete Hox-Cluster characterization of the sterlet (acipenser ruthenus). *Front. Genet.*, 10:776, September 2019.

[52] Karen D Crow, Günter P Wagner, and SMBE Tri-National Young Investigators. Proceedings of the SMBE Tri-National young investigators' workshop 2005. what is the role of genome duplication in the evolution of complexity and diversity? *Mol. Biol. Evol.*, 23(5):887–892, May 2006.

[53] Jacob M Daane, Alex Dornburg, Patrick Smits, Daniel J MacGuigan, M Brent Hawkins, Thomas J Near, H William Detrich, III, and Matthew P Harris. Historical contingency shapes adaptive radiation in antarctic fishes. *Nature Ecology & Evolution*, 3(7):1102–1109, June 2019.

[54] Charles Darwin. *The Origin of Species: By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Cambridge University Press, July 2009.

[55] Matthew P Davis, Nancy I Holcroft, Edward O Wiley, John S Sparks, and W Leo Smith. Species-specific bioluminescence facilitates speciation in the deep sea. *Mar. Biol.*, 161(5):1139–1148, May 2014.

[56] Johan G de Boer, Ryosuke Yazawa, William S Davidson, and Ben F Koop. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics*, 8:422, November 2007.

[57] Rishi De-Kayne, Oliver M Selz, David A Marques, David Frei, Ole Seehausen, and Philine G D Feulner. Genomic architecture of adaptive radiation and hybridization in alpine whitefish. *Nat. Commun.*, 13(1):1–13, August 2022.

[58] Sebastian Deorowicz, Agnieszka Debudaj-Grabysz, and Adam Gudyś. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.*, 6:33964, September 2016.

[59] J Dietrich, M Cella, M Seiffert, H J Bühring, and M Colonna. Cutting edge: signal-regulatory protein beta 1 is a DAP12-associated activating receptor expressed in myeloid cells. *J. Immunol.*, 164(1):9–12, January 2000.

[60] Nazli Dizman and Elizabeth I Buchbinder. Cancer therapy targeting CD47/SIRPα. *Cancers*, 13(24), December 2021.

[61] Steven Dodsworth, Mark W Chase, Laura J Kelly, Ilia J Leitch, Jiří Macas, Petr Novák, Mathieu Piednoël, Hanna Weiss-Schneeweiss, and Andrew R Leitch. Genomic repeat abundances contain phylogenetic signal. *Syst. Biol.*, 64(1):112–126, January 2015.

[62] Alex Dornburg, Sarah Federman, April D Lamb, Christopher D Jones, and Thomas J Near. Cradles and museums of antarctic teleost biodiversity. *Nat Ecol Evol*, 1(9):1379–1384, September 2017.

[63] Alex Dornburg, Rittika Mallik, Zheng Wang, Moisés A Bernal, Brian Thompson, Elspeth A Bruford, Daniel W Nebert, Vasilis Vasiliou, Laurel R Yohe, Jeffrey A Yoder, and Jeffrey P Townsend. Placing human gene families into their evolutionary context. *Hum. Genomics*, 16(1):56, November 2022.

[64] Alex Dornburg and Thomas J Near. The emerging phylogenetic perspective on the evolution of actinopterygian fishes. *Annu. Rev. Ecol. Evol. Syst.*, 52(1):427–452, November 2021.

[65] Alex Dornburg, Jeffrey P Townsend, Matt Friedman, and Thomas J Near. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. *BMC Evol. Biol.*, 14:169, August 2014.

[66] Alex Dornburg, Dustin J Wcisel, Katerina Zapfe, Emma Ferraro, Lindsay Roupe-Abrams, Andrew W Thompson, Ingo Braasch, Tatsuya Ota, and Jeffrey A Yoder. Holosteans contextualize the role of the teleost genome duplica-

tion in promoting the rise of evolutionary novelties in the ray-finned fish innate immune system. *Immunogenetics*, 73(6):479–497, December 2021.

[67] Alex Dornburg, Dustin J Wcisel, Katerina Zapfe, Emma Ferraro, Lindsay Roupe-Abrams, Andrew W Thompson, Ingo Braasch, Tatsuya Ota, and Jeffrey A Yoder. Holosteans contextualize the role of the teleost genome duplication in promoting the rise of evolutionary novelties in the ray-finned fish innate immune system. *Immunogenetics*, 73(6):479–497, December 2021.

[68] Alex Dornburg and Jeffrey A Yoder. On the relationship between extant innate immune receptors and the evolutionary origins of jawed vertebrate adaptive immunity. *Immunogenetics*, 74(1):111–128, February 2022.

[69] Kang Du, Matthias Stöck, Susanne Kneitz, Christophe Klopp, Joost M Woltering, Mateus Contar Adolfi, Romain Feron, Dmitry Prokopov, Alexey Makunin, Ilya Kichigin, Cornelia Schmidt, Petra Fischer, Heiner Kuhl, Sven Wuertz, Jörn Gessner, Werner Kloas, Cédric Cabau, Carole Iampietro, Hugues Parrinello, Chad Tomlinson, Laurent Journot, John H Postlethwait, Ingo Braasch, Vladimir Trifonov, Wesley C Warren, Axel Meyer, Yann Guiguen, and Manfred Schartl. The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. *Nat Ecol Evol*, 4(6):841–852, June 2020.

[70] Entsar Eladl, Rosemarie Tremblay-LeMay, Nasrin Rastgoo, Rumina Musani, Wenming Chen, Aijun Liu, and Hong Chang. Role of CD47 in hematological malignancies. *J. Hematol. Oncol.*, 13(1):96, July 2020.

[71] Francesca Faillaci, Fabiola Milosa, Rosina Maria Critelli, Elena Turola, Filippo Schepis, and Erica Villa. Obese zebrafish: A small fish for a major human health condition. *Animal Model Exp Med*, 1(4):255–265, December 2018.

[72] Guangyi Fan, Yue Song, Liandong Yang, Xiaoyun Huang, Suyu Zhang, Mengqi Zhang, Xianwei Yang, Yue Chang, He Zhang, Yongxin Li, Shanshan Liu, Lili Yu, Jeffery Chu, Inge Seim, Chenguang Feng, Thomas J Near, Rod A Wing, Wen Wang, Kun Wang, Jing Wang, Xun Xu, Huanming Yang, Xin Liu, Nansheng Chen, and Shunping He. Initial data release and announcement of the 10,000 fish genomes project (Fish10K). *Gigascience*, 9(8), August 2020.

[73] Justina X Feng and Nicole C Riddle. Epigenetics and genome stability. *Mamm. Genome*, 31(5-6):181–195, June 2020.

[74] Yongyi Feng, Chunliu Huang, Yingzhao Wang, and Jun Chen. SIRPα: A key player in innate immunity. *Eur. J. Immunol.*, 53(11):e2350375, November 2023.

[75] D J Finnegan. Transposable elements. *Curr. Opin. Genet. Dev.*, 2(6):861–867, December 1992.

[76] Jullien M Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G Clark, Cédric Feschotte, and Arian F Smit. RepeatModeler2 for automated

genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.*, 117(17):9451–9457, April 2020.

[77] Kaien Fujino, Shin-Nosuke Hashida, Takashi Ogawa, Tomoko Natsume, Takako Uchiyama, Tetsuo Mikami, and Yuji Kishima. Temperature controls nuclear import of tam3 transposase in antirrhinum. *Plant J.*, 65(1):146–155, January 2011.

[78] Naoko T Fujito and Masaru Nonaka. Highly divergent dimorphic alleles of the proteasome subunit beta type-8 (PSMB8) gene of the bichir polypterus senegalus: implication for evolution of the PSMB8 gene of jawed vertebrates. *Immunogenetics*, 64(6):447–453, June 2012.

[79] Lars Gabriel, Katharina J Hoff, Tomáš Brůna, Mark Borodovsky, and Mario Stanke. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics*, 22(1):566, November 2021.

[80] James D Galbraith, Alastair J Ludington, Kate L Sanders, Timothy G Amos, Vicki A Thomson, Daniel Enosi Tuipulotu, Nathan Dunstan, Richard J Edwards, Alexander Suh, and David L Adelson. Horizontal transposon transfer and its implications for the ancestral ecology of hydrophiine snakes. *Genes*, 13(2), January 2022.

[81] Rais A Ganai and Erik Johansson. DNA Replication-A matter of fidelity. *Mol. Cell*, 62(5):745–755, June 2016.

[82] Ava Ghezelayagh, Richard C Harrington, Edward D Burress, Matthew A Campbell, Janet C Buckner, Prosanta Chakrabarty, Jessica R Glass, W Tyler McCraney, Peter J Unmack, Christine E Thacker, Michael E Alfaro, Sarah T Friedman, William B Ludt, Peter F Cowman, Matt Friedman, Samantha A Price, Alex Dornburg, Brant C Faircloth, Peter C Wainwright, and Thomas J Near. Prolonged morphological expansion of spiny-rayed fishes following the end-cretaceous. *Nat Ecol Evol*, 6(8):1211–1220, August 2022.

[83] Stella M K Glasauer and Stephan C F Neuhauss. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol. Genet. Genomics*, 289(6):1045–1060, December 2014.

[84] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–652, May 2011.

[85] Zuguang Gu, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. circlize implements and enhances circular visualization in R, 2014.

[86] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman, and Aviv Regev. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.*, 8(8):1494–1512, August 2013.

[87] Shijie Hao, Kai Han, Lingfeng Meng, Xiaoyun Huang, Chengcheng Shi, Mengqi Zhang, Yilin Wang, Qun Liu, Yaolei Zhang, Inge Seim, Xun Xu, Xin Liu, and Guangyi Fan. Three genomes of osteoglossidae shed light on ancient teleost evolution. January 2020.

[88] Seyed Mohammad Gheibi Hayat, Vanessa Bianconi, Matteo Pirro, Mahmoud R Jaafari, Mahdi Hatamipour, and Amirhossein Sahebkar. CD47: role in the immune system and application to cancer therapy. *Cell. Oncol.*, 43(1):19–30, February 2020.

[89] Kathrin Helfenrath, Markus Sauer, Michelle Kamga, Michelle Wisniewsky, Thorsten Burmester, and Andrej Fabrizius. The more, the merrier? multiple myoglobin genes in fish species, especially in gray bichir (polypterus senegalus) and reedfish (erpetoichthys calabaricus). *Genome Biol. Evol.*, 13(7), April 2021.

[90] Uffe Hellsten, Richard M Harland, Michael J Gilchrist, David Hendrix, Jerzy Jurka, Vladimir Kapitonov, Ivan Ovcharenko, Nicholas H Putnam, Shengqiang Shu, Leila Taher, Ira L Blitz, Bruce Blumberg, Darwin S Dichmann, Inna Dubchak, Enrique Amaya, John C Detter, Russell Fletcher, Daniela S Gerhard, David Goodstein, Tina Graves, Igor V Grigoriev, Jane Grimwood, Takeshi Kawashima, Erika Lindquist, Susan M Lucas, Paul E Mead, Therese Mitros, Hajime Ogino, Yuko Ohta, Alexander V Poliakov, Nicolas Pollet, Jacques Robert, Asaf Salamov, Amy K Sater, Jeremy Schmutz, Astrid Terry, Peter D Vize, Wesley C Warren, Dan Wells, Andrea Wills, Richard K Wilson, Lyle B Zimmerman, Aaron M Zorn, Robert Grainger, Timothy Grammer, Mustafa K Khokha, Paul M Richardson, and Daniel S Rokhsar. The genome of the western clawed frog xenopus tropicalis. *Science*, 328(5978):633–636, April 2010.

[91] Alison B Hickman and Fred Dyda. Mechanisms of DNA transposition. In *Mobile DNA III*, pages 529–553. ASM Press, Washington, DC, USA, May 2015.

[92] Alison B Hickman and Fred Dyda. Mechanisms of DNA transposition. *Microbiol Spectr*, 3(2):MDNA3–0034–2014, April 2015.

[93] Simone Hoegg, Henner Brinkmann, John S Taylor, and Axel Meyer. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.*, 59(2):190–203, August 2004.

[94] Kerstin Howe, Matthew D Clark, Carlos F Torroja, James Torrance, Camille Berthelot, Matthieu Muffato, John E Collins, Sean Humphray, Karen McLaren, Lucy Matthews, Stuart McLaren, Ian Sealy, Mario Caccamo, Carol Churcher, Carol Scott, Jeffrey C Barrett, Romke Koch, Gerd-Jörg Rauch, Simon White, William Chow, Britt Kilian, Leonor T Quintais, José A Guerra-Assunção, Yi Zhou, Yong Gu, Jennifer Yen, Jan-Hinnerk Vogel, Tina Eyre, Seth Redmond, Ruby Banerjee, Jianxiang Chi, Beiyuan Fu, Elizabeth Langley, Sean F Maguire, Gavin K Laird, David Lloyd, Emma Kenyon, Sarah Donaldson, Harminder Sehra, Jeff Almeida-King, Jane Loveland, Stephen Trevanion, Matt Jones, Mike Quail, Dave Willey, Adrienne Hunt, John Burton, Sarah Sims, Kirsten McLay, Bob Plumb, Joy Davis, Chris Clee, Karen Oliver, Richard Clark, Clare Riddle, David Elliott, Glen Threadgold, Glenn Harden, Darren Ware, Sharmin Begum, Beverley Mortimore, Giselle Kerry, Paul Heath, Benjamin Phillimore, Alan Tracey, Nicole Corby, Matthew Dunn, Christopher Johnson, Jonathan Wood, Susan Clark, Sarah Pelan, Guy Griffiths, Michelle Smith, Rebecca Glithero, Philip Howden, Nicholas Barker, Christine Lloyd, Christopher Stevens, Joanna Harley, Karen Holt, Georgios Panagiotidis, Jamieson Lovell, Helen Beasley, Carl Henderson, Daria Gordon, Katherine Auger, Deborah Wright, Joanna Collins, Claire Raisen, Lauren Dyer, Kenric Leung, Lauren Robertson, Kirsty Ambridge, Daniel Leongamornlert, Sarah McGuire, Ruth Gilderthorp, Coline Griffiths, Deepa Manthravadi, Sarah Nichol, Gary Barker, Siobhan Whitehead, Michael Kay, Jacqueline Brown, Clare Murnane, Emma Gray, Matthew Humphries, Neil Sycamore, Darren Barker, David Saunders, Justene Wallis, Anne Babbage, Sian Hammond, Maryam Mashreghi-Mohammadi, Lucy Barr, Sancha Martin, Paul Wray, Andrew Ellington, Nicholas Matthews, Matthew Ellwood, Rebecca Woodmansey, Graham Clark, James D Cooper, Anthony Tromans, Darren Grafham, Carl Skuce, Richard Pandian, Robert Andrews, Elliot Harrison, Andrew Kimberley, Jane Garnett, Nigel Fosker, Rebekah Hall, Patrick Garner, Daniel Kelly, Christine Bird, Sophie Palmer, Ines Gehring, Andrea Berger, Christopher M Dooley, Zübeyde Ersan-Ürün, Cigdem Eser, Horst Geiger, Maria Geisler, Lena Karotki, Anette Kirn, Judith Konantz, Martina Konantz, Martina Oberländer, Silke Rudolph-Geiger, Mathias Teucke, Christa Lanz, Günter Raddatz, Kazutoyo Osoegawa, Baoli Zhu, Amanda Rapp, Sara Widaa, Cordelia Langford, Fengtang Yang, Stephan C Schuster, Nigel P Carter, Jennifer Harrow, Zemin Ning, Javier Herrero, Steve M J Searle, Anton Enright, Robert Geisler, Ronald H A Plasterk, Charles Lee, Monte Westerfield, Pieter J de Jong, Leonard I Zon, Christiane Nüsslein-Volhard, Tim J P Hubbard, Hugues Roest Crollius, Jane Rogers, and Derek L Stemple. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496(7446):498–503, April 2013.

[95] Aurélie Hua-Van, Arnaud Le Rouzic, Thibaud S Boutin, Jonathan Filée, and Pierre Capy. The struggle for life of the genome's selfish architects. *Biol. Direct*, 6:19, March 2011.

[96] Lily C Hughes, Guillermo Ortí, Yu Huang, Ying Sun, Carole C Baldwin, Andrew W Thompson, Dahiana Arcila, Ricardo Betancur-R, Chenhong Li, Leandro Becker, Nicolás Bellora, Xiaomeng Zhao, Xiaofeng Li, Min Wang, Chao Fang, Bing Xie, Zhuocheng Zhou, Hai Huang, Songlin Chen, Byrappa Venkatesh, and Qiong Shi. Comprehensive phylogeny of ray-finned fishes (actinopterygii) based on transcriptomic and genomic data. *Proc. Natl. Acad. Sci. U. S. A.*, 115(24):6249–6254, June 2018.

[97] Aurora Irene Idilli, Francesca Precazzini, Maria Caterina Mione, and Viviana Anelli. Zebrafish in translational cancer research: Insight into leukemia, melanoma, glioma and endocrine tumor biology. *Genes*, 8(9), September 2017.

[98] International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, December 2004.

[99] Natalia V Ivanova, Tyler S Zemlak, Robert H Hanner, and Paul D N Hebert. Universal primer cocktails for fish DNA barcoding. *Mol. Ecol. Notes*, 7(4):544–548, July 2007.

[100] P Jiang, C F Lagenaur, and V Narayanan. Integrin-associated protein is a ligand for the P84 neural adhesion molecule. *J. Biol. Chem.*, 274(2):559–562, January 1999.

[101] Shuang Jiang, Danying Cai, Yongwang Sun, and Yuanwen Teng. Isolation and characterization of putative functional long terminal repeat retrotransposons in the pyrus genome. *Mob. DNA*, 7:1, January 2016.

[102] J Jurka, V V Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110(1-4):462–467, 2005.

[103] Subha Kalyaanamoorthy, Bui Quang Minh, Thomas K F Wong, Arndt von Haeseler, and Lars S Jermiin. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*, 14(6):587–589, June 2017.

[104] V V Kapitonov and J Jurka. Chapaev-a novel superfamily of DNA transposons. *Repbase Reports*, 7(9):774–781, 2007.

[105] Masahiro Kasahara, Kiyoshi Naruse, Shin Sasaki, Yoichiro Nakatani, Wei Qu, Budrul Ahsan, Tomoyuki Yamada, Yukinobu Nagayasu, Koichiro Doi, Yasuhiro Kasai, Tomoko Jindo, Daisuke Kobayashi, Atsuko Shimada, Atsushi Toyoda, Yoko Kuroki, Asao Fujiyama, Takashi Sasaki, Atsushi Shimizu, Shuichi Asakawa, Nobuyoshi Shimizu, Shin-Ichi Hashimoto, Jun Yang, Yongjun Lee, Kouji Matsushima, Sumio Sugano, Mitsuru Sakaizumi, Takanori Narita, Kazuko Ohishi, Shinobu Haga, Fumiko Ohta, Hisayo Nomoto, Keiko Nogata, Tomomi Morishita, Tomoko Endo, Tadasu Shin-I, Hiroyuki Takeda, Shinichi

Morishita, and Yuji Kohara. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145):714–719, June 2007.

[106] Erin S Kelleher, Daniel A Barbash, and Justin P Blumenstiel. Taming the turmoil within: New insights on the containment of transposable elements. *Trends Genet.*, 36(7):474–489, July 2020.

[107] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, 37(8):907–915, August 2019.

[108] Yuki Kimura and Masato Nikaido. Conserved keratin gene clusters in ancient fish: An evolutionary seed for terrestrial adaptation. *Genomics*, 113(1 Pt 2):1120–1128, January 2021.

[109] Dusan Kordis, Nika Lovsin, and Franc Gubensek. Phylogenomic analysis of the L1 retrotransposons in deuterostomia. *Syst. Biol.*, 55(6):886–901, December 2006.

[110] A Krogh, B Larsson, G von Heijne, and E L Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.*, 305(3):567–580, January 2001.

[111] E Yu Krysanov and A S Golubtsov. Karyotypes of four fish species from the nile and Omo-Turkana basins in ethiopia. *J. Ichthyol.*, 54(10):889–892, December 2014.

[112] Sudhir Kumar, Michael Suleski, Jack M Craig, Adrienne E Kasprowicz, Maxwell Sanderford, Michael Li, Glen Stecher, and S Blair Hedges. TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.*, 39(8):msac174, August 2022.

[113] S J Lamberth and J K Turpie. The role of estuaries in south african fisheries: Economic importance and management implications. *Afr. J. Mar. Sci.*, 25(1):131–157, June 2003.

[114] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L

Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[115] Anders Larsson. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278, November 2014.

[116] Robert Lehmann, Aleš Kovařík, Konrad Ocalewicz, Lech Kirtiklis, Andrea Zuccolo, Jesper N Tegner, Josef Wanzenböck, Louis Bernatchez, Dunja K Lamatsch, and Radka Symonová. DNA transposon expansion is associated with genome size increase in mudminnows. *Genome Biol. Evol.*, 13(10), October 2021.

[117] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A

Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.

[118] Rosalyn Lo, Katherine E Dougan, Yibi Chen, Sarah Shah, Debashish Bhattacharya, and Cheong Xin Chan. Alignment-Free analysis of Whole-Genome sequences from symbiodiniaceae reveals different phylogenetic signals in distinct regions. *Front. Plant Sci.*, 13:815714, April 2022.

[119] Alexandre Lomsadze, Paul D Burns, and Mark Borodovsky. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.*, 42(15):e119, September 2014.

[120] Michael Lynch. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. U. S. A.*, 104 Suppl 1(Suppl 1):8597–8604, May 2007.

[121] Ravindra Majeti, Mark P Chao, Ash A Alizadeh, Wendy W Pang, Siddhartha Jaiswal, Kenneth D Gibbs, Jr, Nico van Rooijen, and Irving L Weissman. CD47 is an adverse prognostic factor and therapeutic antibody target on human acute myeloid leukemia stem cells. *Cell*, 138(2):286–299, July 2009.

[122] Rittika Mallik, Kara B Carlson, Dustin J Wcisel, Michael Fisk, Jeffrey A Yoder, and Alex Dornburg. A chromosome-level genome assembly of longnose gar, lepisosteus osseus. *G3*, 13(7), July 2023.

[123] Rittika Mallik, Dustin J Wcisel, Thomas J Near, Jeffrey A Yoder, and Alex Dornburg. Investigating the impact of whole genome duplication on transposable element evolution in ray-finned fishes. December 2023.

[124] Mosè Manni, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, 38(10):4647–4654, September 2021.

[125] Aron Marchler-Bauer, Shennan Lu, John B Anderson, Farideh Chitsaz, Myra K Derbyshire, Carol DeWeese-Scott, Jessica H Fong, Lewis Y Geer, Renata C Geer, Noreen R Gonzales, Marc Gwadz, David I Hurwitz, John D Jackson, Zhaoxi Ke, Christopher J Lanczycki, Fu Lu, Gabriele H Marchler, Mikhail Mullokandov, Marina V Omelchenko, Cynthia L Robertson, James S Song, Narmada Thanki, Roxanne A Yamashita, Dachuan Zhang, Naigong Zhang, Chanjuan Zheng, and Stephen H Bryant. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.*, 39(Database issue):D225–9, January 2011.

[126] Carolina A Martinez-Gutierrez and Frank O Aylward. Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. *PLoS Genet.*, 18(5):e1010220, May 2022.

[127] Hanke L Matlung, Katka Szilagyi, Neil A Barclay, and Timo K van den Berg. The CD47-SIRPα signaling axis as an innate immune checkpoint in cancer. *Immunol. Rev.*, 276(1):145–164, March 2017.

[128] M A Matzke and A J Matzke. Polyploidy and transposons. *Trends Ecol. Evol.*, 13(6):241, June 1998.

[129] B McClintock. The significance of responses of the genome to challenge. *Science*, 226(4676):792–801, November 1984.

[130] Masato Mikami, Toshinao Ineno, Andrew W Thompson, Ingo Braasch, Mikio Ishiyama, and Kazuhiko Kawasaki. Convergent losses of SCPP genes and ganoid scales among non-teleost actinopterygians. *Gene*, 811:146091, February 2022.

[131] Elizabeth Christina Miller, Christopher M Martinez, Sarah T Friedman, Peter C Wainwright, Samantha A Price, and Luke Tornabene. Alternating regimes of shallow and deep-sea diversification explain a species-richness paradox in marine fishes. *Proceedings of the National Academy of Sciences*, 119(43):e2123544119, 2022.

[132] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, 37(5):1530–1534, May 2020.

[133] Maria Alessandra Morescalchi, Innocenza Liguori, Lucia Rocco, and Vincenzo Stingo. Karyotypic characterization and genomic organization of the 5S rDNA in erpetoichthys calabaricus (osteichthyes, polypteridae). *Genetica*, 131(2):209–216, October 2007.

[134] Timo Moritz and Ralf Britz. Revision of the extant polypteridae (actinopterygii: Cladistia). *Ichthyol. Explor. Freshw.*, 2019.

[135] Yuuta Moriyama and Kazuko Koshiba-Takeuchi. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Brief. Funct. Genomics*, 17(5):329–338, September 2018.

[136] Mouse Genome Sequencing Consortium, Robert H Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F Abril, Pankaj Agarwal, Richa Agarwala, Rachel Ainscough, Marina Alexandersson, Peter An, Stylianos E Antonarakis, John Attwood, Robert Baertsch, Jonathon Bailey, Karen Barlow, Stephan Beck, Eric Berry, Bruce Birren, Toby Bloom, Peer Bork, Marc Botcherby, Nicolas Bray, Michael R Brent, Daniel G Brown, Stephen D Brown, Carol Bult, John Burton, Jonathan Butler, Robert D Campbell, Piero Carninci,

Simon Cawley, Francesca Chiaromonte, Asif T Chinwalla, Deanna M Church, Michele Clamp, Christopher Clee, Francis S Collins, Lisa L Cook, Richard R Copley, Alan Coulson, Olivier Couronne, James Cuff, Val Curwen, Tim Cutts, Mark Daly, Robert David, Joy Davies, Kimberly D Delehaunty, Justin Deri, Emmanouil T Dermitzakis, Colin Dewey, Nicholas J Dickens, Mark Diekhans, Sheila Dodge, Inna Dubchak, Diane M Dunn, Sean R Eddy, Laura Elnitski, Richard D Emes, Pallavi Eswara, Eduardo Eyras, Adam Felsenfeld, Ginger A Fewell, Paul Flicek, Karen Foley, Wayne N Frankel, Lucinda A Fulton, Robert S Fulton, Terrence S Furey, Diane Gage, Richard A Gibbs, Gustavo Glusman, Sante Gnerre, Nick Goldman, Leo Goodstadt, Darren Grafham, Tina A Graves, Eric D Green, Simon Gregory, Roderic Guigó, Mark Guyer, Ross C Hardison, David Haussler, Yoshihide Hayashizaki, Ladeana W Hillier, Angela Hinrichs, Wratko Hlavina, Timothy Holzer, Fan Hsu, Axin Hua, Tim Hubbard, Adrienne Hunt, Ian Jackson, David B Jaffe, L Steven Johnson, Matthew Jones, Thomas A Jones, Ann Joy, Michael Kamal, Elinor K Karlsson, Donna Karolchik, Arkadiusz Kasprzyk, Jun Kawai, Evan Keibler, Cristyn Kells, W James Kent, Andrew Kirby, Diana L Kolbe, Ian Korf, Raju S Kucherlapati, Edward J Kulbokas, David Kulp, Tom Landers, J P Leger, Steven Leonard, Ivica Letunic, Rosie Levine, Jia Li, Ming Li, Christine Lloyd, Susan Lucas, Bin Ma, Donna R Maglott, Elaine R Mardis, Lucy Matthews, Evan Mauceli, John H Mayer, Megan McCarthy, W Richard McCombie, Stuart McLaren, Kirsten McLay, John D McPherson, Jim Meldrim, Beverley Meredith, Jill P Mesirov, Webb Miller, Tracie L Miner, Emmanuel Mongin, Kate T Montgomery, Michael Morgan, Richard Mott, James C Mullikin, Donna M Muzny, William E Nash, Joanne O Nelson, Michael N Nhan, Robert Nicol, Zemin Ning, Chad Nusbaum, Michael J O'Connor, Yasushi Okazaki, Karen Oliver, Emma Overton-Larty, Lior Pachter, Genís Parra, Kymberlie H Pepin, Jane Peterson, Pavel Pevzner, Robert Plumb, Craig S Pohl, Alex Poliakov, Tracy C Ponce, Chris P Ponting, Simon Potter, Michael Quail, Alexandre Reymond, Bruce A Roe, Krishna M Roskin, Edward M Rubin, Alistair G Rust, Ralph Santos, Victor Sapojnikov, Brian Schultz, Jörg Schultz, Matthias S Schwartz, Scott Schwartz, Carol Scott, Steven Seaman, Steve Searle, Ted Sharpe, Andrew Sheridan, Ratna Shownkeen, Sarah Sims, Jonathan B Singer, Guy Slater, Arian Smit, Douglas R Smith, Brian Spencer, Arne Stabenau, Nicole Stange-Thomann, Charles Sugnet, Mikita Suyama, Glenn Tesler, Johanna Thompson, David Torrents, Evanne Trevaskis, John Tromp, Catherine Ucla, Abel Ureta-Vidal, Jade P Vinson, Andrew C Von Niederhausern, Claire M Wade, Melanie Wall, Ryan J Weber, Robert B Weiss, Michael C Wendl, Anthony P West, Kris Wetterstrand, Raymond Wheeler, Simon Whelan, Jamey Wierzbowski, David Willey, Sophie Williams, Richard K Wilson, Eitan Winter, Kim C Worley, Dudley Wyman, Shan Yang, Shiaw-Pyng Yang, Evgeny M Zdobnov, Michael C Zody, and Eric S Lander. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, December 2002.

[137] Martín Muñoz-López and José L García-Pérez. DNA transposons: nature and

applications in genomics. *Curr. Genomics*, 11(2):115–128, April 2010.

[138] Yoji Murata, Takenori Kotani, Hiroshi Ohnishi, and Takashi Matozaki. The CD47-SIRPα signalling system: its physiological roles and therapeutic application. *J. Biochem.*, 155(6):335–344, June 2014.

[139] Pulak Ranjan Nath, Dipasmita Pal-Nath, Ajeet Mandal, Margaret C Cam, Anthony L Schwartz, and David D Roberts. Natural killer cell recruitment and activation are regulated by CD47 expression in the tumor microenvironment. *Cancer Immunol Res*, 7(9):1547–1561, September 2019.

[140] J Nathans and D S Hogness. Isolation, sequence analysis, and intron-exon arrangement of the gene encoding bovine rhodopsin. *Cell*, 34(3):807–814, October 1983.

[141] Magali Naville, Simon Henriet, Ian Warren, Sara Sumic, Magnus Reeve, Jean-Nicolas Volff, and Daniel Chourrout. Massive changes of genome size driven by expansions of non-autonomous transposable elements. *Curr. Biol.*, 29(7):1161–1168.e6, April 2019.

[142] Thomas J Near, Alex Dornburg, Ron I Eytan, Benjamin P Keck, W Leo Smith, Kristen L Kuhn, Jon A Moore, Samantha A Price, Frank T Burbrink, Matt Friedman, and Peter C Wainwright. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. *Proc. Natl. Acad. Sci. U. S. A.*, 110(31):12738–12743, July 2013.

[143] Thomas J Near, Alex Dornburg, Kristen L Kuhn, Joseph T Eastman, Jillian N Pennington, Tomaso Patarnello, Lorenzo Zane, Daniel A Fernández, and Christopher D Jones. Ancient climate change, antifreeze, and the evolutionary diversification of antarctic fishes. *Proceedings of the National Academy of Sciences*, 109(9):3434–3439, 2012.

[144] Thomas J Near, Alex Dornburg, Masayoshi Tokita, Dai Suzuki, Matthew C Brandley, and Matt Friedman. Boom and bust: ancient and recent diversification in bichirs (polypteridae: Actinopterygii), a relictual lineage of ray-finned fishes. *Evolution*, 68(4):1014–1026, April 2014.

[145] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdiscip. Rev. Comput. Stat.*, 4(2):199–203, March 2012.

[146] Simona Negrini, Vassilis G Gorgoulis, and Thanos D Halazonetis. Genomic instability–an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.*, 11(3):220–228, March 2010.

[147] Xiao-Min Niu, Yong-Chao Xu, Zi-Wen Li, Yu-Tao Bian, Xing-Hui Hou, Jia-Fu Chen, Yu-Pan Zou, Juan Jiang, Qiong Wu, Song Ge, Sureshkumar Balasubramanian, and Ya-Long Guo. Transposable elements drive rapid phenotypic

variation in *Capsella rubella*. *Proceedings of the National Academy of Sciences*, 116(14):6908–6913, 2019.

[148] Peter A Novick, Jeremy D Smith, Mark Floumanhaft, David A Ray, and Stéphane Boissinot. The evolution and diversity of DNA transposons in the genome of the lizard anolis carolinensis. *Genome Biol. Evol.*, 3:1–14, 2011.

[149] Daniel Ocampo Daza, Christina A Bergqvist, and Dan Larhammar. The evolution of oxytocin and vasotocin receptor genes in jawed vertebrates: A clear case for gene duplications through ancestral Whole-Genome duplications. *Front. Endocrinol.*, 12:792644, 2021.

[150] Martin T O'Connell, Travis D Shepherd, Ann M U O'Connell, and Ransom A Myers. Long-term declines in two apex predators, bull sharks (carcharhinus leucas) and alligator gar (atractosteus spatula), in lake pontchartrain, an oligohaline estuary in southeastern louisiana. *Estuaries Coasts*, 30(4):567–574, August 2007.

[151] Emily A O'Connor, Charlie K Cornwallis, Dennis Hasselquist, Jan-Åke Nilsson, and Helena Westerdahl. The evolution of immunity in relation to colonization and migration. *Nat Ecol Evol*, 2(5):841–849, May 2018.

[152] P A Oldenborg, A Zheleznyak, Y F Fang, C F Lagenaur, H D Gresham, and F P Lindberg. Role of CD47 as a marker of self on red blood cells. *Science*, 288(5473):2051–2054, June 2000.

[153] E Olmo, T Capriglione, and G Odierna. Genome size evolution in vertebrates: trends and constraints. *Comp. Biochem. Physiol. B*, 92(3):447–453, 1989.

[154] Bryan Oronsky, Corey Carter, Tony Reid, Franck Brinkhaus, and Susan J Knox. Just eat it: A review of CD47 and SIRP-α antagonism. *Semin. Oncol.*, 47(2-3):117–124, April 2020.

[155] Ovchinnikov YuA. Rhodopsin and bacteriorhodopsin: structure-function relationships. *FEBS Lett.*, 148(2):179–191, November 1982.

[156] M Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, October 1999.

[157] F Paillard, G Sterkers, and C Vaquero. Transcriptional and post-transcriptional regulation of TcR, CD4 and CD8 gene expression during activation of normal human T lymphocytes. *EMBO J.*, 9(6):1867–1872, June 1990.

[158] Nathan Papon, Pauline Lasserre-Zuber, Hélène Rimbert, Romain De Oliveira, Etienne Paux, and Frédéric Choulet. All families of transposable elements were active in the recent wheat genome evolution and polyploidy had no impact on their activity. *Plant Genome*, 16(3):e20347, September 2023.

[159] Christian Parisod, Karine Alix, Jérémy Just, Maud Petit, Véronique Sarilar, Corinne Mhiri, Malika Ainouche, Boulos Chalhoub, and Marie-Angèle Grandbastien. Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol.*, 186(1):37–45, April 2010.

[160] Elyse Parker, Katerina L Zapfe, Jagriti Yadav, Bruno Frédérich, Christopher D Jones, Evan P Economo, Sarah Federman, Thomas J Near, and Alex Dornburg. Periodic environmental disturbance drives repeated ecomorphological diversification in an adaptive radiation of antarctic fishes. *Am. Nat.*, 200(6):E221–E236, December 2022.

[161] Jeremy Pasquier, Cédric Cabau, Thaovi Nguyen, Elodie Jouanno, Dany Severac, Ingo Braasch, Laurent Journot, Pierre Pontarotti, Christophe Klopp, John H Postlethwait, Yann Guiguen, and Julien Bobe. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. *BMC Genomics*, 17:368, May 2016.

[162] Nicole S Paulat, Jessica M Storer, Diana D Moreno-Santillán, Austin B Osmanski, Kevin A M Sullivan, Jenna R Grimshaw, Jennifer Korstian, Michaela Halsey, Carlos J Garcia, Claudia Crookshanks, Jaquelyn Roberts, Arian F A Smit, Robert Hubley, Jeb Rosen, Emma C Teeling, Sonja C Vernes, Eugene Myers, Martin Pippel, Thomas Brown, Michael Hiller, Danny Rojas, Liliana M Dávalos, Kerstin Lindblad-Toh, Elinor K Karlsson, and David A Ray. Chiropterans are a hotspot for horizontal transfer of DNA transposons in mammalia. *Mol. Biol. Evol.*, 40(5):msad092, April 2023.

[163] Jennifer B Phillips and Monte Westerfield. Zebrafish models in translational research: tipping the scales toward advancements in human health. *Dis. Model. Mech.*, 7(7):739–743, July 2014.

[164] Oliver Piskurek and Daniel J Jackson. Tracking the ancestry of a deeply conserved eumetazoan SINE domain. *Mol. Biol. Evol.*, 28(10):2727–2730, October 2011.

[165] Nicholas H Putnam, Brendan L O'Connell, Jonathan C Stites, Brandon J Rice, Marco Blanchette, Robert Calef, Christopher J Troll, Andrew Fields, Paul D Hartley, Charles W Sugnet, David Haussler, Daniel S Rokhsar, and Richard E Green. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.*, 26(3):342–350, March 2016.

[166] P Ráb, M Rábová, K M Reed, and R B Phillips. Chromosomal characteristics of ribosomal DNA in the primitive semionotiform fish, longnose gar lepisosteus osseus. *Chromosome Res.*, 7(6):475–480, 1999.

[167] Narayanan Raghupathy and Dannie Durand. Gene cluster statistics with gene families. *Mol. Biol. Evol.*, 26(5):957–968, May 2009.

[168] Swarajpal S Randhawa and Ravindra Pawar. Fish genomes: Sequencing trends, taxonomy and influence of taxonomy on genome attributes. *J. Appl. Ichthyol.*, 37(4):553–562, August 2021.

[169] Swagat Ray, Sanjiban Chakrabarty, and Arnab Ray Chaudhuri. Editorial: Impact of genome instability on human health. *Front Mol Biosci*, 10:1243968, July 2023.

[170] Gabriel E Rech, Santiago Radío, Sara Guirao-Rico, Laura Aguilera, Vivien Horvath, Llewellyn Green, Hannah Lindstadt, Véronique Jamilloux, Hadi Quesneville, and Josefa González. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in drosophila. *Nat. Commun.*, 13(1):1–16, April 2022.

[171] William B Reinar, Ole K Tørresen, Alexander J Nederbragt, Michael Matschiner, Sissel Jentoft, and Kjetill S Jakobsen. Genomic repeat landscape evolution across the teleost fish lineages. March 2023.

[172] Daphné Reiss, Gladys Mialdea, Vincent Miele, Damien M de Vienne, Jean Peccoud, Clément Gilbert, Laurent Duret, and Sylvain Charlat. Global survey of mobile DNA horizontal transfer in arthropods reveals lepidoptera as a prime hotspot. *PLoS Genet.*, 15(2):e1007965, February 2019.

[173] Liam J Revel. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 2012.

[174] Liam J Revell. phytools 2.0: An updated R ecosystem for phylogenetic comparative methods (and other things). August 2023.

[175] Marco Ricci, Valentina Peona, Etienne Guichard, Cristian Taccioli, and Alessio Boattini. Transposable elements activity is positively related to rate of speciation in mammals. *J. Mol. Evol.*, 86(5):303–310, June 2018.

[176] Fernando Rodriguez and Irina R Arkhipova. Transposable elements and polyploid evolution in animals. *Curr. Opin. Genet. Dev.*, 49:115–123, April 2018.

[177] Iván Rodríguez-Nunez, Dustin J Wcisel, Gary W Litman, and Jeffrey A Yoder. Multigene families of immunoglobulin domain-containing innate immune receptors in zebrafish: deciphering the differences. *Dev. Comp. Immunol.*, 46(1):24–34, September 2014.

[178] Natasha M Rogers, Mingyi Yao, Enrico M Novelli, Angus W Thomson, David D Roberts, and Jeffrey S Isenberg. Activated CD47 regulates multiple vascular and stress responses: implications for acute kidney injury and its management. *Am. J. Physiol. Renal Physiol.*, 303(8):F1117–25, October 2012.

[179] Antonis Rokas, Jennifer H Wisecaver, and Abigail L Lind. The birth, evolution and death of metabolic gene clusters in fungi. *Nat. Rev. Microbiol.*, 16(12):731–744, December 2018.

[180] M Seiffert, C Cant, Z Chen, I Rappold, W Brugger, L Kanz, E J Brown, A Ullrich, and H J Bühring. Human signal-regulatory protein is expressed on normal, but not on subsets of leukemic myeloid cells and mediates cellular adhesion involving its counterreceptor CD47. *Blood*, 94(11):3633–3643, December 1999.

[181] Anna D Senft and Todd S Macfarlan. Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.*, 22(11):691–711, August 2021.

[182] Mathieu Seppey, Mosè Manni, and Evgeny M Zdobnov. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.*, 1962:227–245, 2019.

[183] Antonio Serrato-Capuchina and Daniel R Matute. The role of transposable elements in speciation. *Genes*, 9(5), May 2018.

[184] Camille Sessegolo, Nelly Burlet, and Annabelle Haudry. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biol. Lett.*, 12(8), August 2016.

[185] H Bradley Shaffer, Patrick Minx, Daniel E Warren, Andrew M Shedlock, Robert C Thomson, Nicole Valenzuela, John Abramyan, Chris T Amemiya, Daleen Badenhorst, Kyle K Biggar, Glen M Borchert, Christopher W Botka, Rachel M Bowden, Edward L Braun, Anne M Bronikowski, Benoit G Bruneau, Leslie T Buck, Blanche Capel, Todd A Castoe, Mike Czerwinski, Kim D Delehaunty, Scott V Edwards, Catrina C Fronick, Matthew K Fujita, Lucinda Fulton, Tina A Graves, Richard E Green, Wilfried Haerty, Ramkumar Hariharan, Omar Hernandez, Ladeana W Hillier, Alisha K Holloway, Daniel Janes, Fredric J Janzen, Cyriac Kandoth, Lesheng Kong, A P Jason de Koning, Yang Li, Robert Literman, Suzanne E McGaugh, Lindsey Mork, Michelle O'Laughlin, Ryan T Paitz, David D Pollock, Chris P Ponting, Srihari Radhakrishnan, Brian J Raney, Joy M Richman, John St John, Tonia Schwartz, Arun Sethuraman, Phillip Q Spinks, Kenneth B Storey, Nay Thane, Tomas Vinar, Laura M Zimmerman, Wesley C Warren, Elaine R Mardis, and Richard K Wilson. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biol.*, 14(3):R28, March 2013.

[186] Robert C Sharp, Matthew E Brown, Melanie R Shapiro, Amanda L Posgai, and Todd M Brusko. The immunoregulatory role of the signal regulatory protein family and CD47 signaling pathway in type 1 diabetes. *Front. Immunol.*, 12:739048, September 2021.

[187] Andrew E Shaw, Joseph Hughes, Quan Gu, Abdelkader Behdenna, Joshua B Singer, Tristan Dennis, Richard J Orton, Mariana Varela, Robert J Gifford, Sam J Wilson, and Massimo Palmarini. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. *PLoS Biol.*, 15(12):e2004086, December 2017.

[188] E Sick, A Jeanne, C Schneider, S Dedieu, K Takeda, and L Martiny. CD47 update: a multifaceted actor in the tumour microenvironment of potential therapeutic interest. *Br. J. Pharmacol.*, 167(7):1415–1430, December 2012.

[189] Felipe A Simão, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, October 2015.

[190] Morgan J Smith, Lucia Pastor, Jeremy R B Newman, and Patrick Concannon. Genetic control of splicing at SIRPG modulates risk of type 1 diabetes. *Diabetes*, 71(2):350–358, February 2022.

[191] N G Smith, D J Daugherty, E L Brinkman, M G Wegener, B R Kreiser, A M Ferrara, K D Kimmel, and S R David. Advances in conservation and management of the alligator gar: A synthesis of current knowledge and introduction to a special section. *N. Am. J. Fish. Manage.*, 40(3):527–543, June 2020.

[192] Cibele G Sotero-Caio, Roy N Platt, 2nd, Alexander Suh, and David A Ray. Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.*, 9(1):161–177, January 2017.

[193] Bastiaan Star, Alexander J Nederbragt, Sissel Jentoft, Unni Grimholt, Martin Malmstrøm, Tone F Gregers, Trine B Rounge, Jonas Paulsen, Monica H Solbakken, Animesh Sharma, Ola F Wetten, Anders Lanzén, Roger Winer, James Knight, Jan-Hinnerk Vogel, Bronwen Aken, Oivind Andersen, Karin Lagesen, Ave Tooming-Klunderud, Rolf B Edvardsen, Kirubakaran G Tina, Mari Espelund, Chirag Nepal, Christopher Previti, Bård Ove Karlsen, Truls Moum, Morten Skage, Paul R Berg, Tor Gjøen, Heiner Kuhl, Jim Thorsen, Ketil Malde, Richard Reinhardt, Lei Du, Steinar D Johansen, Steve Searle, Sigbjørn Lien, Frank Nilsen, Inge Jonassen, Stig W Omholt, Nils Chr Stenseth, and Kjetill S Jakobsen. The genome sequence of atlantic cod reveals a unique immune system. *Nature*, 477(7363):207–210, August 2011.

[194] Elizabeth R Stirling. *CD47 Is a Novel Target for Cancer Immunotherapy Treatment*. PhD thesis, Wake Forest University, Ann Arbor, United States, 2022.

[195] Susie Suh, Elliot H Choi, and Natasha Atanaskova Mesinkovska. The expression of opsins in the human skin and its implications for photobiomodulation: A systematic review. *Photodermatol. Photoimmunol. Photomed.*, 36(5):329–338, September 2020.

[196] U Rashid Sumaila, Andrés M Cisneros-Montemayor, Andrew Dyck, Ling Huang, William Cheung, Jennifer Jacquet, Kristin Kleisner, Vicky Lam, Ashley McCrea-Strub, Wilf Swartz, Reg Watson, Dirk Zeller, and Daniel Pauly. Impact of the deepwater horizon well blowout on the economics of US gulf fisheries. *Can. J. Fish. Aquat. Sci.*, 69(3):499–510, March 2012.

[197] Dai Suzuki, Matthew C Brandley, and Masayoshi Tokita. The mitochondrial phylogeny of an ancient lineage of ray-finned fishes (polypteridae) with implications for the evolution of body elongation, pelvic fin loss, and craniofacial morphology in osteichthyes. *BMC Evol. Biol.*, 10:21, January 2010.

[198] Shinichiro Takahashi. Molecular functions of SIRPα and its role in cancer. *Biomed Rep*, 9(1):3–7, July 2018.

[199] Izabella L Tambones, Annabelle Haudry, Maryanna C Simão, and Claudia M A Carareto. High frequency of horizontal transfer in jockey families (LINE order) of drosophilids. *Mob. DNA*, 10:43, November 2019.

[200] Maja Tarailo-Graovac and Nansheng Chen. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, Chapter 4:4.10.1–4.10.14, March 2009.

[201] Felix Teufel, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gíslason, Silas Irby Pihl, Konstantinos D Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.*, 40(7):1023–1025, July 2022.

[202] Andrew W Thompson, M Brent Hawkins, Elise Parey, Dustin J Wcisel, Tatsuya Ota, Kazuhiko Kawasaki, Emily Funk, Mauricio Losilla, Olivia E Fitch, Qiaowei Pan, Romain Feron, Alexandra Louis, Jérôme Montfort, Marine Milhes, Brett L Racicot, Kevin L Childs, Quenton Fontenot, Allyse Ferrara, Solomon R David, Amy R McCune, Alex Dornburg, Jeffrey A Yoder, Yann Guiguen, Hugues Roest Crollius, Camille Berthelot, Matthew P Harris, and Ingo Braasch. The bowfin genome illuminates the developmental evolution of ray-finned fishes. *Nat. Genet.*, 53(9):1373–1384, September 2021.

[203] Marc Tollis, Joshua D Schiffman, and Amy M Boddy. Evolution of cancer suppression as revealed by mammalian comparative genomics. *Curr. Opin. Genet. Dev.*, 42:40–47, February 2017.

[204] Alessandro Torgovnick and Björn Schumacher. DNA repair mechanisms in cancer development and therapy. *Front. Genet.*, 6:157, April 2015.

[205] David E Torres, Bart P H J Thomma, and Michael F Seidl. Transposable elements contribute to genome dynamics and gene expression variation in the fungal plant pathogen verticillium dahliae. *Genome Biol. Evol.*, 13(7), July 2021.

[206] Jeffrey P Townsend, Hayley B Hassler, April D Lamb, Pratha Sah, Aia Alvarez Nishio, Cameron Nguyen, Alexandra D Tew, Alison P Galvani, and Alex Dornburg. Seasonality of endemic COVID-19. *MBio*, 14(6):e0142623, November 2023.

[207] Brandon A Turner, Theresa R Miorin, Nicholas B Stewart, Robert W Reid, Cathy C Moore, and Rebekah L Rogers. Chromosomal rearrangements as a source of local adaptation in island drosophila. September 2021.

[208] Ellen M van Beek, Fiona Cochrane, A Neil Barclay, and Timo K van den Berg. Signal regulatory proteins in the immune system. *J. Immunol.*, 175(12):7781–7787, December 2005.

[209] Richard Van Der Laan, William N Eschmeyer, and Ronald Fricke. Family-group names of recent fishes. *Zootaxa*, 3882:1–230, November 2014.

[210] André Veillette and Jun Chen. SIRPα-CD47 immune checkpoint blockade in anticancer therapy. *Trends Immunol.*, 39(3):173–184, March 2018.

[211] E F Vernon-Wilson, W J Kee, A C Willis, A N Barclay, D L Simmons, and M H Brown. CD47 is a ligand for rat macrophage membrane signal regulatory protein SIRP (OX41) and human SIRPalpha 1. *Eur. J. Immunol.*, 30(8):2130–2137, August 2000.

[212] A Vervoort. Karyotypes and nuclear DNA contents of polypteridae (osteichthyes). *Experientia*, 36(6):646–647, June 1980.

[213] Ricardo Assunção Vialle, Jorge Estefano Santana de Souza, Katia de Paiva Lopes, Diego Gomes Teixeira, Pitágoras de Azevedo Alves Sobrinho, André M Ribeiro-dos Santos, Carolina Furtado, Tetsu Sakamoto, Fábio Augusto Oliveira Silva, Edivaldo Herculano Corrêa de Oliveira, Igor Guerreiro Hamoy, Paulo Pimentel Assumpção, Ândrea Ribeiro-dos Santos, João Paulo Matos Santos Lima, Héctor N Seuánez, Sandro José de Souza, and Sidney Santos. Whole genome sequencing of the pirarucu (arapaima gigas) supports independent emergence of major teleost clades. *Genome Biol. Evol.*, 10(9):2366–2379, July 2018.

[214] Birgit C Viertlboeck, Ramona Schmitt, and Thomas W Göbel. The chicken immunoregulatory receptor families SIRP, TREM, and CMRF35/CD300L. *Immunogenetics*, 58(2-3):180–190, April 2006.

[215] Jean-Nicolas Volff, Laurence Bouneau, Catherine Ozouf-Costaz, and Cécile Fischer. Diversity of retrotransposable elements in compact pufferfish genomes. *Trends Genet.*, 19(12):674–678, December 2003.

[216] Qiu-Hong Wan, Sheng-Kai Pan, Li Hu, Ying Zhu, Peng-Wei Xu, Jin-Quan Xia, Hui Chen, Gen-Yun He, Jing He, Xiao-Wei Ni, Hao-Long Hou, Sheng-Guang Liao, Hai-Qiong Yang, Ying Chen, Shu-Kun Gao, Yun-Fa Ge, Chang-Chang Cao, Peng-Fei Li, Li-Ming Fang, Li Liao, Shu Zhang, Meng-Zhen Wang, Wei Dong, and Sheng-Guo Fang. Genome analysis and signature discovery for diving and sensory properties of the endangered chinese alligator. *Cell Res.*, 23(9):1091–1105, September 2013.

[217] Feng Wang, Yan-Hou Liu, Ting Zhang, Jing Gao, Yangyue Xu, Guang-Yao Xie, Wen-Jie Zhao, Hongda Wang, and Yong-Guang Yang. Aging-associated changes in CD47 arrangement and interaction with thrombospondin-1 on red blood cells visualized by super-resolution imaging. *Aging Cell*, 19(10):e13224, October 2020.

[218] Jiyao Wang, Farideh Chitsaz, Myra K Derbyshire, Noreen R Gonzales, Marc Gwadz, Shennan Lu, Gabriele H Marchler, James S Song, Narmada Thanki, Roxanne A Yamashita, Mingzhang Yang, Dachuan Zhang, Chanjuan Zheng, Christopher J Lanczycki, and Aron Marchler-Bauer. The conserved domain database in 2023. *Nucleic Acids Res.*, 51(D1):D384–D388, January 2023.

[219] Yupeng Wang, Haibao Tang, Jeremy D Debarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-Ho Lee, Huizhe Jin, Barry Marler, Hui Guo, Jessica C Kissinger, and Andrew H Paterson. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, 40(7):e49, April 2012.

[220] Dustin J Wcisel, J Thomas Howard, 3rd, Jeffrey A Yoder, and Alex Dornburg. Transcriptome ortholog alignment sequence tools (TOAST) for phylogenomic dataset assembly. *BMC Evol. Biol.*, 20(1):41, March 2020.

[221] Huamei Wen, Tao Luo, Yali Wang, Siwei Wang, Tao Liu, Ning Xiao, and Jiang Zhou. Molecular phylogeny and historical biogeography of the cave fish genus sinocyclocheilus (cypriniformes: Cyprinidae) in southwest china. *Integr. Zool.*, 17(2):311–325, March 2022.

[222] Hadley Wickham. *ggplot2*. Springer New York.

[223] Wai Yee Wong, Oleg Simakov, Diane M Bridge, Paulyn Cartwright, Anthony J Bellantuono, Anne Kuhn, Thomas W Holstein, Charles N David, Robert E Steele, and Daniel E Martínez. Expansion of a single transposable element family is associated with genome-size increase and radiation in the genus *Hydra*. *Proceedings of the National Academy of Sciences*, 116(46):22915–22917, 2019.

[224] Yong H Woo and Wen-Hsiung Li. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. *Proceedings of the National Academy of Sciences*, 108(8):3306–3311, 2011.

[225] Jeremy J Wright, Spencer A Bruce, Daniel A Sinopoli, Jay R Palumbo, and Donald J Stewart. Phylogenomic analysis of the bowfin (amia calva) reveals unrecognized species diversity in a living fossil lineage. *Sci. Rep.*, 12(1):16514, October 2022.

[226] Hua Yang, Yang Xun, and Hua You. The landscape overview of CD47-based immunotherapy for hematological malignancies. *Biomarker Research*, 11(1):1–22, February 2023.

[227] Yixin Yao and Wei Dai. Genomic instability and cancer. *J Carcinog Mutagen*, 5, 2014.

[228] K C Yeh, S H Wu, J T Murphy, and J C Lagarias. A cyanobacterial phytochrome two-component light sensory system. *Science*, 277(5331):1505–1508, September 1997.

[229] Ella B Yoder, Elyse Parker, Alexandra Tew, Christopher D Jones, and Alex Dornburg. Decoupled spectral tuning and eye size diversification patterns in an antarctic adaptive radiation. September 2022.

[230] Jeffrey A Yoder. Form, function and phylogenetics of NITRs in bony fish. *Dev. Comp. Immunol.*, 33(2):135–144, February 2009.

[231] Daqi Yu, Yandong Ren, Masahiro Uesaka, Alan J S Beavan, Matthieu Muffato, Jieyu Shen, Yongxin Li, Iori Sato, Wenting Wan, James W Clark, Joseph N Keating, Emily M Carlisle, Richard P Dearden, Sam Giles, Emma Randle, Robert S Sansom, Roberto Feuda, James F Fleming, Fumiaki Sugahara, Carla Cummins, Mateus Patricio, Wasiu Akanni, Salvatore D'Aniello, Cristiano Bertolucci, Naoki Irie, Cantas Alev, Guojun Sheng, Alex de Mendoza, Ignacio Maeso, Manuel Irimia, Bastian Fromm, Kevin J Peterson, Sabyasachi Das, Masayuki Hirano, Jonathan P Rast, Max D Cooper, Jordi Paps, Davide Pisani, Shigeru Kuratani, Fergal J Martin, Wen Wang, Philip C J Donoghue, Yong E Zhang, and Juan Pascual-Anaya. Hagfish genome elucidates vertebrate whole-genome duplication events and their evolutionary consequences. *Nat Ecol Evol*, 8(3):519–535, March 2024.

[232] Guangchuang Yu. Preface. `https://yulab-smu.top/treedata-book/`. Accessed: 2023-12-27.

[233] W P Yu, C J Pallen, A Tay, F R Jirik, S Brenner, Y H Tan, and B Venkatesh. Conserved synteny between the fugu and human PTEN locus and the evolutionary conservation of vertebrate PTEN function. *Oncogene*, 20(39):5554–5561, September 2001.

[234] Zihao Yuan, Shikai Liu, Tao Zhou, Changxu Tian, Lisui Bao, Rex Dunham, and Zhanjiang Liu. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genomics*, 19(1):141, February 2018.

[235] Hua-Hao Zhang, Cédric Feschotte, Min-Jin Han, and Ze Zhang. Recurrent horizontal transfers of chapaev transposons in diverse invertebrate and vertebrate animals. *Genome Biol. Evol.*, 6(6):1375–1386, May 2014.

[236] Hua-Hao Zhang, Jean Peccoud, Min-Rui-Xuan Xu, Xiao-Gu Zhang, and Clément Gilbert. Horizontal transfer and evolution of transposable elements in vertebrates. *Nat. Commun.*, 11(1):1–10, March 2020.