

Exon probe sets and bioinformatics pipelines for all levels of fish phylogenomics

Lily C. Hughes^{1,2,3}  | Guillermo Orti^{1,3} | Hadeel Saad¹ | Chenhong Li⁴  | William T. White⁵ | Carole C. Baldwin³ | Keith A. Crandall^{1,2}  | Dahiana Arcila^{3,6,7} | Ricardo Betancur-R⁷

¹Department of Biological Sciences, George Washington University, Washington, DC, USA

²Computational Biology Institute, Milken Institute of Public Health, George Washington University, Washington, DC, USA

³Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA

⁴College of Fisheries and Life Sciences, Shanghai Ocean University, Shanghai, China

⁵CSIRO Australian National Fish Collection, National Research Collections of Australia, Hobart, TAS, Australia

⁶Sam Noble Oklahoma Museum of Natural History, Norman, OK, USA

⁷Department of Biology, University of Oklahoma, Norman, OK, USA

Correspondence

Lily C. Hughes, Department of Organismal Biology and Anatomy, University of Chicago, Chicago, IL, USA.

Email: lilychughes@gmail.com

Funding information

National Science Foundation, Grant/Award Number: DEB 1457426, DEB 1541552, DEB 1541554, DEB 1929248 and DEB 1932759

Abstract

Exon markers have a long history of use in phylogenetics of ray-finned fishes, the most diverse clade of vertebrates with more than 35,000 species. As the number of published genomes increases, it has become easier to test exons and other genetic markers for signals of ancient duplication events and filter out paralogues that can mislead phylogenetic analysis. We present seven new probe sets for current target-capture phylogenomic protocols that capture 1,104 exons explicitly filtered for paralogues using gene trees. These seven probe sets span the diversity of teleost fishes, including four sets that target five hyperdiverse percomorph clades which together comprise ca. 17,000 species (Carangaria, Ovalentaria, Eupercaria, and Syngnatharia + Pelagiaria combined). We additionally included probes to capture legacy nuclear exons and mitochondrial markers that have been commonly used in fish phylogenetics (despite some exons being flagged for paralogues) to facilitate integration of old and new molecular phylogenetic matrices. We tested these probes experimentally for 56 fish species (eight species per probe set) and merged new exon-capture sequence data into an existing data matrix of 1,104 exons and 300 ray-finned fish species. We provide an optimized bioinformatics pipeline to assemble exon capture data from raw reads to alignments for downstream analysis. We show that legacy loci with known paralogues are at risk of assembling duplicated sequences with target-capture, but we also assembled many useful orthologous sequences that can be integrated with many PCR-generated matrices. These probe sets are a valuable resource for advancing fish phylogenomics because targeted exons can easily be extracted from increasingly available whole genome and transcriptome data sets, and also may be integrated with existing PCR-based exon and mitochondrial data.

KEYWORDS

Actinopterygii, evolution, phylogenetics, protein coding, systematics, target capture

1 | INTRODUCTION

Phylogenetic inference relies on the analysis of orthologues—homologous loci that track evolutionary history, not duplication events (Fitch, 1970). Undetected paralogues—gene copies derived

from duplication events—mislead phylogenetic analysis, even with genome-scale data sets including hundreds or thousands of loci (Brown & Thomson, 2017; Philippe et al., 2017). Whole-genome duplication (WGD) events are a major source of duplicated gene copies, and are common in the evolutionary history of plants

(Clark & Donoghue, 2018). But numerous metazoan lineages in the Tree of Life have also experienced WGD, including ancient duplications in hexapods (Li et al., 2018), arachnids (Clarke et al., 2015; Schwager et al., 2017), and the ancestor to modern horseshoe crabs (Kenny et al., 2016). The genomes of all living vertebrates share two early WGD events, and an additional WGD event took place in the ancestor to teleost fishes (Dehal & Boore, 2005; Vandepoele et al., 2004), a lineage that makes up nearly half of the diversity of vertebrate species. While many duplicated gene copies were lost shortly after the teleost WGD (Inoue et al., 2015), up to a quarter of genes in teleost genomes have paralogues as a consequence of this event (Braasch et al., 2015), posing a challenge for molecular phylogenetics of fishes.

Exon markers have played a pivotal role in resolving phylogenetic relationships among ray-finned fishes (Betancur-R et al., 2013, 2017; Hughes et al., 2018; Li et al., 2007; Near et al., 2012; Rabosky et al., 2018). Identification of these exons has typically involved the comparison of a small number of fish model genomes. For example, a suite of 154 exons was identified by Li et al. (2007) using a reciprocal BLAST approach on two genomes, the pufferfish *Takifugu rubripes* and the zebrafish *Danio rerio*, to find “single-copy” conserved exons, and subsequently design PCR primers for amplification and sequencing (Li et al., 2007; nuclear markers optimized for PCR sequencing are hereafter referred to as “legacy markers”). These exons demonstrated their utility for resolving previously enigmatic relationships among fishes (Li et al., 2008), and were the basis for largescale reappraisals of the ray-finned fish Tree of Life (Betancur-R et al., 2013; Near et al., 2012), phylogenetic analysis of the large clade of the “spiny-ray” acanthomorph fishes (Near et al., 2013), and new phylogenetic classifications based on sequence data for more than 2,000 fish species (Betancur-R et al., 2013, 2017). Mitochondrial genomes also have been targeted frequently for sequencing in ray-finned fishes (Iwasaki et al., 2013; Miya et al., 2003; Sato et al., 2018). Most recently, a modest number of legacy markers in combination with mitochondrial data available through GenBank were compiled for one of the largest analyses of a supermatrix with more than 11,000 ray-finned fish species (Rabosky et al., 2018).

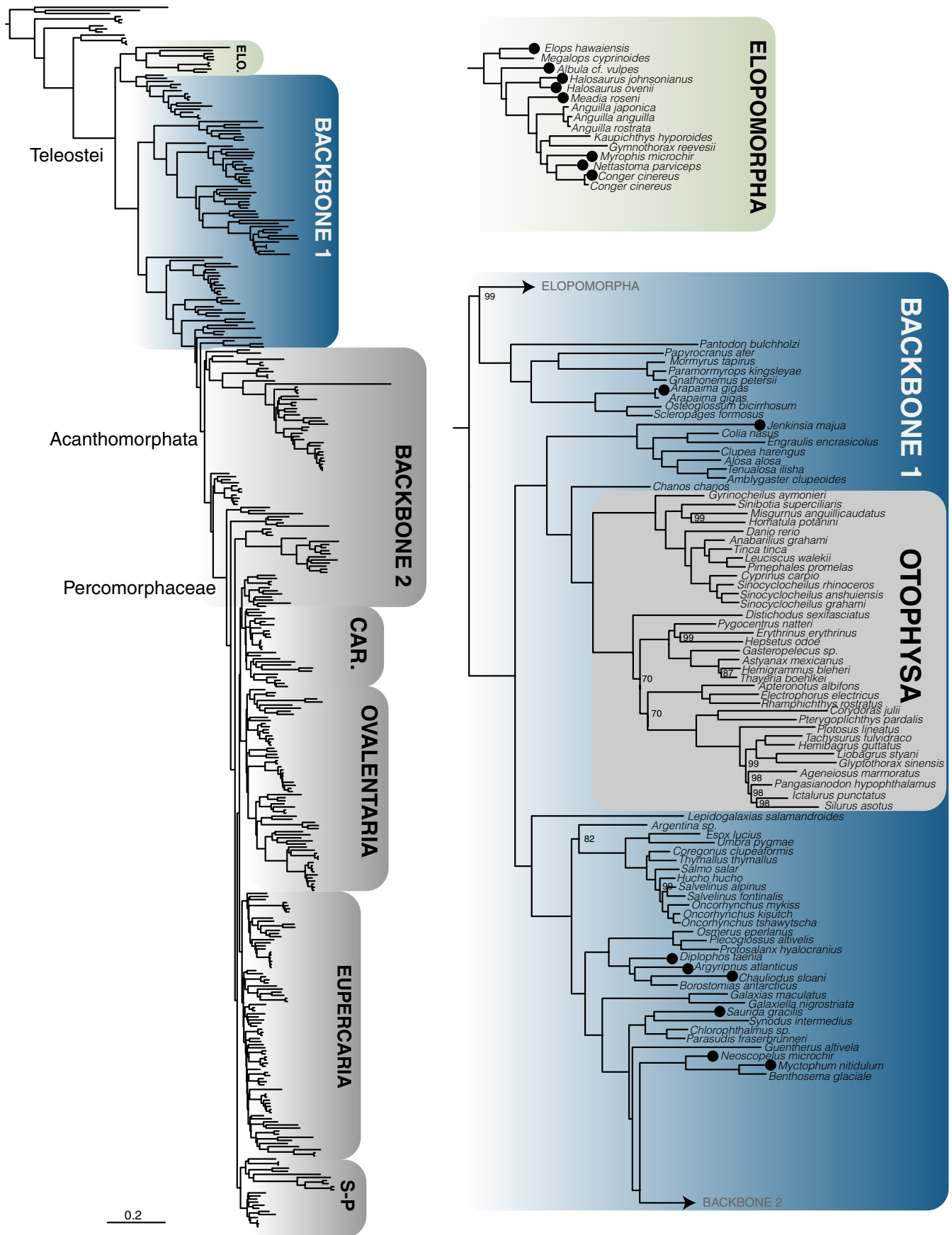
The advent of high-throughput sequencing technologies has drastically increased the number of loci systematists can harness for their groups of interest. But criteria for defining orthology still rely primarily on sequence similarity rather than on more accurate tree-based approaches (Kocot et al., 2013). While sequence capture based on single-stranded RNA probes that enrich genomic DNA libraries for conserved molecular markers have revolutionized phylogenomics (Faircloth et al., 2012; Lemmon et al., 2012), allowing cost-effective sequencing of hundreds or thousands of markers for many taxa, only a few studies have explicitly used tree-based criteria to define orthology for probe design (Owen et al., 2020).

Popular markers used in fish phylogenomic studies include ultra-conserved elements (UCEs) (Alfaro et al., 2018; Chakrabarty et al., 2017; Faircloth et al., 2012, 2013, 2020; Friedman et al., 2019; Harrington et al., 2016; Longo et al., 2017; Roxo et al., 2019), exon capture (Arcila et al., 2017; Betancur-R et al., 2019; Ilves & López-Fernández, 2014; Ilves et al., 2017; Jiang et al., 2019; Song et al., 2017), and anchored hybrid enrichment (AHE) approaches (Dornburg et al., 2017; Eytan et al., 2015; Irisarri et al., 2018; Lemmon et al., 2012; Stout et al., 2016). Still, most genome-scale markers targeted for fish phylogenetics have so far been selected based on the comparison of a limited number of model genomes and some threshold of similarity to define them as “single-copy” (Li et al., 2007). A recent study implementing an explicit tree-based filtering method to test for orthology revealed that one third of the “single-copy” exons > 200 bp in length identified by Jiang et al. (2019) were affected by paralogy, potentially biasing tree inference (Hughes et al., 2018). A set of 1,105 exons free of vertebrate and teleost WGD-derived paralogues identified in the latter study resolved the phylogeny with confidence for more than 300 species of ray-finned fishes. Other markers used for phylogenomic studies such as UCEs (Faircloth et al., 2013), AHE loci (Lemmon et al., 2012) and exons (Arcila et al., 2017; Ilves & López-Fernández, 2014; Jiang et al., 2019; Song et al., 2017) have not been explicitly tested for paralogy using gene-tree-based approaches.

Exon loci have desirable properties for phylogenomics that other markers may lack. They are relatively easy to align, and a number of software programs have been developed for reading frame-aware alignment (Abascal et al., 2010; Ranwez et al., 2011, 2018), avoiding potential homology errors with UCE markers whose alignments become less reliable toward the flanking regions (Edwards et al., 2017). Both protein and nucleotide sequences can be used for phylogenetic inference, making exons useful for deep (Hughes et al., 2018) and shallow phylogenetic scales (Rincon-Sandoval et al., 2019). Exon markers are also easy to integrate with both genomic and transcriptomic data resources for systematists to increase taxon sampling without incurring additional costs.

Because exon markers tend to be more variable across the target region than UCEs or the markers used for AHE, two rounds of in vitro hybridization are optimal for their sequence capture protocols (Li et al., 2013). This improvement in laboratory techniques has resulted in a number of studies that implement exon capture for fish phylogenomics (Arcila et al., 2017; Betancur-R et al., 2019; Ilves & López-Fernández, 2014; Ilves et al., 2017; Kuang et al., 2018; Li et al., 2015; Rincon-Sandoval et al., 2019; Song et al., 2017; Straube et al., 2018; Yin et al., 2019). The increase in genomic resources for fishes also has allowed for the comparison of a larger number of genomes for probe design (Li et al., 2012), and ultimately eight ray-finned fish genomes have been used to

FIGURE 1 Position of “Elopomorpha” and “Backbone 1” probe sets in the maximum likelihood tree based on protein sequences from 394 fish taxa, combining genomes, transcriptomes, and exon capture. Newly sequenced taxa are represented with black dots at the tips. This clade has a specific probe set with a different but overlapping set of exons designed in an earlier study, Arcila et al. (2017), but is not specifically targeted by Backbone 1. Bootstrap values are 100, unless otherwise noted [Colour figure can be viewed at wileyonlinelibrary.com]



identify > 17,000 “single-copy” exons (Song et al., 2017) using a modification of the reciprocal BLAST approach of Li et al. (2007). A subset (4,434) of these exons were optimized for capture across all ray-finned fishes (Jiang et al., 2019). Increasing taxonomic specificity of probes should increase the capture efficiency of loci, thus increasing the percentage of data present in phylogenomic matrices. Yet many resources for sequence capture are targeted toward broad taxonomic scales in fishes, such as actinopterygians (Faircloth et al., 2013; Jiang et al., 2019), or acanthomorphs (Alfaro et al., 2018), although a few probe sets have been designed to target more specific groups including cichlids (Ilves & López-Fernández, 2014), and otophysans (Arcila et al., 2017; Faircloth et al., 2020).

Here we present a new experimental protocol to obtain sequence data across the diversity of fishes for a set of over 1,100 exons filtered for paralogues using gene tree-based filtering approaches. We provide seven new probe sets for exon capture that are designed to enrich genomic libraries for different taxonomic groups, from the early branching teleosts to the major groups within percomorphs, the massive radiation comprising more than 17,000 species. These are the first probe sets to specifically target order- or supraordinal-level clades across the fish diversity (e.g., elopomorphs, carangarians, eupercarians, syngnatharians, and pelagiarans). These probe sets target the same set of ~1,100 exon loci, but the specific sequences of the probes are tailored to capture more efficiently within taxonomic brackets. We have also included probes for other legacy exon loci (e.g., Dettai & Lecointre, 2005; Li et al., 2007; Lopez et al., 2004; Lovejoy et al., 2004) and mitochondrial DNA (mtDNA) markers that have been sequenced for a large number of fishes through PCR-Sanger sequencing methods to facilitate integration of new high-throughput sequencing results with existing phylogenetic data sets. We also provide a bioinformatic pipeline to assemble and filter sequence alignments of these exons from Illumina reads.

2 | MATERIALS AND METHODS

2.1 | Nuclear exon probes

Sequences for probe design came from exon alignments derived from a database of 303 bony fish genomes and transcriptomes (Hughes et al., 2018; Sun et al., 2016). Briefly, the *EvoMarkers* pipeline (Jiang et al., 2019; Li et al., 2012, 2015) was used to identify 1,721 single-copy exons in eight ray-finned fish genomes (*Lepisosteus oculatus*, *Anguilla japonica*, *Danio rerio*, *Gadus morhua*, *Oreochromis niloticus*, *Oryzias latipes*, *Tetraodon nigroviridis*, and *Gasterosteus aculeatus*). These exons were mined from 295 other genomes and transcriptomes using *nhmmer* (Wheeler & Eddy, 2013) in *HMMER* v3.1b2, and exons with paralogues were filtered by testing for duplications in gene trees via topology tests (see Hughes et al., 2018 for full details).

A total of 1,105 exons were retained after filtering for loci with paralogues. We generated seven probe sets for these exons based on different underlying references for our target groups (following

Betancur-R et al., 2017). These include (a) Elopomorpha (~1,000 species, including true eels and tarpons) (Figure 1); (b) early branching teleosts from Osteoglossomorpha (bonytongues) to Myctophiformes (lanternfishes)—hereafter paraphyletic “Backbone 1” (Figure 1); (c) Acanthomorphata (from paracanthopterygians (e.g., cods, oarfish) to Anabantaria (e.g., swamp eels, gouramies)—hereafter paraphyletic “Backbone 2”) (Figure 2); and four specific sets aimed for some of the most species-rich clades of Percomorphaceae, including (d) Carangaria (~1,100 species, including flatfishes and jacks) (Figure 3); (e) Ovalentaria (~5,600 species, including clownfishes, cichlids, flying fishes) (Figure 2); (f) Eupercaria (~6,800 species, including surgeonfishes, pufferfishes, and groupers) (Figure 3); and (g) Syngnatharia-Pelagiaria (~1,000 species, including tunas, seahorses, and pipefishes) (Figure 3). The large freshwater Otophysa clade (>10,000 species, including catfishes, knifefishes, and tetras) is not included in Backbone 1 (Figure 1), largely because it was targeted earlier by a more specific probe set designed for the clade by other exon-capture fish studies (Arcila et al., 2017; Betancur-R et al., 2019), though 143 exons are shared between the two. We designed probe sets for different subsets of taxa from these 1,105 alignments that initially consisted of 303 species that span the diversity of bony fishes (Hughes et al., 2018), as explained above. One particularly long exon included highly divergent sequences that were difficult to align and was ultimately excluded from the final probe sets (a total of 1,104 target exons remained). Each of the seven probe set references were comprised of four to eight of the most phylogenetically-distant taxa in the target clade, depending on the phylogenetic breadth the probe set covers (Table 1). We ranked preferred taxa within each of these groups to form the basis for the probe set references (Table 1), and if all preferred taxa were missing from a group, we took the next longest sequence in the alignment for the clade of interest. This means that some exons may have unique taxa representing them in their reference set.

We also included baits for nuclear markers popular in fish phylogenetics (referred to as “legacy” markers) to better connect sequence data sets produced by targeted amplicon sequencing approaches (Bybee et al., 2011). Several of these widely used markers were already included as part of the “paralogy-tested” 1,105 exons from Hughes et al. (2018), including RAG1 (Lopez et al., 2004), RAG2 (Lovejoy et al., 2004), FICD (Li et al., 2011), PANX2, GCS1, GLYT (Li et al., 2007), VCP1P (Betancur-R et al., 2013), and MLL (Dettai & Lecointre, 2005). A total of 19 additional legacy markers that did not meet the paralogy filtering requirements were nonetheless included in the probe sets, mainly markers developed by the Euteleost Tree of Life Project: TBR1, KIAA1239, MYH6, ENC1, PLAGL2, PTCHD1, RIPK4, SH3PX3, SIDKEY, SREB2, ZIC1, SVEP1, GPR61, SLC10A3, UBE3A, and UBE3A-like (Betancur-R et al., 2013; Li et al., 2007, 2011). Additionally, baits were designed for the markers Rhodopsin (Chen et al., 2003), IRBP (Dettai & Lecointre, 2008), and RNF213 (Li, Dettai, et al., 2009), which have been widely used in fish systematics. Due to the long sequences of MYH6 and KIAA1239 (>3,000 bp), references for these markers were shortened to the region typically amplified by PCR primers. The reference sequences used in

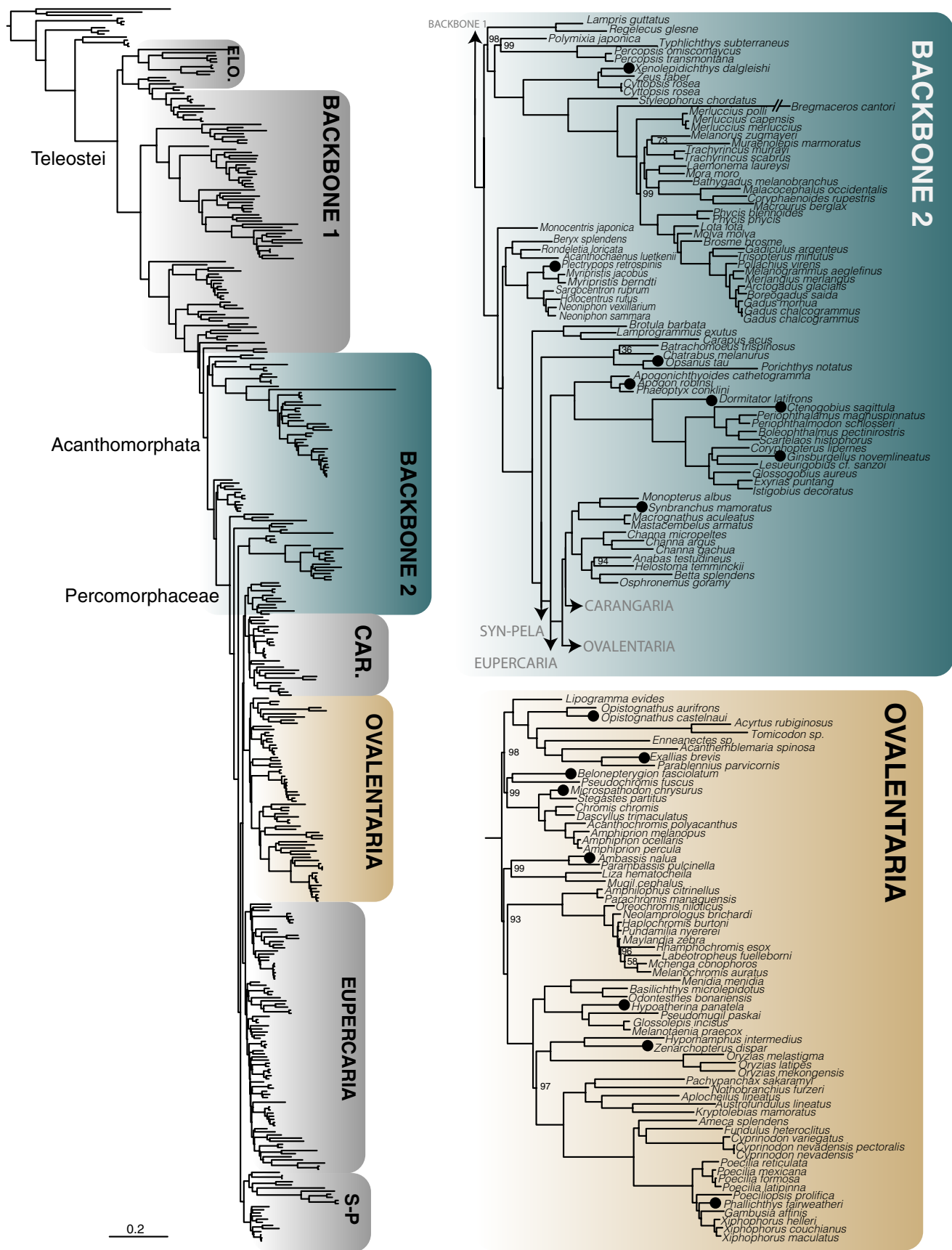


FIGURE 2 Position of “Backbone 2” and “Ovalentaria” probe sets in the maximum likelihood tree based on protein sequences from 394 fish taxa, combining genomes, transcriptomes, and exon capture. Newly sequenced taxa are represented with black dots at the tips. Bootstrap values are 100, unless otherwise noted [Colour figure can be viewed at wileyonlinelibrary.com]

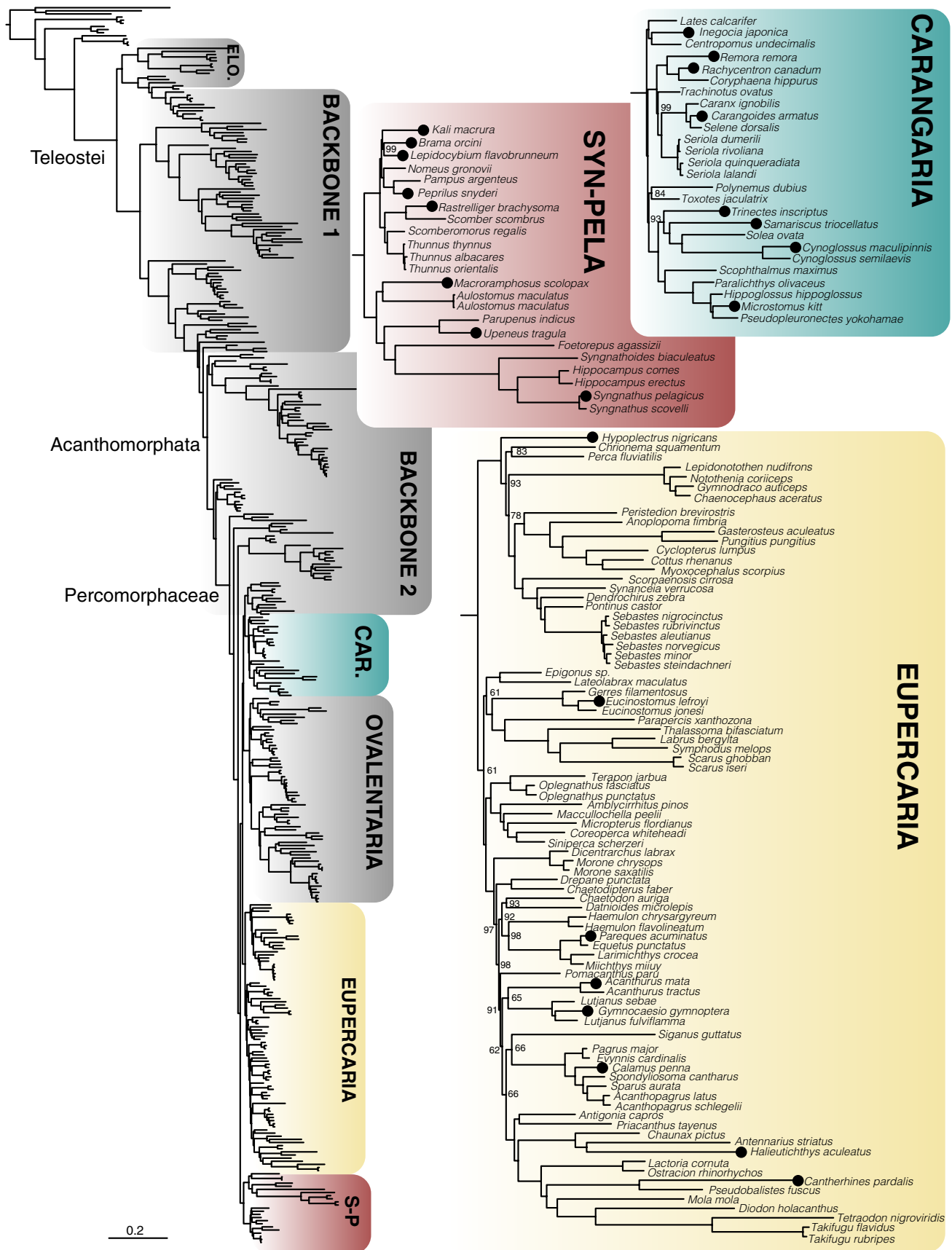


FIGURE 3 Position of "Carangaria," "Syngnatharia-Pelagiaria" (Syn-Pela) and "Eupercaria" probe sets in the maximum likelihood tree based on protein sequences from 394 fish taxa, combining genomes, transcriptomes, and exon capture. Newly sequenced taxa are represented with black dots. Bootstrap values are 100, unless otherwise noted [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Seven probe sets designed for 1,104 conserved exon markers across teleost fishes

Probe set name	Lineages included (preferred lineage in bold)
Elopomorpha	1. Megalopidae 2. Muraenidae 3. Congridae, Chlopsidae 4. Anguillidae
Backbone 1	1. Osteoglossidae, Pantodontidae 2. Notopteridae, Mormyridae 3. Engraulidae, Clupeidae 4. Galaxiidae 5. Argentinidae 6. Stomiidae, Osmeridae, Plecoglossidae, Salangidae 7. Synodontidae, Chlorophthalmidae 8. Myctophidae
Backbone 2	1. Zeidae, Parazenidae 2. Berycidae, Stephanoberycidae, Rondeletiidae 3. Holocentridae 4. Ophidiidae 5. Apogonidae 6. Gobiidae 7. Synbranchiformes, Anabantiformes
Syngnatharia-Pelagiaria	1. Syngnathidae, Callionymidae 2. Mullidae, Aulostomidae 3. Scombridae 4. Nomeidae, Stromateidae
Carangaria	1. Coryphaenidae, Carangidae 2. Cynoglossidae, Paralichthyidae, Pleuronectidae, Scophthalmidae, Soleidae 3. Centropomidae 4. Polynemidae, Toxotidae
Ovalentaria	1. Pseudomugilidae, Melanotaeniidae, Atherinopsidae 2. Aplocheilidae, Nothobranchiidae, Rivulidae, Cyprinodontidae, Fundulidae, Poeciliidae 3. Tripterygiidae, Blenniidae, Chaenopsidae 4. Gobiesocidae 5. Pomacentridae
Eupercaria	1. Anoplopomatidae, Channichthyidae, Cottidae, Gasterosteidae, Nototheniidae, Bathydraconidae, Percidae, Sebastidae 2. Gerreidae, Labridae, Pinguipedidae, Lateolabracidae, Epigonidae 3. Tetraodontidae, Molidae, Chaunacidae, Caproidae, Diodontidae, Antennariidae, Balistidae, Acanthuridae 4. Lutjanidae, Haemulidae, Chaetodontidae, Moronidae, Datnioididae, Ephippidae, Sciaenidae

bait design are available on GitHub (<https://github.com/lilychughes/FishLifeExonCapture/tree/master/ProbeSets>).

Probe sequences of 120 bp in length were initially generated with the `py_tiler.py` script packaged in PHYLUC for each of the four to eight taxa selected for probe design (Table 1) (Faircloth, 2015; Faircloth et al., 2012). Probes were mapped against the consensus sequences of the alignments from Hughes et al. (2018) in Geneious Pro version 8.1 (<http://www.geneious.com>) to examine the distribution of probes across loci. Visual inspection of the distribution of probes initially revealed highly uneven coverage of probes across longer loci. To have the probes cover the reference alignments more evenly, we applied a staggering strategy by tiling probes every 20 bp across each locus for each of the four to eight taxa (Table 1), so that probes from the first species spanned the first 0–120 bp, and probes from the second species spanned from 20–140, and so on. This strategy ultimately improved tiling density and resulted in more even coverage for longer loci in silico. The probe staggering design was generated via custom scripts (Jake Enk, Arbor Biosciences). Probes

that had more than 25% repeats detected on the RepeatMasker.org database were eliminated. Probes were filtered for potential self-hybridization. Four probe sets (Backbone 1, Backbone 2, Elopomorpha and Ovalentaria) had relatively higher GC content, and probes were reduced to 90 bp in length for these sets. Each of our eight probe sets was designed with a MyBaits1 custom probe set with approximately 20,000 biotinylated probes for each set (Arbor Biosciences, Ann Arbor, Michigan). Probe sets are publicly available at Arbor Biosciences, Ann Arbor, MI.

2.2 | Mitochondrial probes

In addition to exon probes, we designed and synthesized a separate, fish-universal probe set to capture four of the most popular mitochondrial DNA (mtDNA) gene markers used in fish systematics: COI (cytochrome c oxidase subunit I), CYTB (cytochrome *b*), and 12S and 16S ribosomal DNA. The goal of maintaining separate mtDNA and

nDNA probe sets is to equilibrate nDNA/mtDNA molar ratios by applying spiking dilutions of the mtDNA probe set after capturing the exon markers (mtDNA:nDNA dilution ratios = 1:1,000). Probes for these four mtDNA genes were individually designed using mtDNA genomes or single sequences from NCBI that span the diversity of ray-finned fishes (*Amia calva*, *Danio rerio*, *Elops saurus*, *Harengula clupeola*, *Harengula jaguana*, *Oryzias latipes*, *Osteoglossum bicirrhosum*, *Polypterus ansorgii*, *Salmo trutta*, *Takifugu vermicularis*, and *Zeus faber*). A total of 7,000 oligonucleotide baits (120 bp long) tiling over the mtDNA genes with 2x density were designed using the `py_tiler.py` script (Faircloth, 2015; Faircloth et al., 2012). We did not target other high-copy nonmitochondrial genes like 28S rDNA, which may have required an additional probe set and spiking dilution.

2.3 | Library preparation and sequencing

Eight fish species were newly sequenced for each bait set (Table 2; 56 total species sequenced). DNA extractions were performed on the GenePrep (Autogen Inc.) following the manufacturer's instructions at the Laboratory of Analytical Biology at the Smithsonian Institution National Museum of Natural History in Washington, D.C. DNA was eluted in 50 µl of Autogen R9 Buffer. Quality control was performed by running 1 µl of eluted DNA on a 1.0% agarose gel stained with GelRed (Biotium) and visually inspecting whether bands of high molecular weight DNA were visible. Library preparation was performed at Arbor Biosciences in Ann Arbor, Michigan, using a dual-round capture protocol (Li et al., 2013), with an 8-plex capture design. Paired-end sequencing of 100 bp reads was performed at the University of Chicago Genomics Facility on a HiSeq 4000. Samples were multiplexed with 192 in a lane, with sequencing runs contain-

ing samples for other projects not used here.

2.4 | Bioinformatics pipeline for exon assembly

We developed a bioinformatics pipeline based around the software aTRAM 2.0 (Allen et al., 2017) with five major steps before multiple sequence alignment (Figure 4). Raw FASTQ files were quality trimmed with Trimmomatic version 0.36 (Bolger et al., 2014), removing low quality sequences and adapter contamination with the parameters "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:31". Trimmed reads were then mapped against a master file containing all sequences used for bait design for any of the seven probe sets using BWA-MEM (Li & Durbin, 2009). SAMtools version 1.8 was used to remove optical PCR duplicates and separate the reads that mapped to each of the exons (Li, Handsaker, et al., 2009). Mapped reads were then assembled individually by exon using Velvet (Zerbino & Birney, 2008), and the longest contig produced by Velvet was used as a reference sequence for aTRAM version 2.0 (Allen et al., 2017) to extend contigs, using Trinity version 2.8.5 as the assembler (Grabherr et al., 2011).

Redundant contigs with 100% identity produced by aTRAM were removed with CD-HIT version 4.8.1 using CD-HIT-EST (Fu et al., 2012; Li & Godzik, 2006). Open reading frames for remaining contigs were identified with Exonerate (Slater & Birney, 2005), using a reference sequence checked by eye for each exon, and any contigs that did not contain the open reading frame were filtered out. If only a single contig contained the open reading frame, the exon passed all filters and was used for multiple sequence alignment. If multiple contigs contained the open reading frame, the reading frames were compared with CD-HIT-EST, using a 99% identity threshold to account for potential allelic variation. If the comparison with CD-HIT-EST resulted in a single contig, that contig passed filters and was used for phylogenetic analysis; more divergent sequences were flagged and excluded from downstream analysis. Unlike another tool recently published to assemble exon-capture data, ASSEXON (Yuan et al., 2019), our pipeline is fully automated and does not require using third-party packages (e.g., Geneious) as part of the assembly (<https://github.com/lilychughes/FishLifeExonCapture>).

2.5 | Alignment and phylogenomic analysis

Target-capture data were combined with genomic and transcriptomic data from Hughes et al. (2018) along with 36 additional recently published genomes that were mined for orthologous exon sequences using nhmmer (Wheeler & Eddy, 2013). Sequences for each exon were aligned with MACSE version 2.03 (Ranwez et al., 2018) after cleaning out potentially nonhomologous fragments with the `-cleanNonHomologousSequences` option. Alignment edges composed of more than 60% missing data as well as insertions that occurred only in a single taxon were removed with custom scripts (AlignmentCleaner.py, <https://github.com/lilychughes/FishLifeExonCapture>). A total of 1,104 nuclear exons filtered for paralogues were concatenated using geneStitcher.py (<https://github.com/ballesterus/Utensils>).

A concatenated protein matrix was analysed under maximum likelihood (ML) with IQ-TREE version 1.6.9 (Nguyen et al., 2015), using the best-fitting model for the entire matrix as determined using ModelFinder (Kalyaanamoorthy et al., 2017). A concatenated nucleotide matrix was partitioned into first, second, and third codon positions, with the best-fitting model applied to each partition. Concatenated matrices contained only the 1,104 loci that have been filtered for paralogy; the legacy markers were excluded from ML analyses.

2.6 | Paralogues in legacy markers

Nineteen nuclear markers commonly used with Sanger-sequencing methods for fish phylogenetics that had been previously excluded for having suspected paralogues, were reincluded in our probe sets to better connect novel sequence-capture data with large existing data sets. For the 18 legacy markers that were re-included (TBR1,

TABLE 2 Species sequenced for each probe set and the number of reads and loci assembled. Estimated clade ages (Hughes et al., 2018) are indicated

Probe set	Family	Taxon	Collection No.	Paired-end reads	Loci
Elopomorpha (196 Ma)	Albulidae	<i>Albula cf. vulpes</i>	^a USNM 421848	1,801,676	642
	Congridae	<i>Conger cinereus</i>	^b CSIRO GT7882	3,427,134	731
	Elopidae	<i>Elops hawaiiensis</i>	USNM 403422	3,020,903	759
	Halosauridae	<i>Halosaurus johnsonianus</i>	USNM 405058	2,034,292	606
	Halosauridae	<i>Halosaurus ovenii</i>	USNM 407039	4,743,893	719
	Nettastomatidae	<i>Nettastoma parviceps</i>	CSIRO GT6156	1,127,769	631
	Ophichthidae	<i>Myrophis microchir</i>	USNM 435225	1,362,661	586
	Synphobranchidae	<i>Meadia roseni</i>	CSIRO GT6877	2,734,565	735
Backbone1 (251 Ma)	Clupeidae	<i>Jenkinsia majua</i>	^c UPR FL0045	5,391,027	407
	Gonostomatidae	<i>Diplophos taenia</i>	USNM 405007	7,017,309	703
	Myctophidae	<i>Myctophum nitidulum</i>	USNM 405014	1,913,709	607
	Neoscopelidae	<i>Neoscopelus microchir</i>	USNM 407030	2,050,980	668
	Osteoglossidae	<i>Arapiama gigas</i>	USNM 440586	3,024,165	742
	Sternoptychidae	<i>Argyripnus atlanticus</i>	USNM 405229	2,328,143	651
	Stomiidae	<i>Chauliodus sloani</i>	USNM 405061	3,058,984	377
	Synodontidae	<i>Saurida gracilis</i>	^d STRI BFT11840	2,623,498	600
Backbone2 (144 Ma)	Apogonidae	<i>Apogon robinsi</i>	UPR FL0162	818,383	787
	Batrachoididae	<i>Opsanus tau</i>	STRI BFT09764	924,435	633
	Eleotridae	<i>Dormitator latifrons</i>	STRI BFT02768	1,584,877	671
	Gobiidae	<i>Ctenogobius sagittula</i>	STRI BFT18404	1,525,545	467
	Gobiidae	<i>Ginsburgellus novemlineatus</i>	UPR FL0141	1,265,097	597
	Grammicolepididae	<i>Xenolepidichthys dalgleishi</i>	USNM 407099	3,208,789	788
	Holocentridae	<i>Plectrypops retrospinis</i>	UPR FL0166/ MZUPRRP-I-00357	3,237,186	920
	Synbranchidae	<i>Synbranchus marmoratus</i>	STRI BFT05012	1,762,538	681
Carangaria (65 Ma)	Achiridae	<i>Trinectes inscriptus</i>	USNM 414275	979,351	988
	Carangidae	<i>Carangoides armatus</i>	USNM 435427	1,983,024	1,034
	Cynoglossidae	<i>Cynoglossus maculipinnis</i>	USNM 437669	1,387,678	868
	Echeneidae	<i>Remora remora</i>	USNM 405009	1,385,700	994
	Pleuronectidae	<i>Microstomus kitt</i>	USNM T5415	597,910	927
	Rachycentridae	<i>Rachycentron canadum</i>	USNM T3521	1,428,271	1,026
	Samaridae	<i>Samariscus triocellatus</i>	USNM 391219	216,271	637
	Sphyraenidae	<i>Inegocia japonica</i>	USNM T10332	2,968,482	1,012

(Continues)

TABLE 2 (Continued)

Probe set	Family	Taxon	Collection No.	Paired-end reads	Loci
Ovalentaria (95 Ma)	Ambassidae	<i>Ambassis nalua</i>	USNM 403430	2,666,551	1,017
	Atherinidae	<i>Hypoatherina panatela</i>	USNM 437959	945,584	878
	Blenniidae	<i>Exallias brevis</i>	USNM 390993	2,475,300	920
	Opistognathidae	<i>Opistognathus castelnaui</i>	USNM 435841	1,117,986	840
	Plesiopidae	<i>Belonepterygion fasciolatum</i>	USNM 432574	1,453,409	956
	Poeciliidae	<i>Phallichthys fairweatheri</i>	STRI BFT06906	1,984,926	911
	Pomacentridae	<i>Microspathodon chrysurus</i>	STRI BFT13594	2,156,316	985
	Zenarchopteridae	<i>Zenarchopterus dispar</i>	STRI BFT07992	1,744,848	916
Eupercaria (105 Ma)	Acanthuridae	<i>Acanthurus mata</i>	USNM 403159	2,439,554	1,005
	Gerreidae	<i>Eucinostomus lefroyi</i>	UPR FL0004/ MZUPRRP-I-00223	2,257,350	994
	Lutjanidae	<i>Gymnoaesio gymnoptera</i>	USNM 435461	1,944,110	953
	Monacanthidae	<i>Cantherhines pardalis</i>	USNM 435717	3,649,643	788
	Ogcocephalidae	<i>Halieutichthys aculeatus</i>	USNM 433145	930,425	765
	Sciaenidae	<i>Pareques acuminatus</i>	UPR FL0151/ MZUPRRP-I-00347	1,553,389	948
	Serranidae	<i>Hypoplectrus nigricans</i>	UPR FL0324/ MZUPRRP-I-00497	1,632,142	943
	Sparidae	<i>Calamus penna</i>	UPR FL0063/ MZUPRRP-I-00281	3,576,529	929
SynPela (96 Ma)	Bramidae	<i>Brama orcini</i>	USNM 403327	2,359,571	1,027
	Centriscidae	<i>Macroramphosus scolopax</i>	USNM 405231	2,823,407	1,034
	Chiasmodontidae	<i>Kali macrura</i>	USNM T2229	643,335	973
	Gempylidae	<i>Lepidocybium flavobrunneum</i>	USNM 407069	2,559,203	1,065
	Mullidae	<i>Upeneus tragula</i>	USNM 403208	3,042,224	918
	Scombridae	<i>Rastrelliger brachysoma</i>	USNM 409000	2,423,976	1,016
	Stromateidae	<i>Peprilus snyderi</i>	USNM 421333	1,581,129	1,044
	Syngnathidae	<i>Syngnathus pelagicus</i>	USNM 423115	1,244,405	867

^aUnited States National Museum (Smithsonian Institution), Washington, DC.

^bCommonwealth Scientific and Industrial Research Organisation, Tasmania, Australia.

^cZoological Museum at University of Puerto Rico-Rio Piedras, San Juan, PR. Specimens without a MZUPRRP number have no voucher due to their small size, only a field number beginning with FL.

^dSmithsonian Tropical Research Institute, Panama.

MYH6, KIAA1239, PLAGL2, PTCHD1, RIPK4, SH3PX3, SIDKEY, SREB2, ZIC1, SVEP1, GPR61, IRBP, RNF213, Rhodopsin, SLC10A3, UBE3A, and UBE3A-like), we integrated our newly sequenced data with the matrices from Betancur-R et al. (2013) for TBR1, MYH6,

KIAA1239, PLAGL2, PTCHD1, RIPK4, SH3PX3, SIDKEY, SREB2, ZIC1. For the remaining genes, we pulled a selection of sequences from GenBank for SVEP1, GPR61, IRBP, RNF213, Rhodopsin, SLC10A3, UBE3A, and UBE3A-like to align with our new data. All

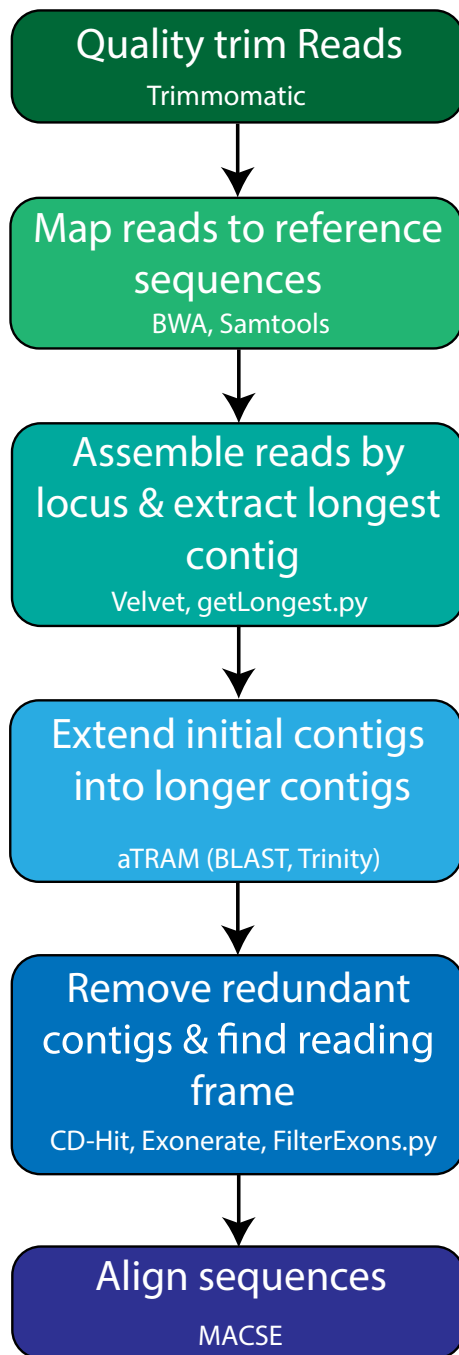


FIGURE 4 Major steps of bioinformatics pipeline and software used to transform data from Illumina reads to multiple sequence alignments. Scripts and pipeline tutorial are available on GitHub (<https://github.com/lilychughes/FishLifeExonCapture>) [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.com)]

GenBank accession numbers can be found on the sequence labels of these gene trees available on FigShare (Hughes et al., 2020). We inferred gene trees in IQ-TREE 1.6.9, partitioning by codon position and using the ModelFinder algorithm to determine the best-fitting sequence model for each partition. Target-capture-derived sequences falling out in unexpected positions or clades were BLASTed against the NCBI nucleotide database to determine their identity.

3 | RESULTS

3.1 | Capture efficiency of nuclear exons

The number of reads per sample varied substantially, from 216,271 to 7,017,309 (Table 2). However, the number of reads per sample was not correlated with the number of loci assembled per species ($r^2 = .0043$; $p = .27$). Capture efficiency (measured as the average number of exons assembled per species) varied across probe sets and across samples (Figure 5a), showing a strong negative correlation with clade (or paraphyletic “grade”) age ($r^2 = .82$; $p = .003131$; Figure 5b). Two probe sets designed to capture markers in paraphyletic groups that include more anciently diverging lineages (~251–144 Ma) with larger phylogenetic diversity (Backbone 1 and 2; Figures 1–2) tended to have lower capture efficiency (Table 2; Figure 5a), with an average of 52% of the loci captured for Backbone 1 and 61% on average captured for Backbone 2. This was also the case for the rather ancient Elopomorpha clade (196 Ma), for which samples sequenced had only a 60% capture rate. Probe sets designed for more recently diverged percomorph clades (~105–66 Ma) had much higher numbers of loci assembled on average, with 83% for Carangaria, 82% for Ovalentaria, 81% for Eupercaria, and 89% for the Syngnatharia + Pelagiaria clade. We looked at a number of other properties of the single-copy loci including number of parsimony informative sites, alignment length, GC content, and average melting temperature of the probes for each locus, but these properties did not seem to have a significant association with the probability of capture in this data set (Figures S1–S5). Exon alignments ranged from a minimum of 60% up to 93.9% sequence identity across the eight model genomes originally used to discover single-copy markers, but the legacy markers tended to have >80% sequence identity (Figure 5c).

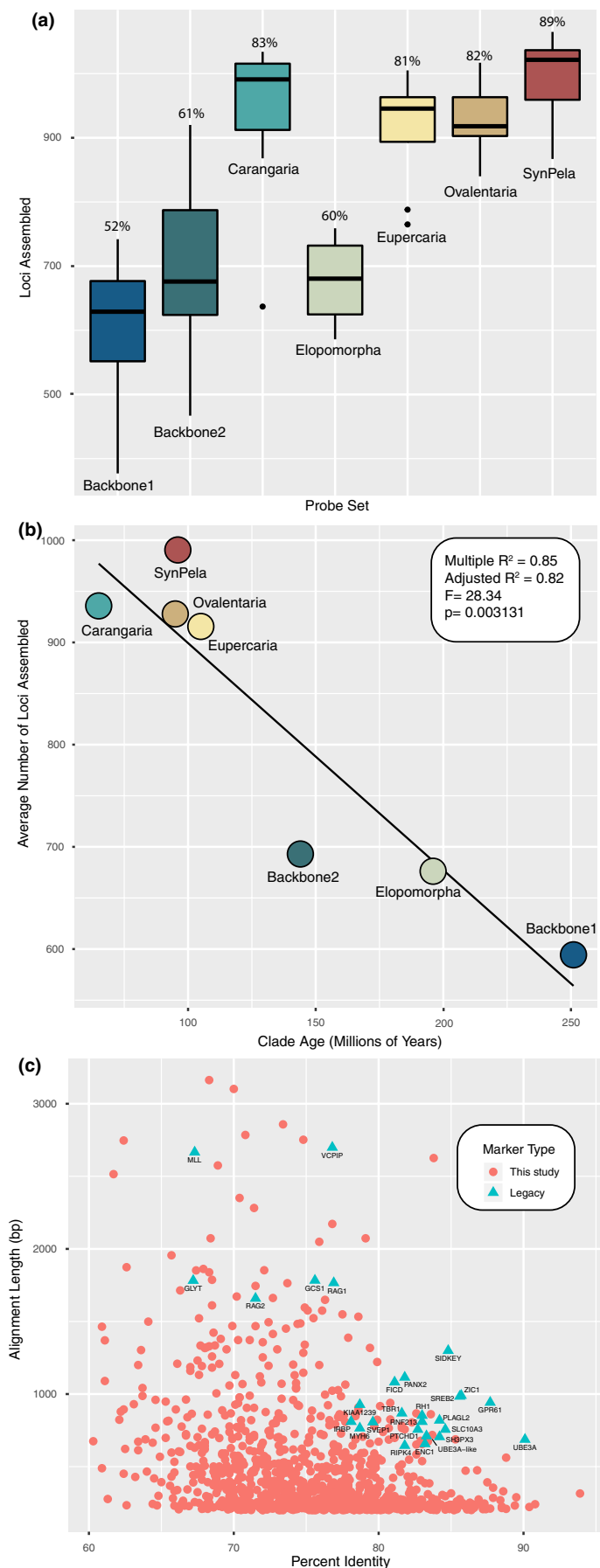
3.2 | Mitochondrial gene capture

Mitochondrial genes for which we designed probes (12S, 16S, COI, ATPase6, and CYTB) tended to have the best rate of capture. Complete sequences of 12S, 16S, and COI assembled for all taxa, while ND6, for which we did not design probes, was only represented for 29 taxa, the lowest of any mitochondrial coding gene. CYTB, which was included in the probe set, assembled for 56 of 58 total taxa, but ATPase6 had a relatively poor capture rate, only assembling for 36 species.

3.3 | Phylogenomic analysis

Combining the new data for 56 samples collected through exon-capture plus 38 additional recently published fish genomes with the data set of Hughes et al. (2018) generated a matrix with 394 taxa representing all major groups of ray-finned fishes with three lobe-finned (sarcopterygian) outgroups, with a final length of 549,861 bp (183,287 amino acid sites). The entire matrix had 72% present data, excluding loci suspected of having paralogues (Table 3). The average

FIGURE 5 (a) Box-and-whisker plots showing the variation in the number of exons captured for all taxa within each probe set. Eight newly sequenced taxa were used for each probe set. Percentages above each box refer to the average percentage of exons assembled by probe set. (b) The average number of loci captured for each of the seven probe sets plotted against the age of the most recent common ancestor of that particular group, showing higher capture of loci for younger clades. (c) Percent sequence identity and length across alignments of eight model genomes (*L. oculatus*, *A. japonica*, *D. rerio*, *G. morhua*, *O. latipes*, *O. niloticus*, *G. aculeatus*, and *T. nigroviridis*) for exons presented in this study (pink circles) compared to legacy markers (blue triangles) [Colour figure can be viewed at wileyonlinelibrary.com]



locus alignment length for genes included in the matrix was 499 bp (range: 129–5,055 bp), and the average number of parsimony-informative sites per locus was 340 (range: 75–3,435).

ModelFinder selected JTT + I+G for the protein matrix, and GTR + I+G for the first two codon positions, with TVM + I+G for the third codon position. Both topologies inferred with IQ-TREE matched previous results obtained by Hughes et al. (2018), with newly added taxa placed in their expected phylogenetic placements (Figures 1–3). Tree files and phylogenomic matrices are available on FigShare (Hughes et al., 2020).

3.4 | Paralogues in legacy markers

We visually examined gene trees for evidence of paralogues that had been assembled for 19 markers included in our probe set. Nine of these loci assembled one or more paralogues when we applied our new pipeline on the raw reads obtained with sequence-capture instead of a single orthologous locus (Table 3). One locus (PTCHD1) could only be assembled for seven of the 56 newly sequenced samples, which made the paralogy assessment difficult. Annotation of the paralogous sequences was determined by blasting assembled contigs against the NCBI nucleotide database.

TABLE 3 Legacy markers and their paralogues assembled from exon-capture data. Loci that did not assemble paralogues in this data set are denoted by “—”

Locus ID	Gene name	Paralogues assembled
E1541	TBR1	—
E1728	KIAA1239	—
E1730	MYH6	MYH7, MYH6-like, MYHb
E1732	ENC1	ENC2
E1735	PLAGL2	PLAGL1
E1736	PTCHD1	(Failed to assemble)
E1737	RIPK4	—
E1738	SH3PX3 (SNX33)	SNX18
E1739	SIDKEY	—
E1740	SREB2 (GPR85)	GPR173
E1741	ZIC1	ZIC4
E1746	SVEP1	—
E1747	GPR61	—
E1748	IRBP	—
E1749	RNF213	RNF213b-like, RNF213a-like
E1750	RH	EXORH
E1751	SLC10A3	—
E1752	UBEA3	—
E1753	UBEA3-like	UBEA3

4 | DISCUSSION

4.1 | Probe sets for exon capture across deep phylogenetic divergences

We present resources for capturing conserved exon sequences for all groups of teleost fishes, including probe sets for early branching lineages (“Backbone 1”), Elopomorpha, Acanthomorphata (“Backbone 2”), and multiple major percomorph radiations including Syngnatharia + Pelagiaria, Carangaria, Ovalentaria, and Eupercaria. The exon markers presented here have been explicitly tested for orthology using a large database of 303 bony fish species (Hughes et al., 2018), and have been screened for paralogues derived from ancient vertebrate or teleost-specific whole genome duplication events. Capture efficiency is strongly correlated with the phylogenetic span of taxa used to design the probes, with probe sets designed for relatively younger (~105–66 Ma) percomorph clades (Syngnatharia-Pelagiaria, Carangaria, Ovalentaria, Eupercaria) capturing 200–300 more loci on average than those designed for more ancient (~251–144 Ma) and/or taxonomically disparate clades (Elopomorpha and Backbones 1 and 2; Figure 5b). Estimates for the divergence times of major percomorph series vary, but the youngest estimates place their origin near the Cretaceous-Paleogene boundary, 66 million years ago (Alfaro et al., 2018). Conversely, the paraphyletic taxonomic groups spanned by Backbone 1 diverged in the Permian or Triassic, and the late Jurassic to early Cretaceous for Backbone 2 (Betancur-R et al., 2013; Hughes et al., 2018; Near et al., 2013). The larger number of nucleotide substitutions accumulated across older clades causes the probes to have less affinity for the targeted DNA regions in vitro, and we noticed a substantial increase in the number of loci captured for those clades younger than 100 million years (Figure 5b). We examined other characteristics of the loci and probes that could be useful for other researchers to consider when designing their own probe sets for exon capture. These properties included number of parsimony informative sites, alignment length, GC content, and average melting temperature of the probes for a particular locus, but none of them were substantially correlated with capture efficiency in any of the seven probe sets (Supporting Information). While loci that failed to capture for a particular probe set tended to be shorter, and have higher probe melting temperatures, we were also successful at capturing many loci with these same characteristics. These results suggest that selecting more closely related taxa for probe set design provides a useful strategy to improve capture efficiency for projects targeting more specific clades.

Despite the variation in the number of loci assembled, all samples with exon-capture data were resolved in their expected clades, and the ML topologies at major nodes matched that of the ML topologies of Hughes et al. (2018). One family, Clupeidae, was not monophyletic, but this result has been reported in previous phylogenetic analyses (Betancur-R et al., 2017), and may reflect the need of taxonomic revision or insufficient taxonomic sampling rather than underlying phylogenetic estimation error arising from the exon-capture data.

These markers appear to be informative for deep divergences in fishes, and the backbone of the ray-finned fish tree largely matches inferences based on legacy gene markers (Betancur-R et al., 2013; Near et al., 2012), though many areas of the tree have only been investigated with sparse taxonomic sampling and will require more thorough investigation with additional sequencing. While deep divergences are the focus of this paper, conserved exon markers have also been shown to contain information appropriate for shallower divergences at the phylogeographic level (Rincon-Sandoval et al., 2019). The flanking intron regions, which are highly variable, have been removed for the analyses presented here, but we include a branch of our bioinformatic pipeline to additionally use the flanking intron sequences for projects with a more recent evolutionary focus.

4.2 | Exon markers can be integrated with existing and future data sets

Taxonomic sampling is critical for accurate phylogenetic analysis (Betancur-R et al., 2019; Heath et al., 2008), and sequence capture methods are a cost-effective approach for increasing taxonomic sampling across a large number of loci (Lemmon & Lemmon, 2013). But both whole-genome sequences (Malmström et al., 2016; Musilova et al., 2019) and transcriptome sequences (e.g., Dai et al., 2018; Hughes et al., 2018) are becoming available for an increasing diversity of fish species. These exon markers can be easily mined from public transcriptome or genome data as they become available, increasing taxonomic sampling for the group of interest without duplicating sequencing efforts. Taxonomically dense super-matrix approaches in fishes (e.g., Rabosky et al., 2018) primarily rely on exon sequences deposited in NCBI. Currently, there are more than 20,000 sequences of RAG1 for teleosts available in NCBI (as of 20 March 2019), more than 35,000 teleost rhodopsin sequences, and even larger numbers for mitochondrial genes like CYTB (>130,000 sequences). This is a rich resource that can be combined with exon-capture data for the probe sets described here to reduce missing data that are often rampant in super matrix approaches but still produce taxonomically dense trees (Cho et al., 2011).

4.3 | Paralogues in sequence capture data sets

Many nuclear exon fragments that have been in wide use in fish phylogenetics for more than a decade do not appear to have paralogues, and new sequence capture data could be easily integrated from genes like RAG1, RAG2, PANX2, MLL, VCPIP, GLYT, GCS1, and FICD. Many of these exons were defined as "single-copy" based on the comparisons of the relatively few fish genomes available at the time (Li et al., 2007). But the specificity of primers designed for nested PCR approaches to amplify and sequence these loci has been a successful strategy to obtain orthologous genes for phylogenetic inference in fishes (Betancur-R et al., 2013; Li et al., 2007, 2008,

2010; Near et al., 2012, 2013; Wainwright et al., 2012). Shotgun sequencing of enriched libraries, in contrast, is a more challenging approach for assembling orthologous genes, since sequence reads of two or more paralogous copies may be sequenced by this approach and need to be separated using bioinformatic pipelines. Nineteen legacy markers included in our probe set previously had been excluded from downstream phylogenetic analyses due to paralogy issues detected either by comparing additional genomes or by topology tests of gene trees (Table 3). Due to high similarity in certain parts of the coding region to the reference coding sequence used in Exonerate, more than half of these assembled paralogous loci passed through to the alignment stage. Often it was only the paralogue that was assembled, and the assembly of multiple contigs was not a reliable way to flag paralogy. The pipeline implemented here (Figure 4) attempts to remove redundant contigs with CD-HIT at a 99% similarity threshold across the reading frame when multiple contigs assemble, but exons that fail this test are not passed onto the alignment stage. Paralogues of ENC1, MYH6, ZIC1 and other genes known to be duplicated (Table 3) passed on to the alignment stage, and do not appear to have assembled the orthologous sequence. However, a majority of the sequences assembled orthologous exons. With additional scrutiny for paralogues using gene trees, these data are still quite useful for integration with older data sets.

ACKNOWLEDGEMENTS

This research was supported by National Science Foundation (NSF) grants NSF-DEB-1929248 and NSF-DEB-1932759 to R.B.R., NSF-DEB-1541554 and NSF-DEB-1457426 to G.O., NSF-DEB-1541552 to C.C.B., and NSF-DEB-2015404 to D.A. We are grateful to Jake Enk (Arbor Biosciences) for his assistance with probe design. We thank Rose Peterson and Victoria Rodriguez for assistance with laboratory work, and Diane Pitassy for access to tissues at the USNM. All data processing and phylogenetic analysis were conducted on the Pegasus HPC cluster at George Washington University and the HPC facility at University of Puerto Rico-Rio Piedras (funded by INBRE Grant Number P20GM103475). We thank five anonymous reviewers for their comments on this manuscript.

AUTHOR CONTRIBUTIONS

L.C.H., R.B.R., G.O., K.A.C., C.L., and D.A. contributed to the design of the study. R.B.R., C.C.B., and W.W. provided tissues. L.C.H., H.S., C.L., and R.B.R. analysed the data. L.C.H., G.O., and R.B.R. wrote the paper with input from all authors.

DATA AVAILABILITY STATEMENT

Raw reads for newly sequenced exon-capture data are archived on NCBI under Bioproject number PRJNA605876. Newick files and phylogenomic matrices are available on FigShare (<https://doi.org/10.6084/m9.figshare.11844783>), and pipeline scripts to analyse data and a tutorial are available on GitHub (<https://github.com/lilychughes/FishLifeExonCapture>). The protein tree topology will be made available on Open Tree of Life.

ORCID

Lily C. Hughes  <https://orcid.org/0000-0003-4006-4036>

Chenhong Li  <https://orcid.org/0000-0003-3075-1756>

Keith A. Crandall  <https://orcid.org/0000-0002-0836-3389>

REFERENCES

- Abascal, F., Zardoya, R., & Telford, M. J. (2010). TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research*, 38, W7–W13. <https://doi.org/10.1093/nar/gkq291>
- Alfaro, M. E., Faircloth, B. C., Harrington, R. C., Sorenson, L., Friedman, M., Thacker, C. E., Oliveros, C. H., Černý, D., & Near, T. J. (2018). Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. *Nature Ecology & Evolution*, 2(4), 688–696. <https://doi.org/10.1038/s41559-018-0494-6>
- Allen, J. M., Boyd, B., Nguyen, N.-P., Vachaspati, P., Warnow, T., Huang, D. I., Grady, P. G. S., Bell, K. C., Cronk, Q. C. B., Mugisha, L., Pittendrigh, B. R., Soledad Leonardi, M., Reed, D. L., & Johnson, K. P. (2017). Phylogenomics from Whole Genome Sequences Using aTRAM. *Systematic Biology*, 66, 786–798. <https://doi.org/10.1093/sysbio/syw105>
- Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L. J., Ko, K. D., Sabaj, M. H., Lundberg, J., Revell, L. J., & Betancur-R, R. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, 1, 20. <https://doi.org/10.1038/s41559-016-0020>
- Betancur-R, R., Arcila, D., Vari, R. P., Hughes, L. C., Oliveira, C., Sabaj, M. H., & Ortí, G. (2019). Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: The monophyly of characiform fishes. *Evolution*, 73, 329–345. <https://doi.org/10.1111/evo.13649>
- Betancur-R, R., Wiley, E. O., Arratia, G., Acero, A., Bailly, N., Miya, M., Lecointre, G., & Ortí, G. (2017). Phylogenetic classification of bony fishes. *BMC Evolutionary Biology*, 17, 162. <https://doi.org/10.1186/s12862-017-0958-3>
- Betancur-R, R., Broughton, R. E., Wiley, E. O., Carpenter, K., López, J. A., Li, C., Holcroft, N. I., Arcila, D., Sanciangco, M., Cureton, J. C. II, Zhang, F., Buser, T., Campbell, M. A., Ballesteros, J. A., Roa-Varon, A., Willis, S., Borden, W. C., Rowley, T., Reneau, P. C., ... Ortí, G. (2013). The Tree of life and a new classification of bony fishes. *PLoS Currents*, 5, 1–45. <https://doi.org/10.1371/currents.tol.53ba26640df0ccee75bb165c8c26288>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Braasch, I., Peterson, S. M., Desvignes, T., McCluskey, B. M., Batzel, P., & Postlethwait, J. H. (2015). A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 324, 316–341.
- Brown, J. M., & Thomson, R. C. (2017). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology*, 66, 517–530.
- Bybee, S. M., Bracken-Grissom, H., Haynes, B. D., Hermansen, R. A., Byers, R. L., Clement, M. J., Udall, J. A., Wilcox, E. R., & Crandall, K. A. (2011). Targeted amplicon sequencing (TAS): A scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution*, 3, 1312–1323. <https://doi.org/10.1093/gbe/evr106>
- Chakrabarty, P., Faircloth, B. C., Alda, F., Ludt, W. B., McMahan, C. D., Near, T. J., Dornburg, A., Albert, J. S., Arroyave, J., Stiassny, M. L. J., Sorenson, L., & Alfaro, M. E. (2017). Phylogenomic systematics of ostariophysan fishes: Ultraconserved elements support the surprising non-monophyly of characiformes. *Systematic Biology*, 66, 881–895. <https://doi.org/10.1093/sysbio/syx038>
- Chen, W.-J., Bonillo, C., & Lecointre, G. (2003). Repeatability of clades as a criterion of reliability: A case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. *Molecular Phylogenetics and Evolution*, 26, 262–288. [https://doi.org/10.1016/S1055-7903\(02\)00371-8](https://doi.org/10.1016/S1055-7903(02)00371-8)
- Cho, S., Zwick, A., Regier, J. C., Mitter, C., Cummings, M. P., Yao, J., Du, Z., Zhao, H., Kawahara, A. Y., Weller, S., Davis, D. R., Baixeras, J., Brown, J. W., & Parr, C. (2011). Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? *Systematic Biology*, 60, 782–796. <https://doi.org/10.1093/sysbio/syr079>
- Clark, J. W., & Donoghue, P. C. J. (2018). Whole-genome duplication and plant macroevolution. *Trends in Plant Science*, 23, 933–945. <https://doi.org/10.1016/j.tplants.2018.07.006>
- Clarke, T. H., Garb, J. E., Hayashi, C. Y., Arensburg, P., & Ayoub, N. A. (2015). Spider transcriptomes identify ancient large-scale gene duplication event potentially important in silk gland evolution. *Genome Biology and Evolution*, 7, 1856–1870. <https://doi.org/10.1093/gbe/evv110>
- Dai, W., Zou, M., Yang, L., Du, K., Chen, W., Shen, Y., Mayden, R. L., & He, S. (2018). Phylogenomic perspective on the relationships and evolutionary history of the major otocephalan lineages. *Scientific Reports*, 8, 205. <https://doi.org/10.1038/s41598-017-18432-5>
- Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate (P Holland, Ed.). *PLoS Biology*, 3, e314. <https://doi.org/10.1371/journal.pbio.0030314>
- Detta, A., & Lecointre, G. (2005). Further support for the clades obtained by multiple molecular phylogenies in the acanthomorph bush. *Comptes Rendus - Biologies*, 328, 674–689. <https://doi.org/10.1016/j.crv.2005.04.002>
- Detta, A., & Lecointre, G. (2008). New insights into the organization and evolution of vertebrate IRBP genes and utility of IRBP gene sequences for the phylogenetic study of the Acanthomorpha (Actinopterygii: Teleostei). *Molecular Phylogenetics and Evolution*, 48, 258–269. <https://doi.org/10.1016/j.ympev.2008.04.003>
- Dornburg, A., Townsend, J. P., Brooks, W., Spriggs, E., Eytan, R. I., Moore, J. A., Wainwright, P. C., Lemmon, A., Lemmon, E. M., & Near, T. J. (2017). New Insights on the sister lineage of percomorph fishes with an anchored hybrid enrichment dataset. *Molecular Phylogenetics and Evolution*, 110, 27–38. <https://doi.org/10.1016/j.ympev.2017.02.017>
- Edwards, S. V., Cloutier, A., & Baker, A. J. (2017). Conserved nonexonic elements: A novel class of marker for phylogenomics. *Systematic Biology*, 66, 1028–1044. <https://doi.org/10.1093/sysbio/syx058>
- Eytan, R. I., Evans, B. R., Dornburg, A., Lemmon, A. R., Moriarty Lemmon, E., Wainwright, P. C., & Near, T. J. (2015). Are 100 enough? Inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. *BMC Evolutionary Biology*, 15, 113. <https://doi.org/10.1186/s12862-015-0415-0>
- Faircloth, B. C. (2015). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, 32, 786–788. <https://doi.org/10.1093/bioinformatics/btv646>
- Faircloth, B. C., Alda, F., Hoekzema, K., Burns, M. D., Oliveira, C., Albert, J. S., Melo, B. F., Ochoa, L. E., Roxo, F. F., Chakrabarty, P., Sidlauskas, B. L., & Alfaro, M. E. (2020). A target enrichment bait set for studying relationships among ostariophysan fishes. *Copeia*, 108, 47–60. <https://doi.org/10.1643/CG-18-139>
- Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, 61, 717–726. <https://doi.org/10.1093/sysbio/sys004>
- Faircloth, B. C., Sorenson, L., Santini, F., & Alfaro, M. E. (2013). A phylogenomic perspective on the radiation of ray-finned fishes based upon

- targeted sequencing of ultraconserved elements (UCEs). *PLoS One*, 8, e65923. <https://doi.org/10.1371/journal.pone.0065923>
- Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19, 99–113. <https://doi.org/10.2307/2412448>
- Friedman, M., Feilich, K. L., Beckett, H. T., Alfaro, M. E., Faircloth, B. C., Černý, D., Miya, M., Near, T. J., & Harrington, R. C. (2019). A phylogenomic framework for pelagiarian fishes (Acanthomorpha: Percomorpha) highlights mosaic radiation in the open ocean. *Proceedings of the Royal Society B: Biological Sciences*, 286, 20191502.
- Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28, 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29, 644–652. <https://doi.org/10.1038/nbt.1883>
- Harrington, R. C., Faircloth, B. C., Eytan, R. I., Smith, W. L., Near, T. J., Alfaro, M. E., & Friedman, M. (2016). Phylogenomic analysis of carangimorph fishes reveals flatfish asymmetry arose in a blink of the evolutionary eye. *BMC Evolutionary Biology*, 16, 224. <https://doi.org/10.1186/s12862-016-0786-x>
- Heath, T. A., Hedtke, S. M., & Hillis, D. M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution*, 46, 239–257.
- Hughes, L. C., Ortí, G., Huang, Y. U., Sun, Y., Baldwin, C. C., Thompson, A. W., Arcila, D., Betancur-R, R., Li, C., Becker, L., Bellora, N., Zhao, X., Li, X., Wang, M., Fang, C., Xie, B., Zhou, Z., Huang, H., Chen, S., ... Shi, Q. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 6249–6254. <https://doi.org/10.1073/pnas.1719358115>
- Hughes, L. C., Ortí, G., Saad, H., Li, C., White, W. T., Baldwin, C. C., Crandall, K. A., Arcila, D., & Betancur-R, R. (2020). Data from: Exon probe sets and bioinformatics pipelines for all levels of fish phylogenomics. *FigShare*, <https://doi.org/10.6084/m9.figshare.11844783>
- Illves, K. L., & López-Fernández, H. (2014). A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. *Molecular Ecology Resources*, 14, 802–811. <https://doi.org/10.1111/1755-0998.12222>
- Illves, K. L., Torti, D., & López-Fernández, H. (2017). Exon-based phylogenomics strengthens the phylogeny of Neotropical cichlids and identifies remaining conflicting clades (Cichliformes: Cichlidae: Cichlinae). *Molecular Phylogenetics and Evolution*, 118, 232–243.
- Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K., & Nishida, M. (2015). Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 14918–14923. <https://doi.org/10.1073/pnas.1507669112>
- Irisarri, I., Singh, P., Koblmüller, S., Torres-Dowdall, J., Henning, F., Franchini, P., Fischer, C., Lemmon, A. R., Lemmon, E. M., Thallinger, G. G., Sturmbauer, C., & Meyer, A. (2018). Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nature Communications*, 9, 3159. <https://doi.org/10.1038/s41467-018-05479-9>
- Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M., & Nishida, M. (2013). MitoFish and MitoAnnotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution*, 30, 2531–2540. <https://doi.org/10.1093/molbev/mst141>
- Jiang, J., Yuan, H., Zheng, X., Wang, Q., Kuang, T., Li, J., Liu, J., Song, S., Wang, W., Cheng, F., Li, H., Huang, J., & Li, C. (2019). Gene markers for exon capture and phylogenomics in ray-finned fishes. *Ecology and Evolution*, 9, 3973–3983. <https://doi.org/10.1002/ece3.5026>
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermini, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14, 587–589. <https://doi.org/10.1038/nmeth.4285>
- Kenny, N. J., Chan, K. W., Nong, W., Qu, Z., Maeso, I., Yip, H. Y., Chan, T. F., Kwan, H. S., Holland, P. W. H., Chu, K. H., & Hui, J. H. L. (2016). Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity*, 116, 190–199. <https://doi.org/10.1038/hdy.2015.89>
- Kocot, K. M., Citarella, M. R., Moroz, L. L., & Halanynch, K. M. (2013). PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary Bioinformatics*, 9, 429–435. <https://doi.org/10.4137/EBO.S12813>
- Kuang, T., Tornabene, L., Li, J., Jiang, J., Chakrabarty, P., Sparks, J. S., Naylor, G. J. P., & Li, C. (2018). Phylogenomic analysis on the exceptionally diverse fish clade Gobioidae (Actinopterygii: Gobiiformes) and data-filtering based on molecular clocklikeness. *Molecular Phylogenetics and Evolution*, 128, 192–202. <https://doi.org/10.1016/j.ympev.2018.07.018>
- Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, 61, 727–744. <https://doi.org/10.1093/sysbio/sys049>
- Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44, 99–121. <https://doi.org/10.1146/annurev-ecolsys-110512-135822>
- Li, B., Dettai, A., Cruaud, C., Couloux, A., Desoutter-Meniger, M., & Lecointre, G. (2009). RNF213, a new nuclear marker for acanthomorph phylogeny. *Molecular Phylogenetics and Evolution*, 50, 345–363. <https://doi.org/10.1016/j.ympev.2008.11.013>
- Li, C., Corrigan, S., Yang, L., Straube, N., Harris, M., Hofreiter, M., White, W. T., & Naylor, G. J. P. (2015). DNA capture reveals transoceanic gene flow in endangered river sharks. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 13302–13307. <https://doi.org/10.1073/pnas.1508735112>
- Li, C., Hofreiter, M., Straube, N., Corrigan, S., & Naylor, G. J. P. (2013). Capturing protein-coding genes across highly divergent species. *BioTechniques*, 54, 321–326. <https://doi.org/10.2144/000114039>
- Li, C., Lu, G., & Ortí, G. (2008). Optimal data partitioning and a test case for ray-finned fishes (actinopterygii) based on ten nuclear loci (T Buckley, Ed.). *Systematic Biology*, 57, 519–539. <https://doi.org/10.1080/10635150802206883>
- Li, C., Ortí, G., Zhang, G., & Lu, G. (2007). A practical approach to phylogenomics: The phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evolutionary Biology*, 7, 44. <https://doi.org/10.1186/1471-2148-7-44>
- Li, C., Ortí, G., & Zhao, J. (2010). The phylogenetic placement of siniperid fishes ("Perciformes") revealed by 11 nuclear loci. *Molecular Phylogenetics and Evolution*, 56, 1096–1104. <https://doi.org/10.1016/j.ympev.2010.05.017>
- Li, C., Ricardo, B.-R., Smith, W. L., & Ortí, G. (2011). Monophyly and interrelationships of Snook and Barramundi (Centropomidae sensu Greenwood) and five new markers for fish phylogenetics. *Molecular Phylogenetics and Evolution*, 60, 463–471. <https://doi.org/10.1016/j.ympev.2011.05.004>
- Li, C., Riethoven, J. J. M., & Naylor, G. J. P. (2012). EvolMarkers: A database for mining exon and intron markers for evolution, ecology

- and conservation studies. *Molecular Ecology Resources*, 12, 967–971. <https://doi.org/10.1111/j.1755-0998.2012.03167.x>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>
- Li, Z., Tiley, G. P., Galuska, S. R., Reardon, C. R., Kidder, T. I., Rundell, R. J., & Barker, M. S. (2018). Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 4713–4718. <https://doi.org/10.1073/pnas.1710791115>
- Longo, S. J., Faircloth, B. C., Meyer, A., Westneat, M. W., Alfaro, M. E., & Wainwright, P. C. (2017). Phylogenomic analysis of a rapid radiation of misfit fishes (Synnathiformes) using ultraconserved elements. *Molecular Phylogenetics and Evolution*, 113, 33–48. <https://doi.org/10.1016/j.ympev.2017.05.002>
- Lopez, J. A., Chen, W.-J., & Ortí, G. (2004). Esociform phylogeny. *Copeia*, 3, 449–464. <https://doi.org/10.1643/CG-03-087R1>
- Lovejoy, N. R., Iranpour, M., & Collette, B. B. (2004). Phylogeny and jaw ontogeny of beloniform fishes. *Integrative and Comparative Biology*, 44, 366–377. <https://doi.org/10.1093/icb/44.5.366>
- Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., Baalsrud, H. T., Nederbragt, A. J., Hanel, R., Salzburger, W., Stenseth, N. C., Jakobsen, K. S., & Jentoft, S. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, 48, 1204–1210. <https://doi.org/10.1038/ng.3645>
- Miya, M., Takeshima, H., Endo, H., Ishiguro, N. B., Inoue, J. G., Mukai, T., Satoh, T. P., Yamaguchi, M., Kawaguchi, A., Mabuchi, K., Shirai, S. M., & Nishida, M. (2003). Major patterns of higher teleostean phylogenies: A new perspective based on 100 complete mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, 26, 121–138. [https://doi.org/10.1016/S1055-7903\(02\)00332-9](https://doi.org/10.1016/S1055-7903(02)00332-9)
- Musilova, Z., Cortesi, F., Matschiner, M., Davies, W. I. L., Patel, J. S., Stieb, S. M., de Busserolles, F., Malmstrøm, M., Tørresen, O. K., Brown, C. J., Mountford, J. K., Hanel, R., Stenkamp, D. L., Jakobsen, K. S., Carleton, K. L., Jentoft, S., Marshall, J., & Salzburger, W. (2019). Vision using multiple distinct rod opsins in deep-sea fishes. *Science*, 364, 588–592. <https://doi.org/10.1126/science.aav4632>
- Near, T. J., Dornburg, A., Eytan, R. I., Keck, B. P., Smith, W. L., Kuhn, K. L., Moore, J. A., Price, S. A., Burbrink, F. T., Friedman, M., & Wainwright, P. C. (2013). Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 12738–12743. <https://doi.org/10.1073/pnas.1304661110>
- Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., Davis, M. P., Wainwright, P. C., Friedman, M., & Smith, W. L. (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 13698–13703. <https://doi.org/10.1073/pnas.1206625109>
- Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32, 268–274. <https://doi.org/10.1093/molbev/msu300>
- Owen, C. L., Stern, D. B., Hilton, S. K., & Crandall, K. A. (2020). Hemiptera phylogenomic resources: Tree-based orthology prediction and conserved exon identification. *Molecular Ecology Resources*, 20, 13180. <https://doi.org/10.1111/1755-0998.13180>
- Philippe, H., de Vienne, D. M., Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, 283. <https://doi.org/10.5852/ejt.2017.283>
- Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T. J., Coll, M., & Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, 559, 392–395. <https://doi.org/10.1038/s41586-018-0273-1>
- Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., & Delsuc, F. (2018). MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, 35, 2582–2584. <https://doi.org/10.1093/molbev/msy159>
- Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple alignment of coding SEquences accounting for frameshifts and stop codons. *PLoS One*, 6, e22594. <https://doi.org/10.1371/journal.pone.0022594>
- Rincon-Sandoval, M., Betancur-R, R., & Maldonado-Ocampo, J. A. (2019). Comparative phylogeography of trans-Andean freshwater fishes based on genome-wide nuclear and mitochondrial markers. *Molecular Ecology*, 28, 1096–1115. <https://doi.org/10.1111/mec.15036>
- Roxo, F. F., Ochoa, L. E., Sabaj, M. H., Lujan, N. K., Covain, R., Silva, G. S. C., Melo, B. F., Albert, J. S., Chang, J., Foresti, F., Alfaro, M. E., & Oliveira, C. (2019). Phylogenomic reappraisal of the Neotropical catfish family Loricariidae (Teleostei: Siluriformes) using ultraconserved elements. *Molecular Phylogenetics and Evolution*, 135, 148–165. <https://doi.org/10.1016/j.ympev.2019.02.017>
- Sato, Y., Miya, M., Fukunaga, T., Sado, T., & Iwasaki, W. (2018). MitoFish and MiFish pipeline: A mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Molecular Biology and Evolution*, 35, 1553–1555. <https://doi.org/10.1093/molbev/msy074>
- Schwager, E. E., Sharma, P. P., Clarke, T., Leite, D. J., Wierschin, T., Pechmann, M., Akiyama-Oda, Y., Esposito, L., Bechsgaard, J., Bilde, T., Buffry, A. D., Chao, H., Dinh, H., Doddapaneni, H. V., Dugan, S., Eibner, C., Extavour, C. G., Funch, P., Garb, J., ... McGregor, A. P. (2017). The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biology*, 15, 62. <https://doi.org/10.1186/s12915-017-0399-x>
- Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, 1–11.
- Song, S., Zhao, J., & Li, C. (2017). Species delimitation and phylogenetic reconstruction of the siniperids (Perciformes: Siniperidae) based on target enrichment of thousands of nuclear coding sequences. *Molecular Phylogenetics and Evolution*, 111, 44–55. <https://doi.org/10.1016/j.ympev.2017.03.014>
- Stout, C. C., Tan, M., Lemmon, A. R., Lemmon, E. M., & Armbruster, J. W. (2016). Resolving Cypriniformes relationships using an anchored enrichment approach. *BMC Evolutionary Biology*, 16, 244. <https://doi.org/10.1186/s12862-016-0819-5>
- Straube, N., Li, C., Mertzen, M., Yuan, H., & Moritz, T. (2018). A phylogenomic approach to reconstruct interrelationships of main clupeocephalan lineages with a critical discussion of morphological apomorphies. *BMC Evolutionary Biology*, 18, 158. <https://doi.org/10.1186/s12862-018-1267-1>
- Sun, Y., Huang, Y. U., Li, X., Baldwin, C. C., Zhou, Z., Yan, Z., Crandall, K. A., Zhang, Y., Zhao, X., Wang, M., Wong, A., Fang, C., Zhang, X., Huang, H., Lopez, J. V., Kilfoyle, K., Zhang, Y., Ortí, G., Venkatesh, B., & Shi, Q. (2016). Fish-T1K (Transcriptomes of 1,000 Fishes) Project: Large-scale transcriptome data for fish evolution studies. *GigaScience*, 5, 18. <https://doi.org/10.1186/s13742-016-0124-7>
- Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A., & Van de Peer, Y. (2004). Major events in the genome evolution of vertebrates: Paraneome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 1638–1643. <https://doi.org/10.1073/pnas.0307968100>

- Wainwright, P. C., Smith, W. L., Price, S. A., Tang, K. L., Sparks, J. S., Ferry, L. A., Kuhn, K. L., Eytan, R. I., & Near, T. J. (2012). The evolution of pharyngognath: A phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. *Systematic Biology*, 61, 1001–1027. <https://doi.org/10.1093/sysbio/sys060>
- Wheeler, T. J., & Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29, 2487–2489. <https://doi.org/10.1093/bioinformatics/btt403>
- Yin, G., Pan, Y., Sarker, A., Baki, M. A., Kim, J. K., Wu, H., & Li, C. (2019). Molecular systematics of Pampus (Perciformes: Stromateidae) based on thousands of nuclear loci using target-gene enrichment. *Molecular Phylogenetics and Evolution*, 140, 106595. <https://doi.org/10.1016/j.ympev.2019.106595>
- Yuan, H., Atta, C., Tornabene, L., & Li, C. (2019). Assexon: Assembling exon using gene capture data. *Evolutionary Bioinformatics*, 15, 117693431987479. <https://doi.org/10.1177/1176934319874792>

- Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18, 821–829. <https://doi.org/10.1101/gr.074492.107>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Hughes LC, Ortí G, Saad H, et al. Exon probe sets and bioinformatics pipelines for all levels of fish phylogenomics. *Mol Ecol Resour.* 2021;21:816–833. <https://doi.org/10.1111/1755-0998.13287>