

ORIGINAL RESEARCH

Gene markers for exon capture and phylogenomics in ray-finned fishes

Jiamei Jiang  | Hao Yuan | Xin Zheng | Qian Wang | Ting Kuang | Jingyan Li | Junning Liu | Shuli Song | Weicai Wang | Fangyuan Cheng | Hongjie Li | Junman Huang | Chenhong Li

Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources (Shanghai Ocean University), Ministry of Education, Shanghai, National Demonstration Center for Experimental Fisheries Science Education (Shanghai Ocean University), Shanghai, China

Correspondence

Chenhong Li, Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution; Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources (Shanghai Ocean University), Ministry of Education, Shanghai; National Demonstration Center for Experimental Fisheries Science Education (Shanghai Ocean University), Shanghai, China. Email: chli@shou.edu.cn

Funding information

The Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning; the Innovation Program of Shanghai Municipal Education Commission

Abstract

Gene capture coupled with the next-generation sequencing has become one of the preferred methods of subsampling genomes for phylogenomic studies. Many exon markers have been developed in plants, sharks, frogs, reptiles, fishes, and others, but no universal exon markers have been tested in ray-finned fishes. Here, we identified a suite of “single-copy” protein-coding sequence (CDS) markers through comparing eight fish genomes, and tested them empirically in 83 species (33 families and nine orders or higher clades: Acipenseriformes, Lepisosteiformes, Elopomorpha, Osteoglossomorpha, Clupeiformes, Cypriniformes, Gobiaria, Carangaria, and Eupercaria; sensu Betancur et al. 2013). Sorting the markers according to their completeness and phylogenetic decisiveness in taxa tested resulted in a selection of 4,434 markers, which were proven to be useful in reconstructing phylogenies of the ray-finned fishes at different taxonomic levels. We also proposed a strategy of refining baits (probes) design a posteriori based on empirical data. The markers that we have developed may greatly enrich the batteries of exon markers for phylogenomic study in ray-finned fishes.

KEYWORDS

Actinopterygii, bait design, nuclear gene markers, phylogenomics, population genomics, target enrichment

1 | INTRODUCTION

Next-generation sequencing (NGS) drastically reduced the cost of sequencing a genome, so that reconstructing phylogenetic relationships using whole genomes became feasible (Jarvis et al., 2014). However, sequencing whole genomes is still costly and sometimes unnecessary. Subsampling genome sequences has gained popularity in phylogenetics and population genomics in recent years (Emerson et al., 2010; Faircloth et al., 2012; Lemmon, Emme, & Lemmon, 2012;

Li, Hofreiter, Straube, Corrigan, & Naylor, 2013; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). There are mainly two different genome subsampling tools. One is associated with restriction site-related markers, such as restriction site-associated DNA (RAD; Baird et al., 2008) and double digest RADseq (ddRAD) markers (Peterson et al., 2012), which could be used to produce sequences from a tremendous number of anonymous loci that are particularly useful in studying population genomics or species-level phylogeny (Davey & Blaxter, 2010). The other method is gene capture, also known as

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2019 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd.

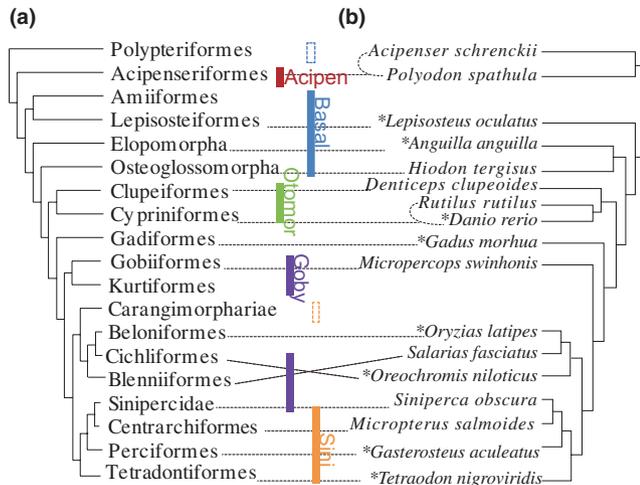


FIGURE 1 (a) Phylogenetic relationships among 21 groups of ray-finned fish (Betancur et al., 2013; Hughes et al., 2018). The vertical bars indicate different projects carried in the author's laboratory. The unfilled vertical bars indicate groups that captured <3,000 loci. (b) Maximum likelihood tree of 17 representative ray-finned fishes based on 4,434 exon loci, all nodes have a 100 bootstrap value. The connected dotted lines between two trees indicate the corresponding taxa. Eight species names marked with stars indicate the fishes used in finding the target markers

target enrichment to capture and sequence target loci, which often result in less missing data than the restriction site-related methods do (Collins & Hrbek, 2015), and the target loci can be applied across highly divergent taxonomic groups (Faircloth et al., 2012; Lemmon et al., 2012; Li et al., 2013). Benefitting from the advantages of two methods, hybrid approaches (Ali et al., 2016; Hoffberg et al., 2016) have also been developed resulting in less missing data and higher coverage compared with traditional RADseq approaches.

Gene capture is based on hybridizing RNA/DNA baits (probes) to DNA libraries of targeted species and enriching sequences similar to the baits for subsequent high-throughput sequencing. Two popular methods, Ultraconserved Element (UCE) captures (Faircloth et al., 2012) and Anchored Hybrid Enrichment (AHE; Lemmon et al., 2012), were developed to retrieve single-copy highly conserved elements in the genome along with variable flanking regions. A third method, exon capture was designed explicitly to capture single-copy coding sequences across moderate to highly divergent species (Bi et al., 2012; Hedtke, Morgan, Cannatella, & Hillis, 2013; Li et al., 2013). Exons have been more commonly used for phylogenetics than anonymous noncoding regions, and evolution of protein-coding sequences has been well studied. Furthermore, lowered stringency in hybridization and washing steps enables baits to hybridize with more distant sequences, so it solves the problem that divergent baits and targeted exons may produce missing data (Cosart et al., 2011 and Mason, Li, Helgen, & Murphy, 2011).

Exon capture markers have been developed in plants (Chamala et al., 2015; Mandel et al., 2014; Weitemier et al., 2014), invertebrates (Hugall, O'Hara, Hunjan, Nilsen, & Moussalli, 2016; Mayer et al., 2016; Teasdale, Kohler, Murray, O'Hara, & Moussalli, 2016; Yuan et

al., 2016), and many vertebrate groups, including sharks and skates (Li et al., 2013), frogs (Hedtke et al., 2013; Portik, Smith, & Bi, 2016), skink lizards (Bragg, Potter, Bi, & Moritz, 2016), and others. As the most diverse group of vertebrates with more than 30,000 described species (Nelson, Grande, & Wilson, 2016), many studies applied target enrichment to investigate the phylogenetic relationships of ray-finned fishes (Actinopterygii), but most of them focused on using UCE markers (Chakrabarty et al., 2017; Faircloth, Sorenson, Santini, & Alfaro, 2013; Gilbert et al., 2015; Harrington et al., 2016; Hulseley, Zheng, Faircloth, Meyer, & Alfaro, 2017; Longo et al., 2017; McGee et al., 2016). As a complementary approach, many exon markers have been reported previously for some ray-finned fishes. Ilves and Lopez-Fernandez (2014) developed 923 exon markers for cichlids based on genome sequence of *Oreochromis niloticus*. Arcila et al. (2017) tested 1,051 exon markers on the Otophysii. We also developed 17,817 single-copy nuclear coding sequence (CDS) markers and applied those in the siniperid fish in a previous study (Song, Zhao, & Li, 2017). However, those markers have not been tested on other groups and may not work well across all ray-finned fishes. Hughes et al. (2018) selected 1,721 exon markers >200 bp from the 17,817 markers and retrieved their sequences from hundreds of transcriptomic and genomic datasets in silico, although they did not verify their utility in wet laboratory experiments.

Selecting target markers and designing baits that are effective across a wide range of species is the first major challenge when applying the gene capture method. Many considerations are taken into baits design, such as uniqueness and conservativeness of markers, length and complexity of markers, and genetic distance between baits and target sequences (Bi et al., 2012; Campana, 2017; Faircloth, 2017; Faircloth et al., 2012; Gilbert et al., 2015; Hugall et al., 2016; Lemmon et al., 2012; Li et al., 2013; Mayer et al., 2016). However, all these measures are usually taken a priori, and few studies have been done to refine baits design after gene capture to improve the baits set for future experiments (but see Branstetter, Longino, Ward, & Faircloth, 2017).

In this study, we tested the 17,817 CDS previously identified as a part of a separate study (Song et al., 2017) and screen these markers to identify the best ones for inferring phylogeny across all major clades of ray-finned fish. We chose phylogenetically decisive markers based on the results of pilot experiments and refined the bait design to improve evenness of reads coverage across all loci. Finally, we tested phylogenetic usefulness of selected markers in ray-finned fishes at both order level and species level. Our goal is to provide a set of common exon markers for gene capture and phylogenomic studies in the ray-finned fishes.

2 | MATERIALS AND METHODS

2.1 | Identification of original marker sets and collecting preliminary data for candidate markers

The markers were identified through comparing eight fish genomes (Figure 1a) using a bioinformatics tool, EvoMarkers (Li, Riethoven, &

Naylor, 2012; Supporting information Appendix S1: Figure S1). Two sets of single-copy markers were generated and used to design baits to capture species in five different research projects conducted in the authors' laboratory, including works on early-branching actinopterygians lineages (Basal), acipenseriforms (Acipen), otomorphs (Otomor), gobioids (Goby), and siniperids (Sini) (Supporting information Appendix S1: Figure S2). One set of markers was designed based on *Oreochromis niloticus* including 17,817 loci (used in "Goby" and "Sini" projects). The other one was identified from *Lepisosteus oculatus* comprising 13,843 loci (used in "Basal," "Acipen" and "Otomor" projects).

Thousands of the candidate CDS markers were tested empirically as pilot experiments in 83 actinopterygian species (99 individuals, 33 families of nine orders or higher clades), covering major clades of ray-finned fishes (Supporting information Table S1).

According to suggestion of the manufacturer, biotinylated RNA baits (MYcroarray, Ann Arbor, Michigan) were synthesized with 2× tiling. Because thymine and adenine have fewer hydrogen bonds with its complementary nucleotide compared with cytosine and guanine, the 3' end of the baits was padded with "Ts" if baits were shorter than 120 bp. For the baits designed on *O. niloticus*, loci longer than 100 bp were targeted. For the baits designed on *L. oculatus*, loci longer than 120 bp were targeted.

Total genomic DNA was extracted from fin or muscle tissue of samples using a Tissue DNA kit (Omega Bio-tek, Norcross, GA, USA) and quantified using a NanoDrop 3300 Fluorospectrometer (Thermo Fisher Scientific, Wilmington, DE, USA). Samples of 350–500 ng genomic DNA were sheared to ~250 bp using a Covaris E220 Focused-ultrasonicator (Covaris, Woburn, USA). Subsequently, sheared DNA was used to construct libraries. Blunt-end repair, adapter ligation, fill-in, prehybridization PCR, and double exon enrichment steps mainly followed the protocol of cross-species gene capture (Li et al., 2013). The enriched libraries were amplified in 25 µl PCR reactions with a forward primer that included 8 bp custom designed indices, a reverse primer, and KAPA HiFi taq ready mix (Kapa Biosystems, Wilmington, MA, USA). Custom indices with at least two nucleotide differences among the indices were designed following Meyer and Kircher (2010). The concentration of products was measured using a NanoDrop 3300 Fluorospectrometer. The products were pooled in equimolar concentrations and sequenced on an Illumina HiSeq 2500 platform (Illumina, Inc, San Diego, CA, USA) with other samples from the same or other projects at Annoroad (Beijing, China).

Read assembling followed the pipeline of Yuan et al. (2016) except that contigs and respective homologous bait sequences were translated into amino acid sequences before comparison. The raw reads were parsed to respective files for each species according to the 8 bp indices on the P7 adaptor using BclToFastq (Illumina, Inc). The remaining adaptor sequence on the 3 primer end and low quality bases were trimmed from raw reads using Trim_galore v0.4.1 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with default parameters. Then, PCR duplicates were filtered and parsed to homologous bait sequences. Reads were separately assembled into contigs using Trinity v2.0.6 (Grabherr et al., 2011) with

default parameters. Overlapped contigs were further assembled using Geneious v7.1.5 (Kearse et al., 2012). Each contig was translated into amino acid sequences and compared with the amino acid sequences of the original baits using the Smith–Waterman algorithm (Smith & Waterman, 1981). The most similar match to each bait was selected as putative target sequence. The section of the target sequence covering the bait sequence in the alignment was identified as the exon, and the remaining was considered flanking sequence. To identify potential paralogs in retrieved sequences, we used BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) to align them against the genomes of *O. niloticus* or *L. oculatus*. Sequences with the best BLAST hit not in target region of the genomes were recognized as potential paralogs and excluded from further analysis. All steps were automated using custom Perl scripts except the further assembly of overlapped contigs in Geneious. The final output includes two fasta files: coding sequences with and without flanks.

2.2 | Selecting the best markers and refining the baits design based on gene capture results

Based on results of the pilot experiments, exon markers resulting in less missing data were selected, and the baits were evaluated and redesigned with the regions with extraordinarily high read depth were masked (Figure 2). The assembled sequences from different projects were merged (*merge.pl*). Briefly, taxa that had more than 3,000 genes captured were kept (*select.pl*). Subsequently, a Perl script *dec.pl* was used to pick phylogenetically decisive loci. Phylogenetic decisiveness means that the datasets should contain all taxa whose relationships are addressed (Dell'Ampio et al., 2014). In our case, the decisive taxonomic groups included eight major clades of the ray-finned fishes: Acipenseriformes, Lepisosteiformes, Elopomorpha, Osteoglossomorpha, Otomorpha, Gobiaria, Ovalentaria, and Eupercaria. The Polypteridae was excluded in bait design, because both species of the polypterids sampled had fewer than 3,000 targets captured.

From our pilot experiments, we found that partial regions of some target loci had extraordinarily high number of reads mapped, which consumed a large proportion of the total data collected. Those regions escaped RepeatMasker (Smit, Hubley, & Green, 1996–2004) checking in original baits design and wasted a lot of sequencing reads, so we excluded those regions to refine the design of baits. To find those problematic regions, the selected decisive data were parsed to different files by species name (*parsefast.pl*). Then, the raw reads of each species were mapped to the assembled reference sequences of each species using BWA (Li & Durbin, 2009). The read depth data were extracted from the mapping results using SAMtools (Li & Durbin, 2009) and a custom Perl script (*mapdepth.pl*). Regions with extraordinary high read depth, that is, 100 times greater than adjacent regions were identified and labeled with lowercase letters (*pickbaits.pl*). Loci with these regions were discarded if their length were shorter than 120 bp excluding the masked regions. Longer loci were separated into multiple regions for bait design with the masked regions excluded. To test the utility of refined markers, baits were redesigned based on the result of the pilot experiments and

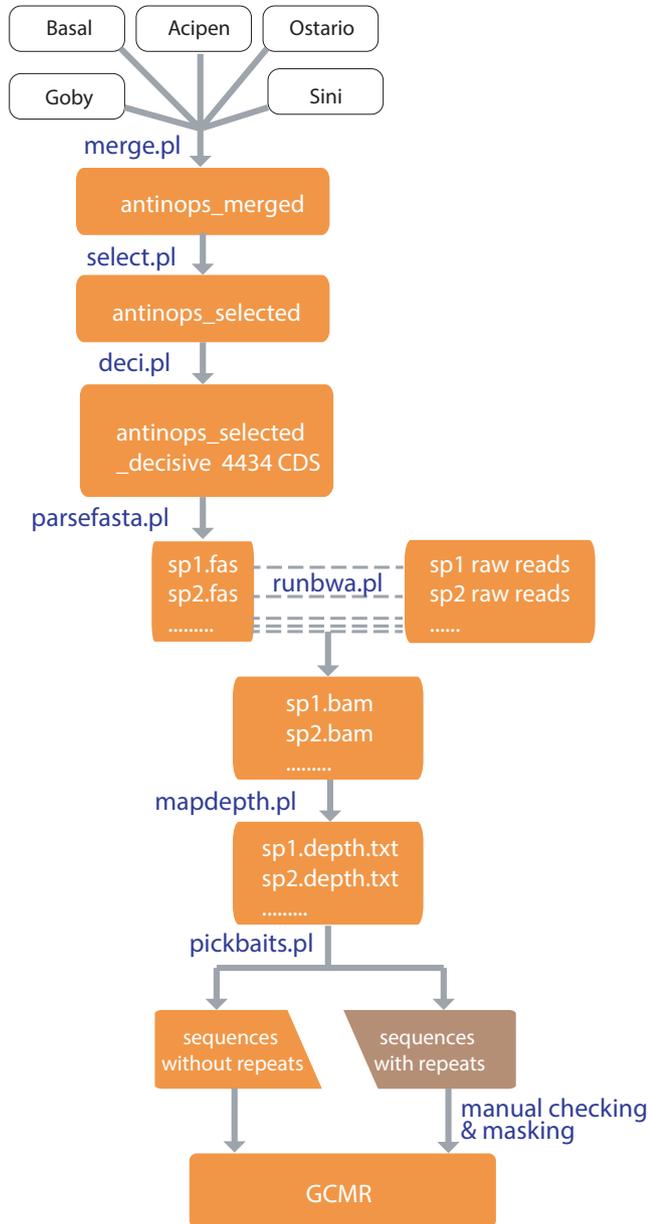


FIGURE 2 Pipeline of screening for markers with less missing data and better phylogenetic decisiveness and posterior baits refining. I. Merge data from different project (merge.pl); II. select loci with less missing data and high phylogenetic decisiveness (gcmr_select.pl; gcmr_deci.pl); III. find and mask region with extraordinary read depth for bait redesign (parsefasta.pl; runbwa.pl; mapdepth.pl; gcmr_pickbaits.pl). The posterior baits refining steps are optional when empirical data from pilot gene capture are available. GCMR stands for gene capture marker refinement

used to enrich and assemble sequences of *Rhinogobius giurinus* for a new round following the aforementioned pipeline.

2.3 | Testing phylogenetic usefulness of the markers selected and efficacy of the new baits

A phylogeny of 17 species of ray-finned fishes, including nine species with gene captured data and eight species with sequence

data extracted from genomes, was reconstructed. Each individual locus was aligned using Mafft v7 (Katoh & Standley, 2013) with default parameter settings (mafft_AA.pl). The aligned AA sequences were translated back to DNA sequences via a custom Perl script aa2dna_align.pl. Statistics were summarized from 4,434 aligned loci of nine captured samples and eight species with available genomes. Sequence statistics including average length of coding regions, average GC content, and average pairwise distance (p-dist) were calculated using a custom Perl script (statistics.pl) and R package “ape” (Paradis, Claude, & Strimmer, 2004). Consistency index (CI) and retention index (RI) were calculated using PAUP* v4.0a (Swofford 2003). Due to the high variability of flanking regions, only the coding regions without flanks were used for phylogenetic inference. All aligned loci were concatenated using a custom Perl script (concatnexus.pl). Then, concatenated maximum likelihood (ML) trees were constructed using the ML method implemented in ExaML v3 (Kozlov, Aberer, & Stamatakis, 2015). Concatenated alignments were partitioned by codon and then used to reconstruct the tree under the GTRGAMMA model with 100 bootstrap replicates to assess nodal support.

To test the utility of selected markers for studies at the species level, we reconstructed a species tree of four species of freshwater sleepers (*Odontobutis*, Gobiiformes), whose relationships are unresolved (Ren & Zhang, 2007; Zhong et al., 2017). The species tree was reconstructed based on exon capture data of the chosen markers, including five individuals of each species of *Odontobutis sinensis*, *O. potamophila*, and *O. yaluensis* and one individual of *O. haifengensis*. Two individuals of *Perccottus glenii* were used as the outgroup. ASTRAL v4.11.1 (Mirarab & Warnow, 2015) was used to infer the species tree. An informative unrooted tree cannot be inferred from loci with less than 4 taxa, so these loci were excluded from analyses. Remaining gene trees of each locus were reconstructed using RAxML HPC-PTHREAD (Stamatakis, 2006) under GTRGAMMA model. Then, they were summarized into species tree using ASTRAL with default parameter settings. Multi-locus bootstrapping would result in high bootstrap supports even with high discordance among gene trees if there is a sufficient number of genes (Sayyari & Mirarab, 2016), so bootstrap supports were not accessed and branch supports were measured as quartet support instead. Quartet support is the frequency of quartets in gene trees supporting the topology of the species tree and is accessed by implementing option “-t 1” in ASTRAL. A concatenated ML tree was constructed as well. The coding region of each locus was aligned using Mafft v7.294b (Katoh & Standley, 2013) with default parameter setting. Then, aligned loci were concatenated to reconstruct ML trees using RAxML HPC-PTHREAD (Stamatakis, 2006) under GTRGAMMA model. Nodal support was accessed with 100 bootstrap replicates.

Principal component analysis (PCA) was carried out to visualize inter- and intraspecific genetic variation among individuals of the four *Odontobutis* species. As input for the PCA, single nucleotide polymorphisms (SNPs) were extracted from coding regions of the *Odontobutis* data. Loci having data in more than two species were processed with a SNP calling procedure. Reference sequences of

filtered loci were generated from aligned sequences based on majority consensus rule using a custom Perl script (*consensus.pl*). Trimmed reads were mapped to the reference using BWA v0.7.15-r1140 (Li & Durbin, 2009). Picard MarkDuplicates (<http://broadinstitute.github.io/picard/>) was used to mark duplicates. Then GATK Best Practices of germline short variant discovery recommendations (Van der Auwera et al., 2013; DePristo et al., 2011; McKenna et al., 2010) were followed to do local realignment, base quality score recalibration, SNPs discovery, and genotyping across all samples in concert using standard hard filtering parameters by GATK-3.2.2 (McKenna et al., 2010). Indels were discarded, and only one of the best SNPs of each locus was selected for downstream analyses to fulfill the assumption of linkage disequilibrium. The vcf file was converted to genotype data file format for PCA by a custom script *vcftosnps.pl*. PCA was performed with R package *ade4* (Dray & Dufour, 2007) to unravel variability among 16 *Odontobutis* samples by the *dudi.pca* function.

The new baits refined based on empirical data were compared with the baits designed a priori. The raw reads of each species were mapped to the respective assemblies using BWA (Li & Durbin, 2009). The read depth data were extracted from the mapping results using *SAMtools* (Li & Durbin, 2009) and a custom Perl script (*mapdepth.pl*). Reads coverage of each locus was calculated by dividing total length of captured reads by length of the locus. The evenness of read coverage was summarized from a custom designed parameter RC50. Loci were sorted by their read coverage in descending order, and then, the number of loci used half of the total reads is the RC50. Higher RC50 reflects better evenness of read coverage. Read coverage was calculated using a custom Perl script (*coverage.pl*), and RC50 was estimated using *Excel*. The comparison of a priori and a posteriori bait designing was done on capture results for a goby species (*R. giurinus*). Finally, to help researchers to design baits using reference species that are closer to their organism of interest than the eight model fishes that we used, we developed a pipeline of retrieving sequences of the target loci from user-provided genomes (Supporting information Appendix S3).

2.4 | Investigate the variability of flanking regions

Since the variability of flanking sequences among different families was too high, we only investigate the variability of flanking regions of the 16 individuals of *Odontobutis*. Sequences with long insertions or deletions, unalignable sequences, and very short flanking sequences (<20 bp) were filtered using a custom Perl script (*flank_filter.pl*, see detailed parameters in Appendix S3). A custom Perl script (*flank_pdis.pl*) was used to summarize length of flanking regions and p-dist between all pairs of flanking sequences for both filtered and unfiltered flanking regions. SNPs were extracted from filtered coding and flanking regions using GATK following the aforementioned procedure. A number of SNPs in coding and flanking regions were counted with a custom Perl script (*snps_num.pl*). All custom Perl scripts can be found online in Supporting information Appendix S3.

3 | RESULTS

3.1 | Single-copy protein-coding markers for ray-finned fishes

The number of loci captured in the pilot experiments ranged from 435 to 11,534 in different samples. All but four samples had more than 3,000 loci captured (Supporting information Appendix S1: Figure S2). The samples that did the worst in gene capture experiment included two polypteriforms (*Erpetoichthys calabaricus* and *Polypterus endlicheri*), one sturgeon (*Acipenser ruthenus*), and the Waigeo barramundi (*Psammoperca waigiensis*). After combining the data from all five projects, excluding taxa with fewer than 3,000 loci captured and selecting for phylogenetic decisive loci, we obtained 4,434 CDS markers of 2,261 genes. The information of the target loci and sequences of the eight model fish species can be found online in Supporting information Appendix S2.

3.2 | Phylogenetic usefulness of selected markers

The average length of the coding region of the chosen markers was 236 bp (94–4,718 bp). GC content ranged from 37% to 69% with an average of 55%. Average p-dist among the 17 species varied from 0.06 to 0.50 substitutions per site, with an overall average of 0.19. Average consistency index (CI) was 0.60 (0.43–0.93), and average retention index (RI) was 0.52 (0.47–0.62) (Supporting information Appendix S1: Figure S3). Maximum likelihood (ML) analyses concatenating 4,434 loci resulted in a well-resolved tree of major ray-finned fish clades, and all nodes had 100 bootstrap support values (Figure 1). The resulting phylogenetic tree is consistent with recent studies (Betancur et al., 2013; Faircloth et al., 2013; Hughes et al., 2018), except that the Elopomorpha and the Osteoglossomorpha were found sister to each other and Beloniformes were found more closely related to Blenniiformes than to Cichliformes. Because our data only involved a handful of taxa, those inconsistent results should be investigated with better taxon sampling with our exon markers.

There were 4,296 of 4,434 loci captured at least in one *Odontobutis* sample. A total of 1,630 loci were captured in all samples. The average length of target regions was 265 bp (120–5,637 bp). A concatenated ML tree was reconstructed for the four Chinese *Odontobutis* species with *P. glenii* as outgroups, which was well resolved with 100 bootstrap support values for each node. *Odontobutis haifengensis* was sister to the rest of the *Odontobutis* species. *O. yaluensis* was grouped with *O. potamophila*, and *O. sinensis* was placed as sister to them. Individuals of the same species were clustered together. A species tree was also reconstructed with four *Odontobutis* species and *P. glenii*, with a normalized quartet score of 0.64. For the topology of species tree, *O. yaluensis* was also grouped with *O. potamophila*, but the placement of *O. haifengensis* and *O. sinensis* was different (Supporting information Appendix S1: Figure S4). We extracted 36,440 single nucleotide polymorphisms (SNPs) sites from coding regions (35 SNPs per kb) of the 16 *Odontobutis* samples, and only one of

the best SNPs from each locus was used for PCA. The PCA showed clear genetic differentiation on interspecific level. Individuals of *O. sinensis* were well separated from other species. Individuals of *O. yaluensis* and *O. potamophila* partially overlapped with each other with respect to PC1 (Supporting information Appendix S1: Figure S5).

3.3 | Gene capture marker refinement

We examined the results of gene capture experiments using original baits. We found that 26 loci of *R. giurinus* had extremely high number of reads mapped. We manually checked those loci and found that all regions with high reads depth had low complexity. We redesigned the baits and carried a new round of gene capture experiment. The gene capture results from new baits had higher RC50 which reflected higher evenness of reads coverage among different loci than the results from the original baits. The reads depths of most of loci using refined baits were higher than the ones using original baits (Figure 3).

3.4 | Variability of the flanking regions of *Odontobutis*

Length of flanking regions ranged from 0 to 2,271 bp and centered around 800 bp (Figure 4a). P-dist among them ranged from 0 to 0.84. After filtering unalignable and uninformative short flank regions, p-dist varied from 0 to 0.57 (Figure 4b). The number of SNPs in the flanking regions was 73,097 (50 per kb), more abundant than in coding regions (36,440 SNPs, 35 per kb).

4 | DISCUSSIONS

4.1 | Exon capture

Protein-coding sequences are more commonly used than noncoding flank regions in phylogenetic analysis. Models of molecular evolution of coding sequences are well studied. Up to 20 popular exon

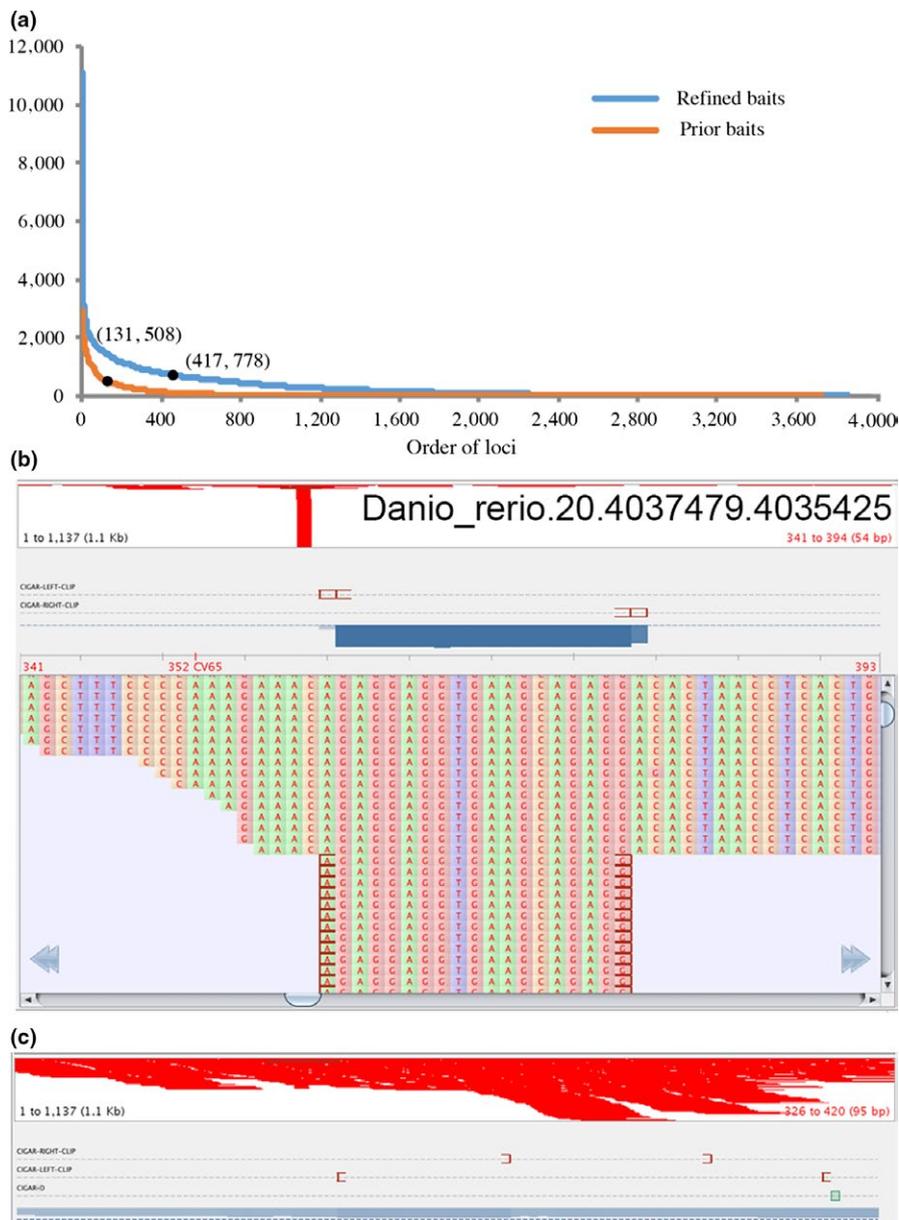


FIGURE 3 Comparison on evenness of read coverage between results of gene capture using the baits designed a priori (a, blue curve) and the baits refined posteriorly (a, orange curve). (b, c) are screenshots from visualizing the read depth of the locus *Danio_riero.20.4037479.4035425* using Tablet v1.16.09.06. In this example, the result using baits designed a priori (b) is much worse than the result using refined baits (c)

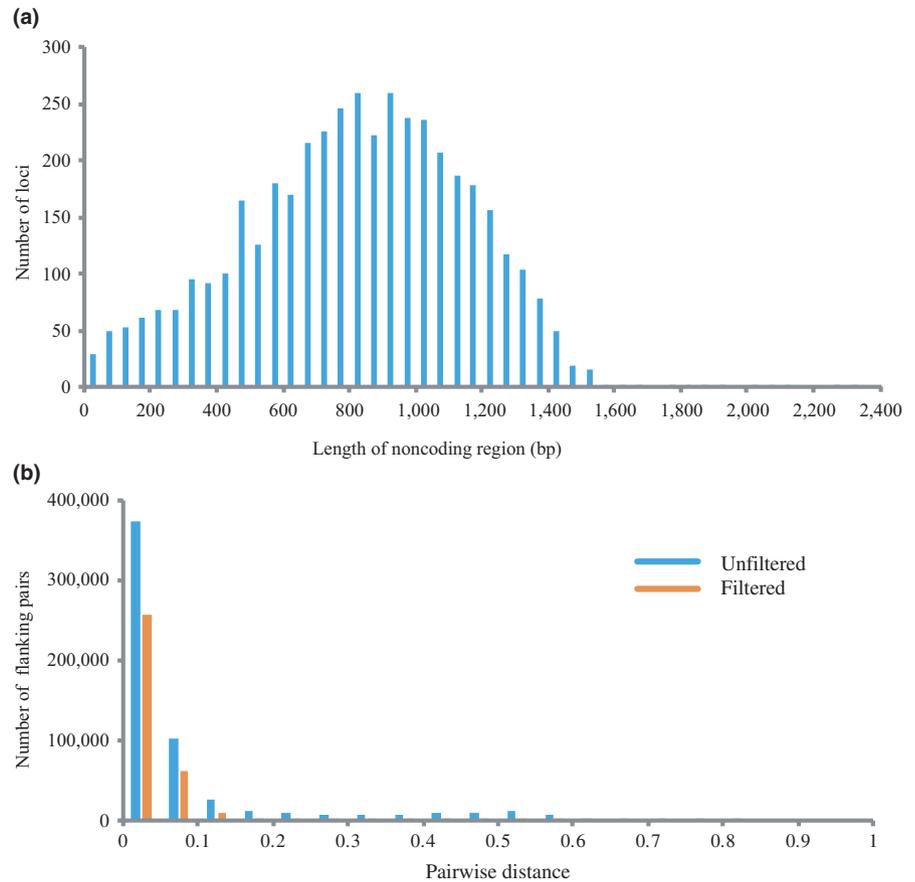


FIGURE 4 Length distribution of unfiltered flanking sequences of 4,296 loci, from 16 *Odontobutis* individuals (a). Pairwise distance distribution of all pairs of unfiltered (blue bar) and filtered (orange bar) flanking sequences (b)

markers, so called “legacy markers” have been used in landmark molecular phylogenetic studies to resolve the tree of life of fishes, long before the target-capture methods were developed (Betancur et al., 2013; Broughton, Betancur, Li, Arratia, & Orti, 2013; Near et al., 2013, 2012). Our experiments showed that the markers selected and the baits designed were effective in studying phylogenetic relationship of major groups of the ray-finned fishes, and closely related species as well. The numerous numbers of markers developed here may provide more power to solve the remaining difficult questions in tree of ray-finned fishes.

4.2 | A posteriori marker design

The simple repeats in the markers were detected and masked using RepeatMasker by the manufacturer, MYcroarray (Ann Arbor, Michigan) before synthesizing the baits. However, repeats with some variations or complex repeats could not be detected with RepeatMasker, which resulted in a high read depth in some regions (Figure 3b). Extremely high read depth suggested that repetitive regions were enriched to a high degree, which could cause problems in subsequent read assembly, and waste sequencing resources. Based on the sequencing results, we masked these unusual regions in subsequent baits refinement, which produced more even depth for the targeted loci (Figure 3b). If a pilot study is planned before a large-scale experiment, we recommend applying our method to refine baits design to improve the efficacy of the baits.

4.3 | Orthology checking and data filtering

Problem of mistakenly using paralogous genes for phylogenetic reconstruction is exacerbated with phylogenomic data, and currently, there is no ideal method to validate orthology of loci assembled from NGS data (Chakrabarty et al., 2017; McCormack, Hird, Zellmer, Carstens, & Brumfield, 2013). The targeted loci we selected for are “single-copy” (Li et al., 2012), which may have less chance to be paralogous than members of gene families, (Li, Orti, Zhang, & Lu, 2007). In addition, we performed a “re-blast” step in data processing pipeline to identify and exclude potential paralogs (Yuan et al., 2016). Nonetheless, both methods cannot guarantee orthology of targeted sequences due to the third round of whole-genome duplication event in teleosts and slow and steady loss of some paired genes over the subsequent 250 My (Inoue, Sato, Sinclair, Tsukamoto, & Nishida, 2015). Tree-based methods such as filtering the loci a posteriori based on known monophyly of taxa could be used to alleviate the problem of paralogy.

4.4 | Phylogenetic utilities of selected markers at species level

The species tree and the concatenated tree reconstructed from 16 *Odontobutis* with two *P. glenii* samples as outgroups showed that the placement of *O. yaluensis* and *O. potamophila* in the two trees was the same, while the position of *O. haifengensis* and *O.*

sinensis was conflicting. We found that quartet supports of 3 possible quadri partitions of the clade of *O. yaluensis* and *O. potamophila*, *O. haifengensis*, *O. sinensis*, and *P. glenii* were 0.39 for topology represented in Supporting information Appendix S1: Figure S5 and 0.31, 0.30 for other two topologies. Close quartet supports for 3 topologies indicated severe incomplete lineage sorting among selected loci, which resulted in the incongruent placement of *O. haifengensis* and *O. sinensis* in species tree and concatenated tree. Nonetheless, concatenated tree still had high bootstrap supports for each node, which indicated high bootstrap value may not reliably reflect accuracy of resulting tree. This finding was also reported in several previous studies (Belfiore, Liu, & Moritz, 2008; Kubatko & Degnan, 2007; Weisrock et al., 2012). For coalescence-based methods, accuracy can be measured based on concordance between resulting tree and gene trees (Larget, Kotha, Dewey, & Ane, 2010; Sayyari & Mirarab, 2016). So, we recommend a coalescence-based method to reconstruct species trees and measure accuracy of it with congruence between species trees and the given gene trees. Overall, our results of *Odontobutis* species using the 4,434 loci suggested that those markers can be applied in species-level applications.

4.5 | Variability of the flanking regions in *Odontobutis*

Although we targeted coding regions, flanking sequences were also captured by hitchhiking. Length of flanking sequences was highly correlated with the size of sheared genomic DNA. We could break DNA into longer pieces during library construction if longer flanks were preferred, but inserts >1 kb may sabotage the performance of Illumina sequencing. Some of flanks were nonfunctional sequences and may be less constrained by purifying selection. After filtering unalignable and uninformative short flanking sequences, the amount of remaining data was dramatically decreased, but there were still more SNPs in flanking regions than in coding regions, suggesting that flanking regions may be useful in phylogenetic studies at the species level.

5 | CONCLUSION

In this work, we developed 4,434 empirically tested protein-coding markers that are useful in reconstructing phylogenies of the ray-finned fishes at different taxonomic levels. We also provided researchers with resources for applying those markers in their group of interest: (a) the target sequences of the 4,434 loci for all eight model species which users can use to design baits; (b) a user-friendly pipeline for users to retrieve target sequences of the 4,434 loci from species of their interest if they provide new genome sequences or transcriptomes; and (c) a pipeline for users to refine their baits design a posteriori based on empirical data. These tools could advance phylogenomic studies in ray-finned fishes, the most diverse vertebrate group.

ACKNOWLEDGEMENTS

This work was supported by the Innovation Program of Shanghai Municipal Education Commission, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning and the Shanghai Collaborative Innovation for Aquatic Animal Genetics and Breeding. We thank Shanghai Oceanus Supercomputing Center (SOSC) for providing computational resource and Dr Luke Tornabene from the University of Washington for editing the manuscript.

CONFLICT OF INTEREST

None declared.

AUTHORS CONTRIBUTION

Jiamei Jiang, Hao Yuan, Xin Zheng, Qian Wang, Ting Kuang, Jingyan Li, Junning Liu, Shuli Song, Weicai Wang, Fangyuan Cheng, Hongjie Li, Junman Huang, and Chenhong Li J Jiang analyzed data and wrote the manuscript with support from H Yuan and C Li; X Zheng carried out the experiment and assembled the database with Q Wang, T Kuang, J Li, J Liu, S Song, W Wang, F Cheng, and H Li; J Huang helped in data analyses and wrote scripts for mining target gene sequences from published genomes; and C Li conceived the original idea and supervised all projects.

DATA ACCESSIBILITY

Information of 4,434 target loci and respective sequences for all eight model fishes can be found in Appendix S2. The pipeline and scripts for reads assembly, ray-finned fishes baits design, and refinement are in Appendix S3. The fastq files of raw reads have been deposited in NCBI Sequence Read Archive (SRA) with accession number SRP162615. Accession number of samples referenced from pilot studies can be found in Supporting information Table S2. Sequence alignments in nexus format and input files for analysis were lodged in Dryad with <https://doi.org/10.5061/dryad.41j28n0>.

ORCID

Jiamei Jiang  <https://orcid.org/0000-0002-6969-1917>

REFERENCES

- Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD capture (rapture): flexible and efficient sequence-based genotyping. *Genetics*, 202(2), 389–400. <https://doi.org/10.1534/genetics.115.183665>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2)
- Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L. J., Ko, K. D., ... Betancur-R., R. (2017). Genome-wide interrogation advances

- resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, 1(2), 10. <https://doi.org/10.1038/s41559-016-0020>
- Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 3(10), e3376. <https://doi.org/10.1371/journal.pone.0003376>
- Belfiore, N. M., Liu, L., & Moritz, C. (2008). Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). *Systematic Biology*, 57(2), 294–310. <https://doi.org/10.1080/10635150802044011>
- Betancur-R., R., Broughton, R. E., Wiley, E. O., Carpenter, K., López, J. A., Li, C., ... Orti, G. (2013). The tree of life and a new classification of bony fishes. *PLoS Currents*, 5, <https://doi.org/10.1371/currents.tol.53ba26640df0c8ae75bb165c8c26288>
- Bi, K., Vanderpool, D., Singhal, S., Linderth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, 13, 403. <https://doi.org/10.1186/1471-2164-13-403>
- Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources*, 16(5), 1059–1068. <https://doi.org/10.1111/1755-0998.12449>
- Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017). Enriching the ant tree of life: Enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, 8(6), 768–776. <https://doi.org/10.1111/2041-210x.12742>
- Broughton, R. E., Betancur, R. R., Li, C., Arratia, G., & Orti, G. (2013). Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. *PLoS Curr*, 5, <https://doi.org/10.1371/currents.tol.2ca8041495ffafd0c92756e75247483e>
- Campana, M. G. (2017). BaitTools: Software for hybridization capture bait design. *Molecular Ecology Resources*, <https://doi.org/10.1111/1755-0998.12721>
- Chakrabarty, P., Faircloth, B. C., Alda, F., Ludt, W. B., McMahan, C. D., Near, T. J., ... Alfaro, M. E. (2017). Phylogenomic systematics of Ostariophysan fishes: Ultraconserved elements support the surprising non-monophyly of characiformes. *Systematic Biology*, 66(6), 881–895. <https://doi.org/10.1093/sysbio/syx1038>. doi:10.1093/sysbio/syx038
- Chamala, S., Garcia, N., Godden, G. T., Krishnakumar, V., Jordon-Thaden, I. E., De Smet, R., ... Soltis, P. S. (2015). Markerminer 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences*, 3(4), 1400115.
- Collins, R. A., & Hrbek, T. (2015). An in silico comparison of reduced-representation and sequence-capture protocols for phylogenomics. *bioRxiv preprint first posted online Nov. 21, 2015*. <https://doi.org/10.1101/032565>
- Cosart, T., Beja-Pereira, A., Chen, S. Y., Ng, S. B., Shendure, J., & Luikart, G. (2011). Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, 12, 8. <https://doi.org/10.1186/1471-2164-12-347>
- Davey, J. W., & Blaxter, M. L. (2010). RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, 9(5–6), 416–423. <https://doi.org/10.1093/bfpg/elq031>
- Dell'Ampio, E., Meusemann, K., Szucsich, N. U., Peters, R. S., Meyer, B., Borner, J., ... Misof, B. (2014). Decisive data sets in phylogenomics: Lessons from studies on the phylogenetic relationships of primarily wingless insects. *Molecular Biology and Evolution*, 31(1), 239–249. <https://doi.org/10.1093/molbev/mst196>
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–498. <https://doi.org/10.1038/ng.806>
- Dray, S., & Dufour, A.-B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4), 1–20.
- Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E., & Holzapfel, C. M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 107(37), 16196–16200. <https://doi.org/10.1073/pnas.1006538107>
- Faircloth, B. C. (2017). Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods in Ecology and Evolution*, 8(9), 1103–1112. <https://doi.org/10.1111/2041-210x.12754>
- 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/syb004>
- 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(6), 7. <https://doi.org/10.1371/journal.pone.0065923>
- Gilbert, P. S., Chang, J., Pan, C., Sobel, E. M., Sinshimer, J. S., Faircloth, B. C., & Alfaro, M. E. (2015). Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. *Molecular Phylogenetics and Evolution*, 92, 140–146. <https://doi.org/10.1016/j.ympev.2015.05.027>
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 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, 14. <https://doi.org/10.1186/s12862-016-0786-x>
- Hedtke, S. M., Morgan, M. J., Cannatella, D. C., & Hillis, D. M. (2013). Targeted enrichment: Maximizing orthologous gene comparisons across deep evolutionary time. *PLoS ONE*, 8(7), e67908. <https://doi.org/10.1371/journal.pone.0067908>
- Hoffberg, S. L., Kieran, T. J., Catchen, J. M., Devault, A., Faircloth, B. C., Mauricio, R., & Glenn, T. C. (2016). RADcap: Sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources*, 16(5), 1264–1278. <https://doi.org/10.1111/1755-0998.12566>
- Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R., & Moussalli, A. (2016). An exon-capture system for the entire class Ophiuroidea. *Molecular Biology and Evolution*, 33(1), 281–294. <https://doi.org/10.1093/molbev/msv216>
- Hughes, L. C., Orti, G., Huang, Y. u., Sun, Y., Baldwin, C. C., Thompson, A. W., ... 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(24), 6249–6254. <https://doi.org/10.1073/pnas.1719358115>
- Hulse, C. D., Zheng, J., Faircloth, B. C., Meyer, A., & Alfaro, M. E. (2017). Phylogenomic analysis of Lake Malawi cichlid fishes: Further evidence that the three-stage model of diversification does not fit. *Molecular Phylogenetics and Evolution*, 114, 40–48. <https://doi.org/10.1016/j.ympev.2017.05.027>
- Ilves, K. L., & Lopez-Fernandez, H. (2014). A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. *Molecular Ecology Resources*, 14(4), 802–811. <https://doi.org/10.1111/1755-0998.12222>
- 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(48), 14918–14923. <https://doi.org/10.1073/pnas.1507669112>
- Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., ... Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215), 1320–1331. <https://doi.org/10.1126/science.1253451>
- Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>
- Kozlov, A. M., Aberer, A. J., & Stamatakis, A. (2015). ExaML version 3: A tool for phylogenomic analyses on supercomputers. *Bioinformatics*, 31(15), 2577–2579. <https://doi.org/10.1093/bioinformatics/btv184>
- Kubatko, L. S., & Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1), 17–24. <https://doi.org/10.1080/10635150601146041>
- Larget, B. R., Kotha, S. K., Dewey, C. N., & Ane, C. (2010). BUCKY: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26(22), 2910–2911. <https://doi.org/10.1093/bioinformatics/btq539>
- Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, 61(5), 727–744. <https://doi.org/10.1093/sysbio/sys049>
- Li, C., Hofreiter, M., Straube, N., Corrigan, S., & Naylor, G. J. (2013). Capturing protein-coding genes across highly divergent species. *BioTechniques*, 54(6), 321–326. <https://doi.org/10.2144/000114039>
- 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(1), 44. <https://doi.org/10.1186/1471-2148-7-44>
- Li, C., Riethoven, J. J., & Naylor, G. J. (2012). EvolMarkers: A database for mining exon and intron markers for evolution, ecology and conservation studies. *Molecular Ecology Resources*, 12(5), 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(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- 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 (Syngnathiformes) using ultraconserved elements. *Molecular Phylogenetics and Evolution*, 113, 33–48. <https://doi.org/10.1016/j.ympev.2017.05.002>
- Mandel, J. R., Dikow, R. B., Funk, V. A., Masalia, R. R., Staton, S. E., Kozik, A., ... Burke, J. M. (2014). A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the compositae. *Applications in Plant Sciences*, 2(2), 1300085. <https://doi.org/10.3732/apps.1300085>
- Mason, V. C., Li, G., Helgen, K. M., & Murphy, W. J. (2011). Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Research*, 21(10), 1695–1704. <https://doi.org/10.1101/gr.120196.111>
- Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., ... Niehuis, O. (2016). BaitFisher: A software package for multispecies target DNA enrichment probe design. *Molecular Biology and Evolution*, 33(7), 1875–1886. <https://doi.org/10.1093/molbev/msw056>
- McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, 66(2), 526–538. <https://doi.org/10.1016/j.ympev.2011.12.007>
- McGee, M. D., Faircloth, B. C., Borstein, S. R., Zheng, J., Hulseley, C. D., Wainwright, P. C., & Alfaro, M. E. (2016). Replicated divergence in cichlid radiations mirrors a major vertebrate innovation. *Proceedings of the Royal Society B-Biological Sciences*, 283(1822), 6. <https://doi.org/10.1098/rspb.2015.1413>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., ... DePristo, M. a. (2010). The Genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- Meyer, M., & Kircher, M. (2010). Illumina Sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010(6), pdb.prot5448. <https://doi.org/10.1101/pdb.prot5448>
- Mirarab, S., & Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12), i44–52. <https://doi.org/10.1093/bioinformatics/btv234>
- Near, T. j., Dornburg, A., Eytan, R. i., Keck, B. p., Smith, W. i., Kuhn, K. i., ... 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(31), 12738–12743. <https://doi.org/10.1073/pnas.1304661110>
- Near, T. j., Eytan, R. i., Dornburg, A., Kuhn, K. i., Moore, J. a., Davis, M. p., ... Smith, W. i. (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(34), 13698–13703. <https://doi.org/10.1073/pnas.1206625109>
- Nelson, J. S., Grande, T. C., & Wilson, M. V. H. (2016). *Fishes of the World* (5th edn). New York, NY: Wiley.
- Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2), 289–290. <https://doi.org/10.1093/bioinformatics/btg412>
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, 7(5), e37135. <https://doi.org/10.1371/journal.pone.0037135>
- Portik, D. M., Smith, L. L., & Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Molecular Ecology Resources*, 16(5), 1069–1083. <https://doi.org/10.1111/1755-0998.12541>
- Ren, G., & Zhang, Q. (2007). Molecular phylogeny of the genus *Odontobutis* based upon partial sequences of mitochondrial 12S rRNA genes. *Acta Hydrobiologica Sinica*, 31(04), 473–478.
- Sayyari, E., & Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution*, 33(7), 1654–1668. <https://doi.org/10.1093/molbev/msw079>
- Smit, A., Hubley, R., & Green, P. (1996–2004). RepeatMasker Open-3.0. Retrieved from <http://www.repeatmasker.org/>
- Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195–197. [https://doi.org/10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5)
- 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>
- Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed

- models. *Bioinformatics*, 22(21), 2688–2690. <https://doi.org/10.1093/bioinformatics/btl446>
- Swofford, D. L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts: Sinauer Associates.
- Teasdale, L. C., Kohler, F., Murray, K. D., O'Hara, T., & Moussalli, A. (2016). Identification and qualification of 500 nuclear, single-copy, orthologous genes for the Eupulmonata (Gastropoda) using transcriptome sequencing and exon capture. *Molecular Ecology Resources*, 16(5), 1107–1123. <https://doi.org/10.1111/1755-0998.12552>
- Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., & del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 11.10.1–11.10.33.
- Weisrock, D. W., Smith, S. D., Chan, L. M., Biebouw, K., Kappeler, P. M., & Yoder, A. D. (2012). Concatenation and concordance in the reconstruction of mouse lemur phylogeny: An empirical demonstration of the effect of allele sampling in phylogenetics. *Molecular Biology and Evolution*, 29(6), 1615–1630. <https://doi.org/10.1093/molbev/mss008>
- Weitemier, K., Straub, S. C., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., & Liston, A. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, 2(9), 1400042. <https://doi.org/10.3732/apps.1400042>
- Yuan, H., Jiang, J., Jimenez, F. A., Hoberg, E. P., Cook, J. A., Galbreath, K. E., & Li, C. (2016). Target gene enrichment in the cyclophyllidean cestodes, the most diverse group of tapeworms. *Molecular Ecology Resources*, 16(5), 1095–1106. <https://doi.org/10.1111/1755-0998.12532>
- Zhong, L., Wang, M., Li, D., Tang, S., Zhang, T., Bian, W., & Chen, X. (2017). Complete mitochondrial genome of *Odontobutis haifengensis* (Perciformes, Odontobutiae): A unique rearrangement of tRNAs and additional non-coding regions identified in the genus *Odontobutis*. *Genomics*, 110(6), 382–388. <https://doi.org/10.1016/j.ygeno.2017.12.008>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Jiang J, Yuan H, Zheng X, et al. Gene markers for exon capture and phylogenomics in ray-finned fishes. *Ecol Evol*. 2019;9:3973–3983. <https://doi.org/10.1002/ece3.5026>