


RESEARCH ARTICLE

Open Access

Genome of the four-finger threadfin *Eleutheronema tetradactylum* (Perciformes: Polynemidae)



Zhe Qu^{1*†}, Wenyan Nong^{1†}, Yifei Yu^{1†}, Tobias Baril^{2†}, Ho Yin Yip¹, Alexander Hayward^{2*} and Jerome H. L. Hui^{1*} 

Abstract

Background: Teleost fish play important roles in aquatic ecosystems and aquaculture. Threadfins (Perciformes: Polynemidae) show a range of interesting biology, and are of considerable importance for both wild fisheries and aquaculture. Additionally, the four-finger threadfin *Eleutheronema tetradactylum* is of conservation relevance since its populations are considered to be in rapid decline and it is classified as endangered. However, no genomic resources are currently available for the threadfin family Polynemidae.

Results: We sequenced and assembled the first threadfin fish genome, the four-finger threadfin *E. tetradactylum*. We provide a genome assembly for *E. tetradactylum* with high contiguity (scaffold N50 = 56.3 kb) and high BUSCO completeness at 96.5%. The assembled genome size of *E. tetradactylum* is just 610.5 Mb, making it the second smallest perciform genome assembled to date. Just 9.07–10.91% of the genome sequence was found to consist of repetitive elements (standard RepeatMasker analysis vs custom analysis), making this the lowest repeat content identified to date for any perciform fish. A total of 37,683 protein-coding genes were annotated, and we include analyses of developmental transcription factors, including the *Hox*, *ParaHox*, and *Sox* families. MicroRNA genes were also annotated and compared with other chordate lineages, elucidating the gains and losses of chordate microRNAs.

Conclusions: The four-finger threadfin *E. tetradactylum* genome presented here represents the first available genome sequence for the ecologically, biologically, and commercially important clade of threadfin fish. Our findings provide a useful genomic resource for future research into the interesting biology and evolution of this valuable group of food fish.

Keywords: *Eleutheronema tetradactylum*, Fish, Genome, Transcription factor, MicroRNA, Repetitive element

* Correspondence: quzhe@cuhk.edu.hk; alex.hayward@exeter.ac.uk; jeromehui@cuhk.edu.hk

[†]Zhe Qu, Wenyan Nong, Yifei Yu and Tobias Baril co-first authors

¹School of Life Sciences, Simon F.S. Li Marine Science Laboratory, State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong, China

²Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Penryn, Cornwall, Exeter TR10 9FE, UK



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Teleostei is the most species-rich and diverse group of vertebrates, with ~ 30,000 species, accounting for around half of all extant vertebrate species [1]. In addition to their great diversity, teleosts play important ecological roles in aquatic ecosystems, and are of great relevance as a source of protein in both wild fisheries and aquaculture. Consequently, the study of teleost genomes is of considerable importance from evolutionary, ecological and applied perspectives.

The teleost order Perciformes comprises more than 10,000 species and represents the single largest group of vertebrates, with representatives in almost every aquatic ecosystem on earth [2]. Many perciform fish are highly important commercially, and the group is fast becoming an important model system for vertebrate genomics, with the number of genome sequences available growing rapidly.

The fourfinger threadfin, *Eleutheronema tetradactylum*, belongs to the perciform family Polynemidae, commonly known as threadfins. A distinguishing feature of threadfins is the pectoral fin, which is composed of two distinct sections, an upper normally shaped fin, and a lower section consisting of up to seven long, threadlike independent rays, which are believed to act as sensory probes for locating food in muddy habitats [3]. Like other threadfins, *E. tetradactylum* has the ability to tolerate a wide range of salinities, and is often found in estuaries and rivers, as well as its main coastal marine habitat over shallow sand or mud flats [4, 5]. Threadfins are important for commercial and sport fisheries across a wide socioeconomic spectrum, and are marketed as fresh, frozen, dried or salted fish. *E. tetradactylum* is a particularly valued species, not least because of its large size compared to other threadfins (~ 2 m maximum length). However, *E. tetradactylum* is believed to be declining rapidly across much of its tropical Indo-West Pacific range, and is classified as endangered by the IUCN [6]. Like many fish, *E. tetradactylum* is a protandrous hermaphrodite that can undergo sex change at different ages [4, 7].

To date, no genomic resources are available for the threadfin family Polynemidae. To address this and provide the first reference genome for Polynemidae, we sequenced and assembled a draft genome for the fourfinger threadfin, *Eleutheronema tetradactylum*. Here we describe the *E. tetradactylum* genome, to provide resources that can facilitate a better understanding of this poorly studied fish lineage. We compare findings to those for other perciform fish, and perform a range of specific analyses on transposable element content, and key developmental loci, including Hox genes and microRNAs.

Results

Genome assembly metrics

The genomic DNA from a single individual of *E. tetradactylum* was isolated and sequenced using the 10X Genomics platform. The final assembly size was 610.5 Mb, which is close to the estimated genome size (~ 630 Mb, Supplementary Figure S1). The scaffold N50 length was 56.3 kb (Table 1), and among the 36,746 scaffolds, the longest scaffold was 568.4 kb. To estimate genome completeness, we performed a BUSCO assessment [8]. We found that 96.5% of BUSCO genes were detected (88.2% of which were complete, and 8.3% of which were fragmented) (Table 1, Supplementary Table S1).

A total of 38,490 gene models were predicted for *E. tetradactylum*, including 37,683 protein-coding genes and 807 tRNAs. Orthologous genes of the fourfinger threadfin were compared to the genomes of goldfish, salmon, zebrafish and human. A total of 8777 gene groups are shared in these five vertebrates (Fig. 1a), while 12,071 gene groups are conserved among the four fish species. In addition, we carried out an orthologous gene comparison between the fourfinger threadfin and all other available percoid fish genomes (Fig. 1b). The number of protein coding genes of these 12 percoid fish species ranges considerably, from 20,541 to 37,683 (Supplementary Table S2).

Repeat content

To assess the repeat content of the fourfinger threadfin genome, we generated a de novo repeat library using RepeatModeler. Following this, we applied two approaches for repeat annotation: (i) a standard RepeatMasker analysis, and (ii) a custom implementation including repeat defragmentation and removal of overlapping annotations. The standard RepeatMasker analysis identified a very low repeat content of just 9.07% (Table 2), while the custom implementation identified a slightly higher, but still very low repeat content of

Table 1 Summary of genome assembly metrics for *E. tetradactylum*

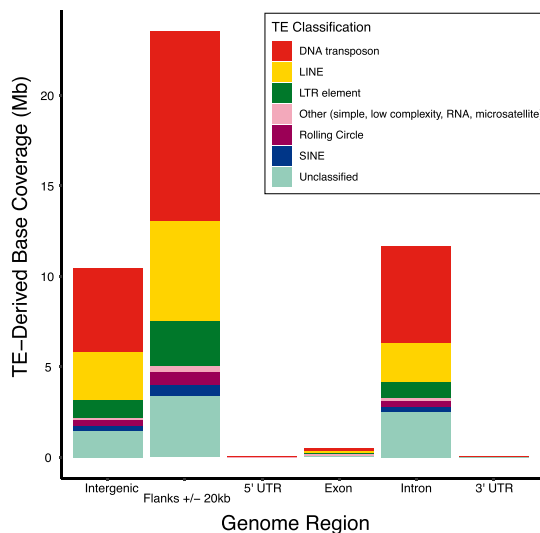
Common name Species	Four-finger threadfin <i>Eleutheronema tetradactylum</i>
Accession number	WFKG00000000
Number of scaffolds	36,746
Assembly size	610,497,648
Scaffold N50	56,314
Largest scaffold	568,375
Number of genes (protein-coding gene)	38,490 (37,683)
Gap content	14,703
BUSCOs (Complete)	96.5% (88.2%)

Repeat analysis of the four-finger threadfin

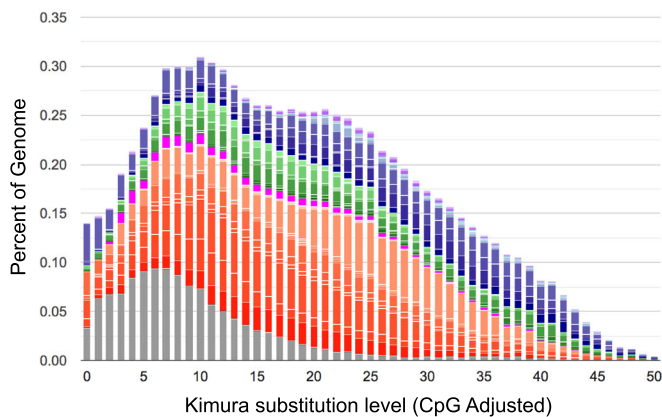
Adult four-finger threadfin
(*Eleutheronema tetradactylum*)



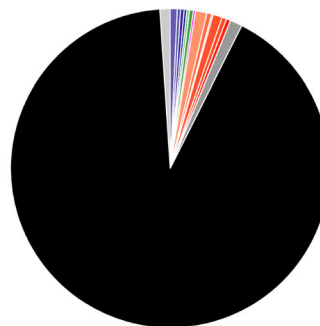
Repeat Localities



Repeat Landscape



Repeat Content



- | | | | |
|----------------|----------------|--------------|---------------|
| SINE/IRNA-CR1 | LINE/R2 | LTR | DNA/Merlin |
| SINE/IRNA-7SL | LINE/Dong-R4 | LTR/ERVL | DNA |
| SINE/U | LINE/Jockey-1 | LTR/Gypsy | DNA/Maverick |
| SINE/MIR | LINE/Proto2 | LTR/Copia | DNA/Kolobok |
| SINE/IRNA-V | LINE/L2 | LTR/Pao | DNA/hAT |
| SINE/Deu | LINE/Rex-Babar | LTR/Ngaro | DNA/Harbinger |
| SINE/IRNA-RTE | LINE/CR1 | LTR/DIRS | DNA/Crypton |
| SINE/IRNA-Alu | LINE/RTE | RC/Helitron | DNA/CMC |
| SINE/IRNA | LINE | DNA/Dada | Other |
| SINE/Alu | LINE/L1 | DNA/Zator | Unknown |
| SINE/7SL | LTR/ERVK | DNA/TcMar | |
| SINE/5S | LTR/ERV | DNA/PiggyBac | |
| SINE | LTR/ERV1 | DNA/P | |
| Retroposon/SVA | | DNA/MULE | |

Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Comparison of orthologous gene groups between *E. tetradactylum* and other vertebrates. **a** Shared and unique orthologous gene groups in four species of Teleostei and human. **b** Genome-wide comparison of orthologous genes among various percid fish. Abbreviations: Cau: goldfish *Carassius auratus*; Dre: zebrafish *Danio rerio*; Ete/Etet: four-finger threadfin *Eleutheronema tetradactylum*; Hsa: human *Homo sapiens*; Ssa: Atlantic salmon *Salmo salar*; Carg: *Channa argus*; Ecra: *Etheostoma cragini*; Espe: *Etheostoma spectabile*; Lcro: *Larimichthys crocea*; Lmac: *Lateolabrax maculate*; Nalb: *Nibeia albiflora*; Pcha: *Parachaenichthys charcoti*; Pfla: *Perca flavescens*; Pflu: *Perca fluviatilis*; Sluc: *Sander lucioperca*; Ssin: *Sillago sinica*

10.91%, with the slight increase a consequence of merging fragmented repeats into longer repeat models during this approach (Table 2).

In the *E. tetradactylum* genome, repeats are much more prevalent in gene flanks (regions within 20 kb upstream and 20 kb downstream of annotated host genes), compared to intergenic regions or introns (Fig. 2: Repeat Localities). This potentially suggests recruitment of repeats for host associated purposes through donation of coding or regulatory sequence (e.g. [9, 10]), although very little evidence exists of repeats directly inserted into gene regulatory 5' and 3' UTRs (Fig. 2: Repeat locality). Alternatively, this pattern suggests the involvement of other at present unclear genomic processes that have resulted in an uneven distribution of repeats.

Transposable elements (TEs) account for the vast majority of repeats annotated in the four-finger threadfin genome (Table 2). Other categories of repeat (simple, small RNA, satellite, and low complexity) account for just 0.32–1.01% of the genome (Table 2, Fig. 2: Repeat content). Among TEs, the largest contribution of sequence comes from DNA transposons (4.27–5.31%), followed by LINEs (1.88–1.99%), and LTR elements (1.01–1.33%) (Table 2). These figures are broadly similar to those reported for the yellow drum *Nibeia albiflora* [11], another perciform fish with a similarly low TE content.

Examination of the repeat landscape generated for the four-finger threadfin implies that there has been a steady

decrease in transposon activity over recent time periods, since there is a notable decrease in repeats separated by low levels of divergence, which correspond to more recent copies (Fig. 2: Repeat landscape plot). This pattern is primarily a consequence of a reduction in the relative activity of DNA transposons and LTR elements, since levels of divergence for LINEs appears to have remained relatively stable.

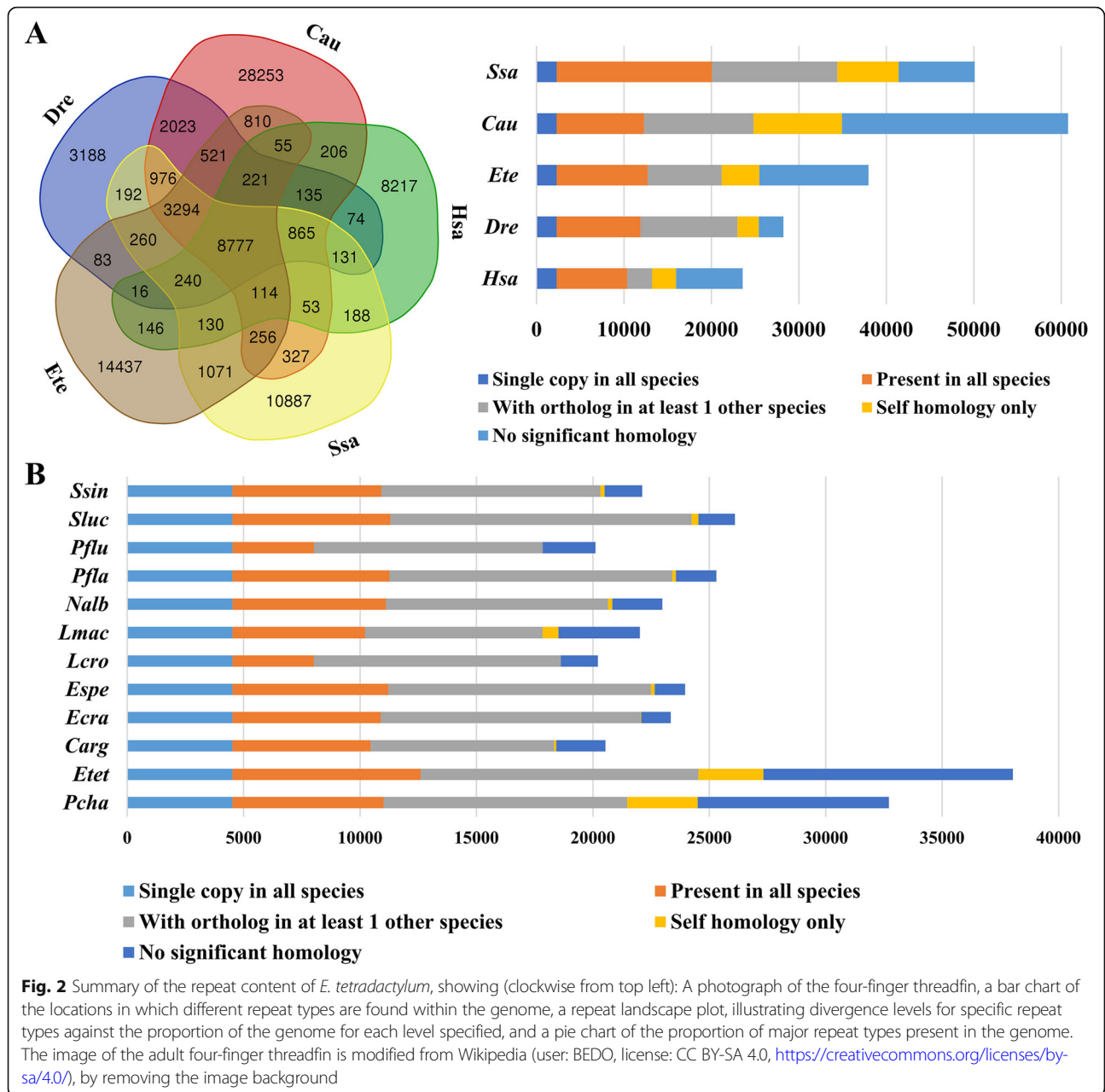
Hox, ParaHox, and sox genes of *E. tetradactylum*

Hox cluster genes encode a group of transcription factors that control the anteroposterior axis during development [12, 13]. A total of 59 *Hox* genes were recovered in the *E. tetradactylum* genome (Fig. 3a-b, Supplementary data S1, Figure S2). *HoxA*, *B*, *C*, *D* clusters with conserved microRNAs (*mir-10* and *mir-196*) were revealed (Fig. 3b). Retention of *HoxA7*, extra copies of *HoxB* genes, as well as the loss of *HoxD13a* were observed in *E. tetradactylum* (Fig. 3b). These data suggest dynamic *Hox* gene gains and losses have occurred during fish evolution.

As for the *Hox* evolutionary sister group, the *ParaHox* genes including 2 *Gsx*, 1 *Pdx*, and 3 *Cdx* genes could be identified on 6 different scaffolds (Supplementary data S1, Figure S3). This situation mirrors what has been found in other teleosts, which have broken *ParaHox* clusters with secondary gene losses having occurred after whole genome duplication [14, 15].

Table 2 Table summarising repeat content in the *E. tetradactylum* genome, detailing the number of elements, overall length, and genomic proportion, for each major repeat type, for a standard RepeatMasker analysis (left) and a custom annotation approach (right, see methods for details)

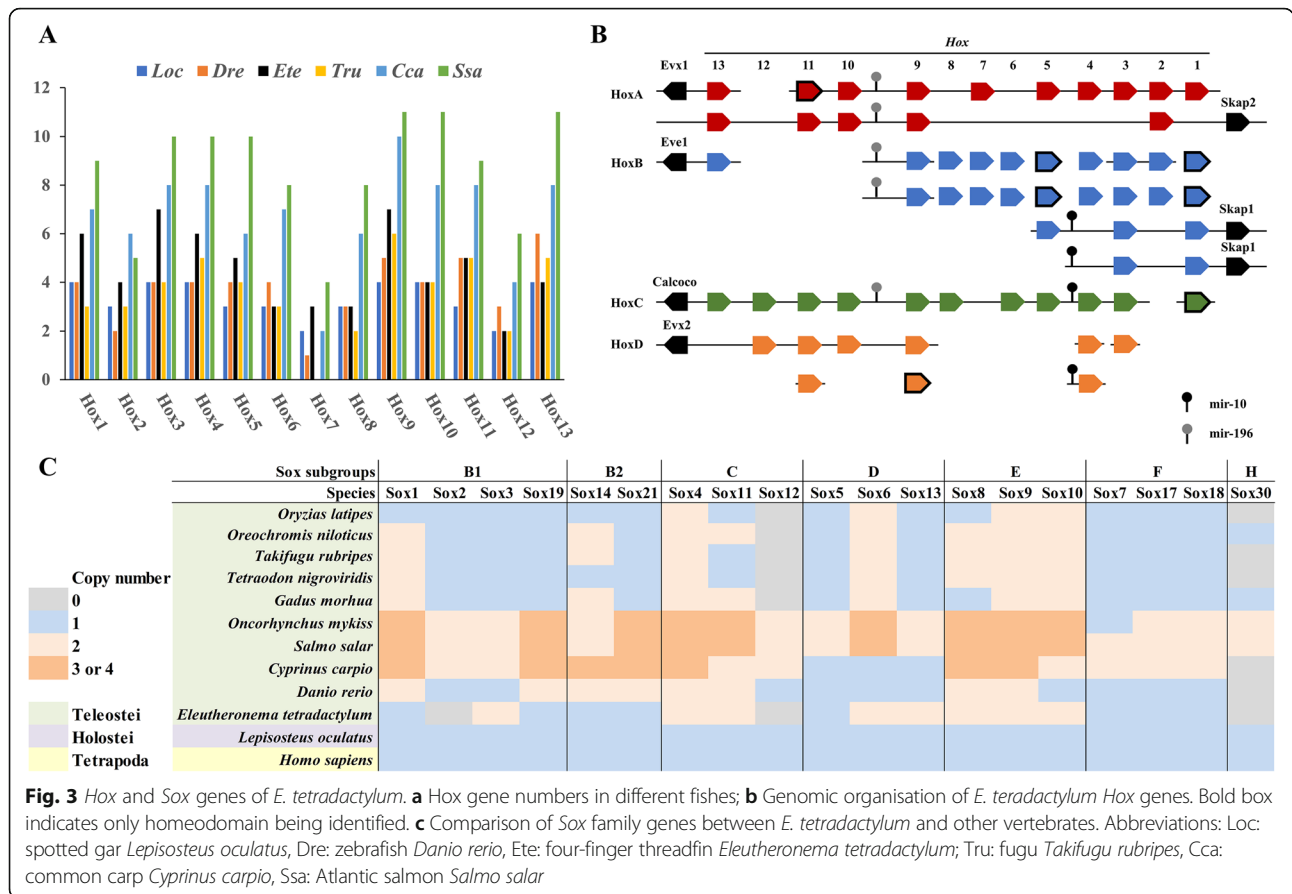
Repeat Class	Traditional Repeat Annotation (RepeatMasker)			Refined Conservative Repeat Annotation		
	No. elements	Total Length (Mb)	Percentage sequence (%)	No. elements	Total Length (Mb)	Percentage sequence (%)
Retroelement	155419	19.14	3.14	150453	21.92	3.59
<i>SINE</i>	13534	1.46	0.24	12936	1.63	0.27
<i>LINE</i>	60949	11.49	1.88	60294	12.18	1.99
<i>LTR element</i>	80936	6.19	1.01	77223	8.11	1.33
DNA transposon	298755	26.04	4.27	276340	32.42	5.31
Rolling-circle	15702	1.28	0.21	15557	2.05	0.34
Unclassified	32821	7.88	1.29	29890	7.80	1.28
Other	11396	1.00	0.16	19703	2.40	0.39
Total repeats	514093	55.34	9.07	491943	66.59	10.91



The expansion of key transcription factors is proposed to act as a critical genetic driver for the evolution of vertebrate innovations [16]. The *Sox* gene family encodes transcription factor members that contain the high mobility group box (HMG box) DNA binding domain, which is conserved throughout the metazoans, and play vital roles in various developmental processes [17, 18]. A total of 26 *Sox* genes from 6 subgroups were identified in *E. tetradactylum* genome, but not *Sox30*, which is similar to findings for the zebrafish and pufferfish (Fig. 3c, Supplementary Figure S4).

MicroRNAs

MicroRNAs are an important group of post-transcriptional regulators with a key role in development. A total of 356 microRNA genes, including 126 bilaterian-conserved microRNA species, were identified in *E. tetradactylum* (Supplementary data S2). Comparing *E. tetradactylum* microRNA content with that of other chordates available in miRBase [19], MirGeneDB [20] and other relevant genomes [21], 36 chordate-conserved microRNAs were identified (Fig. 4a). In particular, 33 and 19 microRNAs appear to have emerged in the vertebrate ancestor and gnathostome ancestor respectively (Fig. 4a, Supplementary data S2).



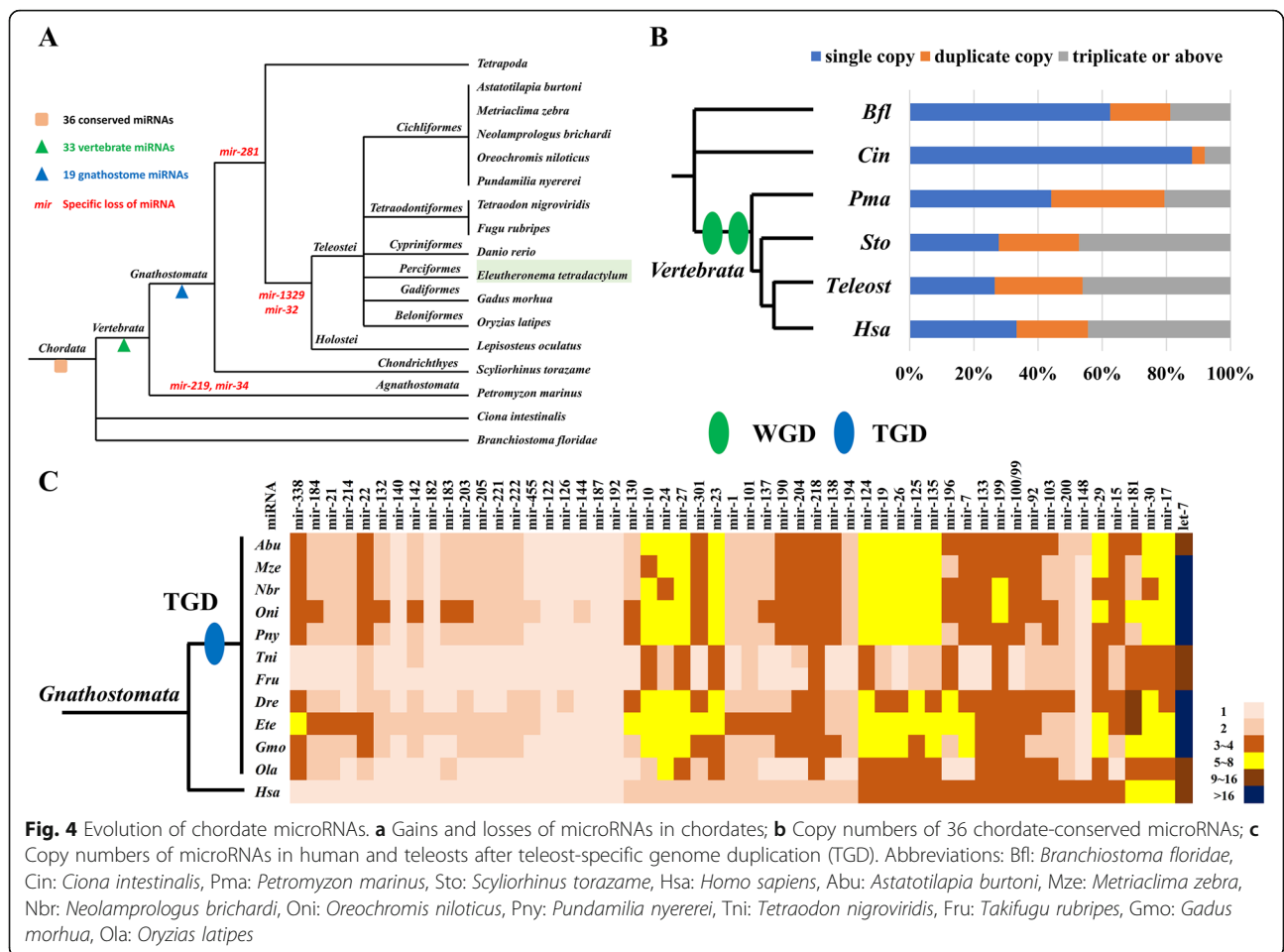
The last common ancestor of vertebrates underwent two rounds of whole genome duplication (WGD), with an additional round of WGD that occurred in the ancestor of teleosts (i.e. the teleost specific genome duplication or ‘TGD’) [22–29]. In the salmonid lineage and the lineage containing the last common ancestor of carp and goldfish, an additional fourth WGD (4R) occurred [30, 31]. Consequently, copy numbers of the 36 chordate-conserved microRNAs discussed above were investigated, to investigate patterns in the retention of duplicates (Fig. 4b). About 60% of microRNAs have multiple copies in different vertebrates (post-WGD). While in more distant vertebrate relatives the pattern appears to differ, for example in the sea squirt and amphioxus, 88 and 62.5% of their microRNAs exist as single copies respectively (Fig. 4b). This is likely to be influence of WGDs in vertebrate lineages. Intriguingly, comparison of microRNA copies between human and teleosts (which were subject to the TGD), also revealed that more than 50% of microRNAs have more paralogues in teleost species than in human (Fig. 4c). For example, 36 out of 54 (67%) conserved microRNAs possess additional duplicates in *E. tetradactylum* compared to human (Fig. 4c). Our findings suggest that WGD events have a profound influence on the evolution of microRNA complements,

and thus that microRNA landscapes may be a useful indicator of WGD events.

Discussion

Considerable variation exists among the genome sizes of perciform fish, with the genome size of the threadfin lying very much at the lower end. At the upper end of the spectrum for perciform fish, the pikeperch (Percidae: *Sander lucioperca*) has an estimated genome size of 1014 Mb, and an assembly genome size of ~900 Mb [32], and the red sea bream (Sparidae: *Pagrus major*) has an estimated genome size of ~806 Mb, and an assembly genome size of 829.3 Mb. Meanwhile, at the lower end of the spectrum, the yellow drum (Sciaenidae: *Nibea albiflora*) has an estimated genome of 573–581 Mb and an assembly genome size of 596 Mb [11], and the Chinese sillago (Sillaginidae: *Sillago sinica*) has an estimated genome size of ~524 Mb, and an assembly genome size of just 534 Mb [33]. Comparison of various percid fish genomes showed that *E. tetradactylum* possesses the largest number of protein-coding genes, but the smallest average protein size, which may contribute to its small overall genome size, despite its large number of genes.

Nguinkal et al. [32] examined repeat content among the genomes of nine perciform fish species, and found it



to vary from 13.8% in the yellow drum (*Nibea albiflora*) to 39.8% in the pikeperch (*Sander lucioperca*). Thus, among perciform fish, the four-finger threadfin has an especially low repeat content. One of the key determinants of genome size in eukaryotes is repeat content [34], and in line with this, a strong relationship between genome size and repeat content is reported for perciform fish [32]. Consequently, the repeat content of the four-finger threadfin conforms to its relatively small genome size of ~ 610 Mb. However, the genome assembly of the yellow drum is even smaller (~ 565 Mb) than the four-finger threadfin genome presented here, but its repeat content is higher at 13.8% [11]. Thus, it appears that the repeat content of the four-finger threadfin is especially low, both among fish genomes [35], and compared to other vertebrate genomes more generally. Reasons for the low repeat content of the threadfin genome are unclear at present, and further elucidation of this finding represents an interesting avenue for future study. Further, it remains unclear what factors have led to certain expansions of repeats, particularly DNA transposable elements, in other perciform fish such as the pikeperch [32].

Hox cluster genes encode a group of transcription factors that control the anteroposterior axis during development [12, 13]. The *E. tetradactylum* contains *HoxA7*, while cyprinid fish (such as zebrafish, common carp and goldfish) and pufferfish have lost *HoxA7* [36–38]. Moreover, additional copies of *HoxB* genes, as well as the loss of *HoxD13a*, were observed in *E. tetradactylum*. These data suggest dynamic *Hox* gene gains and losses have occurred during fish evolution, and have undoubtedly helped to shape the wide diversity of body forms observable among teleost fish, and not least structures such as the charismatic threadfins displayed by fish in the family Polynemidae.

Different to majority of invertebrates, vertebrate ancestor has gone through 2R WGD and teleost ancestor has further experienced an extra TGD, which contributed to the greatly successful radiation and diversification of their genetic complexity [22–29]. By comparing microRNA contents in different chordate lineages (Fig. 4), the emergence/gain of microRNAs well reflected the functional consequences of vertebrate WGD, and the paralogue numbers of conserved microRNAs in various chordate lineages are also in

line with those WGD events. Our findings suggest that microRNA landscape could be a useful indicator of WGD event.

Conclusions

This study provides a new genomic resource of the four-finger threadfin *E. tetradactylum*, and represents the first available genome sequence for the biologically interesting and economically important threadfin fishes (family Polynemidae). In particular, our analyses help to facilitate studies on the developmental biology of the four-finger threadfin, and comparative developmental genomics analyses among perciform fish more generally. With their great diversity, wide variation in body forms, and specialized morphological adaptations, in combination with rapidly increasing genomic resources, the perciform fishes are fast emerging as a vertebrate group with great potential to further the study of developmental genomics. Additionally, we provide a detailed analysis of the repeat content of the four-finger threadfin genome, highlighting the very low repeat content present. It is currently unclear why repeat content varies so greatly among perciform fish, and what mechanisms drive largescale reductions in repeat content in certain genomes. However, the analysis of unsampled lineages from across phylogenetic diversity, as undertaken here, lays important groundwork for further exploration and elucidation of these patterns. Overall, the analyses and genomic resources provided here provide a starting point for further advances in our understanding of the genomics of the unusual threadfin fishes.

Methods

Genomic DNA extraction, sequencing and assembly

Frozen flesh of a single individual of *E. tetradactylum* was obtained from a Sai Kung market at Sai Kung, Hong Kong. Genomic DNA was isolated from muscle tissue using the PureLink Genomic DNA Kit (Invitrogen), and species identity was confirmed with *COI* barcoding. The DNA sample was sent to Novogene (Hong Kong) for library preparation and sequencing on the Illumina HiSeq X system. Chromium WGS reads were assembled using Supernova (v2.1.1) with default parameters (<https://support.10xgenomics.com/de-novo-assembly/software/pipelines/latest/using/running>), and the Supernova pseudohap assembly output was used for further analysis. Genome size estimation was analyzed using a k-mer-based statistical approach in the GenomeScope webtool [39]. Completeness of genome assembly was examined by BUSCO (v4.0.0, metazoa_odb10, actinopterygii_odb10, 8).

Repetitive elements annotation

Repetitive elements were identified as previously described pipeline [40, 41] with the *chordata* RepBase dataset [42]. Subsequently, the resulting de novo repeat library was utilised to identify repetitive elements using

RepeatMasker [43], by implementing two approaches. Firstly, a standard RepeatMasker analysis was performed. Secondly, repeat models were maximised using an automated process implemented in RepeatCraft [44] under strict merge parameters with LTR_FINDER v1.0.5 [45] and the LTR_FINDER_Parallel wrapper [46] to defragment repeat segments. For loci where RepeatMasker annotations overlapped (i.e where the same sequence was annotated as different repeat families), only the longest repeat was selected. This is a conservative approach that ensures TE content estimates are not inflated by counting the same bases multiple times, and facilitates a one-to-one matching of sequence with repeat family identity. A revised summary table was constructed using the revised repeat counts for the second approach, which are presented alongside the bare RepeatMasker results. Rstudio v1.2.1335 [47] with R v3.5.1 [48] and ggplot2 ver. 3.2.1 [49] was used to generate all the plots.

Genome and microRNA annotation

Raw sequencing reads from 4 transcriptome datasets were downloaded from the Sequence Read Archive (SRA) (SRR7899951, SRR7899952, SRR7899953 and SRR7899954) for gene model prediction using Trimmomatic [50], Funannotate [51], Trinity [52] and PASA [53] as previously described procedures and parameters [41, 54].

Precursor sequences of microRNAs of known chordate species were retrieved from both miRbase and MirGeneDB (Supplementary data S3, [19, 20]), and used to search for homologous sequences in the *E. tetradactylum* genome using BLASTN with the following parameters: -r 5 -q 4 -G 8 -E 6 -e value 1. Results were also manually inspected for good sequence conservation and hairpin folding by CentroidFold [55].

Gene family and phylogenetic analyses

Gene models of goldfish (*Carassius auratus*), zebrafish (*Danio rerio*), salmon (*Salmo salar*) and human (*Homo sapiens*) were download from Goldfish genome project (<https://research.nhgri.nih.gov/goldfish>) and NCBI (GCF_000002035.6, GCF_000233375.1, GCF_000001405.39) respectively, and further compared with *E. tetradactylum* using all-against-all BLASTP alignment (*E*-value of 10^{-5}) and OrthoMCL (v2.0.9, inflation value of 1.5, [56]). The links to the gene models of 12 percid fish are shown in Supplementary Table S3. The gene models were then compared with *E. tetradactylum* using the same parameters.

For analysis of *Hox*, *ParaHox*, and *Sox* genes, reference sequences were obtained from HomeoDB ([57], <http://homeodb.zoo.ox.ac.uk/>), NCBI (<https://www.ncbi.nlm.nih.gov/>), Uniprot (<https://www.uniprot.org/>) and relevant genomes in Ensembl database (<https://asia.ensembl.org/index.html>) as queries to carry out tBLASTn [58]

searches to retrieve protein coding gene sequences from the *E. tetradactylum* genome (reference sequences used in this study could be found in Supplementary data S3). Each putatively identified gene was also compared to sequences in the NCBI nr database. Further, protein sequences were aligned to other known members of putative gene families using MAFFT [59], and phylogenetic trees were constructed using MEGA [60] and displayed using iTOL [61].

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-020-07145-1>.

Additional file 1. (XLSX 49 kb)

Additional file 2: Table S1. Results of BUSCO assessment. **Table S2.** Protein-coding genes in various percid fish. **Table S3.** Details of the genome, protein, and gff files of species used in this study. **Figure S1.** Genome size estimation of *E. tetradactylum*. **Figure S2.** Phylogenetic analysis of Hox genes by Neighbor-Joining (A) and Maximum Likelihood (B) method. **Figure S3.** Phylogenetic analysis of ParaHox genes by Neighbor-Joining (A) and Maximum Likelihood (B) method. **Figure S4.** Phylogenetic analysis of Sox family genes by Maximum Likelihood (A) and Neighbor-Joining (B) method. (XLSX 49 kb)

Abbreviations

WGD: Whole genome duplication; TGD: Teleost specific genome duplication; WGS: Whole genome sequencing; Cau: Goldfish *Carassius auratus*; Dre: Zebrafish *Danio rerio*; Ete/ Etet: Four-finger threadfin *Eleutheronema tetradactylum*; Hsa: Human *Homo sapiens*; Ssa: Atlantic salmon *Salmo salar*; Carg: *Channa argus*; Ecr: *Etheostoma cragini*; Espe: *Etheostoma spectabile*; Lcro: *Larimichthys crocea*; Lmac: *Lateolabrax maculata*; Nalb: *Nibeia albiflora*; Pcha: *Parachaenichthys charcoti*; Pfla: *Perca flavescens*; Pflu: *Perca fluviatilis*; Sluc: *Sander lucioperca*; Ssin: *Sillago sinica*; Loc: spotted gar *Lepisosteus oculatus*; Tru/ Fru: *Fugu Takifugu rubripes*; Cca: Common carp *Cyprinus carpio*; Bfl: *Branchiostoma floridae*; Cin: *Ciona intestinalis*; Pma: *Petromyzon marinus*; Sto: *Scyliorhinus torazame*; Abu: *Astatotilapia burtoni*; Mze: *Metriaclima zebra*; Nbr: *Neolamprologus brichardi*; Oni: *Oreochromis niloticus*; Pny: *Pundamilia nyererei*; Tni: *Tetraodon nigroviridis*; Gmo: *Gadus morhua*; Ola: *Oryzias latipes*

Acknowledgements

Not applicable.

Authors' contributions

ZQ performed the microRNA, Hox, ParaHox and Sox analyses. WN carried out the genome assembly, gene model prediction and orthologous gene comparison. YY annotated Hox and ParaHox genes. TB and AH carried out repeat content analyses. HYY provided logistics arrangement of the project. JHLH conceived and supervised the study. All authors revised and contributed to the final version of the text.

Funding

This work was supported by the TUYF Charitable Trust and the 8th Hong Kong University Grant Council Matching Scheme. AH is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (BB/N020146/1). TB is supported by a studentship from the BBSRC-funded South West Biosciences Doctoral Training Partnership (BB/M009122/1). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The genome assembly data can be found at NCBI with accession number WFKG00000000. Datasets of SRR7899951, SRR7899952, SRR7899953, SRR7899954 were retrieved from NCBI (<https://www.ncbi.nlm.nih.gov/>). Gene models of goldfish (*Carassius auratus*), zebrafish (*Danio rerio*), salmon (*Salmo*

salar) and human (*Homo sapiens*) were downloaded from Goldfish genome project (<https://research.nhgri.nih.gov/goldfish>) and NCBI (<https://www.ncbi.nlm.nih.gov/>, GCF_000002035.6, GCF_000233375.1, GCF_000001405.39). The sources for obtaining gene models of 12 percid fish are shown in Supplementary Table S3, and the microRNA precursor sequences of known chordate species and reference sequences of *Hox*, *ParaHox* and *Sox* genes are shown in Supplementary data S3.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors have approved the publication.

Competing interests

The authors declare no conflicts of interest.

Received: 6 July 2020 Accepted: 12 October 2020

Published online: 19 October 2020

References

- Ravi V, Venkatesh B. The divergent genomes of Teleosts. *Annu Rev Anim Biosci.* 2018;6:47–68. <https://doi.org/10.1146/annurev-animal-030117-014821>.
- Nelson JS. *Fishes of the World*. 4th ed. Hoboken: Wiley; 2006. ISBN 978-0-471-25031-9.
- Hastings P, Walker H, Galland G. *Fishes: a guide to their diversity*. Oakland: University of California Press; 2015.
- Ballagh AC, Welch DJ, Newman SJ, Allsop Q, Stapley JM. Stock structure of the blue threadfin (*Eleutheronema tetradactylum*) across northern Australia derived from life-history characteristics. *Fish Res.* 2012;121–122:63–72.
- Motomura H. Threadfins of the world (family Polynemidae): an annotated and illustrated catalogue of polynemid species known to date. Food Agriculture Organisation Species Catalogue for Fishery Purposes. 2004; 117(3):151. <http://www.fao.org/3/y5398e/y5398e00.htm>.
- Motomura H, Matsuura K, Bishop J, Kaymaram F. *Eleutheronema tetradactylum*. The IUCN Red List of Threatened Species. 2015:e.T46087646A57168342. <https://www.iucnredlist.org/species/46087646/57168342>.
- Shihab I, Gopalakrishnan A, Vineesh N, Muktha M, Akhilesh KV, Vijayagopal P. Histological profiling of gonads depicting protandrous hermaphroditism in *Eleutheronema tetradactylum*. *J Fish Biol.* 2017;90:2402–11.
- Seppely M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol.* 2019;1962:227–45.
- Joly-Lopez Z, Bureau TE. Exaptation of transposable element coding sequences. *Curr Opin Genet Dev.* 2018;49:34–42.
- Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet.* 2012;46:21–42.
- Han Z, Li W, Zhu W, Sun S, Ye K, Xie Y, Wang Z. Near-complete genome assembly and annotation of the yellow drum (*Nibeia albiflora*) provide insights into population and evolutionary characteristics of this species. *Ecol Evol.* 2019;9(1):568–75. <https://doi.org/10.1002/ece3.4778>.
- Ferrier DEK, Holland PWH. Ancient origin of the Hox gene cluster. *Nat Rev Genet.* 2001;2(1):33–8.
- Mallo M, Alonso CR. The regulation of Hox gene expression during animal development. *Development.* 2013;140:3951–63. <https://doi.org/10.1242/dev.068346>.
- Mulley JF, C-h C, Holland PWH. Breakup of a homeobox cluster after genome duplication in teleosts. *Proc Natl Acad Sci.* 2006;103(27):10369–72.
- Siegel N, Hoegg S, Salzburger W, Braasch I, Meyer A. Comparative genomics of ParaHox clusters of teleost fishes: gene cluster breakup and the retention of gene sets following whole genome duplications. *BMC Genomics.* 2007; 8(1):312.
- Shimeld SM, Holland PWH. Vertebrate innovations. *Proc Natl Acad Sci.* 2000; 97:4449–52. <https://doi.org/10.1073/pnas.97.9.4449>.
- Bowles J, Schepers G, Koopman P. Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Dev Biol.* 2000;227(2):239–55.
- Wegner M. From head to toes: the multiple facets of sox proteins. *Nucleic Acids Res.* 1999;27(6):1409–20.

19. Kozomara A, Griffiths-Jones S. MiRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 2014;42(D1):68–73.
20. Fromm B, Domanska D, Høyse E, Ovchinnikov V, Kang W, Aparicio-Puerta E, Johansen M, Flatmark K, Mathelier A, Hovig E, et al. MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.* 2019;48(D1):132–41.
21. Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet.* 2016;48(4):427–37.
22. Ohno S. *Evolution by gene duplication.* Berlin: Springer-Verlag; 1970. p. 160.
23. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 2005;3(10):e314.
24. Taylor JS, Braasch I, Frickey T, Meyer A, van de Peer Y. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.* 2003;13:382–90. <https://doi.org/10.1101/gr.640303>.
25. Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature.* 2004;431(7011):946–57.
26. Amores A, Force A, Yan Y-L, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang Y-L, et al. Zebrafish Hox clusters and vertebrate genome evolution. *Science.* 1998;282(5394):1711.
27. Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics.* 2011;188:799–808.
28. Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol.* 2004;21:1146–51.
29. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, et al. The medaka draft genome and insights into vertebrate genome evolution. *Nature.* 2007;447(7145):714–9.
30. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al. The Atlantic salmon genome provides insights into rediploidization. *Nature.* 2016;533(7602):200–5.
31. Chen Z, Omori Y, Koren S, Shirokita T, Kuroda T, Miyamoto A, Wada H, Fujiyama A, Toyoda A, Zhang S, et al. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. *Sci Adv.* 2019;5(6):eaav0547.
32. Nguinkal JA, Brunner RM, Verleih M, Rebl A, de Los R-PL, Schäfer N, Hadlich F, Stüeken M, Wittenburg D, Goldammer T. The First Highly Contiguous Genome Assembly of Pikeperch (*Sander lucioperca*), an Emerging Aquaculture Species in Europe. *Genes (Basel).* 2019;10(9). <https://doi.org/10.3390/genes10090708>.
33. Xu S, Xiao S, Zhu S, Zeng X, Luo J, Liu J, Gao T, Chen N. A draft genome assembly of the Chinese sillago (*Sillago sinica*), the first reference genome for Sillaginidae fishes. *Gigascience.* 2018;7(9):gij108. <https://academic.oup.com/gigascience/article/7/9/gij108/5094561>.
34. Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene.* 2012;509:7–15.
35. Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, Liu Z. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genomics.* 2018;19(1):141.
36. Amores A, Suzuki T, Yan Y-L, Pomeroy J, Singer A, Amemiya C, Postlethwait JH. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res.* 2004;14(1):1–10.
37. Malmström M, Britz R, Matschiner M, Tørresen OK, Hadiaty RK, Yaakob N, Tan HH, Jakobsen KS, Salzburger W, Rüber L. The Most developmentally truncated fishes show extensive Hox gene loss and miniaturized genomes. *Genome Biol Evol.* 2018;10(4):1088–103.
38. Xu P, Zhang X, Wang X, Li J, Liu G, Kuang Y, Xu J, Zheng X, Ren L, Wang G, et al. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nat Genet.* 2014;46(11):1212–9.
39. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* 2017;33:2202–4.
40. Nong W, Law STS, Wong AYP, Baril T, Swale T, Chu LM, Hayward A, Lau DTW, Hui JHL. Chromosomal-level reference genome of the incense tree *Aquilaria sinensis*. *Mol Ecol Resour.* 2020;20(4):971–9.
41. Qu Z, Nong W, So WL, Barton-Owen T, Li Y, Leung TCN, Li C, Baril T, Wong AYP, Swale T, et al. Millipede genomes reveal unique adaptations during myriapod evolution. *PLoS Biol.* 2020;18(9):e3000636.
42. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 2005;110:462–7.
43. Smit AFA, Hubley RR, Green PR. RepeatMasker Open-4.0. <http://repeatmasker.org>. 2013.
44. Wong WY, Simakov O. Repeat craft: a meta-pipeline for repetitive element de-fragmentation and annotation. *Bioinformatics.* 2018;35:1051–2.
45. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 2007;35:W265–8.
46. Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA.* 2019;10.
47. Racine JS. Rstudio: a platform-independent IDE for R and Sweave. *J Appl Econ.* 2013;27:167–72.
48. TeamRC. R: A language and environment for statistical computing. 2013.
49. Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag; 2016. p. 260.
50. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–20.
51. Palmer J, Stajich J. Funannotate: eukaryotic genome annotation pipeline; 2018.
52. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat Protoc.* 2013;8(8):1494–512.
53. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Robin CR, Wortman JR. Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 2008;9.
54. Nong W, Cao J, Li Y, Qu Z, Sun J, Swale T, Yip HY, Qian PY, Qiu J-W, Kwan HS, et al. Jellyfish genomes reveal distinct homeobox gene clusters and conservation of small RNA processing. *Nat Commun.* 2020;11(1):3051.
55. Sato K, Hamada M, Asai K, Mituyama T. CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res.* 2009;37:W277–80.
56. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003;13:2178–89.
57. Zhong YF, Holland PWH. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol Dev.* 2011;13:567–8.
58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
59. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30:772–80. <https://doi.org/10.1093/molbev/mst010>.
60. Kumar S, Stecher G, Li M, Niyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol.* 2018;35:1547–9. <https://doi.org/10.1093/molbev/msy096>.
61. Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 2019;47(W1):W256–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

