




Chromosome-Level Genome Assembly of Chinese Sucker (*Myxocyprinus asiaticus*) Reveals Strongly Conserved Synteny Following a Catostomid-Specific Whole-Genome Duplication

Trevor J. Krabbenhoft ^{1,*}, Daniel J. MacGuigan ², Nathan J.C. Backenstose², Hannah Waterman², Tianying Lan^{2,†}, Jessie A. Pelosi³, Milton Tan ⁴, and Simen R. Sandve⁵

¹Department of Biological Sciences and the RENEW Institute, University at Buffalo, USA

²Department of Biological Sciences, University at Buffalo, USA

³Department of Biology, University of Florida, USA

⁴Illinois Natural History Survey, University of Illinois at Urbana-Champaign, USA

⁵Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway

[†]Present address: Arbor Biosciences, Ann Arbor, MI, USA

*Corresponding author: E-mail: tkrabben@buffalo.edu.

Accepted: 9 August 2021

Abstract

Fishes of the family Catostomidae (“suckers”; Teleostei: Cypriniformes) are hypothesized to have undergone an allopolyploidy event approximately 60 Ma. However, genomic evidence has previously been unavailable to assess this hypothesis. We sequenced and assembled the first chromosome-level catostomid genome, Chinese sucker (*Myxocyprinus asiaticus*), and present clear evidence of a catostomid-specific whole-genome duplication (WGD) event (“Cat-4R”). Our results reveal remarkably strong, conserved synteny since this duplication event, as well as between *Myxocyprinus* and an unduplicated outgroup, zebrafish (*Danio rerio*). Gene content and repetitive elements are also approximately evenly distributed across homeologous chromosomes, suggesting that both sub-genomes retain some function, with no obvious bias in gene fractionation or subgenome dominance. The Cat-4R duplication provides another independent example of genome evolution following WGD in animals, in this case at the extreme end of conserved genome architecture over at least 25.2 Myr since the duplication. The *M. asiaticus* genome is a useful resource for researchers interested in understanding genome evolution following WGD in animals.

Key words: polyploidy, genome stability, genome architecture, allopolyploidy, fractionation, fish.

Significance

We present a high quality, chromosome-level genome assembly for a catostomid fish (“suckers”), the Chinese sucker, *Myxocyprinus asiaticus*. In addition to the three rounds of whole-genome duplication shared with other teleost fishes, we provide the first genome-wide evidence for a catostomid-specific polyploidy event (“Cat-4”). Despite this genome duplication having occurred tens of millions of years ago, the overall genome structure is remarkably stable following the duplication, in contrast to other polyploid fish lineages. These results are an example of extreme genome stability following whole-genome duplication and provide useful evolutionary context to other animal polyploidy events.

Introduction

The fish family Catostomidae (“suckers”) is comprised of approximately 80 cypriniform species, of which only one is extant in Asia, with all others being North American endemics. The Asian species, Chinese sucker (*Myxocyprinus asiaticus*) is

recognized as the sole member of the catostomid subfamily Myxocyprininae (Harris and Mayden 2001; Tan and Armbruster 2018), and is phylogenetically sister to all other extant sucker species, having diverged an estimated ~63 Ma (Bagley et al. 2018). Although *M. asiaticus* is the only surviving

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

catostomid species in Asia (aside from secondary colonization by *Catostomus catostomus*), numerous fossil species have been described across both Asia and North America (Miller and Smith, 1981; Grande et al. 1982; Liu and Chang, 2009).

In addition to the teleost-specific whole-genome duplication (WGD) that occurred around 320–350 Ma (Christoffels et al. 2004; Vandepoele et al. 2004), catostomid fishes underwent an additional ancestral WGD event (“Cat-4R”). Initial evidence for this WGD stemmed from karyotypic and allozyme data (Uyeno and Smith 1972; Ferris and Whitt 1977, 1980). The lack of tetrasomic inheritance in allozyme data and chromosome pairing behavior during meiosis was interpreted to imply that Cat-4R was an allopolyploidy event (reviewed in Ferris [1984]). Subsequent literature has generally assumed this to be the case, but without direct genome-wide sequence data to test that hypothesis.

We sequenced and assembled the first chromosome-level catostomid fish genome, Chinese sucker, *M. asiaticus*, to provide a resource with which to evaluate these hypotheses. This assembly complements a growing list of noncatostomid, cypriniform genomes available, including Tibetan loach (*Triplophysa tibetana*; Yang et al. 2019), zebrafish (*Danio rerio*; Howe et al. 2013), common carp (*Cyprinus carpio*; Xu et al. 2014), and other cyprinids. Our primary aims were to 1) characterize genome structure in this species, 2) identify homeologous chromosome pairs and compare duplicated gene and repeat content among homeologs, and 3) evaluate the degree of neutral genetic differentiation (synonymous substitution rates, K_s) between subgenomes and between Chinese sucker and within and between related species.

Results and Discussion

DNA and Transcriptome Sequencing

A total of 9,069,422 Oxford Nanopore reads were generated, containing 148.3 Gb of sequence data (approximately 62× genome coverage). Nanopore sequences had a read N50 of ~37.9 kb, with the longest read >388 kb (supplementary figs. S1 and S2, Supplementary Material online). An additional 202.3 GB (~80× coverage) of Illumina PE150 Novaseq6000 shotgun genomic reads were generated and used in assembly polishing (supplementary table S1, Supplementary Material online). A total of 281.9 million transcriptome reads were produced across the 12 tissues (84.6 Gb), of which 276.9 million passed QC filtering (83.1 Gb) and were used in de novo transcriptome assembly for gene model evidence (supplementary table S2, Supplementary Material online). These data contained 6.2–7.9 Gb sequence per tissue after filtering.

Genome Assembly

The initial assembly resulted in a contig N50 of 4.19 MB across 1,910 contigs (table 1). Hi-C scaffolding produced 50 scaffolds ranging from 27.7 to 72.2 Mb in length (supplementary

table S3, Supplementary Material online), with scaffold N50 = 49.2 Mb and a final assembly 2.58 Gb in length (within the range of expectation based on flow cytometry; table 1). The final assembly was high quality as evidenced by 98.3% complete BUSCO (table 1 and supplementary fig. S3, Supplementary Material online), and lacked microbial or other contamination based on Kracken2 searches. A total of 54.7% of complete BUSCO genes were duplicated, further supporting the Cat-4R duplication. RepeatModeler identified high levels of repeat content across the genome (53.98% of total assembly), including 21.49% DNA transposons (table 1).

Genome Architecture and Evolution

MAKER3 annotations included a total of 57,229 predicted genes (table 1). We found 87% (49,687) of the predicted genes had at least one BLAST hit against a database of *D. rerio* peptides, with 68% (33,881) of hits >70% sequence similarity and >50 amino acid residues. Also, 69% (39,586) of the predicted genes contained a total of 82,674 putative protein domains. Genome annotation results, Circos plots and dotplots from self–self syntenic mapping revealed clear evidence of the Cat-4R genome duplication (fig. 1A and C). Homeologous chromosomes are readily apparent for all 25 pairs of chromosomes, with remarkably strong, conserved synteny across all pairs. Syntenic mapping between *Myxocyprinus* and zebrafish also showed remarkable stability of genome architecture (fig. 1B) between the two species despite around 100 Myr divergence between the two lineages (Bagley et al. 2018). Again, nearly all 25 chromosome pairs showed strong, shared synteny between the two species. The lone exception is the long arm of zebrafish chromosome 4, which lacks a homologous region in *Myxocyprinus* (fig. 1B, red asterisk). This arm contains a previously identified sex-associated region known to have arisen from a zebrafish lineage-specific expansion (Howe et al. 2013). Homeologous chromosomes tended to be remarkably similar in length, share similar numbers of annotated genes, and have similar numbers of repeats per megabase length (supplementary fig. S4, Supplementary Material online). These results suggest lack of obvious fractionation bias (i.e., loss of duplication genes and no clear subgenome dominance). Synonymous substitution plots (K_s) of duplicate gene pairs also provide evidence for the Cat-4R WGD event. The large peak in the histogram around $K_s = 0.184$ corresponds to the Cat-4R duplication (fig. 1D). Genes with K_s values near this value tend to be found in the duplicated blocks in figure 1C (blue–purple colors). In addition to the Cat-4R peak, a smaller peak corresponding to the teleost-specific duplication (3R) is also seen in figure 1D (red colors). Given a mean estimate of 335 Myr since the 3R duplication and modal K_s for the 3R peak of 2.441 we estimated a synonymous substitution rate of 0.007288 substitutions per site per million years. Using this rate, we estimate 25.2 Myr since the Cat-4R genome duplication.

Table 1

Genome Assembly Statistics

Assembly	Number of Contigs (initial assembly)	1,920
	Longest contig (initial assembly) (Mb)	28.88
	Contig N50 (initial assembly) (Mb)	4.19
	Final assembly length (Gb)	2.58
	Number of scaffolds (chromosomes)	50
	Number of contigs (final assembly)	904
	Contig N50 (final assembly) (Mb)	5.5
	Scaffold N50 (Mb)	49.2
	<i>N</i> (%)	0.003
	GC (%)	39.02
BUSCO v5 (actinopterygii_odb10)	Complete (%)	98.3
	Complete single copy (%)	44.6
	Complete duplicated (%)	53.7
	Fragmented (%)	0.7
	Missing (%)	1.0
Repetitive elements	Total (%)	53.98
	SINEs (%)	0.25
	LINEs (%)	4.60
	LTR elements (%)	7.86
	DNA transposons (%)	21.49
	Unclassified (%)	19.77
Annotation	Predicted genes	57,229
	Mean [median] gene length (bp)	13,532.7 [8,018]
	Mean [median] exon length (bp)	195.9 [128]
	Mean [median] intron length (bp)	1,735.3 [541]
	Mean [median] exons per gene	7.9 [5]
	Mean [median] introns per gene	6.9 [4]

In summary, the *Myxocyprinus* genome assembly provides a useful resource and displays remarkably strong, conserved synteny and genome architecture following a WGD event at least 25.2 Ma. Future work should focus on mechanisms maintaining this apparent genome stability and evaluate the extent of subgenome dominance, biased fractionation, and fate of duplicated genes after the Cat-4R duplication. Additionally, more work is needed to reconcile the Cat-4R time estimates with dates based on the fossil record and molecular phylogenetics (e.g., Bagley et al. 2018). This first catostomid genome assembly adds to a growing list of polyploid animal genomes to complement the wealth of research in plant and fungi systems.

Materials and Methods

One juvenile *M. asiaticus* individual was obtained from Mystic Blue Aquarium Store (Buffalo, NY, USA; [supplementary fig. S1, Supplementary Material](#) online). The fish was sacrificed with an overdose of MS-222 and tissues were immediately dissected and placed in RNA-later and stored at -80°C ([supplementary table S1, Supplementary Material](#) online). Total RNA was extracted using the RNA-Mini kit (Invitrogen Inc.) with Trizol and DNase treatment following the manufacturer's recommendations. RNA from the 12 tissues was sent to Novogene,

Inc. (Davis, CA, USA) for library preparation and mRNA sequencing on the NovaSeq6000 platform. Skeletal muscle tissue was mechanically homogenized in liquid nitrogen and high-molecular-weight (HMW) DNA was extracted using Qiagen Genomic-tip 500/G kit following the manufacturer's protocol. HMW DNA was quantified on the Qubit 3.0 and used in ligation kit library preparation (LSK-109) and sequenced using 12 R9.4 flow cells on the Oxford Nanopore Technologies GridION platform. Base-calling was conducted using Guppy (v3.0.7; nanoporetech.com). Additionally, Illumina PE150 Novaseq6000 shotgun genomic reads were generated by Novogene (Davis, CA, USA) for assembly polishing.

Genome Assembly and Scaffolding

The expected genome size was approximately 1.98 Gb (2.02 pg; Suzuki 1992) or 2.69 Gb (2.75 pg; Zhu et al. 2012) and with $n = 50$ chromosomes based on flow cytometry and karyotyping. Raw nanopore reads passing Guppy QC filters were filtered by discarding reads $< 1,000$ bp. The remaining reads were concatenated and mapped pairwise using Minimap2 (v2.17; Li 2018) with the $-x$ ava-ont flag. Miniasm (v0.3-r179; Li 2016) was used to produce an initial assembly based on the Minimap2 results. Consensus sequences from the Miniasm assembly were generated using two

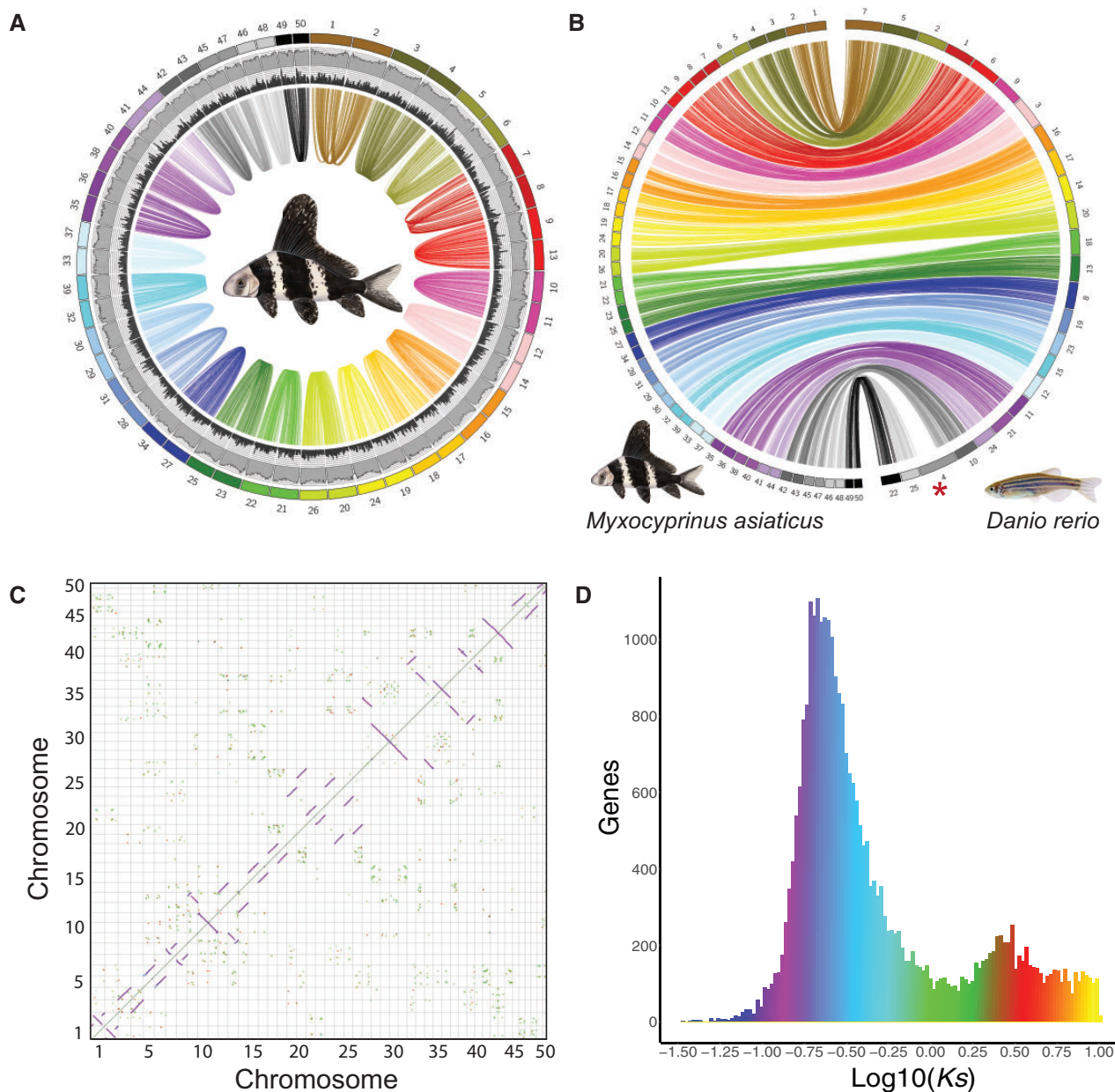


FIG. 1.—Circos plots illustrating *Myxocyprinus* self-self synteny mapping (panel A) or *Myxocyprinus* versus zebrafish (panel B). Arcs reflect blocks with shared synteny between chromosomes and barplots in panel A reflect gene content (inner black ring) or repeats per MB (outer gray ring). Dotplot comparisons of *Myxocyprinus* self-self (panel C) and synonymous substitution histogram (K_s ; panel D). Panel C is a mirror image above and below the diagonal. The blue peak in panel D reflects the Cat-4R duplication, whereas the red peak represents the teleost 3R duplication; point colors in panel C correspond to panel D values. The red asterisk in panel B denotes the gap in synteny with the long arm of zebrafish chromosome 4, a zebrafish-specific sex-associated region.

rounds of Racon (v1.4.11; Vaser et al. 2017) and polished with two rounds of Pilon (v1.23; Walker et al. 2014) using the Illumina reads. Purge_haplotigs (Roach et al. 2018) was used to identify and remove possible haplotigs and high coverage areas. Microbial and other contamination was searched for in the final assembly using Kraken2 (v2.1.0; Wood et al. 2019) using archaea, bacteria, plasmid, viral, human, fungi,

plant, and protozoa databases. BUSCO (v5.1.2; Seppey et al. 2019) was used to assess completeness of the final assembly (without gene models) using actinopterygii_odb10 and to assess the relative fraction of duplicated genes. Dovetail Genomics Inc. (Scotts Valley, CA, USA) produced and sequenced Chicago and Hi-C libraries for genome scaffolding using skeletal muscle tissue as starting material.

Transcriptome Assembly and Genome Annotation

Raw sequence data were run through TrimGalore (Martin, 2011) with default settings to remove all excess adaptors. The RNA-seq data was pooled across the 12 tissues and assembled de novo using the Trinity pipeline (v2.6.6; Grabherr et al. 2011) with the `–trimmomatic` flag. Gene models were generated using the MAKER3 pipeline (v.3.01.03 Holt and Yandell 2011; Campbell et al. 2014). First, repeats were identified and soft masked using RepeatModeler v2.0.1 (Flynn et al. 2020). Coding sequences from the Uniprot-SwissProt database (Poux et al. 2017) present in the custom repeat library were identified and discarded using BLAST (Altschul et al. 1990). The custom repeat library was combined with the RepBase v.23.09 vertebrate database (Bao et al. 2015) and a GFF file was generated using RepeatMasker v.4.1.1 (Smit et al. 2015). Additionally, we supplied MAKER3 with the transposable element library included with RepeatMasker for additional repeat masking. All repeats were soft-masked (`softmask = 1`) for annotation with MAKER3.

Three rounds of genome annotation were performed with MAKER3. First, we aligned evidence to the reference genome. For evidence, we supplied proteomes from two close relatives of the Chinese sucker: *C. carpio* (RefSeq GCF_000951615.1) and *D. rerio* (RefSeq GCF_000002035.6). We also used the de novo Trinity-assembled transcriptome as evidence. The first round of MAKER3 with evidence alignment was used to train the gene prediction software SNAP v2013-02-16 (Korf 2004) and Augustus v.19.12.2006 (Stanke et al. 2006). To avoid poor training performance, only gene models with a minimum length of 50 amino acids and a minimum annotation edit distance of 0.25 were included to train SNAP. Augustus training was performed using BUSCO v.3.1.0 (Simão et al. 2015), specifying zebrafish (*D. rerio*) as the initial gene model and using the BUSCO vertebrata_odb9 database.

One round of ab initio annotation with MAKER3 was performed using SNAP and Augustus gene models, SNAP and Augustus were retrained, and a final round of ab initio annotation was performed. Both rounds of ab initio gene prediction used all protein, RNA, and repeat alignments as evidence. Annotation statistics were generated using custom scripts and genestats (<https://gist.github.com/darencard/fcb32168c243b92734e85c5f8b59a1c3>; last accessed August 18, 2021). To identify putative functions of predicted proteins, a BLAST v.2.2.31+ (Camacho et al. 2009) search was performed using *D. rerio* peptides (RefSeq GCF_000002035.6) as the reference database. Results were filtered to include only matches with >70% sequence similarity and matches longer than 50 amino acid residues. In addition, InterProScan v.5.52-86.0 (Jones et al. 2014) was used to assign putative protein domain information to the predicted genes.

Syntenic Mapping and K_s Plots

A self–self syntenic map was produced using CoGe (Comparative Genomics, genomeevolution.org; Lyons et al. 2008). CoGe’s SynMap2 was used to identify syntenic gene pairs with LASTZ searches and the DAGChainer algorithm (Haug-Baltzell et al. 2017), requiring a maximum distance between two matches (`–D`) = 20 and minimum of five aligned pairs (`–A`). Circos (Krzywinski et al. 2009) was used to generate plots based on synteny using the DAGChainer alignments file produced by CoGe SynMap2 (Haug-Baltzell et al. 2017) for self–self *Myxocyprinus* comparisons and *Myxocyprinus* versus zebrafish. In both instances, DAGChainer results were filtered to retain only one LASTZ hit per query with the highest percent sequence identity. Dotplots were produced in a similar manner using CoGe SynMap2 and points were colored by mean synonymous substitution values (K_s) for each syntenic gene. Synonymous substitutions were calculated using CodeML of the PAML package (Yang 2007), as implemented in CoGe.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Victor Albert for helpful discussions about polyploidy and comments on the manuscript and Tom Dowling and Jerry Smith for their insight on catostomid evolutionary history. This research was conducted in accordance with IACUC protocol #201900161. The *Myxocyprinus* illustration was a commission created by Julie Johnson (Life Science Studios). Funding was provided by the University at Buffalo, laboratory startup account to T.J.K.

Data Availability

Data presented in this article are available from NCBI (Project PRJNA739167; BioSamples SAMN19774987–SAMN19774992).

Literature Cited

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215(3):403–410.
- Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* 6:11.
- Bagley JC, Mayden RL, Harris PM. 2018. Phylogeny and divergence times of suckers (Cypriniformes: Catostomidae) inferred from Bayesian total-evidence analyses of molecules, morphology, and fossils. *PeerJ.* 6:e5168.

- Camacho C, et al. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*. 10(1):421.
- Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. *Curr Protoc Bioinformatics*. 48(1):4–11.
- Christoffels A, et al. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol*. 21(6):1146–1151.
- Ferris SD, 1984. Tetraploidy and the evolution of the catostomid fishes. In: Turner B, editor. *Evolutionary genetics of fishes*. Boston: Springer. p. 55–93.
- Ferris SD, Whitt GS. 1977. Loss of duplicate gene expression after polyploidisation. *Nature* 265(5591):258–260.
- Ferris SD, Whitt GS. 1980. Genetic variability in species with extensive gene duplication: the tetraploid catostomid fishes. *Am Nat*. 115(5):650–666.
- Flynn JM, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*. 117(17):9451–9457.
- Grabherr MG, et al. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol*. 29(7):644–652.
- Grande L, Eastman JT, Cavender TM. 1982. *Amyzon gosiutensis*, a new catostomid fish from the Green River Formation. *Copeia* 1982(3):523–532.
- Harris PM, Mayden RL. 2001. Phylogenetic relationships of major clades of Catostomidae (Teleostei: Cypriniformes) as inferred from mitochondrial SSU and LSU rDNA sequences. *Mol Phylogenet Evol*. 20(2):225–237.
- Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E. 2017. SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* 33(14):2197–2198.
- Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. 12(1):1–14.
- Howe K, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496(7446):498–503.
- Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240.
- Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res*. 19(9):1639–1645.
- Li H. 2016. Minimap and minimiasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 32(14):2103–2110.
- Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18):3094–3100.
- Liu J, Chang MM. 2009. A new Eocene catostomid (Teleostei: Cypriniformes) from northeastern China and early divergence of Catostomidae. *Sci China Ser D Earth Sci*. 52(2):189–202.
- Lyons E, et al. 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol*. 148(4):1772–1781.
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet J*. 17(1):10–12.
- Miller RR, Smith GR. 1981. Distribution and evolution of *Chasmistes* (Pisces: Catostomidae) in western North America. *Occasional Papers of the University of Michigan Museum of Zoology*. No. 696, 46 pp.
- Poux S, et al. 2017. The UniProt Consortium, On expert curation and scalability: uniProtKB/Swiss-Prot as a case study. *Bioinformatics* 33(21):3454–3460.
- Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*. 19(1):1–10.
- Seppy M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M, editor. *Gene prediction*. New York: Humana Press. p. 227–245.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.
- Smit AFA, Hubley R, Green P. 2015. RepeatMasker open-4.0. Available from: <http://www.repeatmasker.org>. Accessed August 18, 2021.
- Smith GR. 1981. Late cenozoic freshwater fishes of North America. *Annu Rev Ecol Syst*. 12(1):163–193.
- Stanke M, et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*. 34(Suppl 2):W435–W439.
- Suzuki A. 1992. Karyotype and DNA contents of the Chinese catostomid fish, *Myxocyprinus asiaticus*. *Chromosome Inf Serv*. 53:25–27.
- Tan M, Armbruster JW. 2018. Phylogenetic classification of extant genera of fishes of the order Cypriniformes (Teleostei: Ostariophysi). *Zootaxa* 4476(1):6–34.
- Uyeno T, Smith GR. 1972. Tetraploid origin of the karyotype of catostomid fishes. *Science* 175(4022):644–646.
- Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci USA*. 101(6):1638–1643.
- Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 27(5):737–746.
- Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9(11):e112963.
- Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 20(1):1–13.
- Xu P, et al. 2014. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nat Genet*. 46(11):1212–1219.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.
- Yang X, et al. 2019. Chromosome-level genome assembly of *Triplophysa tibetana*, a fish adapted to the harsh high-altitude environment of the Tibetan Plateau. *Mol Ecol Resour*. 19(4):1027–1036.
- Zhu D, et al. 2012. Flow cytometric determination of genome size for eight commercially important fish species in China. *In Vitro Cell Dev Biol Anim*. 48(8):507–517.

Associate editor: Bonnie Fraser