A chromosome-level, haplotype-resolved genome assembly and annotation for the Eurasian minnow (Leuciscidae - *Phoxinus phoxinus*) provide evidence of haplotype diversity

Oriowo Temitope Opeyemi[1#], Chrysostomakis Ioannis[1#], Martin Sebastian[1], Kukowka Sandra[1], Brown Thomas[2], Winkler Sylke[2], Myers Eugene W[2,3], Böhne Astrid[1], Stange Madlen*[1]

[1] Leibniz Institute for the Analysis of Biodiversity Change – Bonn, Germany

[2] Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany

[3] Okinawa Institute of Science and Technology, Okinawa, Japan

shared first authorship

*Corresponding author

Correspondence: m.stange@leibniz-lib.de

ABSTRACT

In this study we present an in-depth analysis of the Eurasian Minnow (*Phoxinus phoxinus*) genome, highlighting its genetic diversity, structural variations, and evolutionary adaptations. We generated an annotated haplotype-phased, chromosome-level genome assembly (2n = 25) by integrating high-fidelity (HiFi) long reads and chromosome conformation capture data (Hi-C). We achieved a haploid length of 940 Mbp for haplome one and 929 Mbp for haplome two with high N50 values of 36.4 Mb and 36.6 Mb and BUSCO scores of 96.9% and 97.2%, indicating a highly complete genome.

We detected notable heterozygosity (1.43%) and a high repeat content (approximately 54%), primarily consisting of DNA transposons, which contribute to genome rearrangements and variations. We found substantial structural variations within the genome, including insertions, deletions, inversions, and translocations. These variations affect genes enriched in functions such as dephosphorylation, developmental pigmentation, phagocytosis, immunity, and stress response.

Protein annotation identified 30,980 mRNAs and 23,497 protein-coding genes with a high completeness score, providing further support for our genome's high contiguity. We performed a gene family evolution analysis by comparing our proteome to ten other teleost species, which identified immune system gene families that prioritise histone-based disease prevention over NLR-based immune responses.

Additionally, demographic analysis indicates historical fluctuations in the effective population size of *P. phoxinus*, likely correlating with past climatic changes.

This annotated, phased reference genome provides a crucial resource for resolving the taxonomic complexity within the genus *Phoxinus* and highlights the importance of haplotype-phased assemblies in understanding haplotype diversity in species characterised by high heterozygosity.

Keywords: *Phoxinus, Haplotype-phased genome assembly, comparative genomics, structural variants, immune system, annotation, transposable elements*

Introduction

Species of the genus *Phoxinus* show considerable morphologic and genetic differences (Kottelat & Freyhof, 2007). For a long time, *Phoxinus phoxinus* (Linnaeus, 1758) - as its common name Eurasian minnow suggests - was thought to be a widespread species throughout Europe. However, the taxonomy of the genus was revealed to be more complex (Kottelat, 2007; Kottelat & Freyhof, 2007) including substantial cryptic species diversity (Palandačić et al., 2015, 2017, 2020). At present, 22 distinct genetic *Phoxinus* lineages have been distinguished, comprising at least 13 recognised species (Palandačić et al., 2020). The neotype locality for *P. phoxinus* was placed on the Agger stream (Kottelat, 2007), a tributary of the river Sieg in Germany that feeds into the Rhine river. The designation of a neotype locality is used when the holotype material and sources used to describe a species are missing. Prior to 2012, minnows were a rare occurrence in that area (Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen, 2020). Taxonomically problematic is the fact that the connection between the species name *P. phoxinus* and its assigned lineage "10" (Knebelsberger et al., 2015) is uncertain. This is due to the lack of genetic analysis of the neotype (NRM-55108) designated by Kottelat, (2007), which comes from an area with multiple present lineages (5b, 10, 12; defined by a combination of morphological and single-gene markers) (Palandačić et al., 2020, 2022). Several studies and researchers are currently dealing with resolving the taxonomic and phylogenetic conundrum this genus poses (Bogutskaya et al., 2023; Corral-Lou et al., 2019; Denys et al., 2020; Palandačić et al., 2017, 2020; Reier et al., 2022; Vucić et al., 2018). Until now, these studies suffer from the lack of a high quality reference genome, which would make it possible to go beyond the analysis of single genes.

Due to the previously unrecognised species diversity, most of what is known about the natural history and behaviour of Eurasian minnows is attributed to *P. phoxinus* or its synonyms. Certain presumptions about minnows are assumed to be universally applicable for the entire genus. In the following we will refer to Eurasion minnows including all cryptic species in the *Phoxinus* genus. For instance, Eurasian minnows are small, schooling freshwater fishes that occupy oxygen-rich, cold streams of flowing waters, lake-shores with wave-movements, or highland lakes (Frost, 1943; Mills, 1987, 1988). They assume a complex role in the river or lake ecosystem as they feed on algae, plant debris, molluscs, crustaceans, and insects (Frost, 1943). *Phoxinus* species interact with smaller animals such as water snails and insect larvae, but also larger fishes such as co-occurring salmonids (Museth et al., 2010). However, seemingly contradictory behaviours have been reported as well. According to Frost (1943), *Phoxinus* spp. spawn over clean gravel and shallow, flowing waters. However, populations from the Baltic Sea reportedly reside and spawn in brackish habitats (Svirgsden et al., 2018). While Eurasian Minnows are sometimes reported to be sensitive to pollution (Bagge & Hakkari, 1992), they are also tolerant of polluted waters (King et al., 2011). They might even benefit from increased primary production in eutrophic waters (King et al., 2011; Moiseenko et al., 2009). In terms of dispersal and migration behaviour, Eurasian minnows from the Volga were less frequently found in migratory habitats (Pavlov & Mikheev, 2017), or demonstrated a tendency for downstream movement in a mesocosm experiment conducted with specimens from Wales (Jones et al., 2021). In contrast, Eurasian minnows were found to migrate upstream during breeding season (Frost, 1943; Svirgsden et al., 2018) as well as through fishways (Jansen et al., 1999) in populations from England, Estonia, and Germany, respectively. The aforementioned reported differences or the seemingly broad occurrence pattern could be due to the fact that studies reporting on *P. phoxinus* or "the Eurasian Minnow" were conducted on different *Phoxinus* species.

A crucial step towards informed and state-of-the-art biodiversity, phylogenetic, and comparative genomic research is the generation of haplotype-resolved, chromosome-level reference genomes. Phasing of haplomes, the haplotype-resolved elements in reference genome assemblies, is now easily facilitated with the latest development of chromosome conformation capture, namely Hi-C (Belton et al., 2012; Lin et al., 2018) and its integration into assembly pipelines, such as Hifiasm (H. Cheng et al., 2021). Resultantly, the number of studies acknowledging genetic variation between haplomes, is rapidly

increasing including e.g. Cavendish banana (Huang et al., 2023), African cassava (Mansfeld et al., 2021), Australian lime (Nakandala et al., 2023), carnation (H. Jiang et al., 2023), cattle (Low et al., 2020), pearl oyster (Takeuchi et al., 2022), and ocellated puffer fish (Zeng et al., 2023).

We conducted comparisons between the two *Phoxinus* haplomes to identify both sequence and structural variations present in a *P. phoxinus* genome. When examining sequence variation, heterozygosity, polymorphism, and allelic diversity were considered (Carley et al., 2022; DeWoody et al., 2021). Structural variation can be examined within and among genomes by investigating the copy number, orientation and location of coding and non-coding sequences on the chromosome (Wellenreuther et al., 2019). Copy number variation comprises deletions, insertions, and duplications, while orientational variation encompasses inversions. Local variation can be attributed to translocations. Structural variation at the genome level has been associated with variation in both phenotypes and complex traits (Mérot et al., 2020; Weischenfeldt et al., 2013).

Ultimately, this reference genome serves two purposes. First, it is to aid the *Phoxinus* community with the necessary basis for phylogenetic unravelling. Second, this haplotype-resolved, chromosome-level reference genome demonstrates once more that genetic diversity contains an individual layer, namely, the intra-individual diversity on the haplome level that needs to find attention in the scientific community.

Methods

Specimen Collection and Sampled Tissues

Two *P. phoxinus* specimens originally from the river Sieg close to the neotype locality, North-Rhine Westphalia, Germany, were sacrificed using Ethyl 3-aminobenzoate methanesulfonate salt (MS-222; Sigma-Aldrich) under permission per § 11 Abs. 1 Nr. 1 b, Sa and 8d Tierschutzgesetz (TierSchG) issued by the Amt für Umwelt, Verbraucherschutz und Lokale Agenda, Lebensmittelüberwachung und Veterinärdienste, under Zeichen 56.2. Tissues from fin, brain, liver, spleen, muscle, heart, skin, and testis were immediately sampled and flash-frozen in liquid nitrogen (individual identifiers ZFMK-TIS-60719 [male] and ZFMK-TIS-60720 [male]. Tissues (brain, liver, spleen, muscle, heart, gonad) from individual ZFMK-TIS-60720 were used for the genome assembly; tissues (brain, gill, gonad, liver, muscle, skin, spleen) from ZFMK-TIS-60719 were used for RNA-sequencing and annotation.

Extraction of Long Genomic DNA

Long genomic DNA (gDNA) of *P. phoxinus* was extracted with the circulomics Nanobind Tissue Big DNA kit (part number NB-900-701-01, protocol version Nanobind Tissue Big DNA Kit Handbook v1.0 (11/19)) according to the manufacturer's instructions. In brief, 11 mg of spleen tissue were minced into small slices on a clean and cold surface. Tissues were finally homogenised with the TissueRuptor II device (Qiagen) making use of its maximal settings. After complete tissue lysis, the remaining cell debris were removed, and the gDNA was bound to circulomics Nanobind discs in the presence of Isopropanol. High molecular weight (HMW) gDNA was eluted from the nanobind discs in EB. The integrity of the HMW gDNA was determined by pulse field gel electrophoresis using the Pippin Pulse™ device (SAGE Science) and the Agilent Femtopulse. The majority of the gDNA was between 20 and more than 400 kb in length. All pipetting steps of long gDNA have been done very carefully with wide-bore pipette tips. HMW gDNA was further purified after extraction with 1x AMPure beads.

PacBio HiFi Library Preparation and Sequencing

A long insert HiFi library was prepared as recommended by Pacific Biosciences according to the "Guidelines for preparing HiFi SMRTbell libraries using the SMRTbell Express Template Prep Kit 2.0" (PN 101-853-100, version 03). In summary, 10 µg of purified HMW gDNA was sheared twice, aiming for 25 and 20 kb fragments, respectively, with the MegaRuptor™ device (Diagenode). The sheared gDNA was enriched with 1x AMPure beads according to the manufacturer's instructions. 6 µg sheared gDNA have been used for Pacbio HiFi library preparation. The PacBio SMRTbell™ library was size-selected for fragments larger than 6 kb with the BluePippin™ device according to the manufacturer's instructions. The final HiFi library had a fragment size of 12.8 kb on the Agilent Fragment Analyzer on the High sensitivity 50kb large fragment kit.

60 pM of the size selected library were run on two Sequel II SMRT cells with the SEQUEL II sequencing kit 2.2 for 30 hours on the SEQUEL II; pre-extension time was 2 hours. PacBio CCS reads (read quality > 0.99) were created from the subreads.bam files using PacBio's ccs command line tool (version 6.3.0) and further refined by applying the tool DeepConsensus (v0.2, on PacBio reads within 98.0-99.5% read accuracy) (Baid et al., 2023). Finally, reads containing the PacBio adapter sequence were filtered out by applying a Blastn search (v2.9) (Altschul et al., 1990) providing the PacBio adapter sequence and the following arguments: ``reward 1 -penalty -5 -gapopen 3 -gapextend 3 -dust no -soft_masking false -evalue 700 -searchsp 1750000000000 -outfmt 7". This resulted in a total yield of 37,3 Gb data with a mean read length of 10.7 kb and a read N50 of 10.2 kb.

156 **ARIMA HiC Chromatin Conformation Capturing**

157 Chromatin conformation capturing was done making use of the HiC+ Kit (Material Nr. A410110) and
158 followed the user guide for animal tissues (ARIMA-HiC 2.0 kit Document Nr: A160162 v00). In brief, 40
159 mg flash-frozen powdered muscle tissue was cross-linked chemically. The cross-linked chromatin was
160 digested with a restriction enzyme cocktail consisting of four restriction enzymes. The 5'-overhangs
161 were filled in and labelled with biotin. Spatially proximal digested DNA ends were ligated, and, finally,
162 the ligated biotin-containing fragments were enriched and went for Illumina library preparation, which
163 followed the ARIMA user guide for Library preparation using the Kapa Hyper Prep kit (ARIMA Document
164 Part Number A160139 v00). The barcoded HiC libraries run on an S4 v1.5 XP flow cell of a NovaSeq6000
165 with 300 cycles.

166 **RNA Extraction and Sequencing**

167 RNA was extracted from nitrogen flash-frozen brain, gill, gonad, liver, muscle, skin, and spleen
168 (specimen identifier ZFMK-TIS-60719) tissues. All tissues except for spleen and liver were extracted
169 using the RNeasy Mini Kit (Qiagen, Cat. No. 74104) suitable for tissue amounts of 0.5 to 30 mg. Spleen
170 and liver were extracted using QIAamp DNA Micro Kit (Qiagen) suitable for starting material of less than
171 5 mg. All tissues were homogenised using PowerBead Tubes Ceramic (Qiagen, diameter 2.8mm; Cat. No.
172 13114-50) on a PowerLyzer™ 24 Bench Top Bead-Based Homogenizer (MO BIO Laboratories, Inc). Before
173 use, 180 µL 1 M dithiothreitol (DTT) was added to 4.5 ml lysis buffer RLT. 350 µL prepared RLT buffer
174 was added to each tissue in a PowerBead Tube (Qiagen) and one cycle of 45 seconds at 3500 rpm and
175 another for 45 seconds at 4200 rpm on the PowerLyzer™ Homogenizer with a 5-minute pause at -20°C
176 were performed. After homogenisation, the solution was centrifuged for 1 minute at 14800 rpm. The
177 lysate including tissue debris was transferred into a fresh 1.5 mL tube (LoBind®) and again centrifuged at
178 14800 rpm for 3 minutes at room temperature (RT). Finally, the lysate was transferred to a gDNA
179 Eliminator column (provided in the QIAamp Kits) and centrifuged at 10,000 rpm for 30 sec at RT.
180 Washing and elution steps followed the standard protocol of the Micro and Mini Kit, respectively.
181 Samples were eluted into 14 µL (spleen, liver) and 30 µL (all other tissues) RNAse-free water during 1
182 minute at 10,000 rpm centrifugation. RNA library preparation and sequencing were performed at
183 Biomarker Technologies (BMKGene) GmbH on an illumina Novaseq 6000 S4 chip.

184 **K-mer based Genome Size and Heterozygosity Estimation**

185 To estimate the genome size and heterozygosity of the genome, a k-mer-based approach was used.
186 K-mers were counted in the HiFi reads using Jellyfish (v2.3.0) (Marçais & Kingsford, 2011) with
187 parameters count *-C -m*, with *-C* indicating that the input reads from both the forward and reverse
188 strand are being used, and -m specifying a k-mer length; we specified k-mer lengths of 19, 23, 25 and 30.
189 The resulting k-mer counts were exported into a k-mer count histogram using the *histo* command of
190 Jellyfish for subsequent analysis in GenomeScope (v2.0) (Ranallo-Benavidez et al., 2020) for estimation
191 of genome size and heterozygosity.

192 ***De novo* genome Assembly and Scaffolding**

193 Initial contigs for two representative Haplomes were generated using HiFiasm (v0.16.1-r375) (H.
194 Cheng et al., 2021) with parameters *--hic -l2* and then purge-dups (v1.2.3) (Guan et al., 2020) was run on
195 each produced haplotype assembly.

196 Initial scaffolding of the primary assembly was performed by mapping HiC reads to the primary
197 contigs using bwa-mem (v0.7.17-r1198-dirty) (H. Li, 2013) and mappings were filtered following the VGP
198 arima mapping pipeline (https://github.com/VGP/vgp-assembly/tree/master/pipeline/salsa). The final
199 bed file was given to yahs (v1.1a) (Zhou et al., 2022) for scaffolding. Scaffolds were then manually
200 curated into chromosomes using higlass (v2.1.11) (Kerpedjiev et al., 2018) with Cooler (v0.9.1)
201 (Abdennur & Mirny, 2020) to visualise the HiC data (Supplementary Figure S1 - S2).

Finally, the two haplotype assemblies were polished by mapping the HiFi reads to the assemblies using pbmm2 (v1.3.0) [https://github.com/PacificBiosciences/pbmm2.git] with arguments --preset CCS -N 1 and variants called using deepvariant (v0.2.0) with --model_type=PACBIO. Errors were corrected in the assembly by filtering the vcf file given by deepvariant (v0.2.0) (Poplin et al., 2018) with bcftools view (v1.12) (Danecek et al., 2021) with arguments *-i 'FILTER=\"PASS\" && GT=\"1/1\"'* and a consensus called with bcftools consensus. Assembly completeness was assessed using BUSCO (v5.4.7) (Simão et al., 2015) ran in genome mode with the actinopterygii_odb10 database.

Annotation of Repetitive Elements

Repeat sequences were identified and annotated in both haplomes of the *P. phoxinus* reference genome by combining homology-based and *de novo* approaches. We identified low-complexity repeats and tandem repeats with DustMasker from BLAST+ (v.2.9) from the NCBI C++ toolkit (Morgulis et al., 2006) and Tandem Repeat Finder (v4.10.0) (Benson, 1999), respectively. We de novo identified Transposable Elements (TEs) with Repeatmodeler (v2.0.1) (Smit and Hubley, 2008) to create a species-specific library, and we combined the identified species-specific transposable elements (TEs) with the *Danio rerio* (zebrafish - Cypriniformes) repeats library from Dfam 3.1 (Hubley et al., 2016) to generate a custom combined repeats library.

Using this custom combined library, repetitive elements were identified and annotated in our genome with RepeatMasker (v.4.1.0) (https://repeatmasker.org/). Divergence of our repeat annotation from the consensus repeats sequence was estimated with the *calcDivergenceFromAlign.pl* script implemented in RepeatMasker using the *.align* output file of the previous masking step. The repeats landscape was plotted with the *createRepeatLandscape.pl* script, also in RepeatMasker. Output files of Duskmasker, Tandem Repeats Finder and RepeatMasker were converted to *gff3* format. We merged the positions of repeats with Bedtools (v2.29.2) (Quinlan, 2014) before our genome was softmasked with these identified repeats using Bedtools (v2.29.2) in preparation for protein prediction.

Protein Annotation

Structural annotation of the protein-coding genes in the P. phoxinus genome was carried out with a combination of ab initio, homology, and transcriptome-based protein prediction models. The sequencing reads (NCBI accession numbers SRR26699630 to SRR26699636) from the seven RNA samples from brain, gill, gonad, liver, muscle, skin, and spleen were mapped to the reference genome with STAR (v2.7.10b) (Dobin et al., 2013), the resulting mapping files (.bam) were merged with samtools (v1.10) (H. Li et al., 2009).

BRAKER3 (Brůna et al., 2021; Gabriel et al., 2023; Hoff et al., 2016), the latest genome annotation pipeline in the BRAKER (Hoff et al., 2019) suite was used for protein annotation. Using our assembled transcripts as input, BRAKER3 predicts protein-coding genes with GeneMarkS-T (Tang et al., 2015) and then selects proteins with the best similarity scores and the quality of *ab initio* predictions with GeneMark-ETP (Brůna et al., 2020). GeneMark-ETP then produces three groups of hints, i) hints with transcript and protein similarity support, ii) hints with transcript and *ab initio* support and finally, iii) hints with only protein similarity support using ProtHint. All these hints are used to create a set of high-confidence genes. AUGUSTUS (Stanke et al., 2006) is trained with these gene sets and does an *ab initio* prediction of another genome-wide gene set before TSEBRA (Gabriel et al., 2021) combines the results of AUGUSTUS and GeneMark-ETP to produce a final set of high-confidence genes. The final set was used as input to BUSCO (v5.4.7), which was run in protein mode with the actinopterygii_odb10 database.

Structural annotation of the resulting predicted genes was done with the *blastp* algorithm of DIAMOND (v2.1.8) (Buchfink et al., 2021). Our predicted proteins were blasted against protein databases of Swissprot edition 2023-05-13 (Boeckmann et al., 2003), The Protein Data Bank (Berman et al., 2000) Pdbaa database edition 2023-05-13 and TrEMBL edition 2023-05-14 (Boeckmann et al., 2003).

The blast was done with an e-value cut-off of < $1*10^{-5}$. Functional annotation of these protein domains was done with eggNOG-mapper (v2.0) (Cantalapiedra et al., 2021).

Protein sequences of *D. rerio* (zebrafish, Cypriniformes), GCF_000002035.6 (Howe et al., 2013); *Sinocyclocheilus rhinocerous* (rhinoceros golden-line barbel, Cypriniformes), GCF_001515625.1 (J. Yang et al., 2016); *Sinocyclocheilus grahami* (golden-line barbel, Cypriniformes), GCF_001515645.1 (J. Yang et al., 2016); *Cyprinus carpio* (common carp, Cypriniformes), GCF_000951615.1 (Xu et al., 2014); *Carassius auratus* (goldfish, Cypriniformes), GCF_003368295.1 (Z. Chen et al., 2019); *Pimephales promelas* (fathead minnow, Cypriniformes), GCF_016745375.1 (Martinson et al., 2022); *Megalobrama amblycephala* (Wuchang bream, Cypriniformes), GCF_018812025.1(H. Liu et al., 2021); *Ctenopharyngodon idella* (grass carp, Cypriniformes), GCF_019924925.1 (Wu et al., 2022); *Myxocyprinus asiaticus* (Chinese high-fin banded shark, Cypriniformes), GCF_019703515.2 (Krabbenhoft et al., 2021); and the outgroup, *Ictalurus punctatu*s (channel catfish, Siluriformes), GCF_001660625.3 (Waldbieser et al., 2023) were retrieved from NCBI RefSeq database (O'Leary et al., 2016) on 2023-07-12. We considered *Pimephales promelas* (Fathead minnow) as a representative of the most closely related taxonomic species to *P. phoxinus,* for which a well-documented whole genome assembly and annotation is readily accessible. AGAT (v1.1.0) (Dainat, 2021) was used to extract the longest gene isoforms against which the protein sequences of the *P. phoxinus* were blasted with DIAMOND (v2.1.8) (Buchfink et al., 2021), to identify homologous sequences.

Genome-wide Comparison between Haplomes

Sequence Variation

Genome-wide average heterozygosity was estimated for both haplomes to assess sequence variation using the 25 largest scaffolds representing the 25 chromosomes. To do so, raw HiFi reads were mapped to both haplomes with minimap2 (v2.1) (H. Li, 2018) with preset *-ax asm20*, which optimises alignments between sequences with less than 20% divergence. The resulting alignment files were filtered for mapping quality lower than 20 with the view *-q 20* command and sorted with the *sort* command using Samtools (v1.10) (H. Li et al., 2009). Duplicates were removed with the markdup -r command as implemented in Sambamba (v0.7.1) (Tarasov et al., 2015). Furthermore, we generated windows of 100 kb with the *makewindows* command implemented in Bedtools (v2.31.0) with *-w 100000.* Then, we estimated depth in 100 kb for both haplomes using *coverage -mean* as implemented in Bedtools (v2.31.0) (Quinlan, 2014) in combination with the file specifying coordinates of 100kb windows created.

To estimate heterozygosity, site allele frequency likelihood was estimated using the GATK model for genotype likelihoods in combination with the *-doSAF 1* parameter in ANGSD v.0.940 (Korneliussen et al., 2014), the site frequency spectrum was estimated with the *realSFS* parameter also in ANGSD. Finally, heterozygosity was estimated from sample allele frequencies in non-overlapping windows of 1 Mb. The number of heterozygous alleles was counted within each Mb and then converted to the number of heterozygous alleles per kb and plotted per chromosome across the genome using ggplot2 (v3.1.3), to visually inspect patterns of heterozygosity across the entire genome. Patterns of heterozygosity often reflect standing genetic variation which may be linked to demographic history of a species (Guevara et al., 2021; Miller et al., 2014; Nadachowska-Brzyska et al., 2016).

Demographic History of P. phoxinus

The demographic history of *P. phoxinus* was reconstructed using Pairwise Sequentially Markovian Coalescent (PSMC) (v0.6.5) as implemented by Li & Durbin, (2011). Variants were called from both haplomes with a combination of samtools (v1.17) *mpileup* (H. Li et al., 2009) and Bcftools (v1.17) *call* (Danecek et al., 2021) commands. The resulting variant file was converted to a diploid consensus file with vcfutils.pl vcf2fq accessory implemented in the PSMC package. The resulting fastq files were converted into the input fasta format for PSMC using the fqpsmcfa accessory. The PSMC model was run with the following parameters: -N25 -t15 -r5 -p "4+25*2+4+6" and 100 bootstraps, a generation time of

295 3 years (Museth et al., 2010), which is the maturation age of the *P. phoxinus* and an assumed mutation
296 rate of 3.51 x10^{-9} (Tian et al., 2022). Results of the run were plotted using ggplot2 (v3.1.3).

297 *Structural Variation*

298 We aligned both haplomes with minimap2 (v2.26) (H. Li, 2018) using the *-asm5* preset setting.
299 Alignments were processed with Synteny and Rearrangements Identifier (SyRI) (v.1.6.3) (Goel et al.,
300 2019) to detect structural variation between both haplomes and the results were visualised using the R
301 package plotsr (v.1.1.1) (Goel & Schneeberger, 2022). Genes overlapping Insertions, deletions and
302 Inversions extracted using bedtool (v2.29.2) (Quinlan, 2014). Subsequently, functional enrichment of
303 these genes was achieved with g:profiler (Raudvere et al., 2019) using the zebrafish database option.
304 Finally, the results were visualised by go-figure (v1.0.2) (Reijnders & Waterhouse, 2021).

305 **Identification of Orthologs and Gene Family Evolution**

306 From the above described proteomes, amino acids shorter than 30 bp were filtered out, and
307 Orthofinder (v2.5.4) (Emms & Kelly, 2019) with default settings was run to infer orthologous gene
308 families. Orthofinder created orthogroup gene trees and then inferred a rooted species tree from gene
309 duplication events with the STRIDE approach (Emms & Kelly, 2017) using just single-copy orthologs. The
310 species tree underwent conversion into an ultrametric tree utilising Orthofinder's *make_ultrametric.py*
311 script, employing the parameter -r 142. This value corresponds to approximately 142 million years,
312 which aligns with the estimated median divergence time between Siluriformes and Cypriniformes, as
313 suggested by TimeTree (Kumar et al., 2017). This estimate is calculated by considering the median of
314 divergence times that have been publicly estimated using molecular sequence data analysis.
315 Orthologous gene counts from the Orthofinder run were filtered to remove ortholog groups that are
316 overrepresented in a particular species; this was done with the *clade_and_size_filter.py* script in CAFE
317 (v5.0.0) (Mendes et al., 2021). Gene family evolution analysis was carried out in CAFE5 (v5.0.0) using the
318 ultrametric tree and filtered orthologous gene counts as input. Genes that were confirmed to be
319 species-specific, expanding and contracting in *P. phoxinus,* were blasted against the TrEMBL database.
320 The GO terms derived from the TrEMBL database were then used as input into REViGO (Supek et al.,
321 2011) to remove redundant terms and summarise GO results.

Results and Discussion

K-mer Based Genome Size and Heterozygosity Estimation

We estimated the haploid length of the *P. phoxinus* genome to be 805.8 Mbp with a unique content of 62.1% and a heterozygosity of 1.43%. This estimate deviates from the determined C-value of 1.15 pg (corresponding to 1.13 Gbp) (genomesize database; https://genomesize.com/). Running genome size estimates using 19, 23 ,25 and 30 k-mer lengths, showed slight variation in the estimated genome sizes and heterozygosities (Supplementary Table S1, Figure S3). We chose the 19-mer length due to a lower error rate in comparison to other k-mer lengths. The estimated heterozygosity of 1.43% is high compared to other published k-mer-based heterozygosity estimates in cyprinid fishes. For example, *Onychostoma macrolepis* has an estimated heterozygosity of 0.29% (L. Sun et al., 2020), *Myxocyprinus asiaticus* (Chinese high-fin banded shark) of 0.20% (Krabbenhoft et al., 2021), *Gymnocypris przewalskii* (Przewalskii's naked carp) of 0.96% (Tian et al., 2022), and *Poropuntius huangchuchieni* of 0.68% (L. Chen et al., 2021). To the best of our knowledge, the highest k-mer-based heterozygosity estimate for a cyprinid is 1.82% estimated in *Gymnocypris eckloni* (F. Wang et al., 2022), while the lowest estimate of 0.1% was found in *Triplophysa tibetana* (X. Yang et al., 2019). This suggests the estimated k-mer-based heterozygosity of *P. phoxinus* is on the high side within cyprinids.

The clear bimodal coverage distribution visible from the k-mer plot (Figure 1) is indicative of the diploidy of the *P. phoxinus* genome, with the first peak at around 18-fold coverage corresponding to heterozygous sites and the second peak at around 36-fold coverage corresponding to homozygous genome positions.

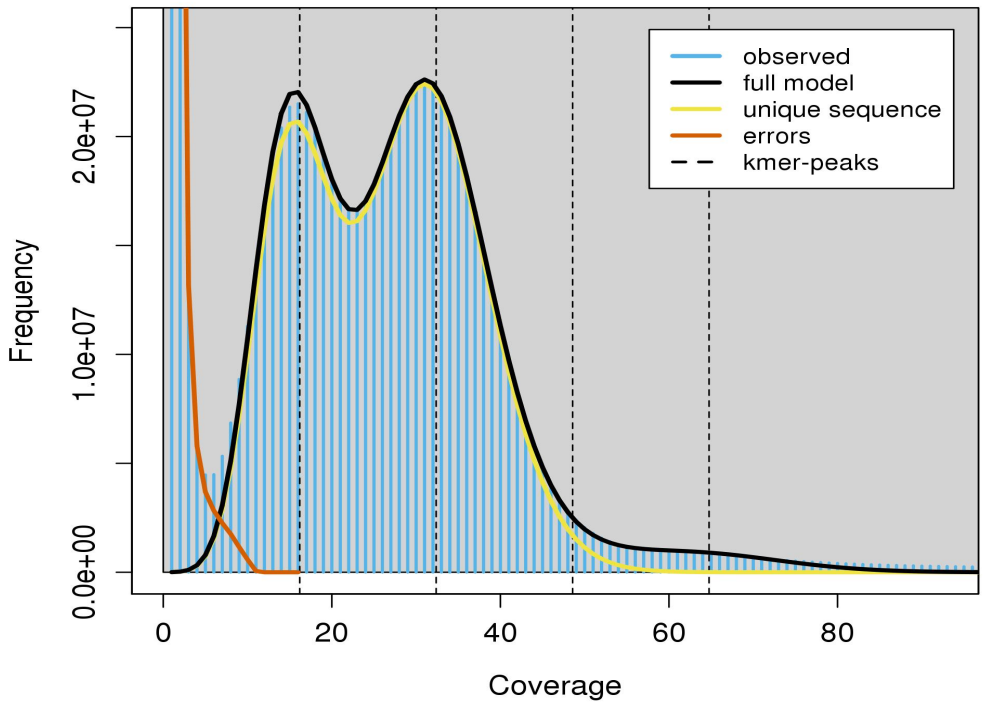


Figure 1: K-mer spectra profile of *Phoxinus phoxinus.* Generated from raw PacBio HiFi reads with GenomeScope2. The y-axis shows the k-mer counts and the x-axis shows sequencing depth. The clear bimodal pattern observed indicates a diploid genome with high heterozygosity of 1.43%. The haploid genome size was estimated to be around 805 Mb.

346

347 **Genome Assembly and Scaffolding**

348 We generated a haplotype resolved assembly of *P. phoxinus*, with a total assembly size of 940 Mbp
349 and 929 Mbp for haplotype 1 (Hap1) and haplotype 2 (Hap2), respectively. The initial assembly of each
350 haplotype contained 394 primary contigs, which were further placed into 101 (N50: 36.4 Mb) and 81
351 (N50: 36.6 Mb) scaffolds, respectively, by mapping HiC reads to the primary contigs. Integrating Hi-C
352 data, we ultimately scaffolded 99.6% of the Hap1 assembly and 99.5% of the Hap2 assembly each into
353 25 chromosomes (Supplementary Table S2). BUSCO analysis revealed the *P. phoxinus* assemblies had
354 only 2.3% and 2.1% missing genes. The assembly statistics for both haplomes (Table 1) denote high
355 contiguity and high completeness. The 11 Mb difference in size between both haplomes is a result of
356 structural variations between haplomes.

357 **Table 1.** Summary of *P. phoxinus* haplotype 1 and haplotype 2 genome assembly statistics.

Assembly Statistics	Haplotype 1	Haplotype 2
Assembly Size	940 Mb	929 Mb
Contigs	394	394
Scaffolds	101	81
N50 Contigs	3.85 Mb	4.59 Mb
N90 Contigs	869 Kb	1.05 Mb
N50 Scaffolds	36.4 Mb	36.6 Mb
N90 Scaffolds	30.7 Mb	30.3 Mb
% of Assembly in Scaffolds	99.6%	99.5%
Complete BUSCOs	3528 (96.9%)	3540 (97.2%)
Complete and Single-copy BUSCOs	3488 (95.8%)	3495 (96.0%)
Complete and duplicated BUSCOs	40 (1.1%)	45 (1.2%)
Fragmented BUSCOs	30 (0.8%)	25 (0.7%)
Missing BUSCOs	82 (2.3%)	75 (2.1%)
Total BUSCO Search	3640	3640

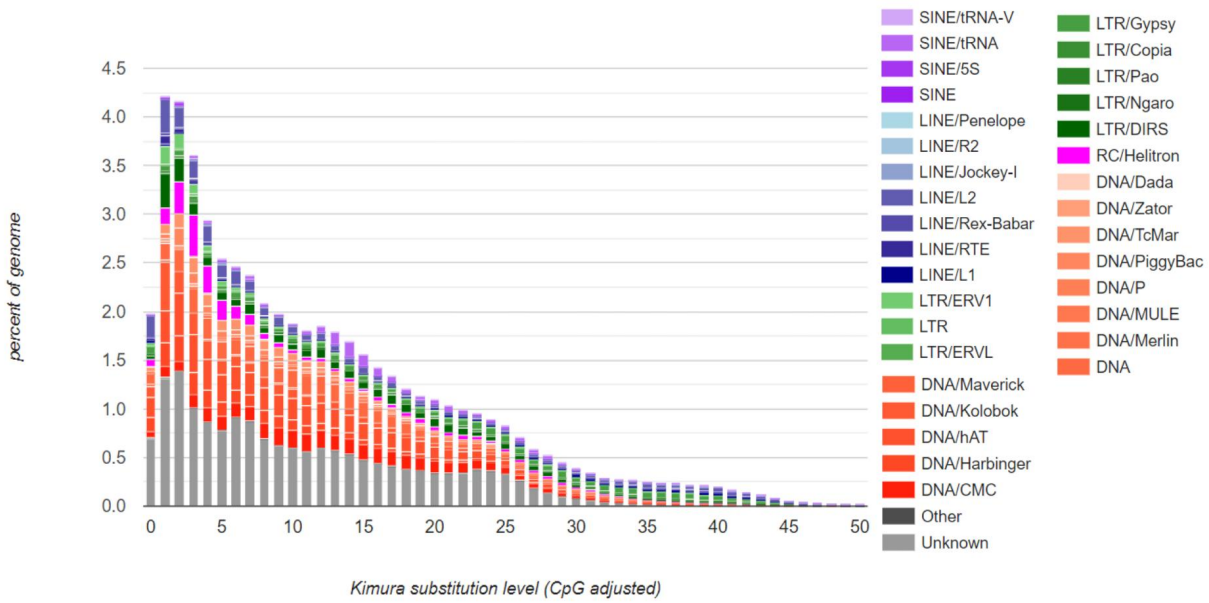
358

359 **Repeats Annotation**

360 Teleost fish genomes are known for their highly diverse repeat content (Gao et al., 2016) and the
361 assembly of the *P. phoxinus* genome follows this precedent. Annotation of the assembly using our
362 custom combined repeats library revealed that around 506 Mb (53.86 %) of the *P. phoxinus* genome is
363 made up of repetitive elements. This is higher than reported to date from other species of the family
364 cyprinids like fathead minnow (43.27%) (Martinson et al., 2022), grass carp (43.26%) (Wu et al., 2022),
365 common carp (31.3%) (Xu et al., 2014) and goldfish (39.6%) (Z. Chen et al., 2019). The repeat content is
366 slightly lower than in zebrafish, which contains around 54.47% repetitive elements (Howe et al., 2013).

367 From the identified TEs, DNA transposons were the most abundant repeat elements in the genome,
368 making up 21.25% (199 Mb) of the genome. SINEs were relatively poorly represented at 1.03%, in line
369 with the expectations derived from other published fish genomes (Shao et al., 2019; Sotero-Caio et al.,
370 2017). DNA transposon content is similar to that of the fathead minnow genome, which contains an

371 estimated 21.47% (Martinson et al., 2022). Retroelements made up 11%; 6.16% LTRs and 4.15% LINEs.
372 Overall, we annotated 46.59% of the *P. phoxinus* genome to be total interspersed repeats, the
373 remaining repetitive elements were annotated as 0.47% small RNA, 1.50% satellites, 2.66% simple
374 repeats, and 0.17% low complexity repeats (full details of all repetitive elements are described in Table
375 S3 and S4).



376

377 **Figure 2: Repeat landscape of the *P. phoxinus* genome.** Distribution of transposable elements (TEs) in the *Phoxinus phoxinus*
378 genome across different Kimura substitution levels (in %). TEs on the left side of the histogram represent recently active TEs
379 with low divergence from the consensus sequence of TEs. TEs towards the right side of the histogram represent ancient TEs
380 with higher degrees of divergence from the consensus TE sequences.

381 Repeat landscapes illustrate how TEs cluster in correlation to the Kimura substitution rate, i.e. how
382 quickly TEs in a genome have diverged from the consensus TE sequence (Shao et al., 2019). The Kimura
383 substitution level is correlated with the age of transposition activity: low substitution levels indicate
384 recent transposition events, higher substitution levels suggest old transposition events. The repeat
385 landscape of the annotated *P. phoxinus* genome (Figure 2) indicates a high level of active, i.e., recent,
386 transposition bursts (Watanabe et al., 2014). In fish genomes, active transposons accumulate at a faster
387 rate than they decline (Shao et al., 2019), leading to an abundance of active transposons with low
388 kimura substitution levels (Böhne et al., 2012; Gao et al., 2016; Sotero-Caio et al., 2017), an exact
389 pattern observed in TEs in the *P. phoxinus* genome. The repeat landscape for haplotype 2 can be found
390 in Supplementary Figure S4 .

391 Active transposition and high repeat content in the *P. phoxinus* genome may result in
392 rearrangements and variations in genome size (Kojima, 2019). Transposable elements can also function
393 as recombination agents, generating structural variations such as deletions, insertions, translocations
394 and inversions (Carvalho & Lupski, 2016; Cerbin & Jiang, 2018; Deininger et al., 2003), which can be
395 sources of standing genetic variation present in an individual.

396

397 **Protein Annotation**

398 Protein prediction, based solely on transcripts derived from our RNA-seq data mapped to the Hap1
399 and Hap2 genome assembly (Supplementary Table S5), generated 40,364 gene models and a mRNA
400 count of 42,408, which covered 47.7% and 53.6% of the genome, respectively. The *ab-initio* prediction
401 resulted in 24,420 genes, covering 49.9% of the genome. After filtering and generating a final high-
402 confidence gene set, the final annotation contained 30,980 mRNAs, of which 23,397 were marked as
403 high-confidence genes in Hap 1, and 29,614 mRNAs, with 23,191 high-confidence genes in Hap 2 (full
404 summary statistics of protein annotation for Hap1 and Hap2 in supplementary Table S6 and S7). Using
405 the predicted gene content of Hap1 of our *P. phoxinus* assembly in comparison to its closest relative
406 with available whole-genome resources, the fathead minnow genome contains more predicted genes
407 (26,150), (Martinson et al., 2022). We ran BUSCO with the actinopterygii_odb10 database and obtained
408 a total of 3,484 (95.7%) complete BUSCOs, 3,435 (94.4%) complete and single-copy BUSCOS, 49 (1.3%)
409 complete and duplicated BUSCOs, 20 (0.5%) fragmented BUSCOs and 136 (3.8%) missing BUSCOs.
410 Structural annotation of the predicted proteins resulted in 22,761 (97.28%) hits in at least 1 database,
411 with 19,871 (84.56%) proteins shared across all four investigated databases (Table 2). These statistics
412 hint at a well-assembled and annotated *P. phoxinus* genome. Homology search of genes in *P. phoxinus*
413 against zebrafish, fathead minnow, goldfish, common carp and grass carp had 21,206 (90.6%) genes
414 annotated across all 5 species (Supplementary Figure S5)

415 **Table 2.** Functional annotation of genes in *P.phoxinus* genome across four different databases: TrEMBL, Swissprot, PDBAA,
416 and eggNOG

Database	Number	Percentage (%)
Swissprot	19,978	85.39
TrEMBL	22,761	97.28
egg-NOG	21,296	91.02
PDBAA	19,978	85.39
Gene overlap	19,871	84.93
Total Annotated genes	22,761	97.28

417

418 **Genome-wide Comparison of Structural and Sequence Variation between Haplomes**

419 *Sequence variation follows universal pattern of eukaryotes*

420 Genome-wide estimates and patterns of mean heterozygosity were similar between both haplomes,
421 with an estimate of 0.7866% in Hap1 and 0.7926% in Hap2. Genome-wide heterozygosity estimates
422 were much lower than k-mer-based estimates and this may be due to the difference between the two
423 approaches, with the genome-wide approach being more accurate as it makes estimates from direct
424 observation of different loci that have been confidently mapped, and the k-mer-based approach being
425 more sensitive to low coverage and error-prone regions. The distribution of heterozygosity across the
426 chromosomes varied, but the pattern found on each chromosome was similar (Figure 3). The regions of
427 highest heterozygosity were always found at both ends of each chromosome and the lowest
428 heterozygosity was found in the central region of each chromosome.

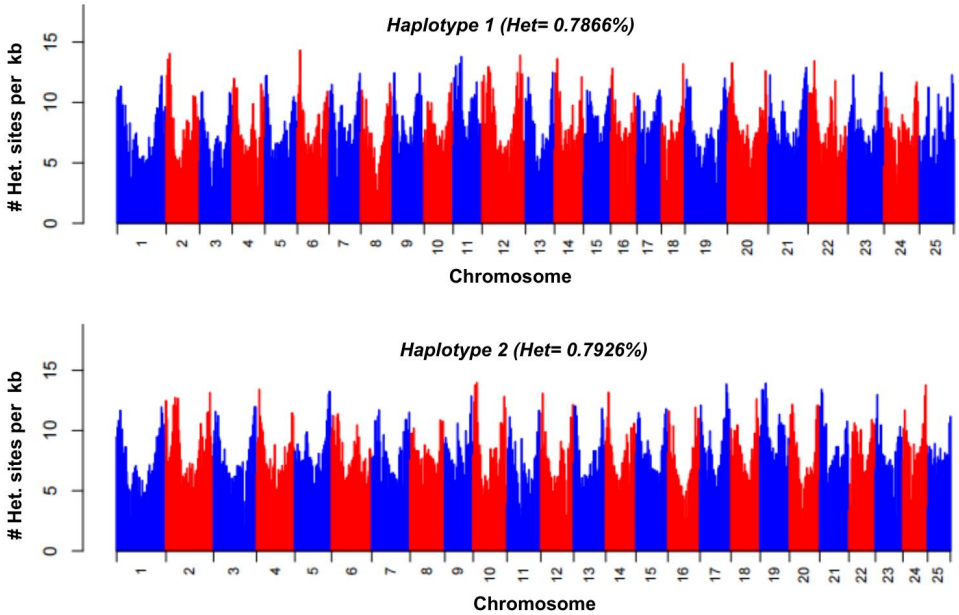


Figure 3: Distribution of heterozygosity along the 25 chromosomes in *P. phoxinus* shown for haplomes 1 (upper plot) and 2 (lower plot). Mean heterozygosity estimates were calculated in 1Mb bins and converted to the number of heterozygous sites per kb across all 25 chromosomes with alternating colours to differentiate adjacent chromosomes.

It is possible that the regions of high heterozygosity are linked to telomeric regions of the chromosome, which generally have an elevated recombination rate (Nachman, 2002), which in turn can be linked to higher genetic diversity (Jaramillo-Correa et al., 2010). In fishes, this was evidenced, e.g., in the Atlantic silverside (*Menidia menidia*) (Tigano et al., 2021) and brown trout (*Salmo trutta)* (Leitwein et al., 2016) genomes. This contrasts with regions of reduced heterozygosity found at central locations potentially linked to centromeres, which in eukaryotes are known as "cold-spots" for recombination, hence have low levels of recombination (Yan & Yu, 2022).

Genomes with high heterozygosity can pose assembly challenges due to multiple scaffold formation from the same sequence (Kajitani et al., 2014; Pryszcz & Gabaldón, 2016). However, implementation of long range information such as Hi-Fidelity long reads and Hi-C contact maps can aid in overcoming these difficulties by reducing sequence redundancy (Kronenberg et al., 2021). Our high-quality assembly, indicating augmented levels of standing genetic variation, is a testament to this approach. Previous studies have also indicated abundant levels of genetic variation in the *Phoxinus* genus (Bogutskaya et al., 2020; Reier et al., 2022), potentially enabling more expeditious responses to environmental changes and even greater potential for adaptation across diverse environments (Lai et al., 2019; Matuszewski et al., 2015).

Demographic History of *P. phoxinus*

Climatic fluctuations have always had an effect on the distribution and survival of species (Hewitt, 2000; J. Li et al., 2021). For aquatic species in the temperate zone, effects of climate fluctuations are even more pronounced as a result of glaciation cycles, which led to fluctuating sea water levels (Munnecke et al., 2010). A PSMC analysis enabled us to track fluctuations in effective population size *(Ne)* of *P. phoxinus* from more than 10 Million years ago (Mya) until around 10 thousand years ago (kya).

Figure 4 shows a gradual and continual rise in *Ne* with time, from roughly 20,000 individuals about 10 million years ago to around 25,000 individuals approximately 800,000 years ago. This incremental increase coincides with the Neogene period, which is divided into the Miocene and the Pliocene epochs

- both known for their global warmth and minimum ice volumes (Böhme et al., 2011; Haywood et al., 2000; Hui et al., 2018; Steinthorsdottir et al., 2021). The climate conditions prevailing during this timeframe suggest the existence of an environment favourable for the propagation and development of freshwater species. Following a small increase in *Ne*, the population declined to fewer than 20,000 individuals between 800 kya and 100 kya, during the mid-Pleistocene period, which was marked by severe quaternary glaciation and low global sea levels. This period also saw changes in global ice volume, sea levels, and temperatures (Ehlers et al., 2006; Ruddiman et al., 1989; Y. Sun et al., 2019). During this epoch, temperatures were on average 11°C lower than current values (Hughes et al., 2007) and lowland glaciation periods were frequent (Rose, 2009). As a result, species either became extinct, migrated to southern Europe seeking refuge, or underwent adaptation (Feliner, 2011).

This reduction in *Ne* during the Last Glacial Maximum is generally observed in fish (Jardim de Queiroz et al., 2022; J. Li et al., 2021) and in birds (Nadachowska-Brzyska et al., 2015), especially in European species (Yuan et al., 2018). In *P. phoxinus*, the end of Last Glacial Maximum was followed by a very sharp increase in *Ne* from around 70 kya, peaking at nearly a million estimated individuals, indicating that *P. phoxinus* was able to repopulate previously covered freshwater habitats across Europe. The high *Ne* also corroborates the high heterozygosity estimates for this genome.

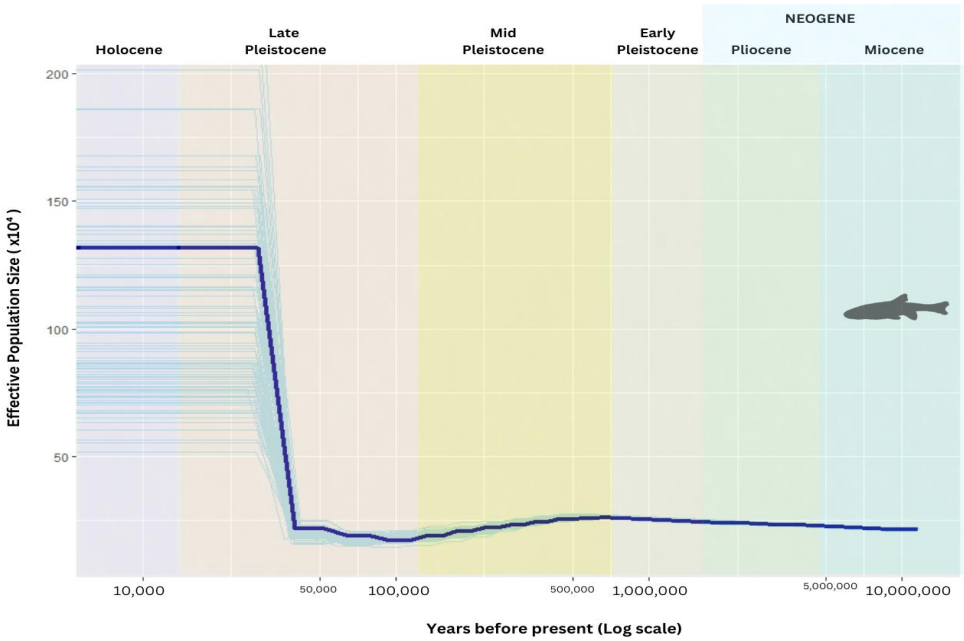


Figure 4. Inferred demographic history of *P. phoxinus* from PSMC analysis. The time ranges cover the Miocene, Pliocene, Early to Late Pleistocene and Holocene, shown in different colours. The x-axis shows years before present on a logarithmic scale and the y-axis shows the estimated effective population size. Bootstrap results are shown as transparent lines. Haplome 2 showed a similar curve.

Large non-syntenic regions between P. phoxinus haplomes

A comparison of both haplomes of *P. phoxinus* with SyRI revealed around 815 to 820 Mb of syntenic regions between both haplomes (Table 3, Figure 5). Our SyRI analysis also identified 1,572 translocations of between 5.34 Mb and 5.36 Mb in both haplomes; 214 duplications totalling 571 kb in Hap1 and 667 duplications of around 1.8 Mb total length in Hap2; 215 inversions with a total length of around 10.8 Mb in Hap1 and 11.1 Mb in Hap2 (Table 3, Figure 5).

487

488 **Table 3.** Identity**,** number, and total lengths of identified structural variations between haplomes 1 and
489 2 in the *P. phoxinus* reference genome.

Structural Variation	Count	Length haplotype 1 [bases]	Length haplotype 2 [bases]
Syntenic regions	1457	813,594,094	820,681,234
Inversions	215	10,843,815	11,161,414
Translocations	1572	5,342,064	5,361,801
Duplications (Haplotype 1)	214	57,1045	-
Duplications (Haplotype 2)	667	-	1,776,709
Insertions	447,578	-	8,108,033
Deletions	447,699	8,113,615	-

490

491 Inversions represented the dominant structural variation within the *P. phoxinus* genome in terms of
492 total size in either haplome (Table 3). A total of 215 inversions were identified, with the largest eight
493 inversions making up 72% (8 Mb) of the estimated 11 Mb of inversions present in the *P. phoxinus*
494 genome. Using Hap1 as a reference for inversion size, chromosome 6 contained the largest inversion
495 with a size of 1.8 Mb; closely followed by a 1.7 Mb inversion on chromosome 2; a 1.5 Mb inversion on
496 chromosome 21; a 1.1 Mb inversion on chromosomes 9 and 19, each; and finally an 800 kb inversion on
497 chromosome 3 (Figure 7).

498 We investigated genes enriched in regions of copy number structural variants, specifically deletions
499 and insertions, which can also be characterised as presence-absence variations (PAVs). GO terms
500 enriched in this region of PAVs between both haplotypes include terms such as “dephosphorylation”,
501 “peptidyl-serine phosphorylation”, and “protein processing”, which have been linked with the activation
502 of sperm motility in numerous fish species (Dumorné et al., 2018; Zilli et al., 2008, 2017). Some GO
503 terms related to development were also enriched within PAVs, such as “regulation of developmental
504 growth” and “developmental pigmentation” (Supplementary Tables S8-S9 and Figure S6).

505 Genes enriched within locational structural variants, in this instance inversions, were primarily
506 associated with immunity and transposable elements in hAT, DDE, and L1 families . These GO terms are
507 “phagocytosis”, “phagosome maturation involved in apoptotic cell clearance” and “phagolysosome
508 assembly involved in apoptotic cell clearance”. Phagocytosis is a conserved immune defence mechanism
509 across multicellular organisms that prevents infection and invasion of the body by pathogens (Esteban
510 et al., 2015; Neumann et al., 2001) (Supplementary Tables S10-S11 and Figure S7).

511 Our haplotype-resolved genome allows for the identification of structural variants, such as inversions,
512 translocations, duplications, deletions, and insertions, between both haplomes of our reference genome.
513 To explore intra-individual structural variation, haplotype-resolved reference genomes have been
514 created in various organisms, particularly in plant species (S.-P. Cheng et al., 2021; L. Jiang et al., 2022; H.
515 Sun et al., 2022). It has also been utilised to a lesser degree to uncover structural variation in animals,
516 for example, oysters (Takeuchi et al., 2022), where transposable elements played a large role in the
517 structural variation between both haplomes, Additionally, in turkeys, structural variations have affected
518 genes involved in disease susceptibility and fertility (Barros et al., 2023).

519

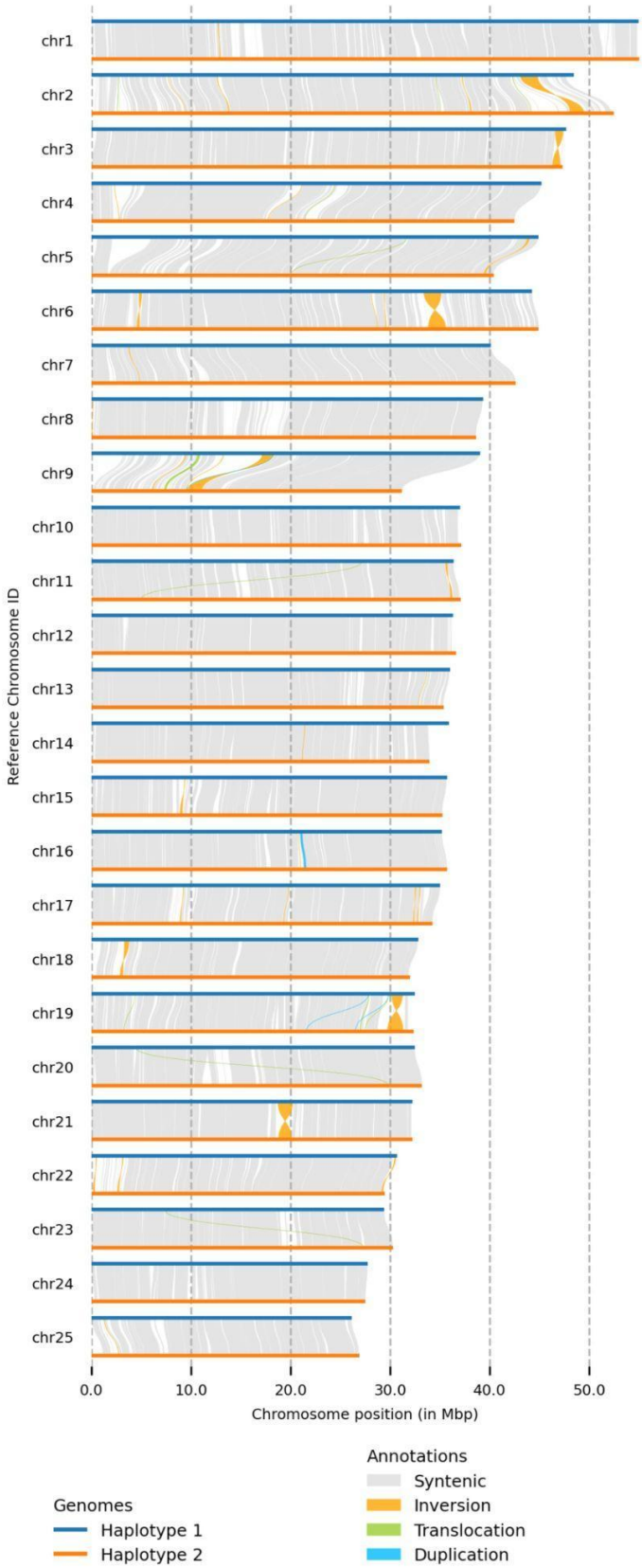


Figure 5: **Synteny plot of all 25 chromosomes of Hap1 (top, blue line) and Hap2 (bottom, orange line) of the *P. phoxinus* genome assembly**. Aligned regions between both haplomes are shown in blocks of grey, unaligned regions are shown by gaps of white. Inversions are represented in orange, translocations in green, duplications in light-blue.

Haplotype-resolved genome assembly revealed high level of within-genome variation in P. phoxinus

Analysis of mean genome wide heterozygosity and structural variation, as outlined in the preceding sections, revealed a high level of variation present in the *P. phoxinus* haplomes. This is a possible reflection of high fecundity and large population size, expected in most freshwater fish species (Karjalainen et al., 2016; Röpke et al., 2021). The detection of genetic diversity in highly heterozygous species can be compromised by either un-phased genome-assembly approaches or haplotype-merged assemblies, which would ultimately lead to a significant underestimation of true genetic diversity (Nakandala et al., 2023; Takeuchi et al., 2022). Collapsed reference genomes restrict assembly accuracy and exploration of heterozygous regions of the genome, missing out on potentially interesting intra-individual variation (Low et al., 2020).

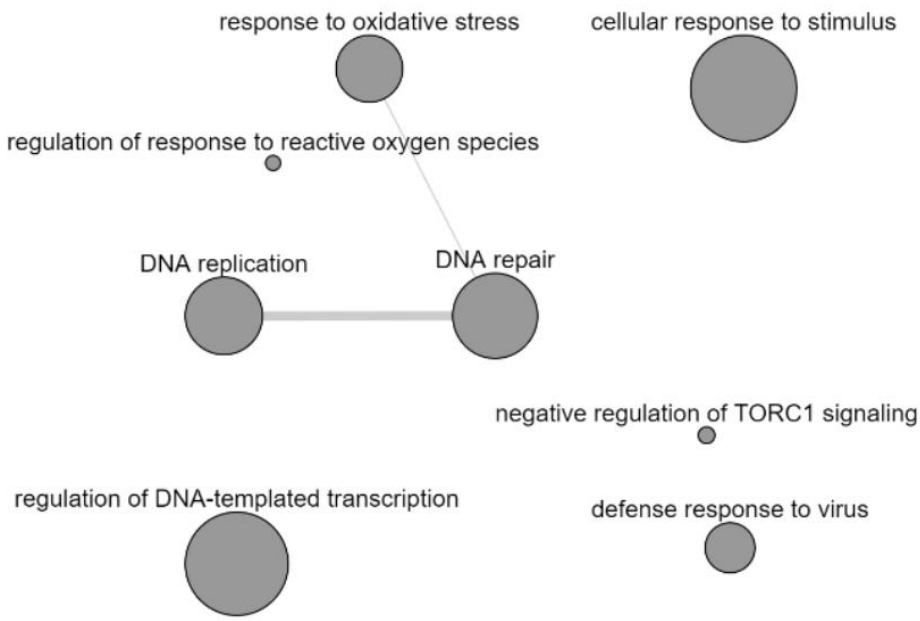
Gene Family Evolution - Histone Gene Families Retracted, NLRC Genes Expanded in *P. phoxinus*

Inferring gene orthology from sequence data is a fundamental approach for understanding evolutionary relationships between genes and species through their evolutionary history (Tekaia, 2016). A total of 396,455 genes across all 11 species described in the protein annotation section were included and 381,263 (96.12%) of these genes were successfully assigned to 27,672 orthogroups, i,e. gene-sets in those species or in a subset of these species, that derived from their respective last common ancestor (Supplementary Table S12). The total number of orthogroups comprising proteins from all species was 14,914 (Table 4). Single-copy orthogroups, also known as 1-to-1 orthogroups, contain genes that have evolved from a common ancestral gene but are present in a single copy across all proteomes used in an ortholog analysis (Creevey et al., 2011; Simão et al., 2015; Tekaia, 2016). Our orthology inference analysis revealed the presence of 301 single-copy orthogroups which provided a solid basis for phylogenetic studies.

Table 4. Summary statistics of orthology inference analysis of *P. phoxinus* and 10 other teleost species.

Overall Summary Statistic	
Number of species	11
Number of genes	396,455
Number of genes in orthogroups	381,263
Number of unassigned genes	15,192
Percentage of genes in orthogroups	96.2
Percentage of unassigned genes	3.8
Number of orthogroups	27,672
Number of species-specific orthogroups	2319
Number of genes in species-specific orthogroups	7975
Percentage of genes in species-specific orthogroups	2.0
Mean orthogroup size	13.8
Median orthogroup size	15.0
Number of single-copy orthogroups	301
Number of orthogroups with all species present	14,914

549 There were 273 genes distributed across 72 *P. phoxinus*-specific orthogroups; a homology blast of
550 these genes in the TrEMBL database returned only results for 130 genes, many of which were
551 uncharacterised proteins (Supplementary Table S13). Only 54 of these genes annotated from the
552 TrEMBL database returned GO annotations and these annotations were clustered into eight major GO
553 terms (Figure 6). The presence of GO terms "response to oxidative stress" and "regulation of response
554 to reactive oxygen", suggests that *P. phoxinus* is particularly sensitive to changes in water quality.
555 Observational studies have found conflicting evidence, with Eurasian minnows ranging from sensitive
556 (Bagge & Hakkari, 1992) to insensitive to intermediate pollution (King et al., 2011). Those findings may
557 be attributed to the fact that different species of *Phoxinus* were studied. Cyprinoidea in general have
558 been thought to be sensitive to oxidative stress, demonstrated in Amur minnow (*Phoxinus lagowskii*) (Q.
559 Wang et al., 2019; Yao et al., 2022) and Chinese rare minnow (*Gobiocypris rarus*) (Z.-H. Li et al., 2014),
560 while the fathead minnow (*Pimephales promelas*) has been used as a bioindicator of water quality
561 (Elizalde-Velázquez et al., 2020).



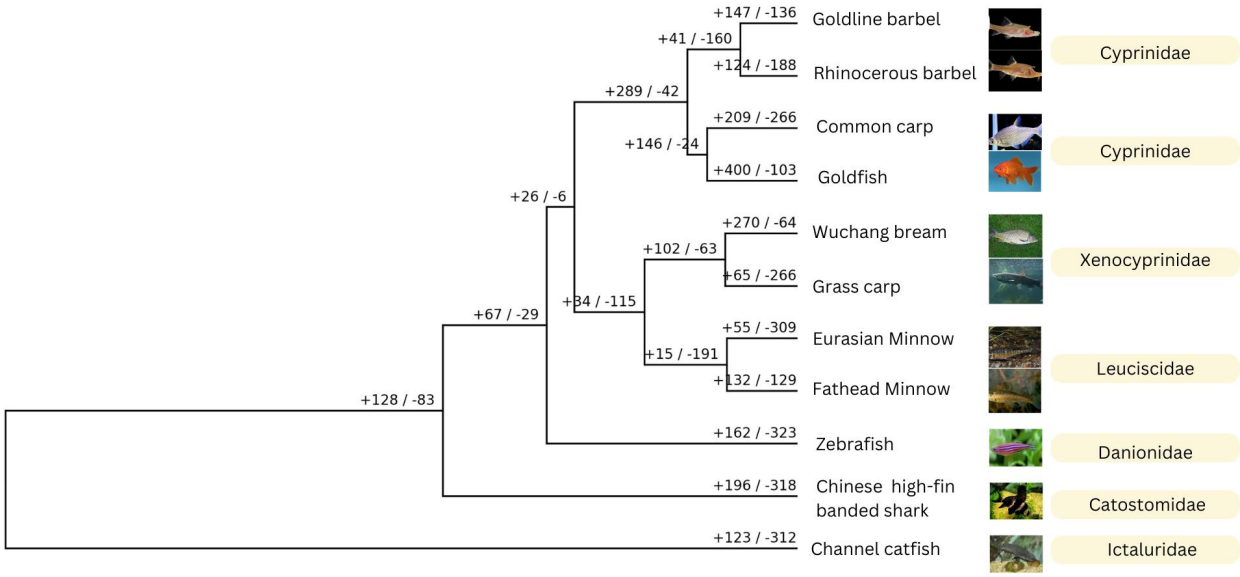
562

563 **Figure 6: Gene ontology** network of 54 functionally annotated *P. phoxinus* specific genes into eight major GO terms: "DNA
564 replication", "DNA repair", "regulation of DNA-templated transcription", "response to oxidative stress", "defense response to
565 virus", "cellular response to stimulus", "regulation of response to reactive oxygen species" and "negative regulation of TORC1
566 signalling". Bubble size indicates the frequency of the GO term in the gene ontology annotation database i.e larger bubbles
567 correspond to more general terms. Edges on the graph connect highly similar or linked GO terms to each other.

568 Gene family changes are expressed as duplication and loss of genes among related species and are
569 based on mutation, fixation, and retention rates of the genes in question (Demuth & Hahn, 2009). Gene
570 family expansion might be an indication of positive selection (Demuth & Hahn, 2009). An example is the
571 expansion of Myxovirus resistance (Mx) genes in the rainbow trout, providing special antiviral resistance
572 for their anadromous lifestyle (T. Wang et al., 2019). While the immune system is relatively conserved
573 across different jawed vertebrates (Flajnik & Kasahara, 2010; Litman et al., 2010), a wide array of
574 species-specific modifications in the underlying immune pathway were observed in other fishes, such as
575 cods, coelacanths, and sharks (Buonocore & Gerdol, 2016; Malmstrøm et al., 2016; Stein et al., 2007).

576 We analysed gene family expansions and contractions in *P. phoxinus* relative to 10 other teleost
577 species (Supplementary Figure S8). Since the split of *P. phoxinus* and the fathead minnow (Figure 7)

578 from their last common ancestor, *P. phoxinus* counted more significant gene contractions (309 gene
579 families) than significant gene expansions (55 gene families) relative to the fathead minnow, which had
580 132 gene family expansions and 129 gene family contractions. The significantly expanded gene families
581 contained genes such as *HIST1, H2B*, *CENP-T*, *GCRV*, *HIST2H2A*, *HIST2H3A*, *H3F3A*, *TRBV25-1*, which
582 mostly belong to histone PFAM domains (Supplementary Figure S9). Immune genes belonging to the
583 NACHT PFAM domain, were also more abundant in the set of contracted gene families (Supplementary
584 Figure S10, genes like *NLRP12, NLRP3, CD22, CLEC17A, GIMAP7*. The expanding immune gene families
585 mostly belonged to histone gene families, whereas the contracted gene families were mostly
586 represented by NLRC gene families (Supplementary Tables S14 and S15).



587

588 **Figure 7**: **Ultrametric phylogenetic tree of selected teleost species and *P. phoxinus*.** Numbers on the branches represent
589 counts for expanded (+) and contracted (-) gene families. Positive numbers indicate significant gene families that have been
590 expanded since the split from the last common ancestor and negative numbers indicate significant gene families that have
591 been contracted since the split from the last common ancestor.

592 Evolution of the immune system has greatly contributed to speciation in teleosts and a lot of
593 diversity in immune system pathways has been identified in teleosts, like the expanded MHC1 and TLE
594 genes in the Atlantic cod (*Gadus morhua*) (Star et al., 2011) and the absence of MHC class II-mediated
595 immunity in the broad-nosed pipefish (*Syngnathus typhle)* (Haase et al., 2013). In the *P. phoxinus*
596 genome*,* we observed an expansion in immunity-boosting histones *H1, H2B, H3*, and *H4,* particularly
597 *H2A* and *H2B.* Histones are highly conserved proteins involved in the innate host defence against
598 microbes (X. Li et al., 2022), they include core histones like *H2A, H2B, H3*, and *H4*. The final histone type,
599 called *H1,* is generally referred to as a linker histone (X. Li et al., 2022; Westman et al., 2015). Histones
600 are known to play a key role in antibacterial immunity across multiple species (X. Li et al., 2022;
601 Pasupuleti et al., 2012). Histone 2A and 2B have antibacterial activity against different types of bacteria
602 in fish. Histones are particularly implicated in fish mucosal immunity (Ángeles Esteban & Cerezuela,
603 2015), acting as a strong barrier to microbial invasion. This has been confirmed in the Indian major carp
604 (*Labeo catla*) (Nigam et al., 2017), lumpfish (*Cyclopterus lumpus*) (Patel & Brinchmann, 2017), gilthead
605 seabream (*Sparus aurata*) (Jurado et al., 2015), and rainbow trout (*Oncorhynchus mykiss*) (Fernandes et
606 al., 2002; Oscoz et al., 2005). The expansion of histone genes in *P. phoxinus* suggests a strong focus on
607 mucosal protection against bacteria and other pathogens. Assuming, that similar evolutionary trends
608 can be found in other *Phoxinus* species, this is a particularly interesting find, given that there are

populations of *Phoxinus* spp. in the Baltic Sea, some of which are anadromous and some of which are entirely resident in the brackish water (Svirgsden et al., 2018). Such migration habits demand an efficient immune system to cope with the different pathogens and parasite communities present in brackish and freshwater.

In contrast, NLRC genes were found to be reduced in the *P. phoxinus* genome, particularly *NLRP3* and *NLRP12*, which are important inflammatory genes (Morimoto et al., 2021). NLRC genes are majorly involved in immune response to exogenous pathogens like bacterial and parasitic infections (Chang, 2023; Morimoto et al., 2021). NLRC genes are involved in immune response in Nile tilapia (*Oreochromis niloticus*) in *Streptococcus agalactiae* and *Aeromonas hydrophila* infections (Q. Li et al., 2022). NLRC genes were also overexpressed in response to *V. anguillarum* infection in the Mi-iuy croaker (*Miichthys miiuy*) (G. Liu et al., 2023). In summary, the increased presence of genes involved in preventive mucosal defence and decreased amount of genes involved in immune responsiveness, might hint to a change in immune response strategy towards infection prevention.

Conclusion

In this paper, we present a highly contiguous and complete Eurasian minnow, *Phoxinus phoxinus,* genome achieved by combining Pacbio Hifi reads and Hi-C data, along with comprehensive repeats and gene annotation. Demographic history analysis and high observed heterozygosity seem to suggest substantial population and individual diversity, respectively. While the identified diversity could pose a challenge for a collapsed assembly, our phased genome enables a more accurate depiction of intra-individual variation. Consequently, we anticipate this genome will prove an invaluable resource for forthcoming research investigating the intricate taxonomic and demographic complexities of the *Phoxinus* genus. For future studies, it will be of great importance to determine the genetic identity of the *Phoxinus* spp. studied before associating natural history traits with the species name. Similarly, a comprehensive understanding of the behaviour of Eurasian Minnows requires tracing the possible genetic lineages in historical literature, in order to obtain a clearer picture.

Acknowledgments

We thank the Long Read Team of the DRESDEN Concept Genome Center, part of the MPI-CBG and the technology platform of the CMCB at the TU Dresden, supported by DFG (INST 269/768-1).

Author Contributions

MS devised the project. TOO and MS lead the writing and devising of the manuscript. EWM, SW, and TB performed genome sequencing and assembly. IC established the genome repeat and protein annotation pipeline with support from SM. TOO performed genome-wide comparison, gene family evolution and demographic history analyses. SM and AB provided TOO and IC with bioinformatic support. AB contributed with sex-determination and genome assembly expertise. All authors contributed to revising the manuscript.

Data, scripts, code, and supplementary information availability

Scripts used for genome assembly by EMW, SW and TB, scripts used for genome-wide comparison, gene family evolution and demographic history by TOO are available here: https://doi.org/10.5281/zenodo.10210241.

The repeat and protein annotation was implemented using a snakemake pipeline. The code is available here: https://doi.org/10.5281/zenodo.10220349. RNA data for annotation is available under NCBI accession numbers SRR26699630 to SRR26699636. *P. phoxinus* haplomes are available under NCBI Bioproject numbers PRJNA1030283 and PRJNA1030284 both of which are linked to the Euro-fish project under umbrella bioproject number PRJNA768423.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest about the content of the article.

Funding

This research was funded by the Leibniz Association under project SAW-J96/2020 awarded to MS, and by the German Research Foundation under INST 269/768-1 awarded to the technology platform of the CMCB at the TU Dresden.

662

663

664 **References**

665 Abdennur, N., & Mirny, L. A. (2020). Cooler: Scalable storage for Hi-C data and other genomically labeled
666 arrays. *Bioinformatics*, *36*(1), 311–316. https://doi.org/10.1093/bioinformatics/btz540

667 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool.
668 *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

669 Ángeles Esteban, M., & Cerezuela, R. (2015). 4 - Fish mucosal immunity: Skin. In B. H. Beck & E. Peatman
670 (Eds.), *Mucosal Health in Aquaculture* (pp. 67–92). Academic Press.
671 https://doi.org/10.1016/B978-0-12-417186-2.00004-2

672 Bagge, P., & Hakkari, L. (1992). Effects of paper mill effluents on the fish fauna of stony shores of Lake
673 Päijänne. In V. Ilmavirta & R. I. Jones (Eds.), *The Dynamics and Use of Lacustrine Ecosystems* (pp.
674 413–420). Springer Netherlands. https://doi.org/10.1007/978-94-011-2745-5_43

675 Baid, G., Cook, D. E., Shafin, K., Yun, T., Llinares-López, F., Berthet, Q., Belyaeva, A., Töpfer, A., Wenger,
676 A. M., Rowell, W. J., Yang, H., Kolesnikov, A., Ammar, W., Vert, J.-P., Vaswani, A., McLean, C. Y.,
677 Nattestad, M., Chang, P.-C., & Carroll, A. (2023). DeepConsensus improves the accuracy of
678 sequences with a gap-aware sequence transformer. *Nature Biotechnology*, *41*(2), Article 2.
679 https://doi.org/10.1038/s41587-022-01435-7

680 Barros, C. P., Derks, M. F. L., Mohr, J., Wood, B. J., Crooijmans, R. P. M. A., Megens, H.-J., Bink, M. C. A.
681 M., & Groenen, M. A. M. (2023). A new haplotype-resolved turkey genome to enable turkey
682 genetics and genomics research. *GigaScience*, *12*, giad051.
683 https://doi.org/10.1093/gigascience/giad051

684 Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi–C: A
685 comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268–276.
686 https://doi.org/10.1016/j.ymeth.2012.05.001

687 Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*,

27(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, *31*(1), 365–370. https://doi.org/10.1093/nar/gkg095

Bogutskaya, N. G., Diripasko, O. A., & Palandačić, A. (2023). Novel data support validity of Phoxinus chrysoprasius (Pallas, 1814) (Actinopterygii, Leuciscidae). *European Journal of Taxonomy*, *861*, 1–20. https://doi.org/10.5852/ejt.2023.861.2061

Bogutskaya, N. G., Jelić, D., Vucić, M., Jelić, M., Diripasko, O. A., Stefanov, T., & Klobučar, G. (2020). Description of a new species of Phoxinus from the upper Krka river (Adriatic Basin) in Croatia (Actinopterygii: Leuciscidae), first discovered as a molecular clade. *Journal of Fish Biology*, *96*(2), 378–393. https://doi.org/10.1111/jfb.14210

Böhme, M., Winklhofer, M., & Ilg, A. (2011). Miocene precipitation in Europe: Temporal trends and spatial gradients. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *304*(3), 212–218. https://doi.org/10.1016/j.palaeo.2010.09.028

Böhne, A., Zhou, Q., Darras, A., Schmidt, C., Schartl, M., Galiana-Arnoux, D., & Volff, J.-N. (2012). Zisupton—A Novel Superfamily of DNA Transposable Elements Recently Active in Fish. *Molecular Biology and Evolution*, *29*(2), 631–645. https://doi.org/10.1093/molbev/msr208

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108. https://doi.org/10.1093/nargab/lqaa108

Brůna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, *2*(2), lqaa026.

https://doi.org/10.1093/nargab/lqaa026

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), Article 4. https://doi.org/10.1038/s41592-021-01101-x

Buonocore, F., & Gerdol, M. (2016). Alternative adaptive immunity strategies: Coelacanth, cod and shark immunity. *Molecular Immunology*, *69*, 157–169. https://doi.org/10.1016/j.molimm.2015.09.003

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Carley, L. N., Morris, W. F., Walsh, R., Riebe, D., & Mitchell-Olds, T. (2022). Are genetic variation and demographic performance linked? *Evolutionary Applications*, *15*(11), 1888–1906. https://doi.org/10.1111/eva.13487

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, *17*(4), Article 4. https://doi.org/10.1038/nrg.2015.25

Cerbin, S., & Jiang, N. (2018). Duplication of host genes by transposable elements. *Current Opinion in Genetics & Development*, *49*, 63–69. https://doi.org/10.1016/j.gde.2018.03.005

Chang, M. X. (2023). Emerging mechanisms and functions of inflammasome complexes in teleost fish. *Frontiers in Immunology*, *14*. https://www.frontiersin.org/articles/10.3389/fimmu.2023.1065181

Chen, L., Li, B., Chen, B., Li, C., Zhou, Z., Zhou, T., Yang, W., & Xu, P. (2021). Chromosome-level genome of Poropuntius huangchuchieni provides a diploid progenitor-like reference genome for the allotetraploid Cyprinus carpio. *Molecular Ecology Resources*, *21*(5), 1658–1669. https://doi.org/10.1111/1755-0998.13365

Chen, Z., Omori, Y., Koren, S., Shirokiya, T., Kuroda, T., Miyamoto, A., Wada, H., Fujiyama, A., Toyoda, A., Zhang, S., Wolfsberg, T. G., Kawakami, K., Phillippy, A. M., NISC Comparative Sequencing Program, Mullikin, J. C., & Burgess, S. M. (2019). De novo assembly of the goldfish (Carassius

740 auratus) genome and the evolution of genes after whole-genome duplication. *Science Advances*,
741 *5*(6), eaav0547. https://doi.org/10.1126/sciadv.aav0547
742 Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly
743 using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), Article 2.
744 https://doi.org/10.1038/s41592-020-01056-5
745 Cheng, S.-P., Jia, K.-H., Liu, H., Zhang, R.-G., Li, Z.-C., Zhou, S.-S., Shi, T.-L., Ma, A.-C., Yu, C.-W., Gao, C.,
746 Cao, G.-L., Zhao, W., Nie, S., Guo, J.-F., Jiao, S.-Q., Tian, X.-C., Yan, X.-M., Bao, Y.-T., Yun, Q.-Z., …
747 Mao, J.-F. (2021). Haplotype-resolved genome assembly and allele-specific gene expression in
748 cultivated ginger. *Horticulture Research*, *8*, 188. https://doi.org/10.1038/s41438-021-00599-8
749 Corral-Lou, A., Perea, S., Aparicio, E., & Doadrio, I. (2019). Phylogeography and species delineation of
750 the genus Phoxinus Rafinesque, 1820 (Actinopterygii: Leuciscidae) in the Iberian Peninsula.
751 *Journal of Zoological Systematics and Evolutionary Research*, *57*(4), 926–941.
752 https://doi.org/10.1111/jzs.12320
753 Creevey, C. J., Muller, J., Doerks, T., Thompson, J. D., Arendt, D., & Bork, P. (2011). Identifying Single
754 Copy Orthologs in Metazoa. *PLOS Computational Biology*, *7*(12), e1002269.
755 https://doi.org/10.1371/journal.pcbi.1002269
756 Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T.,
757 McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools.
758 *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008
759 Deininger, P. L., Moran, J. V., Batzer, M. A., & Kazazian, H. H. (2003). Mobile elements and mammalian
760 genome evolution. *Current Opinion in Genetics & Development*, *13*(6), 651–658.
761 https://doi.org/10.1016/j.gde.2003.10.013
762 Demuth, J. P., & Hahn, M. W. (2009). The life and death of gene families. *BioEssays*, *31*(1), 29–39.
763 https://doi.org/10.1002/bies.080085
764 Denys, G. P. J., Dettai, A., Persat, H., Daszkiewicz, P., Hautecœur, M., & Keith, P. (2020). *Revision of*
765 *Phoxinus in France with the description of two new species (Teleostei, Leuciscidae)*.

https://doi.org/10.26028/CYBIUM/2020-443-003

DeWoody, J. A., Harder, A. M., Mathur, S., & Willoughby, J. R. (2021). The long-standing significance of genetic diversity in conservation. *Molecular Ecology*, *30*(17), 4147–4154. https://doi.org/10.1111/mec.16051

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dumorné, K., Figueroa, E., Cosson, J., Lee-Estevez, M., Ulloa-Rodríguez, P., Valdebenito, I., & Farías, J. G. (2018). Protein phosphorylation and ions effects on salmonid sperm motility activation. *Reviews in Aquaculture*, *10*(3), 727–737. https://doi.org/10.1111/raq.12198

Ehlers, J., Astakhov, V., Gibbard, P., Mangerud, J., & Svendsen, J. (2006). Late Pleistocene Glaciations in Europe. *Encyclopedia of Quaternary Science*, 1085–1095.

Elizalde-Velázquez, A., Crago, J., Zhao, X., Green, M. J., & Cañas-Carrell, J. E. (2020). In vivo effects on the immune function of fathead minnow (Pimephales promelas) following ingestion and intraperitoneal injection of polystyrene nanoplastics. *Science of The Total Environment*, *735*, 139461. https://doi.org/10.1016/j.scitotenv.2020.139461

Emms, D. M., & Kelly, S. (2017). STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution*, *34*(12), 3267–3278. https://doi.org/10.1093/molbev/msx259

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Esteban, M. Á., Cuesta, A., Chaves-Pozo, E., & Meseguer, J. (2015). Phagocytosis in Teleosts. Implications of the New Cells Involved. *Biology*, *4*(4), Article 4. https://doi.org/10.3390/biology4040907

Feliner, G. N. (2011). Southern European glacial refugia: A tale of tales. *TAXON*, *60*(2), 365–372. https://doi.org/10.1002/tax.602007

FERNANDES, J. M. O., KEMP, G. D., MOLLE, M. G., & SMITH, V. J. (2002). Anti-microbial properties of histone H2A from skin secretions of rainbow trout, Oncorhynchus mykiss. *Biochemical Journal*,

368(2), 611–620. https://doi.org/10.1042/bj20020980

Flajnik, M. F., & Kasahara, M. (2010). Origin and evolution of the adaptive immune system: Genetic events and selective pressures. *Nature Reviews Genetics*, *11*(1), Article 1. https://doi.org/10.1038/nrg2703

Frost, W. E. (1943). The Natural History of the Minnow, Phoxinus phoxinus. *Journal of Animal Ecology*, *12*(2), 139–162. https://doi.org/10.2307/1374

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2023). *BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA* (p. 2023.06.10.544449). bioRxiv. https://doi.org/10.1101/2023.06.10.544449

Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: Transcript selector for BRAKER. *BMC Bioinformatics*, *22*(1), 566. https://doi.org/10.1186/s12859-021-04482-0

Gao, B., Shen, D., Xue, S., Chen, C., Cui, H., & Song, C. (2016). The contribution of transposable elements to size variations between four teleost genomes. *Mobile DNA*, *7*(1), 4. https://doi.org/10.1186/s13100-016-0059-7

Goel, M., & Schneeberger, K. (2022). plotsr: Visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*, *38*(10), 2922–2926. https://doi.org/10.1093/bioinformatics/btac196

Goel, M., Sun, H., Jiao, W.-B., & Schneeberger, K. (2019). SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, *20*(1), 277. https://doi.org/10.1186/s13059-019-1911-0

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Guevara, E. E., Webster, T. H., Lawler, R. R., Bradley, B. J., Greene, L. K., Ranaivonasy, J., Ratsirarson, J., Harris, R. A., Liu, Y., Murali, S., Raveendran, M., Hughes, D. S. T., Muzny, D. M., Yoder, A. D.,

Worley, K. C., & Rogers, J. (2021). Comparative genomic analysis of sifakas (Propithecus) reveals selection for folivory and high heterozygosity despite endangered status. *Science Advances*, *7*(17), eabd2274. https://doi.org/10.1126/sciadv.abd2274

Haase, D., Roth, O., Kalbe, M., Schmiedeskamp, G., Scharsack, J. P., Rosenstiel, P., & Reusch, T. B. H. (2013). Absence of major histocompatibility complex class II mediated immunity in pipefish, Syngnathus typhle: Evidence from deep transcriptome sequencing. *Biology Letters*, *9*(2), 20130044. https://doi.org/10.1098/rsbl.2013.0044

Haywood, A. M., Sellwood, B. W., & Valdes, P. J. (2000). Regional warming: Pliocene (3 Ma) paleoclimate of Europe and the Mediterranean. *Geology*, *28*(12), 1063–1066. https://doi.org/10.1130/0091-7613(2000)28<1063:RWPMPO>2.0.CO;2

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, *405*(6789), Article 6789. https://doi.org/10.1038/35016000

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, *32*(5), 767–769. https://doi.org/10.1093/bioinformatics/btv661

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-Genome Annotation with BRAKER. In M. Kollmar (Ed.), *Gene Prediction: Methods and Protocols* (pp. 65–95). Springer. https://doi.org/10.1007/978-1-4939-9173-0_5

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G.-J., White, S., … Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), Article 7446. https://doi.org/10.1038/nature12111

Huang, H.-R., Liu, X., Arshad, R., Wang, X., Li, W.-M., Zhou, Y., & Ge, X.-J. (2023). Telomere-to-telomere haplotype-resolved reference genome reveals subgenome divergence and disease resistance in triploid Cavendish banana. *Horticulture Research*, uhad153.

https://doi.org/10.1093/hr/uhad153

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89. https://doi.org/10.1093/nar/gkv1272

Hughes, P. D., Woodward, J. C., & Gibbard, P. L. (2007). Middle Pleistocene cold stage climates in the Mediterranean: New evidence from the glacial record. *Earth and Planetary Science Letters*, *253*(1), 50–56. https://doi.org/10.1016/j.epsl.2006.10.019

Hui, Z., Zhang, J., Ma, Z., Li, X., Peng, T., Li, J., & Wang, B. (2018). Global warming and rainfall: Lessons from an analysis of Mid-Miocene climate data. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *512*, 106–117. https://doi.org/10.1016/j.palaeo.2018.10.025

Jansen, W., Kappus, B., Böhmer, J., & Beiter, T. (1999). Fish communities and migrations in the vicinity of fishways in a regulated river (Enz, Baden-Württemberg, Germany). *Limnologica*, *29*(4), 425–435. https://doi.org/10.1016/S0075-9511(99)80050-4

Jaramillo-Correa, J. P., Verdú, M., & González-Martínez, S. C. (2010). The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. *BMC Evolutionary Biology*, *10*(1), 22. https://doi.org/10.1186/1471-2148-10-22

Jardim de Queiroz, L., Doenz, C. J., Altermatt, F., Alther, R., Borko, Š., Brodersen, J., Gossner, M. M., Graham, C., Matthews, B., McFadden, I. R., Pellissier, L., Schmitt, T., Selz, O. M., Villalba, S., Rüber, L., Zimmermann, N. E., & Seehausen, O. (2022). Climate, immigration and speciation shape terrestrial and aquatic biodiversity in the European Alps. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1980), 20221020. https://doi.org/10.1098/rspb.2022.1020

Jiang, H., Zhang, X., Leng, L., Gong, D., Zhang, X., Liu, J., Peng, D., Wu, Z., & Yang, Y. (2023). A chromosome-scale and haplotype-resolved genome assembly of carnation (Dianthus caryophyllus) based on high-fidelity sequencing. *Frontiers in Plant Science*, *14*, 1230836. https://doi.org/10.3389/fpls.2023.1230836

Jiang, L., Lin, M., Wang, H., Song, H., Zhang, L., Huang, Q., Chen, R., Song, C., Li, G., & Cao, Y. (2022).

Haplotype-resolved genome assembly of Bletilla striata (Thunb.) Reichb.f. To elucidate medicinal value. *The Plant Journal*, *111*(5), 1340–1353. https://doi.org/10.1111/tpj.15892

Jones, P. E., Champneys, T., Vevers, J., Börger, L., Svendsen, J. C., Consuegra, S., Jones, J., & Garcia de Leaniz, C. (2021). Selective effects of small barriers on river-resident fish. *Journal of Applied Ecology*, *58*(7), 1487–1498. https://doi.org/10.1111/1365-2664.13875

Jurado, J., Fuentes-Almagro, C. A., Guardiola, F. A., Cuesta, A., Esteban, M. Á., & Prieto-Álamo, M.-J. (2015). Proteomic profile of the skin mucus of farmed gilthead seabream (Sparus aurata). *Journal of Proteomics*, *120*, 21–34. https://doi.org/10.1016/j.jprot.2015.02.019

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., & Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, *24*(8), 1384–1395. https://doi.org/10.1101/gr.170720.113

Karjalainen, J., Urpanen, O., Keskinen, T., Huuskonen, H., Sarvala, J., Valkeajärvi, P., & Marjomäki, T. J. (2016). Phenotypic plasticity in growth and fecundity induced by strong population fluctuations affects reproductive traits of female fish. *Ecology and Evolution*, *6*(3), 779–790. https://doi.org/10.1002/ece3.1936

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J. M., Ouellette, S. B., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B. H., Pfister, H., Mirny, L. A., Park, P. J., & Gehlenborg, N. (2018). HiGlass: Web-based visual exploration and analysis of genome interaction maps. *Genome Biology*, *19*(1), 125. https://doi.org/10.1186/s13059-018-1486-1

King, J. L., Marnell, F., Kingston, N., Rosell, R., Boylan, P., Caffrey, J. M., FitzPatrick, Ú., Gargan, P. G., Kelly, F. L., O'Grady, M. F., Poole, R., Roche, W. K., & Cassidy, D. (2011). *Ireland Red List No. 5: Amphibians, Reptiles & Freshwater Fish* (National Parks and Wildlife Service, Department of Arts, Heritage and the Gaeltacht). https://archive.nationalredlist.org/amphibians-reptiles-freshwater-fish-in-ireland/

Knebelsberger, T., Dunz, A. R., Neumann, D., & Geiger, M. F. (2015). Molecular diversity of Germany's

896 freshwater fishes and lampreys assessed by DNA barcoding. *Molecular Ecology Resources*, *15*(3),
897 562–572. https://doi.org/10.1111/1755-0998.12322
898 Kojima, K. K. (2019). Structural and sequence diversity of eukaryotic transposable elements. *Genes &*
899 *Genetic Systems*, *94*(6), 233–252. https://doi.org/10.1266/ggs.18-00024
900 Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation
901 Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4
902 Kottelat, M. (2007). Three new species of Phoxinus from Greece and southern France (Teleostei:
903 Cyprinidae). *Ichthyological Exploration of Freshwaters*, *18*(2), 145–162.
904 Kottelat, M., & Freyhof, J. (2007). *Handbook of European freshwater fishes* (Kottelat, Cornol, & Freyhof,
905 Eds.). Kottelat.
906 Krabbenhoft, T. J., MacGuigan, D. J., Backenstose, N. J. C., Waterman, H., Lan, T., Pelosi, J. A., Tan, M., &
907 Sandve, S. R. (2021). Chromosome-Level Genome Assembly of Chinese Sucker (Myxocyprinus
908 asiaticus) Reveals Strongly Conserved Synteny Following a Catostomid-Specific Whole-Genome
909 Duplication. *Genome Biology and Evolution*, *13*(9), evab190.
910 https://doi.org/10.1093/gbe/evab190
911 Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., Porubsky, D., Kuhn,
912 K., Mueller, K. A., Low, W. Y., Hiendleder, S., Fedrigo, O., Liachko, I., Hall, R. J., Phillippy, A. M.,
913 Eichler, E. E., Williams, J. L., Smith, T. P. L., Jarvis, E. D., … Kingan, S. B. (2021). Extended
914 haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*,
915 *12*(1), Article 1. https://doi.org/10.1038/s41467-020-20536-y
916 Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines,
917 Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819.
918 https://doi.org/10.1093/molbev/msx116
919 Lai, Y.-T., Yeung, C. K. L., Omland, K. E., Pang, E.-L., Hao, Y., Liao, B.-Y., Cao, H.-F., Zhang, B.-W., Yeh, C.-F.,
920 Hung, C.-M., Hung, H.-Y., Yang, M.-Y., Liang, W., Hsu, Y.-C., Yao, C.-T., Dong, L., Lin, K., & Li, S.-H.
921 (2019). Standing genetic variation as the predominant source for adaptation of a songbird.

Proceedings of the National Academy of Sciences, *116*(6), 2152–2157. https://doi.org/10.1073/pnas.1813597116

Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen. (2020). *Fachinformationssystem FischInfo* [dataset].

Leitwein, M., Gagnaire, P.-A., Desmarais, E., Guendouz, S., Rohmer, M., Berrebi, P., & Guinand, B. (2016). Genome-wide nucleotide diversity of hatchery-reared Atlantic and Mediterranean strains of brown trout Salmo trutta compared to wild Mediterranean populations. *Journal of Fish Biology*, *89*(6), 2717–2734. https://doi.org/10.1111/jfb.13131

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997 [q-bio.GN].

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), Article 7357. https://doi.org/10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Bian, C., Yi, Y., Yu, H., You, X., & Shi, Q. (2021). Temporal dynamics of teleost populations during the Pleistocene: A report from publicly available genome data. *BMC Genomics*, *22*(1), 490. https://doi.org/10.1186/s12864-021-07816-7

Li, Q., Jiang, B., Zhang, Z., Huang, Y., Xu, Z., Chen, X., Huang, Y., Jian, J., & Yan, Q. (2022). Involvement and characterization of NLRCs and pyroptosis-related genes in Nile tilapia (Oreochromis niloticus) immune response. *Fish & Shellfish Immunology*, *130*, 602–611. https://doi.org/10.1016/j.fsi.2022.09.041

Li, X., Ye, Y., Peng, K., Zeng, Z., Chen, L., & Zeng, Y. (2022). Histones: The critical players in innate

immunity. *Frontiers in Immunology*, *13*. https://www.frontiersin.org/articles/10.3389/fimmu.2022.1030610

Li, Z.-H., Chen, L., Wu, Y.-H., Li, P., Li, Y.-F., & Ni, Z.-H. (2014). Effects of Mercury on Oxidative Stress and Gene Expression of Potential Biomarkers in Larvae of the Chinese Rare Minnow Gobiocypris Rarus. *Archives of Environmental Contamination and Toxicology*, *67*(2), 245–251. https://doi.org/10.1007/s00244-014-0034-6

Lin, D., Hong, P., Zhang, S., Xu, W., Jamal, M., Yan, K., Lei, Y., Li, L., Ruan, Y., Fu, Z. F., Li, G., & Cao, G. (2018). Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture. *Nature Genetics*, *50*(5), 754–763. https://doi.org/10.1038/s41588-018-0111-2

Litman, G. W., Rast, J. P., & Fugmann, S. D. (2010). The origins of vertebrate adaptive immunity. *Nature Reviews Immunology*, *10*(8), Article 8. https://doi.org/10.1038/nri2807

Liu, G., Xin, S., Geng, S., Zheng, W., Xu, T., & Sun, Y. (2023). Identification of a novel fusion gene NLRC3-NLRP12 in miiuy croaker (Miichthys miiuy). *Fish & Shellfish Immunology*, *136*, 108697. https://doi.org/10.1016/j.fsi.2023.108697

Liu, H., Chen, C., Lv, M., Liu, N., Hu, Y., Zhang, H., Enbody, E. D., Gao, Z., Andersson, L., & Wang, W. (2021). A Chromosome-Level Assembly of Blunt Snout Bream (Megalobrama amblycephala) Genome Reveals an Expansion of Olfactory Receptor Genes in Freshwater Fish. *Molecular Biology and Evolution*, *38*(10), 4238–4251. https://doi.org/10.1093/molbev/msab152

Low, W. Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D. M., Rosen, B. D., Kronenberg, Z. N., Kingan, S. B., Tseng, E., Thibaud-Nissen, F., Martin, F. J., Billis, K., Ghurye, J., Hastie, A. R., Lee, J., Pang, A. W. C., Heaton, M. P., Phillippy, A. M., … Williams, J. L. (2020). Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-15848-y

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., Baalsrud, H. T., Nederbragt, A. J., Hanel, R., Salzburger, W., Stenseth, N. C., Jakobsen, K. S., & Jentoft, S. (2016).

Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, *48*(10), Article 10. https://doi.org/10.1038/ng.3645

Mansfeld, B. N., Boyher, A., Berry, J. C., Wilson, M., Ou, S., Polydore, S., Michael, T. P., Fahlgren, N., & Bart, R. S. (2021). Large structural variations in the haplotype-resolved African cassava genome. *The Plant Journal*, *108*(6), 1830–1848. https://doi.org/10.1111/tpj.15543

Martinson, J. W., Bencic, D. C., Toth, G. P., Kostich, M. S., Flick, R. W., See, M. J., Lattier, D., Biales, A. D., & Huang, W. (2022). De Novo Assembly of the Nearly Complete Fathead Minnow Reference Genome Reveals a Repetitive but Compact Genome. *Environmental Toxicology and Chemistry*, *41*(2), 448–461. https://doi.org/10.1002/etc.5266

Matuszewski, S., Hermisson, J., & Kopp, M. (2015). Catch Me if You Can: Adaptation from Standing Genetic Variation to a Moving Phenotypic Optimum. *Genetics*, *200*(4), 1255–1274. https://doi.org/10.1534/genetics.115.178574

Mendes, F. K., Vanderpool, D., Fulton, B., & Hahn, M. W. (2021). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*, *36*(22–23), 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends in Ecology & Evolution*, *35*(7), 561–572. https://doi.org/10.1016/j.tree.2020.03.002

Miller, J. M., Malenfant, R. M., David, P., Davis, C. S., Poissant, J., Hogg, J. T., Festa-Bianchet, M., & Coltman, D. W. (2014). Estimating genome-wide heterozygosity: Effects of demographic history and marker type. *Heredity*, *112*(3), Article 3. https://doi.org/10.1038/hdy.2013.99

Mills, C. A. (1987). The life history of the minnow Phoxinus phoxinus (L.) in a productive stream. *Freshwater Biology*, *17*(1), 53–67. https://doi.org/10.1111/j.1365-2427.1987.tb01028.x

Mills, C. A. (1988). The effect of extreme northerly climatic conditions on the life history of the minnow, Phoxinus phoxinus (L.). *Journal of Fish Biology*, *33*(4), 545–561. https://doi.org/10.1111/j.1095-8649.1988.tb05498.x

Moiseenko, T. I., Sharov, A. N., Vandish, O. I., Kudryavtseva, L. P., Gashkina, N. A., & Rose, C. (2009). Long-term modification of Arctic lake ecosystems: Reference condition, degradation under toxic impacts and recovery (case study Imandra Lakes, Russia). *Limnologica*, *39*(1), 1–13. https://doi.org/10.1016/j.limno.2008.03.003

Morgulis, A., Gertz, E. M., Schäffer, A. A., & Agarwala, R. (2006). A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *Journal of Computational Biology*, *13*(5), 1028–1040. https://doi.org/10.1089/cmb.2006.13.1028

Morimoto, N., Kono, T., Sakai, M., & Hikima, J. (2021). Inflammasomes in Teleosts: Structures and Mechanisms That Induce Pyroptosis during Bacterial Infection. *International Journal of Molecular Sciences*, *22*(9), Article 9. https://doi.org/10.3390/ijms22094389

Munnecke, A., Calner, M., Harper, D. A. T., & Servais, T. (2010). Ordovician and Silurian sea–water chemistry, sea level, and climate: A synopsis. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *296*(3), 389–413. https://doi.org/10.1016/j.palaeo.2010.08.001

Museth, J., Borgstrøm, R., & Brittain, J. E. (2010). Diet overlap between introduced European minnow (Phoxinus phoxinus) and young brown trout (Salmo trutta) in the lake, Øvre Heimdalsvatn: A result of abundant resources or forced niche overlap? *Hydrobiologia*, *642*(1), 93–100. https://doi.org/10.1007/s10750-010-0162-6

Nachman, M. W. (2002). Variation in recombination rate across the genome: Evidence and implications. *Current Opinion in Genetics & Development*, *12*(6), 657–663. https://doi.org/10.1016/S0959-437X(02)00358-1

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology*, *25*(5), 1058–1072. https://doi.org/10.1111/mec.13540

Nadachowska-Brzyska, K., Li, C., Smeds, L., Zhang, G., & Ellegren, H. (2015). Temporal Dynamics of Avian Populations during Pleistocene Revealed by Whole-Genome Sequences. *Current Biology*, *25*(10), 1375–1380. https://doi.org/10.1016/j.cub.2015.03.047

Nakandala, U., Masouleh, A. K., Smith, M. W., Furtado, A., Mason, P., Constantin, L., & Henry, R. J. (2023). Haplotype resolved chromosome level genome assembly of Citrus australis reveals disease resistance and other citrus specific genes. *Horticulture Research*, *10*(5), uhad058. https://doi.org/10.1093/hr/uhad058

Neumann, N. F., Stafford, J. L., Barreda, D., Ainsworth, A. J., & Belosevic, M. (2001). Antimicrobial mechanisms of fish phagocytes and their role in host defense. *Developmental & Comparative Immunology*, *25*(8), 807–825. https://doi.org/10.1016/S0145-305X(01)00037-4

Nigam, A. K., Kumari, U., Mittal, S., & Mittal, A. K. (2017). Evaluation of antibacterial activity and innate immune components in skin mucus of Indian major carp, Cirrhinus mrigala. *Aquaculture Research*, *48*(2), 407–418. https://doi.org/10.1111/are.12889

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(Database issue), D733–D745. https://doi.org/10.1093/nar/gkv1189

Oscoz, J., Leunda, P. M., Campos, F., Escala, M. C., García-Fresca, C., & Miranda, R. (2005). Spring diet composition of Rainbow Trout, Oncorhynchus mykiss (Walbaum, 1792) in the Urederra river (Spain). *Annales de Limnologie - International Journal of Limnology*, *41*(1), 27–34. https://doi.org/10.1051/limn/2005003

Palandačić, A., Bravničar, J., Zupančič, P., Šanda, R., & Snoj, A. (2015). Molecular data suggest a multispecies complex of Phoxinus (Cyprinidae) in the Western Balkan Peninsula. *Molecular Phylogenetics and Evolution*, *92*, 118–123. https://doi.org/10.1016/j.ympev.2015.05.024

Palandačić, A., Kruckenhauser, L., Ahnelt, H., & Mikschi, E. (2020). European minnows through time: Museum collections aid genetic assessment of species introductions in freshwater fishes (Cyprinidae: Phoxinus species complex). *Heredity*. https://doi.org/10.1038/s41437-019-0292-1

Palandačić, A., Naseka, A., Ramler, D., & Ahnelt, H. (2017). Contrasting morphology with molecular data:

An approach to revision of species complexes based on the example of European Phoxinus (Cyprinidae). *BMC Evolutionary Biology*, *17*(1), 184. https://doi.org/10.1186/s12862-017-1032-x

Palandačić, A., Witman, K., & Spikmans, F. (2022). Molecular analysis reveals multiple native and alien Phoxinus species (Leusciscidae) in the Netherlands and Belgium. *Biological Invasions*, *24*(8), 2273–2283. https://doi.org/10.1007/s10530-022-02784-9

Pasupuleti, M., Schmidtchen, A., & Malmsten, M. (2012). Antimicrobial peptides: Key components of the innate immune system. *Critical Reviews in Biotechnology*, *32*(2), 143–171. https://doi.org/10.3109/07388551.2011.594423

Patel, D. M., & Brinchmann, Monica. F. (2017). Skin mucus proteins of lumpsucker (Cyclopterus lumpus). *Biochemistry and Biophysics Reports*, *9*, 217–225. https://doi.org/10.1016/j.bbrep.2016.12.016

Pavlov, D. S., & Mikheev, V. N. (2017). Downstream migration and mechanisms of dispersal of young fish in rivers. *Canadian Journal of Fisheries and Aquatic Sciences*, *74*(8), 1312–1323. https://doi.org/10.1139/cjfas-2016-0298

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), Article 10. https://doi.org/10.1038/nbt.4235

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *44*(12), e113. https://doi.org/10.1093/nar/gkw294

Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics*, *47*(1), 11.12.1-11.12.34. https://doi.org/10.1002/0471250953.bi1112s47

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-14998-3

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic*

Acids Research, *47*(W1), W191–W198. https://doi.org/10.1093/nar/gkz369

Reier, S., Kruckenhauser, L., Snoj, A., Trontelj, P., & Palandačić, A. (2022). The minnow *Phoxinus lumaireul* (Leuciscidae) shifts the Adriatic–Black Sea basin divide in the north-western Dinaric Karst region. *Ecohydrology*, *15*(6). https://doi.org/10.1002/eco.2449

Reijnders, M. J. M. F., & Waterhouse, R. M. (2021). Summary Visualizations of Gene Ontology Terms With GO-Figure! *Frontiers in Bioinformatics*, *1*. https://www.frontiersin.org/articles/10.3389/fbinf.2021.638255

Röpke, C., Pires, T. H. S., Zuanon, J., Freitas, C. E. C., Hernandes, M. C., Souza, F., & Amadio, S. (2021). Growth–reproduction trade-off and fecundity regulate population stability in Amazon floodplain fishes. *Freshwater Biology*, *66*(6), 1101–1109. https://doi.org/10.1111/fwb.13702

Rose, J. (2009). Early and Middle Pleistocene landscapes of eastern England. *Proceedings of the Geologists' Association*, *120*(1), 3–33. https://doi.org/10.1016/j.pgeola.2009.05.003

Ruddiman, W. F., Raymo, M. E., Martinson, D. G., Clement, B. M., & Backman, J. (1989). Pleistocene evolution: Northern hemisphere ice sheets and North Atlantic Ocean. *Paleoceanography*, *4*(4), 353–412. https://doi.org/10.1029/PA004i004p00353

Shao, F., Han, M., & Peng, Z. (2019). Evolution and diversity of transposable elements in fish genomes. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-51888-1

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sotero-Caio, C. G., Platt, R. N., II, Suh, A., & Ray, D. A. (2017). Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biology and Evolution*, *9*(1), 161–177. https://doi.org/10.1093/gbe/evw264

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(suppl_2), W435–W439. https://doi.org/10.1093/nar/gkl200

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., Rounge, T. B., Paulsen, J., Solbakken, M. H., Sharma, A., Wetten, O. F., Lanzén, A., Winer, R., Knight, J., Vogel, J.-H., Aken, B., Andersen, Ø., Lagesen, K., Tooming-Klunderud, A., … Jakobsen, K. S. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, *477*(7363), 207–210. https://doi.org/10.1038/nature10342

Stein, C., Caccamo, M., Laird, G., & Leptin, M. (2007). Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biology*, *8*(11), R251. https://doi.org/10.1186/gb-2007-8-11-r251

Steinthorsdottir, M., Coxall, H. K., de Boer, A. M., Huber, M., Barbolini, N., Bradshaw, C. D., Burls, N. J., Feakins, S. J., Gasson, E., Henderiks, J., Holbourn, A. E., Kiel, S., Kohn, M. J., Knorr, G., Kürschner, W. M., Lear, C. H., Liebrand, D., Lunt, D. J., Mörs, T., … Strömberg, C. a. E. (2021). The Miocene: The Future of the Past. *Paleoceanography and Paleoclimatology*, *36*(4), e2020PA004037. https://doi.org/10.1029/2020PA004037

Sun, H., Jiao, W.-B., Krause, K., Campoy, J. A., Goel, M., Folz-Donahue, K., Kukat, C., Huettel, B., & Schneeberger, K. (2022). Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature Genetics*, *54*(3), Article 3. https://doi.org/10.1038/s41588-022-01015-0

Sun, L., Gao, T., Wang, F., Qin, Z., Yan, L., Tao, W., Li, M., Jin, C., Ma, L., Kocher, T. D., & Wang, D. (2020). Chromosome-level genome assembly of a cyprinid fish Onychostoma macrolepis by integration of nanopore sequencing, Bionano and Hi-C technology. *Molecular Ecology Resources*, *20*(5), 1361–1371. https://doi.org/10.1111/1755-0998.13190

Sun, Y., Yin, Q., Crucifix, M., Clemens, S. C., Araya-Melo, P., Liu, W., Qiang, X., Liu, Q., Zhao, H., Liang, L., Chen, H., Li, Y., Zhang, L., Dong, G., Li, M., Zhou, W., Berger, A., & An, Z. (2019). Diverse manifestations of the mid-Pleistocene climate transition. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-018-08257-9

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of

Gene Ontology Terms. *PLOS ONE*, *6*(7), e21800. https://doi.org/10.1371/journal.pone.0021800

Svirgsden, R., Rohtla, M., Albert, A., Taal, I., Saks, L., Verliin, A., & Vetemaa, M. (2018). Do Eurasian minnows (Phoxinus phoxinus L.) inhabiting brackish water enter fresh water to reproduce: Evidence from a study on otolith microchemistry. *Ecology of Freshwater Fish*, *27*(1), 89–97. https://doi.org/10.1111/eff.12326

Takeuchi, T., Suzuki, Y., Watabe, S., Nagai, K., Masaoka, T., Fujie, M., Kawamitsu, M., Satoh, N., & Myers, E. W. (2022). A high-quality, haplotype-phased genome reconstruction reveals unexpected haplotype diversity in a pearl oyster. *DNA Research*, *29*(6), dsac035. https://doi.org/10.1093/dnares/dsac035

Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research*, *43*(12), e78. https://doi.org/10.1093/nar/gkv227

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, *31*(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Tekaia, F. (2016). Inferring Orthologs: Open Questions and Perspectives. *Genomics Insights*, *9*, GEI.S37925. https://doi.org/10.4137/GEI.S37925

Tian, F., Liu, S., Zhou, B., Tang, Y., Zhang, Y., Zhang, C., & Zhao, K. (2022). Chromosome-level genome of Tibetan naked carp (Gymnocypris przewalskii) provides insights into Tibetan highland adaptation. *DNA Research*, *29*(4), dsac025. https://doi.org/10.1093/dnares/dsac025

Tigano, A., Jacobs, A., Wilder, A. P., Nand, A., Zhan, Y., Dekker, J., & Therkildsen, N. O. (2021). Chromosome-Level Assembly of the Atlantic Silverside Genome Reveals Extreme Levels of Sequence Diversity and Structural Genetic Variation. *Genome Biology and Evolution*, *13*(6), evab098. https://doi.org/10.1093/gbe/evab098

Vucić, M., Jelić, D., Žutinić, P., Grandjean, F., & Jelić, M. (2018). Distribution of Eurasian minnows (Phoxinus: Cypriniformes) in the Western Balkans. *Knowledge & Management of Aquatic Ecosystems*, *419*, 11. https://doi.org/10.1051/kmae/2017051

1156 Waldbieser, G. C., Liu, S., Yuan, Z., Older, C. E., Gao, D., Shi, C., Bosworth, B. G., Li, N., Bao, L., Kirby, M.
1157 A., Jin, Y., Wood, M. L., Scheffler, B., Simpson, S., Youngblood, R. C., Duke, M. V., Ballard, L.,
1158 Phillippy, A., Koren, S., & Liu, Z. (2023). Reference genomes of channel catfish and blue catfish
1159 reveal multiple pericentric chromosome inversions. *BMC Biology*, *21*(1), 67.
1160 https://doi.org/10.1186/s12915-023-01556-8
1161 Wang, F., Wang, L., Liu, D., Gao, Q., Nie, M., Zhu, S., Chao, Y., Yang, C., Zhang, C., Yi, R., Ni, W., Tian, F.,
1162 Zhao, K., & Qi, D. (2022). Chromosome-level assembly of Gymnocypris eckloni genome.
1163 *Scientific Data*, *9*(1), Article 1. https://doi.org/10.1038/s41597-022-01595-w
1164 Wang, Q., Luo, S., Ghonimy, A., Chen, Y., Guo, Z., Liu, H., & Zhang, D. (2019). Effect of dietary l-carnitine
1165 on growth performance and antioxidant response in Amur minnow (Phoxinus lagowskii
1166 Dybowskii). *Aquaculture Nutrition*, *25*(4), 749–760. https://doi.org/10.1111/anu.12892
1167 Wang, T., Liu, F., Tian, G., Secombes, C. J., & Wang, T. (2019). Lineage/species-specific expansion of the
1168 Mx gene family in teleosts: Differential expression and modulation of nine Mx genes in rainbow
1169 trout Oncorhynchus mykiss. *Fish & Shellfish Immunology*, *90*, 413–430.
1170 https://doi.org/10.1016/j.fsi.2019.04.303
1171 Watanabe, K., Koga, H., Nakamura, K., Fujita, A., Hattori, A., Matsuda, M., & Koga, A. (2014).
1172 Spontaneous germline excision of Tol1, a DNA-based transposable element naturally occurring
1173 in the medaka fish genome. *Genome*, *57*(4), 193–199. https://doi.org/10.1139/gen-2014-0011
1174 Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural
1175 variation: Insights from and for human disease. *Nature Reviews Genetics*, *14*(2), Article 2.
1176 https://doi.org/10.1038/nrg3373
1177 Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of
1178 structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*,
1179 *28*(6), 1203–1209. https://doi.org/10.1111/mec.15066
1180 Westman, J., Papareddy, P., Dahlgren, M. W., Chakrakodi, B., Norrby-Teglund, A., Smeds, E., Linder, A.,
1181 Mörgelin, M., Johansson-Lindbom, B., Egesten, A., & Herwald, H. (2015). Extracellular Histones

Induce Chemokine Production in Whole Blood Ex Vivo and Leukocyte Recruitment In Vivo. *PLoS Pathogens*, *11*(12), e1005319. https://doi.org/10.1371/journal.ppat.1005319

Wu, C.-S., Ma, Z.-Y., Zheng, G.-D., Zou, S.-M., Zhang, X.-J., & Zhang, Y.-A. (2022). Chromosome-level genome assembly of grass carp (Ctenopharyngodon idella) provides insights into its genome evolution. *BMC Genomics*, *23*(1), 271. https://doi.org/10.1186/s12864-022-08503-x

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., Xu, J., Zheng, X., Ren, L., Wang, G., Zhang, Y., Huo, L., Zhao, Z., Cao, D., Lu, C., Li, C., Zhou, Y., Liu, Z., Fan, Z., … Sun, X. (2014). Genome sequence and genetic diversity of the common carp, Cyprinus carpio. *Nature Genetics*, *46*(11), Article 11. https://doi.org/10.1038/ng.3098

Yan, A., & Yu, H. (2022). New insights into centromeres from Arabidopsis Col-CEN assembly. *Trends in Genetics*, *38*(5), 416–418. https://doi.org/10.1016/j.tig.2022.02.001

Yang, J., Chen, X., Bai, J., Fang, D., Qiu, Y., Jiang, W., Yuan, H., Bian, C., Lu, J., He, S., Pan, X., Zhang, Y., Wang, X., You, X., Wang, Y., Sun, Y., Mao, D., Liu, Y., Fan, G., … Shi, Q. (2016). The Sinocyclocheilus cavefish genome provides insights into cave adaptation. *BMC Biology*, *14*(1), 1. https://doi.org/10.1186/s12915-015-0223-4

Yang, X., Liu, H., Ma, Z., Zou, Y., Zou, M., Mao, Y., Li, X., Wang, H., Chen, T., Wang, W., & Yang, R. (2019). Chromosome-level genome assembly of Triplophysa tibetana, a fish adapted to the harsh high-altitude environment of the Tibetan Plateau. *Molecular Ecology Resources*, *19*(4), 1027–1036. https://doi.org/10.1111/1755-0998.13021

Yao, T., Wang, S., Xu, K., Zhang, X., Zhang, T., Wang, J., Wang, Z., & Mu, W. (2022). Biochemical, histological, and gene expression analysis in brain and heart in Phoxinus lagowskii under sustained and diel-cycling hypoxia. *Journal of the World Aquaculture Society*, *53*(4), 860–878. https://doi.org/10.1111/jwas.12901

Yuan, Z., Huang, W., Liu, S., Xu, P., Dunham, R., & Liu, Z. (2018). Historical demography of common carp estimated from individuals collected from various parts of the world using the pairwise sequentially markovian coalescent approach. *Genetica*, *146*(2), 235–241.

https://doi.org/10.1007/s10709-017-0006-7

Zeng, Q., Zhou, Z., He, Q., Li, L., Pu, F., Yan, M., & Xu, P. (2023). Chromosome-level haplotype-resolved genome assembly for Takifugu ocellatus using PacBio and Hi-C technologies. *Scientific Data*, *10*(1), Article 1. https://doi.org/10.1038/s41597-023-01937-2

Zhou, C., McCarthy, S., & Durbin, R. (2022). *YaHS: Yet another Hi-C scaffolding too*. https://doi.org/10.1101/2022.06.09.495093

Zilli, L., Schiavone, R., Storelli, C., & Vilella, S. (2008). Molecular Mechanisms Determining Sperm Motility Initiation in Two Sparids (Sparus aurata and Lithognathus mormyrus). *Biology of Reproduction*, *79*(2), 356–366. https://doi.org/10.1095/biolreprod.108.068296

Zilli, L., Schiavone, R., & Vilella, S. (2017). Role of protein phosphorylation/dephosphorylation in fish sperm motility activation: State of the art and perspectives. *Aquaculture*, *472*, 73–80. https://doi.org/10.1016/j.aquaculture.2016.03.04

1220

1221

1222 **Appendix**

1223 Supplementary figures S1 - S8 and tables S1 - S5: https://doi.org/10.5281/zenodo.10221159

1224

1225 Supplementary tables S6 - S15: https://doi.org/10.5281/zenodo.10220266

1226 **Supplementary figures**

1227 - Figure S1: Heatmap of haplome 1 Hi-C assembly with darker blocks indicating higher intensity of
1228 sequence interaction
1229 - Figure S2: Heatmap of haplome 2 Hi-C assembly with darker blocks indicating higher intensity of
1230 sequence interaction -
1231 - Figure S3: GenomeScope plots of 19, 23, 25 and 30 mer analysis
1232 - Figure S4: Repeats landscape of annotated repeats in Haplotype 2
1233 - Figure S5: Shared homologs between Phoxinus phoxinus and zebrafish, fathead minnow,
1234 goldfish, grass carp and common carp
1235 - Figure S6: Gene ontology of genes in Insertions/Deletions
1236 - Figure S7: Gene ontology of genes in Inversions
1237 - Figure S8: Expansion/contraction profile including all gene families (significant and non-
1238 significant families)

1239

1240 **Supplementary tables**

1241 - Table S1: Summary statistics of 19, 23, 25 and 30 mer genome profiles
1242 - Table S2: Genome assembly statistics of 25 chromosomes of both haplomes
1243 - Table S3: Statistics of repeat sequences annotated in haplome 1
1244 - Table S4: Statistics of repeat sequences annotated in haplome 2
1245 - Table S5: Mapping statistics of RNA sequence data used for protein annotation in both
1246 haplomes
1247 - Table S6: Summary statistics of annotated proteins in haplome 1
1248 - Table S7: Summary statistics of annotated proteins in haplome 2
1249 - Table S8: List of genes present in insertion/deletions
1250 - Table S9: Gene ontology of genes present in insertions/deletions
1251 - Table S10: List of genes present in inversions
1252 - Table S11: Gene ontology of genes present in inversions
1253 - Table S12: Summary statistics of orthology between *Phoxinus phoxinus* proteomes and selected
1254 teleosts
1255 - Table S13: List of genes identified by orthology analysis as *Phoxinus phoxinus*-specific
1256 - Table S14: List of genes significantly expanding in *Phoxinus phoxinus* genome
1257 - Table S15: List of genes significantly contracting in *Phoxinus phoxinus* genome

1258