Peer-reviewed & recommended by PCI Genomics

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, Museum Koenig Bonn – Bonn, Germany

[2] Max Planck Institute of Molecular Cell Biology and Genetics – 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 = 50) by integrating high-fidelity (HiFi) long reads and chromosome conformation capture data (Hi-C). We achieved a haploid size of 940 Megabase pairs (Mbp) for haplome one and 929 Mbp for haplome two with high scaffold N50 values of 36.4 Mb and 36.6 Mb and BUSCO scores of 96.9% and 97.2%, respectively, indicating a highly complete genome assembly.

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.

In the annotation of protein-coding genes, 30,980 mRNAs and 23,497 protein-coding genes were identified with a high completeness score, which further underpins the high contiguity of our genome assemblies. 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 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 lineages present (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 suffered from the lack of a high-quality reference genome, which would make it possible to go beyond the analysis of single-gene markers.

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 Eurasian minnows including all cryptic species in the *Phoxinus* genus. 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. Yet, 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 the 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 facilitated with the latest development of chromosome conformation capture sequencing, namely Hi-C (Belton et al., 2012; Lin et al., 2018) and its integration into assembly pipelines (H. Cheng et al., 2021). Resultantly, the number of studies acknowledging genetic variation between haplomes is rapidly increasing, e.g. in Cavendish banana (Huang et al., 2023), ginger (S.-P. Cheng et al., 2021), carnation (L. Jiang et al., 2022),

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), turkey (Barros et al., 2023), 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 an individual *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 has two major implications. Firstly, it should provide the *Phoxinus* research community with the necessary basis for the phylogenetic unravelling of this complex genus. Secondly, 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 male *P. phoxinus* specimens from the Museum Koenig Bonn live exhibition, 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 until use. A proxy specimen voucher from the neotype locality (River Agger, GPS: 50.841358, 7.201399) sampled on 11 May 2022 is available at the ichthyological collection, Museum Koenig Bonn, under ICH-132418 for the body and at the LIB Biobank, Museum Koenig Bonn, under accession number ZFMK-TIS-64936 for fin tissue, while additional fin tissue from the sampled individuals are deposited for future reference at LIB Biobank under accession numbers ZFMK-TIS-60719 and ZFMK-TIS-60720. Tissues (brain, liver, spleen, muscle, heart, gonad) from individual ZFMK-TIS-60720 were used for long read libraries to generate the genome assembly; tissues (brain, gill, gonad, liver, muscle, skin, spleen) from ZFMK-TIS-60719 were used for RNA-sequencing and subsequent genome annotation.

Extraction of High Molecular Weight Genomic DNA

High molecular weight genomic DNA (HMW 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, spleen tissue was 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. 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 kilobases (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 50 kb 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 v.6.3.0 (https://github.com/PacificBiosciences/ccs.git) command line tool and further refined by applying the tool DeepConsensus v0.2 (Baid et al., 2023), 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

159 -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
160 N50 of 10.2 kb.

161 **Chromatin Conformation Capturing Library Preparation and Sequencing**

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

171 **RNA Extraction and Sequencing**

172 RNA was extracted from flash-frozen brain, gill, gonad, liver, muscle, skin, and spleen (specimen
173 identifier ZFMK-TIS-60719) tissues. All tissues except for spleen and liver were extracted using the RNeasy
174 Mini Kit (Qiagen, Cat. No. 74104) suitable for tissue amounts of 0.5 to 30 mg. Spleen and liver were
175 extracted using QIAamp DNA Micro Kit (Qiagen) suitable for starting material of less than 5 mg. All tissues
176 were homogenised using PowerBead Tubes Ceramic (Qiagen, diameter 2.8mm; Cat. No. 13114-50) on a
177 PowerLyzer™ 24 Bench Top Bead-Based Homogenizer (MO BIO Laboratories, Inc). Before use, 180 µL 1 M
178 dithiothreitol (DTT) was added to 4.5 ml lysis buffer RLT. 350 µL prepared RLT buffer was added to each
179 tissue in a PowerBead Tube (Qiagen) and one cycle of 45 seconds at 3500 rpm and another for 45 seconds
180 at 4200 rpm on the PowerLyzer™ Homogenizer with a 5-minute pause at -20°C were performed. After
181 homogenisation, the solution was centrifuged for 1 minute at 14,800 rpm. The lysate including tissue
182 debris was transferred into a fresh 1.5 mL tube (LoBind®) and again centrifuged at 14,800 rpm for 3
183 minutes at room temperature (RT). Finally, the lysate was transferred to a gDNA Eliminator column
184 (provided in the QIAamp Kits) and centrifuged at 10,000 rpm for 30 sec at RT. Washing and elution steps
185 followed the standard protocol of the Micro and Mini Kit, respectively. Samples were eluted into 14 µL
186 (spleen, liver) and 30 µL (all other tissues) RNAse-free water during 1 minute at 10,000 rpm centrifugation.
187 RNA library preparation and sequencing were performed at Biomarker Technologies (BMKGene) GmbH
188 on an llumina Novaseq 6000 S4 chip.

189 **K-mer Based Genome Size and Heterozygosity Estimation**

190 To estimate the genome size and heterozygosity of the genome, a k-mer-based approach was used. K-
191 mers were counted in the HiFi reads using Jellyfish v2.3.0 (Marçais & Kingsford, 2011) with parameters
192 count *-C -m*, with *-C* indicating that the input reads from both the forward and reverse strand are being
193 used, and *-m* specifying a k-mer length; we specified k-mer lengths of 19, 23, 25 and 30. The resulting k-
194 mer counts were exported into a k-mer count histogram using the *histo* command of Jellyfish for
195 subsequent genome size and heterozygosity estimation in GenomeScope v2.0 (Ranallo-Benavidez et al.,
196 2020).

197 ***De novo* Genome Assembly and Scaffolding**

198 Initial contigs for two representative haplomes were generated using Hifiasm v0.16.1-r375 (H. Cheng
199 et al., 2021) with parameters --h1 --h2 -l2. Next, purge-dups v1.2.3 (Guan et al., 2020) was run on each
200 produced haplotype assembly. The purging parameter "-l2" (purge all haplotigs at a similarity threshold
201 of 75%) was chosen over the default "-l3" (purge all haplotigs at a similarity threshold of 55%) based on
202 BUSCO and contiguity scores (Table 1). Assembly completeness for each produced haplotype assembly
203 was assessed using BUSCO v5.4.7 (Simão et al., 2015) run in genome mode with the actinopterygii_odb10
204 database.

Table 1: Assembly statistics for haplomes 1 and 2 (Hap1, Hap2) from a parameter sweep using differing purging parameters in Hifiasm, followed by the use of purge-dups. BUSCO scores shown are for Complete (C), Single Copy (S), Duplicated (D), Fragmented (F), and Missing (M) genes.

Assembly	BUSCO C	BUSCO S	BUSCO D	BUSCO F	BUSCO M	Assembly size (bp)	No. contigs	Contig N50 (bp)
hifiasm -l2 Hap1	3529	3461	68	31	80	971,720,047	1,130	4,025,542
+purge-dups	3516	3476	40	32	92	939,827,819	556	4,091,346
hifiasm -l2 Hap2	3551	3479	72	29	60	970,047,104	734	4,642,256
+purge-dups	3534	3491	43	28	78	929,446,752	485	5,136,536
hifiasm -l3 Hap1	3549	3480	69	28	63	975,430,134	1,153	4,449,420
+purge-dups	3522	3481	41	27	91	928,353,632	507	4,580,342
hifiasm -l3 Hap2	3532	3461	71	32	76	964,775,996	729	4,275,860
+purge-dups	3522	3481	41	33	85	938,372,303	515	4,317,088

Initial scaffolding of the primary assembly was performed by mapping HiC reads to the primary contigs using bwa-mem v0.7.17-r1198-dirty (H. Li, 2013) and mappings were filtered following the VGP Arima mapping pipeline (https://github.com/VGP/vgp-assembly/tree/master/pipeline/salsa). The final bed file was given to yahs v1.1a (Zhou et al., 2022) for scaffolding. Scaffolds were then manually curated into chromosomes using higlass v2.1.11 (Kerpedjiev et al., 2018) with Cooler v0.9.1 (Abdennur & Mirny, 2020) to visualise the HiC data (Supplementary Figures 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 (Poplin et al., 2018) with --model_type=PACBIO. Errors were corrected in the assembly by filtering the vcf file given by deepvariant with bcftools *view* v1.12 (Danecek et al., 2021) with arguments *-i 'FILTER=\"PASS\" && GT=\"1/1\"'* and a consensus called with bcftools *consensus*. To determine whether any contaminant sequences were present in the genome, we screened for adapter, vector, and foreign sequences using NCBI Foreign Contamination Screen (FCS) v0.2.1 and FCS-GX v0.3.0 (Astashyn et al., 2024). Resulting hits indicative of contaminants were removed from the assembled sequences.

Additionally, we estimated the correctness of the assembly by counting the k-mers of the HiFi reads using FastK (https://github.com/thegenemyers/FASTK.git) with options “-v -t1” and “kmer=31” and then running Merqury.FK v.1 (https://github.com/marbl/merqury.git) on both phased haplomes. The analysis yields Merqury's consensus Quality Value (QV), which is estimated by comparing the read and assembly

k-mer counts and then transformed to a log-scaled probability of base-call errors. A higher QV indicates a more accurate assembly.

To detect any possible biases in the input data that might have led to misassemblies, we aligned the raw PACBIO reads, the HiFi and HiC reads back to each phased assembly. For the PACBIO raw data we used pbmm2 v.1.13 (https://github.com/PacificBiosciences/pbmm2.git) with options "--preset "SUBREAD" --sort-memory 20G" for each SMRT cell. For HiFi reads, we used minimap2 v. 2.26 (Li, 2018) with options "minimap2 -acyL --secondary=no --MD --eqx -x map-hifi -k 20" and for the HiC we used bwa-mem2 v.2.2.1 (Vasimuddin et al., 2019) with option "-M". All resulting mapping files were sorted with SAMtools v1.17 (Li et al., 2009) *sort* and the mapping quality was estimated with qualimap v.2.3 (Okonechnikov et al., 2016) with options "-c -outformat pdf --java-mem-size=10G".

A mitochondrial assembly was created using MitoHiFi v2.0 (Uliano-Silva et al., 2023) using the DeepConsensus PacBio Reads and *Phoxinus phoxinus* reference mitochondrial genome NC_020358.1 as input. Any partial mitochondrial contigs remaining in the assembly were removed based on mapping synteny to the fully assembled mitogenome.

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 using DustMasker from BLAST+ v. 2.9 of 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. The output files of Dustmasker and Tandem Repeats Finder were converted to gff3 format with the custom perl script repeat_to_gff.pl v. 1.0 and the output files of RepeatMasker were converted to *gff3* format with RepeatMasker's rmOutTogff3.pl script and edited with a custom perl command to add a unique ID to each entry. We merged the positions of repeats with Bedtools v2.29.2 (Quinlan, 2014) before our genome was softmasked with these identified repeats using Bedtools in preparation for the prediction of protein-coding genes.

Annotation of Protein Coding Genes

The annotation of protein-coding genes in the *P. phoxinus* genome was carried out with a combination of *ab initio*, protein similarity, and transcriptome-based protein prediction models. The sequencing reads (NCBI accession numbers SRR26699630 to SRR26699636) from seven RNA samples from brain, gill, gonad, liver, muscle, skin, and spleen were mapped to both haplomes 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 then used for annotation of protein-coding genes. With the RNA mapping files as input, BRAKER3 uses StringTie2 (Kovaka et al., 2019) to create a draft transcriptome, which then serves as the basis to predict protein-coding genes with GeneMarkS-T (Tang et al., 2015). The genes with the best similarity scores and quality of *ab initio* predictions are then selected with GeneMark-ETP (Brůna et al., 2020). Then, GeneMark-ETP uses the protein, RNA and ab initio-predicted genes to produce 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 were used to create a set of high-confidence genes. AUGUSTUS (Stanke et al., 2006) was trained with these gene sets and performed an *ab initio* prediction of another genome-wide gene set before TSEBRA (Gabriel et al., 2021) combined 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.

Functional annotation of the resulting predicted genes was done with the *blastp* algorithm implemented in DIAMOND v2.1.8 (Buchfink et al., 2021). Our predicted protein-coding genes were blasted against protein databases of Swiss-Prot, edition 2023-05-13 (The UniProt Consortium, 2023), the Protein Data Bank (PDB) database, edition 2023-05-13 (RCSB.org, Berman et al., 2000) and TrEMBL, edition 2023-05-14 (The UniProt Consortium, 2023) 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 (Cantalapiedra et al., 2021).

Protein sequences of *Danio 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 included *Pimephales promelas* (Fathead minnow) as the species most closely related to *P. phoxinus,* for which a well-documented whole genome assembly and annotation are available. To identify similar sequences, first AGAT v1.1.0 (Dainat, 2021) was used to extract the longest gene isoforms, against which the protein-coding gene set of *P. phoxinus* was then mapped with DIAMOND v2.1.8 (Buchfink et al., 2021).

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 (Quinlan, 2014) with *-w 100000.* Then, we estimated depth in 100 kb windows for both haplomes using *coverage -mean* as implemented in Bedtools.

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 Megabase (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 (Wilkinson, 2011), to visually inspect patterns of heterozygosity across the entire genome. Patterns of heterozygosity often reflect standing genetic variation which may be linked to the 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 (.vcf) 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 3 years (Museth et al., 2010), which is the maturation age of the *P. phoxinus* and an assumed mutation rate of 3.51 $x10^{-9}$ (Tian et al., 2022). The results of the run were plotted using ggplot2.

Structural Variation

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

Identification of Orthologs and Gene Family Evolution

As input to Orthofinder v2.5.4 (Emms & Kelly, 2019) with default settings, we used the same set of proteomes from species that we downloaded and described in the section on annotation of protein-coding genes, after filtering out amino acids shorter than 30 base-pairs (bp). Orthofinder created orthogroup gene trees and then inferred a rooted species tree from gene duplication events with the STRIDE approach (Emms & Kelly, 2017) using just single-copy orthologs. The species tree underwent conversion into an ultrametric tree utilising Orthofinder's *make_ultrametric.py* script, employing the parameter -r 142. This value corresponds to approximately 142 million years, which aligns with the estimated median divergence time between Siluriformes and Cypriniformes, as suggested by the *TimeTree* database (Kumar et al., 2017). This estimate is calculated by taking into account the median of published estimated divergence times using molecular sequence data analysis (Kumar et al., 2017; Kumar & Hedges, 2011). Orthologous gene counts from the Orthofinder run were filtered to remove orthologous groups that are overrepresented in a particular species using the *clade_and_size_filter.py* script in CAFE v5.0.0 (Mendes et al., 2021). Gene family evolution analysis was carried out in CAFE v5.0.0 using the ultrametric tree and the filtered orthologous gene counts as input. Genes that were confirmed to be species-specific, expanding, and contracting in *P. phoxinus,* were mapped against the TrEMBL database with diamond *blastp*. The GO terms derived from the TrEMBL database were then used as input into REViGO (Supek et al., 2011) to remove redundant terms and summarise GO results.

Results and Discussion

K-mer Based Genome Size and Heterozygosity Estimation

Genome size estimations using k-mer lengths 19, 23, 25, and 30, showed slight differences in the estimated genome sizes and heterozygosities (Figure S3). We chose the 19-mer length due to a lower error rate in comparison to other k-mer lengths. We estimated a haploid genome size of *P. phoxinus* of 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/). The estimated heterozygosity of 1.43% is high compared to other published k-mer-based heterozygosity estimates in Cypriniformes. 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 cypriniform 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). The clear bimodal coverage distribution seen in the k-mer plot (Figure 1) is indicative of the diploidy of the *P. phoxinus* genome, with the first peak at approximately 18-fold coverage corresponding to heterozygous sites and the second peak at approximately 36-fold coverage corresponding to homozygous genome positions.

Utilising the proportion of the assembly sequence supported by HiFi reads, Merqury calculated a Quality Value (QV) of 58.9 (corresponding to 99.99871% accuracy) for haplome 1 (Hap1) and 58.8 (99.99868% accuracy) for haplome 2 (Hap2). Aligning the PacBio raw subreads, HiFi, and Illumina HiC reads to both haplomes revealed comparable coverage levels (600±378, 39.03±28, 83.6268±132-fold, respectively, for Hap1; 39±25, 84±82, 605±353-fold, respectively, for Hap2) and mapping quality scores (50, 52, 42, respectively, for Hap1; 50, 52, 43, respectively, for Hap2). These results further indicate that the assembly is well-phased with minimal assembly errors (Supplementary Table S1).

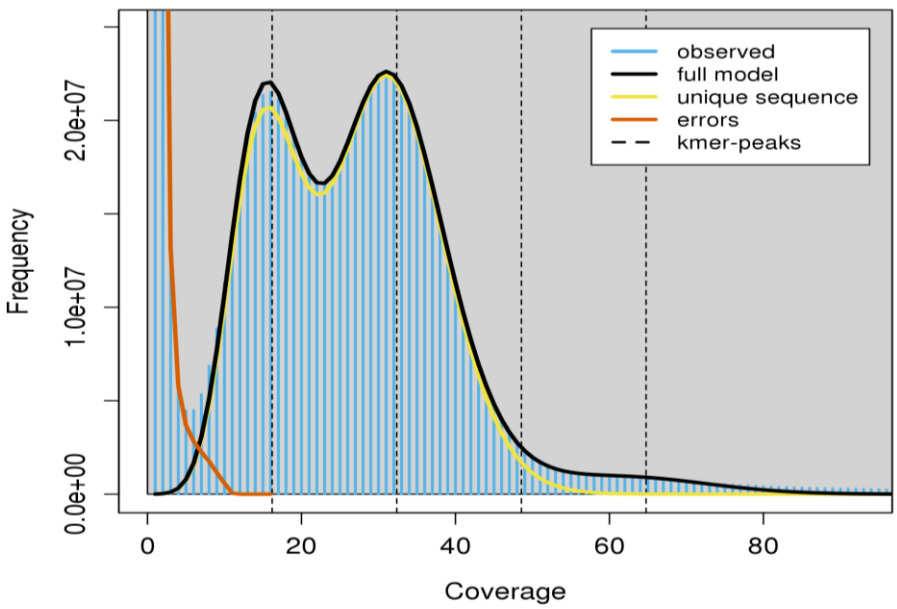


Figure 1: K-mer spectra profile of *Phoxinus phoxinus* generated from raw PacBio HiFi reads with GenomeScope2. The y-axis shows the k-mer counts and the x-axis shows sequencing depth. The clear bimodal pattern observed indicates a diploid genome with high heterozygosity of 1.43%. The haploid genome size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

388 **Genome Assembly and Scaffolding**

389 We generated a haplotype-resolved assembly of *P. phoxinus*, with a total assembly size of 940 Mbp
390 and 929 Mbp for Hap1 and Hap2, respectively. The initial assembly of each haplome contained 394
391 primary contigs, which were further placed into 101 (N50: 36.4 Mb) and 81 (N50: 36.6 Mb) scaffolds,
392 respectively, by mapping HiC reads to the primary contigs. Integrating Hi-C data, we ultimately scaffolded
393 99.6% of the Hap1 assembly and 99.5% of the Hap2 assembly each into 25 chromosomes (Supplementary
394 Table S2). BUSCO analysis revealed the *P. phoxinus* haplomes had only 2.3% and 2.1% missing genes. The
395 assembly statistics for both haplomes (Table 2) denote high contiguity and high completeness. The 11 Mb
396 difference in size between both haplomes is a result of structural variations between haplomes.

397 **Table 2.** Summary of *P. phoxinus* haplome 1 (Hap1) and haplome 2 (Hap2) genome assembly statistics.

Assembly Statistics	Hap1	Hap2
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

398

399 **Annotation of Repetitive Elements**

400 Teleost fish genomes are characterised by a wide range of repetitive elements (Gao et al., 2016). The
401 genome of *P. phoxinus* follows this pattern, with approximately 506 Mbp (53.86%) of its assembly
402 consisting of repetitive elements. This is higher than reported to date in other cyprinid species such as
403 fathead minnow (43.27%) (Martinson et al., 2022), grass carp (43.26%) (Wu et al., 2022), common carp
404 (31.3%) (Xu et al., 2014) or goldfish (39.6%) (Z. Chen et al., 2019). However, the repeat content is slightly
405 lower than that of zebrafish, which contains around 54.47% repetitive elements (Howe et al., 2013).

406 From the identified TEs, DNA transposons were the most abundant repeat elements in the genome,
407 making up 21.25% (approximately 199 Mb). DNA transposon content is similar to that of the fathead
408 minnow genome, which contains an estimated 21.47% (Martinson et al., 2022). SINEs were relatively
409 poorly represented at 1.03%, in line with other published fish genomes (Shao et al., 2019; Sotero-Caio et
410 al., 2017). Retroelements made up 11%, LTRs 6.16% and LINEs 4.15%. Overall, we annotated 46.59% of
411 the *P. phoxinus* genome to be interspersed repeats, the remaining repetitive elements were annotated
412 as 0.47% small RNA, 1.50% satellites, 2.66% simple repeats, and 0.17% low complexity repeats (full details
413 of all repetitive elements are described in Tables S3 and S4).

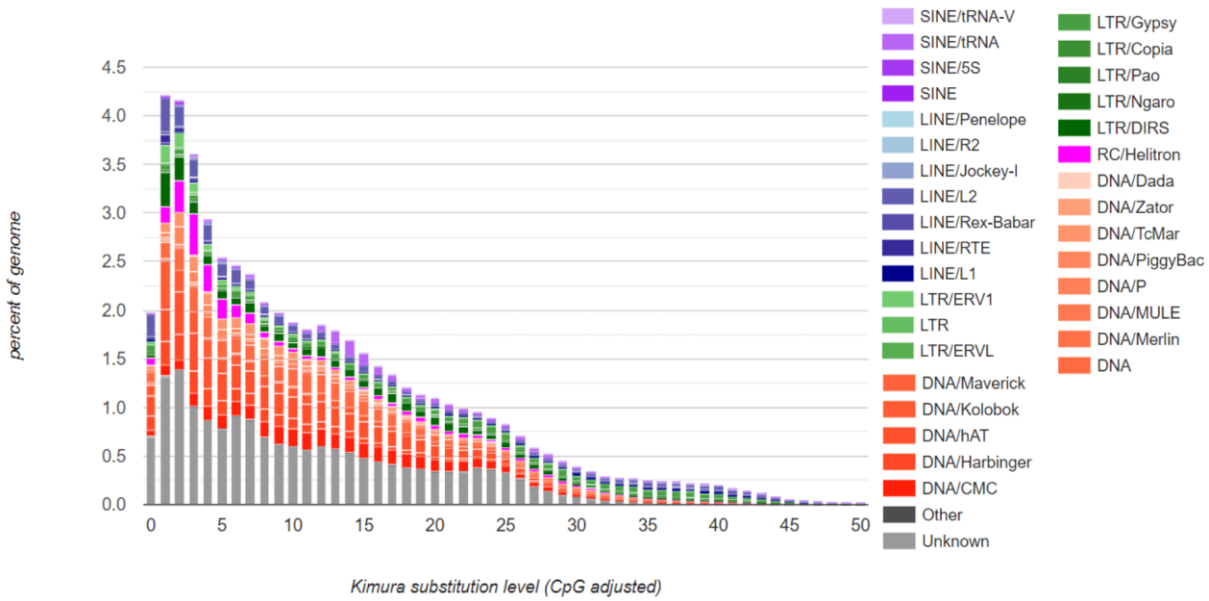


Figure 2: Repeat landscape of transposable elements (TEs) in the *Phoxinus phoxinus* genome across different Kimura substitution levels (in %). TEs on the left side of the histogram represent recently active TEs with low divergence from the consensus sequence of TEs, while TEs towards the right side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

Repeat landscapes illustrate how TEs cluster in correlation to the Kimura substitution rate, i.e. how quickly TEs in a genome have diverged from the consensus TE sequence (Shao et al., 2019). The Kimura substitution level is correlated with the age of transposition activity: low substitution levels indicate recent transposition events; higher substitution levels suggest old transposition events. The repeat landscape of the annotated *P. phoxinus* genome (Figure 2) indicates a high level of active, i.e., recent, transposition bursts (Watanabe et al., 2014). In fish genomes, active transposons accumulate at a faster rate than they decline (Shao et al., 2019), leading to an abundance of active transposons with low Kimura substitution levels (Böhne et al., 2012; Gao et al., 2016; Sotero-Caio et al., 2017), an exact pattern observed in TEs in the *P. phoxinus* genome. For the repeat landscape of Hap2 we refer to Supplementary Figure S4.

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

Annotation of Protein Coding Genes

Prediction of protein-coding genes, based solely on transcripts derived from our RNA-seq data mapped to the haplomes (Supplementary Table S5), generated 40,364 gene models and a mRNA count of 42,408 in Hap1 and 38,754 gene models, and 40,716 mRNAs in Hap2. The *ab-initio* prediction resulted in 24,420 and 35,806 gene models in Hap1 and Hap2, respectively. After filtering and generating a final high-confidence gene set, the final annotation contained 30,980 mRNAs, of which 23,397 were marked as high-confidence genes in Hap 1; and 29,614 mRNAs, with 23,191 high-confidence genes in Hap 2 (full summary statistics of protein annotation for Hap1 and Hap2 in Supplementary Tables S6 and S7). We ran BUSCO with the actinopterygii_odb10 database and obtained a total of 3,484 (95.7%) complete BUSCOs, 3,435 (94.4%) complete and single-copy BUSCOS, 49 (1.3%) complete and duplicated BUSCOs, 20 (0.5%) fragmented BUSCOs and 136 (3.8%) missing BUSCOs. The predicted proteins were functionally annotated using a combination of similarity mapping to protein databases (Swiss-Prot, TrEMBL and PDB) and

446 orthology assignment (eggNOG), resulting in 22,761 annotated genes (Table 3). These statistics hint at a
447 well-assembled and annotated *P. phoxinus* genome. Similarity search of genes in *P. phoxinus* against
448 zebrafish, fathead minnow, goldfish, common carp and grass carp resulted in 21,206 (90.6%) annotated
449 genes across all 5 species (Supplementary Figure S5). Comparing the predicted gene content of Hap1 to
450 its closest relative with available whole-genome resources, the fathead minnow contains more predicted
451 genes (26,150) (Martinson et al., 2022). Still, for *P. phoxinus*, the complete BUSCO score of 95.7% (3,484)
452 was higher compared to the fathead minnow's 94.3% (Martinson et al., 2022).

453 **Table 3.** Functional annotation statistics for genes in *P. phoxinus* using TrEMBL, Swiss-Prot, PDB, and
454 eggNOG indicate that 97.28% (22,761) of genes were successfully annotated, with 84.93% (19,871)
455 present in all four databases.

Annotation Database	Number of annotated genes identified in Hap1	Percentage of annotated genes (%)
Swiss-Prot	19,978	85.39
TrEMBL	22,761	97.28
egg-NOG	21,296	91.02
PDB	19,978	85.39

456

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

458 *Sequence variation follows universal pattern of eukaryotes*

459 Previously, genomes with high heterozygosity presented assembly challenges due to multiple scaffold
460 formation from the same sequence (Kajitani et al., 2014; Pryszcz & Gabaldón, 2016), which resulted in
461 merging of sequence diversity into a single reference genome. However, the implementation of long-
462 range information such as Hi-Fidelity long reads and Hi-C contact maps can aid in overcoming these
463 difficulties by resolving the intra-individual sequence variation (Kronenberg et al., 2021). Our high-quality
464 assembly, indicating increased levels of standing genetic variation, demonstrates the effectiveness of this
465 approach. Previous studies have already indicated abundant levels of genetic variation in the *Phoxinus*
466 genus (Bogutskaya et al., 2020; Reier et al., 2022).

467 Genome-wide estimates and patterns of mean heterozygosity were similar between both haplomes,
468 with an estimate of 0.7866% in Hap1 and 0.7926% in Hap2. Assembly-based heterozygosity estimates
469 were lower than k-mer-based estimates. A similar behaviour was reported in the Genomescope release
470 paper (Vurture et al., 2017). In the present case, the difference between the two approaches used is likely
471 the cause: the genome-wide approach generates heterozygosity estimates from direct observation of the
472 confidently mapped loci as well as the thereof derived SNPs; the k-mer-based approach, however,
473 additionally incorporates structural variants and is more sensitive to low coverage and error-prone
474 regions and hence results in higher heterozygosity estimates. Heterozygosity estimates between the
475 chromosomes varied, however, the pattern found on each chromosome was similar with regions of
476 highest heterozygosity always found at the ends of chromosomes and the lowest heterozygosity in the
477 central region (Figure 3).

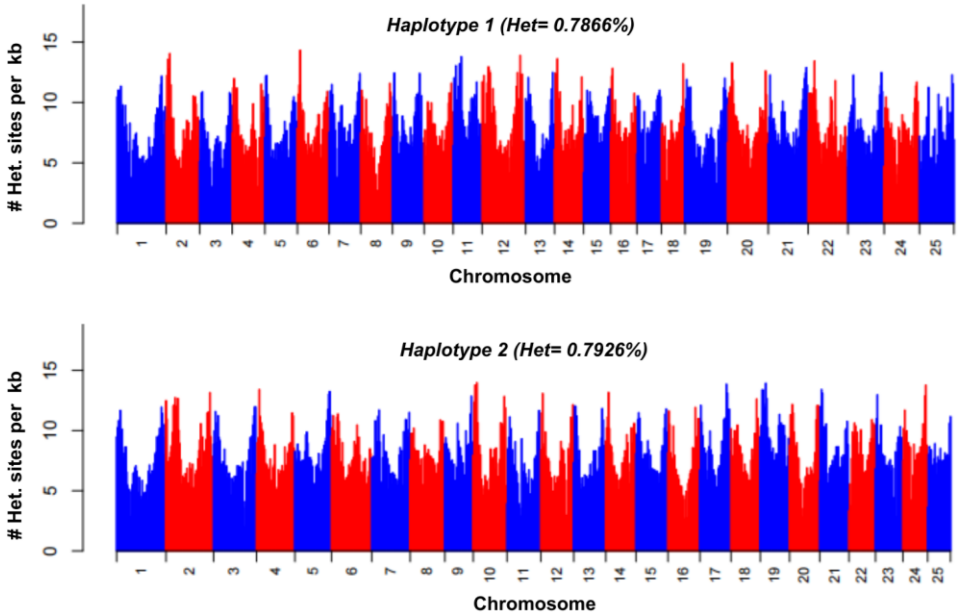


Figure 3: Distribution of heterozygosity along the 25 chromosomes in *P. phoxinus* shown for Hap1 (upper plot) and Hap2 (lower plot). Mean heterozygosity estimates were calculated in 1Mb bins and converted to the number of heterozygous sites per kb across all 25 chromosomes with alternating colours to differentiate adjacent chromosomes.

Regions of high heterozygosity and high recombination rates may be linked to telomeric regions of the chromosome, which in turn can be associated with higher genetic diversity (Jaramillo-Correa et al., 2010). In fishes, this pattern was observed in the genomes of the Atlantic silverside (*Menidia menidia*) (Tigano et al., 2021), brown trout (*Salmo trutta)* (Leitwein et al., 2016) and sticklebacks (*Gasterosteus* spp.) (Sardell et al., 2018). This pattern is consistent with the pattern of heterozygosity observed in the *P. phoxinus* genome. Regions of reduced heterozygosity found at central locations may be associated with centromeres, which are known as "cold spots" for recombination in eukaryotes (Yan & Yu, 2022).

Demographic History of P. phoxinus

Climatic fluctuations have always affected the distribution and survival of species (Hewitt, 2000; J. Li et al., 2021). For aquatic species in the temperate zone, the effects of climate fluctuations due to glaciation cycles, which led to fluctuating sea water levels, were particularly pronounced (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 approximately 10 thousand years ago (kya).

We found a gradual and continuous rise in *Ne* with time, from roughly 20,000 individuals about 10 mya to around 25,000 individuals approximately 800 kya (Figure 4). 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 fishes. 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. A major glaciation in this period is the Saalian glaciation of Northern and North Central Europe from approximately 400 – 150 kya (Lang et al., 2018; Lauer & Weiss, 2018). 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 warmer territories in southern Europe seeking refuge, or underwent adaptation (Feliner, 2011).

The reduction in *Ne* during mid-Pleistocene glaciations (Saalian glaciation) is generally observed in fishes (Brahimi et al., 2016; Gharrett et al., 2023), mammals (Fedorov et al., 2008), and birds (Kozma et al., 2016), especially in European species (Yuan et al., 2018). In *P. phoxinus*, the end of the mid-Pleistocene from around 40 kya was followed by a gradual and eventually sharp increase in *Ne*, peaking at nearly a million estimated individuals, indicating that *P. phoxinus* was able to repopulate previously covered freshwater habitats across Europe.

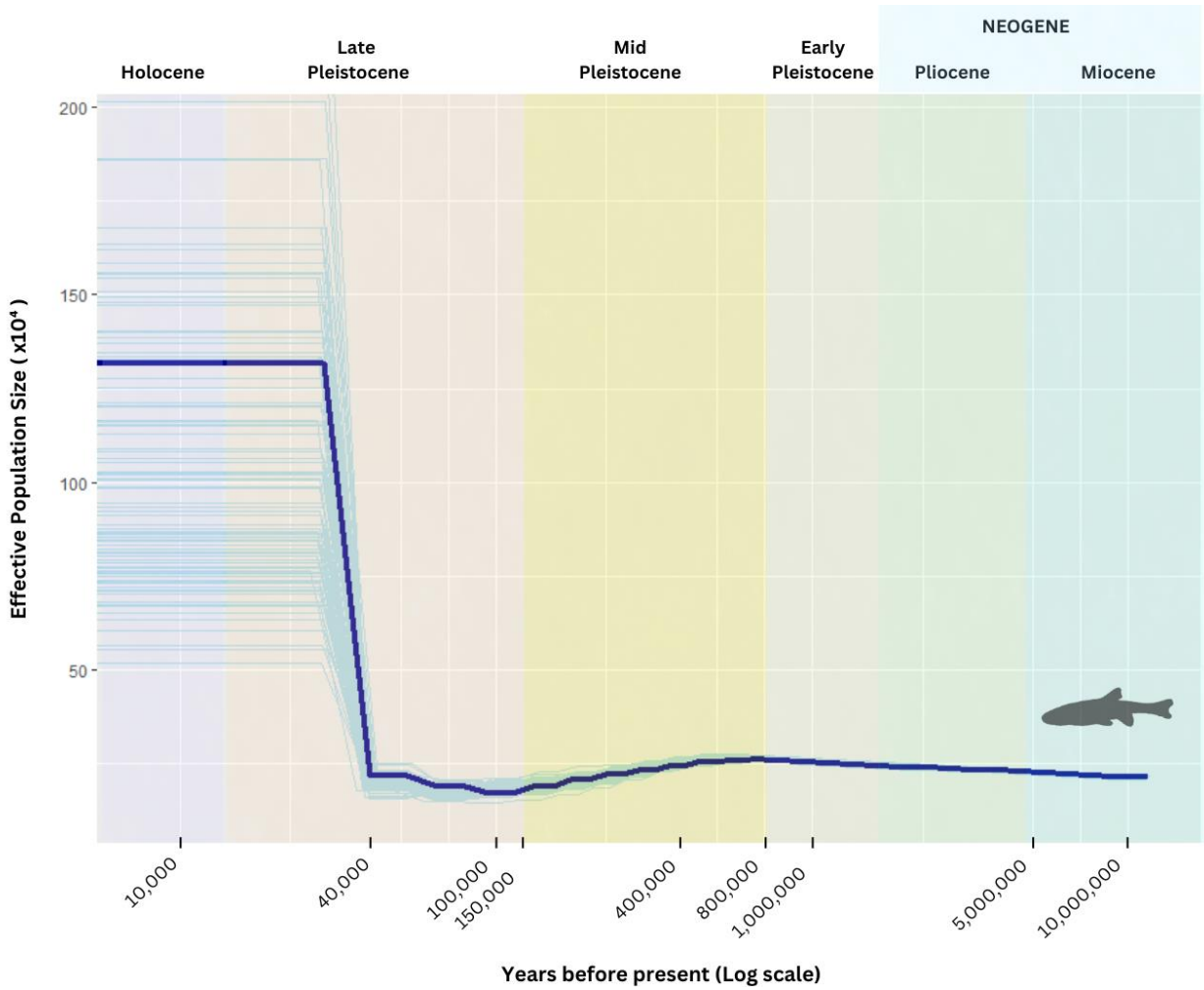


Figure 4. Inferred demographic history of *P. phoxinus* from PSMC analysis based on Hap1. The time ranges cover the Miocene, Pliocene, Early to Late Pleistocene, and Holocene, each shown in different colours. The x-axis shows years before present on a logarithmic scale and the y-axis shows the estimated effective population size. Bootstrap results are shown as transparent lines. Hapl2 showed a similar curve.

Large non-syntenic regions between P. phoxinus haplomes

A comparison of both haplomes of *P. phoxinus* revealed around 815 to 820 Mb of syntenic regions (Table 4, Figure 5). We also identified 1,572 translocations of 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 10.8 Mb in Hap1 and 11.1 Mb in Hap2 (Table 4, Figure 5).

Table 4. Identity, number, and total lengths of identified structural variations between Hap1 and Hap2 in the *P. phoxinus* reference genome.

Structural Variation	Count	Length Hap1 [bases]	Length Hap2 [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 (Hap1)	214	57,1045	-
Duplications (Hap2)	667	-	1,776,709
Insertions	447,578	-	8,108,033
Deletions	447,699	8,113,615	-

528

529 Inversions represented the dominant structural variation within the *P. phoxinus* genome in terms of
530 total size in either haplome (Table 4). The largest eight inversions made up 72% (8 Mb) of the estimated
531 11 Mb of inversions present in the *P. phoxinus* genome. Using Hap1 as a reference for inversion size,
532 chromosome 6 contained the largest inversion with a size of 1.8 Mb; closely followed by a 1.7 Mb
533 inversion on chromosome 2; a 1.5 Mb inversion on chromosome 21; a 1.1 Mb inversion on chromosomes
534 9 and 19, each; and finally, an 800 kb inversion on chromosome 3 (Figure 7).

535 We investigated what type of genes were enriched in regions of copy number structural variants,
536 specifically deletions and insertions, which can also be characterised as presence-absence variations
537 (PAVs). GO terms overrepresented in PAVs between both haplomes included terms such as
538 "dephosphorylation", "peptidyl-serine phosphorylation", and "protein processing", which have been
539 linked with the activation of sperm motility in numerous fish species (Dumorné et al., 2018; Zilli et al.,
540 2008, 2017). GO terms related to development were also enriched within PAVs, such as "regulation of
541 developmental growth" and "developmental pigmentation" (Supplementary Tables S8-S9 and Figure S6).
542 GO terms overrepresented within inversions, were primarily associated with immunity and transposable
543 elements in hAT, DDE, and L1 families: namely "phagocytosis", "phagosome maturation involved in
544 apoptotic cell clearance" and "phagolysosome assembly involved in apoptotic cell clearance".
545 Phagocytosis is a conserved immune defence mechanism across multicellular organisms that prevents
546 infection and invasion of the body by pathogens (Esteban et al., 2015; Neumann et al., 2001)
547 (Supplementary Tables S10-S11 and Figure S7).

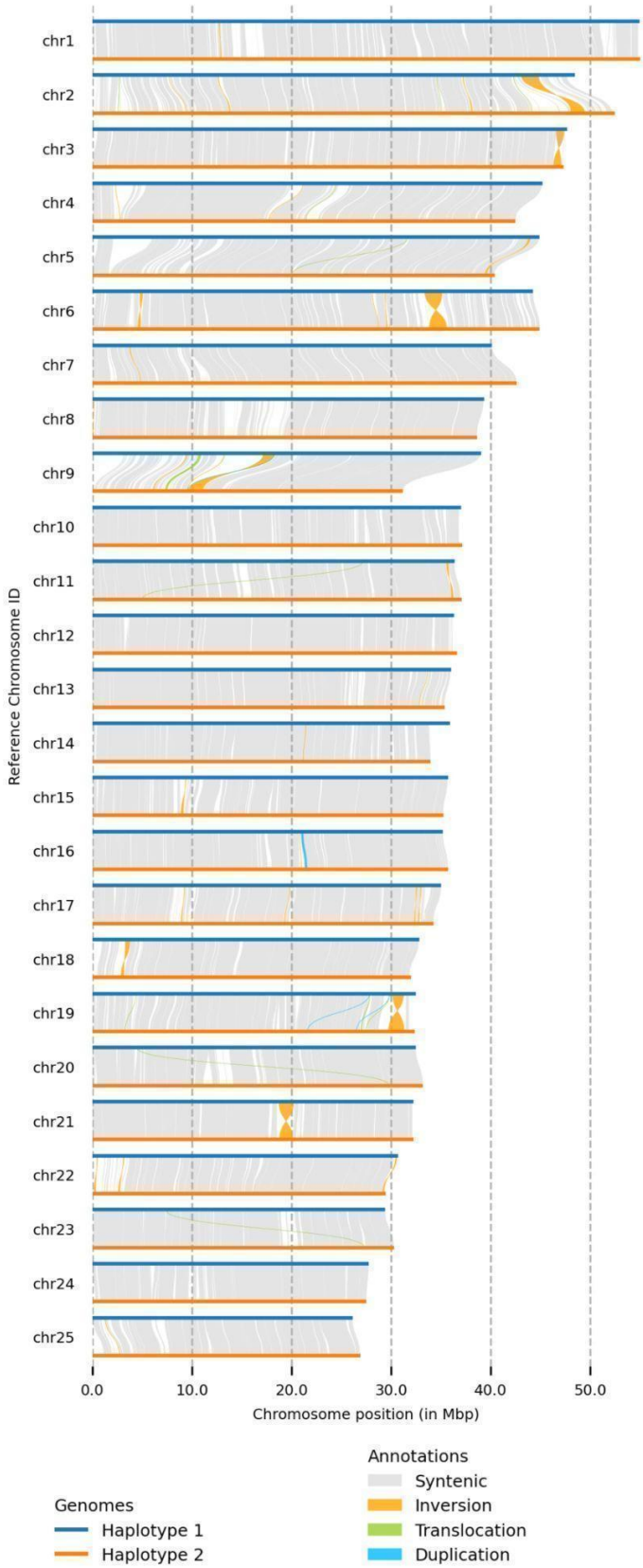


Figure 5: **Synteny plot of all 25 chromosomes of the *P. phoxinus* Hap1 (top, blue line) and Hap2 (bottom, orange line) of assemblies**. Aligned regions between both haplomes are shown in blocks of grey, unaligned regions are shown in gaps of white. Inversions are represented in orange, translocations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

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

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

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

563 Inferring gene orthology from sequence data is a fundamental approach for understanding
564 evolutionary relationships between genes and species through their evolutionary history (Tekaia, 2016).
565 A total of 396,455 genes across all 11 species described in the section on annotation of protein-coding
566 genes were included and 381,263 (96.12%) of these genes were successfully assigned to 27,672
567 orthogroups, i,e. gene sets in those species or in a subset of these species, that derived from their
568 respective last common ancestor (Supplementary Table S12). The total number of orthogroups
569 comprising protein-coding genes from all species was 14,914 (Table 5). Single-copy orthogroups, also
570 known as 1-to-1 orthogroups, contain genes that have evolved from a common ancestral gene and are
571 present in a single copy across all proteomes used in an ortholog analysis (Creevey et al., 2011; Simão et
572 al., 2015; Tekaia, 2016). Our orthology inference analysis revealed the presence of 301 single-copy
573 orthogroups which provided a solid basis for subsequent phylogenetic studies.

574 **Table 5.** 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

575

576 There were 273 genes distributed across 72 *P. phoxinus*-specific orthogroups; a sequence similarity
577 search of these genes to the TrEMBL database returned results for only 130 of these genes, many of which

were uncharacterised proteins (Supplementary Table S13). Only 54 of the successfully mapped genes annotated in the TrEMBL database returned GO annotations, which clustered into eight major GO terms (Figure 6). The presence of GO terms "response to oxidative stress" and "regulation of response to reactive oxygen", suggests that *P. phoxinus* has adaptive potential to changes in water quality. Observational studies have found conflicting evidence, with Eurasian minnows ranging from sensitive (Bagge & Hakkari, 1992) to insensitive to intermediate pollution (King et al., 2011). Those findings may be attributed to the fact that different species of *Phoxinus* were studied. Cyprinoidea in general have been thought to be sensitive to oxidative stress, demonstrated in Amur minnow (*Phoxinus lagowskii*) (Q. Wang et al., 2019; Yao et al., 2022) and Chinese rare minnow (*Gobiocypris rarus*) (Z.-H. Li et al., 2014), while the fathead minnow (*Pimephales promelas*) has been used as a bioindicator of water quality (Elizalde-Velázquez et al., 2020).

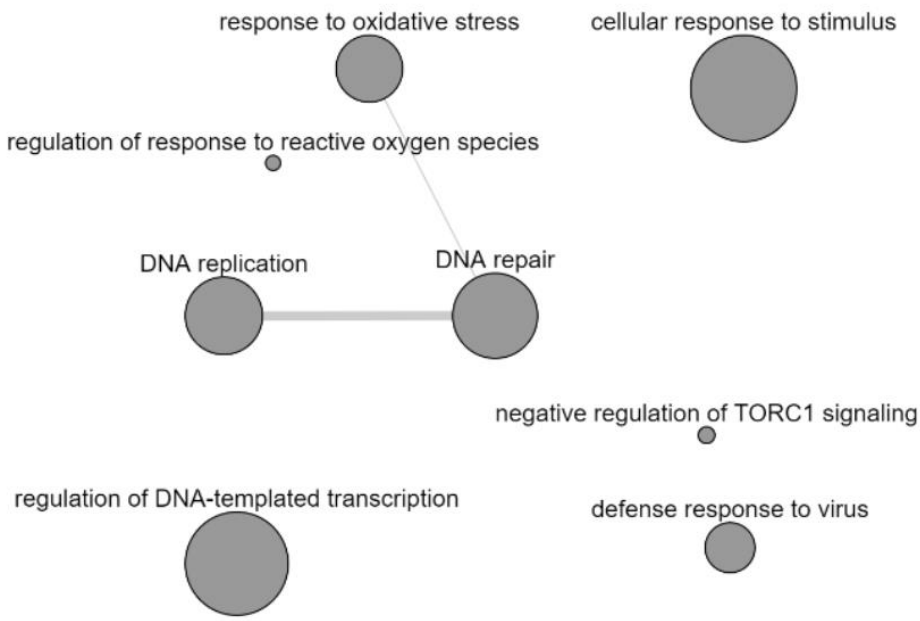
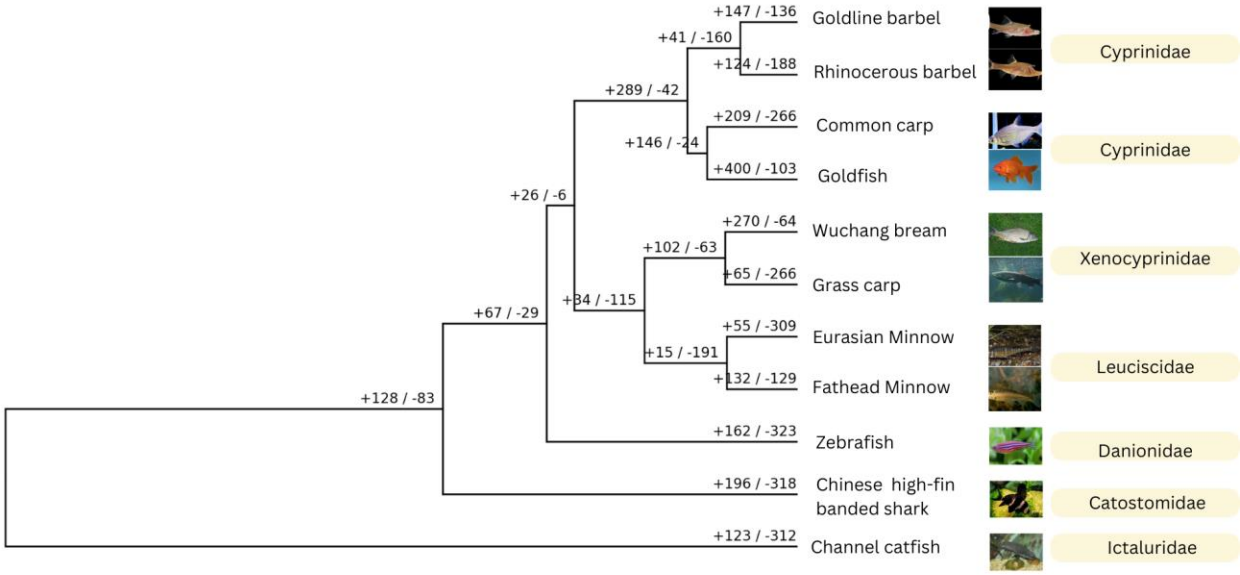


Figure 6: Gene ontology (GO) network of 54 functionally annotated *P. phoxinus*-specific genes in eight major GO terms: "DNA replication", "DNA repair", "regulation of DNA-templated transcription", "response to oxidative stress", "defense response to virus", "cellular response to stimulus", "regulation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

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

We analysed gene family expansions and contractions in *P. phoxinus* relative to 10 other teleost species (Supplementary Figure S8). Since the split of *P. phoxinus* and the fathead minnow from their last common ancestor, *P. phoxinus* counted more significant gene contractions (309 gene families) than

607 significant gene expansions (55 gene families) relative to the fathead minnow, which had 132 gene family
608 expansions and 129 gene family contractions (Figure 7). The significantly expanded gene families
609 contained genes such as *HIST1, H2B*, *CENP-T*, *GCRV*, *HIST2H2A*, *HIST2H3A*, *H3F3A*, *TRBV25-1*, which mostly
610 belong to histone PFAM domains (Supplementary Figure S9). Immune genes belonging to the NACHT
611 PFAM domain were also more abundant in the set of contracted gene families (Supplementary Figure S10,
612 genes like *NLRP12, NLRP3, CD22, CLEC17A, GIMAP7)*. The expanding immune gene families mostly
613 belonged to histone gene families, whereas the contracted gene families were mostly represented by
614 NLRC gene families (Supplementary Tables S14 and S15).



615

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

620 Evolution of the immune system has greatly contributed to speciation in teleosts and a lot of diversity
621 in their immune system pathways has been identified, like the expanded MHC1 and TLE genes in the
622 Atlantic cod (*Gadus morhua*) (Star et al., 2011) and the absence of MHC class II-mediated immunity in the
623 broad-nosed pipefish (*Syngnathus typhle)* (Haase et al., 2013). In the *P. phoxinus* genome, we observed
624 an expansion in immunity-boosting histones *H1, H2B, H3*, and *H4,* particularly *H2A* and *H2B.* Histones are
625 highly conserved proteins involved in the innate host defence against microbes (X. Li et al., 2022); they
626 include core histones like *H2A, H2B, H3*, and *H4*. The final histone type, called *H1,* is generally referred to
627 as a linker histone (X. Li et al., 2022; Westman et al., 2015). Histones are known to play a key role in
628 antibacterial immunity across multiple species (X. Li et al., 2022; Pasupuleti et al., 2012). Histone 2A and
629 2B have antibacterial activity against different types of bacteria in fish. Histones play a distinctive role in
630 mucosal immunity in fish (Ángeles Esteban & Cerezuela, 2015), acting as a strong barrier to microbial
631 invasion. This has been confirmed in the Indian major carp (*Labeo catla*) (Nigam et al., 2017), lumpfish
632 (*Cyclopterus lumpus*) (Patel & Brinchmann, 2017), gilthead seabream (*Sparus aurata*) (Jurado et al., 2015),
633 and rainbow trout (*Oncorhynchus mykiss*) (Fernandes et al., 2002; Oscoz et al., 2005). The expansion of
634 histone genes in *P. phoxinus* suggests a strong focus on mucosal protection against bacteria and other
635 pathogens. Assuming, that similar evolutionary trends can be found in other *Phoxinus* species, this is a
636 particularly interesting find, given that there are populations of *Phoxinus* spp. in the Baltic Sea, some of
637 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. Histones also play important roles in gene expression, DNA replication, and DNA damage repair (Best et al., 2018; Seal et al., 2022), in addition to their involvement in immune activity. It is important to note that histones are not restricted to immune activity alone. In fish, they have also been linked to spermatogenesis and can be an indicator of sperm quality (Herráez et al., 2017).

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 the 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*) to *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 number of genes involved in immune responsiveness might hint to a change in immune response strategy towards infection prevention. All the identified genetic diversity is 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).

Conclusion

In this paper, we present a highly contiguous and complete genome assembly for the Eurasian minnow, *Phoxinus phoxinus,* achieved by combining Pacbio Hifi reads and Hi-C data, along with a comprehensive repeats and gene annotation. Demographic history analysis and high observed heterozygosity and structural variation seem to suggest substantial population and individual diversity, respectively. While the identified sequence diversity could pose a challenge for a collapsed assembly, our phased genome enables a more accurate depiction of intra-individual variation. Consequently, we anticipate that 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

MSt devised the project. TOO, MSt, AB, and IC led the writing and devising of the manuscript. EWM, SW, and TB performed genome sequencing and assembly. IC established the genome repeat and protein-coding gene annotation pipeline with support from SM. TOO performed genome-wide comparisons, gene family evolution, and demographic history analyses. SK performed DNA extractions and provided laboratory expertise. SM and AB provided TOO and IC with bioinformatic support. AB contributed with genome assembly expertise and input on study design. All authors contributed to revising the manuscript.

Data, scripts, code, and supplementary information availability

Scripts used for genome assembly by EMW, SW and TB, and scripts used for assembly, genome-wide comparison, gene family evolution and demographic history by TOO are available here: https://doi.org/10.5281/zenodo.12191118.

The repeat and protein annotation were implemented using a snakemake pipeline. The code is available here: https://doi.org/10.5281/zenodo.11925110. RNA data for annotation is available under NCBI accession numbers SRR26699630 to SRR26699636. *P. phoxinus* haplomes and annotations are available under NCBI Bioproject number PRJNA1040855, which is 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 MSt, and by the German Research Foundation under INST 269/768-1 awarded to the technology platform of the CMCB at the TU Dresden.

References

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

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

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

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

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

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

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

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 auratus) genome and the evolution of genes after whole-genome duplication. *Science Advances*,

5(6), eaav0547. https://doi.org/10.1126/sciadv.aav0547

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), Article 2. https://doi.org/10.1038/s41592-020-01056-5

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., 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., … Mao, J.-F. (2021). Haplotype-resolved genome assembly and allele-specific gene expression in cultivated ginger. *Horticulture Research*, *8*, 188. https://doi.org/10.1038/s41438-021-00599-8

Corral-Lou, A., Perea, S., Aparicio, E., & Doadrio, I. (2019). Phylogeography and species delineation of the genus Phoxinus Rafinesque, 1820 (Actinopterygii: Leuciscidae) in the Iberian Peninsula. *Journal of Zoological Systematics and Evolutionary Research*, *57*(4), 926–941. https://doi.org/10.1111/jzs.12320

Creevey, C. J., Muller, J., Doerks, T., Thompson, J. D., Arendt, D., & Bork, P. (2011). Identifying Single Copy Orthologs in Metazoa. *PLOS Computational Biology*, *7*(12), e1002269. https://doi.org/10.1371/journal.pcbi.1002269

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Deininger, P. L., Moran, J. V., Batzer, M. A., & Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics & Development*, *13*(6), 651–658. https://doi.org/10.1016/j.gde.2003.10.013

Demuth, J. P., & Hahn, M. W. (2009). The life and death of gene families. *BioEssays*, *31*(1), 29–39. https://doi.org/10.1002/bies.080085

Denys, G. P. J., Dettai, A., Persat, H., Daszkiewicz, P., Hautecœur, M., & Keith, P. (2020). *Revision of 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

824 Flajnik, M. F., & Kasahara, M. (2010). Origin and evolution of the adaptive immune system: Genetic
825 events and selective pressures. *Nature Reviews Genetics*, *11*(1), Article 1.
826 https://doi.org/10.1038/nrg2703
827 Frost, W. E. (1943). The Natural History of the Minnow, Phoxinus phoxinus. *Journal of Animal Ecology*,
828 *12*(2), 139–162. https://doi.org/10.2307/1374
829 Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2023). *BRAKER3:*
830 *Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP,*
831 *AUGUSTUS and TSEBRA* (p. 2023.06.10.544449). bioRxiv.
832 https://doi.org/10.1101/2023.06.10.544449
833 Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: Transcript selector for
834 BRAKER. *BMC Bioinformatics*, *22*(1), 566. https://doi.org/10.1186/s12859-021-04482-0
835 Gao, B., Shen, D., Xue, S., Chen, C., Cui, H., & Song, C. (2016). The contribution of transposable elements
836 to size variations between four teleost genomes. *Mobile DNA*, *7*(1), 4.
837 https://doi.org/10.1186/s13100-016-0059-7
838 Goel, M., & Schneeberger, K. (2022). plotsr: Visualizing structural similarities and rearrangements
839 between multiple genomes. *Bioinformatics*, *38*(10), 2922–2926.
840 https://doi.org/10.1093/bioinformatics/btac196
841 Goel, M., Sun, H., Jiao, W.-B., & Schneeberger, K. (2019). SyRI: Finding genomic rearrangements and
842 local sequence differences from whole-genome assemblies. *Genome Biology*, *20*(1), 277.
843 https://doi.org/10.1186/s13059-019-1911-0
844 Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing
845 haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898.
846 https://doi.org/10.1093/bioinformatics/btaa025
847 Guevara, E. E., Webster, T. H., Lawler, R. R., Bradley, B. J., Greene, L. K., Ranaivonasy, J., Ratsirarson, J.,
848 Harris, R. A., Liu, Y., Murali, S., Raveendran, M., Hughes, D. S. T., Muzny, D. M., Yoder, A. D.,
849 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

902 medicinal value. *The Plant Journal*, *111*(5), 1340–1353. https://doi.org/10.1111/tpj.15892

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

906 Jurado, J., Fuentes-Almagro, C. A., Guardiola, F. A., Cuesta, A., Esteban, M. Á., & Prieto-Álamo, M.-J.
907 (2015). Proteomic profile of the skin mucus of farmed gilthead seabream (Sparus aurata).
908 *Journal of Proteomics*, *120*, 21–34. https://doi.org/10.1016/j.jprot.2015.02.019

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

913 Karjalainen, J., Urpanen, O., Keskinen, T., Huuskonen, H., Sarvala, J., Valkeajärvi, P., & Marjomäki, T. J.
914 (2016). Phenotypic plasticity in growth and fecundity induced by strong population fluctuations
915 affects reproductive traits of female fish. *Ecology and Evolution*, *6*(3), 779–790.
916 https://doi.org/10.1002/ece3.1936

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

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

926 Knebelsberger, T., Dunz, A. R., Neumann, D., & Geiger, M. F. (2015). Molecular diversity of Germany's
927 freshwater fishes and lampreys assessed by DNA barcoding. *Molecular Ecology Resources*, *15*(3),

562–572. https://doi.org/10.1111/1755-0998.12322

Kojima, K. K. (2019). Structural and sequence diversity of eukaryotic transposable elements. *Genes & Genetic Systems*, *94*(6), 233–252. https://doi.org/10.1266/ggs.18-00024

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Kottelat, M. (2007). Three new species of Phoxinus from Greece and southern France (Teleostei: Cyprinidae). *Ichthyological Exploration of Freshwaters*, *18*(2), 145–162.

Kottelat, M., & Freyhof, J. (2007). *Handbook of European freshwater fishes* (Kottelat, Cornol, & Freyhof, Eds.). Kottelat.

Krabbenhoft, T. J., MacGuigan, D. J., Backenstose, N. J. C., Waterman, H., Lan, T., Pelosi, J. A., Tan, M., & Sandve, S. R. (2021). Chromosome-Level Genome Assembly of Chinese Sucker (Myxocyprinus asiaticus) Reveals Strongly Conserved Synteny Following a Catostomid-Specific Whole-Genome Duplication. *Genome Biology and Evolution*, *13*(9), evab190. https://doi.org/10.1093/gbe/evab190

Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., Porubsky, D., Kuhn, K., Mueller, K. A., Low, W. Y., Hiendleder, S., Fedrigo, O., Liachko, I., Hall, R. J., Phillippy, A. M., Eichler, E. E., Williams, J. L., Smith, T. P. L., Jarvis, E. D., … Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-020-20536-y

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

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., Hung, C.-M., Hung, H.-Y., Yang, M.-Y., Liang, W., Hsu, Y.-C., Yao, C.-T., Dong, L., Lin, K., & Li, S.-H. (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*,

1006 *48*(10), Article 10. https://doi.org/10.1038/ng.3645

1007 Mansfeld, B. N., Boyher, A., Berry, J. C., Wilson, M., Ou, S., Polydore, S., Michael, T. P., Fahlgren, N., &
1008 Bart, R. S. (2021). Large structural variations in the haplotype-resolved African cassava genome.
1009 *The Plant Journal*, *108*(6), 1830–1848. https://doi.org/10.1111/tpj.15543

1010 Martinson, J. W., Bencic, D. C., Toth, G. P., Kostich, M. S., Flick, R. W., See, M. J., Lattier, D., Biales, A. D.,
1011 & Huang, W. (2022). De Novo Assembly of the Nearly Complete Fathead Minnow Reference
1012 Genome Reveals a Repetitive but Compact Genome. *Environmental Toxicology and Chemistry*,
1013 *41*(2), 448–461. https://doi.org/10.1002/etc.5266

1014 Matuszewski, S., Hermisson, J., & Kopp, M. (2015). Catch Me if You Can: Adaptation from Standing
1015 Genetic Variation to a Moving Phenotypic Optimum. *Genetics*, *200*(4), 1255–1274.
1016 https://doi.org/10.1534/genetics.115.178574

1017 Mendes, F. K., Vanderpool, D., Fulton, B., & Hahn, M. W. (2021). CAFE 5 models variation in evolutionary
1018 rates among gene families. *Bioinformatics*, *36*(22–23), 5516–5518.
1019 https://doi.org/10.1093/bioinformatics/btaa1022

1020 Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A Roadmap for Understanding the
1021 Evolutionary Significance of Structural Genomic Variation. *Trends in Ecology & Evolution*, *35*(7),
1022 561–572. https://doi.org/10.1016/j.tree.2020.03.002

1023 Miller, J. M., Malenfant, R. M., David, P., Davis, C. S., Poissant, J., Hogg, J. T., Festa-Bianchet, M., &
1024 Coltman, D. W. (2014). Estimating genome-wide heterozygosity: Effects of demographic history
1025 and marker type. *Heredity*, *112*(3), Article 3. https://doi.org/10.1038/hdy.2013.99

1026 Mills, C. A. (1987). The life history of the minnow Phoxinus phoxinus (L.) in a productive stream.
1027 *Freshwater Biology*, *17*(1), 53–67. https://doi.org/10.1111/j.1365-2427.1987.tb01028.x

1028 Mills, C. A. (1988). The effect of extreme northerly climatic conditions on the life history of the minnow,
1029 Phoxinus phoxinus (L.). *Journal of Fish Biology*, *33*(4), 545–561. https://doi.org/10.1111/j.1095-
1030 8649.1988.tb05498.x

1031 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

Waldbieser, G. C., Liu, S., Yuan, Z., Older, C. E., Gao, D., Shi, C., Bosworth, B. G., Li, N., Bao, L., Kirby, M.

1188 A., Jin, Y., Wood, M. L., Scheffler, B., Simpson, S., Youngblood, R. C., Duke, M. V., Ballard, L.,
1189 Phillippy, A., Koren, S., & Liu, Z. (2023). Reference genomes of channel catfish and blue catfish
1190 reveal multiple pericentric chromosome inversions. *BMC Biology*, *21*(1), 67.
1191 https://doi.org/10.1186/s12915-023-01556-8
1192 Wang, F., Wang, L., Liu, D., Gao, Q., Nie, M., Zhu, S., Chao, Y., Yang, C., Zhang, C., Yi, R., Ni, W., Tian, F.,
1193 Zhao, K., & Qi, D. (2022). Chromosome-level assembly of Gymnocypris eckloni genome.
1194 *Scientific Data*, *9*(1), Article 1. https://doi.org/10.1038/s41597-022-01595-w
1195 Wang, Q., Luo, S., Ghonimy, A., Chen, Y., Guo, Z., Liu, H., & Zhang, D. (2019). Effect of dietary l-carnitine
1196 on growth performance and antioxidant response in Amur minnow (Phoxinus lagowskii
1197 Dybowskii). *Aquaculture Nutrition*, *25*(4), 749–760. https://doi.org/10.1111/anu.12892
1198 Wang, T., Liu, F., Tian, G., Secombes, C. J., & Wang, T. (2019). Lineage/species-specific expansion of the
1199 Mx gene family in teleosts: Differential expression and modulation of nine Mx genes in rainbow
1200 trout Oncorhynchus mykiss. *Fish & Shellfish Immunology*, *90*, 413–430.
1201 https://doi.org/10.1016/j.fsi.2019.04.303
1202 Watanabe, K., Koga, H., Nakamura, K., Fujita, A., Hattori, A., Matsuda, M., & Koga, A. (2014).
1203 Spontaneous germline excision of Tol1, a DNA-based transposable element naturally occurring
1204 in the medaka fish genome. *Genome*, *57*(4), 193–199. https://doi.org/10.1139/gen-2014-0011
1205 Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural
1206 variation: Insights from and for human disease. *Nature Reviews Genetics*, *14*(2), Article 2.
1207 https://doi.org/10.1038/nrg3373
1208 Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of
1209 structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*,
1210 *28*(6), 1203–1209. https://doi.org/10.1111/mec.15066
1211 Westman, J., Papareddy, P., Dahlgren, M. W., Chakrakodi, B., Norrby-Teglund, A., Smeds, E., Linder, A.,
1212 Mörgelin, M., Johansson-Lindbom, B., Egesten, A., & Herwald, H. (2015). Extracellular Histones
1213 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

1240 Zeng, Q., Zhou, Z., He, Q., Li, L., Pu, F., Yan, M., & Xu, P. (2023). Chromosome-level haplotype-resolved
1241 genome assembly for Takifugu ocellatus using PacBio and Hi-C technologies. *Scientific Data*,
1242 *10*(1), Article 1. https://doi.org/10.1038/s41597-023-01937-2
1243 Zhou, C., McCarthy, S., & Durbin, R. (2022). *YaHS: Yet another Hi-C scaffolding too*.
1244 https://doi.org/10.1101/2022.06.09.495093
1245 Zilli, L., Schiavone, R., Storelli, C., & Vilella, S. (2008). Molecular Mechanisms Determining Sperm Motility
1246 Initiation in Two Sparids (Sparus aurata and Lithognathus mormyrus). *Biology of Reproduction*,
1247 *79*(2), 356–366. https://doi.org/10.1095/biolreprod.108.068296
1248 Zilli, L., Schiavone, R., & Vilella, S. (2017). Role of protein phosphorylation/dephosphorylation in fish
1249 sperm motility activation: State of the art and perspectives. *Aquaculture*, *472*, 73–80.
1250 https://doi.org/10.1016/j.aquaculture.2016.03.0

1251

1252 **Appendix**

1253 Supplementary figures S1 - S8 and tables S2 to S5: https://zenodo.org/doi/10.5281/zenodo.10218893

1254

1255 Supplementary tables S1, S6 to S15: https://zenodo.org/doi/10.5281/zenodo.10220265

1256

1257 **Supplementary figures**

1258 - Figure S1: Heatmap of haplome 1 Hi-C assembly with darker blocks indicating higher intensity of
1259 sequence interaction
1260 - Figure S2: Heatmap of haplome 2 Hi-C assembly with darker blocks indicating higher intensity of
1261 sequence interaction
1262 - Figure S3: GenomeScope plots of 19, 23, 25 and 30 mer analysis
1263 - Figure S4: Repeats landscape of annotated repeats in haplome 2
1264 - Figure S5: Shared homologs between *P. phoxinus* and zebrafish, fathead minnow, goldfish, grass
1265 carp and common carp
1266 - Figure S6: Gene ontology of genes in Insertions/Deletions
1267 - Figure S7: Gene ontology of genes in Inversions
1268 - Figure S8: Expansion/contraction profile including all gene families (significant and non-
1269 significant families)
1270 - Figure S9: Pfam domain distribution of genes expanding in *P. phoxinus* genome
1271 - Figure S10: Pfam domain distribution of genes contracting in *P. phoxinus* genome

1272

1273

1274 **Supplementary tables**

1275 - Table S1: Mapping statistics of both haplomes
1276 - Table S2: Genome assembly statistics of 25 chromosomes of both haplomes
1277 - Table S3: Statistics of repeat sequences annotated in haplome 1
1278 - Table S4: Statistics of repeat sequences annotated in haplome 2
1279 - Table S5: Mapping statistics of RNA sequence data used for annotation of protein-coding genes
1280 in both haplomes
1281 - Table S6: Summary statistics of annotated protein-coding genes in haplome 1
1282 - Table S7: Summary statistics of annotated protein-coding genes in haplome 2
1283 - Table S8: List of genes present in insertion/deletions
1284 - Table S9: Gene ontology of genes present in insertions/deletions
1285 - Table S10: List of genes present in inversions
1286 - Table S11: Gene ontology of genes present in inversions
1287 - Table S12: Summary statistics of orthology between *P. phoxinus* proteomes and selected teleosts
1288 - Table S13: List of genes identified by orthology analysis as *P. phoxinus*-specific
1289 - Table S14: List of genes significantly expanding in *P. phoxinus* genome
1290 - Table S15: List of genes significantly contracting in *P. phoxinus* genome