scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of masu salmon (*Oncorhynchus masou masou*)

Baosheng Wu[1,4], Yepin Yu[1,4], Xiujuan Zhang[1,4], Huiqun Chen[1,2], Chanxia Qin[1], Jiajun Xie[1], Linmiao Li[1], Wenhua Wu[3] ✉ & Jinping Chen[1] ✉

Whole-genome duplication is a major evolutionary force, with salmonids providing a key model system. In this study, we generated a high-quality masu salmon genome using PacBio HiFi, ONT ultra-long and Hi-C sequencing technologies. The final phased assembly includes two haplotypes: HAP1 (2.5 Gb) and HAP2 (2.4 Gb). The contiguity of the assembly is excellent, with scaffold N50 values of 99.28 Mb (HAP1) and 98.82 Mb (HAP2). Notably, the 33 largest scaffolds in each haplotype cover over 99.6% of the genome, approaching chromosome-level scale. The genomic completeness was assessed at 98.3% (HAP1) and 97.7% (HAP2) via BUSCO analysis. Finally, we identified 41,858 and 40,970 protein-coding genes for HAP1 and HAP2, respectively. Furthermore, to enable research into reproductive strategies, we generated 90 transcriptome datasets from the semelparous masu salmon across key post-spawning time points, alongside a comparative dataset of 60 transcriptomes from the iteroparous rainbow trout. Together, these high-quality genomic and transcriptomic resources offer a powerful platform for investigating both post-WGD genome evolution and the molecular basis of semelparity in salmonids.

Background & Summary

Whole-genome duplication (WGD) is a major evolutionary force that has occurred repeatedly throughout the history of vertebrates. Key events include two early rounds shared by all vertebrates (1 R/2 R) and a third round specific to teleost fishes (Ts3R). While these ancient duplications, particularly the ~320 million-year-old Ts3R, have profoundly shaped teleost evolution, the extensive gene loss over time has obscured the evolutionary trajectories of many ohnologues (duplicate genes from WGD), limiting our understanding of their initial functional divergence[1]. In contrast, the more recent salmonid-specific WGD (Ss4R), which occurred approximately 80–95 million years ago, offers a far clearer window into this process[2,3] The high retention rate of ohnologues in salmonid genomes provides an exceptional model system to investigate the roles of WGD in vertebrate adaptation.

Following a duplication event, the resulting paralogs face several potential evolutionary fates[4–7]. Most commonly, one copy accumulates degenerative mutations and becomes a nonfunctionalized pseudogene, a process that resolves dosage redundancy. Alternatively, if the duplication is beneficial, both copies may be preserved by natural selection[4,6–8]. Preserved paralogs can then diverge functionally through two primary, non-exclusive pathways: subfunctionalization, where the ancestral functions are partitioned between the two copies, and neofunctionalization, where one copy acquires a novel function while the other retains the original role. A third possibility is conservation, where both duplicates are maintained to increase gene dosage or functional robustness.

The predominant evolutionary fate of ohnologues remains a topic of active debate, as it appears to vary depending on the specific WGD event and evolutionary lineage[9,10]. For instance, neofunctionalization following the 1 R/2 R events is thought to have been a key driver in the origin of critical vertebrate innovations, such as the liver and lungs[11,12]. In contrast, analyses of the teleost-specific Ts3R suggest that subfunctionalization was the more common outcome[13]. For the salmonid-specific Ss4R, however, emerging evidence points towards a

[1]Guangdong Key Laboratory of Animal Conservation and Resource Utilization, Institute of Zoology, Guangdong Academy of Sciences, Guangzhou, 510260, China. [2]College of Forestry, Soil and Water Conservation, Central South University of Forestry and Technology, Changsha, 410004, China. [3]Heilongjiang River Fisheries Research Institute, Chinese Academy of Fisheries Sciences, Harbin, 150070, China. [4]These authors contributed equally: Baosheng Wu, Yepin Yu, Xiujuan Zhang. ✉e-mail: wuwenhua2008@sina.com; chenjp@giz.gd.cn

	HAP1	HAP2	GCF_036934945.1
Genome size (bp)	2,593,837,940	2,506,373,981	2,663,207,833
Contig number	713	489	21,310
Scaffold number	205	51	17,263
BUSCO score	98.30%	97.70%	98.90%
Contig N50 (bp)	24,672,662	15,978,490	2,468,404
Scaffold N50 (bp)	99,276,845	98,821,334	66,338,746
Contig N90 (bp)	2,574,334	3,371,868	65,025
Scaffold N90 (bp)	51,503,236	49,353,328	107,514
HIFI reads depth (X)	61.06	63.20	0
ONT reads depth (X)	10.9	11.3	33

Table 1. Comparison of assembly statistics for the new masu salmon genome.

prominent role for neofunctionalization, potentially fueling the adaptive radiation of this group[1,14]. This makes salmonids a compelling system to test hypotheses about the link between WGD, neofunctionalization, and lineage-specific adaptation.

Oncorhynchus masou masou, commonly known as the masu salmon, is a Pacific salmon species endemic to East Asia, primarily inhabiting the waters of Japan, Korea, and the Russian Far East[15]. It holds significant economic value for regional aquaculture and fisheries, while also playing a critical ecological role in North Pacific ecosystems due to its diverse life history strategies, which include both anadromous (sea-run) and resident freshwater forms[16]. Consequently, *O. m. masou* is an invaluable model for understanding the genomic consequences of the Ss4R event. It offers unique insights into how gene duplication contributes to species diversification and environmental adaptation.

In this study, combining PacBio HiFi, ONT ultra-long and Phase Genomics Hi-C sequencing, we generate two haplotype genomes (HAP1 and HAP2) for masu salmon. The final assemblies include about 2.5 Gb for HAP1 in 200 scaffolds (scaffold N50 = 99.28 Mb) and 2.4 Gb for HAP2 in 50 scaffolds (scaffold N50 = 98.82 Mb) (Table 1). Both HAP1 and HAP2 are assembled into 33 largest scaffolds, covering 99.6% and 99.7% of their genomes, respectively, which is consistent with their chromosome numbers[17]. The assembled genome size is closer to others in the Salmonidae family (Fig. 1A–C). Using BUSCO analysis, we assessed the genomic completeness of the two haploid genomes, HAP1 and HAP2, which were determined to be 98.3% and 97.7%, respectively.

Our newly generated genome assembly is of substantially higher quality than the recently published reference genome (GCF_036934945.1)[17]. Based on the significantly higher sequencing depth, we successfully filled 20,811 gaps that were present in the reference assembly (Fig. 2A,B). This resulted in a marked enhancement of genome continuity and completeness (Table 1). We also assembled the complete Y chromosome using Hi-C sequence phasing, as opposed to the chimeric mix represented as a single chromosome. In addition, our genome assembly was one of the best (top 3) compared to the 30 other publicly available Salmonidae genomes to date (Fig. 2B and Table 1), given the presence of a significant number of repetitive sequences in the Salmonidae genomes, which poses a challenge to assembly continuity[1]. Additionally, by integrating transcriptome sequencing data with both *ab initio* and homology-based gene prediction approaches, we identified 41,927 protein-coding genes in HAP1 and 41,387 protein-coding genes in HAP2. In the following analyses, HAP2 was used as the reference genome.

Methods

Ethics statement. All experimental procedures involving masu salmon (*Oncorhynchus masou masou*) and rainbow trout (*Oncorhynchus mykiss*) were performed in strict accordance with the guidelines for the care and use of laboratory animals. The study protocol was reviewed and approved by the Animal Care and Use Committee of the Institute of Zoology, Guangdong Academy of Sciences (Approval No. GIZ20240315). All fish were handled following standard ethical guidelines to minimize stress and suffering.

Sample and genomic sequencing. Sequencing samples were obtained from the Cold Water Fish Experiment Station of Bohai Town, Mudanjiang City, Heilongjiang Province, China. All samples were from diploid, landlocked masu salmon (*Oncorhynchus masou masou*) broodstock, with one male selected for genome sequencing.

Genomic DNA was extracted from the selected male masu salmon using the Nanobind PanDNA Kit according to the manufacturer's protocol. The quality and integrity of the extracted DNA was assessed using a Nanodrop spectrophotometer (Thermo Scientific) to measure purity (A260/A280 ratio) and a Qubit fluorometer (Thermo Fisher Scientific) to accurately quantify DNA concentration. In addition, DNA integrity was assessed using gel electrophoresis and a fragment analyser (Agilent) to confirm the presence of high molecular weight DNA suitable for long read sequencing. HiFi libraries were prepared using the SMRTbell® Express Template Prep Kit 2.0 (Pacific Biosciences). Briefly, 5 µg of high molecular weight gDNA was sheared to a target size of approximately 15–20 kb using a Covaris g-TUBE (Covaris). The sheared DNA was then subjected to end repair and ligated with SMRTbell adapters according to the manufacturer's instructions. The resulting SMRTbell libraries were purified using the ProNex size selection system (Promega) to remove DNA fragments shorter than the target size.

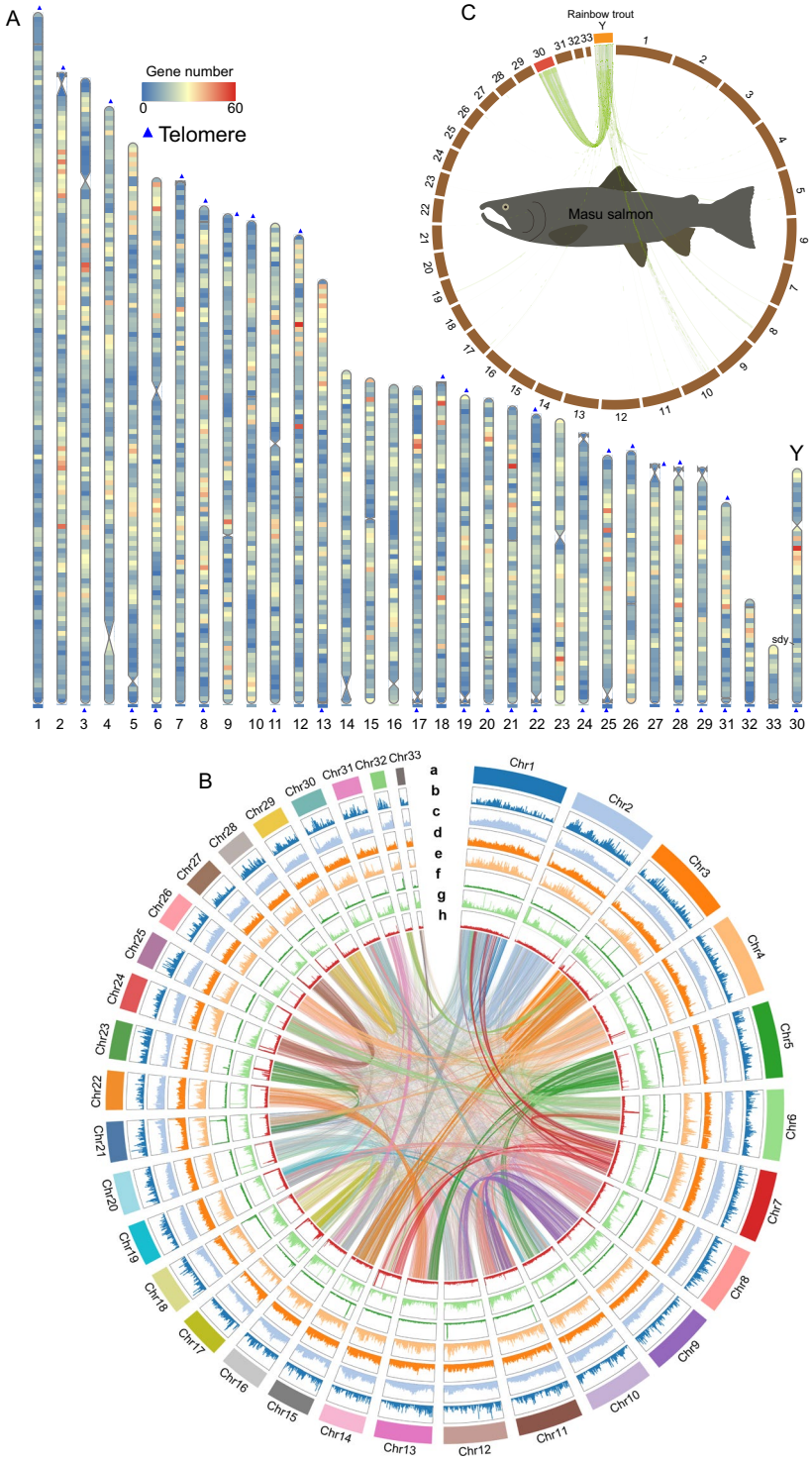

Fig. 1 Genome assembly of masu salmon (*Oncorhynchus masou masou*). (**A**) Chromosomal karyotype and distribution of genomic features. Heat maps show the density of protein-coding genes at 1 Mb intervals along each chromosome. Blue triangles at chromosomal ends indicate putatively identified telomeric regions, while indented regions indicate putative centromeric loci. (**B**) Circos plot illustrates the distribution of genomic components (1 Mb window size). From outer to inner rings: a: chromosomal karyotype; b: protein coding gene density; c: DNA content density; d: long interspersed nuclear element (LINE) density; e: long terminal repeat (LTR) retrotransposon density; f: short interspersed nuclear element (SINE) density; g: other repetitive elements; h: unclassified elements. Interchromosomal links represent syntenic relationships between the two haplotypes (Hap1 and Hap2). (**C**) Syntenic conservation between the Y chromosomes of Masu salmon and rainbow trout (Oncorhynchus mykiss). Homologous regions are connected by lines, highlighting the evolutionary conservation of Y chromosome architecture.

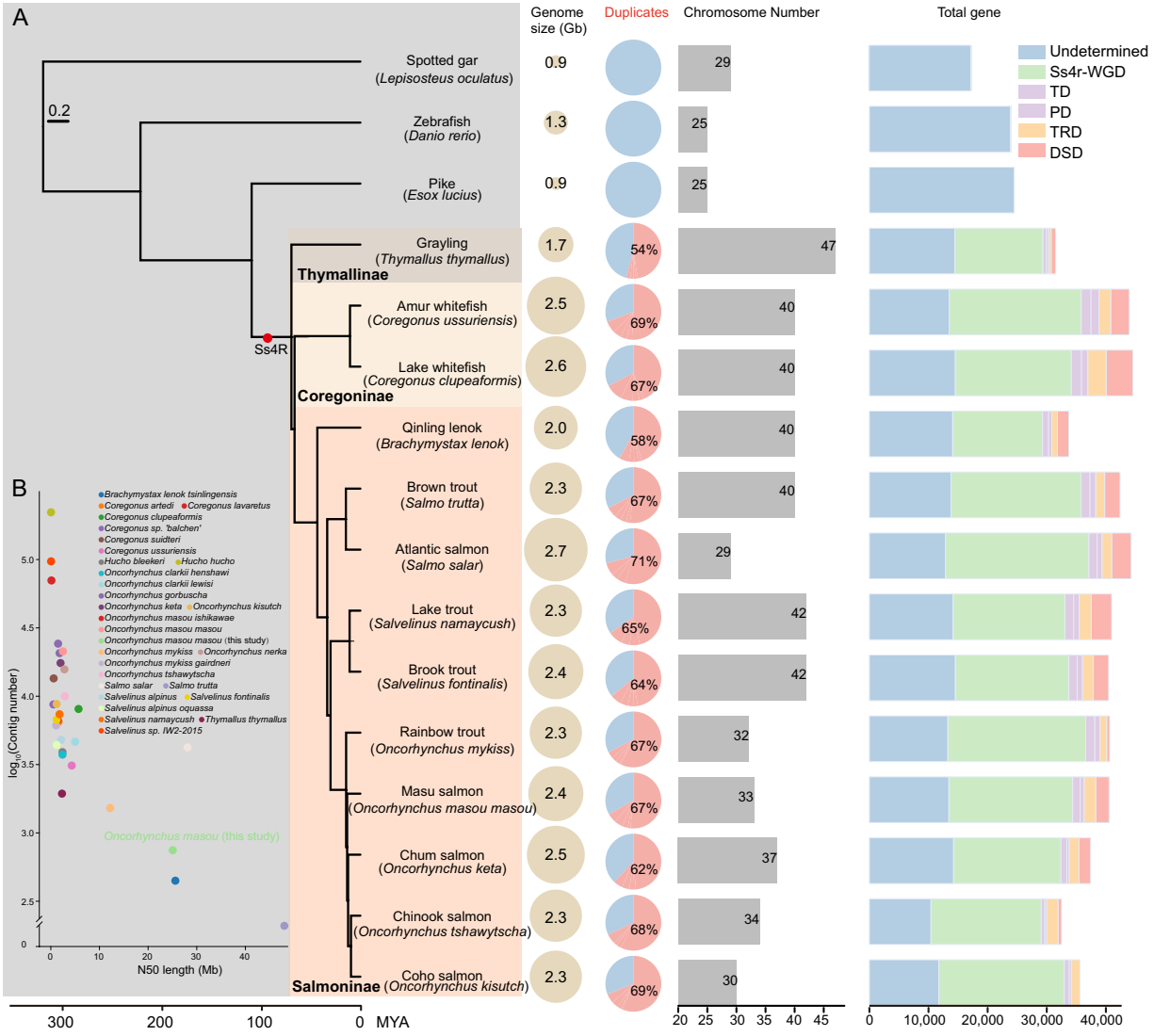

Fig. 2 Genomic landscape of the Salmonidae family. (**A**) Comparative genomic analysis of three Salmonidae subfamilies (Coregoninae, Thymallinae, and Salmoninae). Panels depict, from left to right: divergence time (Mya), genome size (Gb), gene type distribution, chromosome number, and the prevalence of different gene duplication types. Gene duplication mechanisms include whole-genome duplication (WGD), tandem duplication (TD), proximal duplication (PD), transposed duplication (TRD), and dispersed duplication (DSD). (B) Comparison of genome assembly statistics for representative Salmonidae species from public databases alongside the assembled genome of masu salmon from this study.

For ultra-long sequencing, libraries were prepared using the ONT Ligation Sequencing Kit (SQK-LSK114). Approximately 5 µg of DNA was end-repaired, A-tailed and ligated with ONT adapters. Size selection was performed using AMPure XP beads with a low bead-to-sample ratio (0.3×) to retain ultra-long fragments (>50 kb). Sequencing was performed on a PromethION 48 platform using an R10.4.1 flow cell for 72 hours. Base calling was performed using Guppy Base Caller (v6.x) in high accuracy mode. Reads were filtered for quality using Filtlong, and sequencing metrics such as N50 and total yield were assessed using NanoPlot and NanoStat.

Hi-C libraries were prepared using the Arima HiC Kit (Arima Genomics) according to the manufacturer's protocols. Briefly, cross-linked chromatin was first digested with a restriction enzyme cocktail provided in the kit, followed by biotin labelling of the resulting DNA ends. The resulting DNA fragments were then purified and the biotin-labelled fragments were enriched using streptavidin-coated beads. The enriched DNA was then sheared to a target fragment size of 300–500 bp using a Covaris S220 ultrasonicator, followed by library preparation using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs). A PCR amplification step was performed to generate the final Hi-C library. The prepared Hi-C libraries were sequenced on an Illumina NovaSeq 6000 platform in paired-end mode (150 bp reads) to achieve a target depth of 100× genome coverage. The quality of the sequencing reads was assessed using FastQC.

RNA sequencing. To give a resource for the molecular mechanisms underlying semelparity, we performed transcriptome sequencing (RNA-Seq) on ten different tissues (forebrain, midbrain, hindbrain, heart, liver, kidney, eye, muscle, ovary, testis) from the masu salmon collected at three critical time points relative to spawning: pre-mating, 24 hours post-mating, and 96 hours post-mating. For a comparative analysis with an iteroparous (multiple-spawning) reproductive strategy, we generated a parallel dataset from rainbow trout (*Oncorhynchus mykiss)*. This involved collecting the same ten tissues at two time points: immediately before spawning and 24 hours post-spawning, resulting in an additional 60 samples for sequencing.

RNA-seq libraries were prepared using the TruSeq Stranded mRNA Library Prep Kit (Illumina) according to the manufacturer's protocol. Library quality and quantity were assessed using a Bioanalyzer (Agilent) and qPCR, respectively, prior to sequencing. Sequencing was performed on an Illumina NovaSeq. 6000 platform using paired-end reads (150 bp). Raw sequencing reads were first pre-processed to remove adapters and low quality bases using Fastp[18] (v0.20.1). The clean reads were then aligned to the reference genome using HISAT2[19] (v2.2.1), which implements a fast and efficient alignment method based on the Burrows-Wheeler transform. Alignment was performed using default settings, allowing a maximum of two mismatches and preserving the strandedness of the RNA-seq data. After alignment, gene expression levels were quantified using StringTie[20] (v2.1.5). This tool was used to assemble the transcriptome, estimate gene expression levels and generate raw gene counts for downstream analysis. Transcript-level counts were then aggregated to gene-level counts, which were used to calculate expression levels in TPM (Transcripts Per Million).

Genome assembly and annotation. The genome assembly process was carried out by integrating multiple sequencing technologies to produce a high-quality chromosome-level assembly. Initially, a contig-level draft genome was generated using hifiasm version 0.20.0-r639[21] with default parameters. This step used HiFi reads to ensure base-level accuracy, ONT ultra-long reads to resolve repetitive regions, and Hi-C reads to provide essential chromosomal structural information. The contigs were then scaffolded to the chromosome level using HapHiC[22], which used Hi-C data to establish chromosomal linkages and orientations. To improve accuracy, potential misassembles or misplacements of contigs were manually corrected using Juicebox[23], a visualization tool for Hi-C contact maps. Finally, quarTeT v1.2.1[24] was used to fill gaps within the scaffolds and predict centromeric regions, improving both the continuity and structural integrity of the genome assembly. Finally, we used BUSCO v5.5.0[25] with the "actinopterygii_odb10" library to analyses and evaluate the completeness of our genome.

Repeat sequence annotation was performed using RepeatModeler[26] for *de novo* identification of repeat families, followed by RepeatMasker[27] to mask repeat sequences using both the generated repeat library and the Repbase databases. Annotations were cross-validated using bedtools[28] to summarize the repeat distribution across the genome, ensuring accurate identification and masking of repeat elements in the final assembly. Finally, gene structure annotation was performed using a combination of *ab initio* prediction, homology-based methods and RNA-seq evidence.

Phylogenetic analysis and divergence time estimation. A phylogenetic tree was constructed using 1:1 orthologous gene sequence from 16 species. This dataset included the spotted gar (*Lepisosteus oculatus*), zebrafish (*Danio rerio*), and Northern pike (*Esox lucius*) as outgroups, alongside 13 species from the family Salmonidae (*Thymallus thymallus*, *Coregonus ussuriensis*, *Coregonus clupeaformis*, *Brachymystax lenok*, *Salmo trutta*, *Salmo salar*, *Salvelinus namaycush*, *Salvelinus fontinalis*, *Oncorhynchus mykiss*, *Oncorhynchus masou masou*, *Oncorhynchus keta*, *Oncorhynchus tshawytscha*, and *Oncorhynchus kisutch*).

The sequences for each gene were aligned and concatenated to generate a supermatrix of orthologous genes. The phylogenetic tree was inferred using RAxML (v8.2.12)[29] with default parameters under the GTR + G model, utilizing fourfold degenerate synonymous sites (4DTv). Divergence times were subsequently estimated using the MCMCtree program (v4.9), part of the PAML package. Three fossil calibration points from TimeTree were used to constrain node ages: the divergence between spotted gar and zebrafish (180–251 Ma), zebrafish and Northern pike (96–121 Ma), and Northern pike and grayling (51–99 Ma). These calibration points were incorporated into the MCMCtree analysis using a relaxed molecular clock model with Markov Chain Monte Carlo (MCMC) sampling.

Identification and classification of gene duplications. Gene duplications were classified into five categories based on their origin: whole genome duplication (WGD), proximal duplication (PD), transposed duplication (TRD), dispersed duplication (DSD), and tandem duplication (TD). The genomes of 14 salmonid species were analyzed using DupGen_finder[30] to identify these duplication modes (Table 2).

Synteny analysis between different species. We conducted a comprehensive synteny analysis using the JCVI pipeline[31] (Fig. 3). We compared our newly assembled masu salmon genomes against a curated set of species representing key evolutionary nodes: fellow salmonids (rainbow trout, Qinling lenok, grayling), a pre-Ss4R outgroup (Northern pike), more distant teleosts (zebrafish, stickleback), and the non-teleost spotted gar as an ancient anchor. Homologous gene pairs, identified via an all-vs-all DIAMOND protein search (E-value $< 1e^{-5}$), were used as anchors to delineate syntenic blocks. To ensure high-confidence results, we applied stringent post-filtering, retaining only blocks that contained a minimum of four collinear genes and were defined by anchor pairs with a sequence identity of 75% or greater[11].

Data overview. The dataset includes the classification of duplicated genes for all analyzed species. Figure 2A and Table 2 provides a summary of these results, illustrating the proportion of five duplication categories (WGD, TD, PD, TRD, DSD).

	WGD	TD	PD	TRD	DSD
Hucho hucho	9,889	909	304	3,951	4,438
Coho salmon	18,926	1,467	958	2,003	3,451
Oncorhynchus masou masou	20,942	1,267	674	1,970	2,235
Oncorhynchus keta	18,147	1,002	406	1,651	1,970
Salmo salar	24,254	1,400	844	1,639	3,260
Brachymystax tsinlingensis	15,203	1,016	539	954	1,917
Oncorhynchus mykiss	23,314	1,532	905	1,154	467
Salmo trutta	22,013	1,504	970	1,531	2,577
Coregonus ussuriensis	19,615	1,723	1,080	3,110	4,500
Coregonus clupeaformis	19,893	1,610	881	2,715	2,540
Oncorhynchus tshawytscha	18,639	611	325	1,936	572
Oncorhynchus kisutch	21,255	808	369	1,433	2,078
Coregonus ussuriensis	22,327	1,643	1,352	2,000	3,117
Salvelinus fontinalis	19,196	1,399	919	1,825	2,582

Table 2. Summary of gene duplication categories and gene counts identified in salmonid genomes. WGD: whole genome duplication, PD: proximal duplication, TRD: transposed duplication, DSD: dispersed duplication, TD: tandem duplication.

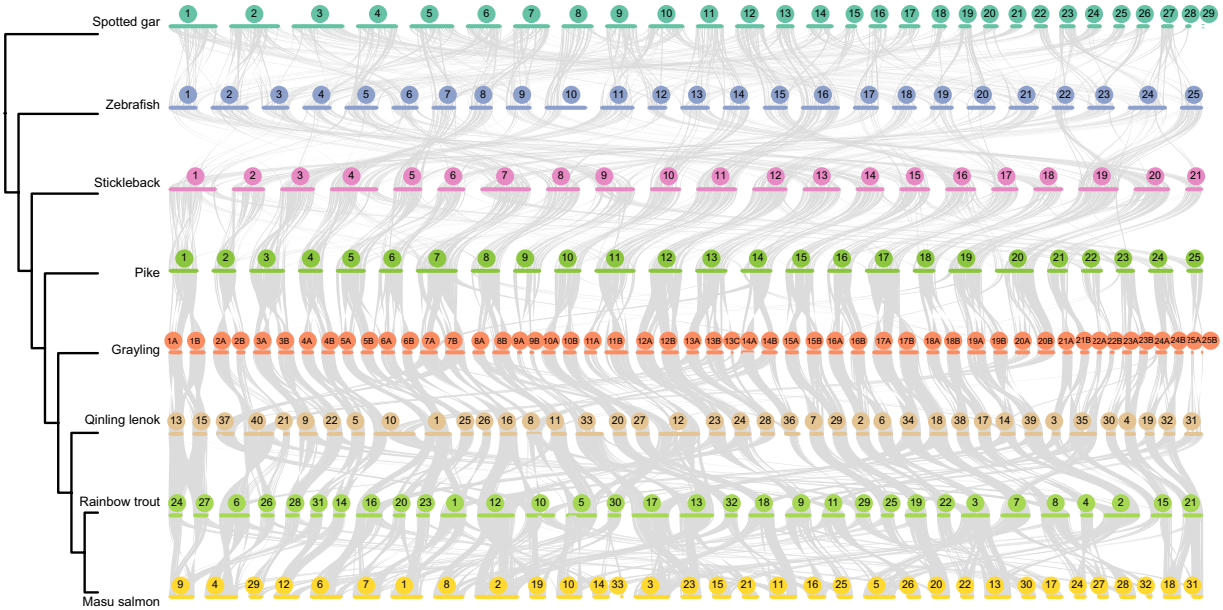

Fig. 3 Comparative synteny analysis between salmonid fishes and phylogenetically related species.

Data Records

All raw data generated in this study have been deposited in the China National GeneBank DataBase (CNGBdb) under accession number CNP0006926 and in the National Center for Biotechnology Information (NCBI) under accession number PRJNA1226994. Raw sequencing data, including HiFi, Hi-C and RNA-seq reads, are available in the NCBI Sequence Read Archive with accession SRR32456020- SRR32456171 and SRR32553755- SRR32553758[32].

HiFi and Hi-C raw reads have been deposited in the NCBI Sequence Read Archive SRR32456170, SRR32456171, SRR32553755-SRR32553758[32]; RNA-seq raw reads have been deposited in the NCBI Sequence Read Archive SRR32456020-SRR32456169[32]. The genome assembly of masu salmon has been submitted to NCBI GenBank with the accession number JBNQWP000000000[33] and JBNQWQ000000000[34]. In addition, the annotation results are accessible in the Figshare database. The genome, annotation files were also deposited in Figshare (https://figshare.com/articles/dataset/masu_salmon/30184666)[35].

Technical Validation

To validate the quality of our assembly, we performed two assessments. Beyond the high completeness scores from BUSCO analysis, we confirmed the assembly's accuracy by mapping the original PacBio HiFi reads to the HAP1 haplotype. This alignment yielded a mapping rate of 99.35%, indicating excellent structural integrity. To assess the assembly consistency, we performed synteny analysis. Figure 3 visualizes the conserved syntenic blocks between the masu salmon assembly and reference species, confirming the structural integrity of the genome assembly.

Data availability

The raw sequencing reads generated during the current study are available in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA1226994 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1226994; https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA1226994) and in the China National GeneBank DataBase (CNGBdb) under the accession number CNP0006926 (https://db.cngb.org/data_resources/project/CNP0006926). The final genome assemblies have been deposited in NCBI GenBank under accession numbers JBNQWP000000000 (https://www.ncbi.nlm.nih.gov/nuccore/JBNQWP000000000.1) and JBNQWQ000000000 (https://www.ncbi.nlm.nih.gov/nuccore/JBNQWQ000000000). The complete dataset, including the final genome assembly and annotation files, is also available in the Figshare repository at https://doi.org/10.6084/m9.figshare.30184666.

Code availability

No specific script was used in this work. The corresponding bioinformatics software and the specific versions of software have been described in Methods.

Received: 30 September 2025; Accepted: 19 February 2026;
Published online: 26 February 2026

References

1. Lien, S. *et al.* The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205, https://doi.org/10.1038/nature17164 (2016).
2. Near, T. J. *et al.* Resolution of ray-finned fish phylogeny and timing of diversification. *Proc Natl Acad Sci USA* **109**, 13698–13703, https://doi.org/10.1073/pnas.1206625109 (2012).
3. Macqueen, D. J. & Johnston, I. A. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc Biol Sci* **281**, 20132881, https://doi.org/10.1098/rspb.2013.2881 (2014).
4. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155, https://doi.org/10.1126/science.290.5494.1151 (2000).
5. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Bio* **3**, research0008.0001, https://doi.org/10.1186/gb-2002-3-2-research0008 (2002).
6. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545, https://doi.org/10.1093/genetics/151.4.1531 (1999).
7. Birchler, J. A. & Yang, H. The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *Plant Cell* **34**, 2466–2474, https://doi.org/10.1093/plcell/koac076 (2022).
8. Li, J.-T. *et al.* The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (*Cyprinus carpio*). *Sci Rep* **5**, 8199, https://doi.org/10.1038/srep08199 (2015).
9. Sandve, S. R., Rohlfs, R. V. & Hvidsten, T. R. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat Genet* **50**, 908–909, https://doi.org/10.1038/s41588-018-0162-4 (2018).
10. Braasch, I., Bobe, J., Guiguen, Y. & Postlethwait, J. H. Reply to: 'Subfunctionalization versus neofunctionalization after whole-genome duplication'. *Nat Genet* **50**, 910–911, https://doi.org/10.1038/s41588-018-0163-3 (2018).
11. Wu, B. *et al.* Single-cell analysis of the amphioxus hepatic caecum and vertebrate liver reveals genetic mechanisms of vertebrate liver evolution. *Nature Ecol Evol* **8**, 1972–1990, https://doi.org/10.1038/s41559-024-02510-9 (2024).
12. Li, Y. *et al.* Origin and stepwise evolution of vertebrate lungs. *Nat Ecol Evol*, https://doi.org/10.1038/s41559-025-02642-6 (2025).
13. Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet* **48**, 427–437, https://doi.org/10.1038/ng.3526 (2016).
14. Robertson, F. M. *et al.* Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol* **18**, 111, https://doi.org/10.1186/s13059-017-1241-z (2017).
15. Yu, J.-N. *et al.* New polymorphic microsatellite markers for the masu salmon (Oncorhynchus masou masou) from Korea and their application to wild and hatchery populations. *Biochem Syst Ecol* **61**, 260–267, https://doi.org/10.1016/j.bse.2015.06.019 (2015).
16. Kendall, N. W. *et al.* Anadromy and residency in steelhead and rainbow trout (*Oncorhynchus mykiss*): a review of the processes and patterns. *Can J Fish Aquat Sci.* **72**, 319–342, https://doi.org/10.1139/cjfas-2014-0192 (2015).
17. Christensen, K. A. *et al.* Masu salmon species complex relationships and sex chromosomes revealed from analyses of the masu salmon (*Oncorhynchus masou masou*) genome assembly. *G3 Genes|Genomes|Genetics* **15**, https://doi.org/10.1093/g3journal/jkae278 (2024).
18. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
19. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
20. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
21. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
22. Zeng, X. *et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nature plants* **10**, 1184–1200, https://doi.org/10.1038/s41477-024-01755-3 (2024).
23. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
24. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res.* **10**, https://doi.org/10.1093/hr/uhad127 (2023).
25. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
26. Saha, S., Bridges, S., Magbanua, Z. V. & Peterson, D. G. Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res.* **36**, 2284–2294, https://doi.org/10.1093/nar/gkn064 (2008).
27. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **25**, 4–10, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
28. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, https://doi.org/10.1093/bioinformatics/btq033 (2010).

29. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313, https://doi.org/10.1093/bioinformatics/btu033 (2014).
30. Qiao, X. *et al*. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20**, 38, https://doi.org/10.1186/s13059-019-1650-2 (2019).
31. Tang, H. *et al*. Synteny and collinearity in plant genomes. *Science* **320**, 486–488, https://doi.org/10.1126/science.1153917 (2008).
32. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP565699 (2025).
33. Wu, B. *Oncorhynchus masou masou*, whole genome shotgun sequencing project. Haplotype 1. *Genbank* https://identifiers.org/ncbi/insdc:JBNQWP000000000 (2025).
34. Wu, B. Oncorhynchus masou masou, whole genome shotgun sequencing project. Haplotype 2. *Genbank* https://identifiers.org/ncbi/insdc:JBNQWQ000000000 (2025).
35. Wu, B. *et al*. Assembly and annotation files of (*Oncorhynchus masou masou*). figshare. Dataset. https://doi.org/10.6084/m9.figshare.30184666.v1.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32500370), Project of Guangdong Academy of Sciences Program (No. 2024GDASZH-2024010101, 2022GDASZH-2022010106), Science and Technology Program of Guangzhou (2025A04J3507), the Natural Science Fund of Guangdong Province (No. 2023A1515011429).

Author contributions

Wenhua Wu and Jinping Chen conceived and designed the study. Baosheng Wu performed data analysis and drafted the manuscript. Yepin Yu, Chanxia Qin and Xiujuan Zhang were responsible for the collection of masu salmon samples and the performance of rainbow trout mating experiments, including sample collection. Huiqun Chen and Linmiao Li performed genome assembly, chromosome typing and genome annotation. Jiajun Xie performed the transcriptome assembly and quantitative analysis.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Correspondence and requests for materials should be addressed to W.W. or J.C.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026