scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-Level Genome Assembly of the Japanese *Zacco platypus* for Comparative Genomics

Jui-Hung Tai[1,2,3,8], Tsung-Han Yu[2,4,8], Feng-Yu Wang[5,8], Yohey Terai[6], Ryoichi Tabata[6], Shih-Pin Huang[3], Tzi-Yuan Wang[3] ✉ & Hurng-Yi Wang[1,2,4,7] ✉

Zacco platypus **is a freshwater minnow widely distributed across East Asia, noted for its high environmental adaptability, strong reproductive capacity, and ability to hybridize across genera. The type specimen was described from Japan, yet no representative genome from the Japanese lineage has been reported to date. Introduced to Taiwan in the 1980s, *Z. platypus* rapidly established a stable population. In this study, we assembled a chromosome-level genome from a Taiwanese population and resequenced a native Japanese individual, providing genomic evidence that the Taiwanese lineage originated from Japan. The assembly and gene annotation achieved BUSCO completeness scores of 98.9% and 96.6%, respectively, representing the highest quality reported among published *Z. platypus* genomes. Furthermore, we identified chromosomal structural variation among populations, and both PCA and genetic distance analyses revealed that the Japanese lineage is distinct from continental populations, indicating the importance of representative genomes across geographic lineages. This high-quality genome provides a valuable resource for future research in comparative genomics, population genetics, hybridization, speciation, and evolutionary biology.**

Background & Summary

Zacco platypus (Temminck & Schlegel, 1846) is a freshwater minnow broadly distributed across East Asia and classified within the family Xenocyprididae, tribe Opsariichthyini. It primarily inhabits rivers and streams in Japan, China, and Korea. Although not originally native to Taiwan, *Z. platypus* is believed to have been unintentionally introduced in the 1980s and subsequently established a stable population[1]. Due to its high environmental adaptability and reproductive capacity[2], *Z. platypus* is one of the most common invasive species found in Taiwanese streams with the potential to negatively impact native freshwater fish assemblages and ecosystem functions.

In Japan, *Z. platypus* has been observed to hybridize naturally with other members of the Opsariichthyini tribe, such as *Nipponocypris* species[3], indicating high mating flexibility and incomplete reproductive isolation. In Taiwan, it has also been documented to hybridize with native Opsariichthyini species of the genus *Opsariichthys*[4,5]. These observations highlight the capacity of *Z. platypus* for intergeneric hybridization and raise concerns regarding its potential to compromise the genetic integrity and reproductive isolation of native species.

Although the type specimen of *Z. platypus* was described from Japan[6], currently available high-quality genome assemblies have been derived exclusively from Chinese and Korean populations[7,8], with genomic data for the Japanese lineage still lacking. Previous genetic studies have revealed pronounced differences between *Z. platypus* from continental Asia and those from Japan[9], further underscoring the need for genomic resources that directly capture the Japanese lineage. To fill this gap, we assembled a chromosome-level genome from a Taiwanese individual and additionally sequenced short-read data from native Japanese individuals, enabling direct comparison of their genetic relationships. Given the close genetic affinity between Taiwanese and Japanese

[1]Genome and Systems Biology Degree Program, National Taiwan University and Academia Sinica, Taipei, Taiwan. [2]Graduate Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan. [3]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan. [4]Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan. [5]Taiwan Ocean Research Institute, National Institutes of Applied Research, Kaohsiung, Taiwan. [6]Lake Biwa Museum, Kusatsu, Shiga, Japan. [7]Department of Entomology, National Taiwan University, Taipei, Taiwan. [8]These authors contributed equally: Jui-Hung Tai, Tsung-Han Yu, Feng-Yu Wang. ✉e-mail: tziyuan@gmail.com; hurngyi@ntu.edu.tw

Reads type	Platform	Individual	Collected year	Collected location	Tissue	Reads number	Data size (Gb)	Accession
DNA long reads	ONT MinION	ZpHy926	2019	Keelung River, Taiwan	Muscle	8,053,443	43.05	SRR30669803
DNA short reads	Illumina X-ten	ZpHy926	2019	Keelung River, Taiwan	Muscle	189,345,098	28.59	SRR30669440
DNA short reads	Illumina NovaSeq6000	Zp_JP02	2018	Biwa Lake, Japan	Fin	153,004,060	22.95	SRR35182058
RNA reads	Illumina NovaSeq6000	ZpHy926	2019	Keelung River, Taiwan	Eye	54,353,574	7.69	SRR30669437
RNA reads	Illumina NovaSeq6000	ZpHy926	2019	Keelung River, Taiwan	Brain	57,656,802	8.09	SRR30669438
RNA reads	Illumina NovaSeq6000	ZpHy926	2019	Keelung River, Taiwan	Olfactory lobe	59,663,636	8.41	SRR30669439
Hi-C reads	Illumina NextSeq2000	ZpZT041905	2023	Xindian River, Taiwan	Brain	253,199,338	38.23	SRR30669902

Table 1. Statistics of the sequencing data used in this study.

populations, the genome generated here serves as a representative reference for the Japanese lineage and fills a critical geographic gap in the genomic resources available for *Z. platypus* in East Asia. This resource provides a foundation for tracing introduction sources, investigating invasion genetics, and examining introgression, hybridization, and the evolution of reproductive isolation.

Methods

Sample collection and sequencing. A larval *Zacco platypus* was collected in 2019 from the Keelung River under Ruifang Bridge, New Taipei City, Taiwan (25°06′22.9″N, 121°48′32.6″E) for DNA and RNA sequencing. Additionally, an adult *Z. platypus* was collected in 2018 from Biwa Lake, Japan. All experiments in this study were performed in accordance with guidelines of the animal ethics committee and were approved by the Institutional Animal Care and Use Committee (IACUC 15-12-923), Academia Sinica.

Approximately 50 mg of muscle tissue was used for genomic DNA extraction with the Genomic DNA Buffer Set (Cat: 19060, Qiagen) and Genomic-tip 20/G (Cat: 10223, Qiagen), following the manufacturer's instructions. Genomic DNA (5–8 μg) for Nanopore sequencing was end-repaired and ligated with the KAPA Hyper Prep Kit (Cat: KR0961, Kapa Biosystems, Wilmington, MA, USA). The genomic DNA library was then premixed with LB and SQB buffer from the Nanopore Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies, UK), loaded onto a flow cell (R9.4.1; FLO-MIN106), and sequenced on MinION devices for 24–72 h.

For Illumina sequencing, genomic DNA was sheared using a Covaris 2 machine and quantified with a Qubit 2.0 Fluorometer. A short-read sequencing library was constructed from 200 ng of sheared gDNA using the TruSeq Nano DNA High Throughput Library Prep Kit (Cat. No. 20015965) and sequenced on an Illumina X-ten platform at Genomics BioScience & Technology (New Taipei City, Taiwan).

For RNA sequencing, tissues including the eye, olfactory lobe, and brain were excised and homogenized with 600 μl of Trizol solution. RNA was extracted with the RNeasy® Plus Mini Kit (Cat: 74134, Qiagen) and quantified using the Agilent 2100 Bioanalyzer. A short-read RNA library was constructed from ~1.5 μg of quantified RNA using the TruSeq Stranded mRNA Library Prep Kit (Cat. No. 20020595) and sequenced on an Illumina NovaSeq. 6000 at Genomics BioScience & Technology.

To assemble the genome to the chromosome level, an additional *Z. platypus* sample was collected in 2023 from the Xindian River, New Taipei City, Taiwan. The sample was anesthetized in 0 °C ice water, and the brain was dissected, rinsed with physiological saline, and stored in liquid nitrogen. The brain tissue was processed at the NGS High Throughput Genomics Core, BRCAS, Academia Sinica, Taiwan, where DNA extraction and library preparation were performed using the Dovetail® Omni-C® Kit (Cat: 21005, Dovetail Genomics®, Cantata Bio.), followed by sequencing on an Illumina NextSeq2000 platform with 150 bp paired-end reads.

In brief, we obtained 43.05 Gb of Nanopore long-read data, 28.59 Gb of Illumina short-read data, 38.23 Gb of Hi-C data, and an average of 7.5 Gb of RNA-seq data per tissue. The details are summarized in Table 1. Raw read quality control was performed with Porechop v0.2.4[10] for long reads and FastQC v0.12.1[11] for short reads.

Genome assembly of *Zacco platypus*. The genome size was estimated using *Jellyfish* based on Illumina short reads. From the 21-mer frequency distribution, the genome size was estimated to be approximately 831 Mb. The genome assembly was generated using Nanopore long reads with the *Flye* assembler v2.9.2-b1786[12] and subsequently polished with Illumina short reads using *Polca*[13] to improve base-level accuracy. The polished draft genome consisted of 2,543 contigs totaling approximately 844.19 Mb, with a contig N50 of 6.17 Mb and a maximum contig length of 18.26 Mb. Potential contamination in the draft genome was assessed using *BlobTools* v1.1.1[14]. A total of 2,313 contigs were assigned to the phylum Chordata, while 230 contigs were classified as "no-hit" (Fig. 1), showing that no evident contamination was detected in the draft assembly.

Using Hi-C data processed through the *Juicer*[15] and *3D-DNA*[16] pipelines, the 2,543 contigs were further anchored into 2,332 contigs with a total length of 847.09 Mb, increasing the scaffold N50 to 31.42 Mb (Table 2). Although the Hi-C data were obtained from a different individual, heterogeneity is expected to be low and unlikely to affect the assembly results, as the Taiwanese *Z. platypus* population was introduced from Japan relatively recently. Consistent with this expectation, 93.38% of the assembled bases were anchored to 24 chromosome-scale scaffolds (Fig. 2), ranging from 23.38 Mb to 46.34 Mb in length (Table 3), while the

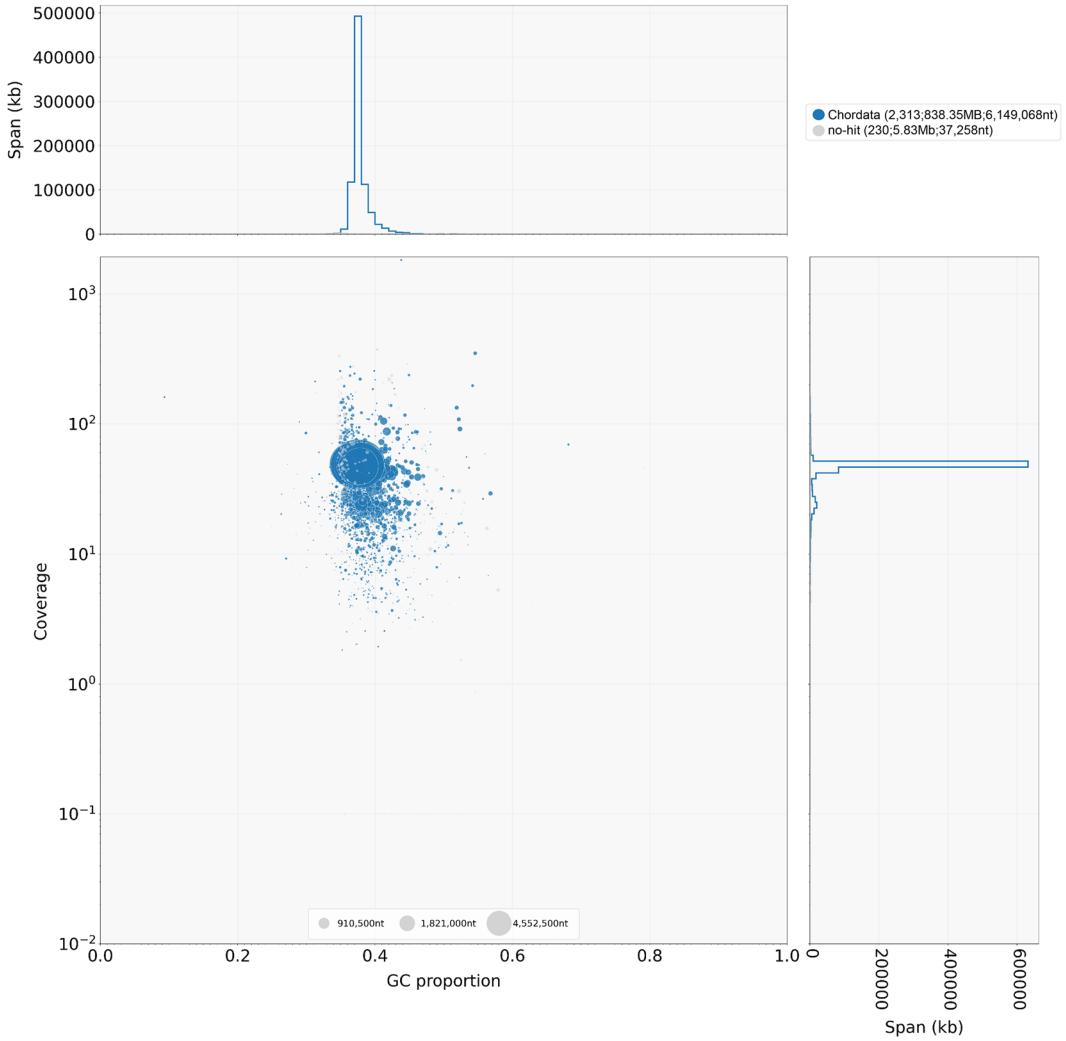

Fig. 1 BlobTools analysis of the unscaffolded assembly. Contigs are plotted by average sequence coverage (y-axis) and GC content (x-axis). Bubble size corresponds to contig length, and colors indicate taxonomic assignments. Histograms display the distribution of GC content and sequence coverage, weighted by contig length. Most of the contigs were classified as Chordata, while a few contigs were assigned as no-hit.

remaining contigs were all shorter than 0.28 Mb. The chromosome numbering follows scaffold length, from the largest (Chr1) to the smallest (Chr24). The final assembly showed a short-read mapping rate of 99.74% with an average coverage depth of 34×, and a complete *BUSCO* v5.8.2[17] score (BUSCOs) of 98.9% (96.0% of genes identified as single-copy orthologs, 2.9% as duplicated, only 0.7% fragmented, and 0.4% missing.). To our knowledge, this genome represents the highest complete BUSCOs reported so far among published *Zacco platypus* genomes (Table 2).

Repeat annotation. Repetitive elements in the genome were identified using a combination of de novo prediction and homology-based approaches. For de novo prediction, *RepeatModeler* v2.0.2[18] was run with default parameters, and its outputs were integrated with those from *LTR_FINDER* v1.07[19], *LTR_FINDER_parallel* v1.2[20], and *LTR_retriever* v2.9.0[21] to enhance the accuracy of long terminal repeat (LTR) detection. For homology-based search, annotated repetitive sequences from two cyprinid species (*Ctenopharyngodon idellus* and *Danio rerio*) were retrieved from FISHTEDB (http://www.fishtedb.org/)[22]. Additional repetitive sequences belonging to the family Cyprinidae were obtained from Dfam (version 3.5, December 22, 2021 release)[23] and Repbase (version 20181026)[24]. All repeat libraries generated from these methods were concatenated and subsequently used as input for *RepeatMasker*[25] to perform genome-wide repeat annotation.

A total of 55.02% of the genome were identified as repetitive sequences. DNA transposons are the most prevalent (25.72%), follow by LTRs (7.57%), LINEs (4.12%) and SINEs (0.25%) (Table 4).

Gene annotation. For high-quality protein-coding gene annotation, we employed both homology-based and transcript-based predictions using *BRAKER2*[26], and subsequently integrated the results with *TSEBRA*[27] by applying the preferred RNA configuration file. For homology-based prediction, protein sequences from diverse

	Taiwan		Korea	China
	Draft assembly	HiC scaffolding	Published Genome	Published Genome
Genome Size	846.39	847.09	838.47	815.00
Number of contigs/scaffolds	2543	2332	316	54
N50	6.17	31.42	33.35	32.32
L50	39	12	11	11
N90	0.17	26.9	28.54	27.68
L90	379	23	22	22
Largest Chromosome/Contig Size	18.26	46.34	47.63	46.87
Smallest Chromosome Size	NA	23.38	24.08	25.28
Depth	—	51x		
Mapping Rate	—	99.74%	NA	NA
Genome BUSCO v5	—	98.90%	98.10%	96.30%

Table 2. Genome assembly statistics for *Zacco platypus*.

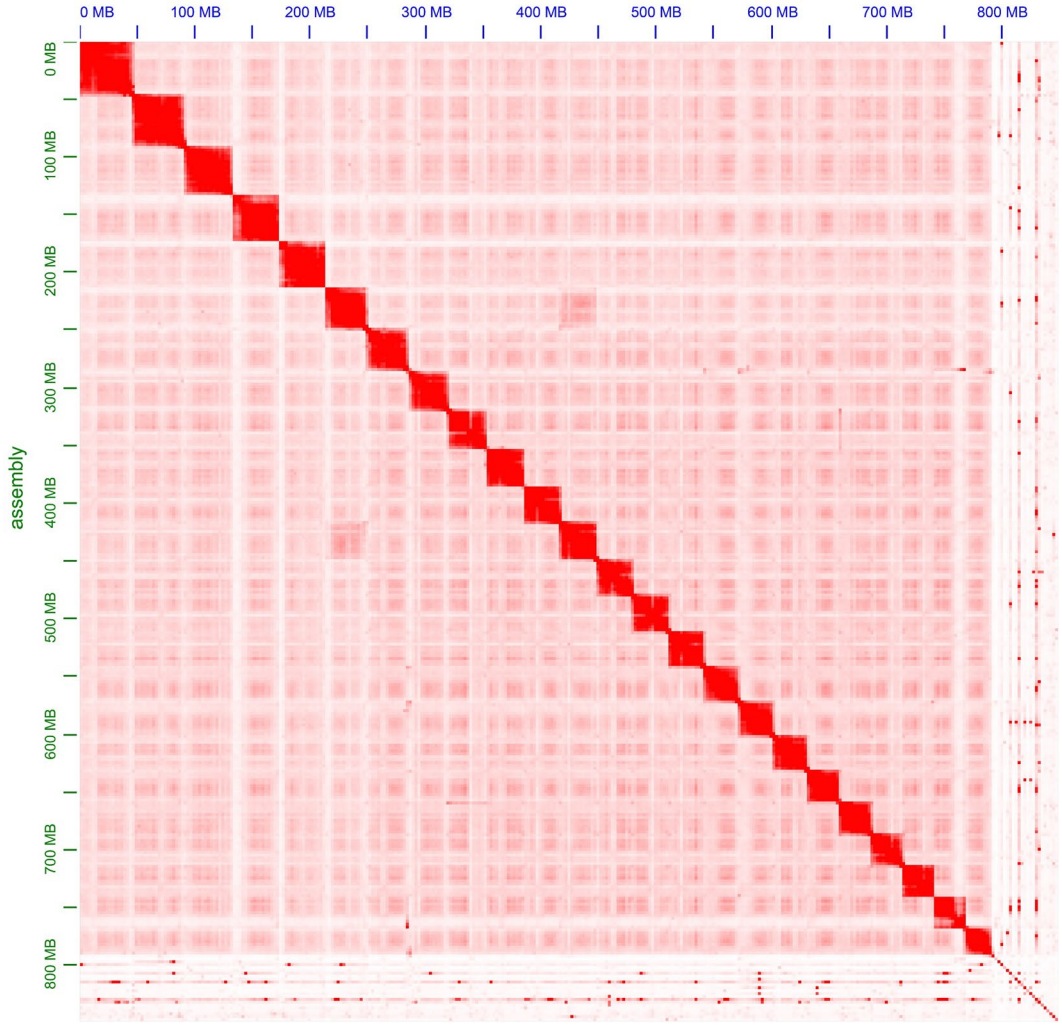

Fig. 2 Heatmap of interaction intensities among chromosome sequences of *Zacco platypus* anchored by Hi-C data. Approximately 93.38% of the assembled bases were anchored to 24 pseudchromosomes.

teleost species were retrieved from GenBank, including *Astyanax mexicanus* (GCA_023375975.1), *Carassius auratus* (GCA_003368295.1), *Chanos chanos* (GCA_902362185.1), *Clupea harengus* (GCA_900700415.2), *Danio rerio* (GCA_000002035.4), *Paramormyrops kingsleyae* (GCA_002872115.1), *Salmo salar* (GCA_905237065.2), and *Scophthalmus maximus* (GCA_022379125.1). For transcript-based prediction, transcriptome data from the eye, olfactory lobe, and brain were incorporated to support gene model construction.

Chr ID	Length	Chr ID	Length
Chr1	46,168,000	Chr13	31,412,848
Chr2	44,612,832	Chr14	31,229,197
Chr3	41,492,984	Chr15	30,382,589
Chr4	40,371,381	Chr16	30,067,470
Chr5	40,040,551	Chr17	30,024,168
Chr6	36,382,142	Chr18	29,511,820
Chr7	35,119,276	Chr19	27,872,422
Chr8	34,996,288	Chr20	27,715,027
Chr9	33,225,031	Chr21	27,259,020
Chr10	32,858,142	Chr22	26,986,294
Chr11	31,510,455	Chr23	26,901,218
Chr12	31,420,721	Chr24	23,415,336
		Mitochondria	16,612

Table 3. The chromosome length of *Zacco platypus*.

	Masked base	Mask percent
DNA transposons	217,872,213	25.72
Retroelements	101,183,321	11.94
LTRs	64,156,129	7.57
LINEs	34,898,721	4.12
SINEs	2,128,471	0.25
Rolling-circles	11,074,486	1.31
Satellites	7,519,109	0.89
Unclassified	128,338,339	15.15
Total bases masked	466,085,046	55.02

Table 4. Statistics of repetitive sequence annotation results.

Function Database	Number of gene	Annotation percentage
NCBI blast	27,060	79.53%
InterProScan	24,217	71.18%
Gene Ontology	24,601	72.30%
KEGG	19,840	58.31%
Annotated	27,249	80.78%
Total	34,024	—

Table 5. Statistics of *Zacco platypus* genome annotation.

The initial prediction identified a total of 34,024 genes, including 36,734 protein-coding transcripts. To reduce potential false positives, we performed functional annotation by aligning the predicted genes against multiple databases, including NCBI, InterProScan[28], Gene Ontology[29], and KEGG[30]. Only genes with functional evidence in at least one of these databases were retained (Table 5). Finally, 27,249 genes were annotated, with an average of 9.6 exons per gene. The predicted gene set achieved a complete BUSCOs of 96.6%, representing the highest reported among published *Zacco platypus* gene annotations, compared with 91.4% and 89.4% complete BUSCOs for the China and Korea populations, respectively.

Mitochondrial genome assembly. Whole mitochondrial genomes were assembled using *NOVOPlasty* v4.3.5[31] with short-read sequences as input. The genome size range was set to 16,000–17,000 bp, and the k-mer size was set to 33, following the default settings. A complete circular mitochondrial genome of 16,612 bp was obtained. The mitochondrial genome sequence was included in the final genome assembly file.

Genome assembly comparisons. *MUMmer* v4.0.0rc1[32] was used to perform whole-genome alignments. Alignments with ≥90% identity and a minimum alignment length threshold of 25 kb were visualized. By comparing the genome structures of the Chinese, Korean, and Taiwanese assemblies (Fig. 3), we observed that, aside from several minor structural variants, most chromosomes were collinear among the three populations. Notably, chromosome 8 in the Korean individual showed a distinct configuration compared with those of the Taiwanese and Chinese individuals, indicating that different populations may harbor distinct chromosomal architectures. This finding demonstrates the importance of generating representative reference genomes from multiple populations. Short reads from Taiwanese, Japanese, Korean, and Chinese were mapped to the assembled genome. Sequencing

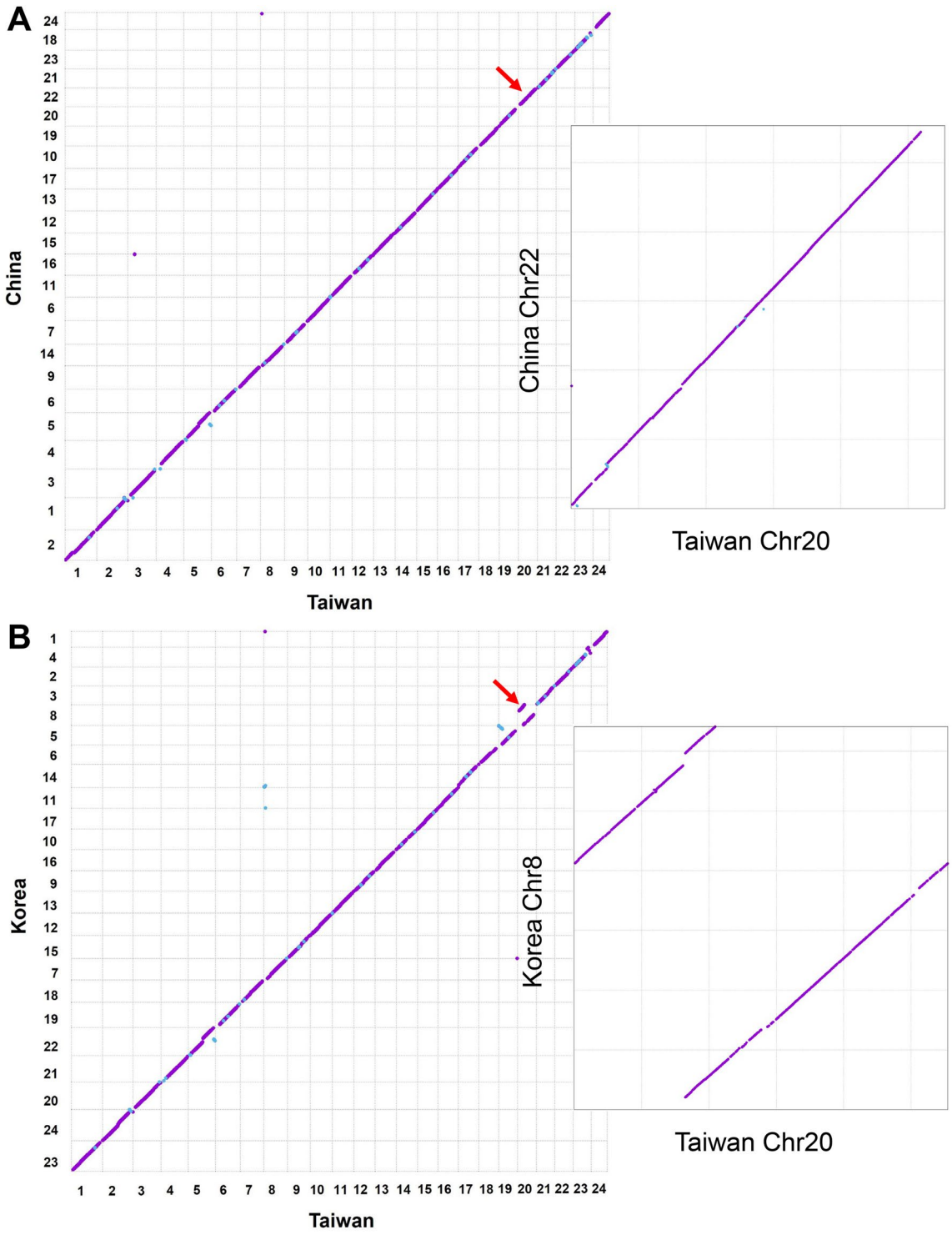

Fig. 3 Whole genome alignments among *Zacco platypus* populations. Comparisons are shown between the Taiwan population and (**A**) the China population and (**B**) the Korea population. The axes represent chromosomes. Purple lines indicate forward alignments, and cyan lines indicate reverse alignments. The panel on the right shows a magnified comparison of Taiwan chromosome 20 with the homologous chromosomes of the other populations.

data for the Korean individual (SRR28903499) and the Chinese individual (SRR26456191) were obtained from NCBI. Only bi-allelic SNPs were retained, and linkage disequilibrium was pruned using *PLINK* v1.9[33] with a 10-kb window, a step size of 1 SNP, and an r^2 threshold of 0.2. PCA was conducted with pruned VCF file[34] using *VCF2PCACluster* v1.41[35]. The results of principal component analysis showed that *Zacco platypus* from Taiwan and Japan clustered closely, consistent with previous findings that demonstrated a Japanese origin[1] (Fig. 4).

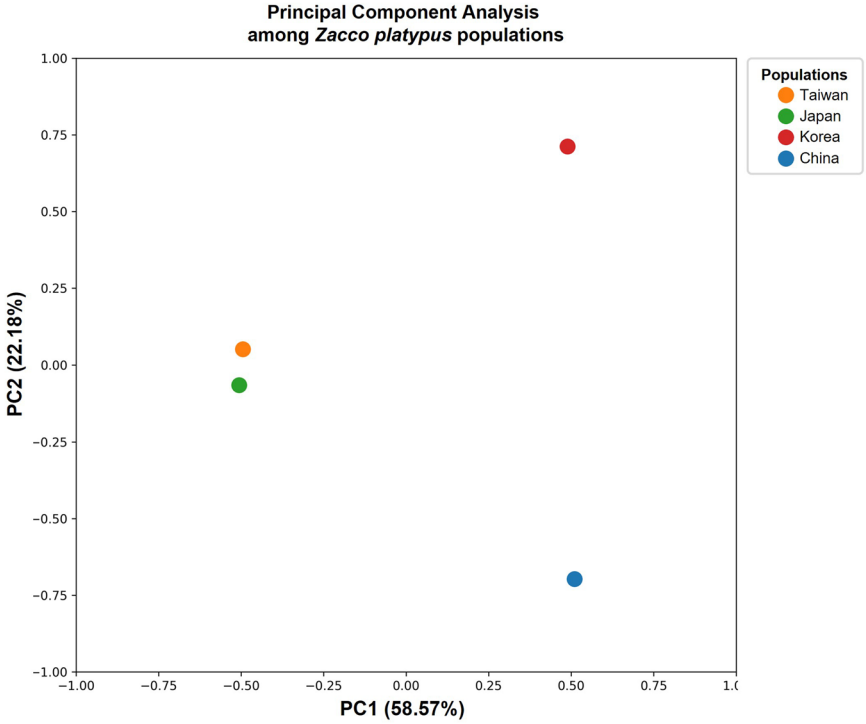

Fig. 4 Relationships among *Zacco platypus* populations. Principal component analysis plot shows genetic relationships among Taiwan (orange), Japan (green), Korea (red), and China (blue) populations are shown along the first two principal components (X-axis: PC1, 58.57%; Y-axis: PC2, 22.18%).

Data Records

All raw sequencing data and the assembled genome have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). The chromosome-level genome assembly is available at GenBank under the accession number JBHGZP000000000[36]. The genome annotation file has been deposited in the *Figshare* database[37]. For Taiwanese *Z. platypus*, the Nanopore long reads is available under accession SRR30669803[38], the Illumina short reads under SRR30669440[39], the Hi-C reads under SRR30669902[40], and the RNA-seq reads under SRR30669437–SRR30669439[41–43]. The Illumina short reads of Japanese *Z. platypus* are available under accession SRR35182058[44].

Technical Validation

Raw read quality control was performed with *Porechop* v0.2.4[10] for long reads and *FastQC* v0.12.1[11] for short reads. Potential contamination in the raw reads was assessed using *BlobTools* v1.1.1[14]. Genome assembly quality and protein-coding gene annotation completeness were evaluated with *BUSCO* v5.8.2 against the *actinopterygii_odb10* lineage dataset.

Data availability

The chromosome-level genome assembly is available at GenBank under the accession number JBHGZP000000000[36]. The genome annotation file has been deposited in the *Figshare* database[37]. For Taiwanese *Z. platypus*, the Nanopore long reads is available under accession SRR30669803[38], the Illumina short reads under SRR30669440[39], the Hi-C reads under SRR30669902[40], and the RNA-seq reads under SRR30669437–SRR30669439[41–43]. The Illumina short reads of Japanese *Z. platypus* are available under accession SRR35182058[44].

Code availability

No specific scripts were developed for this project. All data processing and bioinformatics analyses were conducted using publicly available software, following protocols and manuals provided by each respective tool.

Received: 5 September 2025; Accepted: 10 December 2025;
Published online: 23 December 2025

References

1. Ma, G. C., Watanabe, K., Tsao, H. S. & Yu, H. T. Mitochondrial phylogeny reveals the artificial introduction of the pale chub (Cyprinidae) in Taiwan. *Ichthyol Res* **53**, 323–329, https://doi.org/10.1007/s10228-006-0353-3 (2006).
2. Fu, S.-J., Cao, Z.-D., Yan, G.-J., Fu, C. & Pang, X. Integrating environmental variation, predation pressure, phenotypic plasticity and locomotor performance. *Oecologia* **173**, 343–354, https://doi.org/10.1007/s00442-013-2626-7 (2013).
3. Arao, K. & Shimoyama, J. Hybrids between *Zacco platypus* and *Z. temminckii* from Aichi prefecture. *Japan Sci Rep Toyohashi Mus Nat Hist* **16**, 53–54 (2006).

4. Liao, N. L., Huang, S. P. & Wang, T. Y. Interspecific mating behavior between introduced *Zacco platypus* and native *Opsariichthys evolans* in Taiwan. *Zool Stud* **59**, e6, https://doi.org/10.6620/zs.2020.59-6 (2020).
5. Wang, C.-F., Chang, G.-C., Wang, Y.-Q., Chen, Q.-Y. & Lin, G.-Y. Do introduced *Opsariichthys* and native *Opsariichthys* interbreed naturally?, (New Taipei Municipal Ming Der High School, New Taipei City, Taiwan, 2011).
6. Siebold, P. F. V., Haan, W. D., Schlegel, H. & Temminck, C. J. *Fauna japonica, sive, Descriptio animalium, quae in itinere per Japoniam, jussu et auspiciis, superiorum, qui summum in India Batava imperium tenent, suscepto, annis 1823-1830*. Vol. v.[2] Pisces (Apud Auctorem, 1835).
7. Nam, S.-E. & Rhee, J.-S. Chromosomal-level genome assembly data from the pale chub, *Zacco platypus* (Jordan & Evermann, 1902). *Data in Brief* **55**, 110596, https://doi.org/10.1016/j.dib.2024.110596 (2024).
8. Xu, X. *et al*. A chromosome-level genome assembly of East Asia endemic minnow *Zacco platypus*. *Scientific Data* **11**, 317, https://doi.org/10.1038/s41597-024-03163-w (2024).
9. Perdices, A. & Coelho, M. M. Comparative phylogeography of *Zacco platypus* and *Opsariichthys bidens* (Teleostei, Cyprinidae) in China based on cytochrome b sequences. *Journal of Zoological Systematics and Evolutionary Research* **44**, 330–338, https://doi.org/10.1111/j.1439-0469.2006.00368.x (2006).
10. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics* **3**, https://doi.org/10.1099/mgen.0.000132 (2017).
11. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
12. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540–546, https://doi.org/10.1038/s41587-019-0072-8 (2019).
13. Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology* **16**, e1007981, https://doi.org/10.1371/journal.pcbi.1007981 (2020).
14. Laetsch, D. & Blaxter, M. BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. *F1000Research* **6**, https://doi.org/10.12688/f1000research.12232.1 (2017).
15. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
16. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
17. 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. *Molecular Biology and Evolution* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
18. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
19. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
20. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* **10**, 48, https://doi.org/10.1186/s13100-019-0193-0 (2019).
21. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
22. Shao, F., Wang, J., Xu, H. & Peng, Z. FishTEDB: a collective database of transposable elements identified in the complete genomes of fish. *Database* **2018**, bax106, https://doi.org/10.1093/database/bax106 (2018).
23. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Research* **44**, D81–D89, https://doi.org/10.1093/nar/gkv1272 (2016).
24. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11, https://doi.org/10.1186/s13100-015-0041-9 (2015).
25. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **25**, 4.10.11–14.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
26. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* **3**, lqaa108, https://doi.org/10.1093/nargab/lqaa108 (2021).
27. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566, https://doi.org/10.1186/s12859-021-04482-0 (2021).
28. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
29. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
30. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research* **53**, D672–D677, https://doi.org/10.1093/nar/gkae909 (2025).
31. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research* **45**, e18–e18, https://doi.org/10.1093/nar/gkw955 (2017).
32. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology* **14**, e1005944, https://doi.org/10.1371/journal.pcbi.1005944 (2018).
33. Purcell, S. *et al*. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559–575, https://doi.org/10.1086/519795 (2007).
34. Tai, J.-H. *et al*. The VCF file of four *Zacco platypus* lineages. *figshare*. https://doi.org/10.6084/m9.figshare.30600653 (2025).
35. He, W. *et al*. VCF2PCACluster: a simple, fast and memory-efficient tool for principal component analysis of tens of millions of SNPs. *BMC Bioinformatics* **25**, 173, https://doi.org/10.1186/s12859-024-05770-1 (2024).
36. Wang, T.-Y. *et al*. *Zacco platypus* isolate ZpHy926, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:JBHGZP000000000 (2025).
37. Tai, J.-H. *et al*. Annotation files of Japan lineage *Zacco platypus* from Taiwan. *figshare*. https://doi.org/10.6084/m9.figshare.30011686.v1 (2025).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30669803 (2025).
39. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30669440 (2025).
40. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30669902 (2025).
41. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30669437 (2025).
42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30669438 (2025).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30669439 (2025).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35182058 (2025).

Acknowledgements
This study was supported by grants from the National Science and Technology Council (NSTC), Taiwan (113-2327-B-002-003-, MOST 109-2311-B-002-023-MY3, MOST 105-2311-B-001-064, MOST 106-2311-B-001-022, and MOST 107-2311-B-001-007), and National Taiwan University (113L7223). It was also partially supported by the Taiwan BioGenome Project, funded by Academia Sinica, Taiwan (AS-Grant 23-23). Additional support was provided through the National Key Area International Cooperation Alliance: University Academic Alliance in Taiwan (UAAT) - Kyushu-Okinawa Open University (KOOU) - Medicine and Life Sciences Integrative Program, funded by the Ministry of Education, Taiwan, to promote international collaboration in cutting-edge research.

Author contributions
H.Y.W. and T.Y.W. conceived and designed the study. T.Y.W., S.P.H., F.Y.W., Y.T., R.T., J.H.T. and T.H.Y. collected samples. J.H.T. and T.H.Y. performed the data analysis. T.Y.W. and Y.T. conducted experiments. J.H.T. wrote the manuscript. H.Y.W. and T.Y.W. revised the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to T.-Y.W. or H.-Y.W.