Massive Chromosome Fission and Speciation in Asian River Chub: Integrating Kinetochore Reproduction Theory with Transposable Element Dynamics

Jui-Hung Tai [1,2,3,#], Tzi-Yuan Wang [3,#], Gwo-Chin Ma[4], Yu-Wei Wu[5,6,7], Tsung-Han Yu[8], Chen-Fan Wang[9], Te-Yu Liao[10], Shih-Pin Huang[3], Feng-Yu Wang[11], Yohey Terai[12], Tsung-Han Wu[1], Chun-Hua Hsu[1,13], Hon-Tsen Yu[14], Kwang-Tsao Shao[3,15*], Shu-Miaw Chaw [3*], Hurng-Yi Wang [2,8,16,17*]

[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] Department of Genomic Medicine, Changhua Christian Hospital, Changhua, Taiwan.

[5] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan.

[6] Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan

[7] TMU Research Center for Digestive Medicine, Taipei Medical University, Taipei, Taiwan

[8] Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan

[9] Ming Der Municipal High School, New Taipei City, Taiwan

[10] Department of Oceanography, National Sun Yat-sen University, Kaohsiung, Taiwan

[11] Taiwan Ocean Research Institute, National Applied Research Laboratories, Kaohsiung, Taiwan.

[12] Research Center for Integrative Evolutionary Science, The Graduate University for Advanced Studies, SOKENDAI, Hayama, Kanagawa, Japan

[13] Department of Agricultural Chemistry, National Taiwan University, Taipei, Taiwan

[14] Department of Life Science, National Taiwan University, Taipei, Taiwan

[15] Institute of Marine Biology, National Taiwan Ocean University, Keelung, Taiwan

[16] Graduate Institute of Medical Genomics and Proteomics, College of Medicine, National Taiwan University, Taipei, Taiwan

[17] Department of Entomology, National Taiwan University, Taipei, Taiwan

[#] These authors contribute equally

* KTS, SMC and HYW are the co-corresponding authors

ABSTRACT

Chromosome rearrangements significantly influence speciation, but the role of chromosome fission is paradoxical due to its link with unbalanced segregation. The kinetochore reproduction theory suggests that fissions from kinetochore duplication should allow normal segregation, though empirical support remains scarce. In the Opsariichthyini tribe (Asian river chub), chromosome number vary from 2N=48 in most genera to 76-78 in *Opsariichthys* due to 15 fission events. *Opsariichthys* harbors over 15 species, substantially surpassing the typical three to four species in other genera. This exceptional diversity, coupled with dramatic chromosome fission, makes the Opsariichthyini as an ideal model for studying speciation driven by chromosome fission. We sequenced and annotated genomes from five representative species, finding that significant transposable element (TE) accumulation in centromeres likely contributes to centromeric DNA expansion. This, in turn, may have facilitated neokinetochore formation, potentially leading to chromosome fissions. The fissioned chromosomes showed lower migration rates and higher genetic divergence between species than their intact counterparts, highlighting their role as reproductive barriers that promote speciation. Our findings offer unprecedented support for the kinetochore reproduction theory and elucidate the complex interplay between TE activities, genomic structure, and speciation.

Keywords:
Transposons, chromosome fission, speciation, LINE/L2, LTR/Gypsy, gene flow

INTRODUCTION

Throughout organismal evolution, transposable elements (TEs) have played a recurring role influencing genomic variation and contributing to diversity of their hosts[1]. TEs activities can significantly reshape genome architecture and influence chromosomal dynamics, thereby exerting a global impact on genome organization. This remodeling of the genome often results in chromosomal rearrangements, e.g., fusions, fissions, translocations, and inversions. These alterations can lead to evolutionary novelty and potentially play a role in speciation[2-8].

The chromosomal speciation model suggests that chromosomal rearrangements create reproductive barriers that drive speciation, either by causing underdominance (heterozygous disadvantage) in hybrids or by suppressing recombination [8-11]. However, in species with monocentric chromosomes, such as vertebrates, fissions can result in chromosomes with no centromeres, leading to failed segregation[12]. Therefore, how often and under what circumstances fission act as barriers to gene flow and promote speciation remain unclear. The kinetochore reproduction theory of chromosome fission, initially proposed by Todd [13] and later modified by Kolnicki [14], hypothesized that the accumulation of centromeric DNA can lead to neokinetochore formation. During meiosis, if both chromatids have bipolar attachment in opposition to one another, chromosome fission will occur. Subsequent pericentric inversion may complicate pairing between fissioned and intact homologs, thus creating barriers to gene flow and driving speciation. Although theoretically plausible, this theory lacks substantial empirical evidence to confirm its general applicability [15].

The Opsariichthyini tribe (Asian river chub) (Cypriniformes: Xenocyprididae) represents a monophyletic group of small minnows, widely distributed across Eastern Asia from Russia to northern Vietnam[16,17]. This tribe comprises genera that range from the early-diverging *Parazacco*, *Candidia*, and *Nipponocypris*, to the more recently diverged *Zacco*, and *Opsariichthys*. Among the Opsariichthyini, chromosome numbers are 2N=48, except for *Opsariichthys*, which range from 74-78[18,19]. Interestingly, hybrids between *Zacco* and *Opsariichthys*, as well as between different *Opsariichthys* species, have occasionally been seen in nature environments where they are sympatric[20,21]. Despite chromosome number increase, the DNA content of these species remains nearly identical, suggesting chromosome fissions rather than increases in DNA content as the cause for chromosome changes[22].

In addition to fission, the genus *Opsariichthys* includes over 15 species[23], markedly more than the typical three to four species found in other genera of the same tribe[24]. This prompts an intriguing question about how *Opsariichthys* has managed to reach such high species diversity, given the potential challenges posed by widespread chromosome fission.

Given its unusual tolerance to chromosomal fissions and its high species count, *Opsariichthys* serves as an ideal candidate to explore the chromosomal speciation model.

To investigate relationship between chromosome fission and TE activities in *Opsariichthys*, and to address whether chromosome fissions contribute to accelerated speciation, we sequenced and annotated the genome of *Candidia barbatus* (Cb), *Zacco platypus* (Zp), *O. pachycephalus* (Op), and *O. evolans* (Oe), each including multiple individuals from different populations. For comparative purposes, the genome of *O. bidens* (Ob) was also included in the subsequent analyses [25]. These five species represent the genomic and ecological diversity within the Opsariichthyini tribe. Our findings suggest that chromosome fissions were triggered by lineage-specific TE activity, likely starting from the common ancestor of the Opsariichthyini. Furthermore, fissioned chromosomes experienced reduced migration and increased divergence compared to intact chromosomes, highlighting their role in hindering gene flow and promoting speciation.

RESULTS

Karyotyping, genome assembly and annotation

Chromosome counts were determined by examining multiple individuals from each species. No heteromorphic element was detected between males and females. The chromosome numbers for Cb (*Candidia barbatus*) and Zp (*Zacco platypus*) are 2N = 48, while Op (*Opsariichthys pachycephalus*) and Oe (*O. evolans*) have 2N = 76 and 78, respectively (Fig. 1 and Table 1). Although the chromosome numbers of these species vary widely, their fundamental numbers are similar.

The details of genome assembly for each species are given in Supplementary Text (Genome assembly statistics; See also Table S1-S2). Briefly, the assembled genome sizes were 829.68 Mb for Cb and 847.09 Mb for Zp, with 97.32% and 93.38% of contig lengths, respectively, anchored to 24 pseudo-chromosomes for each species (Fig. S1). For Op and Oe, genome sizes were 916.58 Mb and 920.54 Mb, with 91.06% and 91.21% of contigs length anchored to 38 and 39 pseudo-chromosomes, respectively (Fig. S1). The assembled chromosome numbers are consistent with the karyotyping results and all show more than 96% completeness in BUSCO scores, indicating high-quality assemblies (Table S3).

Transposable elements (TEs) account for 51% to 55% of the opsariichthyin genomes, with DNA transposons being the most prevalent (~25 - 27%), followed by long terminal repeats (LTRs; ~7 - 8%), long interspersed nuclear element (LINEs; ~4 - 5%), and short interspersed nuclear element (SINEs; ~0.2 - 0.3%) (Table S4). Notably, the Opsariichthyini exhibit a significantly higher proportion of LINEs, particularly L2 elements (L2), and a significant increase in Gypsy elements (Gypsy) within their LTR content compared to other cyprinids. Gene annotation statistics reveal 26,894 to 30,730 protein-coding sequences in five opsariichthyin species, with BUSCO scores ranging from 91.10% to 96.70% (Table S5).

Chromosomal evolution is associated with recent TE activities

The syntenic relationships between species demonstrated that in *Opsariichthys*, chromosome numbers have dramatically increased from 24 to either 38 or 39 due to 14 to 15 fission events (Fig. 2). We identified 'breaking regions' in Cb and Zp, where segments of the original chromosome fail to align with sequences in their fissioned homologs (Fig. 3), likely due to structural changes or sequence loss during chromosome fission. Breaking regions are enriched with TEs, particularly Gypsy and L2, with Zp showing a higher Gypsy density than Cb (Fig. 4A). This enrichment extends to the homologous regions of the resultant chromosomes in Op, Oe, and Ob (Fig. 3 and Fig. S2).

Repeat landscapes reveal that recent expansions of L2 and Gypsy (k-distance < 5%; young TEs) are prevalent across all examined opsariichthyin species (Fig. 4B). In the genomes of Cb and Zp, 56% and 43% of L2, respectively, are recently added, increasing to 75% and 72% within the breaking regions (Fig. 4C). Meanwhile, recent additions of Gypsy is more pronounced in Zp, increasing from 17% genome-wide to 31% in the breaking regions. In Cb, only 10% of Gypsy results from recent additions in both genome-wide and breaking regions. Furthermore, in Cb, 14 and 8 of the 15 breaking regions show recent expansions of L2 and Gypsy (Table S7; Fig. S3), respectively, while in Zp, recent expansions occur in all regions (Table S8; Fig. S4). These observations suggest that L2 and Gypsy expansion occurred asynchronously, with L2 expanding before Gypsy.

Chromosome fission within centromere regions

To investigate the mechanism of chromosome fission, we identified putative centromere regions by searching for high order repeats (HORs)[26] and estimating recombination rates along the chromosomes (Table S9). We discovered several HORs in Cb, Zp, and genus *Opsariichthys*. These sequences vary in length (Fig. S5A) but are composed of highly similar repeating units, each approximately 100 bp long with low GC content (Fig. S5A and S5B). Phylogenetic analysis suggests these HORs likely originated from an ancient LINE1 element (Fig. S5C). Notably, these HORs were predominantly found in regions of low recombination rates and gene density (Fig. S3-S4 and Fig. S6-S8), which also enriched with L2 and Gypsy. These features suggest that these regions may be putative centromeres, with HORs serving as centromeric DNA.

We then classified the chromosomes into metacentric, sub-metacentric, and sub-/telocentric categories based on the locations of putative centromere regions[19] (Supplementary Text: Centromere identification). In general, karyotyping results derived *in silico* analysis and microscopic observations are largely consistent (Table 1), demonstrating that the identified putative centromere region identified are authentic.

In Cb, 10 out of 15 breaking regions overlapped with putative centromeres (Fig. S3; Table S7), while in Zp, all 15 coincided with putative centromeres (Fig. S4; Table S8). The difference is primarily due to chromosomal rearrangements, especially pericentric inversions, that occurred during evolution (Fig. 5). Recent L2 expansions were observed in centromeres across all species, with Gypsy expansions prevalent in nearly all centromeres of Zp, Op, Oe, and Ob, but in only 14 of Cb's 24 centromeres (Table S7-8 and S10-12). These patterns, consistent with breaking regions, support earlier L2 expansion than Gypsy.

It is noteworthy that the number of centromeric DNA and their spanning regions are significantly greater in Zp than in Cb for chromosomes that underwent fission (Fig. 6A). In addition, while the number of L2 in the centromere regions is similar between the two species, the number of Gypsy is significantly greater in Zp (Fig. 6B). These observations imply that recent Gypsy expansion might have expanded the actual centromere regions of Zp chromosomes. The duplication of centromeric DNA, associated with expanded centromeres, may lead to neokinetochore formation, generating dicentric chromosome [14]. Therefore, the resulting fissioned chromosomes may each retain one functional kinetochore, ensuring proper segregation.

Insertion preferences and interaction between L2 and Gypsy

Expansion of young L2 in putative centromere regions suggests that the reduced recombination rates in centromeres inhibits transposon removal, leading to their accumulation over time [27,28]. This is evidenced by a strong negative correlation between the presence of young L2 and recombination rates in Zp (r = -0.44) and Cb (r = -0.25) (Fig. S9A). Because centromeres can shift over time (Fig. 5) which may change the recombination landscape within the genome, ancient L2 (k-distance > 5%) may show less correlation with recombination rate. Indeed, the correlation of ancient L2 with recombination rate is weaker in Zp (r = -0.11) and absent in Cb (Fig. S9B).

The expansion of L2 preceding that of Gypsy suggests a potential interdependence between these TEs. Notably, a strong correlation exists between the emergence of young L2 and Gypsy sequences, particularly in Zp (slope = 0.5, r = 0.66) compared to Cb (slope = 0.15, r = 0.42) (Fig. S10). This pattern of L2 expansion preceding Gypsy suggests that the expansion of Gypsy may depend on the genomic environment shaped by L2.

A characteristic mark of LTR insertion in the genome is the presence of a target site duplication (TSD), which result from staggered double-strand breaks during integration [29,30]. Our analysis of preferred insertion sites for Gypsy, using results from LTR_finder and LTR_retriever, identified a conserved palindrome sequence flanking the 5-bp TSD in opsariichthyin species, TNACA-TSD-TGTNA (Fig. S11A). Similar palindrome sequences are present in the L2 of the Opsariichthyini (Fig. S11B), suggesting that each replication and transposition event of L2 generates new target sites that guide the insertion of Gypsy into the same regions.

Mutations in PIWI1 protein associates with L2 expansion

While the recent expansion of Gypsy was also found in other cyprinids (Fig. S12A), the recent expansion of L2 is unique to the Opsariichthyini. This may suggest modifications in transposon silencing mechanisms within the tribe. Given the crucial role of PIWI1 protein in germ cell maintenance and transposon silencing in zebrafish [31], PIWI1 sequences of opsariichthyins were compared with those of other cyprinids (Supplementary Text: Comparative Analysis of Piwi1 Protein).

Within the PAZ domain of PIWI1 protein, an important piRNA binding domain that guides PIWI1 protein to target mRNA, all opsariichthyins possess two unique mutations, G348S and V380I (Fig. S12C). The V380I mutation is of particular interest in piRNA interactions, as it is adjacent to the nucleic acid-binding interface at amino acid 379[32]. Through repeat landscape analysis, other cyprinids carrying V380 do not show L2 expansion (Fig. S12A).

Interestingly, the genus *Triplophysa* (Cypriniformes: Nemacheilidae), an outgroup to cyprinids and carrying V380I mutation, has also experienced recent L2 expansion (Fig. S12B). Besides V380I, *Triplophysa* shares the M378L mutation with *Danionella cerebrum* (Danionidae) (Fig. S12C). Since *D. cerebrum* did not exhibit recent L2 expansion (Fig. S12A), it is evident that the V380I mutation specifically associates with the observed L2 expansion. Given that *Triplophysa* is an outgroup to cyprinids, the V380I mutation observed in both the Opsariichthyini and *Triplophysa* likely arose independently, implying that the L2 expansion events in these species did not originate from a common ancestor. This leads to the inference that the V380I mutation may interfere with the PIWI1 protein's ability to suppress TEs, resulting in the expansion of L2.

Chromosome fission decrease gene flow and increase genetic divergence

According to the chromosome speciation model, chromosomal rearrangements can lead to the formation a partial reproductive barrier, thereby promoting speciation. Moreover, genetic divergence at loci on rearranged chromosomes is likely to be greater than on non-rearranged chromosomes. We used gIMble[33] to calculate the blockwise site frequency spectrum (bSFS) between Zp and Oe, as well as between Zp and Op. We fit various demographic models, including strict isolation (DIV), migration only (MIG), and isolation with migration (IM). The IM model was the best fit (Table S13). Migration near the breaking regions is notably reduced (Fig. 7A and Fig. S13-S14), and gene flow estimates across fission chromosomes are significantly lower on average compared to non-fission chromosomes in both pairs ($p < 0.05$) (Fig. 7B).

We next calculated the genetic distances of single-copy genes between species pairs. On average, genes on fission chromosomes exhibited significantly higher genetic distances compared to those on non-fission chromosomes when analyzing Zp with three species of *Opsariichthys* ($p < 10^{-6}$) (Fig. 7C). In contrast, for the two species with the same chromosome number, Zp and Cb, no significant differences in genetic distance between two groups of chromosomes.

It is generally believed that fission does not drastically reduce recombination. Indeed, not only did fission chromosomes not exhibit reduced recombination rates, but their recombination rates were actually on average higher than those of non-fission chromosomes (Fig. S15) in both Op and Oe. The recombination rates for the two groups of chromosomes are similar in Zp.

Although *Zacco* and *Opsariichthys* exhibit different chromosome numbers and belong to different genera, interspecific mating behavior has been observed[20]. We indeed identified individuals displaying intermediate phenotypes between Zp and Op (refer to as H1 in Fig. 8), as well as between Zp and Oe (H2 in Fig. 8), in their natural habitats. Genome resequencing further revealed that individuals classified as H1 are indeed F1 hybrids of Zp and Op, and those classified as H2 are F1 hybrids of Zp and Oe. Furthermore, mitochondrial DNA from both Zp and Op is present in H1 individuals, and from both Zp and Oe in H2 individuals, indicating that mating between these species is bidirectional (Table S9).

DISCUSSION

Lineage-specific L2 expansion facilitates chromosomal evolution in Opsariichthyini

Through comparative analysis of opsariichthyin (Asian river chub) genomes and gene flow between species, this study provides rare evidence linking TE activities to chromosome evolution and species diversification (Fig. 9A). We first found that the expansion of L2 is specific to the Opsariichthyini within Cyprinoidei(Fig. 4B and Fig. S12A). L2 elements are enriched in the putative centromeres (Fig. 9B), where reduced recombination likely aids their accumulation over time[34,35]. In addition, L2 harbors sequence motifs that are preferential for Gypsy, thereby guiding the insertion of Gypsy into these motifs within centromeric regions (Fig. S11B). Furthermore, LTRs also tend to insert into the 3'UTR of other LTRs, especially Gypsy[36]. These characteristics create a positive feedback loop, leading to a high concentration of both L2 and Gypsy in the centromeres.

Active TE accumulation may trigger mutagenic processes, such as unequal exchange and replication slippage in centromere[37], which may further increase repetitive sequences including centromeric DNA. This is evident in Zp, where chromosomes that eventually underwent fission contain significantly more centromeric DNA than those in Cb. In addition, the centromere regions of these chromosomes have expanded (Fig. 9C), likely due to a recent Gypsy expansion (Fig. 6). According to kinetochore reproduction theory of chromosome fission, accumulation of centromeric DNA can lead to generation of additional kinetochore [13,14] (Fig. 9D). The expansion of centromere regions may further increase the likelihood of additional kinetochore formation.

If none of the extra kinetochores are inactivated, descendants will inherit dicentric chromosomes with two functional kinetochores. During replication, chromatids with monopolar spindle attachment segregate intact, while those with opposing bipolar attachment undergo chromosome fission (Fig. 9D and Fig 1B of ref 14). Of the 15 identified breakpoints, 10 coincide with putative centromeres in Cb (Fig. S3), while the remaining five, initially non-centromeric in Cb, became centromeres in Zp via pericentric inversions (Fig. 5), resulting in all 15 regions perfectly match centromeres in Zp (Fig. S4). Consequently, our results support the kinetochore reproduction theory of chromosome fission. Interestingly, the presence of active TEs associated with chromosome fission has also been observed in primate evolution (Supplementary Text: TEs in the breaking region Macaque chromosome 7)

Chromosome fission through kinetochore reproduction: reducing gene flow and promoting speciation

Rapid chromosome rearrangements are associated with increased species diversity. Between *Zacco* and *Opsariichthys*, gene flow in breaking regions is reduced, with fission chromosomes showing significantly lower migration estimates and higher divergence than non-fission ones (Fig. 7). Notably, within the Xenocyprididae, the average number of species per genus is approximately 3.4[40], but *Opsariichthys* comprises 15 nominal species, with numerous additional species yet to be named[23]. These patterns are consistent with the chromosome speciation model, in which chromosome rearrangements may reduce gene flow, which in turn drives speciation.

Chromosome rearrangements can reduce gene flow either through underdominance selection or by lowering recombination rates on rearranged chromosomes[10]. The observation of higher recombination rates in the fissioned chromosomes of Op and Oe (Fig. S15) suggests that a reduced recombination might not be the primary mechanism driving reductions in gene flow. Consequently, this supports the greater likelihood of underdominance selection as the primary mechanism. However, the underdominance model struggles to account for the prevalence of fissioned chromosomes within species, especially given the detrimental effects of heterokaryotypes.

According to the kinetochore reproduction theory of chromosome fission, each of the two fissioned chromosomes, equipped with a kinetochore, can pair with their unfissioned homologs (forming trivalents) during meiosis in heterokaryotypic individuals (Fig. 9E, left). Furthermore, selection might favor the retention of fissioned over intact homologs, especially if smaller chromosomes are more likely than larger ones to segregate without error during cell division [15,41]. Viable interspecific hybrids between Zp and Op, and between Zp and Oe (Fig. 8), demonstrate that fissioned pairs segregate together from their unfissioned homologs. Therefore, chromosome fission followed by kinetochore reproduction may have no deleterious effect and the fissioned pairs would increase their frequency within local populations rapidly.

After fission, subsequent chromosome rearrangement, such as pericentric inversion, which are common in the Opsariichthin (Fig. 5), may contribute to centromere relocation (Fig. 9E, right). The relocation can cause difficulty in pairing between fissioned chromosomes and their unfissioned homologs, which increase the probability of genetic isolation among populations and eventually leads to speciation (Fig. 9F). This theory also explains the occurrence of very different diploid numbers in the Opsariichthini, despite similar fundamental numbers, and the observation that lower diploid numbers are primitive within the tribe (Supplementary Text: Frequent fission and fusion in *Opsariichthys* genomes).

The potential role of PIWI1 protein mutation in L2 expansion and genome evolution

Understanding the drivers of L2 expansion is crucial to unraveling its role in genome evolution. Although the exact mechanisms are still elusive, the V380I mutation in the PAZ domain of the PIWI1 protein is a key candidate for further investigation. This domain is essential for piRNA's binding and RNA-induced silencing complex (RISC) activation, crucial for transposon repression. A dysfunction in this domain can lead to ineffective RISC formation, impairing transposon silencing[43-45]. Notably, the V380I mutation, located near the nucleic acid-binding interface, correlates with recent L2 expansion across all studied species, including genera *Opsariichthys* and *Triplophysa*.

In the genus *Triplophysa* (Nemacheilidae), despite stable chromosome numbers (N = 24-25), there are notable variations in karyotypes and fundamental numbers[46], associated with L2 expansion (Fig. S12). This genus also shows a dramatic increase in species diversity, with 147 species compared to the family average of 16. This association suggests that the V380I mutation could significantly influence genome architecture and species diversity. Further study of the V380I mutation is necessary, as it may reveal how specific genetic alterations significantly affect genome architecture and species diversity.

CONCLUSION

This study provides unprecedented evidence supporting the kinetochore reproduction theory of chromosome fission and elucidate the complex interplay between TE activities, genomic structure, and speciation. The traditional chromosomal speciation model states that chromosomal rearrangements lead to reproductive isolation and subsequent speciation. However, the role of chromosome fission is paradoxical as heterozygosity for fissions could have large fitness cost due to the difficulty of properly segregating. The kinetochore reproduction theory offers a plausible mechanism for chromosome fission. It posits that the steps leading to fission, such as the production of additional kinetochores and the inheritance of chromosomes with dual functional kinetochores, do not negatively affect cell division. In addition, because the fissioned chromosomes pair readily with their corresponding homologs, fission should not have negative effect in subsequent generations. Despite this, the specific mechanisms driving the duplication of centromeric DNA, which facilitates formation of additional kinetochore, remain unclear. Through comparative analysis of opsariichthyin genomes, we find that lineage-specific TE accumulation in centromeres likely contributes to

centromeric DNA expansion which, may have enabled neokinetochore formation and subsequent chromosome fissions. These fissioned chromosomes showed lower migration rates and higher genetic divergence between species than their intact counterparts, highlighting their role as reproductive barriers that promote speciation.

Online Methods

Chromosomal Analysis

The species under studied are listed in Table 1. Chromosomal analyses were performed on actively dividing cells from tissues such as the anterior kidney, spleen, and liver of both male and female specimens, using the method described by Ojima et al. [19]. Approximately 16 hours before the experiment, a yeast solution (0.05g yeast powder, 0.05g glucose, 2ml ddH2O) was injected into the dorsal musculature. One hour before euthanasia, 0.5% colchicine in 0.8% NaCl solution (8 μL per 1 g of body weight) was administered intraperitoneally. Tissues were minced using a glass tissue grinder, mixed with 3ml of 0.56% KCl, and gently resuspended. The suspension was filtered through cheesecloth to remove debris and transferred to a 15ml tube. After swelling in the hypotonic solution for 30 minutes, the suspension was centrifuged at 1600 rpm for 5 minutes. The supernatant was discarded, and the pellet was resuspended in residual solution. Gradually, 5ml of fixative (ethanol: glacial acetic acid = 3:1) was added. After 20 minutes at room temperature and subsequent centrifugation, the cells were resuspended in fixative. This fixation and centrifugation cycle was repeated once more. Finally, the supernatant was discarded, leaving 1-0.5ml of fixative. To prepare slides, the cell suspension was pipetted onto slides from a height of 20-25 cm above a slide on an 80°C heat slab. Slides were then either heated at 80°C for 2-3 hours or left at room temperature overnight. Slides were treated with Trypsin-EDTA at 37°C for one minute, halted with 95% ethanol, rinsed with ddH2O, and stained with Wright Dye for 1-3 minutes. After washing with tap water, slides were dried and stained with 6% Giemsa solution for 5-10 minutes, then rinsed and examined. Chromosomes were classified into three types, metacentric, submetacentric, and sub-/telocentric-like, following the classification method of Ojima et al [19].

Sample Collection

Species, including *Candidia barbatus* (Cb), *Zacco platypus* (Zb), *Opsariichthys evolans* (Oe) and *Opsariichthys pachycephalus* (Op), used for genome sequencing and assembly are listed in Supplementary Table 1. Tissues for RNA sequencing were dissected from fish body and preserved in RNAlater™ Stabilization Solution (Cat: AM7021, Invitrogen, Thermo Fisher Scientific Inc.) at -80°C freezer. The other tissues were preserved in 100% ethanol at -80°C freezer. 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.

DNA and RNA sequencing

Approximately 30-50 mg of muscle tissue was processed using proteinase K and Buffer G2, which includes RNAse A, from the Genomic DNA Buffer Set (Cat: 19060, Qiagen). DNA extraction was performed using the Genomic-tip 20/G (Cat: 10223, Qiagen) following the manufacturer's instructions.

For Nanopore sequencing, 5-8μg of genomic DNA was prepared by end-repair and ligation using the KAPA Hyper Prep Kit (Cat#KR0961, Kapa Biosystems, Wilmington, MA, USA) according to the supplier's guidelines. The prepared DNA library was combined with LB and SQB buffer from the Nanopore Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies, UK), loaded into a flow cell (R9.4.1; FLO-MIN106), and sequenced using MinION devices over a period of 24-72 hours.

For Illumina sequencing, genomic DNA was fragmented with a Covaris 2 device and its concentration was determined using a Qubit 2.0 Fluorometer. A sequencing library for short reads was generated from 200 ng of fragmented DNA utilizing the TruSeq Nano DNA High Throughput Library Prep Kit (Cat. No.20015965) and subsequently sequenced on an Illumina X-ten by GENOMICS BIOSCIENCE & TECHNOLOGY CO., LTD, Taipei, Taiwan.

For RNA sequencing, tissues (5 to 30 mg) were dissected and homogenized in 600 μl of Trizol for RNA extraction. RNA was then isolated using the RNeasy® Plus Mini Kit (Cat: 74134, Qiagen), and its quality was assessed with the Agilent 2100 Bioanalyzer. A sequencing library for short reads was created from about 1.5μg of RNA with the TruSeq Stranded mRNA Library Prep Kit (Cat. No. 20020595) and sequenced on the Illumina NovaSeq 6000 by the services of GENOMICS.

Genome Assembly and Size Estimation

In this study, we assembled four genomes, Cb, Zp, Oe, and Op. The nanopore long reads were assembled using Flye v2.9.2-b1786 [47] with "-m 10000 -meta" parameter. The resulting assemblies were corrected using POLCA [48] on Illumina short reads.

To assemble a chromosome-level genome, we employed scaffolding techniques using HiC and HiRise. Samples were prepared following the protocol of Lieberman-Aiden, et al. [49]. In brief, chromatin was fixed using formaldehyde. The fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and the blunt ends were subsequently ligated. After ligation, the crosslinks were removed and DNA was purified. Any

unligated biotin was eliminated from the purified DNA, which was then fragmented to approximately 350 bp. Sequencing libraries were generated with NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads prior to PCR enrichment for each library. These libraries were then sequenced on an Illumina HiSeq X platform. After sequencing, Dovetail HiC library reads were processed with Juicer [50] to generate high-resolution contact maps. These maps were then served as the input file for 3d-dna [51] to anchor the address of contigs. The final visualization of the anchored contigs was conducted using Juicebox [52].

Genome size was estimated by Jellyfish [53] with parameter *k=17*. For estimating genome completeness, we blasted with "Actinopterygii" ortholog set, which consists of 3640 core genes, by using Benchmarking Universal Single-Copy Orthologs v5 (BUSCO v5) [54]. The corrected nanopore reads were mapped back to the assembled genome to calculate the mapping rate using BWA v0.7.17-r1188 [55].

Transposable Elements Annotation

Transposable elements (TEs) in the genome were identified using both *de novo* prediction and homology search methods. For *de novo* prediction, we primarily used the repeat library retrieved from RepeatModeler v2.0.2 [56] with default parameters, supplementing it with LTR_FINDER_parallel v1.2 [57], LTR_FINDER v1.07 [58] and LTR_retriever v2.9.0[59]. For the homology search, we initially used the script famdb.py from RepeatMasker v4.1.2-p1 [60] to obtain "Cyprinidae" repetitive sequences from Dfam (ver. 3.5, 22-Dec-2021) [61] and Repbase (ver. 20181026) [62]. Next, we downloaded repetitive sequences of *Ctenopharyngodon idellus* and *Danio rerio* from FISHTEDB (http://www.fishtedb.org/) [63]. After concatenating these libraries, RepeatMasker was used for genome masking. We also utilize the script *calcDivergenceFromAlign.pl* to calculate the Kimura distance for each TEs family and used *createRepeatLandscape.pl* to graph the repeat landscape. To plot the distribution and proportion of TEs across chromosomes, a custom Python script was used with a window size of 50k bp.

Characterizing TSDs and Protein Domains in LTR/Gypsy and LINE/L2

To investigate the insertion sites of LTR/Gypsy (Gypsy), we retrieved both the left and right target site duplications (TSDs) from the output of LTR_retriever. A custom Python script was then used to analyze sequences 30 base pairs upstream and downstream of the TSDs and calculated the nucleotide composition at each position.

To determine whether LINE/L2 (L2) contain TSDs of Gypsy, we followed the protocol of Goubert, et al. [64] to retrieve full-length L2. In brief, we first blast the young L2 sequences (Kimura distance < 5%) obtained from the TEs annotation step, against the reference genome. We then aligned the L2 sequences using MAFFT [65], generated a consequence sequence using *cons* script from EMBOSS [66], and manually trimmed the TSDs using the visualization tool AliView [67]. Sequentially, we used the script *getorf* to identify the potential open reading frames (ORFs), and then using InterProScan [68] to annotate the protein domains.

Gene Annotation

Braker2 [69] was used for the prediction of protein-coding genes. Protein-based and RNA-based prediction were predicted separately. Protein sequences from various fish species, including *Astyanax mexicanus*, *Carassius auratus*, *Chanos chanos*, *Clupea harengus*, *Danio rerio*, *Paramormyrops kingsleyae*, *Salmo salar* and *Scophthalmus maximus*, were downloaded from GenBank (Supplementary Text: Published genome used in the analysis). RNA sequences were derived from sequencing of tissues detailed in Supplementary Table 1. In order to annotate the published genome of Ob (NGDC, GWHBEIO00000000), the RNA sequences of Ob were downloaded from NCBI (SRR15969039). The integration of predictions was facilitated by TSEBRA [70], using the 'pref_braker1.cfg' configuration to prioritize RNA-based predictions. Subsequent filtering of preliminary gene predictions involved two criteria: 1) incompleteness of the prediction, identified by the absence of an initial methionine, a missing stop codon, or the presence of multiple stop codons within the sequence; and 2) sequences shorter than thirty-three amino acids.

Synteny analysis among species

Synteny analysis was conducted by MCscan [71] and Synima [72]. MCscan was used to visualize synteny among species. The fine scale synteny blocks were retrieved from the results of Synima and were further used to identify breaking regions. The breaking regions are the segment of the original chromosome that cannot be readily aligned with the sequences in two fissioned homologous chromosomes. These regions represent genomic areas where disruptions in synteny occur, likely due to structural changes or the loss of genetic material during the chromosome fission process.

Recombination rate estimation

Illumina short reads were mapped to the reference genome using NVIDIA Parabricks

v4.3.1 (fq2bam) [73]. Duplicate reads were removed using samtools rmdup [74]. Bcftools mpileup (-q 20) and call (-m) [74] was used to generate VCF files for each individual. These files were filtered using bcftools view to retain sites with depths either greater than five or less than the mean depth plus two standard deviations. Subsequently, VCF files from the same species were merged using bcftools merge, and variants with data missing in fewer than two individuals were kept using bcftools view (-i 'N_MISSING<2'). The VCF files were phased using using Beagle v5.4 [75]. Recombination rate estimation was carried out with LDhelmet v1.9 [76], following its recommended procedures with default settings. Recombination rates along the chromosomes were smoothed using the locally weighted scatterplot smoothing (LOWESS) method, utilizing the lowess function available in R version 4.3.3 [77] .

Candidate centromere sequences

The procedures for identifying centromere markers was based on the approach of Melters, et al. [26]. In summary, the short read assembler PRICE [78] was utilized to assemble reads from the dataset previously used for genome polishing. The script provided by Melters, et al. [26] was then used to identify high order repeats, which were considered likely candidates for centromere repeats due to their high abundance as tandem repeats in the genome. Due to variations in the length of candidate centromere sequences, they were aligned and visualized using the AliView software [67] to discern the simplest repeat units, with manual adjustments made to the sequences as necessary. These procedures were performed for each species, revealing that the identified centromere sequences were fundamentally similar across all species examined. Subsequently, the putative centromere sequences were blasted against their respective genomes to determine their positions and copy numbers on chromosomes.

Morphological analyses and hybrid identification

Seven metric characters were measured, including maxilla length, body depth, eye diameter, head length, snout length, predorsal length, and caudal peduncle depth. Each measurement was standardized by dividing by the standard length. Subsequently, stepwise discriminant analysis was performed using STATISTICA 6.0 software [79] to establish discriminant functions. Two-dimensional scatterplots were then generated to visualize the relationships between the discriminant functions.

Genomes of Zp, Op, Oe, and individuals identified as hybrids based on morphological analyses were sequenced. Illumina short reads were mapped to the Zp reference genome

using NVIDIA Parabricks v4.3.1 (fq2bam) [73]. Duplicate reads were removed using samtools rmdup [74]. VCF files for each individual were generated and merged following the same procedure used for recombination rate estimation. SNPs that are fixed within the same species but vary between species were retained. These SNPs were analyzed using a triangle plot using R v4.3.3[77].

Migration rate estimation

We used gImble [33] to investigate the impact of chromosome fission on gene flow, particularly focusing on migration rates between Zp and Oe, as well as between Zp and Op. Two individuals were included for each species. Following the guidelines outlined in the gImble protocol, we used *bwa mem* to align reads from various samples to the reference genome of *Zacco platypus*. BAM files generated for the different comparisons were called for variants and subsequently merged into a single VCF file using the *freebayes-parallel* script. We prepared the dataset for gImble analysis with the *gimbleprep* script, and tested various models, including strict isolation (DIV), migration only (MIG), and isolation with migration (IM), using only the intergenic regions. Since gImble assumes a constant rate of unidirectional gene flow, the migration model can be further divided by species A to species B (IM_AB and MIG_AB) and species B to species A (IM_BA and MIG_BA).

Single copy gene and genetic distance

We used Orthofinder 2.5.4 [80] to identify single-copy genes (SCGs) between Zp and each of Cb, Op, Oe, and Ob. For Zp and Cb, we found 13,719 SCGs (Fission = 9,123, Nonfission = 4,596). Zp and Op yielded 13,604 SCGs (Fission = 8,401, Nonfission = 5,203), Zp and Oe resulted in 12,358 SCGs (Fission = 8,200, Nonfission = 4,158), and Zp and Ob had 13,880 SCGs (Fission = 9,220, Nonfission = 4,660). These SCGs were aligned using MACSE v2.06 [81]. The p-distance was calculated using a custom Python script, which divided the number of nucleotide differences by the total number of nucleotides compared.

Data availability

All data in this study has been deposited to NCBI bioproject under accession PRJNA1159521 (Cb), PRJNA1160456 (Zp), PRJNA1192024 (Op), PRJNA1192752 (Oe) and PRJNA1203011 (Hybrids). The details were listed in Supplementary Tables and Supplementary Information.

Competing interests

The authors declare no competing interests.

Contributions

TYW, KTS, SMC, and HYW conceived the idea. JHT, TYW, GCM, THY, CFW, TYL, SPH, FYW, and YT collected samples. TYW, GCM, THW, CHH, and HTY performed the experiment. JHT, TYW, CFW, YWW, and THY performed the analysis. KTS, SMC and HYW supervised the project and secured the fundings. JHT, TYW, THY, KTS, SMC, and HYW wrote the manuscript.

Acknowledgement

The authors wish to express their gratitude to Professor Chi-Chang Liu for providing valuable samples. Additionally, we thank the NGS High Throughput Genomics Core at the Biodiversity Research Center, Academia Sinica, for their support in library preparation and Hi-C sequencing. We are deeply grateful to Professor Shau-Ping Lin and Professor Isheng Jason Tsai for their invaluable advice throughout this research.

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). This study was also facilitated by the National Key Area International Cooperation Alliance: University Academic Alliance in Taiwan (UAAT) - Kyushu-Okinawa Open University (KOOU) - Medicine and Life Sciences Integrative Program. Supported by the Ministry of Education, Taiwan, the program fosters international collaboration in cutting-edge research.

Reference

1 Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C. & Volff, J.-N. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Research* **16**, 203-215, doi:10.1007/s10577-007-1202-6 (2008).

2 Brown, J. D. & O'Neill, R. J. Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *Annu Rev Genomics Hum Genet* **11**, 291-316, doi:10.1146/annurev-genom-082509-141554 (2010).

3 Klein, S. J. & O'Neill, R. J. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res* **26**, 5-23, doi:10.1007/s10577-017-9569-5 (2018).

4 Masly, J. P., Jones, C. D., Noor, M. A., Locke, J. & Orr, H. A. Gene transposition as a cause of hybrid sterility in Drosophila. *Science* **313**, 1448-1450, doi:10.1126/science.1128721 (2006).

5 Volff, J.-N., Körting, C., Meyer, A. & Schartl, M. Evolution and Discontinuous Distribution of Rex3 Retrotransposons in Fish. *Molecular Biology and Evolution* **18**, 427-431, doi:10.1093/oxfordjournals.molbev.a003819 (2001).

6 Ray, D. A., Xing, J., Salem, A.-H. & Batzer, M. A. SINEs of a Nearly Perfect Character. *Syst. Biol.* **55**, 928-935, doi:10.1080/10635150600865419 (2006).

7 Carbone, L. *et al.* Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *PLoS Genet.* **5**, e1000538, doi:10.1371/journal.pgen.1000538 (2009).

8 Ayala, F. J. & Coluzzi, M. Chromosome speciation: humans, Drosophila, and mosquitoes. *Proc Natl Acad Sci U S A* **102 Suppl 1**, 6535-6542, doi:10.1073/pnas.0501847102 (2005).

9 White, M. J. D. *Modes of speciation*. (W.H. Freeman, 1978).

10 Navarro, A. & Barton, N. H. Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution* **57**, 447-459, doi:10.1111/j.0014-3820.2003.tb01537.x (2003).

11 Mackintosh, A. *et al.* Chromosome Fissions and Fusions Act as Barriers to Gene Flow between Brenthis Fritillary Butterflies. *Molecular Biology and Evolution* **40**, doi:10.1093/molbev/msad043 (2023).

12 Ruckman, S. N., Jonika, M. M., Casola, C. & Blackmon, H. Chromosome number evolves at equal rates in holocentric and monocentric clades. *PLOS Genetics* **16**, e1009076, doi:10.1371/journal.pgen.1009076 (2020).

13 Todd, N. B. Karyotypic fissioning and Canid phylogeny. *J. Theor. Biol.* **26**, 445-480, doi:https://doi.org/10.1016/0022-5193(70)90096-2 (1970).

14 Kolnicki, R. L. Kinetochore reproduction in animal evolution: cell biological explanation of karyotypic fission theory. *Proc Natl Acad Sci U S A* **97**, 9493-9497,

doi:10.1073/pnas.97.17.9493 (2000).

15 Godfrey, L. R. & Masters, J. C. Kinetochore reproduction theory may explain rapid chromosome evolution. *Proc Natl Acad Sci U S A* **97**, 9821-9823, doi:10.1073/pnas.97.18.9821 (2000).

16 Huang, S. P., Wang, F. Y. & Wang, T. Y. Molecular Phylogeny of the Opsariichthys Group (Teleostei: Cypriniformes) Based On Complete Mitochondrial Genomes. *Zool Stud* **56**, e40, doi:10.6620/zs.2017.56-40 (2017).

17 Wang, H.-Y., Wang, C.-F., Du, S.-Y. & Lee, S.-C. New insights on molecular systematics of opsariichthines based on cytochrome b sequencing. *J. Fish Biol.* **71**, 18-32, doi:https://doi.org/10.1111/j.1095-8649.2007.01515.x (2007).

18 Yu, X., Zhou, T., Li, K., Li, Y. & Zhou, M. On the karyosystematics of cyprinid fishes and a summary of fish chromosome studies in China. *Genetica* **72**, 225-235, doi:10.1007/BF00116227 (1987).

19 Ojima, Y., Hayashi, M. & Ueno, K. Cytogenetic studies in lower vertebrates. X. Karyotype and DNA studies in 15 species of Japanese Cyprinidae. *The Japanese journal of genetics* **47**, 431-440, doi:10.1266/jjg.47.431 (1972).

20 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, doi:10.6620/zs.2020.59-6 (2020).

21 Wang, H.-Y., Lee, S.-C. & Yu, M.-J. Genetic evidence to clarify the systematic status of the genera Zacco and Candidia. *Zool. Stud.* **36**, 170-177 (1997).

22 Takai, A. Karyotype and Chromosomal Distribution of C-band-Positive Constitutive Heterochromatin in Opsariichthys uncirostris (Cyprinidae, Pisces). *Memoirs of Osaka Shin-ai College* **53**, 1-5 (2019).

23 Peng, X., Zhou, J.-J., Gao, H.-D. & Yang, J.-Q. A new species of Opsariichthys (Teleostei, Cypriniformes, Xenocyprididae) from Southeast China. *ZooKeys* **1214**, doi:10.3897/zookeys.1214.127532 (2024).

24 Froese, R. & Pauly, D. *FishBase*. Vol. 1594 (WorldFish, 2024).

25 Xu, X. *et al*. Chromosome-Level Assembly of the Chinese Hooksnout Carp (Opsariichthys bidens) Genome Using PacBio Sequencing and Hi-C Technology. *Frontiers in Genetics* **12**, doi:10.3389/fgene.2021.788547 (2022).

26 Melters, D. P. *et al*. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology* **14**, R10, doi:10.1186/gb-2013-14-1-r10 (2013).

27 Graham, T. & Boissinot, S. The genomic distribution of L1 elements: the role of insertion bias and natural selection. *J. Biomed. Biotechnol.* **2006**, 75327, doi:10.1155/jbb/2006/75327 (2006).

28 Talbert, P. B. & Henikoff, S. What makes a centromere? *Exp. Cell Res.* **389**, 111895,

doi:https://doi.org/10.1016/j.yexcr.2020.111895 (2020).

29 Linheiro, R. S. & Bergman, C. M. Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLoS One* **7**, e30008, doi:10.1371/journal.pone.0030008 (2012).

30 Craig, N. L. in *Mobile DNA II* 1-11 (2007).

31 Houwing, S. *et al*. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**, 69-82, doi:10.1016/j.cell.2007.03.026 (2007).

32 Ma, J. B., Ye, K. & Patel, D. J. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* **429**, 318-322, doi:10.1038/nature02519 (2004).

33 Laetsch, D. R. *et al*. Demographically explicit scans for barriers to gene flow using gIMble. *PLOS Genetics* **19**, e1010999, doi:10.1371/journal.pgen.1010999 (2023).

34 Pereira, V. Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. *Genome Biol* **5**, R79, doi:10.1186/gb-2004-5-10-r79 (2004).

35 Kofler, R., Betancourt, A. J. & Schlotterer, C. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. *PLoS Genet.* **8**, e1002487, doi:10.1371/journal.pgen.1002487 (2012).

36 Jedlicka, P., Lexa, M., Vanat, I., Hobza, R. & Kejnovsky, E. Nested plant LTR retrotransposons target specific regions of other elements, while all LTR retrotransposons often target palindromes and nucleosome-occupied regions: in silico study. *Mob DNA* **10**, 50, doi:10.1186/s13100-019-0186-z (2019).

37 Balzano, E. & Giunta, S. Centromeres under Pressure: Evolutionary Innovation in Conflict with Conserved Function. *Genes (Basel)* **11**, doi:10.3390/genes11080912 (2020).

38 Giannuzzi, G. *et al*. Hominoid fission of chromosome 14/15 and the role of segmental duplications. *Genome Res* **23**, 1763-1773, doi:10.1101/gr.156240.113 (2013).

39 Rudd, M. K. *et al*. Comparative sequence analysis of primate subtelomeres originating from a chromosome fission event. *Genome Res* **19**, 33-41, doi:10.1101/gr.083170.108 (2009).

40 Froese, R. & Pauly, D. *FishBase 2000: concepts designs and data sources*. Vol. 1594 (WorldFish, 2000).

41 Schubert, I. & Oud, J. L. There Is an Upper Limit of Chromosome Size for Normal Development of an Organism. *Cell* **88**, 515-520, doi:10.1016/S0092-8674(00)81891-7 (1997).

42 Xu, X. *et al*. Three genome assemblies of Opsariichthys bidens from Yangzte River,

Pearl River and Qiantang River basins. *Scientific Data* **11**, 1110, doi:10.1038/s41597-024-03899-5 (2024).

43 Houwing, S. *et al.* A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. *Cell* **129**, 69-82, doi:10.1016/j.cell.2007.03.026 (2007).

44 Tóth, K. F., Pezic, D., Stuwe, E. & Webster, A. The piRNA Pathway Guards the Germline Genome Against Transposable Elements. *Adv. Exp. Med. Biol.* **886**, 51-77, doi:10.1007/978-94-017-7417-8_4 (2016).

45 Gu, S., Jin, L., Huang, Y., Zhang, F. & Kay, M. A. Slicing-independent RISC activation requires the argonaute PAZ domain. *Curr. Biol.* **22**, 1536-1542, doi:10.1016/j.cub.2012.06.040 (2012).

46 Niu, Y., Zhao, Q., Zhao, H., Ludwig, A. & Peng, Z. Karyotype and genome size of an endangered cavefish (Triplophysa rosa Chen & Yang, 2005). *J. Appl. Ichthyol.* **33**, 124-126, doi:https://doi.org/10.1111/jai.13209 (2017).

47 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540-546, doi:10.1038/s41587-019-0072-8 (2019).

48 Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology* **16**, e1007981, doi:10.1371/journal.pcbi.1007981 (2020).

49 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

50 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).

51 Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95, doi:10.1126/science.aal3327 (2017).

52 Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99-101, doi:https://doi.org/10.1016/j.cels.2015.07.012 (2016).

53 Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770, doi:10.1093/bioinformatics/btr011 (2011).

54 Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocals* **1**, e323, doi:10.1002/cpz1.323 (2021).

55 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv: Genomics* (2013).

56 Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451-9457, doi:10.1073/pnas.1921046117 (2020).

57 Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* **10**, 48, doi:10.1186/s13100-019-0193-0 (2019).

58 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265-W268, doi:10.1093/nar/gkm286 (2007).

59 Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons *Plant Physiology* **176**, 1410-1422, doi:10.1104/pp.17.01310 (2018).

60 Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, 4.10.11-14.10.14, doi:10.1002/0471250953.bi0410s25 (2009).

61 Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**, D81-89, doi:10.1093/nar/gkv1272 (2016).

62 Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11, doi:10.1186/s13100-015-0041-9 (2015).

63 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, doi:10.1093/database/bax106 (2018).

64 Goubert, C. *et al.* A beginner's guide to manual curation of transposable elements. *Mobile DNA* **13**, 7, doi:10.1186/s13100-021-00259-7 (2022).

65 Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

66 Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276-277, doi:10.1016/S0168-9525(00)02024-2 (2000).

67 Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276-3278, doi:10.1093/bioinformatics/btu531 (2014).

68 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240, doi:10.1093/bioinformatics/btu031 (2014).

69 Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation

with BRAKER. *Methods Mol Biol* **1962**, 65-95, doi:10.1007/978-1-4939-9173-0_5 (2019).

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

71 Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49-e49, doi:10.1093/nar/gkr1293 (2012).

72 Farrer, R. A. Synima: a Synteny imaging tool for annotated genome assemblies. *BMC Bioinformatics* **18**, 507, doi:10.1186/s12859-017-1939-7 (2017).

73 NVIDIA. (2024).

74 Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, doi:10.1093/gigascience/giab008 (2021).

75 Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics* **108**, 1880-1890, doi:https://doi.org/10.1016/j.ajhg.2021.08.005 (2021).

76 Chan, A. H., Jenkins, P. A. & Song, Y. S. Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLOS Genetics* **8**, e1003090, doi:10.1371/journal.pgen.1003090 (2012).

77 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2024).

78 Ruby, J. G., Bellare, P. & Derisi, J. L. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)* **3**, 865-880, doi:10.1534/g3.113.005967 (2013).

79 STATISTICA (Data Analysis Software System), Version 10. (2011).

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

81 Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Molecular Biology and Evolution* **35**, 2582-2584, doi:10.1093/molbev/msy159 (2018).

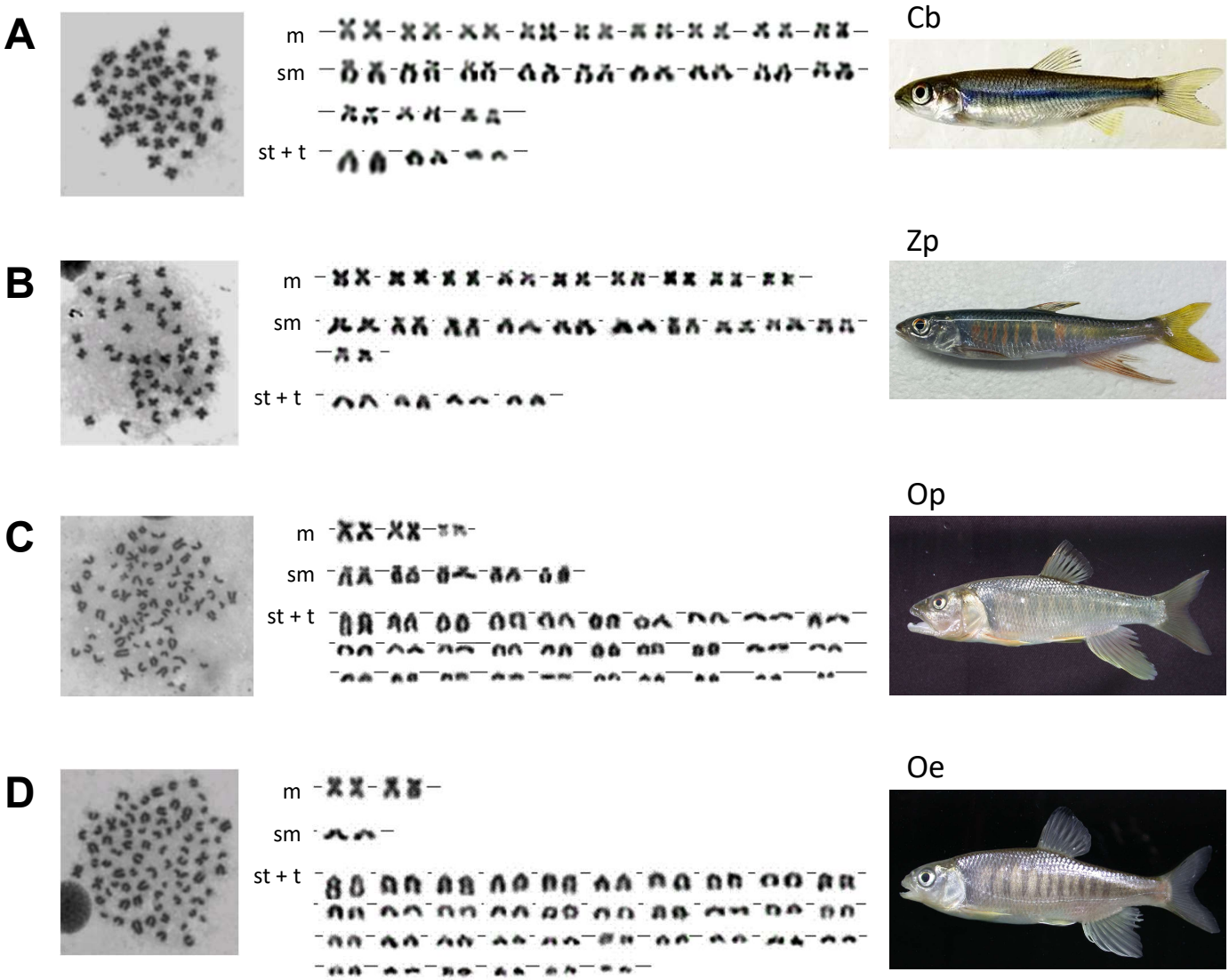


Figure 1. Karyotypes of opsariichthyins examined.
A, Cb (*Candidia barbatus*) **B,** Zp (*Zacco platypus*) **C,** Op (*Opsariichthys pachycephalus*) and **D,** Oe (*O. evolans*). m: metacentric; sm: submetacentric; st: sub-telocentric; t: telocentric.

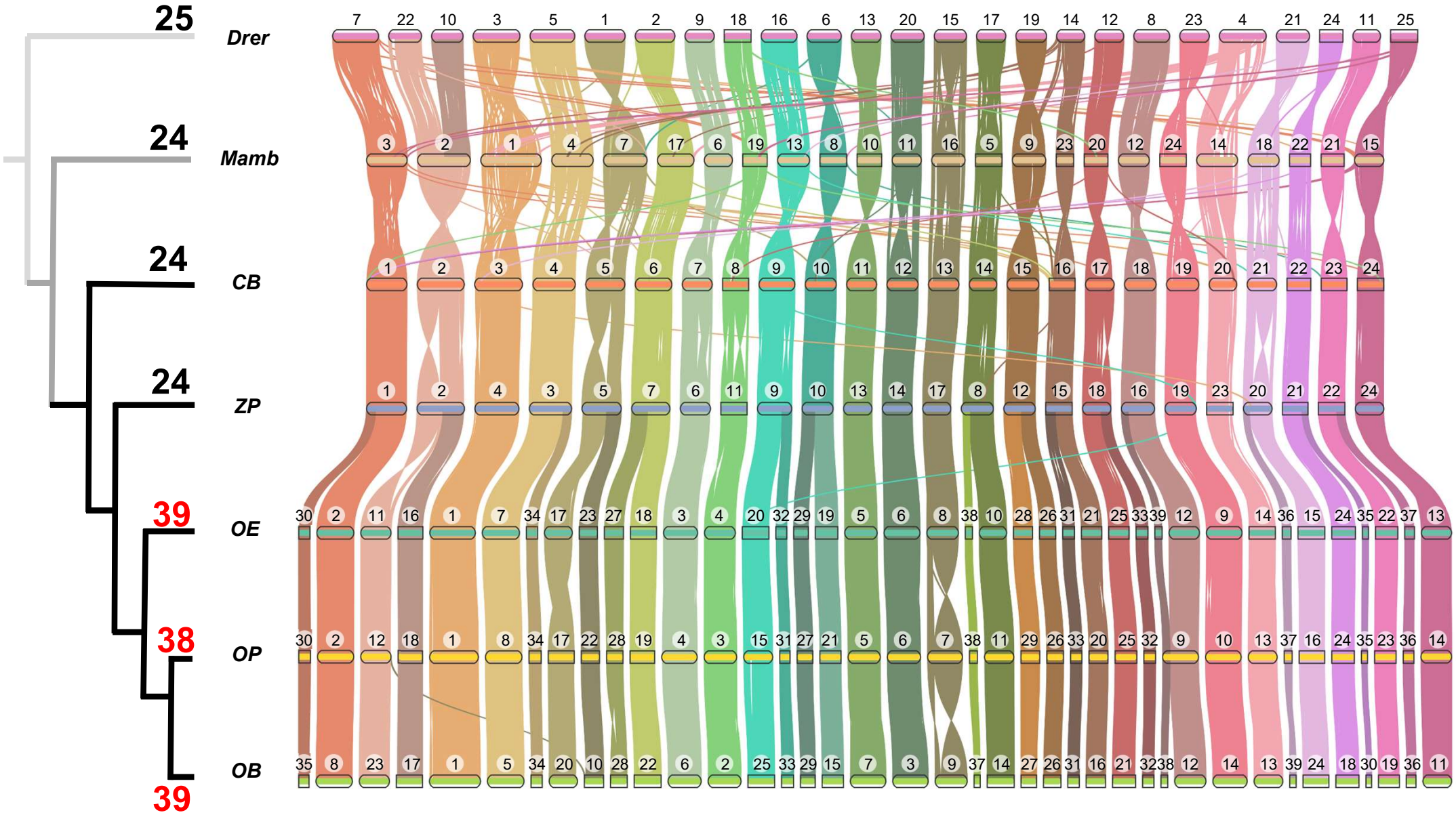


Figure 2. The phylogeny (left panel) and syntenic relationships (right panel) among Opsariichthyini tribe and the outgroups.
Dark black branches represent lineages of the Opsariichthyini tribe.
Drer: *Danio rerio*; Mamb: *Megalobrama amblycephala*; Cb: *Candidia barbatus*; Zp: *Zacco platypus*; Oe: *Opsariichthys evolans*; Op: *O. pachycephalus*; Ob: *O. bidens*

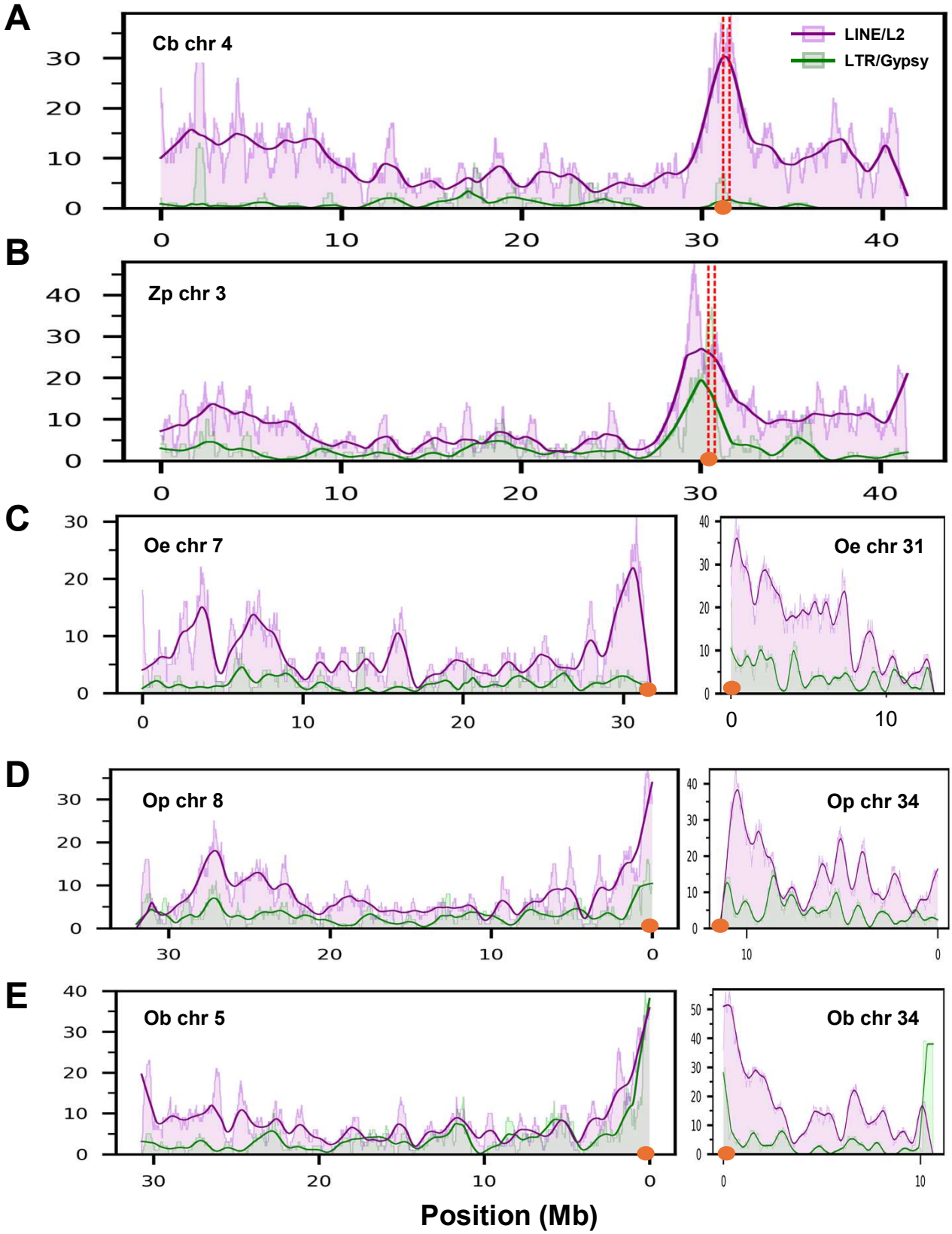


Figure 3. Represented chromosome synteny among studied species highlights L2 and Gypsy enrichment in breaking regions

A, Cb (*C. barbatus*) **B,** Zp (*Z. platypus*) **C,** Oe (*O. evolans*) **D,** Op (*O. evolans*) **E,** Ob (*O. bidens*). The dashed lines in Cb and Zp indicate breaking regions. The recently expanded L2 and Gypsy (k-distance < 5 %), shown in purple and green respectively, are predominantly localized around these breaking regions. The orange solid dot represents the putative centromere region. (Only chromosomes homologous to chromosome 3 of Zp are shown here, the rest are shown in Supplementary Fig. 2)

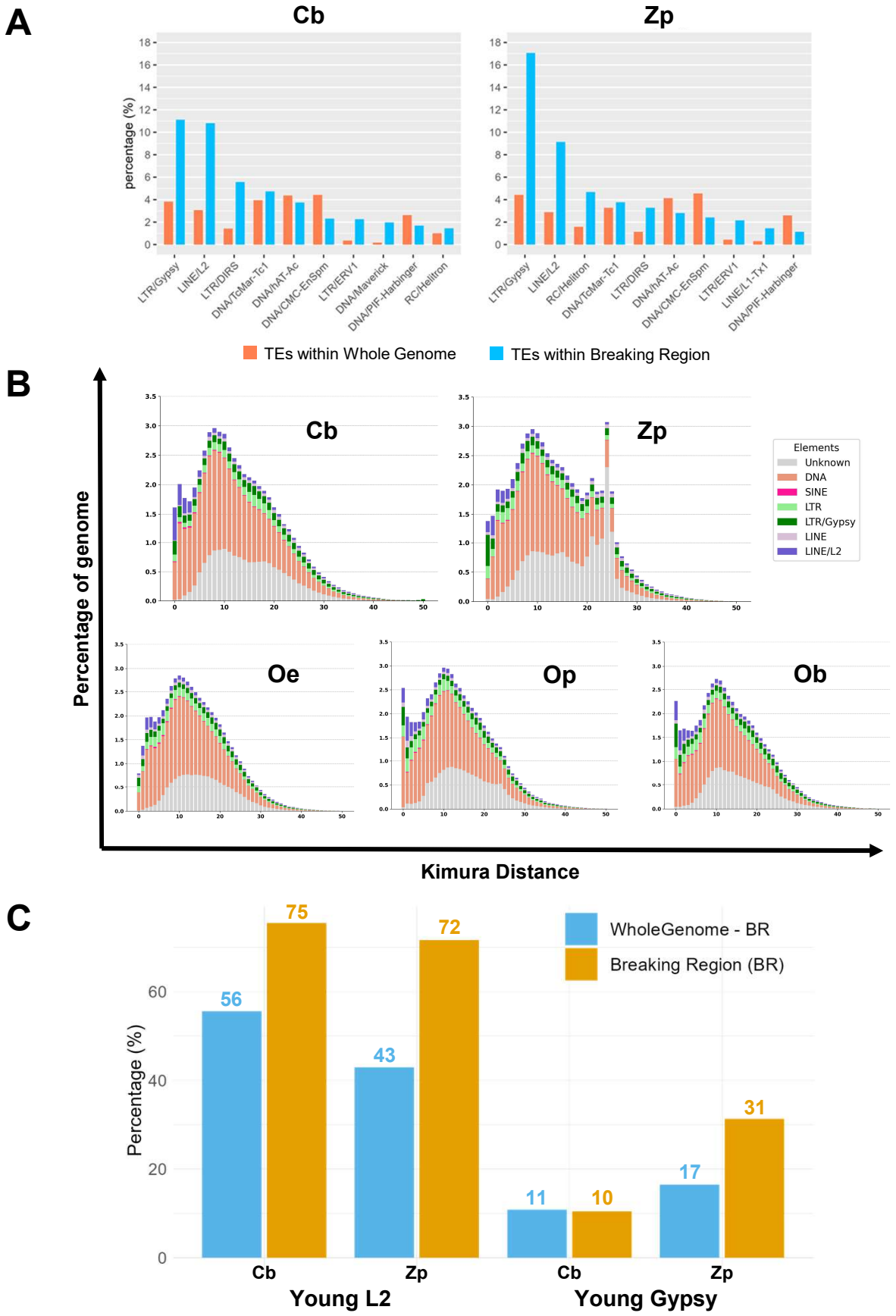


Figure 4. Distribution of transposable elements in the opsariichthyini genomes.
A, Proportion of different TEs in the whole genomes and breaking regions. In Cb (*C. barbatus*), the proportions of L2 and Gypsy in the breaking regions are 3.5- and 2.9-fold higher, respectively, than in the entire genome. In Zp (*Z. platypus*), they are 3.2- and 3.9-fold higher, respectively. **B,** Repeat landscapes among five opsariichthyini species. The y-axis represents the percentage of the genome occupied by each repeat element, and the x-axis represents the sequence divergence (measured by Kimura distance). **C,** Young L2 (Kimura distance < 5%) is enriched within the breaking regions of both Cb and Zp. In contrast, recent expansion of Gypsy within the breaking regions is more pronounced in Zp.

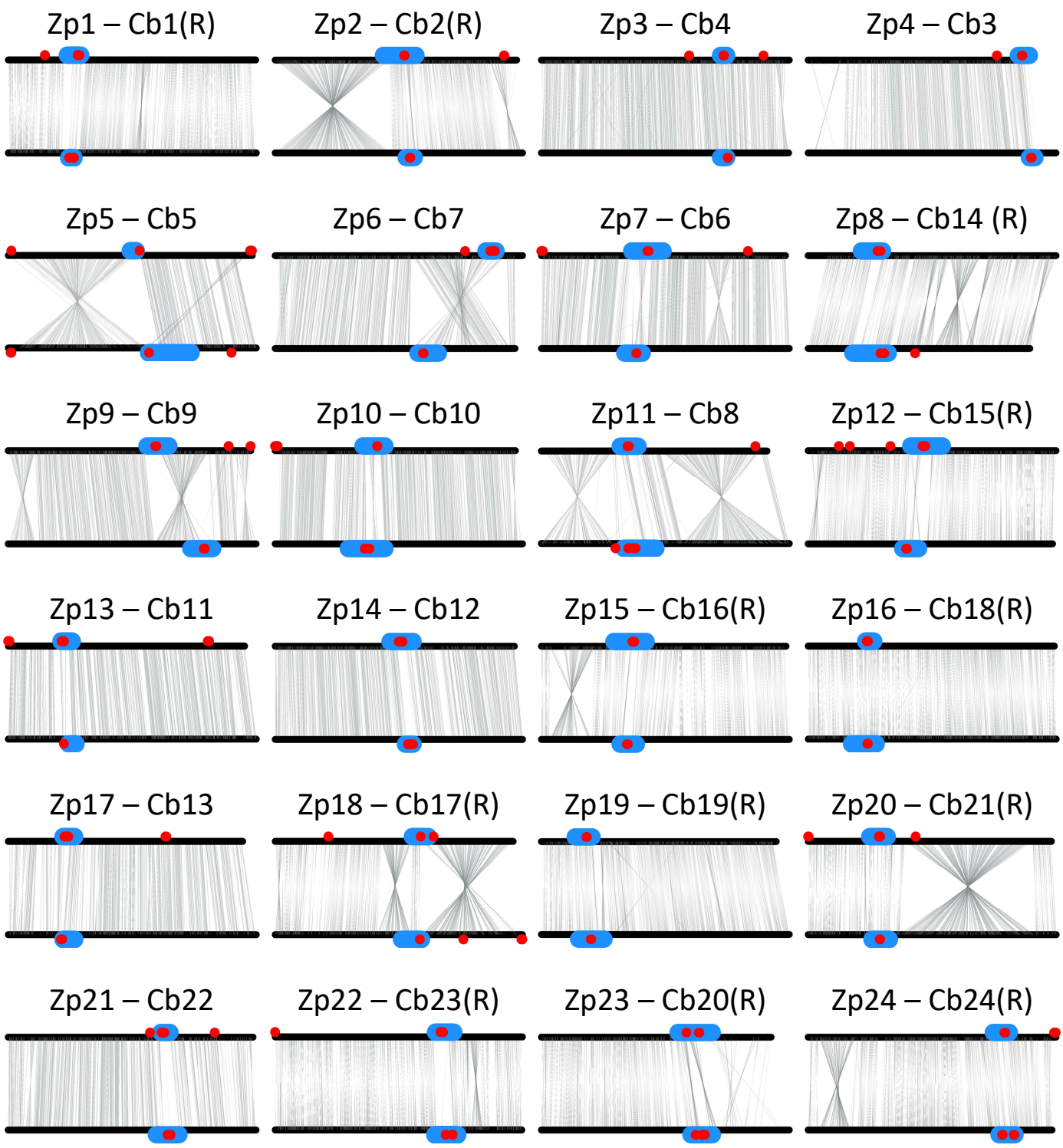


Figure 5. Syntenic relationships between Zp (above) and Cb (below) chromosomes highlight frequent pericentric inversions between species

The red dots are the centromere markers, and blue regions represent the putative centromere regions. The orientations of synteny blocks are based on those in Zp. In Cb, the notation "(R)" indicates that the synteny orientation is reversed.

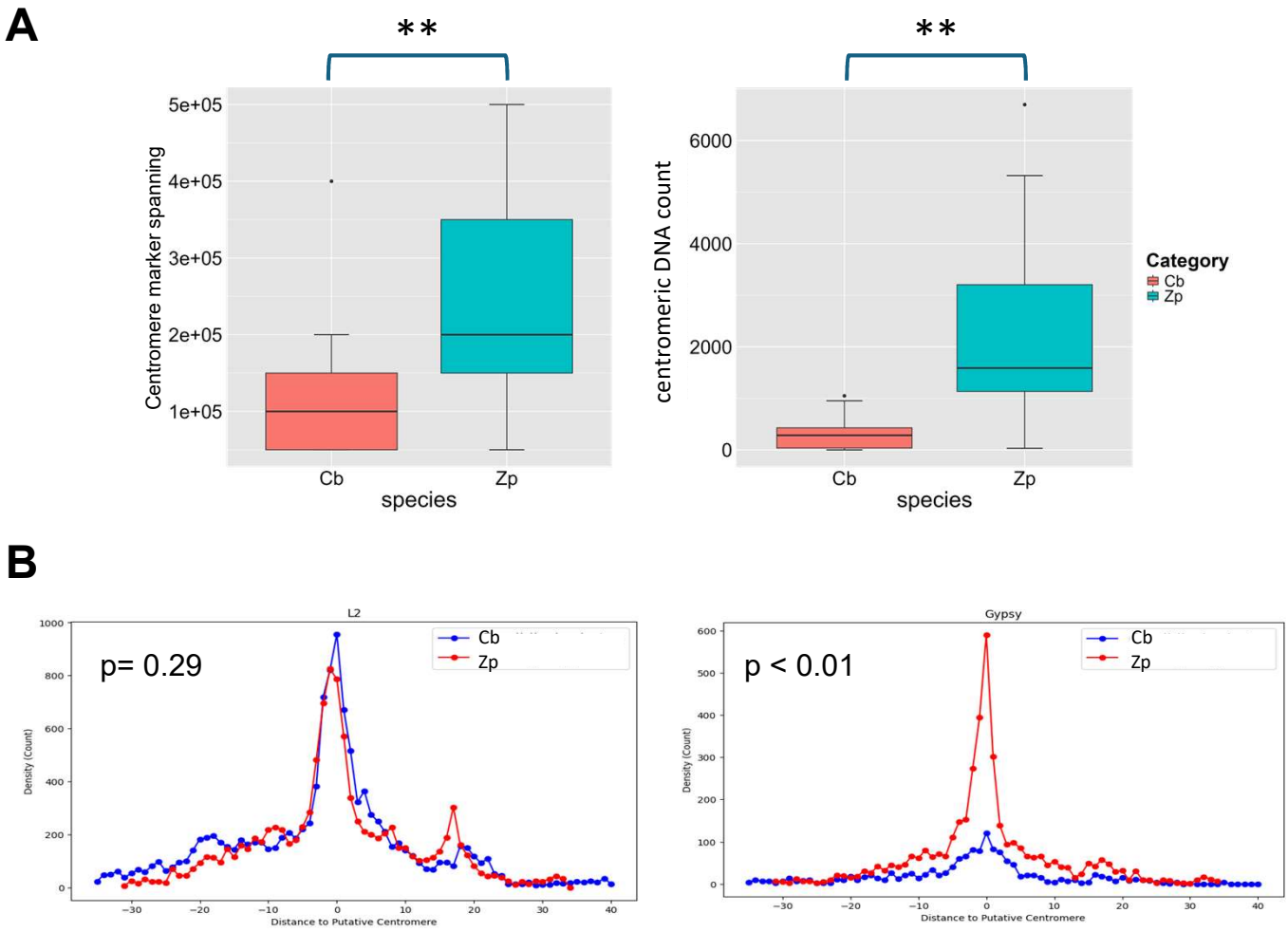


Figure 6. The distribution of centromeric DNA, L2, and Gypsy within putative centromeres of Cb and Zp in fissioned chromosomes.

A, The left panel displays the regions spanned by centromere markers, and the right panel shows their counts in fissioned chromosomes. **B,** The distribution of L2 (left panel) and Gypsy (right panel) around centromeres in fissioned chromosomes. X-axis represents distance to the center of putative centromere regions.

** $p < 0.01$

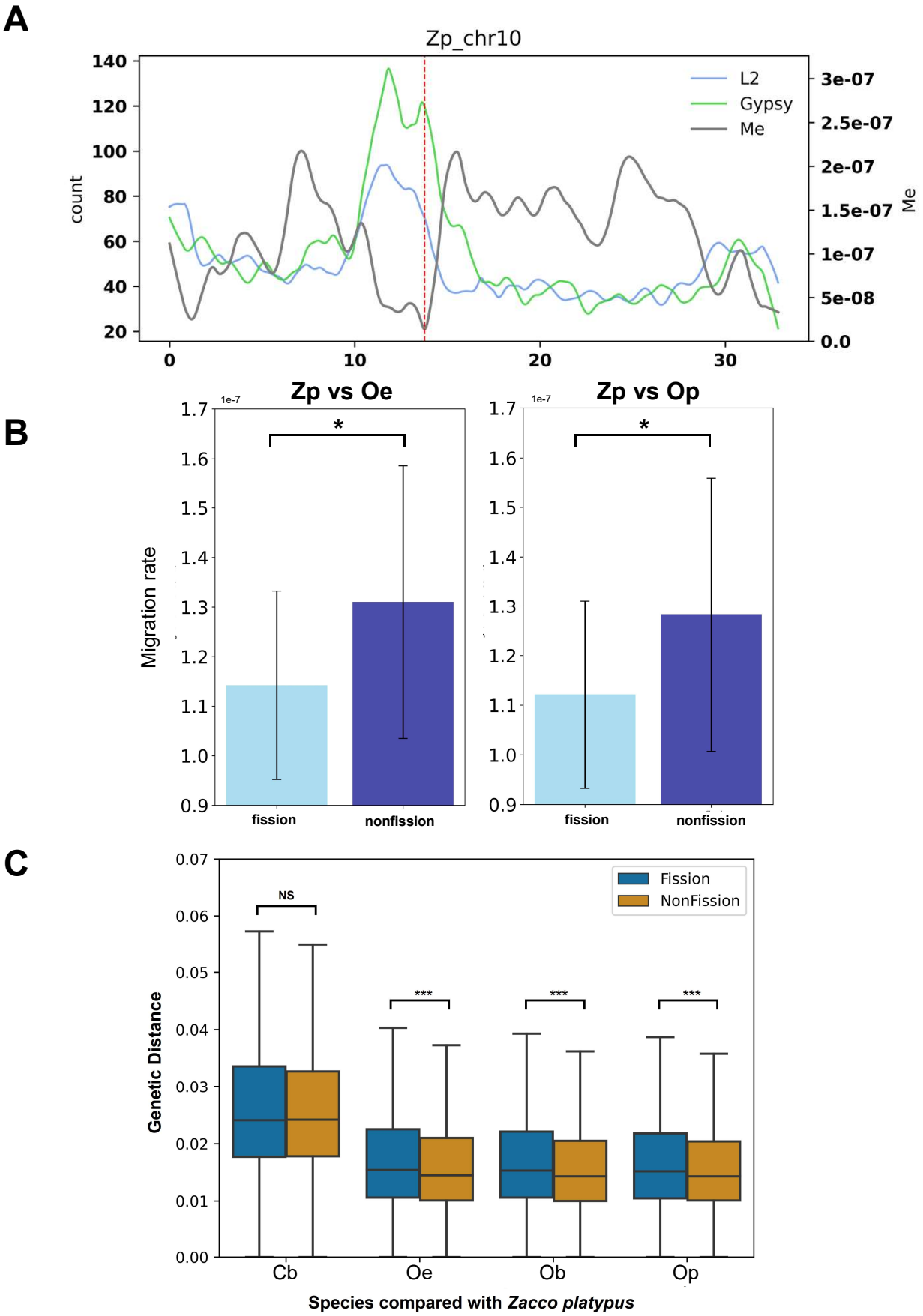
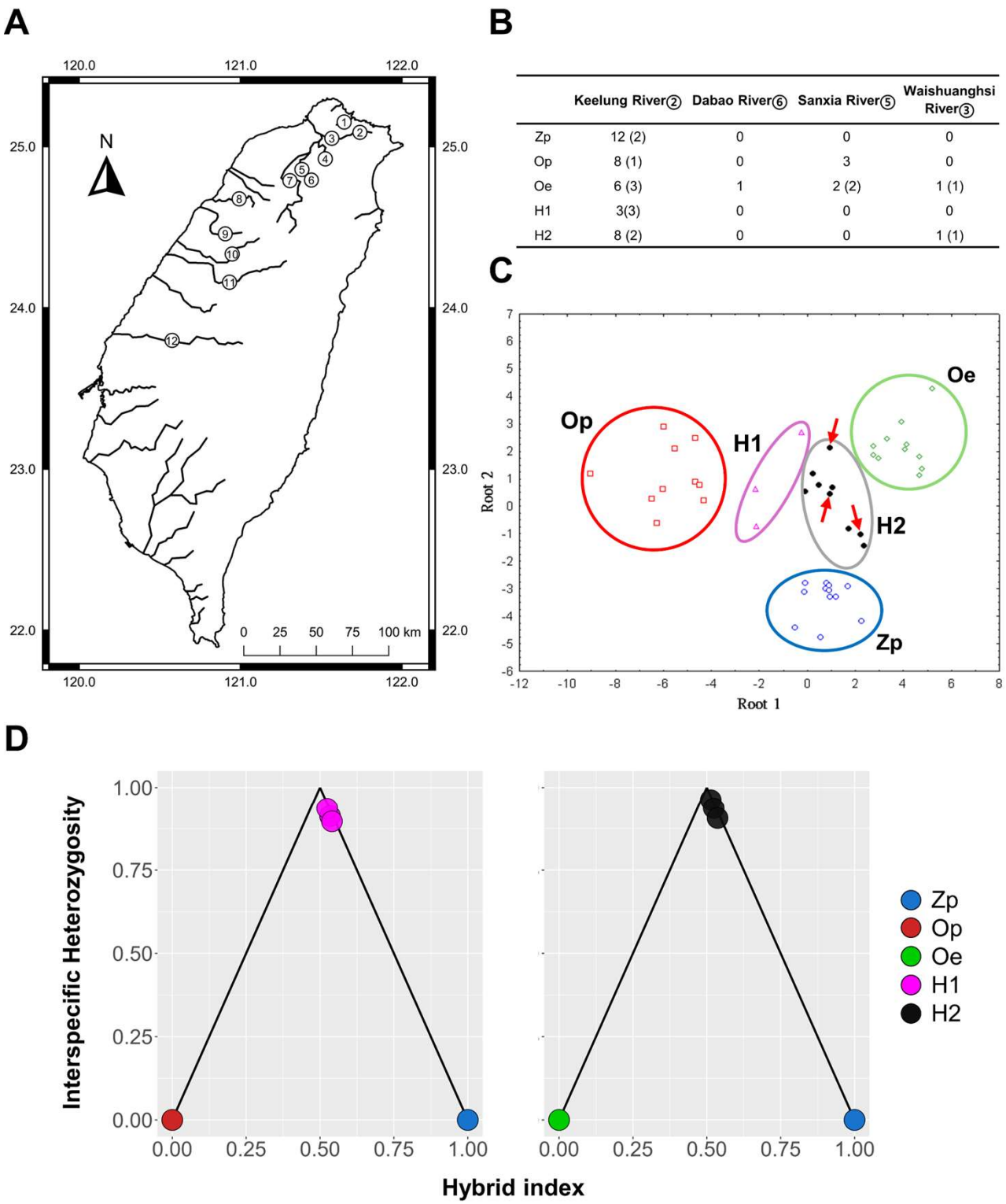


Figure 7. Gene flow was reduced in fissioned chromosomes.

A, Migration rate (Me) between Zp (*Z. platypus*) and Oe (*Opsariichthys evolans*) across various chromosomes, exemplified here by chromosome 10 of Zp, with additional data for other chromosomes presented in Fig. S8-S9. Red dash line indicates breaking regions. **B,** The average Me estimates for fission and non-fission chromosomes between Zp and Oe (left panel) and between Zp and Op (right panel). **C,** Average genetic distance of single copy genes on fission and non-fission chromosomes between Zp and each of Cb, Oe, Ob, and Op.

* $p < 0.05$; *** $p < 10\text{-}6$, NS: not significant.



	Keelung River②	Dabao River⑥	Sanxia River⑤	Waishuanghsi River③
Zp	12 (2)	0	0	0
Op	8 (1)	0	3	0
Oe	6 (3)	1	2 (2)	1 (1)
H1	3(3)	0	0	0
H2	8 (2)	0	0	1 (1)

Figure 8. Hybridization between Zp and Oe and between Zp and Op.
A, The map shows the sampling locations in Taiwan. **B,** Number of specimens collected at each site. Numbers in parentheses indicate those selected for genome sequencing. **C,** Scatter plots of Root 1 and Root 2 discriminant functions, illustrating group differentiation based on morphological traits. Circles: *Zacco platypus*(Zp); Squares: *O. pachycephalus*(Op); Diamonds: *O. evolans*(Oe); Triangles: Intermediate Type I(H1; putative hybrids between Zp and Op); Solid Circles: Intermediate Type II(H2; putative hybrid between Zp and Oe). Red arrows in H2 indicate samples used for genome sequencing. **D,** Triangle plots showing hybrid individuals between Zp and Op (H1; left panel) and between Zp and Oe (H2l right panel). Individuals identified as hybrids in **Figure 7c** are confirmed as F1 hybrids of the respective species. The rivers sampled are numbered from north to south as follows: 1) Masu River, 2) Keelung River, 3) Waishuanhsi River, 4) Xindian River, 5) Sanxia River, 6) Dabao River, 7) Dahan River, 8) Daping River, 9) Houlong River, 10) Daan River, 11) Dajia River, 12) Wu River, 13) Jhoushuei River and 14) Puzih River (See also Supplementary Table 9).

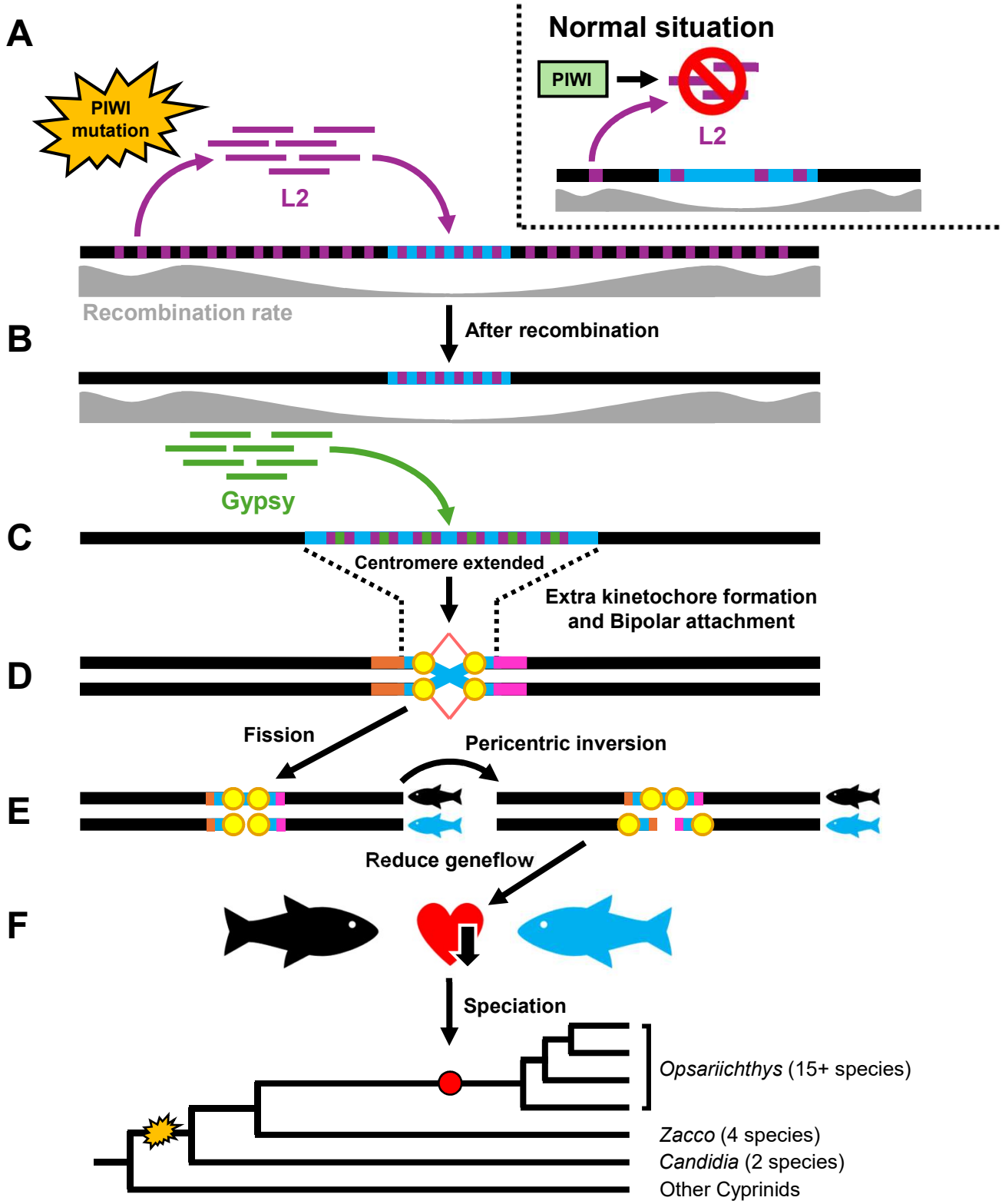


Figure 9. Transposable element expansion facilitates extensive chromosome fissions leading to substantial species diversification

A, Under normal conditions (inset at the top right), the PIWI1 protein suppresses TE activity. However, mutations in the PAZ domain of the *piwi1* gene in the ancestor of the Opsariichthyini enabled the expansion of L2 within the genome. **B,** Reduced recombination within centromeres likely contributed to the accumulation of L2. **C,** L2 elements harbor sequence motifs preferential for Gypsy, guiding its insertion into these motifs within centromeric regions. This led to centromere expansion. **D,** Centromere expansion may facilitate neokinetochore formation, resulting in dicentric chromosomes. During replication, dicentric chromosomes with opposing bipolar attachment undergo chromosome fission. **E,** Each of the two resulting chromosomes equips with a kinetochore, enabling proper segregation (left). Following fission, however, chromosome rearrangement such as pericentric inversion may contribute to centromere relocation (right). **F.** Such rearrangements can act as reproductive barriers, reducing gene flow and driving speciation. Consequently, the species diversity of *Opsariichthys* increased.

Table 1. Comparative karyotype of species used in this study.

Microscopic typing

Species	Sex		Karyotype				
	Male	Female	N	M	SM	ST/T	FN
Cb	1	1	24	9	12	3	45
Zp	2	1	24	9	11	4	44
Oe	2	2	39	2	1	36	42
Op	4	4	38	3	5	30	46
Ob*	-	1	39	2	2	35	43

In silico typing

Species	Sex		Karyotype				
	Male	Female	N	M	SM	ST/T	FN
Cb		1	24	9	11	4	44
Zp	1		24	9	9	6	42
Oe	1		39	2	1	36	42
Op		1	38	2	3	33	43
Ob*		1	39	1	1	37	41

* Xu X et. al. 2024