Complete sequencing of medaka genomes reveals the architecture of centromeric satellites, giant mobile elements, and sex chromosomes

Yoshihiko Suzuki, Kazuki Ichikawa, Yusuke Inoue, et al.

Genome Res. 2026 36: 1696-1707 originally published online July 1, 2026
Access the most recent version at doi:10.1101/gr.281813.125

Resource

Complete sequencing of medaka genomes reveals the architecture of centromeric satellites, giant mobile elements, and sex chromosomes

Yoshihiko Suzuki,[1,14] Kazuki Ichikawa,[1,15] Yusuke Inoue,[2,15] Chie Owa,[1] Haruka Kobayashi,[1] Kenji Morikami,[3] Ryohei Nakamura,[2] Takafumi Ikeda,[4,5] Manabu Fujie,[6] Mayumi Kawamitsu,[6] Nana Arakaki,[6] Eugene W. Myers,[7] Hiroyuki Aburatani,[8] Yusuke Takehana,[9] Masaru Matsuda,[10] Kiyoshi Naruse,[11,12] Shigehiro Kuraku,[3,13] Hiroyuki Takeda,[2,4,5] and Shinichi Morishita[1,14]

[1]*Department of Computational Biology and Medical Sciences, The University of Tokyo, Chiba 277-8561, Japan;* [2]*Department of Biological Sciences, The University of Tokyo, Tokyo 113-0033, Japan;* [3]*Department of Genomics and Evolutionary Biology, National Institute of Genetics, Shizuoka 411-8540, Japan;* [4]*Faculty of Life Sciences, Kyoto Sangyo University, Kyoto 603-8555, Japan;* [5]*Institute for Protein Dynamics, Kyoto Sangyo University, Kyoto 603-8555, Japan;* [6]*Sequencing Section, Core Facilities, Okinawa Institute of Science and Technology, Okinawa 904-0495, Japan;* [7]*Algorithms for Ecological and Evolutionary Genomics Unit, Okinawa Institute of Science and Technology, Okinawa 904-0495, Japan;* [8]*Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo 153-8904, Japan;* [9]*Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology, Shiga 526-0829, Japan;* [10]*Center for Bioscience Research and Education, Utsunomiya University, Tochigi 321-8505, Japan;* [11]*Laboratory of Bioresources, National Institute for Basic Biology, Aichi 444-8585, Japan;* [12]*Department of Basic Biology, School of Life Science, The Graduate University for Advanced Studies (SOKENDAI), Aichi 444-8585, Japan;* [13]*Department of Genetics, The Graduate University for Advanced Studies, Shizuoka 411-8540, Japan*

Medaka (*Oryzias latipes*) is a small freshwater teleost widely used as a vertebrate model organism. Existing medaka reference genomes, however, contain many gaps and unresolved repetitive regions, hindering precise genome annotation and comparative analyses. Here we present one complete and two near-complete genome assemblies for three inbred medaka strains derived from geographically distant populations. These assemblies provide a comprehensive view of highly repetitive sequences and chromosome-scale genome architecture in medaka. The fully resolved centromeres reveal an intriguing sequence organization characterized by short, distinct SF1+3 satellite arrays flanked by larger homogenized repeats. These short arrays are putatively hypomethylated and conserved across all acrocentric chromosomes, suggesting a functional role in centromere stability. The reconstructed 121 copies of the giant mobile element *Teratorn* retain complete genes of both a transposon and a herpesvirus, highlighting its unique persistence and impact on host genomes. Moreover, our assemblies reveal extensive structural divergence of medaka Y Chromosomes, yet identify a small (~24 kb) conserved region encompassing *Dmy* that may suffice for male determination. Collectively, these (near-)complete medaka genomes provide a powerful resource for exploring the biology of uncharacterized repetitive regions and the molecular basis of phenotypic diversity in vertebrates.

[Supplemental material is available for this article.]

The medaka fish (*Oryzias latipes*) has been studied for almost a century and was the first vertebrate in which the XY sex determination system was documented (Aida 1921). With its long research history, medaka has provided numerous insights into evolutionary, developmental, and reproductive biology (Takeda and Shimada 2010). More recently, the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel, 80 inbred strains derived from a wild medaka population (Leger et al. 2022), has provided a powerful resource for dissecting genotype–phenotype relationships, including epistatic and gene-by-environment interactions, thereby complementing human disease GWAS (Gierten et al. 2025; Welz et al. 2026). To further advance such studies, complete genome assemblies are urgently required to capture repetitive and structurally complex regions. In this study, we resequence the ~800 Mb genomes of three medaka inbred strains, Hd-rR, HNI, and HSOK, which originated from wild populations in southern and northern Japan and Korea, respectively (Fig. 1A).

The two Japanese strains (Hd-rR and HNI) diverged ~10 million years (MY) ago (MYA), and the Japanese and Korean (HSOK)

Present addresses: [14]Research Organization of Information and Systems (ROIS), Tokyo 105-0001, Japan; [15]Rhelixa Inc., Tokyo 104-0042, Japan
Corresponding authors: takeda_h@cc.kyoto-su.ac.jp, moris@edu.k.u-tokyo.ac.jp
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.281813.125. Freely available online through the *Genome Research* Open Access option.

36:1696–1707 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/26; www.genome.org

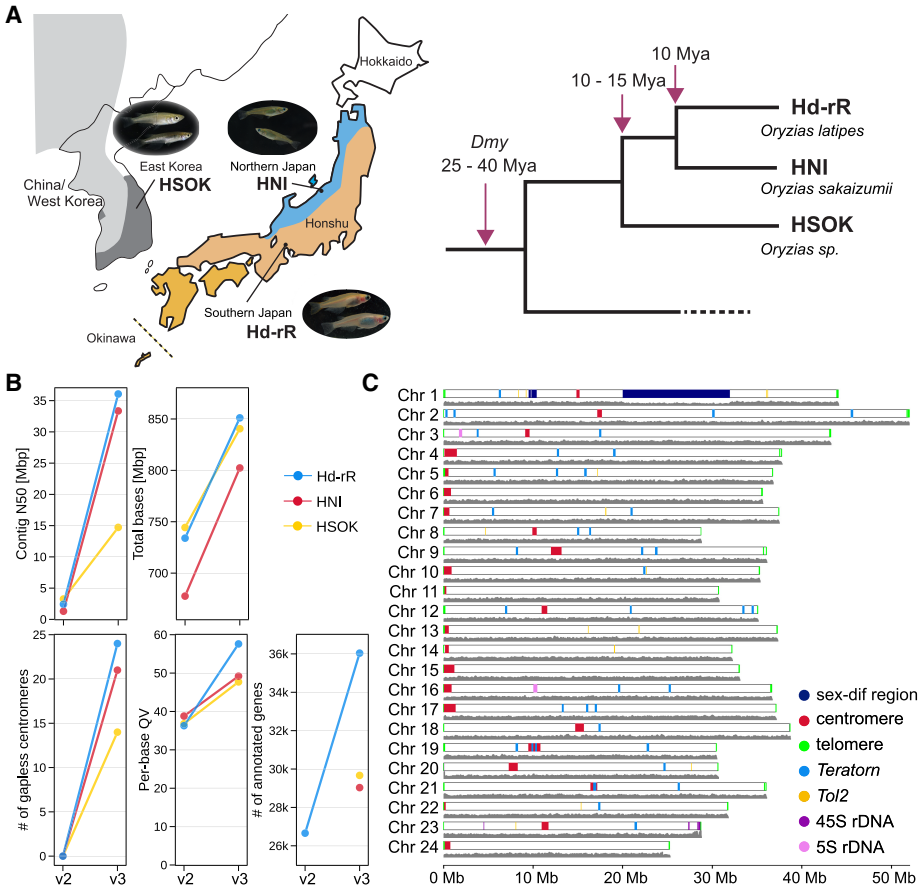

Figure 1. Geographic distribution and assembly improvements of medaka inbred strains. (*A*) Geographic distribution of wild medaka populations. The inbred strains Hd-rR, HNI, and HSOK were established from southern and northern Japanese populations and an eastern Korean population, respectively. Divergence times are based on a previous report (Yamahira et al. 2021), and the timing of the duplication/transposition event of (proto-)*Dmy* is inferred from the distribution of the *Dmy*-mediated sex-determination system within the genus *Oryzias*. (*B*) Assembly statistics on the v3 assemblies compared with the previous v2 assemblies. (*C*) Complete chromosomes of Hd-rR with annotations of important regions. Mean ONT-UL read depth (0–100×) in 10 kb bins is shown *below* each chromosome bar.

strains diverged ~10–15 MYA (Fig. 1A). Despite this level of divergence, these populations can still interbreed under laboratory conditions (Setiamarga et al. 2009; Takehana et al. 2016; Yamahira et al. 2021). Indeed, the two Japanese subpopulations have been proposed as distinct species (Asai et al. 2011) and exhibit marked physiological differences, including pronounced variations in their responses to photoperiod and temperature (Shinomiya et al. 2023). They are considered to be undergoing incipient speciation, representing a moderate evolutionary distance: genetically close enough to permit reliable alignment of noncoding sequences, yet divergent enough to harbor substantial sequence variation underlying phenotypic diversity (Takehana et al. 2003, 2004; Spivakov et al. 2014).

We reported draft genome assemblies of these strains in 2007 and 2017 (Kasahara et al. 2007; Ichikawa et al. 2017). However, it has been challenging to create a gapless, telomere-to-telomere (T2T) genome assembly because of highly repetitive regions including centromeres, telomeres, ribosomal DNA arrays, XY Chromosomes, and the giant mobile element *Teratorn* (Inoue et al. 2017). In this study, using state-of-the-art long-read technologies, Pacific Biosciences (PacBio) HiFi sequencing (Wenger et al. 2019) and Oxford Nanopore Technologies (ONT) ultralong sequencing (Shafin et al. 2020), we reconstruct a complete genome of the Hd-rR strain, as well as near-complete genomes of the HNI and HSOK strains. These (near-)complete medaka genomes uncover a comprehensive view of highly repetitive regions and provide unique insights into three previously inaccessible yet biologically crucial genomic regions: namely, centromeric sequences, giant transposons, and sex chromosomes.

Centromeres consist of tandem satellite arrays in medaka as in many other eukaryotes (Melters et al. 2013). Concerted evolution of centromeric satellite arrays through unequal crossover and gene conversion accelerates both sequence turnover and homogenization (Smith 1976; Schueler and Sullivan 2006), often giving rise to lineage- or species-specific sequence structures. In primates, centromeric sequence evolution follows the layered expansion model, in which new layers of homogenized arrays expand at the active core whereas older layers of divergent arrays are displaced outward (Shepelev et al. 2009; Altemose et al. 2022). Similar gradients of sequence similarity across a centromeric satellite array have also been observed in other animals and plants (Naish et al. 2021; Chen et al. 2023; Chaudhry et al. 2025). However, comprehensive understanding of how rapid sequence evolution is reconciled with conserved centromere function still requires complete assemblies of centromeric sequences from a wider diversity of species (Henikoff et al. 2001). In medaka,

centromeric sequences have been partially (~10%) resolved, providing only a glimpse into their genomic and epigenetic organization (Ichikawa et al. 2017). Comparative analyses of complete centromeric sequences across medaka strains reveal distinctive satellite sequences correlated with putative CpG hypomethylation, offering a basis to compare with other species and explore how centromeric satellites evolve within vertebrates.

Teratorn, a unique ~180 kb active mobile element that originated from the fusion of a *piggyBac*-like DNA transposon and a herpesvirus, is another underexplored component in medaka (Inoue et al. 2017). Although *Teratorn*-like transposons are widely distributed in the teleost lineage (Inoue et al. 2018), their evolutionary dynamics have remained unclear because of assembly difficulties, except for their identification as strong natural drivers of mutagenesis contributing to phenotypic diversity in mutants (Moriyama et al. 2012; Koita et al. 2025) and potentially driving speciation. To understand its impact, it is essential to elucidate the unique biology of its dual transposon-virus nature, particularly how such unique hybrid elements replicate, persist, and coevolve with their hosts.

The medaka DM-domain gene on the Chr Y (*Dmy* gene) was cloned as the first known nonmammalian male-determining gene analogous to *Sry* (Matsuda et al. 2002, 2007). A 43 kb duplication of Chr 9 containing *Dmrt1* (Kondo et al. 2006) was inserted into Chr 1 ~25–40 MYA (Fig. 1A), and the inserted copy soon evolved into the Y-specific (Y-sp) region with *Dmy*, initiating proto-Y Chromosome differentiation. Nevertheless, medaka XY Chromosomes remain cytogenetically homomorphic, unlike heteromorphic mammalian XY Chromosomes that arose 170–210 MYA (Graves 2006). Thus, medaka is considered to be at an early stage of sex chromosome evolution and a model suitable for studying transitions toward heteromorphy and the maintenance of *Dmy*. However, repetitive sequences have hindered contiguous assembly of medaka sex chromosomes.

Here, we first describe overall statistics and structural features of our medaka (near-)T2T assemblies and then discuss their significance and potential as genomic resources through initial analyses of the three aforementioned repetitive regions. These resources provide a foundation for future studies aiming to link genomic variation, including structural variation, to phenotypic diversity.

Results

Genome sequencing of three medaka strains

We extracted genomic DNA from three medaka strains (Hd-rR, HNI, and HSOK) and sequenced it using PacBio HiFi and ONT ultralong (ONT-UL) reads, which have complementary strengths. HiFi provides ~10× lower error rates (~0.1%) compared with ONT-UL (~1%), whereas ONT-UL produces ~10× longer reads (N10 ~150 kb vs. ~20 kb). HiFi libraries were prepared from whole-body samples, whereas ONT-UL libraries were prepared from pooled heart, spleen, and testis tissues to obtain DNA of sufficient length and yield. To facilitate the discrimination of sequences divergent between Chr X and Chr Y (which are also known as Chr 1) in medaka, HiFi reads were sequenced from male and female samples separately for each strain. The combined sequencing depth of HiFi reads was 68.4×–75.9×, and that of ONT-UL reads of size >100 kb amounted to 19.7×–32.1×. Additionally, we generated Hi-C long-range data for assembly validation and curation of Hd-rR and HNI. All sequencing statistics are summarized in Supplemental Table S1.

Genome assembly and gene annotation

For each strain, the initial set of chromosomal scaffolds was generated using Hifiasm (Cheng et al. 2024) and RagTag (Alonge et al. 2022). After manually correcting false inversions using Hi-C contact maps (Supplemental Fig. S1), we performed semiautomated curation of the assemblies to obtain a gapless, T2T complete genome for Hd-rR and nearly complete genomes for HNI and HSOK (Methods), which we refer to as the v3 assemblies to distinguish them from the v2 assemblies reported in 2017 (Ichikawa et al. 2017). In particular, we corrected large tandem repeats such as centromeric satellite arrays, whose sequences are prone to collapse during assembly, and extended telomeric sequences at both chromosomal ends (Supplemental Fig. S2). Additionally, sequences of sex-differentiating regions on Chr 1 and the mitochondrial genomes were identified during the curation process.

After the curation, the base-level quality value (QV) of the Hd-rR assembly improved from 49.1 to 57.6 (Fig. 1B; Supplemental Table S2), whereas the QV was limited to 36.3 in the previous v2 assembly. For HNI and HSOK, their assembly graphs were highly tangled (Supplemental Fig. S3), indicating that some repeats remain unresolved, and the assemblies for these two strains remained incomplete. Nevertheless, we anchored 98.8%/97.5% of the sequences to chromosomes with only 15/83 gaps in HNI/HSOK, compared with about 1000 gaps per strain in the v2 assemblies. The comparatively lower assembly contiguity in HSOK may be attributable to the relatively smaller amount of ONT-UL reads (Supplemental Table S1) and the higher copy number of the large transposon *Teratorn* (see below). In 45S and 5S ribosomal DNA arrays and adjacent megabase-scale tandem repeats, we observed sequence variation and read-depth fluctuations that may reflect sample variation, somatic variation, and/or residual uncertainty (Fig. 1C; Supplemental Figs. S4, S5; Supplemental Discussion). Overall, the present v3 assemblies represent a substantial improvement on reconstructing repetitive elements over the v2 assemblies (Fig. 1C; Supplemental Fig. S6).

We annotated protein-coding genes by integrating multiple gene prediction methods and external evidence (Supplemental Fig. S7). Specifically, gene models of the previous v2 Hd-rR assembly were lifted over to the v3 genomes and merged with de novo predictions with RNA-seq providing homolog hints. In total, 36,040 genes were determined for Hd-rR, and 29,032 and 29,668 genes were predicted for HNI and HSOK, respectively (Fig. 1B; Supplemental Table S3). The higher gene count in Hd-rR than in HNI and HSOK may reflect the larger amount of RNA-seq data available for Hd-rR (387 Gb compared with 69 Gb) and the resulting more comprehensive detection of lowly expressed transcripts. The BUSCO completeness (Manni et al. 2021) ranged from 99.6%–99.8% across the three strains, supporting the high quality of our medaka gene annotations.

Centromere localization and chromosome orientation

In the prior v2 medaka assemblies, only ~10% of centromeric sequences were inferred, whereas the v3 assemblies resolved all 24 centromeres in Hd-rR and yielded 21 and 14 gapless centromeres in HNI and HSOK, respectively. Compared with v2, centromere positions were newly determined or corrected for 42 chromosomes (Supplemental Table S4). The updated centromeric satellite locations were largely shared across the three strains and consistent with previous FISH images in Hd-rR (Ichikawa et al. 2017). Specifically, of the 24 chromosomes, 12 were acrocentric whose centromeres are located near telomere, and the other 12

were nonacrocentric, except that five HNI chromosomes (Chr 5, 6, 7, 10, 16) contained two distant centromeric satellite arrays at submetacentric and acrocentric positions, respectively. These putative pseudodicentric chromosomes were likely generated by large inversions, which were supported by multiple ONT-UL reads (Supplemental Figs. S8, S9; Supplemental Discussion).

Precise identification of centromeric locations enabled p/q-arm assignment and standardization of chromosome orientation across all strains by defining position 1 at the telomeric end of the p-arm. Compared with v2, 12 chromosomes including the sex chromosomes were reverse-complemented in each v3 assembly.

Comparative analysis of centromeric satellite monomers

Leveraging fully assembled centromeric sequences, for each chromosome of the three strains we clustered ~160 nt centromeric satellite monomers into representative monomers based on sequence similarity. As similarly defined in primates (Alexandrov et al. 2001), we further categorized these chromosomal representative monomers into suprachromosomal families (SFs), in which each SF is a group of monomers that share sequence similarity and span multiple chromosomes. We newly identified five SFs (SF1–SF5), whereas only four SFs were inferred from the fragmented v2 assemblies (Supplemental Fig. S10A–D).

In the phylogenetic tree of the 350 representative monomers belonging to SF1–SF5 (Fig. 2A; Supplemental Data File 1), monomers from different strains were intermixed, indicating that all three strains share the same set of SFs. We then tested whether representative monomer sequence similarity varies with strain relationship (intra- vs. interstrain) and chromosome type (acrocentric vs. nonacrocentric) and detected a strong effect of chromosome type on sequence divergence (Freedman–Lane permutation ANOVA, 20,000 permutations, $P = 5.0 \times 10^{-5}$), with no significant effect of strain relationship ($P = 0.190$) and no interaction ($P = 0.683$). Consistently, pairwise comparisons confirmed that monomers were significantly less diverged within acrocentric chromosomes than within nonacrocentric ones (two-sided Wilcoxon test, Bonferroni-corrected for all possible comparisons, $P < 1.0 \times 10^{-50}$) (Fig. 2B) as previously suggested (Ichikawa et al. 2017). We further identified 20 groups of highly similar (>97%) representative monomers shared between different acrocentric chromosomes in a genome, which we refer to as interchromosomal ultraconserved (ICUC) monomers (Fig. 2A,B). All ICUC monomers were found within individual strains and not across strains, as well as exclusively among acrocentric chromosomes (Supplemental Fig. S10E). These results support preferential sequence homogenization among acrocentric chromosomes, likely through interchromosomal gene conversions, as previously suggested in humans (Choo et al. 1989; Guarracino et al. 2023).

The composition of SFs as a set was generally conserved for each chromosome across the three strains, although individual array sizes varied (e.g., from 190 kb in Chr 22 to 1.4 Mb in Chr 4 in Hd-rR) (Fig. 2C), probably owing to unequal crossover and

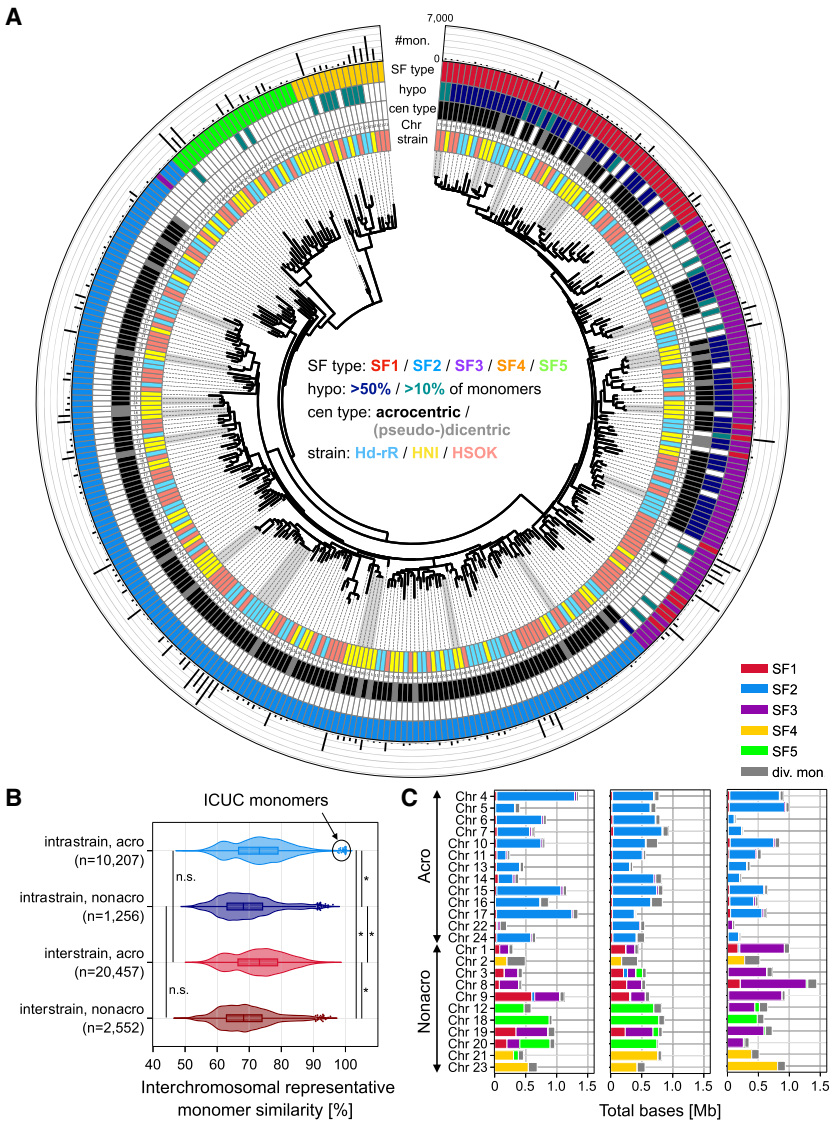

Figure 2. Compositional and phylogenetic landscape of centromeric satellites. (*A*) Phylogenetic tree of 350 representative monomers excluding divergent monomers. The *outermost* track is the monomer count per representative, and the "hypo" track indicates ONT-inferred CpG hypomethylation level (i.e., the proportion of hypomethylated monomers within each representative). Dark gray shades on the dotted lines connecting the tree leaves and the strain track indicate interchromosomal ultraconserved (ICUC) monomers. (*B*) Pairwise interchromosomal monomer sequence similarities for four categories by two aspects: intra- versus interstrain, and within acrocentric versus nonacrocentric chromosomes. Wilcoxon test; (*) $P < 1.0 \times 10^{-50}$, (n.s.) $P > 0.01$. (*C*) Centromeric satellite array lengths by chromosome and SF group. Acrocentric chromosomes are placed on the *upper* side and nonacrocentric chromosomes on the *lower* side.

replication slippage during meiosis (Smith 1976; Levinson and Gutman 1987). Although nonacrocentric chromosomes showed greater variability in SF composition, acrocentric chromosomes were consistently dominated by SF2 monomers with minor fractions of SF1 and SF3.

Positional distributions and methylation patterns in centromeric satellite arrays

Despite their low abundance in acrocentric chromosomes (Fig. 2C), SF1 and SF3 monomers displayed a unique positional and putative epigenomic organization distinct from other reported eukaryotic centromeres (Naish et al. 2021; Altemose et al. 2022; Chen et al. 2023; Chaudhry et al. 2025). Specifically, in the centromeric sequences of acrocentric chromosomes in Hd-rR, short composite arrays containing both SF1 and SF3 (hereinafter referred to as SF1+3 arrays) were consistently interspersed at multiple positions within large SF2 arrays, except on Chr 22 (Fig. 3A). In line with this structural organization, these SF1+3 arrays corresponded predominantly to hypomethylated regions inferred by ONT-UL reads, whereas the large SF2 arrays were almost entirely hypermethylated (Fig. 3A). The assembly of these arrays was validated using ONT-UL reads (Supplemental Fig. S11). In contrast, nonacrocentric centromeres were largely homogenized by a single SF or a composite of SF1+3, supporting independent and rapid concerted evolution of satellite arrays (Fig. 3B). These observations were common in HNI and HSOK as well (Supplemental Figs. S12, S13).

We detected putative CpG hypomethylated regions in centromeric sequences, termed centromere dip regions (CDRs), although future experimental validation will be required. Each chromosome had at least one CDR, whereas some nonacrocentric chromosomes exhibited relatively lower dip levels (Fig. 3A,B). Of note, in acrocentric chromosomes, 93.1% of monomers in the interspersed SF1+3 arrays were hypomethylated at CpG sites, whereas 99.7% of monomers in the large SF2 arrays were hypermethylated (Figs. 2A, 3A; Supplemental Fig. S14). Among the acrocentric SF1 and SF3 monomers, we further identified sequence motifs enriched in hypomethylated monomers (Fig. 3C; Supplemental Fig. S15; Supplemental Table S5). Specifically, GCATATTGTACATAAAAA (18 nt) was the significant motif enriched in hypomethylated SF1 monomers (86.8%, $n=81$, χ^2 test, Bonferroni-corrected, $P=4.8\times10^{-6}$). In SF3, ATAGCGT (7 nt) containing a CpG site was the significant motif, although its enrichment was smaller than that of SF1 (62.9%, $n=72$, χ^2 test, Bonferroni-corrected, $P=3.9\times10^{-4}$). These results suggest that putative CpG hypomethylated regions in acrocentric chromosomes are characterized by SF1+SF3 composite arrays within SF2 arrays and by distinct sequence motifs.

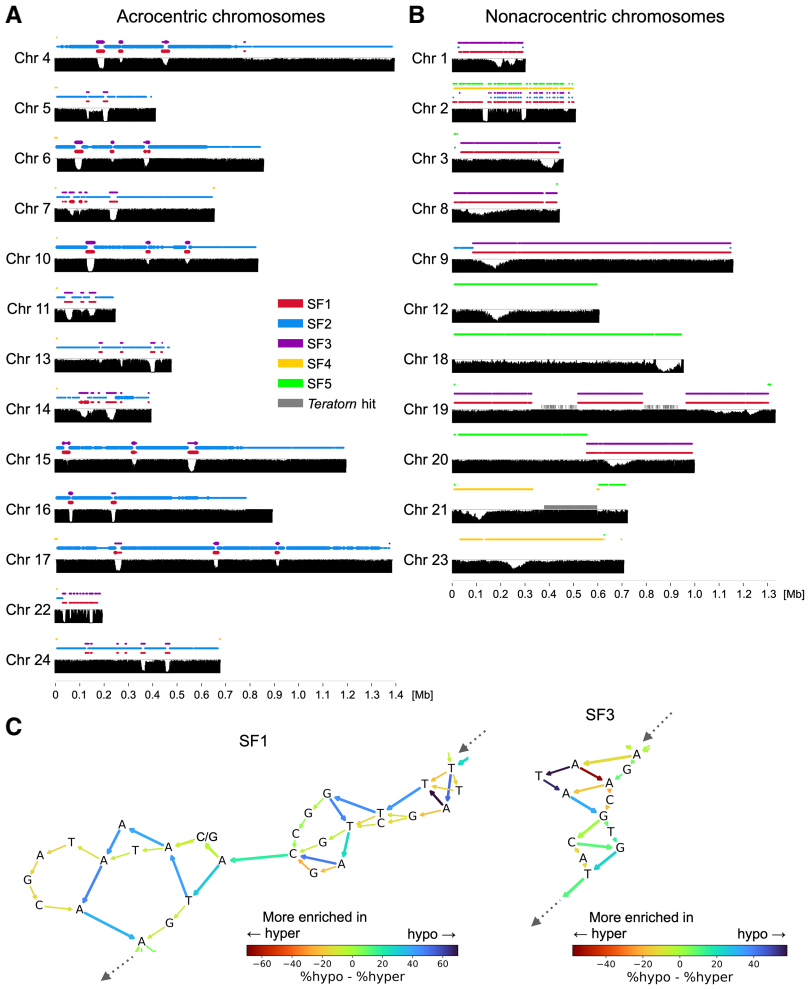

Figure 3. Positional landscape of sequence and CpG methylation status of centromeric satellites in the complete Hd-rR genome. (*A*,*B*) Positional distributions of centromeric monomers colored by SF group, along with average ONT-inferred CpG methylation levels from zero to 100, for acrocentric (*A*) and nonacrocentric (*B*) chromosomes of the T2T Hd-rR assembly. Thick lines indicate ICUC monomers. For HNI and HSOK, see Supplemental Figures S12 and S13. (*C*) Graph representations of representative monomers for SF1 and SF3, for which monomer sequences were progressively aligned and sequence variations appear as branching paths, shown around the sequence motifs enriched in putatively hypomethylated monomers. Edges are colored by relative enrichment in hypo- and hypermethylated monomers.

Structural rearrangements among three medaka strains

We detected large-scale structural variation between the three nearly complete genomes (Fig. 4). In addition to the previously reported large (>10 Mb) inversion on Chr 11 in Hd-rR, we identified large subtelomeric inversions in HNI on five chromosomes causing division of centromeric satellite arrays. We also identified large (~20 Mb) highly nonsyntenic regions in the distal part of Chr 2 (Fig. 4; Supplemental Fig. S8), most of which were unassembled in previous assemblies.

Pairwise comparisons identified 21,495 SVs (insertions, deletions, and inversions >50 bp) between Hd-rR and HNI compared with 18,713 SVs between HNI and HSOK and 18,229 SVs between Hd-rR and HSOK. These SV counts did not simply scale with strain divergence time. Interpreting this pattern is limited by technical factors such as assembly quality and alignment sensitivity that

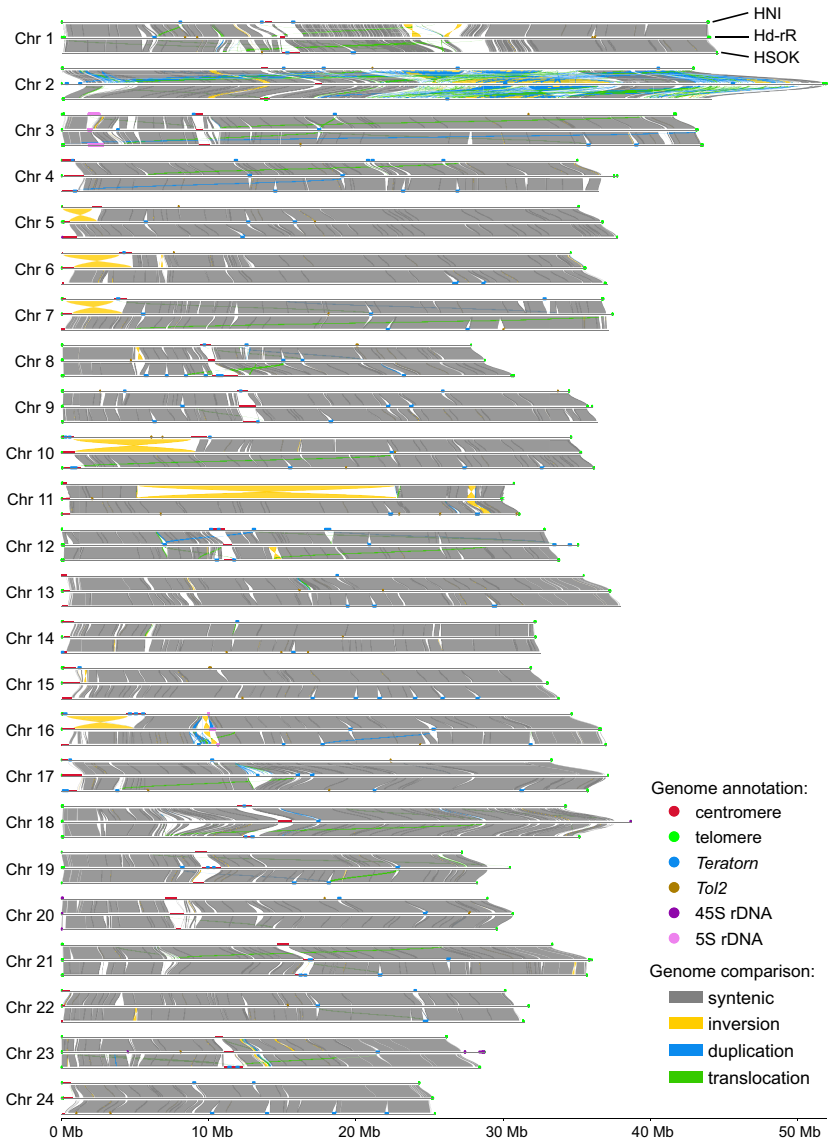

Figure 4. Structural rearrangements and transposons. Structural rearrangements among the three medaka strains with annotations of major repetitive elements. Strains are ordered as HNI, Hd-rR, and HSOK to clearly illustrate that megabase-scale inversions associated with divided centromeric arrays are specific to HNI. For Chr 11, however, the order is Hd-rR, HNI, and HSOK to highlight a large inversion specific to Hd-rR.

vary with genome divergence. Moreover, pairwise genome comparisons cannot unambiguously assign individual SVs to specific evolutionary lineages. Additional high-quality outgroup genomes will be needed to place each SV on a particular branch of the medaka phylogeny.

Evolution of *Teratorn* copies

The giant (180 kb) DNA transposon *Teratorn* has two subtypes, which share ~88% sequence identity within their coding regions, with an ~80 kb inversion in the middle of subtype 2 (Supplemental Fig. S16A). Our new assemblies revealed the complete architecture of *Teratorn*, showing that the copy number and composition of each subtype varied among strains, ranging from 35–84 copies for subtype 1 and one to 15 copies for subtype 2 (Supplemental Fig. S16B; Supplemental Table S6). Overall sequence structures were conserved among copies except for small indels and long-repeat expansion in some copies (Fig. 5; Supplemental Figs. S16C,D, S17).

Phylogenetic analyses demonstrated that *Teratorn* copies generally clustered by strain and followed the topology of host genes (Fig. 5; Supplemental Fig. S18; Supplemental Data File 2), supporting independent expansion and vertical inheritance of *Teratorn* within each strain, although herpesviruses are generally transmitted horizontally without being integrated into the host genome. *Teratorn* could produce virus particles as it contains the whole genome of the alloherpesvirus species (Inoue et al. 2017). Consistent with this, horizontal transfer, although possibly less frequent, was also suggested by the shorter branch length of the phylogenetic trees between Hd-rR and HNI for subtype 2 compared with that of host genes (Fig. 5; Supplemental Fig. S18).

Most *Teratorn* copies retained coding capacity; namely, the genes essential for viral replication and propagation are intact (Supplemental Fig. S19A). In addition, the nonsynonymous-to-synonymous mutation rate (d_N/d_S) of herpesvirus genes was much lower than one for almost all copies (Supplemental Fig. S19B), suggesting that herpesvirus genes have undergone purifying selection. Indeed, there seemed to be subclusters with higher proportion of full-length *Teratorn* copies (especially for subtype 1 in Hd-rR) (Fig. 5), implying that full-length copies exhibit higher propagation capacity. In contrast, medaka *Tol2*, one of the conventional transposons, showed extremely high sequence identity between the copies of each strain and also between the three strains (less than two mutations among 4678 nt aligned regions) (Supplemental Fig. S20) as reported previously (Ichikawa et al. 2017). *Tol2* is thus likely to have recently invaded into medaka populations widely distributed in southeast Asia (Koga et al. 2000). Taken together, although the underlying mechanism remains elusive, the transposase gene and the herpesvirus genes may be coordinately utilized for the propagation of *Teratorn*, presumably through virus-like particles in medaka.

Evolution of the Y Chromosome

In our new assemblies, we clearly defined, for the first time, the Y-differentiating (Y-dif) region as a genomic segment largely homologous to Chr X but showing sequence or structural divergence.

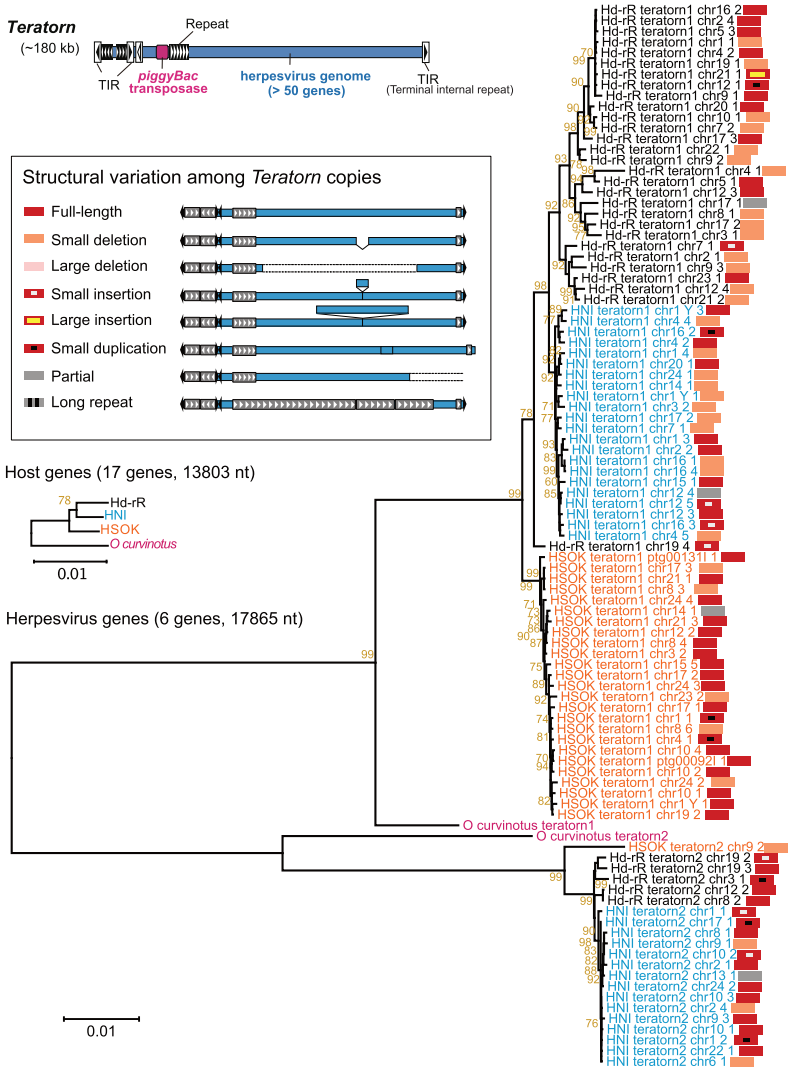

Figure 5. Evolutionary relationships of *Teratorn* copies. Maximum-likelihood trees based on six concatenated herpesvirus genes (17,865 nt in total) and 17 host genes (13,803 nt). Branch values indicate support from 100 replicates. Structural categories of *Teratorn* copies are shown next to *Teratorn* copy names.

The X/Y-dif regions were identified by reduced male read depth in X-dif (Fig. 6A) and by almost no female reads in Y-dif (Supplemental Fig. S21). We focused on Hd-rR and HNI because the HSOK sex chromosomes were less complete, although the sequence accuracy around the *Dmy* locus in HSOK was supported by raw reads (Supplemental Fig. S21E). In HNI, the Y-dif region spans ~23.6 Mb and contains 858 genes, extending beyond the centromere. In Hd-rR, the primary Y-dif region spans ~11.6 Mb and contains 596 genes; it does not extend to the centromere, but instead, two additional small Y-dif regions are present on opposite sides of the centromere (Supplemental Tables S7, S8). The expansion of Y-dif regions along the chromosome is consistent with theoretical predictions (Kirkpatrick and Guerrero 2014). Within the Y-dif region, the portions flanking the Y-sp region exhibit extensive structural divergence, especially on the telomeric side, including large inversions, indels, and long stretches of repetitive sequences (Fig. 6B). Beyond these flanking regions, a certain degree of diversification was broadly detected over the Y-dif region (Fig. 6A), indicating that the structural differentiation extends far beyond the Y-sp region.

A previous study estimated the Y-sp region of HNI to be ~258 kb, having expanded from the original 43 kb segment derived from Chr 9 (Kondo et al. 2006). In our assemblies, we precisely determined the Y-sp region in the three strains (250 kb in HNI, 507 kb in Hd-rR, and 351 kb in HSOK) and identified strain-specific patterns of internal repeats (Supplemental Fig. S22). Despite extensive divergence, comparison of the Y-sp regions revealed a small, structurally conserved region surrounding *Dmy*, spanning 53 kb (from 0.2 kb upstream of exon1 to 3.1 kb downstream of exon6) in HNI, 31 kb in Hd-rR, and 24 kb in HSOK. We refer to these conserved intervals as the candidate "minimum *Dmy* cassette" (Fig. 6C). Although further experimental validation will be required to confirm its regulatory activity, this conserved cassette could be sufficient for *Dmy* to function as the sex-determining gene.

Discussion

We uncovered a unique centromere landscape in medaka that is distinct from other eukaryotes analyzed so far. Particularly, the short SF1+3 composite arrays appear to have been conserved across all acrocentric chromosomes among strains since the divergence of Hd-rR and HNI from HSOK ~10–15 MYA, despite the typically rapid turnover of centromeric sequences. Although colocalization of multiple types of centromeric satellites occurs in other organisms such as primates and plants (Naish et al. 2021; Logsdon et al. 2024), these generally form large single arrays and are not conserved among (acrocentric) chromosomes. The 18 nt and 7 nt motifs enriched in hypomethylated SF1 and SF3 monomers may be analogous to the mammalian CENP-B box (Masumoto et al. 1989; Kipling et al. 1995). Multiple hypomethylated regions observed in medaka are similarly found in humans (Logsdon et al. 2025), potentially serving as substitutes of functional regions. Comparison of the relative positions of CDRs within the entire centromeric arrays among the three strains showed that some chromosomes exhibited similar CDR distributions, whereas others showed more variable patterns (Supplemental Fig. S23). Nevertheless, more complete assemblies will be necessary to assess the extent to which CDR locations are conserved among individuals and among strains. These observations suggest that medaka might possess specific DNA-protein interactions to stabilize kinetochore attachment. Proving this hypothesis will be essential for understanding the function and evolution of vertebrate centromeres and will require further experimental validation.

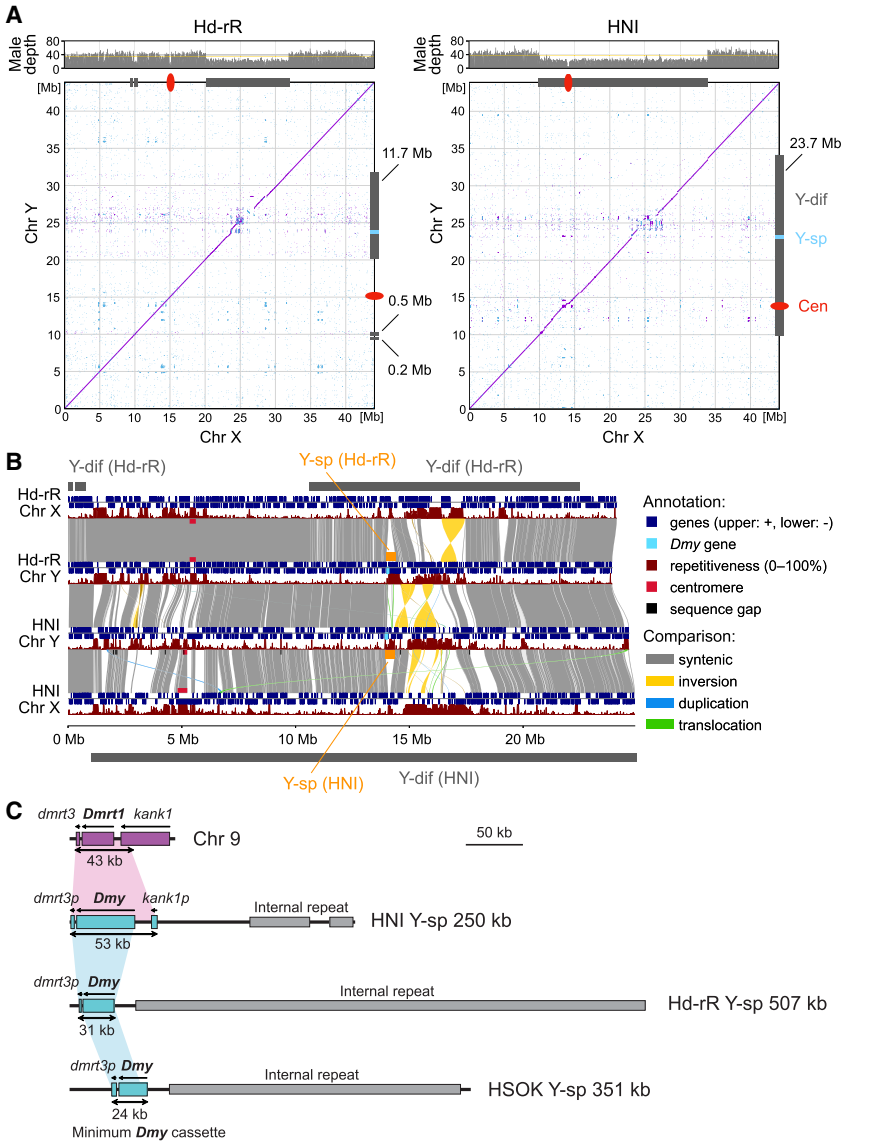

Figure 6. Structural divergence of sex chromosomes and conservation of the *Dmy* locus. (*A*) Dot plots between Chr X and Chr Y for Hd-rR and HNI, with male HiFi read depth. Blue and purple dots indicate forward and reverse complement matches, respectively. Yellow lines indicate global average depth. Reduced male read depth marks Y-dif regions. Centromeres showing reduced HiFi depth are supported by much longer ONT-UL reads (Fig. 1C). (*B*) Structural comparison of X/Y-dif regions (dark gray bars) and Y-sp regions (orange bars) between Hd-rR and HNI. (*C*) Schematic representation of the structural evolution of the Y-sp region and its corresponding region on Chr 9. Blue and purple boxes indicate genes; arrows, transcriptional direction. A 43 kb segment containing *Dmrt1* was duplicated and inserted into the proto-Y Chromosome, giving rise to *Dmy*.

Teratorn has long persisted in medaka through vertical and, possibly, occasional horizontal transmission, which could be facilitated by its dual nature as both a transposon and a herpesvirus. Its persistence stands in sharp contrast to the apparent fate of decay and extinction of the authentic DNA transposons *Tol1* and *Tol2* belonging to the *hAT* superfamily. The long-term survival of *Teratorn* can be explained by (1) the preferential retention of intact copies owing to the requirement of viral particles for propagation (probably because of its large size); (2) the ability to evade host immunity, a common characteristic of herpesviruses; and (3) the benefit to host immunity against an otherwise lethal virus, via preactivation of host innate immunity and/or inhibition of virus entry by receptor blockade. *Teratorn* is known to exert profound effects on the expression of nearby genes, owing to its large size and/or viral activity. A recent report shows that a *Teratorn* insertion in the noncoding region of the *hoxc* cluster alters the *hox* code along the anterior–posterior axis, leading to a phenotypic change in fin shape in medaka (Koita et al. 2025). In future work, we aim to investigate phenotypic differences attributable to *Teratorn*.

We revealed the structural evolution of medaka sex chromosomes over the past 25–40 MY with unprecedented resolution. Despite extensive divergence within the large Y-dif region, likely reflecting suppressed recombination (Matsuda et al. 1999; Kondo et al. 2001), large-scale gene loss or degeneration has not occurred. Indeed, most genes on Chr X and Chr Y are retained as pairs except near the Y-sp region (Supplemental Fig. S24). This contrasts with the three-spined stickleback, whose Chr Y (<26 MY old) is already heteromorphic (Peichel et al. 2020). Collectively, medaka Chr Y represents a transitional stage in sex chromosome evolution. In contrast, the small (24–53 kb) conserved region surrounding *Dmy* provides a focused candidate "minimum *Dmy* cassette" for identifying the key *cis*-regulatory elements that enabled *Dmy* neofunctionalization, which could not be resolved in a previous work (Herpin et al. 2010).

In summary, our complete and near-complete medaka genomes highlight unique aspects of medaka genome biology and provide a precise framework for future vertebrate genome research including comparative, evolutionary, and functional genomics and studies of genotype–phenotype relationships.

Methods

Sample preparation and sequencing

Three inbred medaka strains, Hd-rR (southern Japan), HNI (northern Japan), and HSOK (Korea), were used in this study. All three strains were obtained from the Japan NBRP Medaka resource center (https://shigen.nig.ac.jp/medaka/).

For PacBio HiFi sequencing, we used whole-body samples excluding the head, skin, and gut. Male and female samples were prepared separately for each strain. The numbers of fish used were two males and two females for Hd-rR, two males and three females for HNI, and one male and one female for HSOK.

For ONT-UL read sequencing, we did not use whole-body samples because pilot extractions from whole bodies did not yield

DNA of sufficient length and quantity for ultralong library preparation. We therefore sought tissues suitable for ultra-high-molecular-weight DNA extraction. As a result, a combination of heart, spleen, and testis yielded the best balance between DNA length and total yield, whereas single tissues alone did not provide sufficient amount of DNA. Hd-rR and HNI tissues from both males and females were mixed for the UL DNA library, whereas HSOK tissues were collected only from males. The numbers of fish used were 15 for Hd-rR, 17 for HNI, and 15 for HSOK. Details of sequencing are provided in the Supplemental Methods.

Genome assembly

For each strain, we assembled HiFi and ONT-UL reads of >50 kb jointly using Hifiasm v0.19.6 (Cheng et al. 2024) and Verkko v2.0 (Antipov et al. 2025). Because the strains are inbred, their genomes were treated as effectively haploid except for sex-specific regions. Therefore, we specified the "-l0" option for Hifiasm and the "--haploid" option for Verkko. Because Hifiasm contigs showed higher contiguity, we primarily proceeded with the Hifiasm assembly, curating it using both Verkko contigs and raw reads. Initial chromosomal scaffolds, named v3.0.0, were obtained by scaffolding the Hifiasm contigs using RagTag v2.1.0 (Alonge et al. 2022) based on the v2.2.4 assembly (NCBI Genomes database [https://www.ncbi.nlm.nih.gov/home/genomes/] accession number GCF_002234675.1).

Sex-differentiating sequences were identified based on female and male HiFi read depths. Specifically, sequences whose male read depths were approximately half of the global average depth were classified as sex-differentiating sequences. Among these, sex-differentiating sequences with normal female read depths were categorized into X-dif, whereas those with almost no female read depths were designated as Y-dif. We searched for sex-differentiating sequences from not only primary contigs but also alternative contigs Hifiasm generated. Details of assembly are provided in the Supplemental Methods.

Semiautomated curation of potentially misassembled regions

We defined potentially misassembled regions (PMRs) in an assembly as intervals flagged by sequence gaps, abnormal ONT-UL read depth, or local accumulation of homozygous variants detected from same-strain reads. ONT-UL reads were used preferentially for these metrics because they provide more unambiguous mappings. Additionally, we assessed assembly quality using *k*-mer QVs and the annotations of major important elements such as telomeric and centromeric repeats, rDNA arrays, *Teratorn*, and sex-specific sequences.

We fixed PMRs particularly for completing the Hd-rR genome. First, Verkko contigs were aligned to the scaffolds, and PMRs were replaced with the corresponding Verkko sequence when the number of variants detected becomes smaller. Then we curated most of the remaining PMRs using ONT-UL reads spanning each PMR. For each PMR, we selected the spanning ONT-UL read with the longest flanking sequence as a backbone, aligned other spanning reads to it, calculated a consensus sequence, and substituted the consensus for the PMR. Telomeric sequence extension was performed similarly, except that only one flanking side of each PMR was available. In the end, we curated remaining small homozygous variants detected by either HiFi or ONT-UL reads by taking sequence consensus among mapped reads. We iterated this small variant correction step twice. Details of assembly curation are provided in the Supplemental Methods.

Orienting chromosomes

We categorized each chromosome into one of the metacentric, submetacentric, subtelocentric, acrocentric, and (pseudo-)dicentric. Twelve chromosomes (Chr 1, 2, 7, 8, 9, 11, 12, 13, 18, 19, 22, and 24) were initially oriented from the q-arm to the p-arm and were therefore reverse-complemented. We named these final scaffolds as v3.1 assemblies or simply as v3 assemblies. For compatibility, the same 12 chromosomes in the previous v2.2.4 assemblies were also reverse-complemented, and we named these assemblies as v2.3. In this study, when it is clear whether v2.2.4 or v2.3 is being referred to, it is simply called v2.

Inference of protein-coding regions

Protein-coding genes were predicted from the assemblies by integrating three different programs. LiftOff v1.6.3 (Shumate and Salzberg 2021) was run using the existing protein-coding gene models in GCF_002234675.1_ASM223467v1_genomic.gff associated with the v2 Hd-rR assembly. BRAKER3 v3.0.8 (Gabriel et al. 2024) was run using odb11_actinopterygii_fasta from OrthoDB (Zdobnov et al. 2021) as hints, as well as short-read RNA-seq alignments (NCBI Sequence Read Archive [SRA; https://www.ncbi.nlm.nih.gov/sra] accession number DRP003809) as evidence. Helixer v0.3.3 (Holst et al. 2026) was run using the developer-provided model vertebrate_v0.3_m_0080.h5 with the parameters "--lineage vertebrate --batch-size 16 --subsequence-length 213840."

These three gene sets were merged, and genes encoding peptides shorter than 20 amino acids were removed. Completeness of the annotation was assessed with BUSCO v5.1.2 (Manni et al. 2021) using the actinopterygian BUSCO genes in OrthoDB v10. Details of gene annotation are provided in the Supplemental Methods.

Analysis of centromeric monomers

Centromeric arrays in the v3 assemblies were initially decomposed into monomers with StringDecomposer v1.1.2 (Dvorkina et al. 2020) using the four SF consensus sequences previously defined in the v2 assemblies. For each chromosome, representative monomers were selected by grouping highly similar (>80%) monomers. Representative monomers accounting for >1% of all centromeric monomers on a chromosome were retained.

These chromosome-level representative monomers were compared by hierarchical clustering and principal component analysis. Based on the correspondence with the v2 SFs and the clustering patterns in v3, we redefined medaka centromeric monomers into five SFs. Three v3 SFs corresponded clearly to v2 SF1, SF2, and SF4 and retained these names. The two additional groups represented intermediate monomer families between v2 SF1/SF2 and v2 SF2/SF3 and were designated v3 SF3 and v3 SF5, respectively. The resulting five SF consensus sequences were then used to reannotate centromeric arrays. Monomers with <65% identity to any SF were treated as divergent monomers.

ICUC centromeric monomers were defined as clusters of chromosome-level representative monomers with >97% sequence similarity. Individual monomers were assigned to an ultraconserved monomer class when their similarity to the corresponding representative monomer was >99%.

CpG methylation levels derived from ONT-UL reads were summarized for each centromeric monomer. A raw monomer was classified as hypomethylated or hypermethylated when the average CpG methylation level across its CpG sites was <30% or >70%, respectively. A representative monomer was classified as hypomethylated or hypermethylated when the corresponding class was more frequent among the raw monomers assigned to

it. CDRs were detected from 10 kb sliding-window methylation profiles within each array. Partial order alignment graphs were constructed for SF1 and SF3 from each SF consensus sequence and its representative monomers. Details of centromere analysis are provided in the Supplemental Methods.

Teratorn analysis

Teratorn insertions were screened in the three medaka genomes using 1000 nt terminal sequences from both ends of a *Teratorn* copy as queries. Each *Teratorn* copy was defined as the interval between the left and right terminal inverted repeats (TIRs). Copies overlapping PMRs were removed in HSOK. Subtype 1 and subtype 2 *Teratorn* copies were analyzed separately because their terminal sequences share little sequence identity. Structural categorization was performed based on dot plots between each copy and a representative full-length copy. ORFs were identified using ORFFinder (https://github.com/Chokyotager/ORFFinder).

Sequences of six herpesvirus core genes, *piggyBac*-like transposase, ORF56, *Tol2*, and 17 host genes (Betancur-R. et al. 2013) were obtained from the three medaka genomes and the genome of *Oryzias curvinotus*. Multiple alignments were built up for each gene, and alignments were concatenated for copies containing at least five of the six herpesvirus core genes. Maximum-likelihood trees were constructed from these alignments, with branch support estimated from 100 replicates.

Nonsynonymous-to-synonymous ratios (d_N/d_S) for the six concatenated herpesvirus genes were calculated for each medaka inbred strain and each *Teratorn* subtype. Consensus sequences were generated from multiple alignments, and d_N/d_S values were calculated by comparing each copy with the corresponding consensus sequence using the Nei–Gojobori model. Details of *Teratorn* analysis are provided in the Supplemental Methods.

Comparative analysis of genomes and sex chromosomes

Genome-wide structural rearrangements were computed using SyRI v1.6.3 (Goel et al. 2019) with the parameter "-x asm5" for genome alignments and were visualized using pyGenomeViz (https://github.com/moshi4/pyGenomeViz; v1.5.0).

Y-dif sequences were identified in each strain based on sex-specific HiFi read depths. Y-sp sequences were identified from the insertion signature in the alignment plot between the X- and Y-dif sequences. Internal repeats were identified with RepeatMasker (http://www.repeatmasker.org; v4.1.5). The *Dmy* gene was identified from the gene annotation by mapping the previously known *Dmy* transcript sequence of HNI (NCBI: AB071534.1). *Dmy* amino acid sequences were compared among the v3 assemblies and the previously registered HNI sequence (NCBI: BAB92012.1). Synteny of regions containing X- and Y-dif sequences was calculated with SyRI and visualized with pyGenomeViz. Details of sex chromosome analysis are provided in the Supplemental Methods.

Data access

All raw sequencing reads and genome assemblies generated in this study have been submitted to the DDBJ BioProject database (https://www.ddbj.nig.ac.jp/bioproject/) under accession number PRJDB19938. Supplemental data and code are available on Zenodo (https://doi.org/10.5281/zenodo.15551132) and in the Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank the Information Technology Center, The University of Tokyo for supercomputing facilities. We thank Yuta Suzuki for basecalling reads and Sakuto Yamanaka for critical reading and discussion. This research was supported by the Japan Agency for Medical Research and Development (AMED) CREST grant number JP23gm1110007 to H.T. and S.M., AMED GRIFIN grant number 24tm0424219h0004 to S.M., Japan Society for the Promotion of Science (JSPS) KAKENHI grant number JP22H04925 (PAGS) to S.M., JSPS KAKENHI grant number JP23H02484 and the Kyoto Sangyo University Research grants H2501 to H.T., and JSPS KAKENHI grant number JP24K18091 to Y.S.

Author contributions: S.M. and H.T. conceived and supervised the study. Y.S. performed genome assembly and analyses on centromeres, repeats, and genome comparison with S.M. Y.S. and K.I. performed rDNA analysis. Y.I. performed *Teratorn* analysis. C.O. and H.K. generated HiFi and ONT-UL reads for assembly. K.M. and S.K. performed gene annotation. Y.S., K.M., T.I., Y.T., M.M., K.N., S.K., and H.T. performed analyses on sex chromosomes. R.N., K.N., and M.M. prepared medaka fish samples. M.F., M.K., N.A., E.W.M., and H.A. generated ONT-UL reads for validation. Y.S., S.M., and H.T. wrote the manuscript with input from all authors.

References

Aida T. 1921. On the inheritance of color in a fresh-water fish, *Aplocheilus Latipes* Temmick and Schlegel, with special reference to sex-linked inheritance. *Genetics* **6:** 554–573. doi:10.1093/genetics/6.6.554

Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y. 2001. Alpha-satellite DNA of primates: old and new families. *Chromosoma* **110:** 253–266. doi:10.1007/s004120100146

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* **23:** 258. doi:10.1186/s13059-022-02823-7

Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, et al. 2022. Complete genomic and epigenetic maps of human centromeres. *Science (1979)* **376:** eabl4178. doi:10.1126/science.abl4178

Antipov D, Rautiainen M, Nurk S, Walenz BP, Solar SJ, Phillippy AM, Koren S. 2025. Verkko2 integrates proximity ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. *Genome Res* **35:** 1583–1594. doi:10.1101/gr.280383.124

Asai T, Senou H, Hosoya K. 2011. *Oryzias sakaizumii*, a new ricefish from northern Japan (Teleostei: Adrianichthyidae). *Ichthyol Explor Freshwaters* **22:** 289–299.

Betancur-R R, Broughton RE, Wiley EO, Carpenter K, López JA, Li C, Holcroft NI, Arcila D, Sanciangco M, Cureton JC, et al. 2013. The tree of life and a new classification of bony fishes. *PLoS Curr* **5:** ecurrents.tol.53ba26640df0ccaee75bb165c8c26288. doi:10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288

Chaudhry G, Chen J, Snipes L, Bahl S, Packiaraj J, Lin X, Thakur J. 2025. Genomic and epigenomic maps of mouse centromeres and pericentromeres. *Nat Commun* **16:** 9688. doi:10.1038/s41467-025-64689-0

Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, Hu J, Wang K, Wang C, Xin B, et al. 2023. A complete telomere-to-telomere assembly of the maize genome. *Nat Genet* **55:** 1221–1231. doi:10.1038/s41588-023-01419-6

Cheng H, Asri M, Lucas J, Koren S, Li H. 2024. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat Methods* **21:** 967–970. doi:10.1038/s41592-024-02269-8

Choo KH, Vissel B, Earle E. 1989. Evolution of alpha-satellite DNA on human acrocentric chromosomes. *Genomics* **5:** 332–344. doi:10.1016/0888-7543(89)90066-9

Dvorkina T, Bzikadze AV, Pevzner PA. 2020. The string decomposition problem and its applications to centromere analysis and assembly. *Bioinformatics* **36:** i93–i101. doi:10.1093/bioinformatics/btaa454

Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. 2024. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res* **34:** 769–777. doi:10.1101/gr.278090.123

Gierten J, Welz B, Fitzgerald T, Thumberger T, Agarwal R, Hummel O, Leger A, Weber P, Naruse K, Hassel D, et al. 2025. Natural genetic variation quantitatively regulates heart rate and dimension. *Nat Commun* **16:** 4062. doi:10.1038/s41467-025-59425-7

Goel M, Sun H, Jiao W-B, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20:** 277. doi:10.1186/s13059-019-1911-0

Graves JAM. 2006. Sex chromosome specialization and degeneration in mammals. *Cell* **124:** 901–914. doi:10.1016/j.cell.2006.02.024

Guarracino A, Buonaiuto S, de Lima LG, Potapova T, Rhie A, Koren S, Rubinstein B, Fischer C, Abel HJ, Antonacci-Fulton LL, et al. 2023. Recombination between heterologous human acrocentric chromosomes. *Nature* **617:** 335–343. doi:10.1038/s41586-023-05976-y

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science (1979)* **293:** 1098–1102. doi:10.1126/science.1062939

Herpin A, Braasch I, Kraeussling M, Schmidt C, Thoma EC, Nakamura S, Tanaka M, Schartl M. 2010. Transcriptional rewiring of the sex determining dmrt1 gene duplicate by transposable elements. *PLoS Genet* **6:** e1000844. doi:10.1371/journal.pgen.1000844

Holst F, Bolger AM, Kindel F, Günther C, Maß J, Triesch S, Kiel N, Saadat N, Ebenhöh O, Usadel B, et al. 2026. Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *Nat Methods* **23:** 732–739. doi:10.1038/s41592-025-02939-1

Ichikawa K, Tomioka S, Suzuki Y, Nakamura R, Doi K, Yoshimura J, Kumagai M, Inoue Y, Uchida Y, Irie N, et al. 2017. Centromere evolution and CpG methylation during vertebrate speciation. *Nat Commun* **8:** 1833. doi:10.1038/s41467-017-01982-7

Inoue Y, Saga T, Aikawa T, Kumagai M, Shimada A, Kawaguchi Y, Naruse K, Morishita S, Koga A, Takeda H. 2017. Complete fusion of a transposon and herpesvirus created the *Teratorn* mobile element in medaka fish. *Nat Commun* **8:** 551. doi:10.1038/s41467-017-00527-2

Inoue Y, Kumagai M, Zhang X, Saga T, Wang D, Koga A, Takeda H. 2018. Fusion of *piggyBac*-like transposons and herpesviruses occurs frequently in teleosts. *Zoological Lett* **4:** 6. doi:10.1186/s40851-018-0089-8

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447:** 714–719. doi:10.1038/nature05846

Kipling D, Mitchell AR, Masumoto H, Wilson HE, Nicol L, Cooke HJ. 1995. CENP-B binds a novel centromeric sequence in the Asian mouse *Mus caroli*. *Mol Cell Biol* **15:** 4009–4020. doi:10.1128/MCB.15.8.4009

Kirkpatrick M, Guerrero RF. 2014. Signatures of sex-antagonistic selection on recombining sex chromosomes. *Genetics* **197:** 531–541. doi:10.1534/genetics.113.156026

Koga A, Shimada A, Shima A, Sakaizumi M, Tachida H, Hori H. 2000. Evidence for recent invasion of the medaka fish genome by the *Tol2* transposable element. *Genetics* **155:** 273–281. doi:10.1093/genetics/155.1.273

Koita R, Otake S, Fukaya N, Yamamoto K, Maeno A, Kanno H, Matsuda M, Kawamura A. 2025. The phenotypic variation of *widefins* medaka is due to the insertion of a giant transposon containing a viral genome within the hoxca cluster. *Genetics* **231:** iyaf218. doi:10.1093/genetics/iyaf218

Kondo M, Nagao E, Mitani H, Shima A. 2001. Differences in recombination frequencies during female and male meioses of the sex chromosomes of the medaka, *Oryzias latipes*. *Genet Res* **78:** 23–30. doi:10.1017/S0016672301005109

Kondo M, Hornung U, Nanda I, Imai S, Sasaki T, Shimizu A, Asakawa S, Hori H, Schmid M, Shimizu N, et al. 2006. Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. *Genome Res* **16:** 815–826. doi:10.1101/gr.5016106

Leger A, Brettell I, Monahan J, Barton C, Wolf N, Kusminski N, Herder C, Aadepu N, Becker C, Gierten J, et al. 2022. Genomic variations and epigenomic landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel. *Genome Biol* **23:** 58. doi:10.1186/s13059-022-02602-4

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4:** 203–221. doi:10.1093/oxfordjournals.molbev.a040442

Logsdon GA, Rozanski AN, Ryabov F, Potapova T, Shepelev VA, Catacchio CR, Porubsky D, Mao Y, Yoo D, Rautiainen M, et al. 2024. The variation and evolution of complete human centromeres. *Nature* **629:** 136–145. doi:10.1038/s41586-024-07278-3

Logsdon GA, Ebert P, Audano PA, Loftus M, Porubsky D, Ebler J, Yilmaz F, Hallast P, Prodanov T, Yoo D, et al. 2025. Complex genetic variation in nearly complete human genomes. *Nature* **644:** 430–441. doi:10.1038/s41586-025-09140-6

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* **38:** 4647–4654. doi:10.1093/molbev/msab199

Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T. 1989. A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. *J Cell Biol* **109:** 1963–1973. doi:10.1083/jcb.109.5.1963

Matsuda M, Sotoyama S, Hamaguchi S, Sakaizumi M. 1999. Male-specific restriction of recombination frequency in the sex chromosomes of the medaka, *Oryzias latipes*. *Genet Res* **73:** 225–231. doi:10.1017/S0016672399003754

Matsuda M, Nagahama Y, Shinomiya A, Sato T, Matsuda C, Kobayashi T, Morrey CE, Shibata N, Asakawa S, Shimizu N, et al. 2002. DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* **417:** 559–563. doi:10.1038/nature751

Matsuda M, Shinomiya A, Kinoshita M, Suzuki A, Kobayashi T, Paul-Prasanth B, Lau E, Hamaguchi S, Sakaizumi M, Nagahama Y. 2007. *DMY* gene induces male development in genetically female (XX) medaka fish. *Proc Natl Acad Sci* **104:** 3865–3870. doi:10.1073/pnas.0611707104

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* **14:** R10. doi:10.1186/gb-2013-14-1-r10

Moriyama Y, Kawanishi T, Nakamura R, Tsukahara T, Sumiyama K, Suster ML, Kawakami K, Toyoda A, Fujiyama A, Yasuoka Y, et al. 2012. The medaka *zic1/zic4* mutant provides molecular insights into teleost caudal fin evolution. *Curr Biol* **22:** 601–607. doi:10.1016/j.cub.2012.01.063

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, Mandáková T, Jamge B, Lambing C, Kuo P, et al. 2021. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science (1979)* **374:** eabi7489. doi:10.1126/science.abi7489

Peichel CL, McCann SR, Ross JA, Naftaly AFS, Urton JR, Cech JN, Grimwood J, Schmutz J, Myers RM, Kingsley DM, et al. 2020. Assembly of the threespine stickleback Y chromosome reveals convergent signatures of sex chromosome evolution. *Genome Biol* **21:** 177. doi:10.1186/s13059-020-02097-x

Schueler MG, Sullivan BA. 2006. Structural and functional dynamics of human centromeric chromatin. *Annu Rev Genomics Hum Genet* **7:** 301–313. doi:10.1146/annurev.genom.7.080505.115613

Setiamarga DHE, Miya M, Yamanoue Y, Azuma Y, Inoue JG, Ishiguro NB, Mabuchi K, Nishida M. 2009. Divergence time of the two regional medaka populations in Japan as a new time scale for comparative genomics of vertebrates. *Biol Lett* **5:** 812–816. doi:10.1098/rsbl.2009.0419

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* **38:** 1044–1053. doi:10.1038/s41587-020-0503-6

Shepelev VA, Alexandrov AA, Yurov YB, Alexandrov IA. 2009. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS Genet* **5:** e1000641. doi:10.1371/journal.pgen.1000641

Shinomiya A, Adachi D, Shimmura T, Tanikawa M, Hiramatsu N, Ijiri S, Naruse K, Sakaizumi M, Yoshimura T. 2023. Variation in responses to photoperiods and temperatures in Japanese medaka from different latitudes. *Zoological Lett* **9:** 16. doi:10.1186/s40851-023-00215-8

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37:** 1639–1643. doi:10.1093/bioinformatics/btaa1016

Smith GP. 1976. Evolution of repeated DNA sequences by unequal crossover. *Science (1979)* **191:** 528–535. doi:10.1126/science.1251186

Spivakov M, Auer TO, Peravali R, Dunham I, Dolle D, Fujiyama A, Toyoda A, Aizu T, Minakuchi Y, Loosli F, et al. 2014. Genomic and phenotypic characterization of a wild medaka population: towards the establishment of an isogenic population genetic resource in fish. *G3 (Bethesda)* **4:** 433–445. doi:10.1534/g3.113.008722

Takeda H, Shimada A. 2010. The art of medaka genetics and genomics: What makes them so unique? *Annu Rev Genet* **44:** 217–241. doi:10.1146/annurev-genet-051710-151001

Takehana Y, Nagai N, Matsuda M, Tsuchiya K, Sakaizumi M. 2003. Geographic variation and diversity of the cytochrome *b* gene in Japanese wild populations of medaka, *Oryzias latipes*. *Zoolog Sci* **20:** 1279–1291. doi:10.2108/zsj.20.1279

Takehana Y, Jeon S-R, Sakaizumi M. 2004. Genetic structure of Korean wild populations of the medaka *Oryzias latipes* inferred from allozymic variation. *Zoolog Sci* **21:** 977–988. doi:10.2108/zsj.21.977

Takehana Y, Sakai M, Narita T, Sato T, Naruse K, Sakaizumi M. 2016. Origin of boundary populations in medaka (*Oryzias latipes* species complex). *Zoolog Sci* **33:** 125–131. doi:10.2108/zs150144

Welz B, Pierotti S, Fitzgerald T, Thumberger T, Suzuki R, Watson P, Fuss J, Cordeiro da Trindade T, Defranoux F, Ferreira M, et al. 2026. Discovery and characterization of gene-by-environment and epistatic genetic effects in a vertebrate model. *Cell Genomics* **6:** 101164. doi:10.1016/j.xgen.2026.101164

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37:** 1155–1162. doi:10.1038/s41587-019-0217-9

Yamahira K, Ansai S, Kakioka R, Yaguchi H, Kon T, Montenegro J, Kobayashi H, Fujimoto S, Kimura R, Takehana Y, et al. 2021. Mesozoic origin and 'out-of-India' radiation of ricefishes (Adrianichthyidae). *Biol Lett* **17:** 20210212. doi:10.1098/rsbl.2021.0212

Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva EV. 2021. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res* **49:** D389–D393. doi:10.1093/nar/gkaa1009

Received December 20, 2025; accepted in revised form June 22, 2026.