Research

High resolution genomes of multiple *Xiphophorus* species provide new insights into microevolution, hybrid incompatibility, and epistasis

Yuan Lu,[1,9] Edward Rice,[2,9] Kang Du,[1] Susanne Kneitz,[3] Magali Naville,[4] Corentin Dechaud,[4] Jean-Nicolas Volff,[4] Mikki Boswell,[1] William Boswell,[1] LaDeana Hillier,[5] Chad Tomlinson,[6] Kremitzki Milin,[6] Ronald B. Walter,[7] Manfred Schartl,[1,8] and Wesley C. Warren[2]

[1]*The Xiphophorus Genetic Stock Center, Texas State University, San Marcos, Texas 78666, USA;* [2]*Department of Animal Sciences, Department of Surgery, Institute for Data Science and Informatics, University of Missouri, Bond Life Sciences Center, Columbia, Missouri 65201, USA;* [3]*Biochemistry and Cell Biology, Biozentrum, University of Würzburg, 97074 Würzburg, Germany;* [4]*Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, CNRS UMR 5242, Université Claude Bernard Lyon 1, F-69364 Lyon, France;* [5]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA;* [6]*McDonnell Genome Institute, Washington University, St. Louis, Missouri 63108, USA;* [7]*Department of Life Sciences, Texas A&M University, Corpus Christi, Texas 78412, USA;* [8]*Developmental Biochemistry, Biozentrum, University of Würzburg, 97074 Würzburg, Germany*

Because of diverged adaptative phenotypes, fish species of the genus *Xiphophorus* have contributed to a wide range of research for a century. Existing *Xiphophorus* genome assemblies are not at the chromosomal level and are prone to sequence gaps, thus hindering advancement of the intra- and inter-species differences for evolutionary, comparative, and translational biomedical studies. Herein, we assembled high-quality chromosome-level genome assemblies for three distantly related *Xiphophorus* species, namely, *X. maculatus*, *X. couchianus*, and *X. hellerii*. Our overall goal is to precisely assess microevolutionary processes in the clade to ascertain molecular events that led to the divergence of the *Xiphophorus* species and to progress understanding of genetic incompatibility to disease. In particular, we measured intra- and inter-species divergence and assessed gene expression dysregulation in reciprocal interspecies hybrids among the three species. We found expanded gene families and positively selected genes associated with live bearing, a special mode of reproduction. We also found positively selected gene families are significantly enriched in nonpolymorphic transposable elements, suggesting the dispersal of these nonpolymorphic transposable elements has accompanied the evolution of the genes, possibly by incorporating new regulatory elements in support of the Britten–Davidson hypothesis. We characterized inter-specific polymorphisms, structural variants, and polymorphic transposable element insertions and assessed their association to interspecies hybridization-induced gene expression dysregulation related to specific disease states in humans.

[Supplemental material is available for this article.]

Xiphophorus is widely used for studying many questions in ecology, physiology, fish biology, and evolution, as well as comparative and translational medicine. *Xiphophorus* is a teleost fish genus consisting of 26 species. They are found in a wide range of different geographical regions within Central and South America and show a plethora of distinctive phenotypes, such as pigmentation pattern, presence of nuchal hump, early/late maturation, and body size (Lampert et al. 2010; Shen et al. 2016; Lu et al. 2017b, 2018; Liu et al. 2020). These phenotypes often result from adaptations to mutations or species-specific niches. When such adaptive phenotypes are similar to human health conditions or diseases, one can leverage the adaptive process within the natural population to gain understanding about disease etiology or inborn strategies in controlling the pathological process. One of the best-known examples is an oncogene-driven pigmentation pattern found in *Xiphophorus maculatus* and *Xiphophorus birchmanni* (Wittbrodt et al. 1989; Lu et al. 2017a, 2020b; Powell et al. 2020). In both cases, a pigmentation pattern is observed, which is driven by a mutant ortholog of the human epidermal growth factor receptor (*EGFR*) named *Xiphophorus* melanoma receptor kinase (*xmrk*) (Wittbrodt et al. 1989). The *xmrk* gene harbors two mutations that lead to constitutive proliferation-promoting function of the receptor. In fact, ectopic expression of *xmrk* in medaka, zebrafish, and murine cells led to tumorigenesis, reprogramming, enhanced proliferation, and up-regulation of several proliferative signaling pathways (Wellbrock et al. 2002; Schartl et al. 2010; Mishra et al. 2014; Zheng et al. 2014; Yang et al. 2017; Klotz et al. 2018). However, in both *X. maculatus* and *X. birchmanni*, the oncogenic action of *xmrk* becomes apparent only as nevi-like pigmentation, suggesting there are molecular adaptations

[9]**These authors contributed equally to this work.**
Corresponding authors: y_l54@txstate.edu, phch1@biozentrum.uni-wuerzburg.de, warrenwc@missouri.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277434.122.

counteracting the detrimental effect of *xmrk* (Powell et al. 2020; Lu et al. 2020b).

A unique and powerful feature of the *Xiphophorus* model is the ability to produce viable interspecies hybrids, enabling the possibility of combining alleles that diverged within different species lineages in an inter-species hybrid. Therefore, creating F_1 interspecies hybrids provides a perfect tool for identifying phenotypes that are under control of incompatible genetic interactions (i.e., negative epistasis). Further hybridization experiments using F_1 interspecies hybrids generate viable backcross, intercross, or outcross interspecies hybrids and allow Mendelian segregation and recombination of parental chromosomes to take place within the hybrid cohort. By studying the cosegregation pattern of phenotypes and parental allele inheritance, it is possible to identify, via genetic mapping, loci linked to traits (Wittbrodt et al. 1989; Lu et al. 2017a, 2020b; Powell et al. 2020, 2021; Schartl et al. 2021). Using interspecies hybridization coupled with gene mapping strategies, the *X. maculatus* and *X. birchmanni* alleles that reduce the deleterious effect of *xmrk* (i.e., *X. maculatus rab3d* and *X. birchmanni adgre5*) were identified (Powell et al. 2020; Lu et al. 2020b). These findings have translational importance to characterize molecular mechanisms of the oncogene adaptation process for developing novel strategies in regulating human EGFR. The knowledge advancement using the unique *Xiphophorus* system exemplifies how evolutionary adaptations can provide insight into human medicine (i.e., evolutionary mutant models) (see Albertson et al. 2009; Schartl 2014; Beck et al. 2022).

The phenotypes resulting from incompatible genetic interactions are not limited to the well-documented gross morphological traits but extrapolate to molecular phenotypes. Previous molecular genetic analyses have shown differences in allele-specific regulatory mechanisms between closely related species (e.g., *X. maculatus* and *Xiphophorus couchianus*) and their interspecies hybrids (Lu et al. 2015). Differences between species appeared mainly owing to *cis*-regulatory elements, whereas changes in *trans*-regulatory elements and/or the interaction of both *cis*- and *trans*-effects were also shown to play important roles (Lu et al. 2018). However, genome-wide structural and functional understanding of incompatible loci within the *Xiphophorus* hybrids is understudied, mainly owing to the unavailability of high-quality reference *Xiphophorus* genomes. Chromosome-level genome models that are independently assembled and annotated are indispensable resources to deconvolute the adaptative processes and to determine loci controlling certain phenotypes. The obtained information can then be used to investigate nearby and distant regulatory sequences of such loci and to forward orthologous sequences for translational studies.

In this study, we aimed to establish high-continuity chromosomal assemblies for three representative *Xiphophorus* species (*X. maculatus*, *X. couchianus*, and *X. hellerii*), perform pairwise comparisons between the genomes of these species, and reveal genetic architecture differences that may be associated with overall trait differentiation. In addition, to investigate incompatible loci and associated transcriptional phenotypes within the interspecies hybrid, we surveyed the transcriptomes of reciprocal interspecies hybrids between the three species. Studying closely related genomes can provide essential information on the understudied processes and forces of microevolution and allow for investigating how mutation, migration, genetic drift, and natural selection acted on the evolution of closely related *Xiphophorus* species that comprise a wide range of ecological, physiological, and morphological adaptations (Li et al. 2018).

Results

Genome assembly and annotation

Reference genomes were generated for three *Xiphophorus* species: *X. maculatus*, *X. couchianus*, and *X. hellerii*, each from a single laboratory-reared male or female that descended from line breeding (see Methods) (Fig. 1). Each genome was sequenced and assembled using SMRT sequencing and the HGAP assembler to ungapped sizes ranging from 687 to 730 Mb, similar to the expected genome size of species from this genus (Table 1). Primary scaffolding of the assembled contigs was accomplished with the aid of DNA restriction enzyme site-based imaging (BioNano) for *X. maculatus* and *X. couchianus* and proximity ligation maps (Hi-C) for *X. helleri*. The final genome assemblies display similar overall contiguity metrics to other long-read assembled teleost genomes with a scaffold number and N50 length of 68–102 and 30–32 Mb, respectively (Table 1). All 24 chromosomes were assembled with a 0.4%–0.8% range of unassigned sequences. We found few ordering discrepancies and show significant chromosome-wide synteny between the *Xiphophorus* species (Fig. 2; Supplemental Fig. S1).

Protein-coding genes predicted using the NCBI (NCBI Resource Coordinators 2016) automated pipeline show similar numbers for each species (Table 1). Improvements to earlier *Xiphophorus* gene sets were seen with a range of 1562–1665 new protein-coding genes. The completeness of the gene annotation as assessed by BUSCO ranged from 93.6%–94.6% when aligned to the 15,231 single-copy orthologs in the Cyprinodontiformes set, with between 0.4–1.0% of these duplicated (Table 1). In total, our measures of gene representation in these *Xiphophorus* genomes show high-quality resources for the study of *Xiphophorus* biology and translational studies.

Genome evolution

A phylogenomic reconstruction of the evolutionary relationships of the three *Xiphophorus* species to other teleosts using a gene set of 1425 high-quality orthologs confirmed previous groupings and revealed a split of the *Xiphophorus* branch from egg-

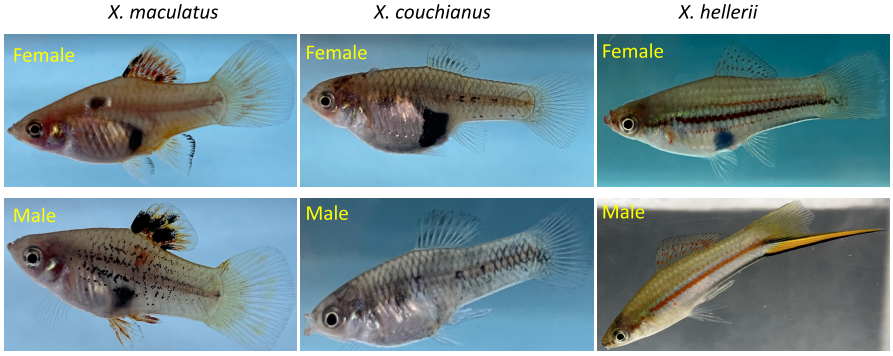

Figure 1. Images of *X. maculatus*, *X. couchianus*, and *X. hellerii*.

Table 1. *Xiphophorus* genome assembly statistics

	X. maculatus	*X. couchianus*	*X. hellerii*	Outgroup (*Oryzias latipes*)
Assembly metrics				
Assembled version	X maculatus-5.0-male	X couchianus 1.0	Xiphophorus helleri-4.1	ASM223467v1
N50 contig (Mb)	9.1	15	7	2.5
N50 scaffold (Mb)	31	30	32	31
Total assembly size (Mb)	701	687	730	733
% Repeat Masked	27.6	27.6	28.2	34.1
Gene annotation metrics				
Protein-coding genes	23,238	22,784	23,921	
Total ncRNA	4696	7497	7597	
mRNAs	43,551	47,063	46,235	
miscRNA	700	1038	1175	
lncRNA	1769	4714	4628	
snoRNA	150	151	153	
snRNA	73	72	76	
Guide RNA	7	7	7	
BUSCO summary of gene representation				
Complete	94.6	94.3	93.6	
Complete and single copy	94.1	93.9	92.6	
Complete and duplicated	0.5	0.4	1	
Fragmented	0.9	0.9	1	
Missing	4.5	4.8	5.4	
Predicted *Xiphophorus* genome heterozygosity and haploid length				
% Heterozygosity	0.04	0.01	0.13	
Haploid length Mb	682	672	690	
Assembled length Mb	701	687	730	

laying cyprinodonts around 52 MYA. The radiation of the genus *Xiphophorus* is estimated to have occurred around 5 MYA (Fig. 3).

Gene family analysis revealed 19 gene families expanded or contracted exclusively in the *Xiphophorus* lineage. Some show expansion or contraction only in one or two species (Table 2).

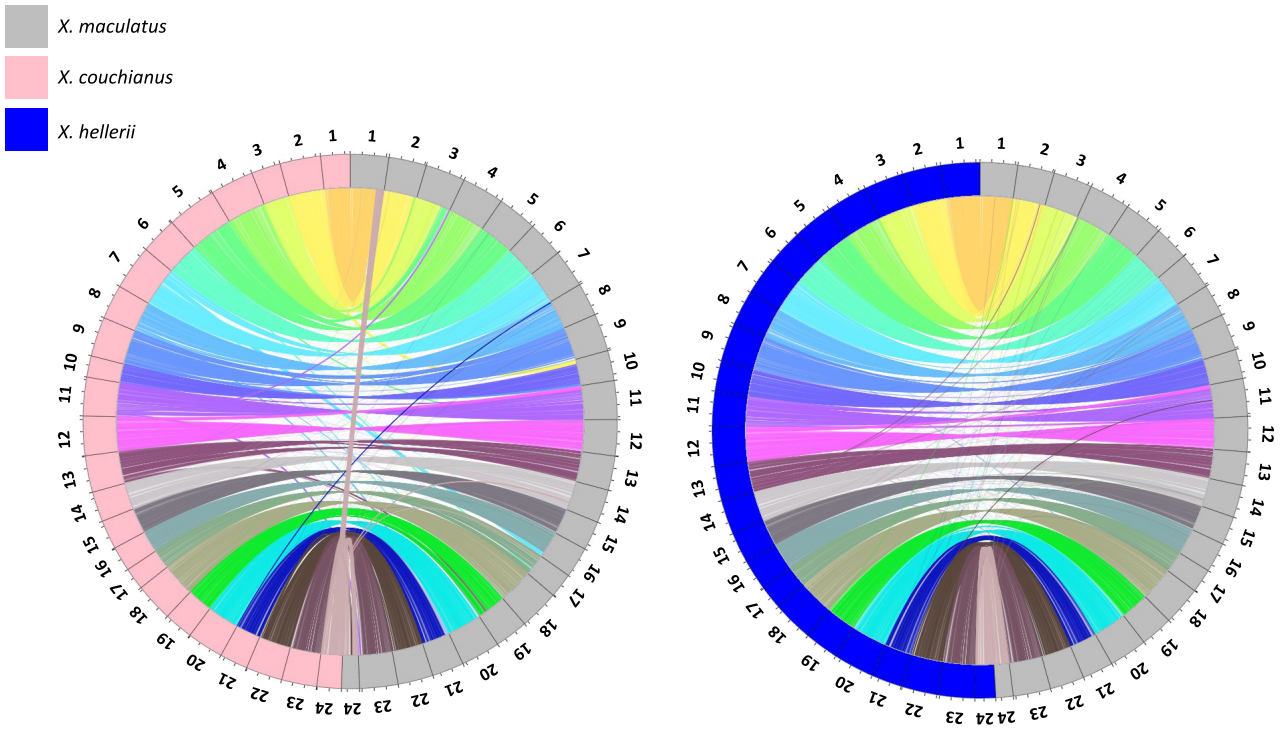

Figure 2. Whole-genome synteny of *Xiphophorus* species. *X. maculatus* (XM; gray), *X. couchianus* (XC; pink), and *X. hellerii* (XH; blue) genome syntenic blocks are plotted as Circos plots. *Outer* rings represent chromosome lengths, with each tick representing 10 Mbp. *Inner* ribbons show gene orthology between XC and XM or between XH and XM. Orthologous genes between species are linked by lines in the Circos plot.

Evolution of novel traits and specific features of *Xiphophorus*, in particular viviparity, may have required changes in protein structures and should be visible as signatures of positive selection. To uncover such genomic traces of natural selection, 8492 one-to-one orthologs from the three *Xiphophorus* species and seven other teleost species were collected and tested for positive selection in the lineage leading to *Xiphophorus* (Supplemental Tables S1, S2). We uncovered 55 genes with site class 3 (marked branch and conserved in rest), and 207 genes with site class 4 (marked branch and relaxed in rest) that are under positive selection (Supplemental Table S3). The positively selected genes were enriched for proteins with functions in transmembrane transport and the extracellular compartment, as well as proteins with epidermal growth factor-like domains (Supplemental Fig. S2).

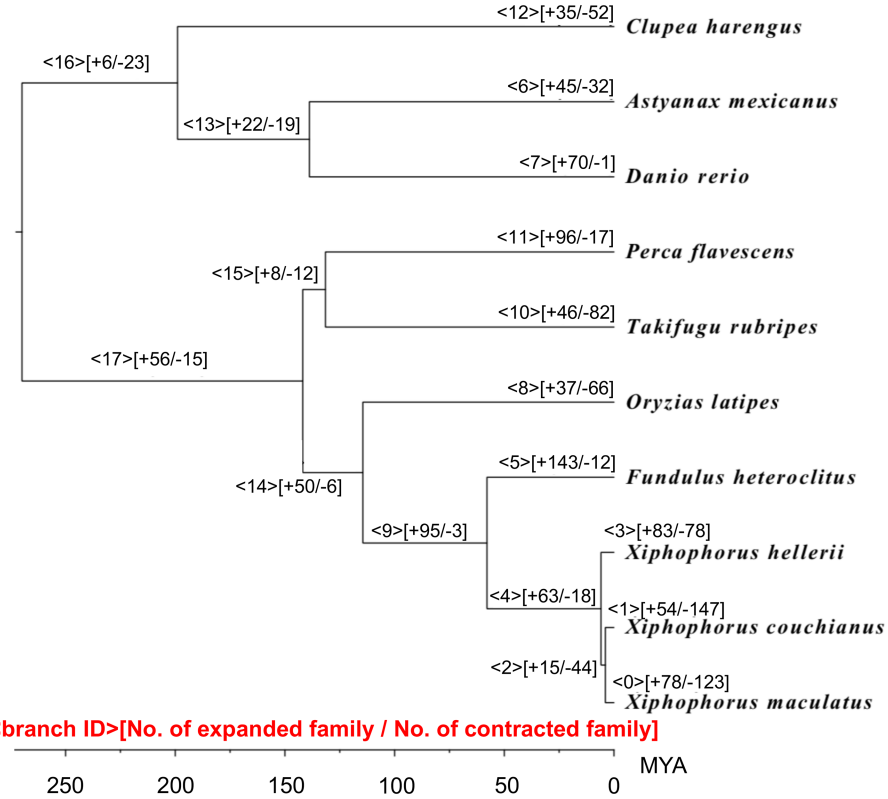

Figure 3. Phylogenetic tree displaying species divergence time and number of gene families expanded or contracted on each branch. Branch IDs are shown in angle brackets; numbers of expanded gene families are shown in square brackets next to plus mark; numbers of contracted gene families are shown in square brackets next to minus mark, for example, <branch ID>[no. of expanded family/no. of contracted family].

Intra-species comparative genomics

To assess intra-species heterozygosity and genomic variants, we sequenced individuals from different genetic backgrounds. These include two inbred *X. maculatus* lines, Jp 163A and Jp 163B; one wild *X. maculatus*; one inbred *X. couchianus*; one random-breeding *X. couchianus* subline; and closed-colony bred *X. hellerii*. Per fish type, we investigated polymorphisms (SNP and short indels) and structural variations (SVs). Nucleotide diversity (π) estimates showed *X. hellerii* had the highest π (0.0078), followed in order by the wild-caught *X. maculatus*, the *X. maculatus* Jp163B inbred strain (6.11×10^{-6}), the *X. maculatus* Jp 163A strain (5.97×10^{-6}), and *X. couchianus* (1.98×10^{-6}). Inbred population generally had π values 300- to 1000-fold lower compared with outbred populations (Supplemental Fig. S3; Supplemental Table S4). Polymorphisms in all assessed populations are more common in gene regulatory regions, rather than coding regions. There are similar numbers of codon disruptive high-impact variants and synonymous or single-amino-acid-changing low-impact variants within the inbred populations. In comparison, high-impact variants are far fewer than low-impact variants in the wild population and closed colony-bred *X. hellerii* (Supplemental Table S5).

Because of the available coverage and the lower sensitivity and accuracy of calling SVs using short reads, our analysis focused on deletions in the range of 500 bp to 100 kbp found by both of two SV callers: LUMPY (Layer et al. 2014) and Manta (Supplemental Figs. S4, S5; Chen et al. 2016). In *X. hellerii*, we found 9638 high-confidence deletions present in at least one haplotype of one individual (Supplemental Table S6). This is equivalent to a total of 5 Mbp long, or 0.7% of the genome. The deletions were further classified as affecting the coding sequence (623 deletions), intronic sequence (4990), or flanking sequence (1585). The numbers of deletions of each category and ploidy were similar across all samples, with more homozygous deletions than heterozygous in each category. This is consistent with the high homogeneity of the sampled laboratory population. In *X. maculatus*, 6003 high-confidence deletions were present in at least one haplotype of one individual (Supplemental Tables S6, S7). There were significantly more deletions, both homozygous and heterozygous, in the wild fish than in the laboratory strains (one-tailed *t*-tests; $P = 1.92 \times 10^{-16}$ for homozygous deletions and $P = 8.78 \times 10^{-14}$ for heterozygous deletions).

Inter-species comparative genomics

To evaluate structural differences between the three *Xiphophorus* species, we used cross-species alignments to find fixed deletions in the genomes of *X. couchianus* and *X. maculatus* compared with *X. hellerii*. A fixed deletion herein is defined as a deletion called by both SV callers and present in all the individuals of a species. We found 79 fixed deletions of coding sequence in *X. maculatus* overlapping with 143 genes (Supplemental Table S8). In *X. couchianus*, we found 143 deletions of coding sequence interacting with 235 genes (Supplemental Table S8); 113 of these genes contain deletions of coding sequence in both *X. maculatus* and *X. couchianus* compared with *X. hellerii*.

In addition, we identified genes that are influenced by high-impact (i.e., disruptive in-frame deletion/insertion, frameshift, start codon lost, stop codon gain/loss, splice acceptor/donor variant) homozygous polymorphisms on gene sequences (i.e., between *X. hellerii* and *X. maculatus*: 3147; between *X. maculatus* and *X. couchianus*: 3400; between *X. hellerii* and *X. couchianus:* 5186) (Supplemental Table S9). Comparing the high-impact

Table 2. Gene family expansion and contraction

ID	*A. mexicanus*	*F. heteroclitus*	*O. latipes*	*P. flavescens*	*T. rubripes*	*X. couchianus*	*X. hellerii*	*X. maculatus*	*D. rerio*	*C. harengus*	Gene symbol	Description
3523	1	1	1	1	1	1	3	11	0	1	*LOC103042955*	Melanocortin receptor 4–like
6454	1	1	1	1	1	2	5	12	1	1	*LOC116710423*	DUF2452 domain-containing protein/ transposon?
6782	1	1	1	1	1	15	8	4	1	1	*LOC114558340*	Zinc finger BED domain-containing protein 4–like
8346	1	1	1	1	1	9	4	3	1	1	*LOC114153686*	Methyltransferase N6AMT1–like
1 24 19	1	1	1	1	1	2	3	8	1	1	*LOC106700416*	Sodium-coupled neutral amino acid transporter 2–like isoform X1
15765	1	2	2	1	3	2	6	4	2	1	*paox*	Peroxisomal N(1)-acetyl-spermine/ spermidine oxidase
18112	1	1	1	1	1	9	16	11	1	1	*oard1*	ADP-ribose glycohydrolase OARD1 isoform X1
20812	0	1	1	1	1	7	1	1	0	1	*LOC114154550*	Riboflavin kinase–like isoform X2
23086	1	1	1	1	1	1	1	7	1	0	*drosha*	Ribonuclease 3 isoform X1
25351	1	1	1	1	1	1	12	3	1	0	*gpr180*	Integral membrane protein GPR180
1743	0	0	0	0	0	0	6	0	1	0	*LOC101886504*	Retrotransposon
5103	1	0	0	0	0	14	7	11	0	0	*LOC103045528*	Kinesin-like protein KIN-14N
5771	1	0	0	0	0	4	0	0	0	0	*LOC114143748*	Zinc finger MYM-type protein 1–like isoform X3
5962	1	0	0	1	0	5	0	4	0	0	*LOC114134843*	Gametogenetin
6395	0	1	0	0	0	12	13	20	1	0	*LOC114145046*	Uncharacterized protein LOC114145046
6720	1	0	0	0	0	1	12	2	0	0	*LOC116707762*	LOW QUALITY PROTEIN: uncharacterized protein LOC116707762
8330	1	0	0	0	0	6	9	5	1	0	*si:dkey-65l23.2*	DUF4806 domain-containing protein/ transposon?
18377	1	0	0	0	0	5	0	1	0	1	*LOC103042417*	TLD domain-containing protein 2
23547	0	0	0	0	0	5	1	0	1	0	*LOC110439524*	AIR1; arginine methyltransferase-interacting protein

homozygous polymorphisms bearing genes to expanded gene families, we found 50% of genes within the expanded families show these variants (Supplemental Table S3). In addition, Gene Ontology (GO) enrichment analyses showed these genes are overrepresented in a cadre of functional categories (Supplemental Table S10).

Transposable element analyses

The transposable element (TE)–thrust hypothesis, based on the work by McClintock and others, hypothesized that TEs are powerful facilitators of evolution (McClintock 1956; Brosius 1991; Fedoroff 1999; Kidwell and Lisch 2001; Bowen and Jordan 2002; Deininger et al. 2003; Kazazian 2004; Biémont and Vieira 2006; Volff 2006; Feschotte and Pritham 2007; Muotri et al. 2007; Böhne et al. 2008; Oliver and Greene 2011). In addition, the Britten–Davidson hypothesis states that TEs can introduce transcription factor binding sites and accelerate evolution (Britten and Davidson 1969). To test these hypotheses, we investigated TEs in the three *Xiphophorus* genomes. TE annotations of *X. maculatus*, *X. couchianus*, and *X. hellerii* showed similar TE coverage in the three species (i.e., *X. maculatus*: 23.2%; *X. couchianus*: 22.5%; *X. hellerii*: 27.4%) (Fig. 4A). The TEs are equally distributed between class I (retrotransposons) and class II elements (DNA transposons) (Fig. 4A). To get an overview of the ancient and present waves of TE expansion, we compared TE insertions between each other in each identified TE family. The obtained "landscapes" (Fig. 4B) reveal three main bursts of transposition in the history of the genomes, with a peak of highly similar insertions (on the right of the graph) indicating recent transposition events (Fig. 4B).

To evaluate the impact of TEs on genome divergence, we searched for polymorphic insertions between pairs of species to identify elements (i.e., TEs that were inserted specifically in one of the two species). Using a TE insertion length >300 nt as a threshold, we identified 6773 between *X. maculatus* and *X. couchianus* and 11,242 between *X. hellerii* and *X. couchianus* polymorphic TEs (Supplemental Table S11). These polymorphic TEs represented between 4.7% and 10% of all insertions >300 nt in a given genome. Between *X. maculatus* and *X. couchianus*, these TEs account for 9.2 Mbp of genomic DNA, or 1.3% of the genome; between *X. hellerii* and *X. couchianus*, they account for 17.6 Mbp of genomic DNA, or 2.5% of the genome. A mapping of *X. maculatus* compared with *X. hellerii* polymorphic elements revealed a global dispersion of TEs along chromosomes, some with accumulations at chromosome extremities, notably for *X. maculatus* (Supplemental Fig. S6).

To assess impacts of the polymorphic TEs on genes, we identified those that localize close to or overlap (i.e., polymorphic TEs that locate <1000 bp to the transcription start site [TSS]) with TSSs in *X. maculatus* and *X. couchianus*. There were 407 in *X. maculatus* and 760 in *X. couchianus* (Supplemental Tables S12, S13; an example, *brca2*, is presented in Supplemental Fig. S7). Among the genes whose promoter regions were identified to contain polymorphic TEs, a large proportion encoded lncRNAs (i.e., 12% in *X. maculatus* and 24% in *X. couchianus*). TSSs for 87 genes in *X. maculatus* and 116 genes in *X. couchianus* were overlapped with polymorphic TEs, suggesting that these TEs can create species-specific alternative transcripts by providing new transcription starts. LncRNAs were even more represented in this subset of data, with 34% and 66% of corresponding genes being lncRNAs in *X. maculatus* and *X. couchianus*, respectively (Supplemental Table S12).

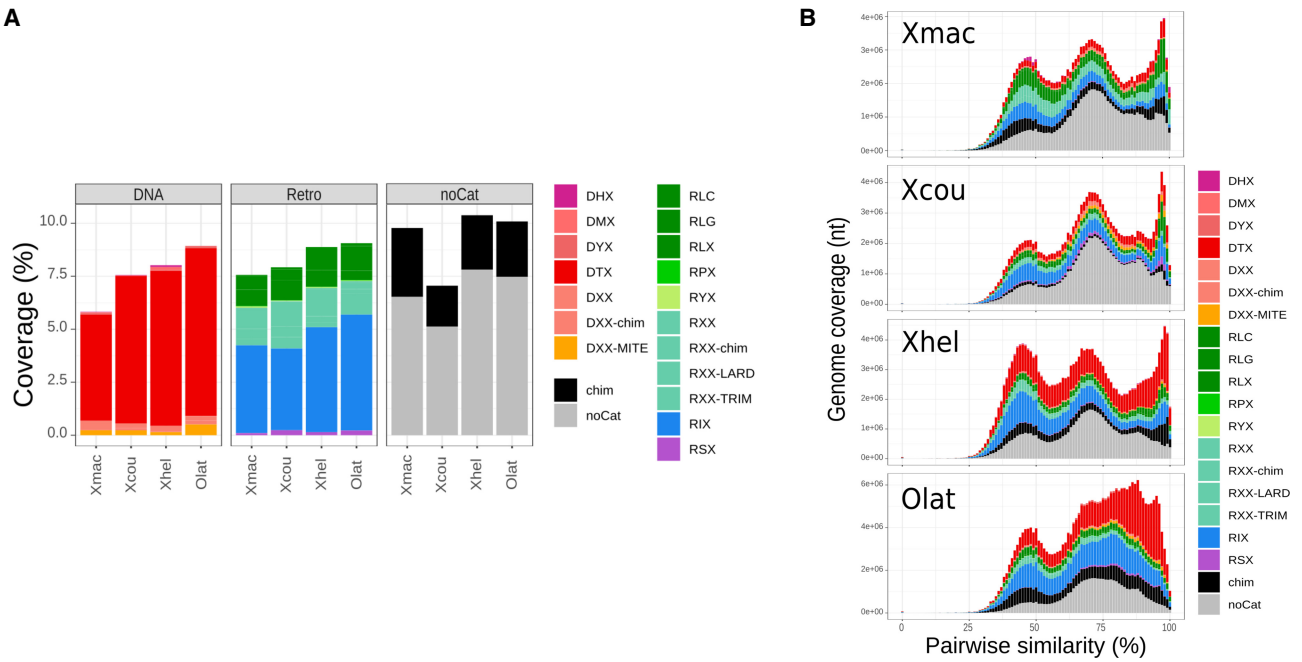

Figure 4. TE analyses in *Xiphophorus*. (Xmac) *X. maculatus*; (Xcou) *X. couchianus*; (Xhel) *X. hellerii*; (Olat) *Oryzias latipes*. (*A*) TE coverages are similar between the surveyed species, and TEs are equally distributed between class II (DNA transposons) and class I (retrotransposons). (DHX) Helitron; (DMX) Maverick; (DYX) Crypton; (DTX) TIR transposon; (DXX) unclassified DNA element; (chim) chimeric element; (RLC) Copia; (RLG) Gypsy; (RLX) LTR retrotransposon; (RPX) Penelope; (RYX) DIRS; (RXX) unclassified retrotransposon; (RIX) LINE; (RSX) SINE. (*B*) TE landscapes in *Xiphophorus* species and medaka species. TE insertions were compared between each other for each identified TE family; the distribution of similarity scores was computed and converted into genomic coverage according to the coverage of the TE family. The *right* extremity of the graph corresponds to recent (highly similar) TE insertions.

To test the hypothesis that the insertion of TEs is a prerequisite for transcription factor binding site divergence and to create evolutionary innovations, we assessed if TEs could participate in the rapid evolution of genes. We searched for polymorphic TEs in the vicinity of genes found positively selected in one of the three species or in all three species together. A total of eight polymorphic TEs were found <1 kb of genes under positive selection in the *X. maculatus* (associated to *slc6a6*, *slc13a1*, *nos1ap*, *rcn3*, and *suclg2*), *X. couchianus* (associated to *LOC114149751*), and *X. hellerii* (associated to *wrap73* and *nat10*) genomes, respectively (Supplemental Table S14). Two hundred nine polymorphic TEs were found <1 kb of genes positively selected in all three genomes. Among these, 50 were under purifying selection in outgroup species, and 159 were under relaxed constraints in outgroup species (Supplemental Table S14).

We subsequently tested the hypothesis that genes under positive selection are associated with polymorphic TEs compared with nonpositively selected genes. In *X. maculatus*, we found a mean of 0.0089 and 0.0132 polymorphic TEs per kilobase of positively selected and nonpositively selected gene, respectively (*t*-test *P*-value = 0.0012). Therefore, there is no enrichment of polymorphic elements near positively selected genes. We also tested if genes found positively selected in all three species are significantly associated with specific TE families of only nonpolymorphic elements (i.e., elements that were inserted before the divergence of the species). Applying 1000 resamples of nonpolymorphic TEs, 20 families were significantly enriched in nonpolymorphic elements in the vicinity of class 3 genes, and 19 and one families were significantly enriched and depleted, respectively, in nonpolymorphic elements in the vicinity of class 4 genes (distance < 1000 bp, *P*-value < 0.01), with only one family in common for the two sets of genes (Supplemental Table S15). Therefore, positively selected gene families are significantly enriched in nonpolymorphic TEs. Among the families enriched in the vicinity of class 3 genes, six were DNA transposons, whereas no particular class appeared near class 4 genes.

In addition, polymorphic TEs were compared with genes belonging to expanded gene families (increased copy number). There are 26 genes involved in both (107 genes in expanded gene families, 6332 genes close to polymorphic TEs), but a hypergeometric test showed the overlap was not significant (*P*-value = 0.299), suggesting TE and expanded gene families are under different pattern of selection.

Inter-species transcriptional incompatibility

Whole-fish transcriptome profiling of reciprocal interspecies hybrid between *X. maculatus*, *X. couchianus*, or *X. hellerii* uncovered 2570 (*X. hellerii*–*X. maculatus* hybrid), 436 (*X. couchianus*–*X. hellerii* hybrid), and 245 (*X. couchianus*–*X. maculatus* hybrid) genes with a dysregulated expression pattern compared with both parental species (i.e., transcriptional incompatibility) (Fig. 5). We compared the dysregulated genes to inter-specific genetic variants of coding sequences, upstream regulatory sequences, SVs, and polymorphic TEs.

Inter-specific polymorphisms within the upstream regulatory sequences, SVs, and TEs are associated to gene expression dysregulation within the hybrids. On average, they contributed to 7.4%, 9.5%, and 7.6% of gene dysregulation, respectively (Fig. 5). The disease types associated with human orthologs of the dysregulated *Xiphophorus* genes were identified. On average, 6% of hybridization-induced dysregulated genes are related to human diseases. The diseases associated with these dysregulated genes are organ diverse (Fig. 6).

Discussion

In this study, we provided the first nearly complete genome representations of multiple species of the *Xiphophorus* genus. Compared with earlier assemblies based on short sequencing reads, all genome quality parameters are much improved (Table 1; Schartl et al. 2013; Shen et al. 2016). *Xiphophorus* genome models have

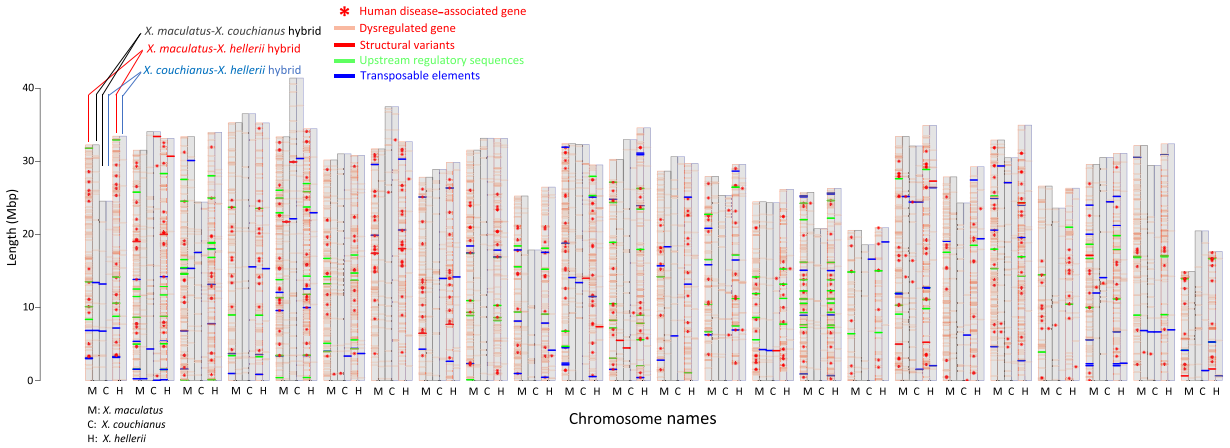

Figure 5. Chromosomal distribution of dysregulated genes in reciprocal *Xiphophorus* hybrids. Chromosomal distribution of hybrid dysregulated genes and their association to structural variants, transposable elements, and upstream regulatory sequences in the parental genomes. (M) *X. maculatus*; (C) *X. couchianus*; (H) *X. hellerii*. Reciprocal F_1 interspecies hybrids are produced between *X. maculatus*, *X. couchianus*, and *X. hellerii*. Dysregulation of the genes is identified by comparing hybrid gene expression to parental species and determined if their expression pattern is different from the parentals (i.e., transgressively expressed in hybrid). A clustered bar graph is used to show the genomic locations of dysregulated genes. Bar height represents chromosome length. Because each species is involved in two types of hybrids, each chromosome per species is split in the middle, with the *left* and *right* halves representing chromosome in the two types of hybrids, as illustrated in the figure. Pink bars represent the chromosomal location of dysregulated genes, with red, green, and blue bars highlighting dysregulated genes adjacent to inter-specific structural variant(s) polymorphisms or showing polymorphisms in upstream regulatory sequences or transposable elements in their −1000 to zero of transcription start site. If the human ortholog of a *Xiphophorus* dysregulated gene is known to be related to disease, the *Xiphophorus* gene is labeled with an asterisk.

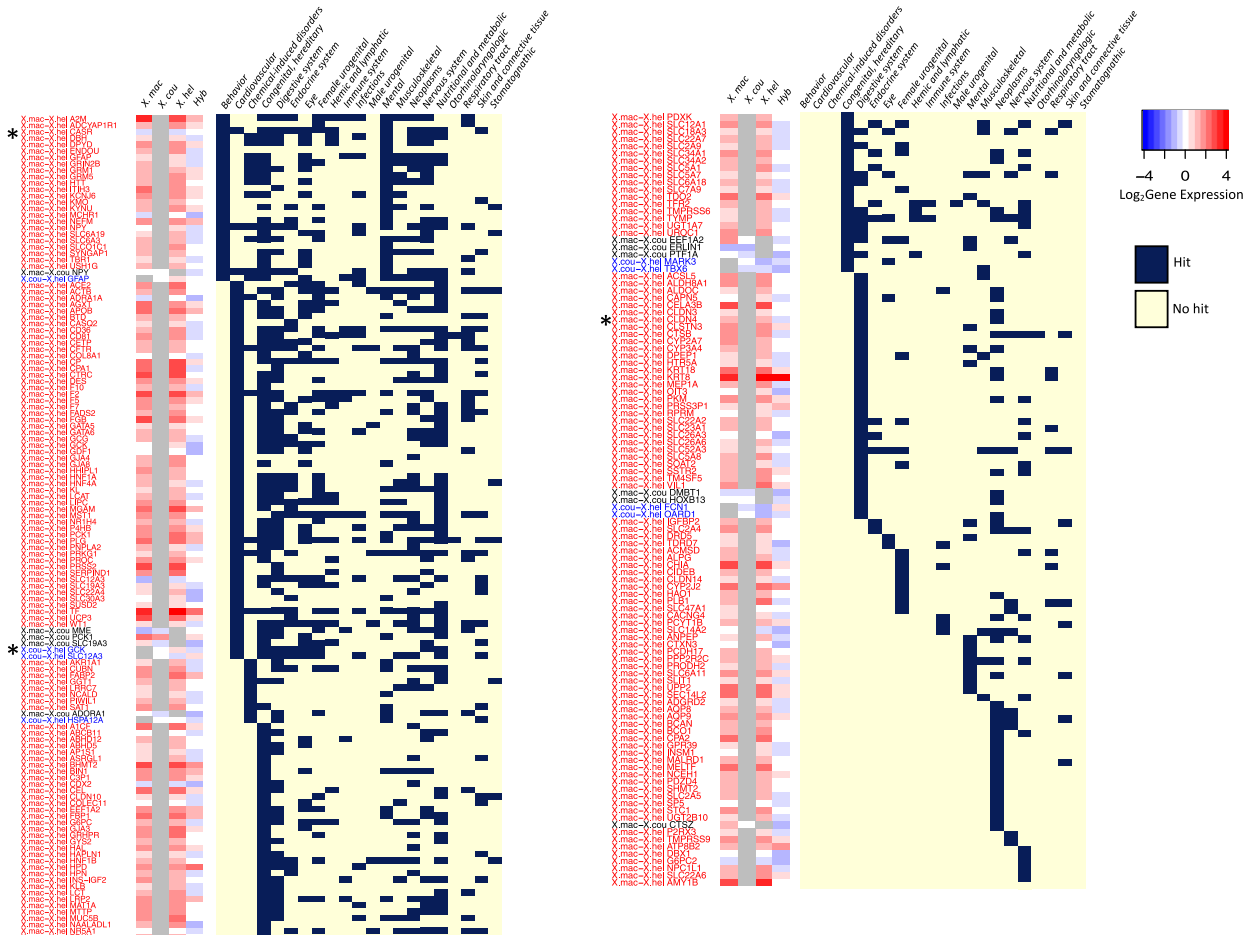

Figure 6. Disease annotation of hybridization-induced dysregulated genes. The expression patterns of disease-related hybridization-dysregulated genes in *X. maculatus* (*X.* mac), *X. couchianus* (*X. cou*), *X. hellerii* (*X. hel*), and interspecies hybrids (*Hyb*) are presented using a heatmap (*left*). Colors represent gene expression levels, with gray meaning "not appliable." Names of each gene are color-coded (red indicates *X. mac–X. hel* hybrids; dark gray, *X. mac–X. cou* hybrids; blue, *X. cou–X. hel* hybrids). The dysregulated genes' associations to diseases are plotted on the *right*. A yellow block means a gene is not associated to a disease type, and a blue block means a gene is associated to a disease type. Asterisks highlight examples of human disease–relevant genes that are dysregulated in *Xiphophorus* hybrids.

been used as standard assemblies in comparing assembly statistics of other species. Using these highly contiguous references allowed us to refine our knowledge of *Xiphophorus* genome evolution, especially in identifying structural adaptations owing to natural or artificial selection, as well as deconvoluting molecular mechanisms of negative epistasis. The earlier version of the *X. maculatus* genome is infested with sequence gaps owing to long repeats. For example, the sequence gaps located in the *xmrk* regulatory region overlap with the promoter regions of a few other genes. In contrast, the new assembly presented in this study has most of the sequence gaps closed, allowing unbiased functional assignment to regions that we were not able to resolve previously. The availability of the latest high-quality *Xiphophorus* genome assemblies led to several discoveries that exemplify the usability of the resources.

Xiphophorus gene family evolution

The phylogenetic tree using the new *Xiphophorus* assemblies reconfirms the evolutionary divergency between the swordtail *X. hellerii* and two platyfish, *X. maculatus* and *X. couchianus* (Cui et al. 2013; Kang et al. 2013). Results show that radiation of the genus *Xiphophorus* occurred in the late Neogen, approximately 5 MYA concurrent with the formation of the Central America ridge connecting the South American plate with North America (Rosen and Bailey 1963). This is consistent with the hypothesis that the genus *Xiphophorus* evolved in the Atlantic drainages around the Trans Mexican Volcanic Belt (Kallman and Kazianis 2006). The more basal position of the southernmost of the three species, *X. hellerii*, may reflect that the Poeciliids have a neotropical origin that colonized Mexico from the south (Rosen and Bailey 1963).

Compared with the egg layers, *Xiphophorus* gene families that are predicted to be involved in processes connected to the viviparity and courting behavior were found to be expanded and/or positively selected (Table 2). The internal fertilization of livebearers has led to the evolution of sperm packets (spermatozeugmata) as an additional morphogenetic step of male gametogenesis. Two expanded gene families, gametogenetins and peroxisomal N(1)-acetyl-spermine/spermidine oxidase *paox*, may play a role in this process. The expansion of the melanocortin receptor 4–like gene represents the known copy number and allelic variation of this gene underlying the polymorphism of puberty and male size in

the genus *Xiphophorus* (Volff et al. 2013; Liu et al. 2020). One of the positively selected genes is *LDL receptor related protein 13*, which in fish is a vitellogenin receptor, a mediator of yolk formation (Reading et al. 2014). Because of the 3- to 4-wk-long intrauterine development of the *Xiphophorus* embryo, the provision with yolk is much increased compared with egg-laying fish, evident from the large eggs of all Poeciliid species. Also, angiopoietin 1, an important regulator of angiogenesis, is positively selected and its possible role in the expansion of gene families involved in vascularization should be discussed in the context of the intricate livebearer yolk sac system.

TEs in *Xiphophorus* evolution

The TE expansion analyses showed there are three main bursts of transposition during the histories of the three genomes, with a peak of highly similar insertions that indicates recent transposition events (Fig. 4B). In comparison, similar analyses drawn on the previous *X. maculatus* genome assembly showed a recent burst of small amplitude compared with more ancient ones (Chalopin et al. 2015). This difference might reflect the refinement of the assembly with the use of long reads that now allows inclusion of more recent insertions. It is also apparent that the three *Xiphophorus* species show some degree of different dynamics in transposition, with *X. hellerii* being distinguished from both the *X. maculatus* and *X. couchianus* (e.g., DTX: TIR transposon). Importantly, polymorphic TE insertions accounted for up to 2% of genome total coverage, underlining their major role in genome evolution. Positively selected gene families are significantly enriched in nonpolymorphic TEs, suggesting the spread of these nonpolymorphic TEs accompanied the evolution of the genes, possibly by introducing new regulatory elements. This observation supports the Britten–Davidson hypothesis, which states TEs can bring transcription factor binding sites and accelerate evolution (Britten and Davidson 1969).

Genome heterozygosity reflects laboratory *Xiphophorus* strain life history

The observed intraspecies heterozygosity for the three species reflects their managed breeding and significant depression in some (i.e., inbred *X. maculatus* is 300-fold lower than its wild counterparts, as well as has an order of magnitude of fewer large deletions in at least one haplotype). The inbred *X. maculatus* line showed a π value of 5.97×10^{-6}, which is much lower than that of inbred *Oryzias latipes* lines, with a heterozygosity accounting for 6.3% of all SNP genotypes ($\pi = 1.34 \times 10^{-3}$) (Fitzgerald et al. 2022), presumably owing to higher inbreeding generations (i.e., 114 generation of inbreeding for *X. maculatus* vs. nine generations of inbreeding for *O. latipes*).

Identification of genome-wide epistasis in *Xiphophorus* interspecies hybrids

Reproductive barriers within *Xiphophorus* species are often formed by efficient prezygotic isolation (Jones et al. 2016; Schumer et al. 2017). However, interspecies hybrids can be produced by enforced mating or artificial insemination, allowing for the assessment of postzygotic isolation. In addition, the F_1 generation hybrid can be further crossed to produce advanced hybrid generations (e.g., intercross, backcross) with mixed parental/ancestral allele genotype, chromosome recombination patterns, and segregating phenotypes. The capability of producing hybrids provides the opportunity to dissect identity, quantity, and mode of loci action affecting complex traits with higher resolution. Therefore, the hybrid system can be used as a forward genetic tool to screen phenotypic changes that result from incompatible genetic interactions. The whole-fish transcriptome comparisons between hybrids and respective parental species identified dysregulated genes associated with diseases of multiple organs, suggesting transcriptional dysregulation in hybrids can be systemic (Fig. 6). It was observed that loci showing dysregulation are not randomly dispersed on chromosomes but rather cluster together (Fig. 5; Supplemental Fig. S8). Considering that previous expression quantitative trait loci studies showed that gene expression is both *cis*- and *trans*-regulated (Lu et al. 2018, 2020a) whereas most known gene regulators from eQTL analyses are *cis*-regulators (Lu et al. 2018), the clustered dysregulation pattern suggests that when the regulatory machinery is interfered by a genome of a different species, the whole cluster of genes is influenced. As proof of concept using the interspecies hybrid to model human disease–relevant epistasis, genes with a known function involved in pathological processes in humans were identified. For examples, *CLDN4* is overexpressed in human ovarian cancer, and its *Xiphophorus* ortholog is dysregulated in the *X. maculatus*–*X. hellerii* hybrid (Litkouhi et al. 2007); *CASR* overexpression is relevant to inflammation, vascular calcification, atherosclerosis, myocardial infarction, hypertension, and obesity, and its *Xiphophorus* ortholog is overexpressed in the *X. maculatus*–*X. hellerii* hybrid (Sundararaman and van der Vorst 2021); and GCK inactivation causes maturity-onset diabetes of the young, and its *Xiphophorus* ortholog is transcriptionally silenced in the *X. couchianus*–*X. hellerii* hybrid (Gloyn 2003). Comparing gene sequences between respective parental species showed these genes show missense mutations in the coding sequences but are not affected by SVs, nor do they show polymorphic TEs or genetic variants within the promotor region. This suggests that the gene expression dysregulation is owing to *trans*-regulators. Nevertheless, the consistency of the molecular phenotypes between human diseases and *Xiphophorus* hybridization-induced gene dysregulation indicates that *Xiphophorus* is a novel model system for exploratory studies to identify gene regulation networks for potential disease control.

In conclusion, the newly assembled high-quality reference genomes for three *Xiphophorus* species provide important information concerning the microevolution of genomes at the species level. The availability of these new resources will promote the utilization of the *Xiphophorus* model system for a wide range of studies. The comparative genomics between the *Xiphophorus* species and dysregulated gene scan performed in reciprocal interspecies hybrids showed that the *Xiphophorus* model system is a unique system for studying disease etiology and for seeking alternative strategies to identify novel therapeutic methods. The knowledge we learned from this study highlights new mechanistic inroads to understand trait variability and stability and leads to new animal models for biomedical research.

Methods

DNA sequencing

High-molecular-weight DNA was isolated from single *Xiphophorus* fishes of each species located at the *Xiphophorus* Genetic Stock Center (XGSC) using the MagAttract kit (Qiagen) according to the manufacturer's protocol. The DNA sample used for *X. maculatus* (southern platyfish) was a male from a laboratory-reared line (strain JP 163A) taken at 114 generations of inbreeding. The *X. couchianus* (northern platyfish) DNA is derived from a female of

Pedigree Xc77(B) in its 77th generation of inbreeding, originally collected from the La Huasteca Canyon, Nuevo Leon, Mexico in 1961, and the *X. hellerii* (orange swordtail) DNA was from a male of the XGSC that originated from the Rio Sarabia and was maintained by brother–sister mating (Fig. 1). All fish were maintained in accordance with an approved institutional animal care and use committee protocol (IACUC 7381). Texas State University has an animal welfare assurance on file with the Office of Laboratory Animal Welfare, National Institute of Health (A4147). Single-molecule real-time (SMRT) sequencing was completed on a Pacific Biosciences (PacBio) RSII instrument, yielding an average read length of ~12 kb. SMRT sequence coverage of more than 50-fold on average was generated using an estimated genome size of 750 Mb. All raw sequences are available under NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA72525 (*X. maculatus*), PRJNA290781 (*X. couchianus*), and PRJNA290782 (*X. hellerii*).

To assess structural variance and polymorphisms within different *Xiphophorus* species or cohorts, four laboratory-maintained inbred *X. maculatus* Jp163A and four *X. maculatus* Jp163B; four wild-caught *X. maculatus* that were collected in 2018 at Rio Jamapa, where the founding fish for both Jp163A and Jp163B lines were captured in 1939; four laboratory-maintained inbred *X. couchianus* (XGSC strain); one laboratory *X. couchianus* that was maintained by closed colony breeding (Wuerzburg strain, WLC 1271); and four laboratory *X. hellerii* that were maintained by crossing males and females of different pedigree were sequenced using the whole-genome shotgun (WGS) method (150-bp pair-ended; Illumina HiSeq X instrument).

Assembly and error correction

For de novo assembly, all SMRT sequences were error-corrected and then assembled with a new revision of FALCON, HGAP4 (SMRT Link v5.0.1.9585) (Chin et al. 2013), followed by contig error correction with Arrow (https://anaconda.org/bioconda/genomicconsensus). Starting with the reference DNA source for each species, additional primary contig polishing of mostly indel errors was performed by aligning ~50× of paired Illumina reads (150 bp length) generated on the Illumina HiSeq X instrument using two successive iterations of Pilon (Walker et al. 2014). Genome size and heterozygosity were estimated using 21-mer default parameter settings within GenomeScope version 1.0 (Vurture et al. 2017).

Assembly scaffolding

To scaffold de novo assembled contigs, we generated BioNano Irys restriction maps for *X. maculatus* and *X. couchianus* that allowed sequence contigs to be ordered and oriented and potential misassemblies to be identified and corrected. We prepared HMW-DNA in agar plugs using a previously established protocol for soft tissues (Lam et al. 2012). Briefly, we followed a series of enzymatic reactions that (1) lysed cells, (2) degraded protein and RNA, and (3) added fluorescent labels to nicked sites using the IrysPrep reagent kit. The nicked DNA fragments were labeled with Alexa Fluor 546 dye, and the DNA molecules were counter-stained with YOYO-1 dye. The labeled DNA fragments were electrophoretically elongated and sized on a single IrysChip, and subsequent imaging and data processing determined the size of each DNA fragment. Finally, a BioNano proprietary algorithm performed a de novo assembly of all labeled fragments >150 kbp into a whole-genome optical map with defined overlap patterns. The individual map was clustered and scored for pairwise similarity, and Euclidian distance matrices were built. Manual refinements were then performed as previously described (Lam et al. 2012). For *X. helleri*, we were unable to recover a high-quality BioNano map; instead, we used a chromosome proximity map. A muscle sample from the SMRT sequenced reference individual was used to lyse the resulting cells. Chromatin was then cross-linked and purified to generate Hi-C libraries as per the protocol instructions in the Phase Genomics kit. Libraries were sequenced from both sides on an Illumina X10 and reads were aligned to the error corrected contigs using BWA V0.7.16 (Li 2014) with strict parameters (-n 0) to prevent mismatches and nonspecific alignments. Only read pairs that aligned to different contigs were used for scaffolding. The Proximo Hi-C pipeline performed chromosome clustering and contig orientation as described previously (Bickhart et al. 2017). A key feature of this multimodule software is the use of SALSA, a process that combines Hi-C and linkage information, which better resolves ambiguous contig orientations (Ghurye et al. 2019). A final manual curation of scaffold order was accomplished with Juicebox (Robinson et al. 2018).

Chromosome builds

Upon chimeric contig correction and completion of the primary scaffolded assembly, we first used Chromonomer (Catchen et al. 2020) to align the X. *maculatus* scaffolds to the genetic linkage map (Amores et al. 2014) and then assigned chromosome coordinates. Using default parameter settings, Chromonomer attempts to find the best set of nonconflicting markers that maximizes the number of scaffolds in the map while minimizing ordering discrepancies. The output is a FASTA file describing the location of scaffolds by chromosome. The X_maculatus-5.0-male chromosome was then used to guide generation of the *X. couchianus* and *X. hellerii* chromosomes. Each assembly was independently aligned to X_maculatus-5.0-male by using NUCmer of the MUMmer4 software (Marçais et al. 2018) and then separately breaking the assembly into 1000-bp nonoverlapping segments to be aligned by BLAT (Kent 2002). Possible breakpoints, defined as where at least 50 kb of sequence aligned to a chromosome other than the primary chromosome for the remainder of the scaffold or where at least 50 kb of sequence aligned to a discontinuous location (>100 kb apart from the neighboring segment), were manually reviewed. Order and orientation were defined initially using the alignments to only the X_maculatus-5.0-male reference. After creation of the chromosomes, chromosomal sequences were again aligned against each other, and careful comparisons were made with any discrepancies subjected to manual review. Importantly, intrachromosomal rearrangements were not altered to only reflect the reference source of alignment; that is, *X. couchianus* chromosomal sequences were not arranged to be a mirror of *X. maculatus*. After these chromosome assignments, any scaffolds that remained were considered unplaced.

Gene annotation

Each assembly was annotated using the previously described NCBI RefSeq workflow (Pruitt et al. 2014), including masking of repeats before ab initio gene predictions and RNA sequencing (RNA-seq) evidence-supported gene model building. Gene annotation relied on an extensive variety of public RNA-seq data from various tissues to improve gene model accuracy. The RefSeq gene annotation reports for each species provide a full accounting of all methodology deployed and their output metrics. The NCBI annotation pipeline of both assemblies included WindowMasker (Morgulis et al. 2006) and RepeatMasker (Smit et al. 2012) steps to delineate and exclude repetitive regions from gene model annotation. The positional coordinates for repeats identified by RepeatMasker are provided in

the BED format at NCBI for each genome. WindowMasker's "nmer" files (counts) were used to regenerate repetitive region BED coordinates (Morgulis et al. 2006).

Structural variant analysis

We used a process described in our previous study and the genome resequencing data to find high-confidence deletions present in the populations of the three species of *Xiphophorus* (Warren et al. 2021). First, we aligned WGS reads to the reference using BWA-MEM v0.7.17 with default options (Li and Durbin 2009). The alignments were postprocessed using the SAMtools v1.14 modules fixmate with flag "-m," sort, and markdup with flag "-r," in that order (Li et al. 2009). The postprocessed alignments were used as input to two structural variant callers: LUMPY (Layer et al. 2014) and Manta (Chen et al. 2016) v1.6.0. To run LUMPY, we used the smoove pipeline (https://github.com/brentp/smoove) v0.2.3 as recommended by LUMPY's documentation. That is, we ran smoove call on each individual sample, then smoove merge to combine all sets of calls from each individual into a merged call set followed by smoove genotype to perform joint genotyping of each sample over the merged call set, and, finally, smoove paste to concatenate the results into a single VCF using the default options for each of these commands. To run Manta, we used the script "configManta.py" with default options and the full list of BAM files from the alignment step to set up the run and then "runWorkflow.py" to run the full Manta workflow with default options. To focus on only the highest-confidence structural variant calls, we limited our subsequent analyses and reported results to only deletions passing all filters in the length range of 500 bp to 100 kbp detected by both LUMPY and Manta. We considered a deletion to be detected by both LUMPY and Manta only if there was a reciprocal overlap of at least 50% of the length of the deletion between the calls made by LUMPY and Manta. We then annotated the merged deletion set based on whether the deletion affected one or more genes' coding sequence, introns, or flanking sequence. We performed this full process of finding deletions on all *X. hellerii* sequenced individuals compared with the *X. hellerii* reference and on all *X. maculatus* sequenced individuals compared with the *X. maculatus* reference. The *X. couchianus* reference was too fragmented to use in this analysis. To find fixed differences between species, we aligned individuals from all three species to the *X. hellerii* reference and performed the same process, defining fixed deletions as homozygous deletions present in every individual of a given species.

Orthology between *Xiphophorus* genomes

Annotated genes were downloaded from RefSeq (*X. maculatus*: GCF_002775205.1_X_maculatus-5.0-male; *X. couchianus*: GCF_001444195.1_X_couchianus-1.0; *X. hellerii*: GCF_003331165.1_X_hellerii-4.1). All-versus-all BLASTN (Altschul et al. 1990) was performed between *X. couchianus*, *X. hellerii*, and *X. maculatus* genes (*e*-value: 1×10^{-10}; best_hit_score_edge: 0.1; best_hit_over hang: 0.1). Gene models of each species are plotted as a Circos plot (Krzywinski et al. 2009) using chromosomal coordinates listed in the gene annotation files (GFF) for each species.

Heterozygosity in *Xiphophorus* genomes

To assess genome heterozygosity of three inbred *Xiphophorus* laboratory lines (i.e., *X. maculatus* Jp163A, *X. maculatus* Jp163B, and *X. couchianus*), laboratory fish that were maintained by inter-strain cross (*X. hellerii*), and a wild population (i.e., wild caught *X. maculatus*), four fish per cohort were resequenced. Short sequencing reads were mapped to corresponding genome assemblies using Bowtie 2 (Langmead and Salzberg 2012) in head-to-head mode (v2.2.4). Single-nucleotide variants (SNVs) and short insertions and deletions (indels) were determined using SAMtools (Li et al. 2009) pileup function, followed by BCFtools (Li et al. 2009; Li 2011) call function or followed by VarScan (Koboldt et al. 2009, 2013). Heterozygous loci were determined by both BCFtools and VarScan genotyping statistics (i.e., BCFtools: GT = 0/1, with Phred scores of alternative genotypes <30; VarScan: $P < 0.01$). Heterozygosity density was calculated per species cohort by binning the quantity of SNVs per 500-kbp sliding window across the whole genome.

Nucleotide diversity

Genotype of polymorphic sites are determined by both BCFtools (Li et al. 2009; Li 2011) and VarScan (Koboldt et al. 2009, 2013). Only loci with determined genotype were included for analyses. Nucleotide diversity (π) is calculated as the total number of heterozygous loci/genome size.

Run of homozygosity analyses

SNVs and indels were determined using SAMtools pileup in variant call format (VCF). The VCF file of each cohort (i.e., *X. maculatus* Jp163A, *X. maculatus* Jp163B, wild *X. maculatus*, *X. couchianus*, and *X. hellerii*) were subsequently used for run of homozygosity (RoH) analyses using BCFtools (v1.12) (Li et al. 2009; Li 2011) with the default setup. Custom R scripts were used for data visualization of autozygous loci within each population.

Annotation of genetic variants

Custom genome databases are established manually using *X. maculatus* genome GCF_002775205.1_X_maculatus-5.0-male and *X. hellerii* genome GCF_003331165.1_X_hellerii-4.1, with corresponding genome annotation files following SnpEff (Cingolani et al. 2012) instructions. Polymorphisms identified between any species pair among *X. maculatus*, *X. couchianus*, and *X. hellerii* were annotated using SnpEff.

TE annotation

TE databases were reconstructed separately in the three genomes using the TEdenovo tool from the REPET pipeline (Quesneville et al. 2005; Flutre et al. 2011; Hoede et al. 2014). As recommended by REPET investigators, we built these banks by using an ~400-Mb subset of each genome. The banks were then filtered using TEannot, another tool of the REPET pipeline, with this step selecting consensi presenting at least one full length copy in the genome. TEannot was then used with the filtered banks to identify TE loci and annotate them in the corresponding genome.

TE landscapes

For each TE family identified by the REPET pipeline, all genomic insertions were retrieved and aligned together using MAFFT (Katoh et al. 2002). The global DNA sequence identity was then computed for each possible pair of sequences, excluding gaps. Landscape graphs were drawn by reporting the total number of pairwise comparisons for a given family to the total genomic density of this family.

Search for polymorphic TEs

Polymorphic elements were searched between pairs of two genomes, with one serving as a reference and another as a target consecutively. For each reference genome, TEs >300 nt were retrieved as well as their 100-bp upstream and downstream flanking sequences.

BLAT (Kent 2002) searches were then managed against the target genome, taking as a query each of the flanking sequences or the element together with its flanking sequences, with the following parameters: tileSize = 12 and minScore = 150. BLAT results were then interpreted in various ways. If the flanking regions were present in only one significant hit each and were found adjacent on the same chromosome or scaffold of the target genome while the whole region (TE plus flanking sequences) did not match entirely anywhere, the TE of the reference genome was considered polymorphic. If the whole segment "element plus flanking regions" presented a significant and complete hit in the target genome, the TE of the reference genome was considered not polymorphic. In all the other cases, and to be the most stringent as possible, we stated the reference element status as "NA," in particular when flanking sequences presented only one significant hit each but on different scaffolds or when flanking sequences corresponded to a repeated TE and presented multiple hits throughout the genome. Even if we could not find a complete match of the segment "TE plus flanking sequences," we considered this output as possibly artifactual and did not consider the TE as polymorphic annotated it as "NA."

Gene family evolution

Gene annotations of *Astyanax mexicanus*, *Clupea harengus*, *Danio rerio*, *Fundulus heteroclitus*, *O. latipes*, *Perca flavescens*, *Takifugu rubripes*, *X. couchianus*, *X. hellerii*, and *X. maculatus* were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) for gene family clustering analysis. For each gene, the longest protein sequence was kept to represent the gene and was pooled together to run an all-versus-all BLAST (Camacho et al. 2009) to calculate pairwise sequence similarity as an H-score (Cho et al. 2013). Based on the H-score, all proteins were clustered into groups (gene families) using Hcluster_sg (Ruan et al. 2008). In our gene family analysis, all the included protein sequences were from NCBI annotations. The annotation does not include protein sequences for pseudogenes, and therefore, there is no pseudogene involved in the analysis.

In groups that have one gene for each species, those genes were identified as one-to-one orthologous genes and were used to reconstruct a species phylogeny tree and to estimate the divergence time between species. One-to-one orthologous genes were first aligned as protein sequences using MAFFT (Nakamura et al. 2018) and then converted into coding sequence alignments using PAL2NAL (Suyama et al. 2006). Protein alignments were trimmed using trimAl (Capella-Gutiérrez et al. 2009) and concatenated and transferred into RAxML v.8.2.9 (Stamatakis 2014) for phylogenetic tree reconstruction. The topology of the tree was further confirmed by MrBayes (Ronquist et al. 2012). Coding sequence alignments were transferred into MCMCTree (Inoue et al. 2010), where species divergence time was estimated. Three fossil calibrations were used: *O. latipes–Tetraodon nigroviridis* (~96.9–150.9 Mya) (Lin et al. 2016), *O. latipes–D. rerio* (~314–332 Mya) (Yamanoue et al. 2006), and Clupeiformes–Cypriniformes (~185–225 Mya) (Near et al. 2012; Hughes et al. 2018).

The gene cluster information and the species tree were then transferred into CAFE5 (https://github.com/hahnlab/CAFE5), in which a gene birth/death rate was set globally in the tree; hence, gene families that are expanded or contracted significantly were identified and evaluated accordingly.

Gene selection

To estimate genes under positive selection in X*iphophorus* species, the protein and cDNA FASTA files for several phylogenetically chosen species of fish were downloaded from NCBI (Supplemental Table S1). Orthologous proteins of all fish were identified using inparanoid (O'Brien et al. 2005) with default settings. For each gene with a protein ortholog across all species, the corresponding protein and cDNA sequences were aligned and converted into a codon alignment using PAL2NAL (version v14) (Suyama et al. 2006). The resulting sequences were aligned by MUSCLE (option: -fastaout) (Edgar 2004), and nonconserved blocks were removed using Gblocks (version 0.91b; options: -b4 10 -b5 n –b3 5 –t = c) (Castresana 2000). The Gblocks output was converted to paml format using an in-house script. The same species tree as established for the gene family dynamics was used. For the phylogenetic analyses by maximum likelihood, the "Environment for Tree Exploration" (ETE3) toolkit (Huerta-Cepas et al. 2016) was used. For the detection of positive selection in X*iphophorus* species, we calculated two branch-site specific models, which involved model bsA1 (neutral) versus model bsA (positive selection) to identify sites under positive selection on a specific branch. To find genes commonly positively selected in all *Xiphophorus* species, the common branch for *X. maculatus*, *X. hellerii*, and *X. couchianus* was marked. To find genes positively selected exclusively in one species, the subbranches for the *Xiphophorus* species were marked separately. Both models were compared using a likelihood ratio test (FDR ≤ 0.05). FDR was calculated using "p.adjust" from the R package "stats." To detect sites under positive selection, naive empirical Bayes (NEB) probabilities for all four classes were calculated for each site. Genes with a probability > 0.95 for either site class 2a (positive selection in marked branch and conserved in rest) or site class 2b (positive selection in marked branch and relaxed in rest) were considered.

Interspecies genome alignment

To predict genomic differences between the three *Xiphoporus* genomes, we aligned the repeatmasked genomes using NUCmer from the MUMmer package (--maxmatch -c 200 -b 700 -l 75) (Delcher et al. 2002). The alignments were filtered using delta-filter, and the Synteny and Rearrangement Identifier (SyRI) (Goel et al. 2019) was used to identify genomic rearrangements from the resulting whole-genome alignments.

Production of interspecies hybrids

The *X. maculatus*, *X. hellerii*, *X. couchianus*, and their reciprocal hybrids (*X. maculatus–X. hellerii* F_1 hybrids, *X. maculatus–X. couchianus* F_1 hybrids, and *X. couchianus–X.* hellerii F_1 hybrids) used in this study were supplied by the *Xiphophorus* Genetic Stock Center. *X. maculatus–X. couchianus* F_1 hybrids were produced by natural breeding between a *X. maculatus* female and *X. couchianus* male, and *X. maculatus–X. hellerii* and *X. couchianus–X. hellerii* F_1 hybrids were produced by artificial insemination.

RNA isolation and RNA sequencing

Total RNA from two whole fishes of each species and hybrids were isolated using Tri Reagent (Sigma-Aldrich). Samples were homogenized in Tri Reagent followed by the addition of 200 µL chloroform, and the samples were vigorously shaken and subjected to centrifugation at 12,000*g* for 15 min at 4°C. Total RNA was further purified using RNeasy mini-RNA isolation kit (Qiagen). Residue DNA was eliminated by performing column DNase digestion for 30 min at 37°C. Total RNA concentration was determined using a spectrophotometer (NanoDrop Technologies). RNA quality was verified on an Agilent Bioanalyzer (Agilent Technologies) to confirm that RIN scores were above 8.0 before sequencing.

RNA sequencing was performed upon libraries constructed using the Illumina TruSeq library preparation system. RNA libraries were sequenced as 125-bp pair-end fragments using the

Illumina HiSeq system (Illumina). Sequencing adaptor sequences were removed from raw sequencing reads. The processed reads were subsequently trimmed and filtered based on quality scores by using a custom filtration algorithm that removes low-scoring sections of each read and preserved the longest remaining fragment.

Identification of dysregulated gene expression in interspecies hybrids

RNA-seq reads were filtered by removing adaptor sequence contamination and removing high error rate base calls. Sequencing reads of parental species were mapped to corresponding reference genomes, and reads of hybrids were mapped to both parental species reference genomes, respectively, using TopHat2 (Kim et al. 2013). Transcriptome profiling was subsequently performed using the Subread package featureCounts (Liao et al. 2014). For gene expression comparison between samples of different genetic backgrounds, orthologs were first identified among the three parental species using reciprocal best hits, followed by converting gene IDs of gene expression count tables to *X. maculatus* orthologs. Hybrid gene expression was quantified by averaging the reads counts mapped to both parental reference genomes. For each type of hybrid, the hybrid gene expression profile was compared to each of the parental profile using the R/Bioconductor package (R Core Team) edgeR (Robinson et al. 2010). Hybridization-induced dysregulated genes were determined if the hybrid gene expression was higher ($\log_2$ fold change > 2; false-discovery rate < 0.05) or lower expressed ($\log_2$ fold change < −2; false-discovery rate < 0.05) than both parental species.

Human disease–associated gene analyses

Human orthologs of dysregulated genes identified from *Xiphophorus* interspecies hybrids were identified using reciprocal best hit. Disease types that are associated with the dysregulated genes were subsequently determined by querying through DisGeNET database (Pinero et al. 2015).

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA700566. All scripts used in this study can be found in Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Milin Kreminski and Tina Graves for BioNano map generation. This work was supported by the National Institutes of Health (NIH) R24OD011198 to W.C.W., NIH R24OD011120 to R.B.W., NIH R24OD031467 to Y.L. and M.S., NIH R15CA223964 to Y.L., and a French-German Collaboration for Joint Projects in Natural, Life and Engineering Sciences (ANR/DFG) joint grant Evobooster to J.-N.V. and M.S. Computational work associated to the study was performed on the learning, exploration, analysis, and processing (LEAP) next-generation high-performance computing cluster at the Texas State University, San Marcos, Texas, and high-performance computing infrastructure was provided by Research Computing Support Services and in part by the National Science Foundation under grant number CNS-1429294 at the University of Missouri, Columbia.

References

Albertson RC, Cresko W, Detrich HW 3rd, Postlethwait JH. 2009. Evolutionary mutant models for human disease. *Trends Genet* **25:** 74–81. doi:10.1016/j.tig.2008.11.006

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410. doi:10.1016/S0022-2836(05)80360-2

Amores A, Catchen J, Nanda I, Warren W, Walter R, Schartl M, Postlethwait JH. 2014. A RAD-tag genetic map for the platyfish (*Xiphophorus maculatus*) reveals mechanisms of karyotype evolution among teleost fish. *Genetics* **197:** 625–641. doi:10.1534/genetics.114.164293

Beck EA, Healey HM, Small CM, Currey MC, Desvignes T, Cresko WA, Postlethwait JH. 2022. Advancing human disease research with fish evolutionary mutant models. *Trends Genet* **38:** 22–44. doi:10.1016/j.tig.2021.07.002

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* **49:** 643–650. doi:10.1038/ng.3802

Biémont C, Vieira C. 2006. Genetics: junk DNA as an evolutionary force. *Nature* **443:** 521–524. doi:10.1038/443521a

Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN. 2008. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res* **16:** 203–215. doi:10.1007/s10577-007-1202-6

Bowen NJ, Jordan IK. 2002. Transposable elements and the evolution of eukaryotic complexity. *Curr Issues Mol Biol* **4:** 65–76. doi:10.21775/cimb.004.065

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science* **165:** 349–357. doi:10.1126/science.165.3891.349

Brosius J. 1991. Retroposons—seeds of evolution. *Science* **251:** 753. doi:10.1126/science.1990437

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10:** 421. doi:10.1186/1471-2105-10-421

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972–1973. doi:10.1093/bioinformatics/btp348

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17:** 540–552. doi:10.1093/oxfordjournals.molbev.a026334

Catchen J, Amores A, Bassham S. 2020. Chromonomer: a tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. *G3 (Bethesda)* **10:** 4115–4128. doi:10.1534/g3.120.401485

Chalopin D, Naville M, Plard F, Galiana D, Volff JN. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol Evol* **7:** 567–580. doi:10.1093/gbe/evv005

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32:** 1220–1222. doi:10.1093/bioinformatics/btv710

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10:** 563–569. doi:10.1038/nmeth.2474

Cho YS, Hu L, Hou H, Lee H, Xu J, Kwon S, Oh S, Kim HM, Jho S, Kim S, et al. 2013. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat Commun* **4:** 2433. doi:10.1038/ncomms3433

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly (Austin)* **6:** 80–92. doi:10.4161/fly.19695

Cui R, Schumer M, Kruesi K, Walter R, Andolfatto P, Rosenthal GG. 2013. Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes. *Evolution (N Y)* **67:** 2166–2179. doi:10.1111/evo.12099

Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. 2003. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* **13:** 651–658. doi:10.1016/j.gde.2003.10.013

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30:** 2478–2483. doi:10.1093/nar/30.11.2478

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792–1797. doi:10.1093/nar/gkh340

Fedoroff NV. 1999. Transposable elements as a molecular evolutionary force. *Ann N Y Acad Sci* **870:** 251–264. doi:10.1111/j.1749-6632.1999.tb08886.x

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41:** 331–368. doi:10.1146/annurev.genet.40.110405.090448

Fitzgerald T, Brettell I, Leger A, Wolf N, Kusminski N, Monahan J, Barton C, Herder C, Aadepu N, Gierten J, et al. 2022. The Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel. *Genome Biol* **23:** 59. doi:10.1186/s13059-022-02623-z

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in *de novo* annotation approaches. *PLoS One* **6:** e16526. doi:10.1371/journal.pone.0016526

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol* **15:** e1007273. doi:10.1371/journal.pcbi.1007273

Gloyn AL. 2003. Glucokinase (*GCK*) mutations in hyper- and hypoglycemia: maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemia of infancy. *Hum Mutat* **22:** 353–362. doi:10.1002/humu.10277

Goel M, Sun H, Jiao WB, 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

Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H. 2014. PASTEC: an automatic transposable element classification tool. *PLoS One* **9:** e91929. doi:10.1371/journal.pone.0091929

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* **33:** 1635–1638. doi:10.1093/molbev/msw046

Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, Arcila D, Betancur RR, Li C, Becker L, et al. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proc Natl Acad Sci* **115:** 6249–6254. doi:10.1073/pnas.1719358115

Inoue JG, Miya M, Lam K, Tay BH, Danks JA, Bell J, Walker TI, Venkatesh B. 2010. Evolutionary origin and phylogeny of the modern holocephalans (Chondrichthyes: Chimaeriformes): a mitogenomic perspective. *Mol Biol Evol* **27:** 2576–2586. doi:10.1093/molbev/msq147

Jones JC, Fruciano C, Keller A, Schartl M, Meyer A. 2016. Evolution of the elaborate male intromittent organ of *Xiphophorus* fishes. *Ecol Evol* **6:** 7207–7220. doi:10.1002/ece3.2396

Kallman KD, Kazianis S. 2006. The genus *Xiphophorus* in Mexico and Central America. *Zebrafish* **3:** 271–285. doi:10.1089/zeb.2006.3.271

Kang JH, Schartl M, Walter RB, Meyer A. 2013. Comprehensive phylogenetic analysis of all species of swordtails and platies (Pisces: genus *Xiphophorus*) uncovers a hybrid origin of a swordtail fish, *Xiphophorus monticolus*, and demonstrates that the sexually selected sword originated in the ancestral lineage of the genus, but was lost again secondarily. *BMC Evol Biol* **13:** 25. doi:10.1186/1471-2148-13-25

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30:** 3059–3066. doi:10.1093/nar/gkf436

Kazazian HH Jr. 2004. Mobile elements: drivers of genome evolution. *Science* **303:** 1626–1632. doi:10.1126/science.1089670

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res* **12:** 656–664. doi:10.1101/gr.229202

Kidwell MG, Lisch DR. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution (N Y)* **55:** 1–24. doi:10.1111/j.0014-3820.2001.tb01268.x

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14:** R36. doi:10.1186/gb-2013-14-4-r36

Klotz B, Kneitz S, Regensburger M, Hahn L, Dannemann M, Kelso J, Nickel B, Lu Y, Boswell W, Postlethwait J, et al. 2018. Expression signatures of early-stage and advanced medaka melanomas. *Comp Biochem Physiol C Toxicol Pharmacol* **208:** 20–28. doi:10.1016/j.cbpc.2017.11.005

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25:** 2283–2285. doi:10.1093/bioinformatics/btp373

Koboldt DC, Larson DE, Wilson RK. 2013. Using VarScan 2 for germline variant calling and somatic mutation detection. *Curr Protoc Bioinformatics* **44:** 15.4.1–15.4.17. doi:10.1002/0471250953.bi1504s44

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645. doi:10.1101/gr.092759.109

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, et al. 2012. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol* **30:** 771–776. doi:10.1038/nbt.2303

Lampert KP, Schmidt C, Fischer P, Volff JN, Hoffmann C, Muck J, Lohse MJ, Ryan MJ, Schartl M. 2010. Determination of onset of sexual maturation and mating behavior by melanocortin receptor 4 polymorphisms. *Curr Biol* **20:** 1729–1734. doi:10.1016/j.cub.2010.08.029

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15:** R84. doi:10.1186/gb-2014-15-6-r84

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27:** 2987–2993. doi:10.1093/bioinformatics/btr509

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30:** 2843–2851. doi:10.1093/bioinformatics/btu356

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760. doi:10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Li J, Huang JP, Sukumaran J, Knowles LL. 2018. Microevolutionary processes impact macroevolutionary patterns. *BMC Evol Biol* **18:** 123. doi:10.1186/s12862-018-1236-8

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656

Lin Q, Fan S, Zhang Y, Xu M, Zhang H, Yang Y, Lee AP, Woltering JM, Ravi V, Gunter HM, et al. 2016. The seahorse genome and the evolution of its specialized morphology. *Nature* **540:** 395–399. doi:10.1038/nature20595

Litkouhi B, Kwong J, Lo CM, Smedley JG, McClane BA, Aponte M, Gao Z, Sarno JL, Hinners J, Welch WR, et al. 2007. Claudin-4 overexpression in epithelial ovarian cancer is associated with hypomethylation and is a potential target for modulation of tight junction barrier function using a C-terminal fragment of clostridium perfringens enterotoxin. *Neoplasia* **9:** 304–314. doi:10.1593/neo.07118

Liu R, Du K, Ormanns J, Adolfi MC, Schartl M. 2020. Melanocortin 4 receptor signaling and puberty onset regulation in *Xiphophorus* swordtails. *Gen Comp Endocrinol* **295:** 113521. doi:10.1016/j.ygcen.2020.113521

Lu Y, Bowswell M, Bowswell W, Yang K, Schartl M, Walter RB. 2015. Molecular genetic response of *Xiphophorus maculatus–X. couchianus* interspecies hybrid skin to UVB exposure. *Comp Biochem Physiol C Toxicol Pharmacol* **178:** 86–92. doi:10.1016/j.cbpc.2015.07.011

Lu Y, Boswell M, Boswell W, Kneitz S, Hausmann M, Klotz B, Regneri J, Savage M, Amores A, Postlethwait J, et al. 2017a. Molecular genetic analysis of the melanoma regulatory locus in *Xiphophorus* interspecies hybrids. *Mol Carcinog* **56:** 1935–1944. doi:10.1002/mc.22651

Lu Y, Klimovich CM, Robeson KZ, Boswell W, Ríos-Cardenas O, Walter RB, Morris MR. 2017b. Transcriptome assembly and candidate genes involved in nutritional programming in the swordtail fish *Xiphophorus multilineatus*. *PeerJ* **5:** e3275. doi:10.7717/peerj.3275

Lu Y, Boswell M, Boswell W, Kneitz S, Klotz B, Savage M, Salinas R, Marks R, Regneri J, Postlethwait J, et al. 2018. Gene expression variation and parental allele inheritance in a *Xiphophorus* interspecies hybridization model. *PLoS Genet* **14:** e1007875. doi:10.1371/journal.pgen.1007875

Lu Y, Olivas TJ, Boswell M, Boswell W, Warren WC, Schartl M, Walter RB. 2020a. Intra-strain genetic variation of platyfish (*Xiphophorus maculatus*) strains determines tumorigenic trajectory. *Front Genet* **11:** 562594. doi:10.3389/fgene.2020.562594

Lu Y, Sandoval A, Voss S, Lai Z, Kneitz S, Boswell W, Boswell M, Savage M, Walter C, Warren W, et al. 2020b. Oncogenic allelic interaction in *Xiphophorus* highlights hybrid incompatibility. *Proc Natl Acad Sci* **117:** 29786–29794. doi:10.1073/pnas.2010133117

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol* **14:** e1005944. doi:10.1371/journal.pcbi.1005944

McClintock B. 1956. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* **21:** 197–216. doi:10.1101/SQB.1956.021.01.017

Mishra RR, Kneitz S, Schartl M. 2014. Comparative analysis of melanoma deregulated miRNAs in the medaka and *Xiphophorus* pigment cell cancer models. *Comp Biochem Physiol C Toxicol Pharmacol* **163:** 64–76. doi:10.1016/j.cbpc.2014.01.002

Morgulis A, Gertz EM, Schäffer AA, Agarwala R. 2006. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22:** 134–141. doi:10.1093/bioinformatics/bti774

Muotri AR, Marchetto MC, Coufal NG, Gage FH. 2007. The necessary junk: new functions for transposable elements. *Hum Mol Genet* **16 Spec No. 2:** R159–R167. doi:10.1093/hmg/ddm196

Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34:** 2490–2492. doi:10.1093/bioinformatics/bty121

NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44:** D7–D19. doi:10.1093/nar/gkv1290

Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, Wainwright PC, Friedman M, Smith WL. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. *Proc Natl Acad Sci* **109:** 13698–13703. doi:10.1073/pnas.1206625109

O'Brien KP, Remm M, Sonnhammer EL. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33:** D476–D480. doi:10.1093/nar/gki107

Oliver KR, Greene WK. 2011. Mobile DNA and the TE-thrust hypothesis: supporting evidence from the primates. *Mob DNA* **2:** 8. doi:10.1186/1759-8753-2-8

Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)* **2015:** bav028. doi:10.1093/database/bav028

Powell DL, García-Olázabal M, Keegan M, Reilly P, Du K, Díaz-Loyo AP, Banerjee S, Blakkan D, Reich D, Andolfatto P, et al. 2020. Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. *Science* **368:** 731–736. doi:10.1126/science.aba5216

Powell DL, Payne C, Banerjee SM, Keegan M, Bashkirova E, Cui R, Andolfatto P, Rosenthal GG, Schumer M. 2021. The genetic architecture of variation in the sexually selected sword ornament and its evolution in hybrid populations. *Curr Biol* **31:** 923–935.e11. doi:10.1016/j.cub.2020.12.049

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, et al. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42:** D756–D763. doi:10.1093/nar/gkt1114

Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* **1:** 166–175. doi:10.1371/journal.pcbi.0010022

R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Reading BJ, Hiramatsu N, Schilling J, Molloy KT, Glassbrook N, Mizuta H, Luo W, Baltzegar DA, Williams VN, Todo T, et al. 2014. Lrp13 is a novel vertebrate lipoprotein receptor that binds vitellogenins in teleost fishes. *J Lipid Res* **55:** 2287–2295. doi:10.1194/jlr.M050286

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. 2018. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst* **6:** 256–258.e1. doi:10.1016/j.cels.2018.01.001

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61:** 539–542. doi:10.1093/sysbio/sys029

Rosen DE, Bailey RM. 1963. The poeciliid fishes (Cyprinodontiformes), their structure, zoogeography, and systematics. *Bull Am Mus Nat Hist* **126:** 1–176.

Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, Heriche JK, Hu Y, Kristiansen K, Li R, et al. 2008. TreeFam: 2008 update. *Nucleic Acids Res* **36:** D735–D740. doi:10.1093/nar/gkm1005

Schartl M. 2014. Beyond the zebrafish: diverse fish species for modeling human disease. *Dis Model Mech* **7:** 181–192. doi:10.1242/dmm.012245

Schartl M, Wilde B, Laisney JA, Taniguchi Y, Takeda S, Meierjohann S. 2010. A mutated EGFR is sufficient to induce malignant melanoma with genetic background-dependent histopathologies. *J Invest Dermatol* **130:** 249–258. doi:10.1038/jid.2009.213

Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volff JN, Lesch KP, et al. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nat Genet* **45:** 567–572. doi:10.1038/ng.2604

Schartl M, Kneitz S, Ormanns J, Schmidt C, Anderson JL, Amores A, Catchen J, Wilson C, Geiger D, Du K, et al. 2021. The developmental and genetic architecture of the sexually selected male ornament of swordtails. *Curr Biol* **31:** 911–922.e4. doi:10.1016/j.cub.2020.11.028

Schumer M, Powell DL, Delclós PJ, Squire M, Cui R, Andolfatto P, Rosenthal GG. 2017. Assortative mating and persistent reproductive isolation in hybrids. *Proc Natl Acad Sci* **114:** 10936–10941. doi:10.1073/pnas.1711238114

Shen Y, Chalopin D, Garcia T, Boswell M, Boswell W, Shiryev SA, Agarwala R, Volff JN, Postlethwait JH, Schartl M, et al. 2016. *X. couchianus* and *X. hellerii* genome models provide genomic variation insight among *Xiphophorus* species. *BMC Genomics* **17:** 37. doi:10.1186/s12864-015-2361-z

Smit A, Hubley R, Green P. 2012. RepeatMasker Open-4.0. http://www.repeatmasker.org/.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313. doi:10.1093/bioinformatics/btu033

Sundararaman SS, van der Vorst EPC. 2021. Calcium-sensing receptor (CaSR), its impact on inflammation and the consequences on cardiovascular health. *Int J Mol Sci* **22:** 2478. doi:10.3390/ijms22052478

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34:** W609–W612. doi:10.1093/nar/gkl315

Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28:** 913–922. doi:10.1002/bies.20452

Volff JN, Selz Y, Hoffmann C, Froschauer A, Schultheis C, Schmidt C, Zhou Q, Bernhardt W, Hanel R, Böhne A, et al. 2013. Gene amplification and functional diversification of melanocortin 4 receptor at an extremely polymorphic locus controlling sexual maturation in the platyfish. *Genetics* **195:** 1337–1352. doi:10.1534/genetics.113.155952

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33:** 2202–2204. doi:10.1093/bioinformatics/btx153

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9:** e112963. doi:10.1371/journal.pone.0112963

Warren WC, Boggs TE, Borowsky R, Carlson BM, Ferrufino E, Gross JB, Hillier L, Hu Z, Keene AC, Kenzior A, et al. 2021. A chromosome-level genome of *Astyanax mexicanus* surface fish for comparing population-specific genetic differences contributing to trait evolution. *Nat Commun* **12:** 1447. doi:10.1038/s41467-021-21733-z

Wellbrock C, Weisser C, Geissinger E, Troppmair J, Schartl M. 2002. Activation of $p59^{Fyn}$ leads to melanocyte dedifferentiation by influencing MKP-1-regulated mitogen-activated protein kinase signaling. *J Biol Chem* **277:** 6443–6454. doi:10.1074/jbc.M110684200

Wittbrodt J, Adam D, Malitschek B, Mäueler W, Raulf F, Telling A, Robertson SM, Schartl M. 1989. Novel putative receptor tyrosine kinase encoded by the melanoma-inducing *Tu* locus in *Xiphophorus*. *Nature* **341:** 415–421. doi:10.1038/341415a0

Yamanoue Y, Miya M, Inoue JG, Matsuura K, Nishida M. 2006. The mitochondrial genome of spotted green pufferfish *Tetraodon nigroviridis* (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. *Genes Genet Syst* **81:** 29–39. doi:10.1266/ggs.81.29

Yang Q, Yan C, Gong Z. 2017. Activation of liver stromal cells is associated with male-biased liver tumor initiation in *xmrk* and *Myc* transgenic zebrafish. *Sci Rep* **7:** 10315. doi:10.1038/s41598-017-10529-1

Zheng W, Li Z, Nguyen AT, Li C, Emelyanov A, Gong Z. 2014. *Xmrk*, *Kras* and *Myc* transgenic zebrafish liver cancer models share molecular signatures with subsets of human hepatocellular carcinoma. *PLoS One* **9:** e91179. doi:10.1371/journal.pone.0091179

Received October 25, 2022; accepted in revised form March 29, 2023.

High resolution genomes of multiple *Xiphophorus* species provide new insights into microevolution, hybrid incompatibility, and epistasis

Yuan Lu, Edward Rice, Kang Du, et al.

Genome Res. 2023 33: 557-571 originally published online May 5, 2023
Access the most recent version at doi:10.1101/gr.277434.122