Article

Gene conversion empowers natural selection in a clonal fish species

https://doi.org/10.1038/s41586-026-10180-9

Received: 12 February 2025

Accepted: 23 January 2026

Published online: 11 March 2026

Check for updates

Edward S. Ricemeyer[1,2,14]✉, Nathan K. Schaefer[3,4,5,14], Kang Du[6], Irene da Cruz[7], Susanne Kneitz[8], Rafael D. Acemel[9], Darío G. Lupiáñez[9], Rachel A. Carroll[1], Rosie Drinkwater[2], Manfred Schartl[10,11,12,15] & Wesley C. Warren[1,13,15]

Sexual reproduction is ancient and ubiquitous despite its obvious disadvantages[1]. Theory predicts that the reassortment of alleles that results from sex is necessary for natural selection to act effectively on individual loci; therefore, a purely clonal organism should rapidly accumulate deleterious mutations and go extinct[2–4]. Nevertheless, many asexual species have existed for longer than theory predicts is possible[5–7], such as the Amazon molly (*Poecilia formosa*), a clonally reproducing fish arising from a single hybridization event more than 100,000 years ago[8–10]. Here we show that although the Amazon molly has accumulated mutations faster than its sexual progenitor species, this has not led to functional mutational decay, defying theoretical expectations. Instead, gene conversion facilitates both adaptive and purifying selection by generating new clonal lineages in which previous mutations are either reverted or fixed, and by resolving hybrid incompatibilities between the ancestral haplotypes. The transition to clonality altered chromatin structure, but the asexual haplotypes of the Amazon molly nonetheless maintain the divergent mutational landscapes of their progenitor species. Together, these results provide new insights into long-standing questions about the trade-offs involved in asexual reproduction.

Genetic recombination during meiosis is a key force shaping genomes and driving organismal evolution[3,4,11]. By decoupling mutations, recombination allows locus-specific selection to spread adaptive mutations and purge deleterious mutations. Meiotic recombination is widespread among eukaryotes and is intimately connected to sexual reproduction.

This understanding has led to the perception that reproduction without recombination, such as clonal reproduction in unisexual lineages, is evolutionarily unstable. This theory predicts several consequences of a switch to asexual reproduction: haplotypes should diverge rapidly (Meselson effect)[6,12,13], deleterious mutations should accumulate because they cannot be efficiently purged (Muller's ratchet)[2,3,14], and positive selection should occur inefficiently because beneficial mutations from different lineages cannot be combined, slowing adaptation (Fisher–Muller hypothesis)[15]. Therefore, asexual lineages should be at a disadvantage relative to sexually reproducing species[11,15–17]. Nonetheless, clonally reproducing multicellular species, although rare[18], are more numerous than historically thought, with several[5,7,10] existing for longer than theoretical predictions allow[19]. Answers to how they defy these grim prognoses should be found in their genomes.

Without high-quality genomic resources, these effects have been difficult to measure. Of the estimated several thousand parthenogenetic animal species, genome assemblies exist for only a few, most of which are highly fragmented[20], limiting their utility. Despite this, preliminary investigations of these genomes have confirmed different theoretical predictions about the evolution of asexual genomes in different species, suggesting that the consequences of parthenogenetic evolution may be lineage-specific[20].

The Amazon molly (*Poecilia formosa*) was the first known clonal vertebrate[8]. Like most of the approximately 100 other known asexual vertebrates[21–23], the Amazon molly arose as an interspecific hybrid[24–27]: all Amazon mollies descend from a single cross between a female *Poecilia mexicana* and a male *Poecilia latipinna*[9,10] that arose at least 100 thousand years ago (ka) near Tampico, Mexico[27,28]. In the ovary of *P. formosa*, the germ cells undergo achiasmatic meiosis, diploid oocytes are generated by apomixis as a result of the failure in synapsis of homologous chromosomes, and the chromosomes of primary oocytes initiate pachytene but do not proceed to bivalent formation and meiotic crossovers[29]. Thus, no recombination or chromosome segregation

[1]Bond Life Sciences Center, University of Missouri, Columbia, MO, USA. [2]Institute of Animal Systems Genomics, Faculty of Veterinary Medicine, Ludwig-Maximilians-Universität, Munich, Germany. [3]The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, San Francisco, CA, USA. [4]Department of Neurology, University of California, San Francisco, San Francisco, CA, USA. [5]Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. [6]Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, China. [7]Developmental Biochemistry, Biocenter, University of Würzburg, Würzburg, Germany. [8]Biochemistry and Cell Biology, Biocenter, University of Würzburg, Würzburg, Germany. [9]Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide/Junta de Andalucía, Seville, Spain. [10]Institute of Pathology, University of Würzburg, Würzburg, Germany. [11]Institute for Molecular Life Sciences, Texas State University, San Marcos, TX, USA. [12]Research Department for Limnology, University of Innsbruck, Mondsee, Austria. [13]Division of Animal Sciences, Department of Surgery, Institute for Data Science and Informatics, University of Missouri, Columbia, MO, USA. [14]These authors contributed equally: Edward S. Ricemeyer, Nathan K. Schaefer. [15]These authors jointly supervised this work: Manfred Schartl, Wesley C. Warren. ✉e-mail: E.Ricemeyer@lmu.de

occurs during meiosis, leading to clonal, all-female offspring[30–32]. The Amazon molly, having arisen at least 100 ka, has survived far beyond its predicted extinction time[33], estimated based on modelling and simulation to be fewer than 10,000 years.

A first reference genome of the Amazon molly has been available since 2018[10], but this short-read assembly is a composite of the two parental haplotypes and thus cannot be used to study the two haplotypes independently. Using long-read sequencing and haploid genome assembly techniques, we introduce chromosome-level haplotype-resolved assemblies and gene annotations of *P. formosa* and its progenitor species *P. latipinna* and *P. mexicana*. These genomes present a unique opportunity to study the parallel evolution of sexual and asexual sister genomes, and to investigate long-standing theoretical predictions about the effects of loss of recombination on genome evolution.

Ancestral reconstruction

To better understand the evolution of the Amazon molly genome, we took advantage of its hybrid genome composition and used the trio-binning assembly technique[34,35] to create separate assemblies of the *P. latipinna* and *P. mexicana*-derived haplotypes of *P. formosa* (hereafter PforHlat and PforHmex, respectively) (Fig. 1a). We also assembled a single reference genome for *P. mexicana* and one for *P. latipinna*, which we refer to hereafter as Plat and Pmex. All genomes were assembled to the chromosome level (Supplementary Data Table 1) and we observe no structural rearrangements between the genomes of the parental species (Extended Data Fig. 1a).

Gene flow between the parental species or introgression from the parental species into *P. formosa* could complicate findings, so we looked for evidence of gene flow using short-read samples from these three species and an outgroup. We found evidence for limited gene flow postdating *P. formosa* speciation from *P. mexicana* into two *P. latipinna* genomes, along with a smaller amount of possible *P. latipinna* introgression into two *P. mexicana* genomes (Extended Data Figs. 1b–f and 2, Supplementary Note 1 and Supplementary Data Tables 2–4). We therefore excluded potentially admixed regions from all downstream analyses that could be confounded by this admixture.

We next used these four assemblies to perform whole-genome assembly-based ancestral reconstruction of the ancient *P. mexicana* and *P. latipinna* genomes (AncMex and AncLat, respectively), as well as the ancestor of *P. mexicana*, *P. formosa* and *P. latipinna* (AncMol; Fig. 1b). Comparing these ancestral genomes to their present-day descendants enabled us to explore long-standing hypotheses about asexual reproduction.

Asexual genomes defy Muller's ratchet

Muller's ratchet predicts that deleterious mutations should quickly accumulate in clonally reproducing genomes[2–4]. Simulation has predicted that this accumulation of deleterious mutations should have driven *P. formosa* to extinction within 10,000 years despite the persistence of the species for more than 100,000 years (refs. 10,33).

To quantify accumulation of deleterious mutations, we computed the ratio of rates of nonsynonymous to synonymous substitution in coding sequence (dN/dS) between sister asexual and sexual branches of the tree. Bacteria in endosymbiotic relationships with their hosts and therefore obligate asexual reproduction show evidence for Muller's ratchet via increased dN/dS compared with their free-living sexually reproducing relatives[36]. To determine whether the switch to asexual evolution in *P. formosa* also had this effect, we annotated the four genome assemblies as well as the ancestral reconstructions AncLat and AncMex, and used the coding sequences to compare dN/dS along asexual and sexual lineages (Fig. 1b). Contrary to the increase in dN/dS expected under Muller's ratchet, we find negligible differences in dN/dS between the sexual and asexual branches (Fig. 2a and

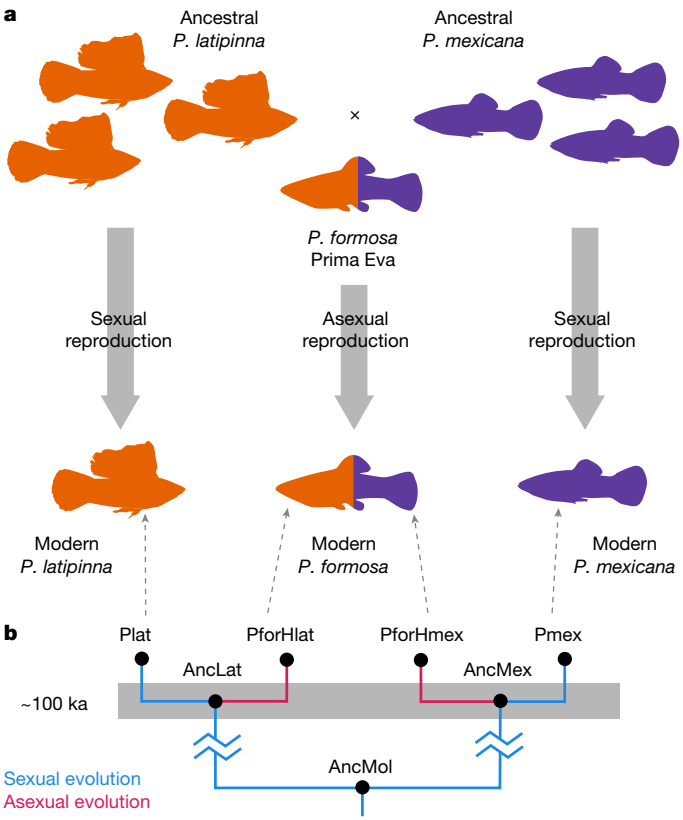


Fig. 1 | Origin and phylogeny of the Amazon molly *P. formosa*. a, Origin of the Amazon molly *P. formosa* from a single hybridization between *P. latipinna* and *P. mexicana* approximately 100 ka. Since then, *P. formosa* has been reproducing asexually, whereas *P. latipinna* and *P. mexicana* have been reproducing sexually. **b**, Genomes assembled from *P. latipinna* (Plat) and *P. mexicana* (Pmex) as well as both haplotypes of *P. formosa*: the one descending from *P. latipinna* (PforHlat) and the one descending from *P. mexicana* (PforHmex). Branches are coloured by sexual versus asexual evolution. PforHlat and PforHmex, despite being present in the same fish, are less related to each other than they are to their sexually reproducing sister species. Ancestral genomes AncLat, AncMex and AncMol were reconstructed. *P. formosa* and *P. latipinna* silhouettes by Kamil S. Jaron (CC0 1.0); *P. mexicana* silhouette by Michael Tobler (CC0 1.0).

Supplementary Note 2). Using population-level data and within-species polymorphism at different codon positions rather than comparison to ancestral reconstruction also shows negligible difference in distributions of total polymorphism between present-day sexual and asexual populations (Fig. 2b and Supplementary Note 2).

We next looked for reduced efficacy of purifying selection in the asexual relative to the sexual genomes by counting heterozygous and homozygous mutations per individual, broken down by predicted variant effect and excluding fixed differences between the parental species (Fig. 2c). Because of the hybrid origin and clonal reproduction of *P. formosa*, all individuals have few non-reference homozygous mutations, regardless of variant effect. The hybrid origin of *P. formosa* results in more heterozygous mutations than its progenitor species, as measured previously[10], but this heterozygosity is reduced by 47% for high-impact (probably deleterious) mutations relative to neutral mutations, demonstrating the action of purifying selection.

Together, these results show little evidence for Muller's ratchet, consistent with previous work in the Amazon molly[10].

Faster divergence of asexual haplotypes

Another proposed effect of the faster mutational accumulation expected under obligate asexual reproduction is the Meselson effect,

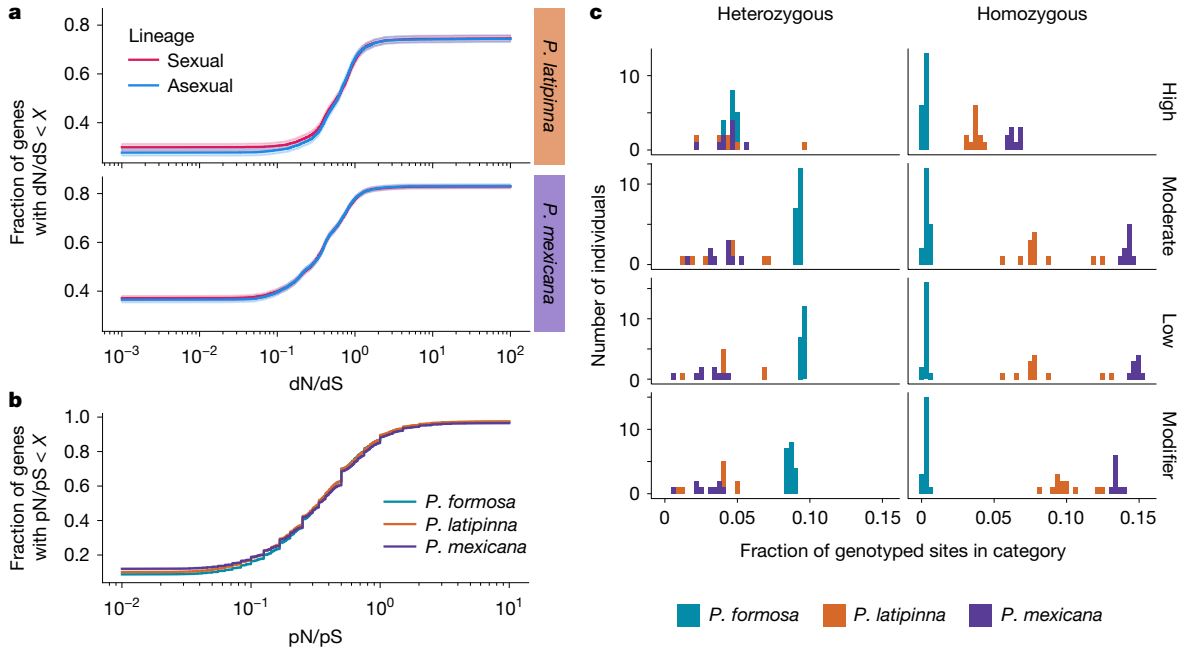


Fig. 2 | Asexual *P. formosa* genomes show at most limited mutational decay compared with sexual progenitor species. a, Cumulative distributions of per-gene ratios of rates of nonsynonymous to synonymous substitution (dN/dS) along sexual and asexual lineages of *P. latipinna* (top) and *P. mexicana* (bottom, curves overlapping). Curve centres represent true dN/dS distributions and 95% confidence intervals shown around curves are based on bootstrapping these distributions with 1,000 resamples. **b**, Cumulative distributions of per-gene ratios of nonsynonymous to synonymous polymorphism (pN/pS) at the population level within each species. The pN/pS distribution of *P. formosa* is negligibly different from those of the sexual progenitor species, indicating similar distributions of total polymorphism in present-day asexual versus sexual species. Curve centres represent true pN/pS distributions and 95% confidence intervals shown around curves are based on bootstrapping with 1,000 resamples. **c**, Incidence of variants with different predicted functional impacts within individuals from each species. For each variant impact category, the number of heterozygous or homozygous variant calls of that type per individual are shown, normalized by the total number of genotyped loci containing variants of that impact type in that individual, with fixed differences between the parental species excluded from *P. formosa*.

in which genomes diverge more rapidly along asexual lineages than along sexual lineages[12,13]. We looked for evidence of this change in divergence in *P. formosa* relative to *P. latipinna* and *P. mexicana* to determine whether the lack of Muller's ratchet results from the absence of increased divergence or a mechanism for eliminating deleterious mutations post hoc.

As predicted by the Meselson effect, we find that PforHlat has accumulated more single-nucleotide variants (SNVs) than Plat in the time since their most recent common ancestor (MRCA; Fig. 3a). Similarly, in the time since PforHmex and Pmex diverged, PforHmex has accumulated more SNVs than Pmex (Fig. 3a). A neighbour-joining tree built using whole-genome pairwise mash distances confirms this finding (Extended Data Fig. 3). This observation of the Meselson effect suggests that the mutational accumulation that is expected to drive both Muller's ratchet and the Meselson effect is indeed occurring, raising the question of how the species has avoided Muller's ratchet despite faster mutational accumulation.

Haplotype-specific divergence

Notably, both sexual and asexual *P. mexicana* genomes (Pmex and PforHmex) are much more diverged from their MRCA AncMex than the sexual and asexual *P. latipinna* genomes (Plat and PforHlat) are from their MRCA AncLat (Fig. 3a). The larger divergence of the *P. mexicana* genomes than the *P. latipinna* genomes from their respective MRCAs could be an artefact of differing coalescent times of the assembled *P. latipinna* and *P. mexicana* individuals to the assembled *P. formosa* individual. However, this difference in branch lengths could also be a result of faster evolution of both *P. mexicana* genomes compared with both *P. latipinna* genomes, an especially intriguing possibility for the asexual *P. mexicana* and *P. latipinna* genomes, which exist as permanently linked haplotypes within each *P. formosa* individual. To test this possibility, we used genotypes from short-read samples of the three species to reconstruct both haplotypes of the MRCA of the 19 *P. formosa* short-read samples (Extended Data Fig. 4), and then computed divergence between both haplotypes of each modern *P. formosa* and the corresponding haplotype of the MRCA. We found that every one of the *P. formosa* individuals has a higher divergence to the MRCA in its *P. mexicana*-ancestry haplotype than in its *P. latipinna*-ancestry haplotype (Fig. 3b and Supplementary Note 3), suggesting faster divergence of the *P. mexicana*-ancestry haplotypes compared with the *P. latipinna*-ancestry haplotypes. Determining whether the sexual *P. mexicana* genome has also diverged faster from the ancestral state than the sexual *P. latipinna* genome, as the tree suggests, will require further data.

In general, the level of divergence of the sequence of a genome from an ancestral state is influenced by many factors, including time, drift (a function of population size), selection and background mutation rate. Because the two haplotypes of *P. formosa* have reproduced clonally and existed in the same fish since the MRCA of the sequenced individuals, the time to MRCA must be the same for both haplotypes, and their population sizes must have remained equal to each other in this time, so there should be no differences in time or drift between the haplotypes. We find no global signal of allele-biased expression (Supplementary Note 4 and Supplementary Data Table 5), making it unlikely that the haplotypes are under different selective pressure genome-wide. Therefore, the two haplotypes are most likely to have different background mutation rates (Supplementary Note 5).

To better understand the cause of the haplotype-specific mutation rates, we considered several possibilities (Supplementary Note 6).

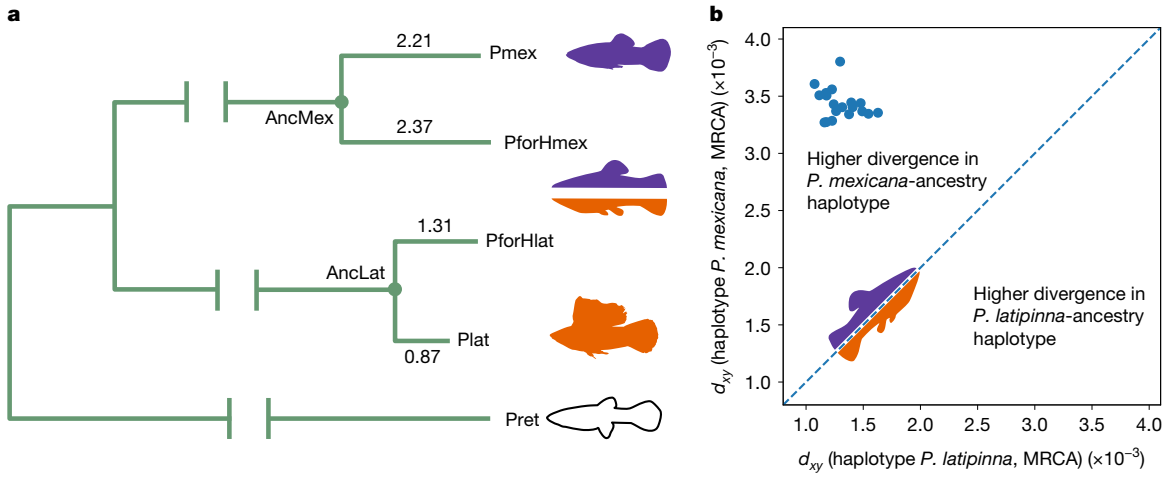


Fig. 3 | Asexual genomes have diverged more than their sexual counterparts, but haplotypes of *P. formosa* have diverged at different rates. a, A tree of the four assemblies plus *Poecilia reticulata* (Pret) as an outgroup. Numbers represent SNVs per kb between ancestral and modern genomes. **b**, Nucleotide divergence (d_{xy}) between each *P. formosa* sample and the MRCA of all *P. formosa* samples, broken down by haplotype. *P. mexicana*-ancestry haplotype of each individual has diverged from the MRCA more than its *P. latipinna*-ancestry haplotype. *P. formosa* and *P. latipinna* silhouettes by Kamil S. Jaron (CC0 1.0); *P. mexicana* silhouette by Michael Tobler (CC0 1.0); *P. reticulata* silhouette by Fidji Berio (CC0 1.0).

GC and microsatellite content are both known to affect mutation rate[37], but we found no major differences in either of these between the two haplotypes of *P. formosa*. Chromatin structure also influences mutation rate[38], but the boundaries of topologically associated domains (TADs) are highly conserved between *P. mexicana* and *P. latipinna*, and between the two haplotypes of *P. formosa*, although the TADs of *P. formosa* are smaller than those of its progenitor species (Extended Data Fig. 5). By contrast, local divergence rates are better conserved between sexual and asexual sister genomes than between the *P. formosa* haplotypes (Extended Data Fig. 6 and Supplementary Data Table 6). Moreover, at the base composition level, we find evidence of a *P. latipinna*-derived shift in mutational spectrum resulting in the mutational spectrum of the *P. mexicana*-origin haplotype of *P. formosa* most closely resembling that of free-living *P. mexicana*, but the *P. latipinna*-origin haplotype appearing intermediate between the *P. mexicana* haplotype and free-living *P. latipinna* (Extended Data Fig. 7 and Supplementary Note 6). Together, these results suggest that haplotype-specific mutation rates are likely to be a function of *cis*-acting differences in sequence context, which underlie 12% of the difference in mutational spectra between the parental species (Supplementary Note 6).

Rare crossing-over recombination

Returning to the question of why *P. formosa* shows signs of at most limited mutational decay despite faster divergence from the ancestral state than its sexual progenitor species, we next looked for evidence of crossing-over recombination in the Amazon molly. Trio-binning assembly is unsuited to detect crossover events between haplotypes, so we created ancestry-blind phased assemblies of *P. formosa* without referencing genomes from the parental species. We refer to these haploid assemblies as PforH1 and PforH2.

We found one location in each haplotype where a *P. formosa* contig switches from higher sequence identity with *P. latipinna* to higher sequence identity with *P. mexicana* or vice versa (Extended Data Fig. 8a,b). Long reads from the assembled *P. formosa* aligned to the assemblies span both breakpoints, confirming these locations as recombination breakpoints, robust against possible misassemblies or assembly phasing errors. Indeed, whole-genome alignment reveals that these two breakpoints are in homologous locations of their respective assemblies, so together, they represent a single crossing-over recombination between the two haplotypes of *P. formosa*. Of note, both of these breakpoints occur at the boundary of a top-level TAD (PforH1: 723 bp away; PforH2: 789 bp away), closer than expected by random chance ($P = 0.0108$, mean permutation distance to closest TAD boundary: 59,369 bp). A population-level analysis of the 19 *P. formosa* samples confirms the rarity of crossing-over (Supplementary Note 7), and shows that linkage disequilibrium remains elevated even between loci 1 Mb apart (Extended Data Fig. 8c).

Gene conversion slowed Muller's ratchet

Gene conversion represents another form of recombination that could counteract Muller's ratchet by eliminating deleterious mutations and increasing haplotype diversity. Gene conversion has previously been detected in Amazon molly[10], and has been hypothesized to slow Muller's ratchet in other asexual species[6,12,39–41], but no direct evidence for this hypothesis has been found so far in any species.

To determine the prevalence and diversity of gene conversion tracts in wild *P. formosa* populations, we used genotypes of the 19 *P. formosa*, 12 *P. latipinna* and 10 *P. mexicana* short-read individuals, looking for series of consecutive variant sites where there is a fixed difference between *P. latipinna* and *P. mexicana* but one or more *P. formosa* individuals is homozygous. We found that gene conversion tracts are widespread, representing on average 46.1 ± 7.14 Mb (mean ± s.d.; 6.26%) of the genome of each individual, and occur in the same locations more often than expected by random chance (permutation $P < 0.001$ of greater overlap; Fig. 4a and Extended Data Fig. 9a). In total, 20.4% of the genome is part of a gene conversion tract in at least one individual.

There is no bias towards gene conversion on one haplotype over the other (Fig. 4b). Hinting at a mechanism, we found that gene conversion breakpoints are enriched for proximity to polyA/T repeats (Extended Data Fig. 9b), which can cause replication fork collapse during the S phase of the cell cycle[42], potentially inducing double-strand breaks that result in gene conversion via homologous recombination[43]. These and other types of homopolymer repeats have been found near double-strand break-mediated gene conversion tracts involved in human disease[44]. This mode of gene conversion appears specific to *P. formosa* based on three observations about the breakpoints: first, they are enriched for fixed sequence differences between *P. formosa* and the parental species; second, they are associated with significant reductions, rather than increases, in population-scaled recombination rates in the parental species; and third, they show no signs of increased

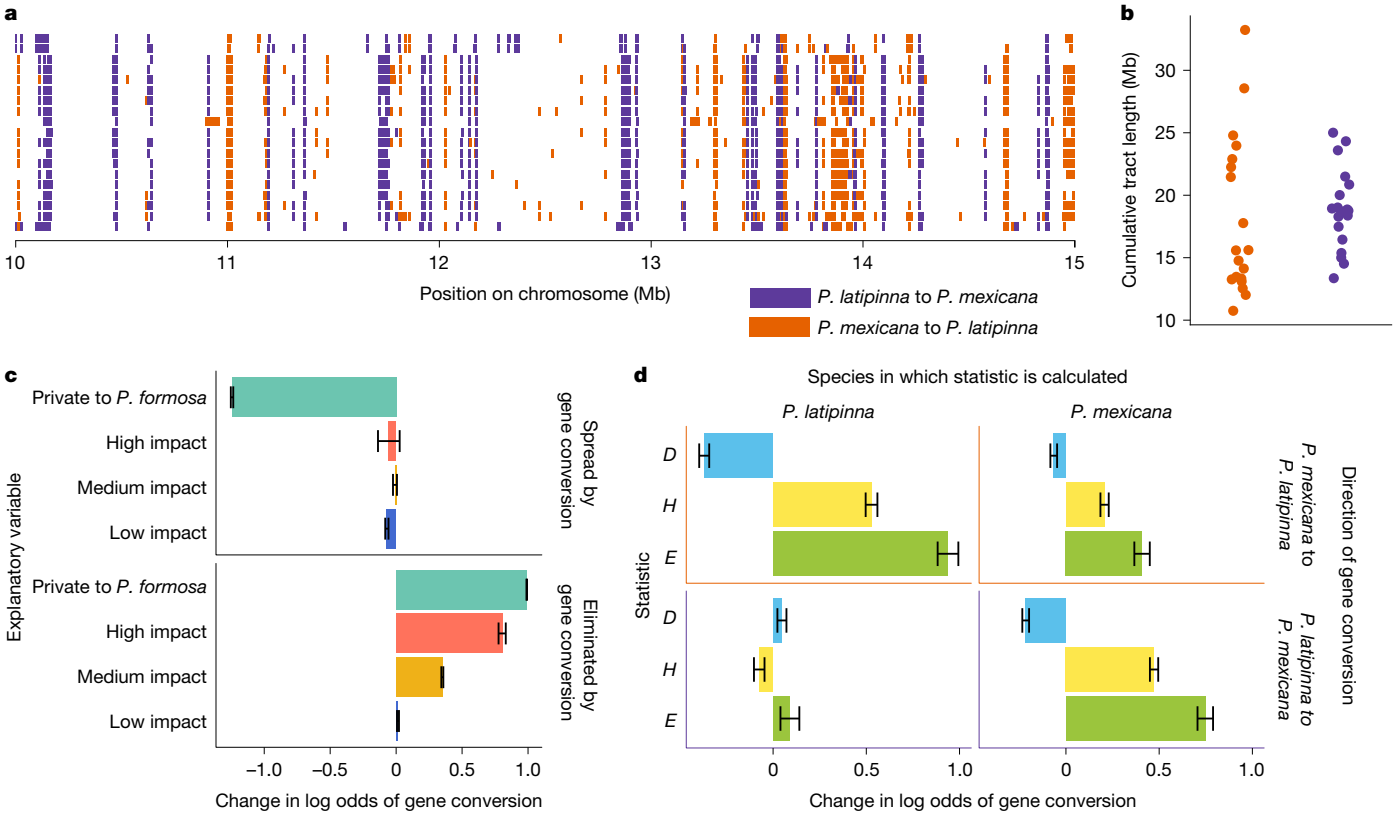


Fig. 4 | Gene conversion slows Muller's ratchet, facilitating both positive and negative selection. a, Gene conversion tracts in 19 *P. formosa* individuals on a segment of contig h2tg000018l, coloured by the direction of conversion. **b**, Cumulative tract lengths per individual of gene conversion tracts, separated by direction of conversion and not including fixed tracts. **c**, Gene conversion is more likely to purge young variants that affect coding sequence. Results from a binomial regression model used to predict the effect of variant impact and whether alleles are private to *P. formosa* (versus shared with *P. latipinna*, *P. mexicana*, or both) on the probability of the allele being spread or eliminated by gene conversion. $n = 19$ biologically independent *P. formosa* individuals, 10 *P. mexicana* individuals and 12 *P. latipinna* individuals. Error bars show 95% confidence intervals for the maximum likelihood estimate as estimated by the regression model and represented by centre bar height. **d**, Results from a binomial regression model used to predict the effect of three site frequency spectrum-based selection statistics calculated within the parental species on the probability of noncoding sequence being converted to one or the other haplotype in *P. formosa*. Negative Tajima's *D*, positive Fay and Wu's *H*, and Zeng's *E* all suggest positive selection[45]. Error bars show 95% confidence intervals for the maximum likelihood estimate as estimated by the regression model and represented by centre bar height.

gene conversion rates in *P. latipinna* and *P. mexicana* (Supplementary Note 8).

Gene conversion tracts are enriched in genic sequence (permutation test $P < 0.001$; observed overlap = 117 Mb; mean randomly sampled overlap = 89.9 Mb), suggesting that gene conversion may have an adaptive role in the genome, making it subject to selection. To investigate this hypothesis, we next looked for loci where gene conversion has acted on the site of a previous mutation that occurred more recently than the MRCA of the 19 sampled *P. formosa* individuals (Extended Data Fig. 9c). Out of 195,406 such gene conversion events, we find 10.6× more events (95% confidence interval: [10.5, 10.8]×) where gene conversion has reverted a derived allele to the ancestral allele (178,605) than where gene conversion has changed an ancestral allele to a derived allele (16,801), suggesting that the two directions of gene conversion have been subject to different selective pressures. Of these sites, 99.7% have no reads supporting the other genotype, consistent with these gene conversion events being heritable germline mutations rather than mosaic somatic mutations.

We next tested whether gene conversion enables purifying selection by considering, for coding sequence variants, the relationship between the predicted functional impact of each variant and its likelihood of being affected by gene conversion (Fig. 4c). Using a binomial regression model, we found that high-impact (frameshift and premature stop) mutations are the most likely to be eliminated by gene conversion (change in log odds ratio: 0.804; 95% confidence interval: [0.777, 0.831]) and unlikely to overwrite another allele (log odds ratio: −0.0553; 95% confidence interval: [−0.138, −0.0269]). Moderate-impact (missense and in-frame deletion) mutations are also more likely to be eliminated (log odds ratio: 0.350; 95% confidence interval: [0.343, 0.358]) than spread (log odds ratio: −0.00870; 95% confidence interval: [−0.0239, 0.00637]) by gene conversion. We also found that gene conversion preferentially fixes standing variation: alleles that are private to *P. formosa* and therefore probably arose after speciation are less likely to be spread (log odds ratio: −1.24; 95% confidence interval: [−1.25, −1.23]) and more likely to be eliminated (log odds ratio: 0.990; 95% confidence interval: [0.988, 0.993]) by gene conversion than mutations arising before speciation.

Turning to noncoding sequence, we also found evidence that gene conversion enables positive selection. For all non-exonic segments within gene conversion tracts, we computed three statistics that can detect selection by contrasting different parts of the site frequency spectrum (Tajima's *D*, Fay and Wu's *H* and Zeng's *E*)[45] in both *P. latipinna* and *P. mexicana*. These statistics quantify evidence that a sequence was subjected to positive selection in its species of origin, suggesting adaptive importance. We used each of these six statistics as explanatory variables in a binomial regression model and found that higher rates of gene conversion in *P. formosa* were associated with values indicative of positive selection (negative *D*, positive *H* and positive *E*) in the parental species whose haplotype was favoured by gene conversion (Fig. 4d).

Despite the enrichment of gene conversion tracts near genes, gene conversion does not facilitate adaptation predominantly through coding substitutions: gene conversion tracts are depleted for fixed coding differences between *P. latipinna* and *P. mexicana* (permutation $P < 0.001$) and for genes showing signs of positive selection between those two species (McDonald–Kreitman neutrality index <1, permutation $P < 0.001$). Nonetheless, coding substitutions within gene conversion tracts are enriched in genes involved in immune system processes, including formation of T cell receptor and immunoglobulin complexes (Supplementary Data Table 7). Components of these complexes undergo somatic recombination, in which high allelic diversity in coding sequence increases the repertoire of potential receptor molecules and thus the likelihood of successful response to pathogens[46]; diverse haplotypes of immune system genes in *P. formosa* were reported previously[10]. This illustrates the power of gene conversion to generate recombinant peptide sequences where they are most biologically useful.

In contrast to the depletion of coding substitutions in gene conversion tracts, we found evidence that gene conversion often modulates noncoding loci that regulate interconnected networks of genes involved in cell adhesion cell–cell signalling. We first identified all noncoding loci within gene conversion tracts, and then ranked the closest gene to each locus both by the number of *P. formosa* genomes with the gene conversion tract and by evidence that positive selection has affected the locus (Fay and Wu's *H* within the parental species of origin). Gene ontology (GO) enrichment analysis revealed similar enrichments for both sets, highlighting cell adhesion, cell–cell signalling and cell migration, among other processes (Supplementary Data Tables 8 and 9).

Notably, these enriched categories are also associated with genes that undergo random cell-by-cell monoallelic expression in humans[47] (hypergeometric $P = 4.86 \times 10^{-23}$, 95% confidence interval log odds ratio (LOR) 1.47–2.09 for genes ranked by gene conversion frequency; $P = 1.82 \times 10^{-17}$, LOR confidence interval 1.56–2.36 for genes ranked by by evidence for selection) and differ from categories of genes for which both alleles are stably expressed in humans (hypergeometric $P = 0.999$, LOR confidence interval: [−5.60, −0.0408] by frequency; $P = 0.985$, LOR confidence interval: [−4.94, 0.618], by selection). Genes subject to random monoallelic expression are thought to be fast-evolving, as both alleles can influence cellular phenotype, and thus be exposed to selection, without buffering one another[47]. Hypothesizing that the mutations that we found may represent a general solution favoured by natural selection to problems posed by hybrid ancestry, we also tested our enriched GO categories for overlap with those found in human genome regions devoid of admixture and incomplete lineage sorting with archaic hominins[48] and again found enrichment (hypergeometric $P = 3.01 \times 10^{-144}$, LOR confidence interval: [3.89, 4.52], by frequency; $P = 1.67 \times 10^{-149}$, LOR confidence interval: [5.50, 7.05], by selection). The set of genes near noncoding loci within high-frequency gene conversion tracts (in 18 out of 19 *P. formosa* genomes) is also enriched for protein–protein interactions ($P = 8.97 \times 10^{-5}$), suggesting that regulatory changes to these genes in the parental species may manifest as hybrid incompatibilities as a result of epistasis, in line with theory[49]. Indeed, a recent *Drosophila* study found that changes in expression level of cell adhesion genes contributed to hybrid male sterility[50]. Together, these findings suggest that gene conversion is capable of modulating regulatory sequence to reconcile incompatible mutations at fast-evolving loci, allowing natural selection to overcome genetic challenges posed by hybridization.

Discussion

Of the estimated several thousand clonally reproducing animals, only a small number have been studied to examine the impact of asexual reproduction on their genomes[20]. The Amazon molly, with its diploid genome and well-established parental species, presents an ideal system for comparing the side-by-side evolution of sexual and asexual genomes. As predicted by Meselson, we found that both haploid asexual genomes of the Amazon molly have diverged from the ancestral state more rapidly than their sexual sister genomes. Prior published examples of the Meselson effect are rare and have relied on inferring this effect based on heterozygosity, rather than directly comparing asexual haplotypes to sexually reproducing outgroups as we have done here[20,40,41].

Notably, contrary to theoretical predictions, the faster divergence along asexual lineages is not reflected in an excess of nonsynonymous mutations or a lack of purifying selection. Instead, we show that gene conversion, which has frequently been hypothesized to have a role in preventing mutational decay in clonally reproducing organisms[10,39,40], is a powerful mechanism counteracting the expected negative effects of asexual reproduction by facilitating both positive and negative selection as well as the resolution of hybrid incompatibilities in the ancestral haplotypes. It remains to be seen whether other long-extant asexual species escape Muller's ratchet through a similar mechanism, and whether haplotype-specific mutation rates are common in hybrid asexual genomes. Future functional genomic studies, especially at the single-cell level, will be instrumental for better characterizing the specific regulatory impacts of gene conversion and their contribution to organismal phenotype. This in turn could help illuminate adaptive traits that underlie the remarkable success and persistence of *P. formosa*.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10180-9.

1. Speijer, D., Lukeš, J. & Eliáš, M. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proc. Natl Acad. Sci. USA* **112**, 8827–8834 (2015).
2. Muller, H. J. Some genetic aspects of sex. *Am. Nat.* **66**, 118–138 (1932).
3. Muller, H. J. The relation of recombination to mutational advance. *Mutat. Res.* **106**, 2–9 (1964).
4. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737–756 (1974).
5. Quattro, J. M., Avise, J. C. & Vrijenhoek, R. C. An ancient clonal lineage in the fish genus *Poeciliopsis* (Atheriniformes: Poeciliidae). *Proc. Natl Acad. Sci. USA* **89**, 348–352 (1992).
6. Welch, M. D. & Meselson, M. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* **288**, 1211–1215 (2000).
7. Schwander, T., Henry, L. & Crespi, B. J. Molecular evidence for ancient asexuality in timema stick insects. *Curr. Biol.* **21**, 1129–1134 (2011).
8. Hubbs, C. L. & Hubbs, L. C. Apparent parthenogenesis in nature, in a form of fish of hybrid origin. *Science* **76**, 628–630 (1932).
9. Stöck, M., Lampert, K. P., Möller, D., Schlupp, I. & Schartl, M. Monophyletic origin of multiple clonal lineages in an asexual fish (*Poecilia formosa*). *Mol. Ecol.* **19**, 5204–5215 (2010).
10. Warren, W. C. et al. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. *Nat. Ecol. Evol.* **2**, 669–679 (2018).
11. Smith, J. M. *The Evolution of Sex* (Cambridge Univ. Press, 1978).
12. Birky, C. W. Jr. Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics* **144**, 427–437 (1996).
13. Judson, O. P. & Normark, B. B. Ancient asexual scandals. *Trends Ecol. Evol.* **11**, 41–46 (1996).
14. Lynch, M., Conery, J. & Burger, R. Mutation accumulation and the extinction of small populations. *Am. Nat.* **146**, 489–518 (1995).
15. Bell, G. *The Masterpiece of Nature: Evolution and Genetics of Sexuality* (Croom Helm, 1982).
16. Smith, J. M. in *Group Selection* (ed. Williams, G. C.) Ch. 9 (Routledge, 1971).
17. Williams, G. C. *Sex and Evolution* (Princeton Univ. Press, 1975).
18. Tree of Sex Consortium Tree of Sex: a database of sexual systems. *Sci. Data* **1**, 140015 (2014).
19. Lynch, M. & Gabriel, W. Mutation load and the survival of small populations. *Evolution* **44**, 1725–1737 (1990).
20. Jaron, K. S. et al. Genomic features of parthenogenetic animals. *J. Hered.* **112**, 19–33 (2021).
21. Dawley, R. M. & Bogart, J. P. *Evolution and Ecology of Unisexual Vertebrates* (New York State Museum, 1989).
22. Avise, J. C. *Clonality: The Genetics, Ecology, and Evolution of Sexual Abstinence in Vertebrate Animals* (Oxford Univ. Press, 2008).
23. Barley, A. J., Nieto-Montes de Oca, A., Manríquez-Morán, N. L. & Thomson, R. C. The evolutionary network of whiptail lizards reveals predictable outcomes of hybridization. *Science* **377**, 773–777 (2022).
24. Avise, J. C., Trexler, J. C., Travis, J. & Nelson, W. S. *Poecilia mexicana* is the recent female parent of the unisexual fish *P. formosa*. *Evolution* **45**, 1530–1533 (1991).

25. Schartl, M., Wilde, B., Schlupp, I. & Parzefall, J. Evolutionary origin of a parthenoform, the Amazon molly *Poecilia formosa*, on the basis of a molecular genealogy. *Evolution* **49**, 827 (1995).
26. Turner, B. J. The evolutionary genetics of a unisexual fish, *Poecilia formosa*. *Prog. Clin. Biol. Res.* **96**, 265–305 (1982).
27. Costa, G. C. & Schlupp, I. Placing the hybrid origin of the asexual Amazon molly (*Poecilia formosa*) based on historical climate data. *Biol. J. Linn. Soc. Lond.* **129**, 835–843 (2020).
28. Costa, G. C. & Schlupp, I. Biogeography of the Amazon molly: ecological niche and range limits of an asexual hybrid species. *Glob. Ecol. Biogeogr.* **19**, 442–451 (2010).
29. Dedukh, D. et al. Achiasmatic meiosis in the unisexual Amazon molly, *Poecilia formosa*. *Chromosome Res.* **30**, 443–457 (2022).
30. Kallman, K. D. Population genetics of the gynogenetic teleost, *Mollienesia formosa* (Girard). *Evolution* **16**, 497–504 (1962).
31. Turner, B. J., Elder, J. F. Jr, Laughlin, T. F. & Davis, W. P. Genetic variation in clonal vertebrates detected by simple-sequence DNA fingerprinting. *Proc. Natl Acad. Sci. USA* **87**, 5653–5657 (1990).
32. Schartl, M. et al. On the stability of dispensable constituents of the eukaryotic genome: stability of coding sequences versus truly hypervariable sequences in a clonal vertebrate, the amazon molly, *Poecilia formosa*. *Proc. Natl Acad. Sci. USA*. **88**, 8759–8763 (1991).
33. Loewe, L. & Lamatsch, D. K. Quantifying the threat of extinction from Muller's ratchet in the diploid Amazon molly (*Poecilia formosa*). *BMC Evol. Biol.* **8**, 88 (2008).
34. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
35. Rice, E. S. et al. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *GigaScience* **9**, giaa029 (2020).
36. Moran, N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **93**, 2873–2878 (1996).
37. Nesta, A. V., Tafur, D. & Beck, C. R. Hotspots of human mutation. *Trends Genet.* **37**, 717–729 (2021).
38. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local determinants of the mutational landscape of the human genome. *Cell* **177**, 101–114 (2019).
39. Omilian, A. R., Cristescu, M. E. A., Dudycha, J. L. & Lynch, M. Ameiotic recombination in asexual lineages of *Daphnia*. *Proc. Natl Acad. Sci. USA* **103**, 18638–18643 (2006).
40. Weir, W. et al. Population genomics reveals the origin and asexual evolution of human infective trypanosomes. *eLife* **5**, e11473 (2016).
41. Brandt, A. et al. Haplotype divergence supports long-term asexuality in the oribatid mite Oppiella nova. *Proc. Natl Acad. Sci. USA* **118**, e2101485118 (2021).
42. Tubbs, A. et al. Dual roles of poly(dA:dT) tracts in replication initiation and fork collapse. *Cell* **174**, 1127–1142.e19 (2018).
43. Stewart, J. A. et al. Noncanonical outcomes of break-induced replication produce complex, extremely long-tract gene conversion events in yeast. *G3* **11**, jkab245 (2021).
44. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8**, 762–775 (2007).
45. Zeng, K., Fu, Y.-X., Shi, S. & Wu, C.-I. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**, 1431–1439 (2006).
46. Mikocziova, I., Greiff, V. & Sollid, L. M. Immunoglobulin germline gene variation and its impact on human disease. *Genes Immun.* **22**, 205–217 (2021).
47. Kravitz, S. N. et al. Random allelic expression in the adult human body. *Cell Rep.* **42**, 111945 (2023).
48. Schaefer, N. K., Shapiro, B. & Green, R. E. An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Sci. Adv.* **7**, eabc0776 (2021).
49. Johnson, N. A. & Porter, A. H. Rapid speciation via parallel, directional selection on regulatory genetic pathways. *J. Theor. Biol.* **205**, 527–542 (2000).
50. Go, A., Alhazmi, D. & Civetta, A. Altered expression of cell adhesion genes and hybrid male sterility between subspecies of *Drosophila pseudoobscura*. *Genome* **62**, 657–663 (2019).

Methods

Animals

The *P. formosa* individuals used in this study were raised at the fish facilities of the Biocenter of the University of Würzburg following approved experimental protocols through an authorization (568/300-1870/13) of the Veterinary Office of the District Government of Lower Franconia, Germany, in accordance with the German Animal Protection Law (TierSchG). The *P. formosa* DNA for whole-genome sequencing was derived from a single adult female from a clonal subline (WLC 7122, *P. formosa* + M) established from fish collected in 1953 by C. P. Haskins near Brownsville, TX, USA. A sibling of this fish was used for Hi-C library preparation.

The *P. mexicana* (Rio Purification bei Nueva Padilla, Tamaulipas, Mexico) and *P. latipinna* (Laguna de Champayan, Tamaulipas, Mexico) individuals used in this study were supplied by the International Stock Center for Livebearing Fishes, University of Oklahoma, for sequencing.

Sample processing

To prepare samples of *P. formosa*, *P. latipinna* and *P. mexicana* for long-read sequencing, we first flash-froze them in liquid nitrogen and stored at −80 °C until ready for isolation, using muscle for *P. formosa*, soft organs (brain, eyes, gills, liver, kidney and spleen) and carcass for *P. mexicana*, and soft organs (brain, eyes, gills, liver, kidney and spleen) for *P. latipinna*. High molecular weight (HMW) DNA extraction for all three fish was performed using the 10x Genomics Demonstrated Protocol: DNA Extraction from Single Insects (10x Genomics) and then purified and concentrated through standard ethanol precipitation. The final DNA quantity was determined using the high-sensitivity Invitrogen Qubit Fluorometer kit (Invitrogen). Final DNA quality was evaluated on a 0.7% agarose gel and visualized using the Uvitec Cambridge Uvidoc HD6 UV Fluorescence and Colorimetry instrument. All isolated DNA was shipped to HudsonAlpha for library construction and long-read sequencing on a PacBio Sequel II in CCS/HiFi mode.

We prepared Hi-C libraries using the Phase Genomics Proximo Hi-C Animal Kit v.2 (Phase Genomics). For *P. mexicana* and *P. latipinna*, we minced and pooled brain and muscle tissues using approximately 50 mg for *P. mexicana* and 120 mg of *P. latipinna*. For *P. formosa*, we used 217 mg of muscle tissue. We then purified the resulting DNA and generated libraries, targeting an insert size of 500–600 bp. We performed fragmentation using the Applied Biosystems Veriti 96-Well Thermal Cycler (Thermo Fisher Scientific) and assessed DNA quality using a fragment analyser. We measured DNA concentration using the high-sensitivity Invitrogen Qubit Fluorometer Kit (Invitrogen). We sequenced the Hi-C libraries on an Illumina NovaSeq 6000 with a target of >150 million reads for each sample.

Assembly

We assembled the initial contigs for *P. latipinna* (Plat) and *P. mexicana* (Pmex) using hifiasm[51] v.0.16.1-r375 with default settings.

For the trio-binning assembly of *P. formosa* (haplotypes PforHlat and PforHmex), we first separated the *P. formosa* HiFi reads into parental bins based on *P. latipinna* and *P. mexicana* short reads from a previous study[10] as previously described for F_1 hybrids[35] using a custom program[52], which internally uses kmc[53] v.3.2.1 for computing sets of *k*-mers and comparing these sets. We used $k = 21$. We then assembled the reads in each bin separately using hifiasm[51] v.0.16.1-r375 with the parameter -l0 to prevent the assembler from treating the assembly as a diploid.

For the ancestry-blind assembly of *P. formosa* (haplotypes PforH1 and PforH2), we assembled all HiFi reads using hifiasm in Hi-C-assisted phasing mode and Hi-C reads from *P. formosa*, with the parameter -hg-size 750 m to help the assembler find the correct coverage peaks due to the genome's high heterozygosity. For downstream analyses in which we binned these contigs by ancestry, we first aligned all hap1 contigs to all hap2 contigs and vice versa using minimap2[54] in asm10 mode. We then created a bipartite graph with all H1 contigs on one side, and all H2 contigs on the other side, and for each contig, drew an edge to the contig in the opposite haplotype with the most aligned bases. Finally, we partitioned the graph into connected components, summed the number of bases in contigs from each haplotype aligned to the *P. latipinna* or *P. mexicana* assembly, and assigned all H1 contigs in the connected component to either the *P. latipinna*-derived haplotype or the *P. mexicana*-derived haplotype, and all H2 contigs to the other haplotype, based on which configuration had a higher concordance using total aligned bases.

We scaffolded the contigs of each assembly using a custom nextflow[55] pipeline[56]. In brief, the pipeline uses chromap[57] v.0.2.3-r407 with the hic preset to align the Hi-C reads to the contigs and YAHS[58] v.1.2a.1 with default options to assemble the contigs into scaffolds based on these alignments. We used Juicebox Assembly Tools[59] to manually curate the assemblies and assigned chromosome names based on synteny with *Xiphophorus maculatus*[60].

Repeat analysis

We annotated repeats in all assemblies using the TETools[61] docker image v.1.7, running BuildDatabase, RepeatModeler and then RepeatMasker with default options on each assembled genome.

Annotation

We annotated the assemblies using a pipeline adapted from a previous study[62]. This pipeline aligns known protein sequences and RNA sequencing (RNA-seq) reads to the assembly to collect homology and transcriptome gene evidence. It also runs AUGUSTUS[63] for ab initio gene prediction. In the end all results are compared and synthesized into a final gene set.

For homology alignment, we collected 505,310 protein sequences from the vertebrate database of Swiss-Prot[64], RefSeq database (proteins with ID starting with 'NP' from 'vertebrate_other') and the NCBI genome annotation of human (GCF_000001405.39_GRCh38), zebrafish (GCF_000002035.6), platyfish (GCF_002775205.1), medaka (GCF_002234675.1), mummichog (GCF_011125445.2), turquoise killifish (GCF_001465895.1), guppy (GCF_000633615.1) and Shortfin molly (GCF_001443325.1). These sequences were aligned to the genome assemblies using GeneWise[65] v.2-4-1 and Exonerate[66] v.2.2.0. GenblastA[67] v.1.0.1 was used to locate rough alignment regions for GeneWise. In each result, when multiple gene models overlapped in one genome region, the one with the highest score was kept.

To collect transcriptome gene evidence, RNA-seq reads were collected from gonad, brain, gill, liver or mixed-organs (SRR13349691, SRR13349691, SRR13349692, SRR13349692, SRR3171725, SRR3171768, SRR3171771, SRR5513066, SRR5513066, SRR5224069, SRR5224069, SRR5224074, SRR5224074, SRR5224079, SRR5224079, SRR7638274 and SRR7638274). We aligned the reads to the assemblies using HISAT[68] v.2.2.1. Gene locations and structures were then phased using StringTie[69] v.2.2.1. In parallel, transcripts were assembled using Trinity[70] v.2.9.1 and aligned to the assembly using splign[71] v.2.1.0 for gene locations and structure phasing.

AUGUSTUS v.3.5.0 was used for ab initio gene prediction after two rounds of training. The first round was trained by BUSCO[72] v.5.3.2 with the parameter -long. The second was done by using high-quality gene models that were commonly agreed upon by Exonerate, Genewise, StraingTie and splign. AUGUSTUS was then run using all the homology and transcriptome gene evidence as hints.

To synthesize all the gene evidence collected above, we screened through each gene locus to compare the homology and ab initio gene evidence. When multiple gene models competed for one gene locus, the one best supported by transcriptome evidence was kept. If the kept gene model possessed poorly supported exons (with low homology identity or no transcriptome support), we screened the eliminated gene models for better supported counterparts and replaced them.

Finally, genes with no transcriptome support and low/no homologous identity were removed.

Genotyping

We genotyped short-read samples twice: once using AncMol as a reference, and once using PforH2 as a reference. We used AncMol-based calls for downstream analyses subject to reference bias, and PforH2-based calls for downstream analyses that required an annotated reference.

We first aligned short-read samples from previous studies[10,73–75], including 19 *P. formosa*, 12 *P. latipinna* and 10 *P. mexicana*, using minimap2[54] v.2.28-r1209 with the sr preset. We then genotyped each sample separately using elprep[76] v.5.1.3, a multithreaded rewrite of GATK that produces the exact same output given the same input, with subcommand sfm and parameters -mark-duplicates -mark-optical-duplicates -sorting-order coordinate -haplotypecaller. We next combined the gVCFs with GATK[77] v.4.5.0.0 subcommand CombineGVCFs and performed joint genotyping with GATK subcommand GenotypeGVCFs, both with default options. Finally, we filtered variant calls down to biallelic sites with genotype quality of at least 10. Code for the custom pipeline we used for this is available with the other code for this project.

Phasing

We first performed read-backed phasing on each individual *P. formosa* using WhatsHap[78] v.2.3 with the phase command and default options. Then, for each phasing set output by WhatsHap, we used the following heuristic to determine which haplotype was derived from *P. latipinna* and which from *P. formosa*: a phasing set must have at least two variants supporting one phasing polarity and none supporting the opposite phasing polarity, where a variant is considered to support a given phasing polarity if all *P. mexicana* calls are homozygous for one allele, all *P. latipinna* calls are homozygous for the other allele, and there are at least two haplotypes from each species with confident calls. For example, for a phased *P. formosa* variant call 0|1, this variant would support the first haplotype being derived from *P. latipinna* and the second haplotype being derived from *P. mexicana* if and only if all *P. latipinna* calls are 0/0 and all *P. mexicana* calls are 1/1. We removed phasing information from all variants in any phasing set with fewer than two variants supporting either polarity, or with one or more variants supporting both polarities. The program written for this purpose is available with other project code.

Assembly-based ancestral reconstruction

To reconstruct genomes at internal nodes of the tree of assembled genomes, we performed ancestral reconstruction using progressive cactus[79] v.2.4.2 with default options and the tree (((Pmex:0.1, PforHmex:0.1) AncMex:2.6, (Plat:0.1, PforHlat:0.1) AncLat:2.6) AncMol:3.3, Pret:6.0);. All of the leaves of this tree are assemblies generated for this study except for Pret, which is a previously published genome of the guppy *P. reticulata*[80] used as an outgroup.

Muller's ratchet

To make genes comparable across annotations, and to map genes from annotations produced for this study to those with curated functional annotations, we downloaded the guppy (*Poecilia reticulata*) transcriptome[81] and aligned all transcript sequences to guppy transcript sequences using the blastn program from NCBI BLAST[82] v.2.3.0. We determined a transcript pair to be homologous if each was the reciprocal highest-scoring hit for the other. We then obtained gene and protein names as well as GO annotations[83] for the guppy transcripts from BioMart[84].

For each pair of transcriptomes compared, we loaded all coding sequences along with their translated protein sequences and adjusted the reading frame of the coding sequence to match the protein sequence where necessary. We then aligned all pairs of homologous amino acid sequences from the annotations being compared using the Biopython v.1.85 pairwise2 module and a BLOSUM62 matrix, deleting codons that aligned to gaps in the other sequence and removing stop codons from the ends of sequences. We then used PAML[85] v.4.10.9 to calculate dN/dS using the Yang and Nielsen 2000 (yn00) algorithm[86].

We used our polymorphism data to calculate other selection-relevant statistics. Using the program ann_codon_pos from the hts_popgen package (https://github.com/nkschaefer/hts_popgen), we inserted codon position information into our variant call set as a VCF tag. Using third codon positions to approximate synonymous sites and using first and second codon positions to approximate nonsynonymous sites, we then compiled McDonald–Kreitman test relevant statistics[87] for each gene and for each species, using the program mk from the hts_popgen package. We used the within-population values computed for pN and pS, and we normalized pN by the number of nonsynonymous sites per gene (two-thirds the number of codons) and normalized pS by the number of synonymous sites per gene (one-third the number of codons).

We computed bootstrapped 95% confidence intervals for differences between median dN/dS and pN/pS for each comparison by resampling genes with replacement 10,000 times.

For the variant impact analysis, we annotated the functional impact of each variant using the variants called against the formosa_hap2_broken.fa genome and corresponding genome annotation using SnpEff version 5.2e[88]. We discarded genes that SnpEff rejected, including those with premature stop codons in the reference annotation, resulting in 14,467 genes for which mutational impact could be predicted. For each *P. latipinna*, *P. mexicana* and *P. formosa* individual, we then counted the number of heterozygous and homozygous-alternate genotypes for variants with functional annotations, counting mutations with each predicted functional impact severity, choosing the highest severity wherever a variant was predicted to affect multiple genes in different ways.

Tree-building

To build the tree of genome assemblies based on ancestral reconstruction, we started with the tree topology input to cactus and calculated branch lengths by running the halBranchMutations command of the hal toolkit[89] v.2.2 on the hal file output by cactus, dividing the total number of SNVs output by the total length of the alignment between each node and its parent to obtain a number of SNVs per kilobase and excluding inferred admixed regions. Total alignment length between each parent–child pair was calculated using halAlignmentDepth.

To build the tree based on pairwise mash distances, we first calculated a distance matrix of mash distances between every pair of leaves in the tree using the triangle command from the mash[90] v.2.3 package with default options and computed a neighbour-joining tree based on this distance matrix using BioPython's Phylo package[91] v.1.85.

Chromatin structure

We aligned Hi-C reads to final assemblies using chromap[57] v.0.2.3-r407 with the hic preset. For *P. formosa*, we concatenated the H1 and H2 ancestry-blind assemblies and aligned reads to the concatenated reference. We then converted the pairs output of chromap to.hic format using the pre command of hic_tools[92] v.3.30 with default options. TADs were identified using DeDoc2[93] on KR-normalized matrices of 10 kb resolution (sliding window = 10 Mb). Insulation scores were calculated with FAN-C[94] v.0.9.28, also on 10 kb resolution matrices, using different window sizes. Insulation score heatmaps (window size = 100 kb) were plotted using custom R scripts compiled in the sorolla package (https://gitlab.com/rdacemel/sorolla). For this task, sorolla wraps functions from EnrichedHeatmap[95] v.1.34.0 package.

To convert TAD boundaries between coordinate systems, we wrote a custom python script (available in project code repository) as a wrapper for halLiftover[89] as distributed with the cactus v.2.4.2 Docker image.

Recombination

We first aligned HiFi reads from *P. formosa*, *P. mexicana* and *P. latipinna* to PforH1 and PforH2 using minimap2[54] v.2.27-r1193 with the map-hifi preset. We then ran samtools[96] v.1.20 subcommand mpileup with options -Q20 -q20 -B. We performed a simple genotype estimation of the output by classifying each position as either hom-ref (>80% of aligned reads matching reference), hom-alt (>80% of aligned reads matching same alternate allele), het (not hom-ref or hom-alt, and the most frequent two bases supported by >90% of reads together), or uncallable (fewer than 10 aligned bases or not matching any of the other three categories). We then calculated a mismatch rate, defined as (count(hom-alt) + 0.5 × count(het))/(called bases), for each window with at least 70% of its positions called. We plotted these window mismatch rates to find places where the *P. mexicana* and *P. latipinna* rates cross each other, and then examined the long-read alignments of all three species to these potential breakpoints using IGV[97] v.2.17.4. We consider a locus to confidently represent a recombination breakpoint if there are multiple *P. formosa* long reads spanning the breakpoint that are more similar to *P. latipinna* on one side of the breakpoint but more similar to *P. mexicana* on the other side of the breakpoint, with at least two reads showing evidence for the switch in each of the two directions (that is, at least two reads similar to *P. latipinna* before the breakpoint and *P. mexicana* after, and at least two reads similar to *P. mexicana* before the breakpoint and *P. latipinna* after).

To examine patterns of linkage disequilibrium at different distances, we randomly sampled from the set of all pairs of biallelic SNVs in genotype calls of *P. formosa* individuals on chromosome AncMolrefChr0. For each sampled pair of SNVs with a minor allele frequency of at least 5% and at least 5 individuals with confident calls at both sites, we calculated r^2 using the Rogers–Huff estimator[98]. We binned these r^2 estimates into distance bins, and then calculated the mean for each bin, as well as a 95% confidence interval by bootstrapping with 1,000 samplings. We performed this analysis at two different resolutions: 1 kb bins for biallelic pairs from 0–100 kb apart with 5% of all pairs sampled, and 10 kb bins from 0–1 Mb apart with 1% of all pairs sampled. The code used to perform this analysis is in the project code repository.

Gene conversion detection

For each *P. formosa* individual, we searched the unphased genotype calls for biallelic SNV sites with genotypes characteristic of gene conversion, where at least one individual from every species has a confident genotype call, all *P. latipinna* are homozygous for one allele, all *P. mexicana* are homozygous for a different allele, and one or more *P. formosa* are homozygous for either the *P. mexicana* allele or the *P. latipinna* allele. We then defined gene conversion tracts as any run of three or more consecutive gene conversion-characteristic variant sites at least 100 bp in length, with all sites supporting gene conversion in the same direction.

To determine the presence of repetitive sequence at gene conversion breakpoints, we created a BED file on formosa_hap2_broken coordinates representing the 50 bp interval around each true gene conversion breakpoint, and then randomly sampled matched intervals using bedtools v.2.30.0 shuffle[99]. We then extracted the sequence corresponding to each interval using bedtools getfasta and wrote a Python program to count occurrences of each type of mono- or dinucleotide repeat in both true and randomly sampled sequences. We treated each sequence the same as its reverse complement, and we also collapsed similar dinucleotide repeats (for example, AC and CA repeats were treated as equivalent).

To find sites that have been affected both by mutation and gene conversion since the MRCA of the 19 *P. formosa* samples, we first built a tree of short-read samples to compare local topologies to the genome-wide topology. To this end, we made a distance matrix of the AncMol-based genotypes using plink[100] v.1.90 with parameters -distance square -maf 0.05 -geno 0.15 -indep-pairwise 50 10 0.1 and then used BioPython's Phylo package[91] to generate a neighbour-joining tree rooted at the midpoint of the longest branch. Owing to the clonal reproduction of *P. formosa*, the tree topology should be consistent across sites in the absence of gene conversion, so any sites deviating from the genome-wide topology represented by the gene conversion tree may be indicative of gene conversion.

Next, at every site in the genome containing an SNV that is biallelic within *P. formosa* and confidently genotyped in at least five individuals, we used Fitch's algorithm[101] to find the most parsimonious set of character changes within the tree consistent with the leaf genotypes. Any path in the resulting annotated tree from the MRCA to a leaf on which a change from the ancestral allele to a derived allele is followed by a change from heterozygosity to homozygosity indicates a site where mutation was followed by gene conversion. For example, in the hypothetical tree (A,(B,(C,D))), the minimum number of state changes consistent with the leaf genotypes A:0/0, B:0/1, C:0/1, D:0/0 is two: (1) a change from 0/0 to 0/1 between the root and the ancestor of B, C, and D, and (2) a change from 0/1 to 0/0 between the ancestor of C and D and the leaf node D. Thus, the full path from the root to D requires a change from 0/0 to 0/1 followed by a change from 0/1 to 0/0, which can be best explained by mutation followed by gene conversion; the only other explanation is back mutation.

We counted sites where gene conversion reverts a mutation to the ancestral allele (for example, 0/0 → 0/1 → 0/0) and sites where gene conversion overwrites the ancestral allele with a derived allele (for example, 0/0 → 0/1 → 1/1). We calculated a 95% confidence interval for the ratio of these two kinds of sites using bootstrapping with 1,000 samples.

An implementation of this algorithm in rust is located in the project code repository.

Gene conversion recurrence

We tested how likely gene conversion tracts were to recur at similar loci using resampling. We first used the bedtools v.2.30.0 multiinter command[99] to find genomic regions where all *P. formosa* genomes were converted to the *P. latipinna* haplotype, or where all were converted to the *P. mexicana* haplotype; these fixed tracts were stored in a BED file to exclude from analyses, as they could also include possible admixed loci in the progenitors of *P. formosa*. Next, we merged all gene conversion tracts (using bedtools merge) and subtracted fixed tracts (using bedtools subtract), and summed the total number of bases. Then, for 1,000 random trials, we re-sampled gene conversion tracts (using bedtools shuffle, excluding sites of fixed tracts) and similarly counted the number of bases in each.

We used a similar technique to assess overlap of gene conversion tracts with genic sequences: we computed the number of bases overlapping between gene conversion tracts, merged across all individuals with fixed tract loci subtracted, and then re-sampled 1,000 times using bedtools shuffle and excluding sites of fixed gene conversion tracts. We report a *P* value describing the number of trials in which the overlap in any trial exceeded observed overlap, divided by the number of trials.

Gene conversion selection

We used a generalized linear model approach to study the connection between variant impact and probability of gene conversion. Considering only non-fixed gene conversion tracts spanning at least two SNPs, we produced a file identifying all sites at which each *P. formosa* individual has undergone gene conversion. We then visited all functional impact-annotated variants on the collapsed *P. formosa* genome (formosa_hap2_broken.fa) that were polymorphic within *P. formosa*. For each such variant, we noted the functional impact, the total number of called *P. formosa* genotypes, the number of *P. formosa* individuals with gene conversion at the SNP and homozygous for the reference allele (for which the annotated allele was lost through gene conversion), the number of *P. formosa* individuals with gene conversion at the

SNP and homozygous for the alternate allele (for which the annotated allele was gained through gene conversion), and whether or not the alternate allele was private to *P. formosa* (as opposed to also present in *P. latipinna* and/or *P. mexicana* genomes).

We then created indicator variables for the three highest levels of variant impact (low, moderate, and high), using the MODIFIER category as the reference level. We also created an indicator variable for *P. formosa*-private alleles, using alleles shared with *P. latipinna* and/or *P. mexicana* as the reference level. We fit a binomial model to the probability of gain and the probability of loss using the glm function in R using glm(cbind(gain, not_gain) ~ low + medium + high + private, family = binomial(link = 'logit')) and glm(cbind(loss, not_loss) ~ low + medium + high + private, family = binomial(link = 'logit')) and computed the effect of each predictor variable on the log odds ratio of gene conversion-mediated gain or loss and the 95% confidence interval from the results.

We then extended this approach to study the relationship between evidence that noncoding sequences were subject to positive selection in their species of origin and the probability of gene conversion. We obtained a set of noncoding sequences subject to high-confidence gene conversion by subtracting exonic sequences from our set of non-fixed gene conversion tracts spanning at least two SNPs, using bedtools[99] v.2.30.0 commands subtract, sort, and merge. For each region, we then computed Tajima's *D*, Fay and Wu's *H*, and Zeng's *E* (ref. 45) using the sfs program in hts_popgen (https://github.com/nkschaefer/hts_popgen), using the two *P. reticulata* genomes as an outgroup. We then merged the resulting BED file with our consensus *P. latipinna* and *P. mexicana* gene conversion intervals (produced using bedtools multiinter) to obtain the direction of gene conversion and the number of *P. formosa* individuals with the gene conversion tract, for each interval.

To obtain a comparably-sized set of noncoding regions not subject to gene conversion, we sampled a set of regions of the same size and number from non-exonic sequence using bedtools shuffle (with -excl set to all exons and -f 0.001) and ran the sfs program from hts_popgen on these as well. We kept the original direction of gene conversion (toward the *P. latipinna* or *P. mexicana* haplotype) for each of these intervals but set the number of *P. formosa* individuals undergoing gene conversion to 0 for each.

Treating segments converted to the *P. latipinna* and *P. mexicana* haplotypes separately, we then fit a binomial generalized linear model to our data in R using the command: glm(cbind(freq, failures) ~ lat.D + mex.D + lat.H + mex.H + lat.E + mex.E, family = binomial(link = 'logit')), where freq was the number of *P. formosa* individuals with gene conversion of the given type at a locus, lat.D, lat.H, and lat.E are Tajima's *D*, Fay and Wu's *H* and Zeng's *E* in *P. latipinna* individuals at the locus, and mex.D, mex.H, and mex.E are Tajima's *D*, Fay and Wu's *H* and Zeng's *E* in *P. mexicana* individuals at the locus. We extracted coefficients and standard errors using the summary() command.

Gene ontology enrichment testing

We used the R package GOfuncR v.1.22.0, which wraps FUNC[102], for GO enrichment testing. We obtained *P. reticulata* gene-GO term mappings from Biomart and used the default GO hierarchy included with GOfuncR, building a new OrgDb object for the *P. reticulata* annotation. To obtain a set of genes in gene conversion tracts impacted by coding substitutions, we created a BED file listing positions of variants in *P. formosa* that fall within first or second codon positions (as annotated using the ann_codon_pos program in the hts_popgen package), noted the affected genes, intersected with gene conversion tracts (using bedtools v.2.30.0 intersect), and converted gene IDs to guppy gene IDs (as determined through reciprocal blastn; see 'Muller's ratchet'). We then tested for GO enrichment using the go_enrich command with test = 'hyper'. To test for GO enrichment near noncoding sequences subject to positive selection in their species of origin, we took all noncoding sequences annotated with site frequency spectrum-based measures of selection (see 'Gene conversion selection' section), identified the closest gene to each tract using bedtools closest, mapped each gene to the highest Fay & Wu's *H* value reported in any nearby tract, converted gene IDs to guppy gene IDs, and used the go_enrich command with test='wilcoxon'). We also repeated this test, but ranked genes by the frequency of enclosed gene conversion tract (determined by running bedtools multiinter on all *P. latipinna*-converted and *P. mexicana*-converted tracts separately, taking the fourth column as frequency) and choosing the maximum frequency per gene. In all GO enrichment tests, we applied a cut-off of 0.05 for the family wise error rate of overrepresentation. Protein–protein interaction (PPI) enrichment testing was done using the web-based UI for the STRING database[103].

To compare enriched GO terms with those reported to correspond to genes undergoing random monoallelic expression, we obtained lists of enriched GO terms for genes that undergo random monoallelic expression in humans, and genes that undergo stable biallelic expression in humans, from a prior study[47]. We then obtained a set of genomic regions in which a panel of modern human genomes was free from alleles introgressed through admixture from, or on lineages incompletely sorted with, archaic hominins, from another prior study[48]. Reasoning that gene regulation probably diverged considerably between humans and *Poecilia*, but that patterns at the gene network level were more likely to hold, we chose to compare lists of enriched GO terms for overlap rather than examine individual genes.

We compared GO term lists using the hypergeometric CDF (phyper in R), using as background 8,024 GO terms that were mapped to *P. reticulata* genes with a clear 1:1 homologue in the *P. formosa* annotation we used, and which also were present in the human GO annotation (September 2022 release). We limited all GO lists to only terms present in all annotations.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All sequence data generated for this project, as well as the assembled genomes, are deposited in the relevant NCBI databases under the following BioProject IDs: PRJNA1398340 (*P. latipinna*), PRJNA1398337 (*P. mexicana*), and PRJNA1398335 and PRJNA1398336 (*P. formosa*). Hi-C reads for *P. latipinna* and *P. mexicana* are available under BioProject ID PRJNA614959. Previously published accessions from the NCBI Sequence Read Archive are listed in Supplementary Table 2.

Code availability

Code used to perform the analyses described in this manuscript is open-source licensed under the GNU General Public License and publicly available on GitHub (https://github.com/esrice/amazon-molly-paper) and at https://doi.org/10.5281/zenodo.17976427 (ref. 104).

51. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
52. Rice, E. S. Trio_binning: programs implementing the trio-binning genome assembly method. *Github* https://github.com/esrice/trio_binning (2022).
53. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
54. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
55. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
56. Rice, E. S. hic-scaffolding-nf: nextflow pipeline for scaffolding genome assemblies with Hi-C reads. *Github* https://github.com/WarrenLab/hic-scaffolding-nf (2022).
57. Zhang, H. et al. Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat. Commun.* **12**, 6566 (2021).
58. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).

59. Dudchenko, O. et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Preprint at *bioRxiv* https://doi.org/10.1101/254797 (2018).
60. Lu, Y. et al. High resolution genomes of multiple Xiphophorus species provide new insights into microevolution, hybrid incompatibility, and epistasis. *Genome Res.* **33**, 557–571 (2023).
61. Dfam Consortium. TETools: Dfam transposable element tools docker container. *Github* https://github.com/Dfam-consortium/TETools (2022).
62. Du, K. et al. Phylogenomic analyses of all species of swordtail fishes (genus *Xiphophorus*) show that hybridization preceded speciation. *Nat. Commun.* **15**, 6609 (2024).
63. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–9 (2006).
64. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **406**, 89–112 (2007).
65. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
66. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
67. She, R., Chu, J. S.-C., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
68. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
69. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
70. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
71. Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* **3**, 20 (2008).
72. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
73. Darolti, I. et al. Extreme heterogeneity in sex chromosome differentiation and dosage compensation in livebearers. *Proc. Natl Acad. Sci. USA* **116**, 19031–19036 (2019).
74. Greenway, R. et al. Convergent evolution of conserved mitochondrial pathways underlies repeated adaptation to extreme environments. *Proc. Natl Acad. Sci. USA* **117**, 16424–16430 (2020).
75. De-Kayne, R. et al. Evolutionary rate shifts in coding and regulatory regions underpin repeated adaptation to sulfidic streams in poeciliid fishes. *Genome Biol. Evol.* **16**, evae087 (2024).
76. Herzeel, C. et al. Multithreaded variant calling in elPrep 5. *PLoS ONE* **16**, e0244471 (2021).
77. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O'Reilly Media, 2020).
78. Martin, M., Ebert, P. & Marschall, T. Read-based phasing and analysis of phased variants with WhatsHap. *Methods Mol. Biol.* **2590**, 127–138 (2023).
79. Armstrong, J. et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
80. Fraser, B. A. et al. Improved reference genome uncovers novel sex-linked regions in the guppy (*Poecilia reticulata*). *Genome Biol. Evol.* **12**, 1789–1805 (2020).
81. Fraser, B. A., Künstner, A., Reznick, D. N., Dreyer, C. & Weigel, D. Population genomics of natural and experimental populations of guppies (*Poecilia reticulata*). *Mol. Ecol.* **24**, 389–408 (2015).
82. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
83. The Gene Ontology Consortium The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
84. Smedley, D. et al. BioMart-biological queries made easy. *BMC Genomics* **10**, 22 (2009).
85. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
86. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
87. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
88. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
89. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).
90. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
91. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
92. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
93. Li, A., Zeng, G., Wang, H., Li, X. & Zhang, Z. DeDoc2 identifies and characterizes the hierarchy and dynamics of chromatin TAD-like domains in the single cells. *Adv. Sci.* **10**, e2300366 (2023).
94. Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol.* **21**, 303 (2020).
95. Gu, Z., Eils, R., Schlesner, M. & Ishaque, N. EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC Genomics* **19**, 234 (2018).
96. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
97. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
98. Rogers, A. R. & Huff, C. Linkage disequilibrium between loci with unknown phase. *Genetics* **182**, 839–844 (2009).
99. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
100. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
101. Fitch, W. M. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**, 406 (1971).
102. Prüfer, K. et al. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8**, 41 (2007).
103. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
104. Ricemeyer, E. S., Schaefer, N. & Acemel, R. D. esrice/amazon-molly-paper: Zenodo DOI release (v0.0.3). *Zenodo* https://doi.org/10.5281/zenodo.17976428 (2025).

Acknowledgements We thank M. Baldwin, L. Frantz, G. Hickey, P. Jern, N. Luscombe, D. Metzler, F. Rheindt and M. Tobler for discussions about the project; G. Schneider and P. Weber for breeding and care of the fish sequenced for these analyses; M. Tobler for permission to use his photo to create the silhouette of *P. mexicana* used in the figures; K. S. Jaron for dedicating silhouettes of *P. latipinna* and *P. formosa* available on PhyloPic to the public domain; F. Berio for dedicating a silhouette of *P. reticulata* available on PhyloPic to the public domain; and I. Schlupp for supplying the *P. mexicana* and *P. latipinna* individuals for reference genome sequencing. Computation for this work was performed on the high performance computing infrastructure 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 MO. The authors gratefully acknowledge the Leibniz Supercomputing Centre for funding this project by providing computing time on its Linux cluster. Research in the Lupiañez lab was funded by the European Research Council (grant no. 101045439, 3D-REVOLUTION) and by the Spanish "Agencia Estatal de Investigación" (grant number PID2022-143253NB-I00/ AEI/10.13039/501100011033/ FEDER, UE). Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions This project was conceived and led by W.C.W. and M.S. with input from E.S.R. N.K.S. performed admixture, selection and gene function-related analyses. R.A.C. performed DNA extraction and library preparation. K.D. annotated genome assemblies. I.d.C. assisted with genome assembly curation. S.K. performed allele-specific expression analysis. R.D. built trees. R.D.A. and D.G.L. analysed chromatin structure. E.S.R. performed all other analyses. E.S.R. and M.S. wrote the first draft of the manuscript, which all authors edited and approved.

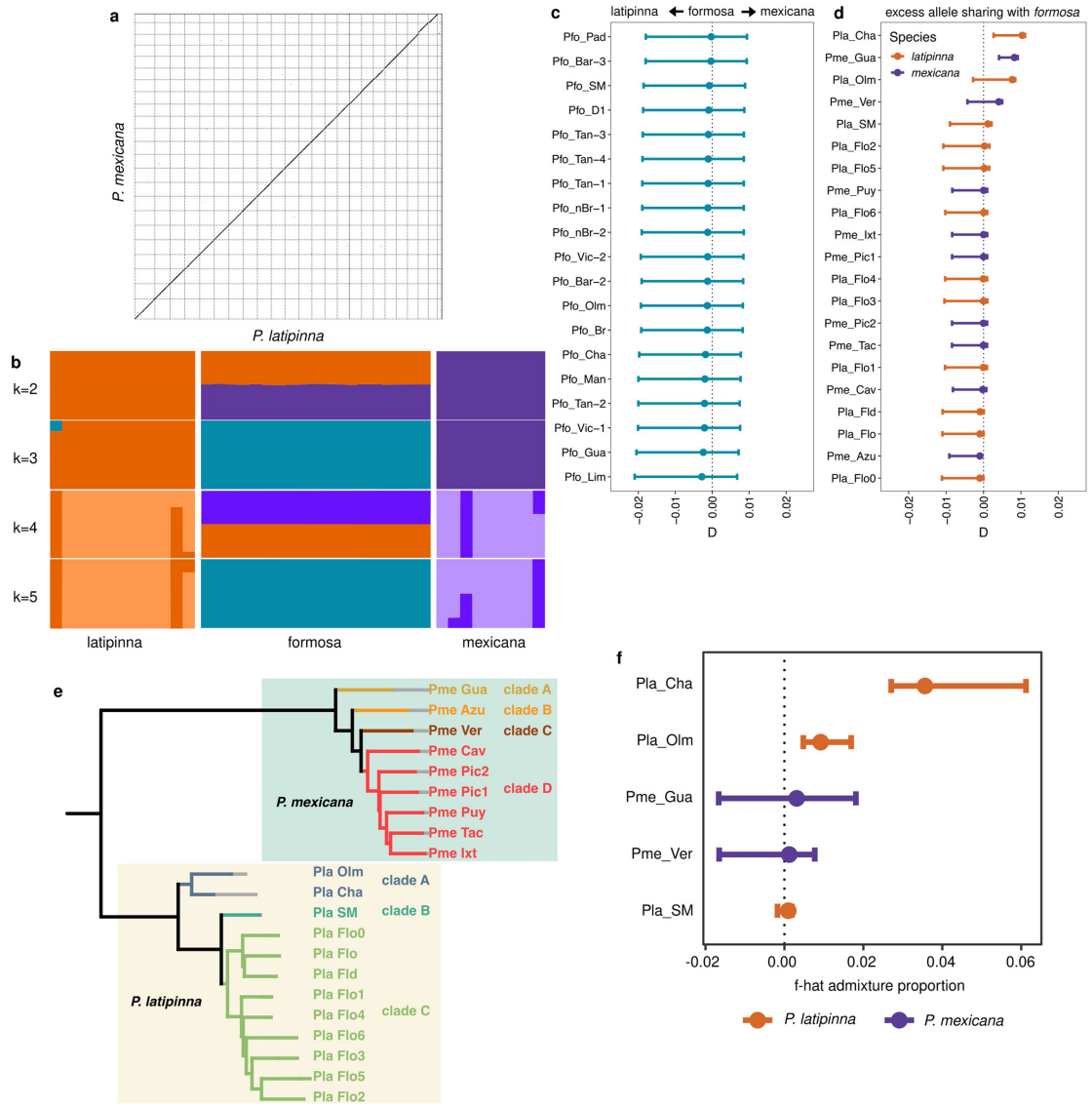
Competing interests The authors declare no competing interests.

Additional information

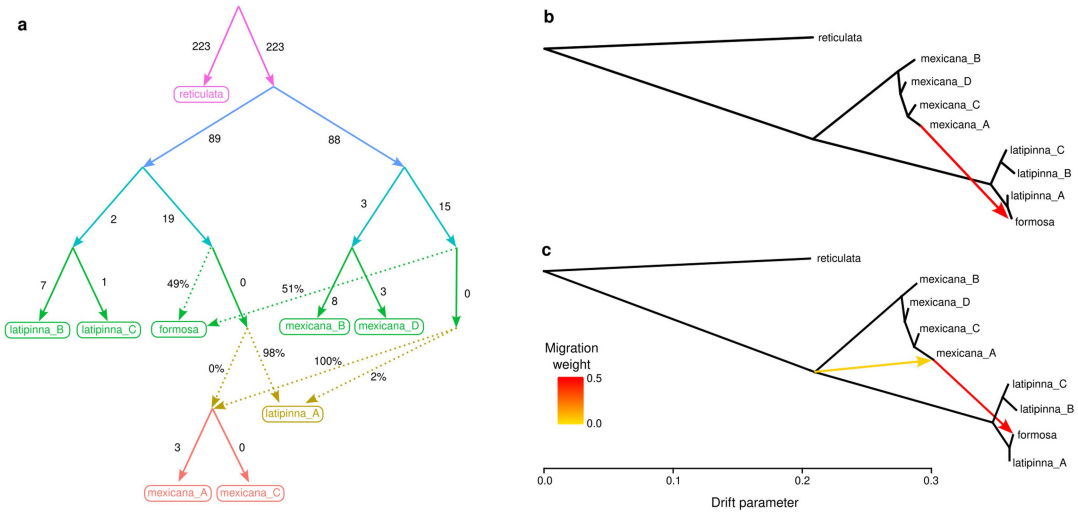
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10180-9.

Correspondence and requests for materials should be addressed to Edward S. Ricemeyer.

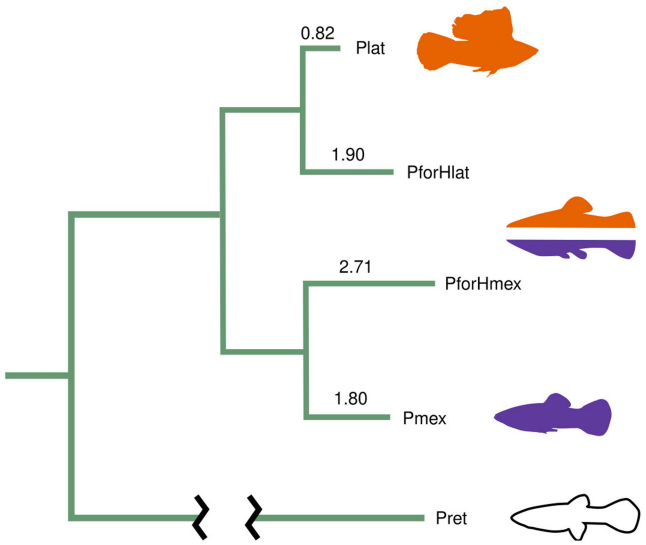
Peer review information *Nature* thanks Daniel Berner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



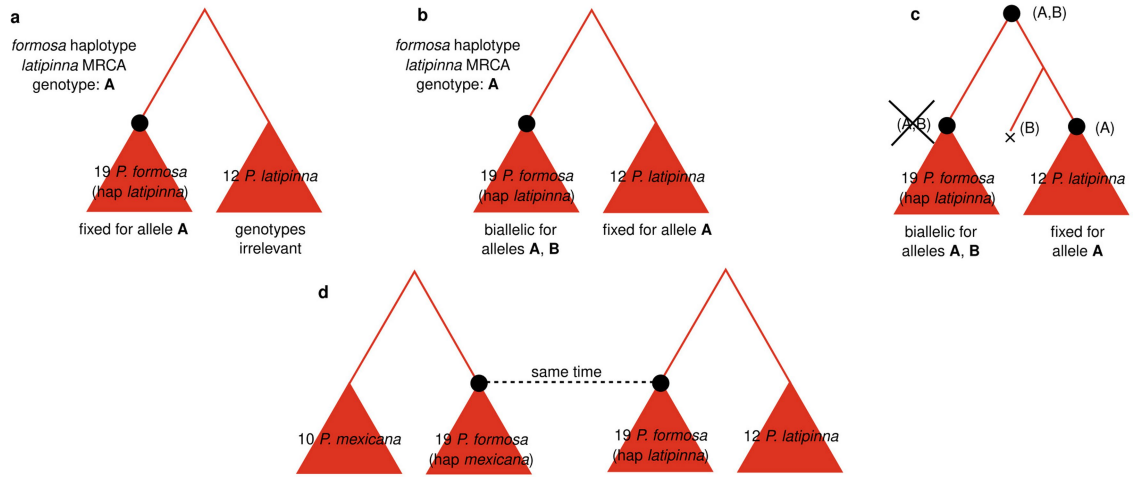
Extended Data Fig. 1 | Gene flow among *P. latipinna*, *P. mexicana* and *P. formosa*. **a**, Gene-order based synteny mapping shows high structural conservation between P. latipinna and P. mexicana. The best path through both annotations was found using a Needleman-Wunsch-like algorithm, which identifies inversions (none found) and translocations (highlighted red). Each dot is a gene present in both annotations. **b**, ADMIXTURE analysis shows signs of at most limited gene flow among the species. **c**, D-statistics of the form D(*latipinna*, *mexicana*, X, *reticulata*), with each *formosa* genome standing in for X and all individuals from the other species included in each calculation. Negative D indicates greater sharing of derived alleles between the shown *formosa* genome and *latipinna* individuals than with *mexicana* individuals, whereas positive D indicates greater sharing between *formosa* and *mexicana* than *formosa* and *latipinna*. Error bars show maximum and minimum values of each genome-wide D calculation including individual X; points are median values. **d**, D-statistics of the form D(X, Y, *formosa*, *reticulata*), where X and Y are either every permutation of two *latipinna* genomes (orange) or two *mexicana* genomes (purple). Positive D indicates greater derived allele sharing between the Y individual and *formosa* than between the X individual and *formosa*, which could result either from gene flow or the Y individual sharing more ancestral drift with the conspecific ancestor of *P. formosa*. **e**, Sub-clades defined for admixture analysis. **f**, $\hat{f}$ statistics for *P. latipinna* and *P. mexicana* individuals showing the highest rates of allele sharing with the opposite species. D-statistics were computed using all conspecific individuals not belonging to the same clade, and with every individual from the opposite population; denominators were computed using every possible combination of two individuals from the opposite population, but with the same P1 individual as the numerator. Points are medians across all possible computations; error bars show minimum and maximum values.



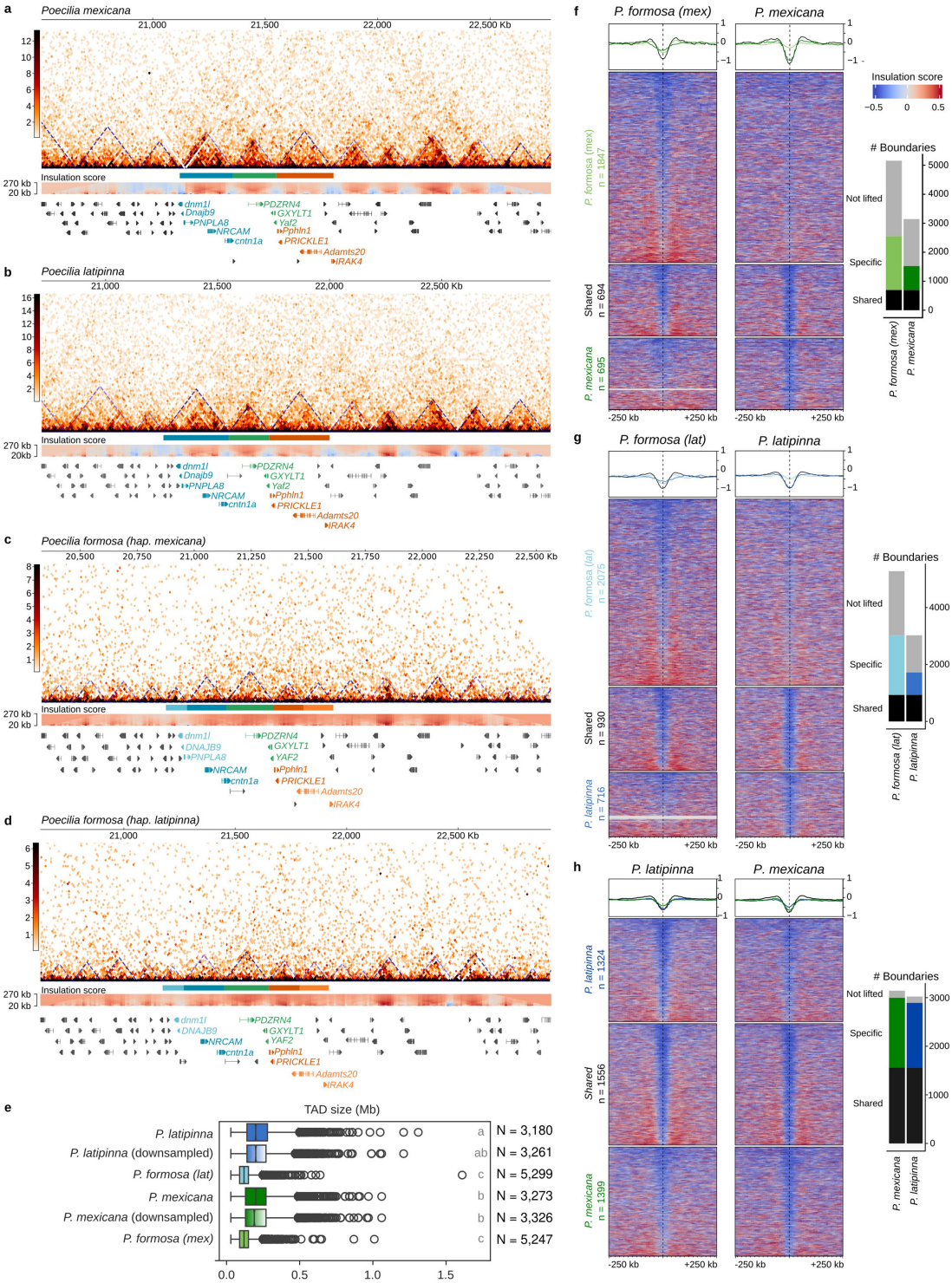
Extended Data Fig. 2 | Admixture graph suggests possible inter-species gene flow into one population of *P. latipinna* and one population of *P. mexicana*. **a**, ADMIXTOOLS2 qpGraph showing the best fit model of those we tested. Solid edges represent parent/child relationships, and solid edge labels quantify drift along those edges. Dotted edges show ancestry derived from multiple groups (admixture), and dotted edge labels quantify the amount of ancestry derived from each group. **b**-**c**, Best fit TreeMix graphs with one (**b**) and two (**c**) gene flow edges.



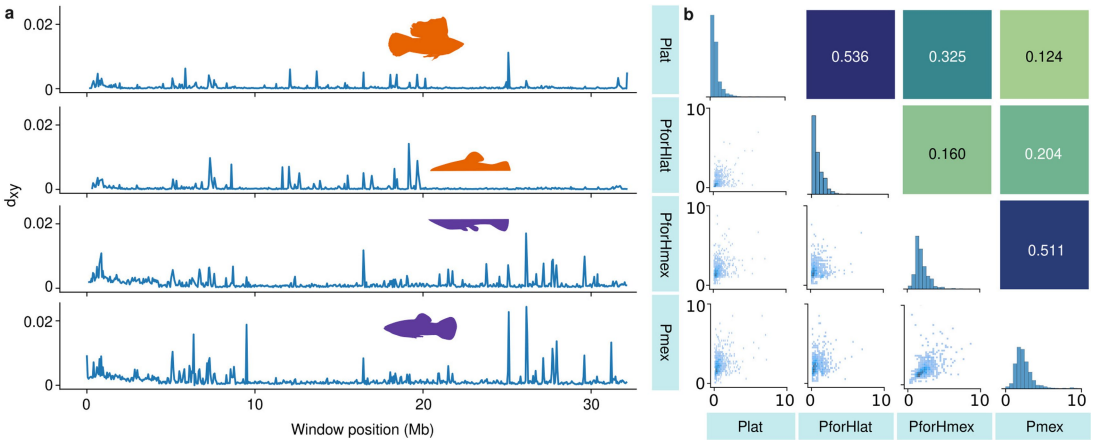
Extended Data Fig. 3 | Neighbor-joining tree of genome assemblies. A neighbor-joining tree using mash distances between each pair among the five assemblies shows the same two patterns as the tree based on ancestral reconstruction: both asexual genomes have diverged more quickly than their sexual counterparts, and the *mexicana*-ancestry genomes are more diverged from their common ancestor than the *latipinna*-ancestry genomes with comparable reproduction strategies. Leaf branches are labeled with mash distances, multiplied by 10^3 for readability. *P. formosa* and *P. latipinna* silhouettes by Kamil S. Jaron (CC0 1.0); *P. mexicana* silhouette by Michael Tobler (CC0 1.0); *P. reticulata* silhouette by Fidji Berio (CC0 1.0).



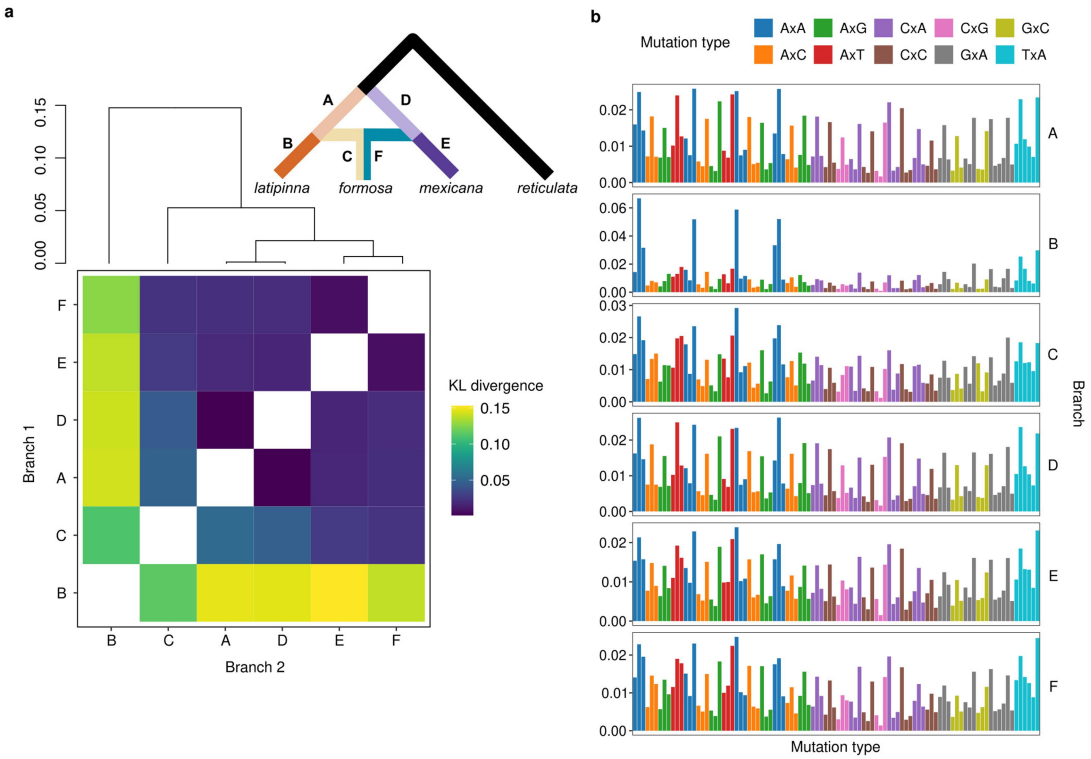
Extended Data Fig. 4 | The single origin and clonal reproduction of *P. formosa* allow confident inference of the ancestral state of both haplotypes of 19 *P. formosa* samples. a, If latipinna-derived haplotypes of all *P. formosa* samples are fixed for allele "A" at a given site, the MRCA of these haplotypes must also be "A" regardless of the genotypes of present *P. latipinna* individuals. **b**, Because of the single origin and clonal reproduction of *P. formosa*, all variation within the species is a result of either differences between the two progenitors or mutations occurring subsequent to speciation; no intraspecies variation from the parental species was inherited by *P. formosa*. Therefore, if a site is variable among the *P. formosa* samples for alleles "A" and "B" but fixed in *P. latipinna* for "A", the MRCA of the *P. formosa* samples must be "A" except in cases of shared variation between *P. mexicana* and *P. latipinna*, which are rare based on high fixation and low gene flow between these two species (see Supplementary Note 1). A counter-example is illustrated in the next panel. **c**, This method is robust to undersampled variation in the sexual outgroup species: even if both "A" and "B" alleles were present in the ancestor of *P. latipinna* and *P. formosa* but only the "A" allele was sampled in *P. latipinna* due to either undersampling or post-speciation fixation of the "A" allele in *P. latipinna*, *P. formosa*, having descended from a single individual, could only have inherited one of these alleles except in cases of shared variation between the parental species. Fst and D-statistic analyses confirm the extremely low level of shared variation among *P. latipinna*, *P. mexicana*, and *P. formosa*. **d**, The MRCA of the 19 *P. formosa latipinna*-derived haplotypes and the 19 *P. formosa mexicana*-derived haplotypes necessarily existed at the same time because they were present in the same fish.



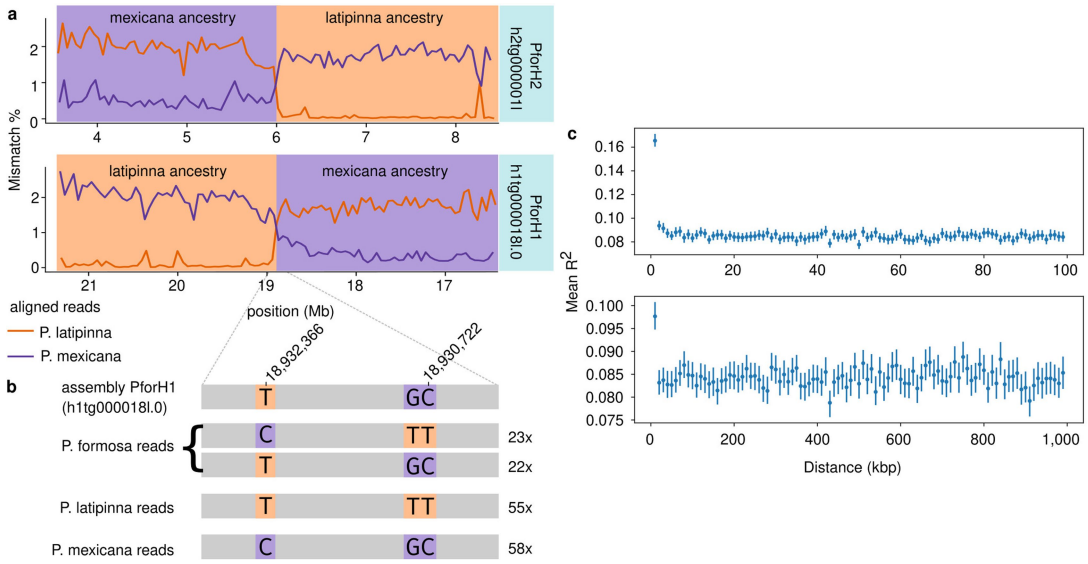
Extended Data Fig. 5 | Chromatin conformation shows that although TAD boundaries are conserved, switch to asexual reproduction increased insulation genome-wide. a-d, KR-normalized Hi-C matrices at 10 kb resolution around syntenic regions of (a) *P. mexicana*, (b) *P. latipinna*, and the *mexicana* (c) and *latipinna* (d) haplotypes of *P. formosa*. Dedoc2 predictions of TADs (see methods) are overlaid as blue discontinuous lines. Syntenic TADs are marked with matching colored rectangles. Split TADs in *P. formosa* (c) and (d) can be distinguished with different tones of the matching colors. Syntenic genes that characterize the syntenic TADs are also colored accordingly. Insulation score heatmaps at different window sizes from 20 to 270 kb are shown below the Hi-C heatmaps. **e**, Boxplots showing the distribution of TAD sizes in the different *Poecilia* species. Results of the TAD analysis after downsampling the total number of contacts to the levels of *P. formosa* in *P. mexicana* and *P. latipinna* are also shown. The results of Bonferroni corrected pairwise Mann-Whitney U-tests are summarized using compact letter display (CLD). Boxplots display the median (center line), the interquartile range (IQR; 25th–75th percentiles), and whiskers extending to the datapoints within 1.5×IQR. P-values are calculated with two-sided wilcoxon tests. Reported p-values were adjusted for multiple testing using Bonferroni. **f-g**, Insulation score heatmaps (100 kb window size) around syntenic boundary positions between either *P. mexicana* (f) and *P. latipinna* and the corresponding subgenomes of *P. formosa*. Boundaries were classified as shared if they were called as boundaries in both species or species-specific if they were only called as boundaries in either of them. A quantification of the shared, specific and not-lifted boundaries in each of the comparisons is shown on the right. **h**, As in f-g but comparing directly the boundaries of the parental *P. latipinna* and *P. mexicana*.



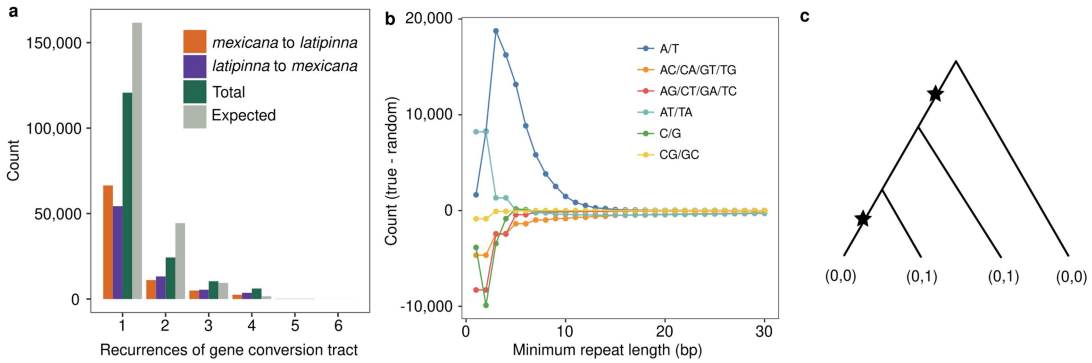
Extended Data Fig. 6 | Local divergence rates are correlated between sister genomes. a, Single-base divergence from ancestral genome within 50 kb windows on LG2 for each assembly, in Pmex coordinates. **b**, Bivariate distributions of genome-wide window divergence rates for each pair of assemblies, with Pearson correlation coefficients above (all correlations $p < 10^{-3}$, values for all comparisons in Supplementary Data Table 6). *P. formosa* and *P. latipinna* silhouettes by Kamil S. Jaron (CC0 1.0); *P. mexicana* silhouette by Michael Tobler (CC0 1.0).



Extended Data Fig. 7 | The mutation spectrum shows a possible derived shift in *P. latipinna* lineages relative to others. On each branch (A-F) shown in the tree, we compared reference genome sequences and counted mutations by considering their trimer context (e.g. AAA to ACA or CGG to CTG), then divided these counts by the total mutations on the branch, to generate a branch-specific mutation spectrum. **a**, Kullback-Leibler (KL) divergence between each pair of mutation spectra are shown in the heatmap. The dendrogram is the result of hierarchically clustering the KL distances. KL divergence is not symmetrical; the heatmap values correspond to $D_{KL}(Branch1||Branch2)$. **b**, Each bar represents the normalized count of a specific mutation type on a specific branch. Boxes correspond to labeled branches in the tree, and mutations are colored by their context (the base immediately upstream and the base immediately downstream of the mutation).



Extended Data Fig. 8 | Long read alignments show rare crossing-over recombination. a-b, Aligning long reads from all three species to both haplotypes of *P. formosa* reveals a crossing-over recombination between the haplotypes. **a**, Comparing mismatched base percentages of long reads from *P. latipinna* vs. *P. mexicana* aligned to both *P. formosa* haplotypes shows locations of two breakpoints where a *P. formosa* contig switches ancestry. These two breakpoints are in homologous positions between the two haplotypes. **b**, Many individual *P. formosa* long reads span the breakpoint, supporting the inference of a crossing-over recombination. **c**, Linkage disequilibrium on AncMol chr0 measured by the Rogers-Huff r^2 estimator at 1 kb bins for biallelic pairs from 0-100 kb apart with 5% of all pairs sampled (top), and 10 kb bins from 0-1 Mb apart with 1% of all pairs sampled (bottom). n = 19 biologically independent *P. formosa* individuals; bootstrapped 95% confidence intervals are shown based on 1,000 resamples. As expected, LD is stable with increasing distance between loci in *P. formosa*. The only exception to this is for loci less than 1 kb apart, which have higher mean linkage disequilibrium than loci pairs in all other distance bins, consistent with gene conversion, which acts on relatively small pieces of the genome, but not crossing-over and reassortment during sexual reproduction, which should affect LD over greater distances.



Extended Data Fig. 9 | Gene conversions arise repeatedly at the same loci and often near polyA/T repeats. a, Gene conversion tracts arise repeatedly at the same loci. For each consensus gene conversion tract, we determined how many times at minimum it must have arisen independently, according to maximum parsimony and the consensus phylogenetic tree using SNP data. We then compared these numbers with those expected, assuming that non-overlapping genomic segments undergo gene conversion following the Poisson distribution. **b**, Numbers of occurrences of each possible type of mononucleotide or dinucleotide repeat within 50 bp of a gene conversion breakpoint, minus numbers of occurrences of the same types of repeats in matched, randomly sampled 50 bp intervals. Each point corresponds to repeats of the given length or longer. **c**, Due to clonal reproduction in *P. formosa*, the ancestry of all parts of the genome must be the same outside of regions of gene conversion; this can be used to detect gene conversion using Fitch's algorithm. In the example tree shown, the most parsimonious set of mutations to explain the genotypes is an earlier mutation from homozygous (0,0) to heterozygous (0,1) followed by a gene conversion back to homozygous (0,0), both represented with stars.

nature portfolio

Corresponding author(s): Edward S. Ricemeyer
Last updated by author(s): Jan 8, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection in this study.
Data analysis	All third-party software used for data analysis in this study is named with version numbers, parameters, and citations to any relevant publications in the Methods or Supplementary Methods. All software written for this study is available in the project's code repository on GitHub at https://github.com/esrice/amazon-molly-paper Third-party software used in this study: hifiasm v0.16.1-r375 kmc v3.2.1 trio_binning v1.0.0 chromap v0.2.3-r407 YAHS v1.2a.1 Juicebox Assembly Tools v2.20.00 TETools v1.7 GeneWise v2.4.1 exonerate v2.2.0 genblasta v1.0.1 HISAT v2.2.1 StringTie v2.2.1 Trinity v2.9.1 splign v2.1.0

AUGUSTUS v3.5.0
BUSCO v5.3.2
minimap2 v2.28-r1209
elprep v5.1.3
GATK v4.5.0.0
WhatsHap v2.3
cactus v2.4.2
BioPython v1.85 (including Phylo)
PAML v4.10.9
hts_popgen commit 8b5d8fc
SnpEff v5.2e
hal toolkit v2.2
mash v2.3
hic_tools v3.30
FAN-C v0.9.28
EnrichedHeatmap v1.34.0
sorolla commit 46c3a1ec
samtools v1.20
IGV v2.17.4
bedtools v2.30.0
GOfuncR v1.22.0
svgenes commit 9e93471
plink v1.9.0-b.8
ADMIXTOOLS2 v2.0.10
Ancestry_HMM v1.0.2
STAR v2.5.1b
blastn v2.3.0
LDHat v2.2a
vcftools v0.1.16
heRho commit 90c27d1
ADMIXTURE v1.3.0
TreeMix v1.13
OptM v0.1.9
scipy v1.15.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All sequence data generated for this project, as well as the assembled genomes, are deposited in the relevant NCBI databases under the following BioProject IDs: PRJNA1398340 (P. latipinna), PRJNA1398337 (P. mexicana), and PRJNA1398335 and PRJNA1398336 (P. formosa). Hi-C reads for P. latipinna and P. mexicana are in PRJNA614959.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No human participants or human data were involved in this study.
Reporting on race, ethnicity, or other socially relevant groupings	No human participants or human data were involved in this study.
Population characteristics	No human participants or human data were involved in this study.
Recruitment	No human participants or human data were involved in this study.
Ethics oversight	No human participants or human data were involved in this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We studied genomes of a set of samples without performing any treatments or manipulations to these individuals.
Research sample	We assembled genomes for one individual each of the three species involved in this study. These individuals were each lab-reared samples from established lines. For short-read samples, we used previously-published wild samples.
Sampling strategy	We used a single individual from each of the three species for creating reference genomes. For short-read data, we used all previously-published short-read data sets of sufficient depth.
Data collection	DNA extraction was performed by RAC.
Timing and spatial scale	We used previously published genomic data from across the spatial ranges of the three species.
Data exclusions	No data were excluded from this study.
Reproducibility	We resampled calculations with bootstrapping where relevant to produce 95% confidence intervals.
Randomization	We did not allocate samples into groups or perform any other randomization.
Blinding	We did not perform any blinding.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	The P. formosa used in this study was a 7-month-old adult female from a clonal subline (WLC 7122, P formosa +M) established from fish collected in 1953 by C. P. Haskins near Brownsville, TX, USA. The P. mexicana used was a 9-month-old adult from a lab-reared population established from fish collected from Rio Purification bei Nueva Padilla, Tamaulipas, Mexico. The P. latipinna was a 9-month-old adult from a lab-reared population established from fish collected from Laguna de Champayan, Tamaulipas, Mexico.
Wild animals	No data from wild animals were collected for this study.

Reporting on sex	P. formosa is a female-only species.
Field-collected samples	No data from field-collected samples were generated for this study.
Ethics oversight	Animal work was performed through authorization 568/300-1870/13 of the Veterinary Office of the District Government of Lower Franconia, Germany, in accordance with the German Animal Protection Law (TierSchG)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	No plants were involved in this study.
Novel plant genotypes	No plants were involved in this study.
Authentication	No plants were involved in this study.