nature ecology & evolution

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41559-018-0473-y

In the format provided by the authors and unedited.

OPEN

Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly

Wesley C. Warren[1]*, Raquel García-Pérez[2,21], Sen Xu[3,21], Kathrin P. Lampert[4,21], Domitille Chalopin[5], Matthias Stöck[6], Laurence Loewe[7], Yuan Lu[8], Lukas Kuderna[2], Patrick Minx[1], Michael J. Montague[9], Chad Tomlinson[1], LaDeana W. Hillier[1], Daniel N. Murphy[10], John Wang[11], Zhongwei Wang[12,13], Constantino Macias Garcia[14], Gregg C. W. Thomas[15], Jean-Nicolas Volff[5], Fabiana Farias[1], Bronwen Aken[10], Ronald B. Walter[8], Kim D. Pruitt[16], Tomas Marques-Bonet[2,17,18], Matthew W. Hahn[15], Susanne Kneitz[12], Michael Lynch[15] and Manfred Schartl[12,19,20]*

[1]McDonnell Genome Institute, School of Medicine, Washington University, St Louis, MO, USA. [2]Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain. [3]Department of Biology, University of Texas at Arlington, Arlington, TX, USA. [4]Department of Animal Ecology, Evolution and Biodiversity, Ruhr-Universität Bochum, Bochum, Germany. [5]Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, CNRS, Université Lyon, Lyon, France. [6]Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany. [7]Laboratory of Genetics and Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA. [8]Department of Chemistry and Biochemistry, Texas State University, San Marcos, TX, USA. [9]Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [10]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK. [11]Biodiversity Research Center, Academica Sinica Taipei, Taipei, Taiwan. [12]Department of Physiological Chemistry, Biocenter, University of Würzburg, Würzburg, Germany. [13]Institute of Hydrobiology, Chinese Academy of Sciences, Beijing, China. [14]Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad Universitaria, Mexico City, Mexico. [15]Department of Biology, Indiana University, Bloomington, IN, USA. [16]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [17]Center for Genomic Regulation (CGR), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [18]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain. [19]Hagler Institute for Advanced Study and Department of Biology, Texas A&M University, College Station, TX, USA. [20]Comprehensive Cancer Center Mainfranken, University Hospital Würzburg, Würzburg, Germany. 21These authors contributed equally: Raquel García-Pérez, Sen Xu and Kathrin P. Lampert. *e-mail: wwarren@genome.wustl.edu; phch1@biozentrum.uni-wuerzburg.de

Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly

Supplementary Information

Note: Supplementary Tables 7, 8, 9, 13, 14, 15, 25, and 35 are available as single Excel files.

Supplementary Note 1: Transposable element analysis

Repeat element annotation was performed to obtain a complete inventory of transposable elements (TEs) for the Amazon molly, as well as for the two parental species. TEs account for about 22-23% of the genomes of *P. formosa, P. mexicana* and *P. latipinna* (Supplementary Table 4). This is in the range of other similar sized fish genomes [1] and in particular of closely related species, the guppy (25%), Southern platyfish (22.5%), Monterrey platyfish (21.8%) and swordtail (21.1%). We found a high diversity of TE families in the Amazon molly genome (Supplementary Table 4), with many families that are also present in other fish genomes but absent from mammals and birds [2]. The *P. formosa* genome contains all major groups of vertebrate DNA transposons and retrotransposons.

Most TE families are represented by relatively low full copy number except a few that show a strong expansion, including two families of hAT transposons (>1000 copies longer than 80% of the reference sequence), two TcMariner families (>600 copies), PiggyBac transposons (about 1000 copies), Rex-Babar non-LTR retrotransposons (>400 copies) and V-SINE retrotransposons (>1200 copies) (Supplementary Table 5). No major difference in TE composition was detected between the genome of *P. formosa* and the parental genomes.

When we analyzed TE transposition history in the genomes of *P. formosa* and its parental species (Figure 2A), it became apparent that most of the superfamilies have a rather ancient transposition history represented by a high percentage of divergent copies, mostly concerning TcMariner elements. However, a burst of transposition occurred in recent timing (K-3/4), mostly involving hAT transposons. The comparison with two other Poecilids, i.e. *X. maculatus* and *P. reticulata*, indicates that the oldest major burst, which is common to all species tested, might have occurred in the Poecilid ancestral genome. The more recent burst is very different between the different poeciliid genomes and is probably genera-specific: a small burst in *X. maculatus*, a very high burst followed by a strong decrease in *P. reticulata* and an intermediate burst in *P. formosa* and parents. Finally, a higher fraction at most recent timing (K-0/1) in *P. formosa* might reflect species-specific transposition event. No major differences in transposition history between *P. formosa* and the parental genomes and other related species was noted.

Based on Kimura plots, TEs appear to be more active in *P. formosa* than in parental genomes (percentage higher in K-0). Presence of TE sequences in the transcriptome can indicate such an activity. TEs make about 1.4% of the Amazon molly transcriptome (Supplementary Table 6), which indicates a somehow lower overall activity compared to the platyfish (5.4%). Indeed, the most highly transcribed TE superfamilies are gypsy retrotransposons and TcMariner transposons. Expression of the Gypsy elements, but not TcMariner, might be the result of their activity rather than of general background expression because their relative fraction is remarkably higher in the transcriptome than in the genome (Figure 2B).

Because long terminal repeat (LTR) retrotransposons are known to play fundamental roles in genome rearrangement after “catastrophic” events such as polyploidization or hybridization [3] we investigated Gypsy element activity by comparing copy-specific insertion sites within *P. formosa* to the parental genomes. Six of the 47 insertions of Gypsy elements are absent from both parental genomes while the remainder is found

at the same genomic position in all three genomes. These *P. formosa*-specific insertions may be explained by post-hybridization transposition events of Gypsy elements.

Despite the absence of recombination, the RNA-mediated silencing machinery provides important mechanisms in epigenetic silencing of TEs preventing them from multiplying. Members of this critical pathway in *P. formosa* were not lost, corrupted or expanding faster than expected relative to other sexual teleost species (Supplementary Table 31).

Supplementary Note 2: Gene selection, gene family expansion and gene duplication

Gene selection

Using analyses of protein coding genes relative to other closely related taxa, it is possible to discover signatures of positive selection along the *P. formosa* lineage by measuring the dN/dS ratio (the ratio of the rate of nonsynonymous substitutions to the rate of synonymous substitutions). We therefore selected 8,286 one-to-one orthologous gene alignments from eight species of teleosts (*Poecilia formosa, Xiphophorus maculatus, Oryzias latipes, Oreochromis niloticus, Tetraodon nigroviridis, Gasterosteus aculeatus, Astynax mexicanus, Danio rerio*). Next, using maximum likelihood ratio tests, we selected genes that showed significant results ($p<0.01$) for both the "free-ratio" and "two-ratio" model comparisons of branch tests. We measured the signal across the entire gene and selected only those genes that were significant for the molly lineage for further biological inference measures. Our analysis revealed 351 genes (4.3%) as significantly diverged. Canonical pathway (KEGG) enrichment analysis found only two pathways of significance ($p<.002$). These two pathways, calcium signaling and neuroactive ligand-receptor interaction, both have unknown phenotypic relevance specific to *P. formosa*. Further assessments of particular nonsynonymous substitutions within the positively selected gene set revealed that 153 genes (43.6%) contained substitutions that made significant impacts on the biological function of the translated protein.

Among total shared loss-of-function variants (LoF) *P. formosa* has a two-fold higher number than either *P. mexicana* or *P. latipinna* (Supplementary Table 13) however, the sexual parents shared LoF variants across all samples from different geographical locations are likely due to false calls or faulty gene models with incorrect consensus bases in the reference causing frameshifts. Once LoF variants shared across all sexual parents within species are removed the occurrence of LoFs in *P. formosa* is lower than either of the sexual parents of LoF types (Supplementary Figure 13A). We find the average number of LoFs per *P. formosa* individual to be 984 while 1,085 and 1,167 for *P. latipinna* and *P. mexicana*, respectively. If we only focus on genes that experienced a likely more damaging event, stop gain or losses, we also see *P. formosa* has a lower count (Supplementary Table 13). We find a total of 63,241 non-synonymous shared variants within *P. formosa* compared to 13,588 and 16,174 within *P. latipinna* and *P. mexicana*, respectively (Supplementary Figure 13 B). Contrary to the overall higher polymorphic sites in *P. formosa,* the average of homozygous non-synonymous base changes is ~10-fold lower in *P. formosa*. Should the asexual genome be carrying parental copies of each allele this observation would be expected if gene conversion

was rare which we have confirmed. Average non-synonymous base substitutions, heterozygous or homozygous allelic state, is only slightly higher in *P. formosa*. A measure of neutral mutations in sexual species is the number of synonomous sites. In *P. formosa* we observe a much higher number of synonomous sites compared to the sexual parents (Supplementary Figure 13 B). We attribute this finding to higher numbers of polymorphic sites at the time of hybridization and their neutral effect on genome fitness. We find no overlapping genes that share putative LoFs within species. Only a few *P. formosa* genes with putative LoFs (n=4) are shared with one of the parents. These putative LoFs assumed to be fixed in each species were possibly inherited by *P. formosa* at the time of it's hybridization.

A total of 254 genes were found to contain LoF variants within the *P. formosa* genome that upon evaluation for GO enrichment revealed significant GO annotation only for the biological processes category. Other categories, i.e. molecular, were not significant. We find eight genes that belong to the neuromuscular process grouping (Supplementary Table 32). *P. formosa* have no known difference in neuromuscular control compared to their sexual parental species. However, proof of this difference will require additional investigations.

Gene family evolution

In order to identify rapidly evolving gene families along the Amazon molly lineage we obtained peptides from *Astyanax mexicanus* (Cavefish), *Danio rerio* (Zebrafish), *Tetraodon nigroviridis* (Green spotted puffer), *Gasterosteus aculeatus* (Stickleback), *Oreochromis niloticus* (Tilapia), *Oryzias latipes* (Medaka), *Xiphophorus maculatus* (Platyfish) and *Poecilia formosa* (Amazon Molly) from ENSEMBL 76 [4]. To ensure that each gene was counted only once, we used only the longest isoform of each protein in each species and all other isoforms were filtered out. We then performed an all-vs-all BLAST [5] search on these filtered sequences. The resulting e-values from the search were used as the main clustering criterion for the MCL program to group peptides into gene families [6]. This resulted in 13,667 clusters. Of these, all single-peptide clusters were filtered out, leaving 8,719 gene families.

We also constructed an ultrametric tree (Supplementary Figure 1) for these 8 species based on ENSEMBL's ortholog identifications. We aligned the nucleotide coding sequences of the 844 one-to-one orthologs with MUSCLE [7]. Then we used RAxML [8] to construct gene trees for each alignment. These trees were combined using the average consensus method [9] implemented in SDM [10]. This resulted in a single species phylogeny with branch lengths in terms of relative number of changes. To obtain branches in terms of absolute time the tree was smoothed using the r8s program. r8s uses a semiparametric, penalized likelihood approach to estimate substitution rates and divergence times across a phylogeny, given at least one calibration time [11]. We calibrated the times on the tree based on the divergence between Zebrafish and Stickleback at 265 million years from TimeTree [12].

With the gene family data and ultrametric phylogeny as input, we estimated gene gain and loss rates (λ) with CAFE v3.0 [13]. This version of CAFE is able to estimate the amount of assembly and annotation error (ε) present in the input data using a distribution across the observed gene family counts and a pseudo-likelihood search.

CAFE is then able to correct for this error and obtain a more accurate estimate of λ. We find an ε of about 0.06, which implies that 6% of gene families have observed counts that are not equal to their true counts. After correcting for this error rate, we find λ = 0.0007. Using the estimated λ value, CAFE infers ancestral gene counts and calculates p- values across the tree for each family to assess the significance of any gene family changes along a given branch. Those branches with low p-values are said to be rapidly evolving. Variations in the number of copies of genes between species have also been proposed to be a driving force of evolutionary change [14]. Both the loss and gain of genes can be adaptive and various methods exist to quantify these changes [15]. With this in mind, we sought to classify gene family evolution in the *P. formosa* and several closely related fish species (See methods). A summary of all gene family changes for all species (n=8) measured in this study (Supplementary Table 33) show *P. formosa* has a comparatively higher number of rapid expansions (110 of 131) compared to sexual fishes.

Given a *P. formosa* unique mode of reproduction, we sought molecular evidence of any rapid gene family turnover supporting this trait. After correcting for multiple testing (Dunn-Sidak correction), we find almost no evidence for rapidly evolving gene families linked to reproductive function. However, among the families without enriched GO terms we do observe a family of matrix metallopeptidases (family 75) which has functions related to reproduction [16,17]. Another family of Diaphanous homologs (681) was found to be rapidly expanding in *P. formosa* and is also associated with reproduction (GO:0007292). Finally, a family of highly conserved Hox proteins (37), which partition segments of the body during development [18,19], was found to be rapidly contracting in the *P. formosa*.

Gene duplication

We found 14.8 Mb, ~1.97%, of the *P. formosa* genome, to be located in segmental duplications (SDs) across the 19 analyzed samples (Supplementary Table 34). For the parental species, we detected 16.6 Mb (2.22%) in the *P. latipinna* genome and 16.5 Mb (2.20%) in the *P. mexicana* genome to be in shared SDs across the respective samples (Supplementary Table 34). A total of 11.5 Mb (78.07% of calls in the *P. formosa* samples) are shared SDs across all three species. Stratifying by species, 85.52% of the *P. formosa* SDs are present in *P. latipinna* and 85.44 % in *P. mexicana.* Across samples the median value of SDs in duplication is 22.5 Mb (range: 21.1-28.1 Mb) for *P. formosa*, 22.9 (range: 22.4-23.8 Mb) for *P. latipinna* and 22.4 Mb (range: 21.9-22.8 Mb) for the *P. mexicana*. The average of shared duplicated regions for each individual are 65.6% and unique SD ranged from 0.59 to 3.1% (Supplementary Table 34).

We intersected the genomic coordinates of the shared duplications on a per species basis with the genomic coordinates of the *P. formosa* genes. We required at least 80% of the gene sequence to fall within a shared duplication to consider a gene as duplicated. This resulted in a total of 283 duplicated genes within SDs among all the *P. formosa* samples (Supplementary Table 35), 210 in the *P. latipinna* and 207 in the *P. mexicana* samples. Altogether, we find 131 genes to be in fixed duplications across all 3 species. In Supplementary Tables 16-21 we include the GO terms significantly enriched (corrected p-value < 0.05) in each of the GO categories analyzed. We find several GO-terms associated with transposition and DNA recombination to be significantly

enriched in the *P. formosa* and both parental species SD regions. We retrieved the 11 unique genes in duplications that contribute to the terms "DNA integration" (GO:0015074), "transposition, DNA-mediated" (GO:0006313), "transposase activity" (GO:0004803) and "DNA recombination" (GO:0006310). We compared absolute copy number in these genes between the *P. formosa* samples and the parental species (Supplementary Figure 14), and find 3 genes to have significantly expanded copy number in the *P. formosa* samples (ENSPFOG00000001300, ENSPFOG00000004306, ENSPFOG00000021805, Wilcoxon test, $p<0.05$). ENSPFOG00000001300 and ENSPFOG00000004306 are Tc1 transposons of the same family and ENSPFOG00000021805 is related to Gypsy transposons. We suggest each most likely represents expansions of certain TEs, which, however, do not contribute to a higher total TE content of *P. formosa* compared to *P. mexicana* and *P. latipinna*. To clarify the putative adaptive role of these unique gene family expansions and contractions occurring within the asexual vertebrate genome, regardless of the molecular mechanism of duplication, further experimental study is required.

Supplementary Note 3: Copy number variation

The asexual reproduction mode of the Amazon molly requires that diploid oocytes are formed without meiosis. In order to assure that homologous chromosomes and not chromatids are separated in meiotic division I the kinetochores of each pair of homologous chromosomes must capture microtubules emanating from the opposite poles of the acentrosomal meiotic spindle. Thereby the bipolarity of the kinetochore of each chromosome has to turn into unipolarity. This process is regulated by a complex of proteins that delay microtubule-kinetochore (K-MT) attachment and homolog separation for hours until anaphase of meiosis I is reached [20], while in mitosis K-MT capture is completed within minutes. Disturbance of this process leads to missegregation of chromosomes and can even turn meiosis I into a normal mitosis [21]. We find that several genes involved in meiotic K-MT interaction show copy number variation in the *P. formosa* genome (Supplementary Table 22). In particular, cyclin dependent kinase 1 (*cdk1*) is present in 3 copies, all of which are expressed in ovaries of *P. formosa*. Cdk1 has been shown to be an essential regular of the delay of K-MT capture and increasing Cdk1 activity in mouse oocytes resulted in premature stable attachments of the spindle [22]. CNV (2 copies or more) is also observed for the noncyclin protein Ringo/Speedy, which substitutes for cyclinB in activation of Cdk1 in oocytes [23]. The biorientation of chromosomes in cell division (*bod1*) gene, which encodes a kinetochore-associated protein involved in regulation of K-MT attachment stability [24], is present in two copies in the *P. formosa* genome, as well as other regulators of kinetochore function, including aurora kinase b (*aurkb*), and securin/pituitary tumor- transforming 1 (*pttg1*). As local lineage-specific gene duplications we also find genes involved in cytokinesis (*sept6*) and progression of meiosis (*mos*). Interestingly, compared to the parental species the gene encoding the centrosomal protein centlein (*cntl*) is under positive selection in *P. formosa* (Supplementary Table 15).

All genes that show CNV and *cntl* are expressed in *P. formosa* ovaries with higher levels in the previtellogenic, immature ovary (Supplementary Table 22).

Supplementary Note 4: Mitochondrial DNA analysis

To elucidate the evolutionary origin and in particular the number of hybridization events represented by the present day individuals of *P. formosa* we produced reference mitochondrial genomes of all three species by long PCR and Sanger sequencing and reconstructed the complete mitochondrial genomes of all population samples.

We first assembled and annotated the complete mitochondrial genomes of one paternal (*P. latipinna*, Pla_Tam-S, origin: North of Tampico,) and maternal (*P. mexicana*, Pme_Nue-S, origin: Rio Purification at Nueva Padilla) ancestor and two *P. formosa* (Pfo_Br-S, origin: Brownsville, Pfo_SM-S, San Marcos). The four complete mitochondrial genomes are deposited on GenBank under accession numbers KT175511-14.

Thirteen different mitochondrial haplotypes were found in the 20 investigated *P. formosa* samples (Supplementary Figure 2). Mitochondrial haplotypes within *P. formosa* differed in two to 16 bases (total alignment length 16,582bp) and were more closely related to the *P. mexicana* haplotype (14 to 26 differences) than to any of the *P. latipinna* haplotypes (1,137 to 1,145 differences) confirming *P. mexicana* as the maternal ancestor. Identical haplotypes were not only found at the same field site (haplotypes A, C, D) but also shared between individuals from different field sites (haplotypes B, E, F). We then calculated phylogenetic trees from the complete mitochondrial genomes of *P. formosa*, *P. mexicana*, and *P. latipinna* allowing for a sequence length of 16,587 bp available (Figure 3) with phylogenetic parameters; GTR model of sequence evolution; BioNJ as a starting tree, combined subtree pruning and regrafting plus nearest neighbor interchange options for tree improvement; otherwise, default parameters (http://atgc.lirmm.fr/ phyml/ for details). This revealed substantial subclustering of the maternal ancestral *P. mexicana,* of which only the northern subspecies (taxonomically classified as *P. mexicana limantouri*) forms a sister clade of *P. formosa* from throughout its range. Importantly, all *P. formosa* are united in a highly supported single clade (Figure 3) confirming a single maternal origin of *P. formosa*.

To make a new estimate on the age of the hybridization event that generated *P.formosa* complete mitochondrial sequences of poeciliid fish including *Xiphophorus maculatus* (GenBank NC011379.1), *X. couchianus* (KT594624.1), *X. helleri* (FJ226476.1), *Gambusia affinis* (NC_004388) and *Poecilia reticulata* (AB898687.1) were aligned with the newly obtained complete mtDNAs of *P. formosa, P. mexicana* and *P. latipinna*. Non-alignable parts of the control region (D-loop) and maximal 123 bp of an INDEL from each of the three *Xiphophorus* and *Gambusia* mitochondrial DNAs were cut out, resulting in a final alignment of 16607 bp.

We calculated Bayesian phylogenetic trees using BEAST v. 1.8.3 [25]. From the recently published fossil-based, multi-species and multi-locus phylogeny by Helmstetter et al. ([26] therein Supplementary Figure 1b), we used two calibration points as priors, provided by this reference as minimal ages: (i) *Xiphophorus helleri* and *Gambusia affinis* diverged from *Poecilia reticulata* 18.87 Mya and (ii) *Gambusia affinis* from *Xiphophorus helleri* 14.9 Mya, assuming normal distributions of the divergence time of 1 My around each node. We estimated the time to the most recent common ancestors of each taxon, *P. latipinna, P. mexicana* and *P. formosa*, assuming a yule birth rate (as most appropriate for speciation models) under a simple JC-substitution model, obtained with jModelTest[27] under a strict molecular clock with free variation of the substitution rate,

starting with a prior of 1, using an MCMC chain of 10 million generations. All multiple-sample taxa were considered to be monophyletic, except *P. mexicana*, the close mitochondrial (maternal) ancestor of *P. formosa*.

The resulting time-calibrated Bayesian tree (Supplementary Figure 3) shows *P. formosa* to be diverged from *P. mexicana* about 97.4 kya (95% CI: 102.27-92.53 ky). The divergence time estimate for the paternal ancestor *P. latipinna* (ca. 5.78 Mya) from *P. mexicana* lies in the range (ca. 4-11 Mya) as also obtained independently by Bagley et al. ([28]: therein Figure S3), analyzing several fragments of mtDNA.

In parthenogenetic reptiles mitochondrial segmental duplications of several kilobases were detected [29-31] and connected to the unisexual reproduction. Either the clonal and hybrid nuclear background may cause cyto-nuclear incompatibilities resulting in improper mtDNA replication [30,31] or the two-fold effective population size of mtDNA in parthenogens relative to a same-sized sexual population may tolerate additional genetic diversity including such duplications[32]. Finally, absence of males, normally selecting for accurate mtDNA-transcription (e.g. by energy-demanding sperm), may increase the persistence of deleterious mtDNA in all-female clones [29]. Our analyses of mitochondrial genomes of *P. formosa* revealed the absence of large mitochondrial duplications, indicating accurate mtDNA replication. Thus, the mitochondrial genome of *P. formosa* – like the nuclear genome – does not show negative consequences of hybrid origin and asexual reproduction.

Supplementary Note 5: Gene conversion

To detect gene conversion events we analyzed all *P. formosa* samples for mutations at fixed different sites in the parental species. For each tentative gene conversion site, we counted the depth of sequencing coverage (Supplementary Figure 15). All the conversion sites in *P. formosa* demonstrate a coverage greater than 10, the genome-wide average coverage for sequenced *P. formosa* isolates. This observation suggests these sites experienced true conversion events, rather than genotypic change from heterozygosity to homozygosity due to deletion of one allele (LOH). Assuming an ancestral heterozygous genotype at a site (e.g., A/G), a mutation event at this site would lead to six possible genotypes (A/T, A/A, A/C, T/G, G/G, or C/G), whereas a gene conversion event would result in two possible genotypes (A/A or G/G). Within our high quality SNP dataset consisting of 53,175 fixed differences we detected 725 such single base homozygous sites. Thus, gene conversions rather than mutations are more likely to be responsible for the homozygous genotype for either parental alleles at these fixed difference sites in *P. formosa*.

Nonetheless, in contrast to the widely observed GC-biased conversion in eukaryotes [33], these gene conversion events are not biased towards G+C nucleotides at AT:GC heterozygous sites (268 AT→GC vs 301 GC→AT, chi square test p = 0.1665).

Within the fixed SNP differences between the parental species, we identified 706 tentative single-base conversion events that occurred in only one *P. formosa* isolate with at least 12 of the remaining isolates being heterozygous. We found homozygosity for 498 fixed sites that are present in more than 12 isolates and thus are assumed to be of ancient origin.

Considering the single F1 hybrid origin of *P. formosa,* such sites strongly indicate phases of selective sweep and/or a demographic bottleneck associated with genetic drift following gene conversion events during early evolution of *P. formosa* after it arose from hybridization of the parental species.

To obtain an estimate on the overall gene conversion rate we considered the 706 single-site conversion events from the total 53175 fixed-difference sites. Assuming an origin of *P. formosa* 100, 000 years ago (4 generation/year), this results in (706/53175)/(100,000 * 4) = 3.32 E-8 per site per generation.

Interestingly, gene conversion events are significantly biased towards the alleles originated from the maternal species *P. mexicana* (n=412, 58%) compared to paternal *P. latipinna* (n=294, 42%; chi-square test $p < 0.0001$).

To test the hypothesis that gene conversion is low in *P. formosa* due to the high heterozygosity of its hybrid genome we calculated Fst as an estimate of divergence between the two parental species for each scaffold, where we found conversion events (Supplementary Figure 16). We find no correlation of divergence and conversion event frequency, and thus no support for this explanation of paucity of gene conversions.

Supplementary Note 6: Paternal introgression

To determine paternal introgression we assessed allele imbalance at heterozygous sites across all scaffolds larger than 100Kb. We then studied the distribution of the reference allele frequency across samples and scaffolds. We identified putative paternal introgression events as bimodal distributions that result from the presence of an extra copy inherited from the male host, what makes the ratio of reads at heterozygous sites deviate from the theoretical 0.5. We confirmed that the predicted copy number throughout the affected scaffold was higher and continuous in the samples where the paternal introgression was detected. We did not observe any evident structural changes in the introgressed genomic material (Supplementary Figure 17). For further confirmation we determined whether the observed imbalance could be consistently traced back to the corresponding paternal species. Sequencing data from five *P. mexicana* and six *P. latipinna* was used to call SNPs. After filtering, we looked for fixed homozygous sites with opposing genotypes in the two species. Then these fixed homozygous sites were intersected with heterozygous variants called in each of the 20 *P. formosa* samples and for each site the percentage of reads carrying the allele present in *P. mexicana* was calculated. In the absence of paternal introgression, frequency should be ~50%. With introgression the imbalanced allele at heterozygous sites should match the allele present in the corresponding paternal species, and thus this frequency would deviate from 50%. In wild-caught isolates, we trace the introgressed sequence to the known male hosts being either *P. latipinna* or *P. mexicana*.

GO term analyses of introgressed genes revealed no significant (corrected p-value < 0.05) functional enrichment (Supplementary Tables 25-26).

To evaluate the relief from genomic decay allowed by paternal introgression, we need to distinguish the following potential scenarios, since we do not know which effect on fitness a particular paternally introgressed sequence will have. To keep this analysis simple, we will pretend that all paternal introgressions are either selectively neutral, deleterious, or advantageous in the sense that they repair preexisting deleterious effects that had been

caused by Muller's ratchet. The mechanistic scenario behind our model is as follows. We infer from the observed frequency of paternal introgressions in the natural population of about 25% (see above) that there's likely to be a steady state where paternal introgressions continue to occur and remain visible as B-chromosomes for some time until the B-chromosomes disappear again. Disappearance could mean complete loss without leaving a trace in the nuclear genome, or it could mean that the nuclear DNA has been replaced via gene conversion and thus a potentially beneficial effect of the paternal introgression has become permanent. We currently cannot determine which of these scenarios are most likely, and will therefore consider all of them. These scenarios suggest changing the parameter estimates that were presented in Loewe and Lamatsch 2008 as follows (previous values reused here, unless changes have been explicitly stated).

A. *No or neutral paternal introgression.* This plot is exactly as the corresponding plot in Loewe and Lamatsch 2008 for all parameter estimations, except that the overall age of the Amazon molly as indicated by the horizontal gray line at about 100,000 years in the past, has been adjusted to the results of this present paper. This plot is shown as a baseline with and without individual based simulation results (for comparison) and represents the case where either no paternal introgression occurs or all paternal introgression is effectively selectively neutral.
B. *Paternal introgression as exclusively deleterious.* Same parameter combination as for A; except that the effective population size N_e has been reduced by 25%, assuming that all Amazon molly in natural populations observed to carry B-chromosomes will ultimately go extinct, due to deleterious effects caused by these paternal introgressions. As argued elsewhere [34,35], the N_e most relevant for Muller's ratchet analyses is the N_e that results from removing all individuals that carry strongly deleterious mutations, which will ultimately be removed deterministically by background selection. This is the case even in asexual populations [36]. Thus, reducing N_e estimates by removing the fraction of individuals with potentially deleterious paternal introgressions seems to be the most appropriate approach.
C. *Paternal introgression as exclusively advantageous.* Same parameter combination as for A; leaving N_e unchanged, we now reduce the deleterious mutation rate U_{sdm}, assuming that there is a constant flux of paternally introgressing genes that repair previously existing deleterious mutations accumulated from Muller's ratchet. To estimate this flux, we assume a steady state quasi-equilibrium that results in the observed frequency of B-chromosomes and most importantly allows us to assume that the total number of introgression events per generation per individual is effectively as frequent as the overall loss of B-chromosomes per generation per individual. Applying a logic similar to that used for quantifying neutral evolution, we translate observations of 1 B-chromosome loss per 10 or 100 generations in the lab into assumed estimates of 0.1 or 0.01 paternal introgressions per individual per generation. It is difficult to see how paternal introgression rates could be much higher in natural environments. These rates do not directly translate into advantageous mutations, because only 1% or 0.1% of the genomic DNA is transferred in a single introgression event, and might not include the sites required for a particular repair. Thus, the probability that a particular introgression will repair a particular deleterious mutation is given by the ratio of the size of the paternal introgression over the size of the genome. As a result, the overall probability of repairing damage from the initial operation of Muller's ratchet in the Amazon molly is at most 0.001 per genome per generation – as inferred from an upper limit of 0.1 paternal introgressions per genome

per individual * 0.01 advantageous ratchet repairing introgressions out of all introgressions. The latter fraction of advantageous introgressions might increase substantially after many slightly deleterious mutations have accumulated throughout the genome, because then the probability of repairing at least 1 of them, will substantially increase. A full analysis of this and other related potential solutions to the genomic decay paradox in the Amazon molly (see [34]) certainly needs further in-depth analysis.

D. *Paternal introgression as mostly deleterious and occasionally advantageous.* Same parameter combination as in A; except that we now combine the cases B and C, and reduce N_e and the mutation rate U_{sdm} accordingly.

The corresponding U-shaped plots for the parameter estimations defined above can be found in Supplementary Figure 7, panels A - D. They show that paternal introgression has a rather small impact on the initial rate of Muller's ratchet in the Amazon molly. While this might change for Amazon molly genomes in advanced stages of genomic decay, it does not seem that our present understanding of paternal introgression reduces genomic decay processes so much that the absence of molecular signatures of decay as reported in this paper is easily explained by paternal introgression. This conclusion is not likely to be changed if a model of paternal introgression can be justified that merely modulates extinction time by some linear combination of introgression rate and introgressing genome size. The predicted U-shaped plots of such models merely move all curves slightly up the y-axis; the resulting effects are minor because the y-axis is logarithmic.

Supplementary Note 7: Heterozygosity analysis

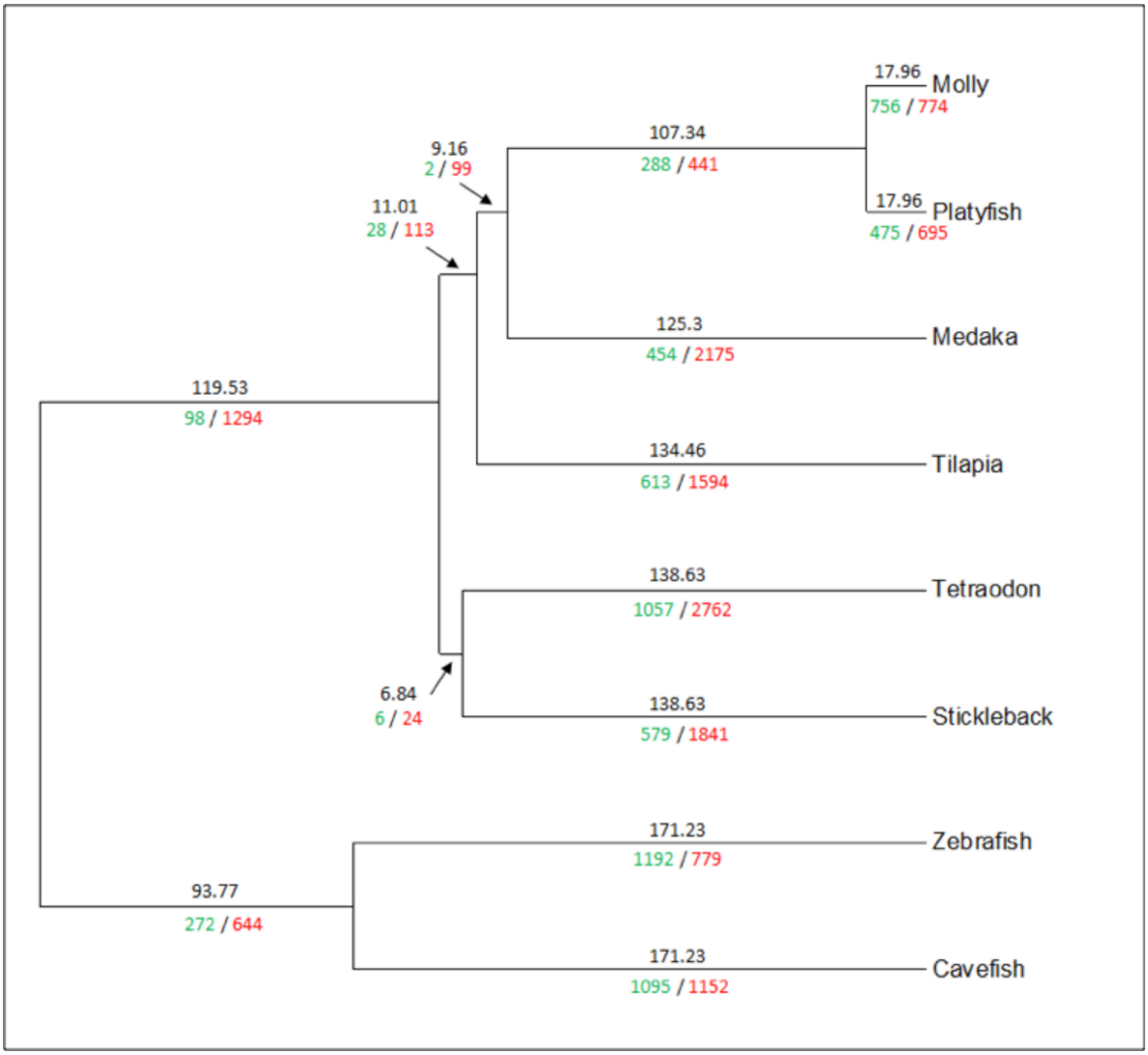
Based on our collection of 293,324 SNPs we find abundance of medium frequency variants and no excess of low and/or high frequency polymorphisms (Tajima's D P. formosa 2.37, P. mexicana 0.82, P. latipinna 0.75).

From the SNP dataset we calculated nucleotide diversity (π) within *P. formosa*, *P. mexicana* and *P. latipinna.* We find a higher π for *P. formosa* (0.324) than for *P. latipinna* (0.2) and *P. mexicana* (0.225). These values are higher than what is usually expected from whole genome level analysis, because only the variable sites, which make up the SNP dataset are considered. Although π is usually/conventionally measured as a "virtual heterozyogosity" assuming random mating, which is not applicable to the mating system of the Amazon molly, it provides another indication for the high heterozygosity maintained in the asexual species.

Using the same SNP dataset, we reconstructed haplotypes for all sequenced isolates of *P. formosa* and the parental species. Taking advantage of the large amount of fixed nucleotide differences between the two parental species, the genotypes for *P. formosa* at such fixed different sites can be accordingly phased into two parental haplotypes, whereas the genotypes at other sites, where fixed difference is not detected, are phased randomly. For the two parental species, the alleles at a site are randomly phased to reconstruct two haplotypes. A neighbor-joining phylogenetic analysis with 100 bootstrap replicates revealed two strongly supported clades, one clade consisting of parental *P. latipinna* haplotypes and *P. formosa* haplotypes derived from *P. latipinna* and the other clade containing parental *P. mexicana* haplotypes and *P. formosa* haplotypes derived from *P. mexicana* (Supplementary Figure 4).

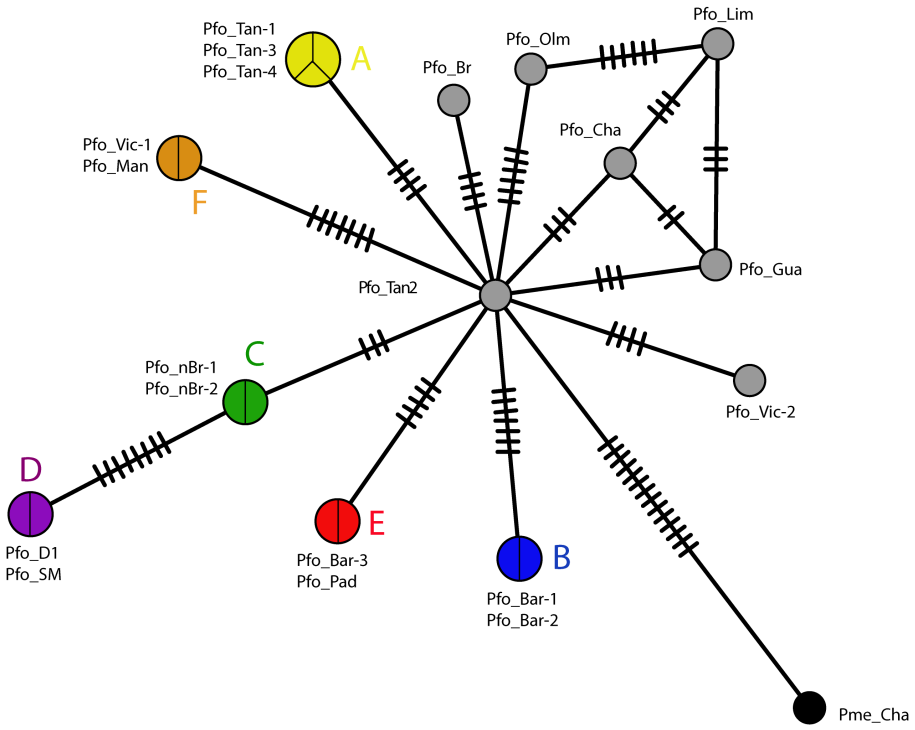
Supplementary Figures

Supplementary Figure 1



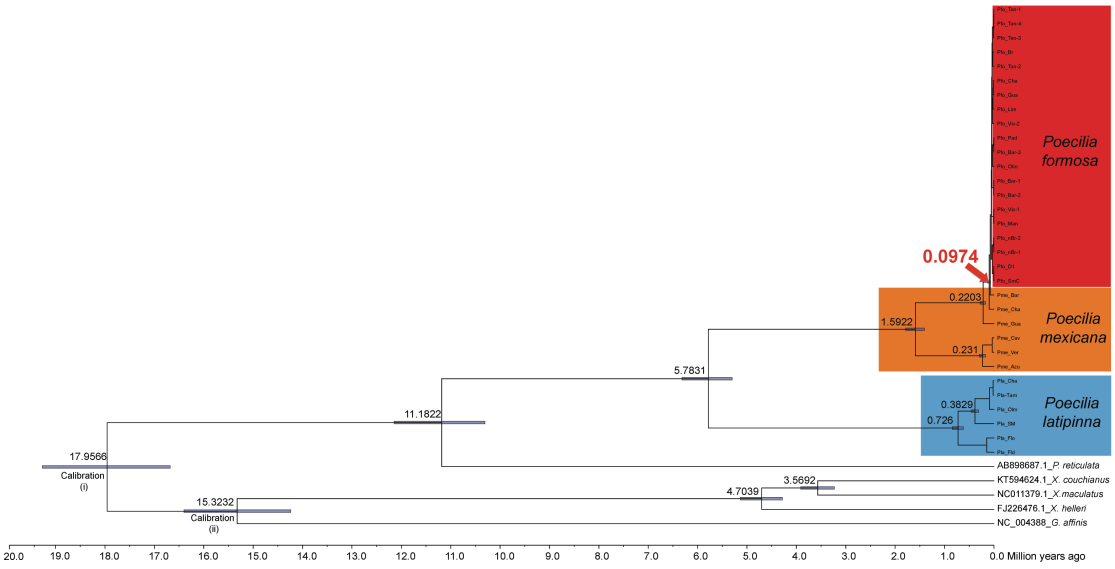
Ultrametric tree with branch lengths in millions of years for the 8 fish species used in this study. Branch lengths appear above the branch while the number of families that have expanded (green) and contracted (red) along that branch appear below it.

Supplementary Figure 2



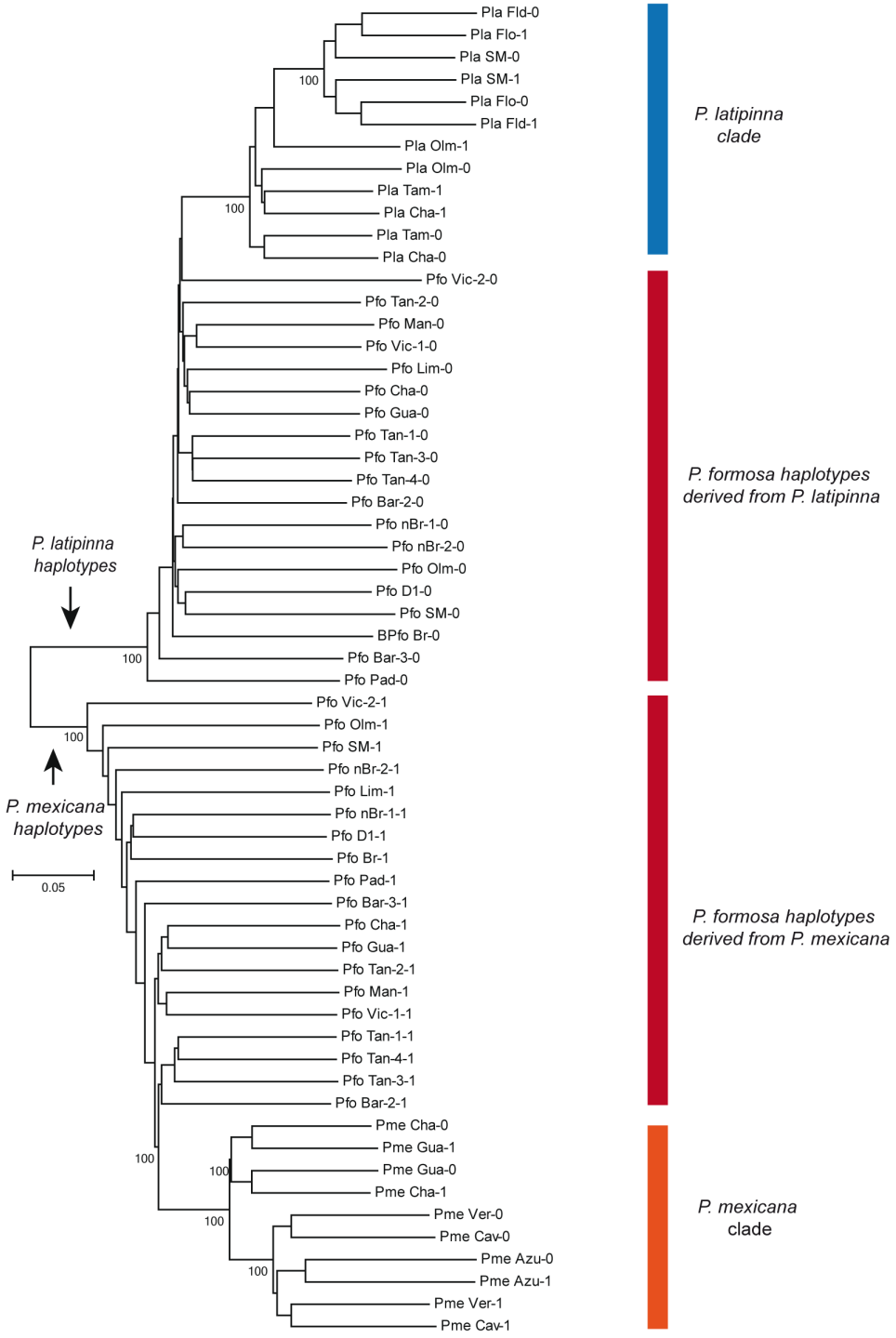
Full length (16,522bp alignment) mitochondrial haplotype Neighbor Joining network. All *P. formosa* (Pfo) haplotypes are shown. As a reference the closest related *P. mexicana* (Pme) haplotype is given. Grey haplotypes were unique, colored haplotypes (large letters) were shared between individuals. Letters indicate field site: Bar – Rio Purificacion (Barretal), Br – Brownsville, nBr - laboratory strain (originally from Brownsville area), Cha - Laguna Champayan, D1 – Brownsville area (Ditch 1), Gua – Rio Guayalejo (Emiliano Zapata), Lim - Rio Guayalejo (El Limon), Man - Mante, Olm – Brownsville area (Olmito), Pad - Rio Purificacion (Nuevo Padilla), SM - San Marcos, Tan - Taninul, Vic - Ciudad Victoria.

Supplementary Figure 3



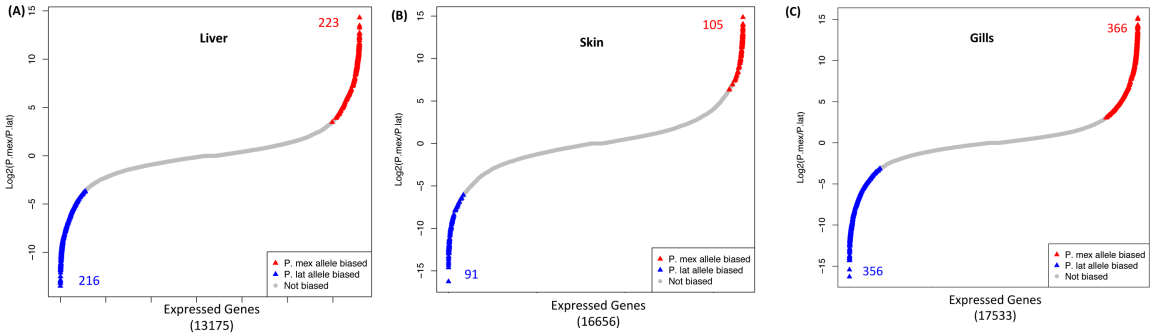
Time-calibrated Bayesian phylogenetic tree obtained with BEAST v.1.8.3. Using independent previously published minimum age estimates for divergence of *Gambusia affinis* from *Xiphophorus hellerii* and the divergence of both from *Poecilia reticulata* for calibration the origin of *Poecilia formosa* is dated to at least 100 000 years ago. Bluish bars on nodes represent 95% confidence intervals.

Supplementary Figure 4



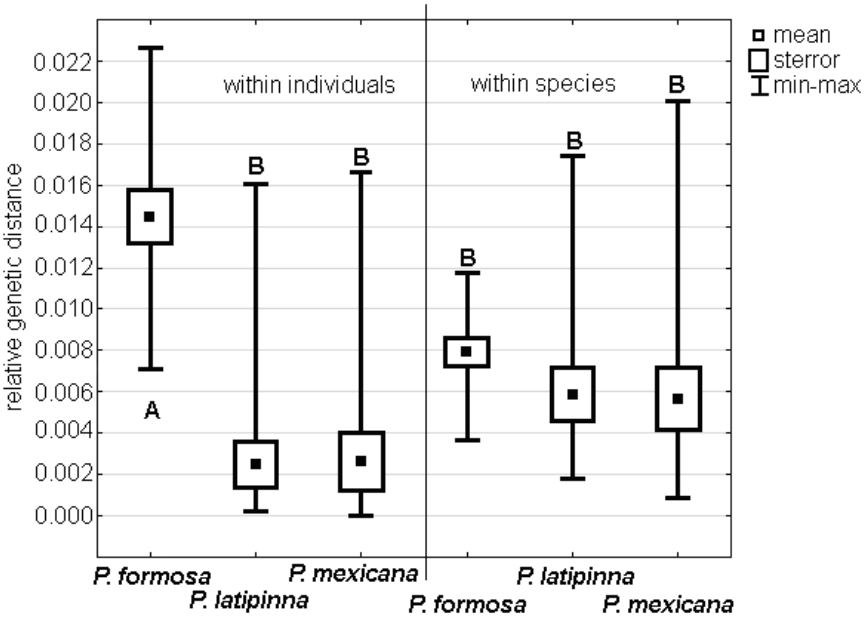
Neighbor-joining phylogeny (unrooted) of the reconstructed haplotypes of *P. formosa*, *P.latipinna,* and *P.mexicana* using uncorrected nucleotide divergence (292,324 sites). Numbers next to the tree represent bootstrap values for major clades. Scale bar indicates 0.05 divergence per site. The notation for each haplotype consists of the sample name followed by either 0 or 1 that indicate two haplotypes within each isolate, respectively. Phylogenetic tree reconstruction using maximum likelihood yielded the same topology.

Supplementary Figure 5

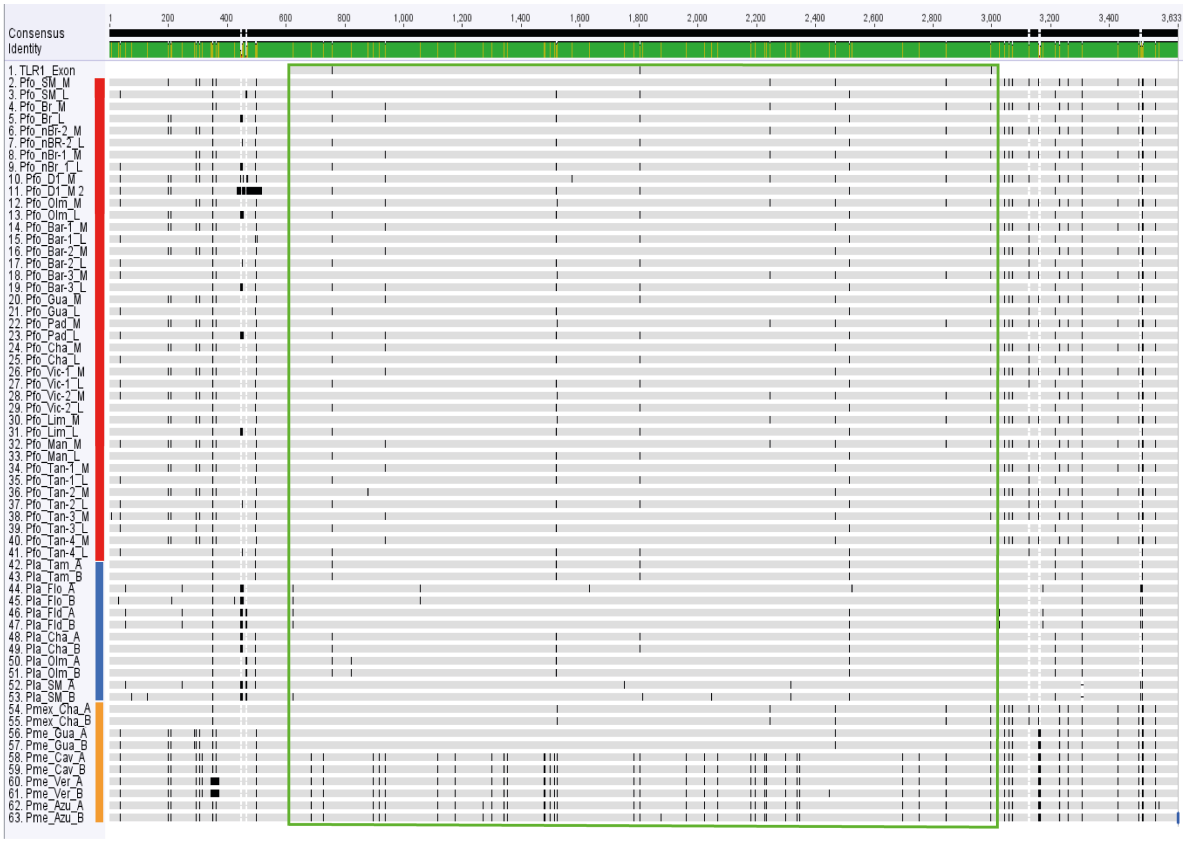


Allele specific gene expression in *P. formosa* liver (A), skin (B) and gills (C). Numbers in blue are genes with expression biased for the P. latipinna allele and number is red are genes with expression biased for the P. mexicana allele.

Supplementary Figure 6

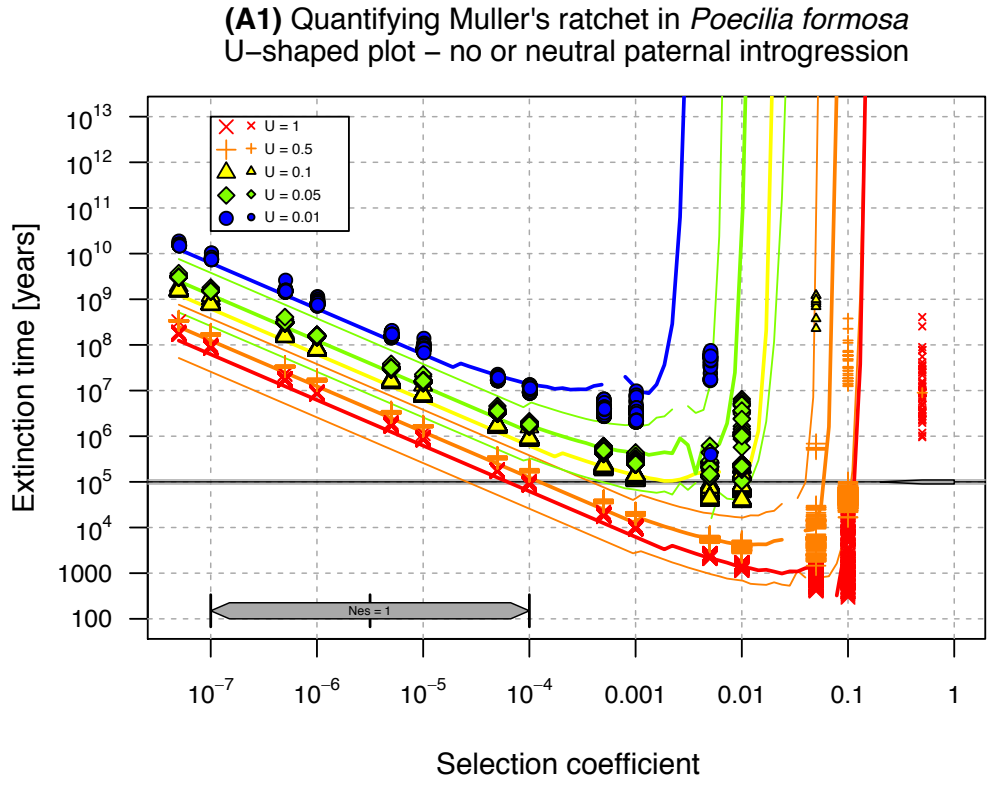


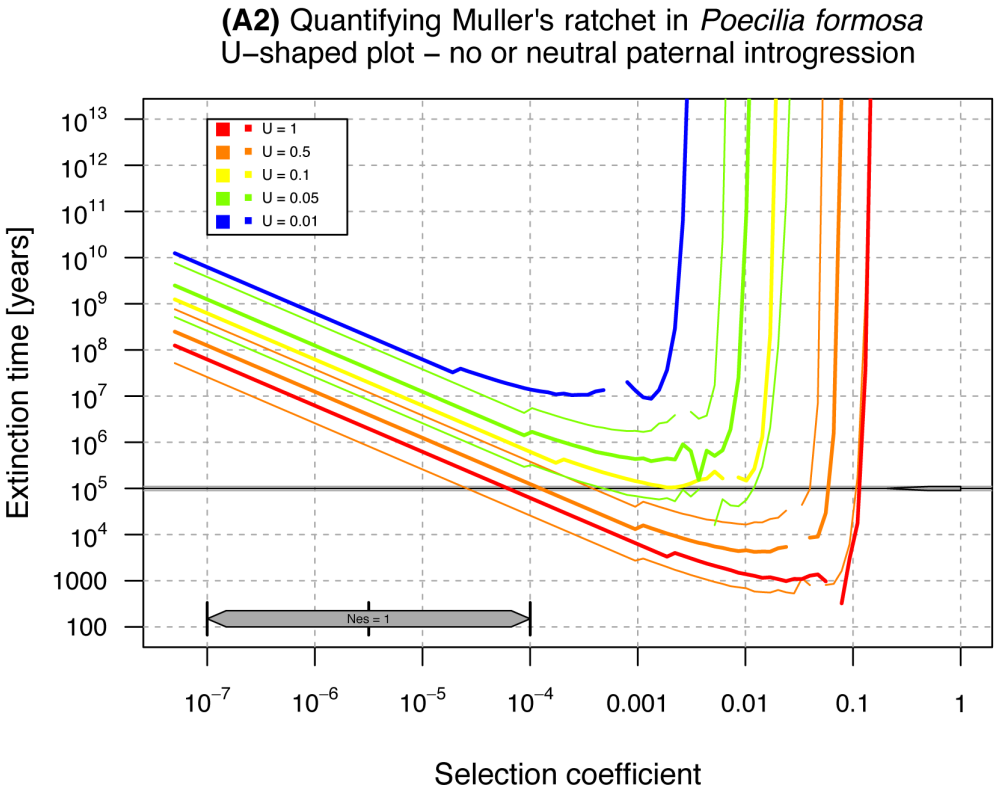
(6A) Box plot of genetic distances in immune genes for all Poecilia species. Mean values for 15 immune genes (see Supplementary Table 20), standard error (sterror) and minimum and maximum values are given. Distances were calculated within individuals and within species (among and within individuals). Different letters (A;B) indicate significant differences based on ANOVA analyses (SQ = 0.001, df = 5, MQ = 0.000, F = 12,63, $p<0.0001$) and Scheffé Post hoc test.



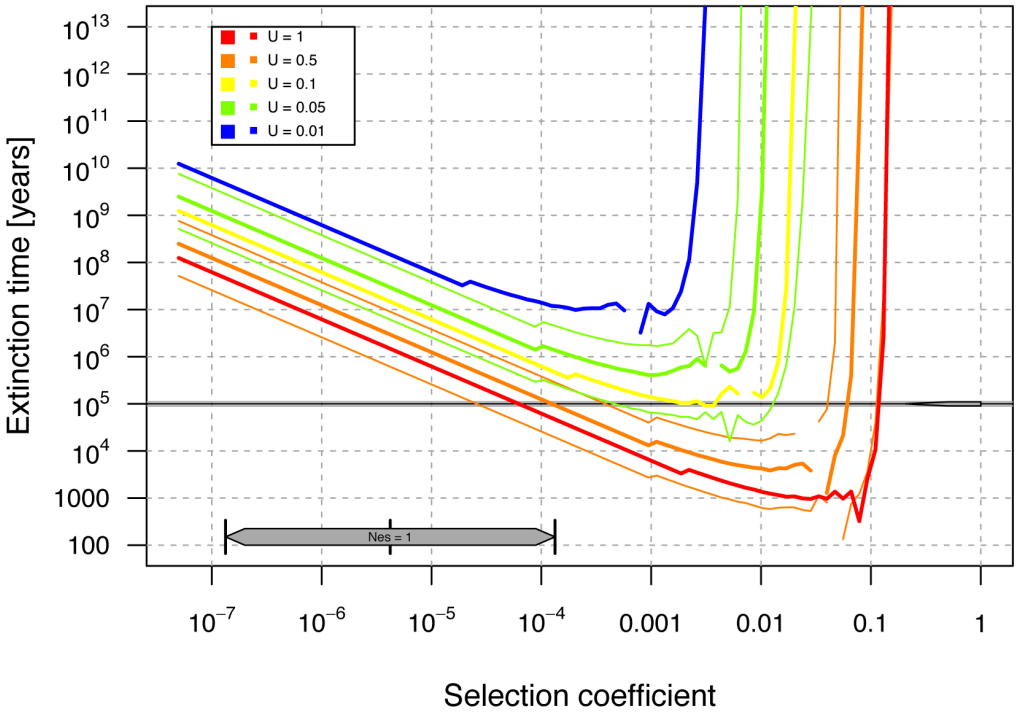
(6B) Full length alignment (exon1 (marked with a green square) and flanking sequences) of the *tlr1* alleles in all Poecilia isolates investigated. Variable sites are shown as black lines. Colored bars indicate species (*P. formosa* - red, *P. latipinna* - blue, *P. mexicana* - orange). *P. formosa* alleles are marked L or M depending on *P. mexicana* (M) or *P. latipinna* (L) origin. Parental alleles in *P. mexicana* and *P. latipinna* are marked as A and B. Individual names are consistent with other graphs and indicate species and origin.

Supplementary Figure 7

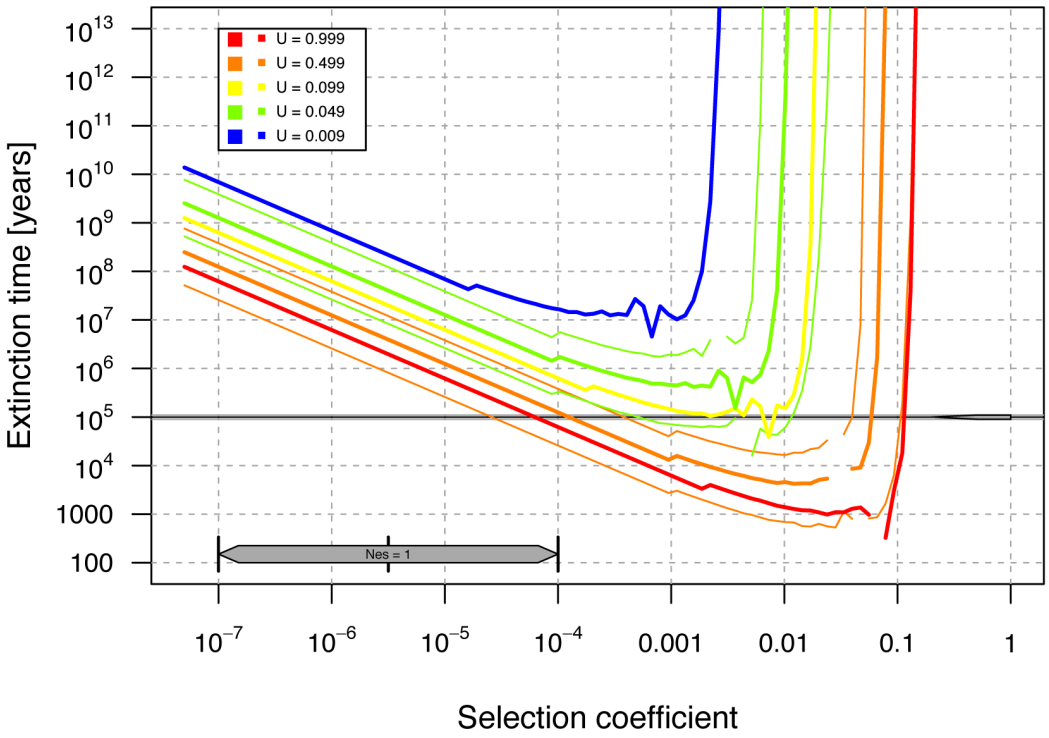


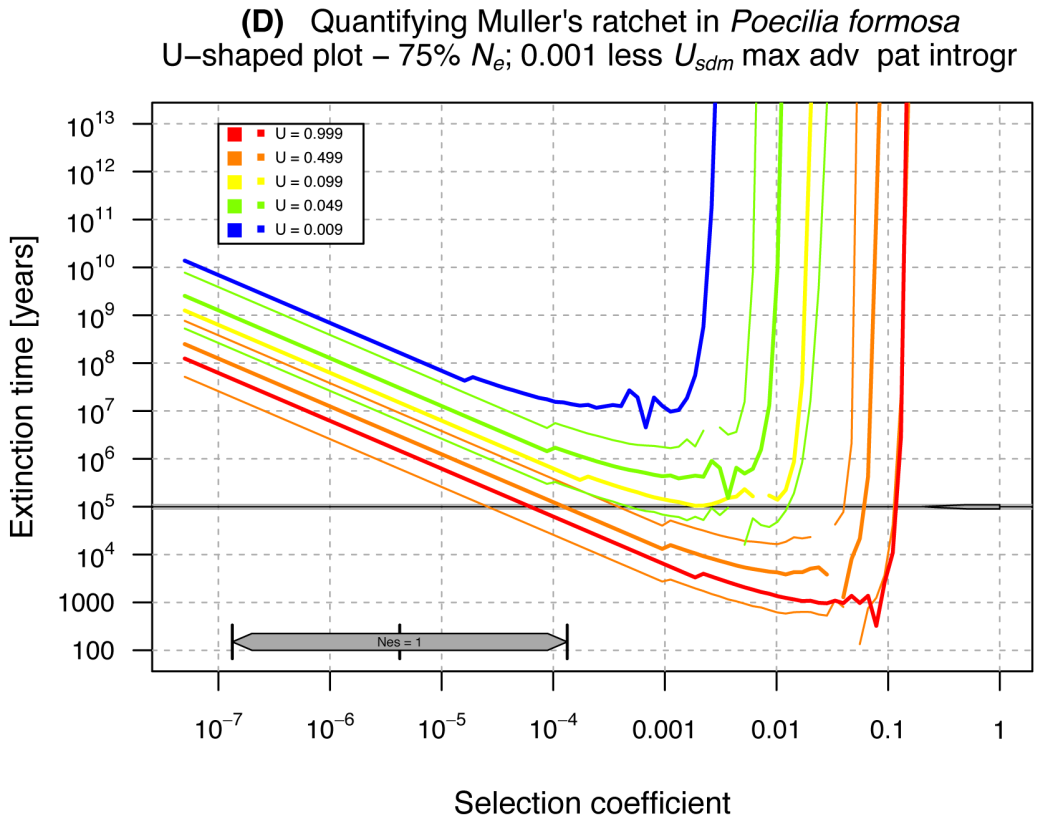


(B) Quantifying Muller's ratchet in *Poecilia formosa*
U–shaped plot – 75% N_e from deleterious paternal introgression



(C) Quantifying Muller's ratchet in *Poecilia formosa*
U–shaped plot – 0.001 less U_{sdm} from max adv pat introgr



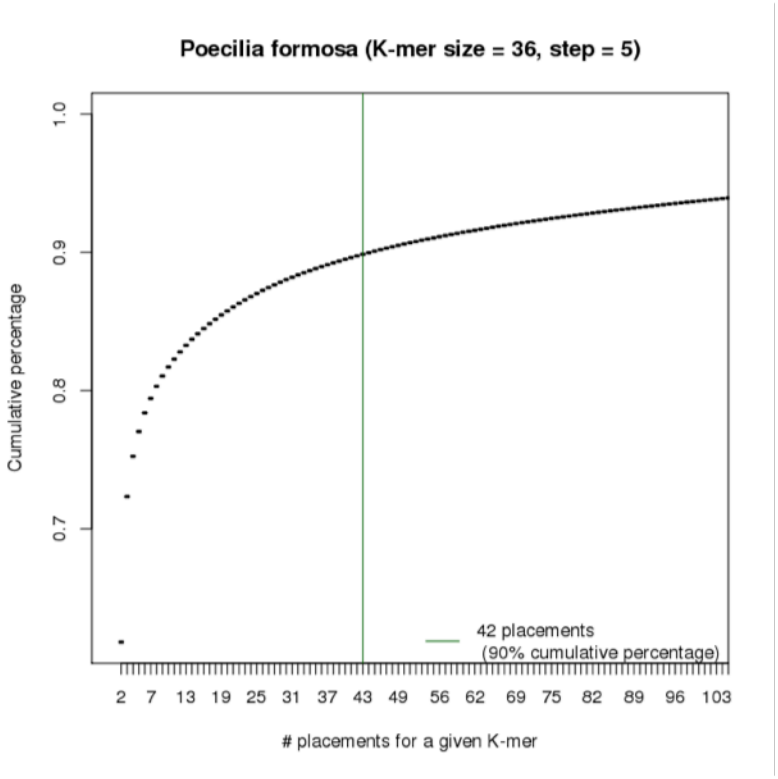


U-Shaped plots quantifying the impact of Muller's ratchet in the presence of paternal introgression in the Amazon molly. U-shaped plots are a way to quantify the danger of extinction caused by Muller's ratchet in asexual populations. Briefly, they use a multiplicative fitness model and an assumed maximal reproduction capacity to predict the number of clicks Muller's ratchet might need to drive the population to extinction by mutational meltdown. This number of clicks is then multiplied with the time it takes Muller's ratchet to click. This click time is predicted from mutation rate (U_{sdm}) the selection coefficient (s, on x-axis), and the effective population size (N_e, indicated indirectly on the x-axis as $1/s$). Given the many uncertainties associated with these analyses, lower and upper limits are used for all respective values; thus, a range of s, U_{sdm} values need to be considered for appropriately interpreting this plot. For more details, see [34]. The five panels shown focus on the main topic of this study by exploring the four potential roles of paternal introgression that are discussed next. In addition to the analytic predictions shown for all other panels, the first panel also shows related individual-based simulation results to clarify expectations for the range of parameter values where the analytic predictions break down.

(A) Basic scenario of no introgression or purely neutral introgression effects. Plots correspond to [35] except for adjustments in the age of the Amazon molly. Predicted extinction time (T_{ex}) (estimated as in [35]) from the largest to the smallest conceivable effective population size, N_e (10^7 to 10^4). Lines indicate the predicted extinction time for different deleterious genomic mutation rates (U_{sdm}) with N_e=316,000, generation time T_{gen}=1 yr and maximal reproductive capacity R_{max}=500 offspring/generation. The thin orange (U_{sdm}=0.5) and green (U_{sdm}=0.05) lines represent the variability of the extinction time for the lower and upper credible deleterious mutation rate estimates as taken from [35] (U_{sdm}=0.05, 0.5) using the corresponding upper and lower limits for N_e (10^7 ,10^4; see bar indicating the range of s values that correspond to the populations N_e values and are chosen such that the product $N_e * s = 1$), T_{gen} (2.5, 0.25) and R_{max} (2,000, 50). The horizontal grey line marks the age of the Amazon molly lineage (100,000 years). (A1) This panel was

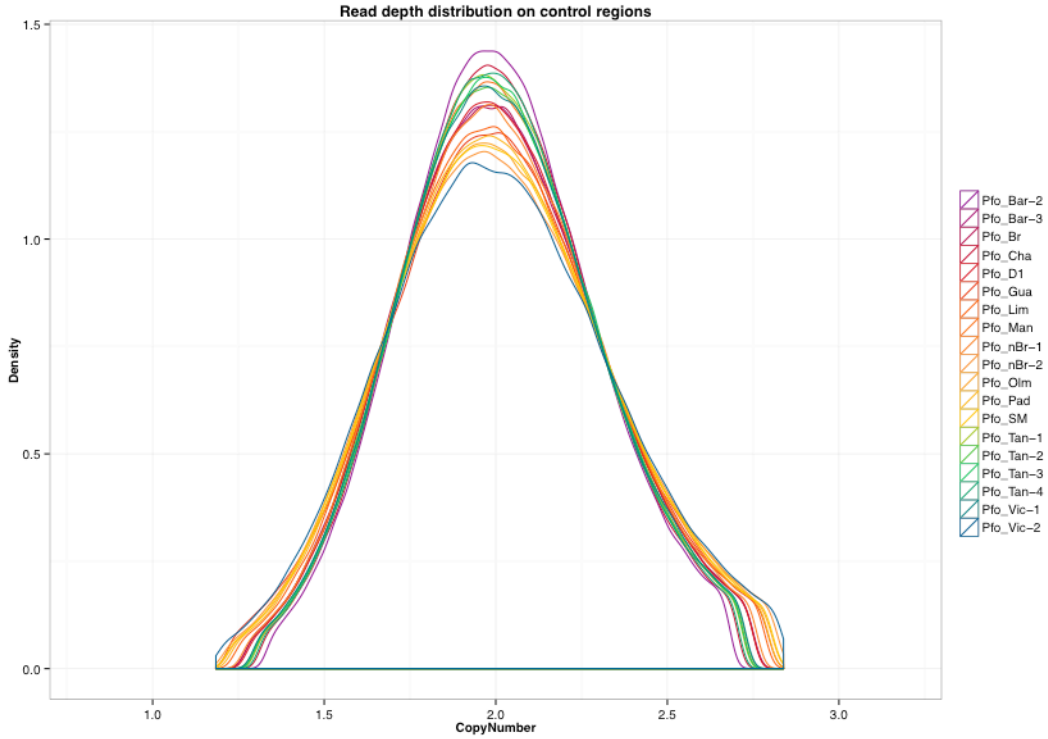
generated by parameter combination A as described above; it includes all evolution@home individual-based simulation results, which were also shown in [35] and bring more certainty to the most interesting intermediate region where analytical predictions break down. (A2) Same plot as (A1), except without the individual-based simulation results, in order to improve visual clarity make it easier to compare this parameter combination to the near identical results for the other parameter combinations shown in the subsequent plots. (B) Paternal introgression effects when exclusively deleterious; see online methods for details. (C) Paternal introgression effects when exclusively advantageous during the initial stages of Muller's ratchet operation; see online methods for details. (D) Combination of the scenarios B and C.

Supplementary Figure 8



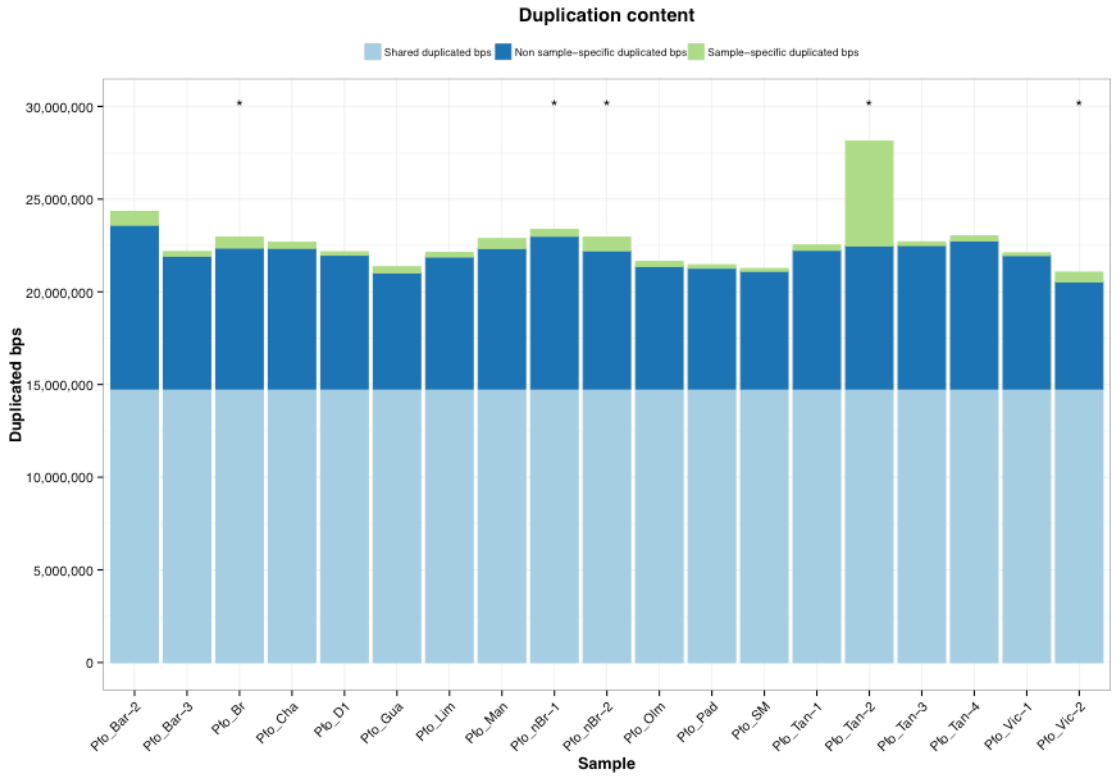
Cumulative distribution of 36 base pairs kmers mapping approach applied to detect potentially hidden repeats.

Supplementary Figure 9



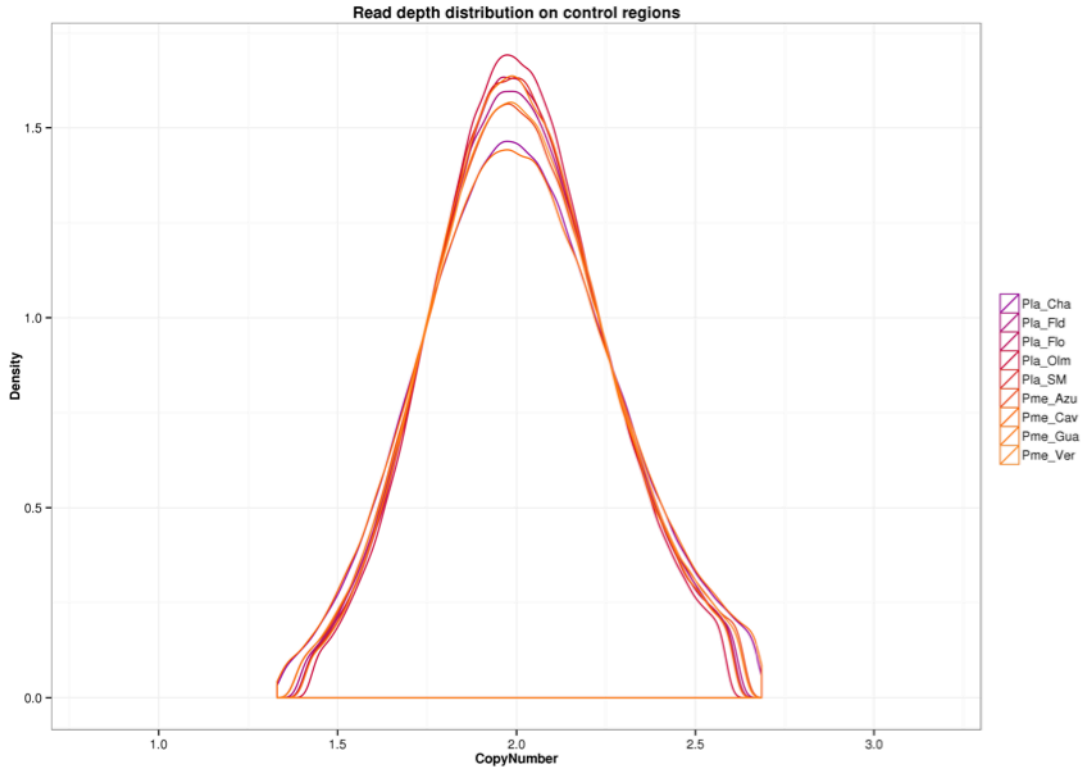
Copy number distribution in control regions for *P. formosa* samples

Supplementary Figure 10



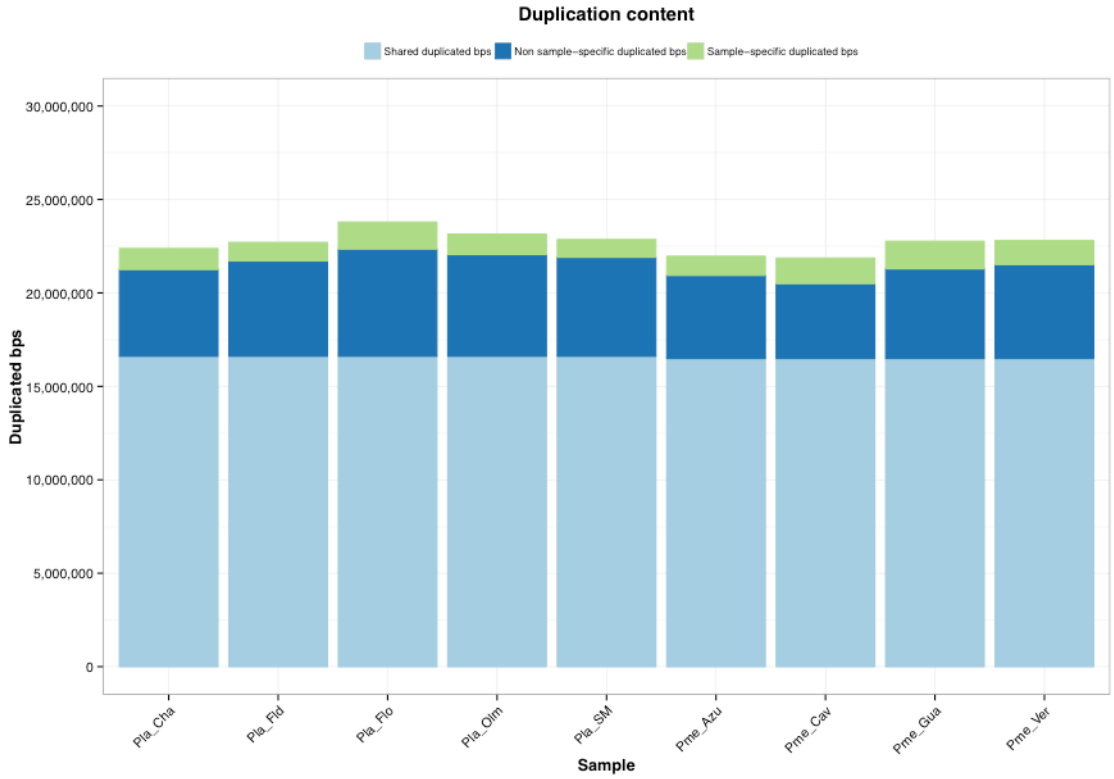
Per sample duplication content in the *P. formosa* samples. Asterisks mark the samples with paternally introgressed scaffolds.

Supplementary Figure 11



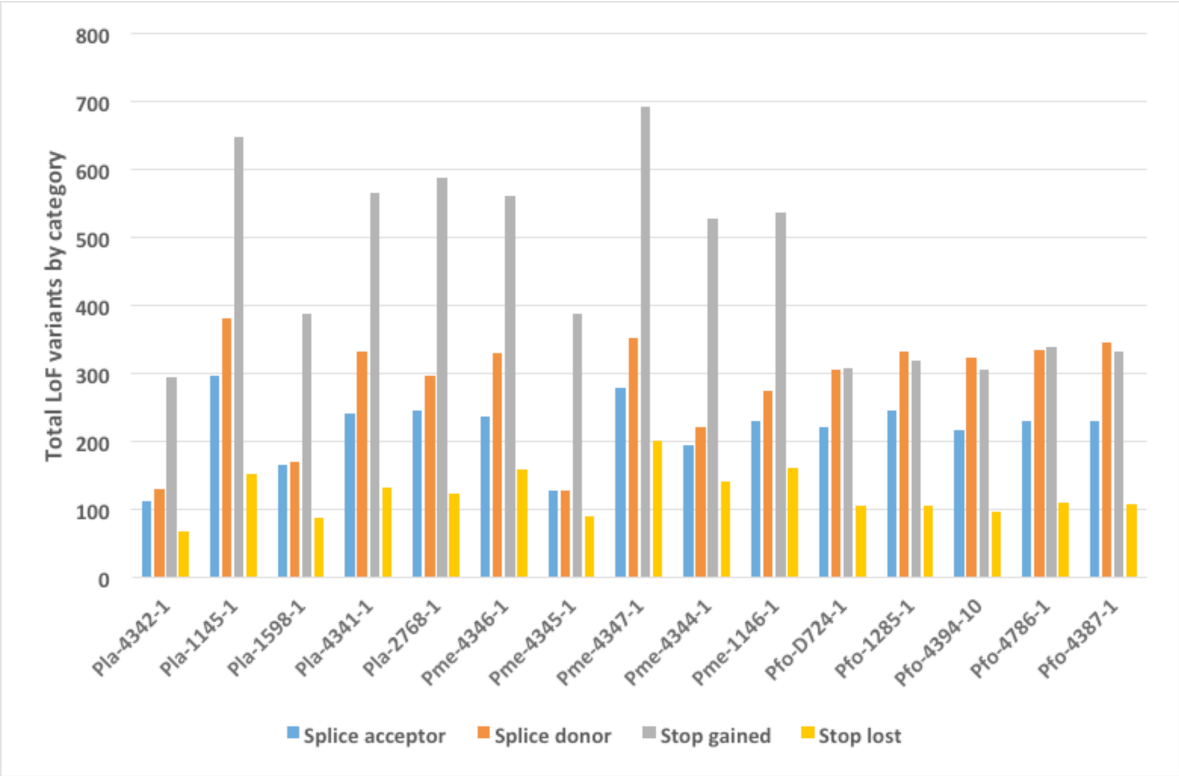
Copy number distribution in control regions of *P. latipinna* and *P. mexicana* samples.

Supplementary Figure 12

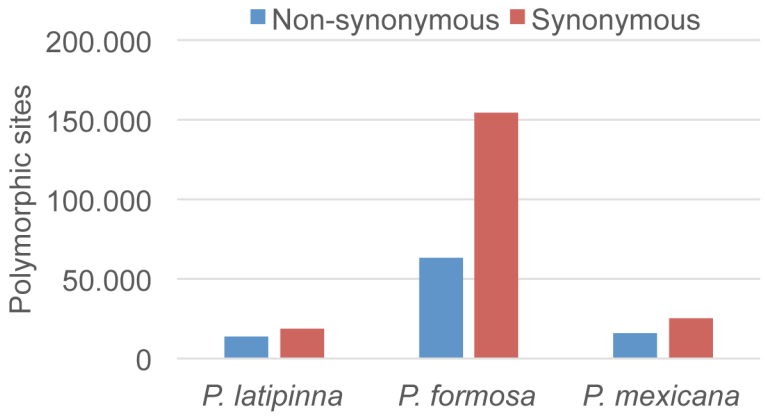


Per sample duplication content in the *P. latipinna* and *P. mexicana* samples.

Supplementary Figure 13

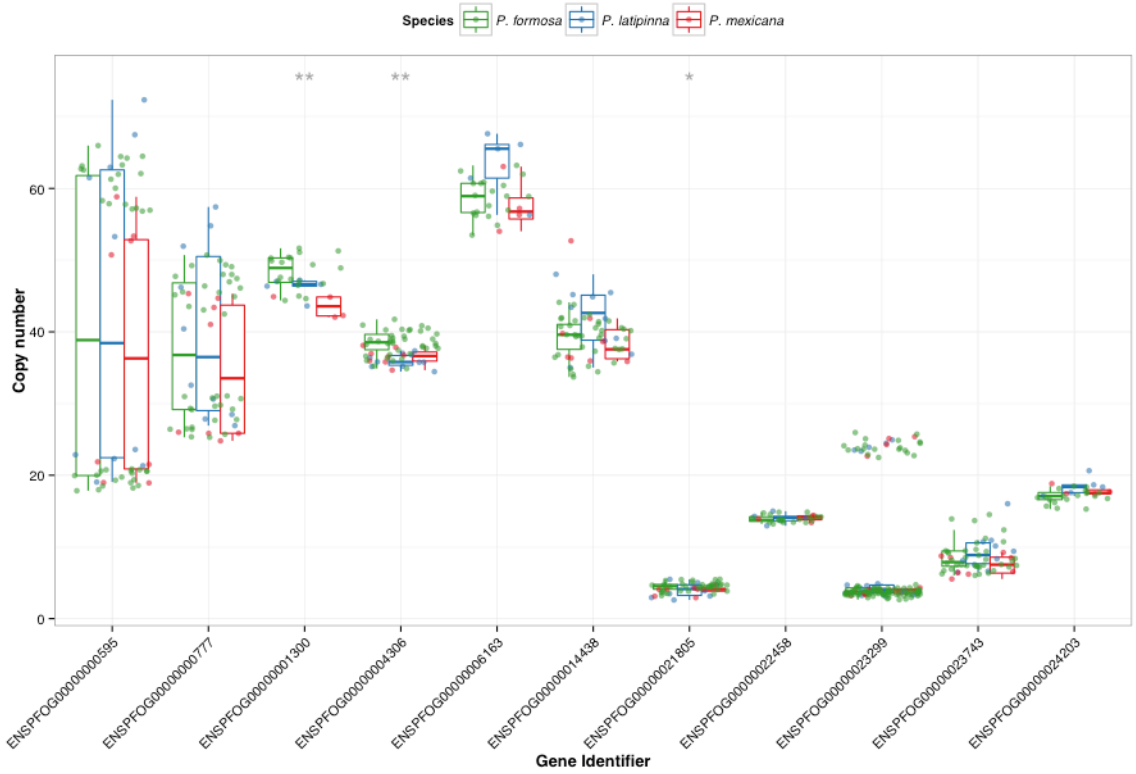


(A) Total LoFs per individual across species by variant classification. All LoF variants shared across all samples for the parental genomes were removed as these likely constitute false positives due to frameshifted gene models due to indels and very low likelihood these variants were fixed in the diverse populations selected.



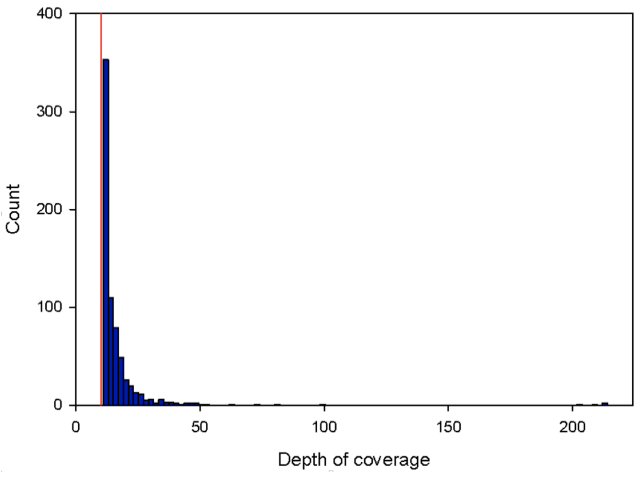
(B) Total non-synonymous and synonomous variants estimated within the asexual and sexual parental genomes.

Supplementary Figure 14



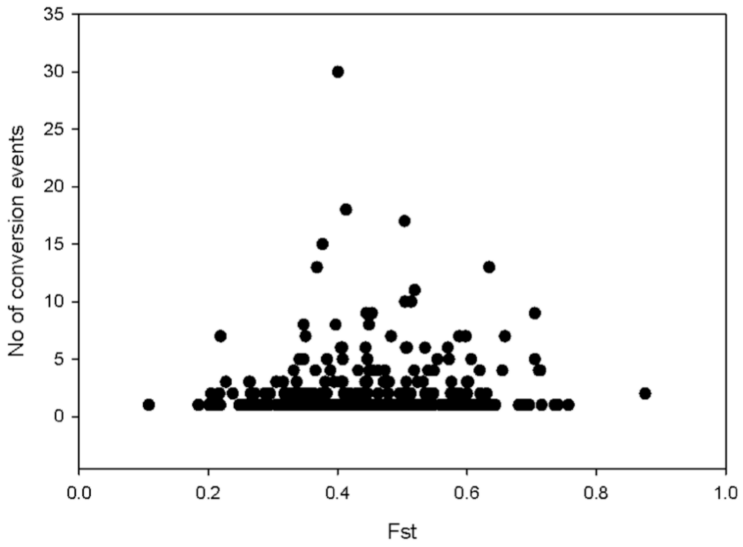
Expansion of absolute copy number in the 11 genes related to GO terms “DNA integration” (GO:0015074), “transposition, DNA-mediated” (GO:0006313), “transposase activity” (GO:0004803) and “DNA recombination” (GO:0006310) in *P. formosa* relative to the parental species.

Supplementary Figure 15



Histogram of sequence coverage of tentative gene conversion events in *P. formosa*. The red vertical line indicates the 10x sequence coverage.

Supplementary Figure 16



Plot of number of conversion events per scaffold vs Fst.

Supplementary Figure 17

Copy number in 1kb repeat-free windows along scaffold KI519701.1. Each panel corresponds to a different sample. Each dot represents the copy number in one 1kb-repeat-free window. Tracks for gaps, repeats, genes and segmental duplications (SDs) are included at the top of each panel. In grey, the samples with no introgression are shown. In teal, the sample (Pfo_Tan-2) carrying the introgressed scaffold (KI519701.1) is depicted. We observed an increased and constant copy number (~3) throughout the entire scaffold (KI519701.1) in sample Pfo_Tan-2, with no evidence of duplications/deletions. The same result was obtained for all other introgressed scaffolds.

References

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

2 Volff, J. N., Bouneau, L., Ozouf-Costaz, C. & Fischer, C. Diversity of retrotransposable elements in compact pufferfish genomes. *Trends Genet* **19**, 674-678, doi:10.1016/j.tig.2003.10.006 (2003).

3 Bennetzen, J. L. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev* **15**, 621-627, doi:10.1016/j.gde.2005.09.010 (2005).

4 Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res* **42**, D749-755, doi:10.1093/nar/gkt1196 (2014).

5 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

6 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).

7 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113, doi:10.1186/1471-2105-5-113 (2004).

8 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:10.1093/bioinformatics/btl446 (2006).

9 Lapointe, F. J. & Cucumel, G. The average consensus procedure: Combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology* **46**, 306-312, doi:Doi 10.2307/2413625 (1997).

10 Criscuolo, A., Berry, V., Douzery, E. J. & Gascuel, O. SDM: a fast distance-based approach for (super) tree building in phylogenomics. *Syst Biol* **55**, 740-755, doi:10.1080/10635150600969872 (2006).

11 Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301-302 (2003).

12 Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971-2972, doi:10.1093/bioinformatics/btl505 (2006).

13 Han, M. V., Thomas, G. W., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* **30**, 1987-1997, doi:10.1093/molbev/mst100 (2013).

14 Ohno, S. *et al.* Microchromosomes in holocephalian, chondrostean and holostean fishes. *Chromosoma* **26**, 35-40 (1969).

15 Demuth, J. P. & Hahn, M. W. The life and death of gene families. *Bioessays* **31**, 29-39, doi:10.1002/bies.080085 (2009).

16 Curry, T. E., Jr. & Osteen, K. G. The matrix metalloproteinase system: changes, regulation, and impact throughout the ovarian and uterine reproductive cycle. *Endocr Rev* **24**, 428-465, doi:10.1210/er.2002-0005 (2003).

17 Hulboy, D. L., Rudolph, L. A. & Matrisian, L. M. Matrix metalloproteinases as mediators of reproductive function. *Mol Hum Reprod* **3**, 27-45 (1997).

18 McGinnis, W., Levine, M. S., Hafen, E., Kuroiwa, A. & Gehring, W. J. A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes. *Nature* **308**, 428-433 (1984).

19 Scott, M. P. & Weiner, A. J. Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. *Proc Natl Acad Sci U S A* **81**, 4115-4119 (1984).
20 Schuh, M. & Ellenberg, J. Self-organization of MTOCs replaces centrosome function during acentrosomal spindle assembly in live mouse oocytes. *Cell* **130**, 484-498, doi:10.1016/j.cell.2007.06.025 (2007).
21 Miller, M. P., Unal, E., Brar, G. A. & Amon, A. Meiosis I chromosome segregation is established through regulation of microtubule-kinetochore interactions. *Elife* **1**, e00117, doi:10.7554/eLife.00117 (2012).
22 Davydenko, O., Schultz, R. M. & Lampson, M. A. Increased CDK1 activity determines the timing of kinetochore-microtubule attachments in meiosis I. *J Cell Biol* **202**, 221-229, doi:10.1083/jcb.201303019 (2013).
23 Ruiz, E. J., Hunt, T. & Nebreda, A. R. Meiotic inactivation of Xenopus Myt1 by CDK/XRINGO, but not CDK/cyclin, via site-specific phosphorylation. *Mol Cell* **32**, 210-220, doi:10.1016/j.molcel.2008.08.029 (2008).
24 Porter, I. M., Schleicher, K., Porter, M. & Swedlow, J. R. Bod1 regulates protein phosphatase 2A at mitotic kinetochores. *Nat Commun* **4**, 2677, doi:10.1038/ncomms3677 (2013).
25 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969-1973, doi:10.1093/molbev/mss075 (2012).
26 Helmstetter, A. J. *et al.* Viviparity stimulates diversification in an order of fish. *Nat Commun* **7**, 11271, doi:10.1038/ncomms11271 (2016).
27 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**, 772, doi:10.1038/nmeth.2109 (2012).
28 Bagley, J. C. *et al.* Assessing species boundaries using multilocus species delimitation in a morphologically conserved group of neotropical freshwater fishes, the Poecilia sphenops species complex (Poeciliidae). *PLoS One* **10**, e0121139, doi:10.1371/journal.pone.0121139 (2015).
29 Fujita, M. K., Boore, J. L. & Moritz, C. Multiple origins and rapid evolution of duplicated mitochondrial genes in parthenogenetic geckos (Heteronotia binoei; Squamata, Gekkonidae). *Mol Biol Evol* **24**, 2775-2786, doi:10.1093/molbev/msm212 (2007).
30 Moritz, C. & Brown, W. M. Tandem duplications in animal mitochondrial DNAs: variation in incidence and gene content among lizards. *Proc Natl Acad Sci U S A* **84**, 7183-7187 (1987).
31 Zevering, C. E., Moritz, C., Heideman, A. & Sturm, R. A. Parallel origins of duplications and the formation of pseudogenes in mitochondrial DNA from parthenogenetic lizards (Heteronotia binoei; Gekkonidae). *J Mol Evol* **33**, 431-441 (1991).
32 Hartl, D. & Clark, A. *Principles of population genetics*. (Sinauer Associates, 2007).
33 Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**, 285-311, doi:10.1146/annurev-genom-082908-150001 (2009).
34 Loewe, L. Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA. *Genet Res* **87**, 133-159, doi:10.1017/S0016672306008123 (2006).
35 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, doi:10.1186/1471-2148-8-88 (2008).
36 Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-1303 (1993).

Supplementary Tables

Supplementary Table 1. A summary of assembly metrics

Species	Estimated genome size[2]	Assembled genome size	Estimated gap size	N50 contig (kb)[3]	N50 scaffold (Mb)[3]	Gene count
P. formosa	830 Mb	714 Mb	34 Mb	57	1.57	23,615[4]; 25,474
P. latipinna	970 Mb	680 Mb	135 Mb	33	0.28	25,220
P. mexicana	960 Mb	680 Mb	121 Mb	39	0.27	25,341
P. reticulata[1]	770 Mb	664 Mb	67 Mb	42	5.2	22,897

[1] All guppy *Poecilia reticulata* assembly or gene annotation metrics are derived from the NCBI assembly or Refseq databases.

[2] The estimated genome size for each species was derived from the average of individual estimates quoted in the literature and summarized in the animal genome size database (http://www.genomesize.com/index.php).

[3] N50 contig or scaffold length is defined as the length N for which 50% of all bases in the assembly sequences L are in a sequence of length N >L.

[4] Ensembl gene counts; all others are NCBI origins.

Supplementary Table 2. Read types

Library type	Species	Estimated coverage	Read length	Instrument
200bp fragment	*P. formosa*	39x	100bp	HiSeq2000
2kb paired	*P. formosa*	41x	100bp	HiSeq2000
4kb paired	*P. formosa*	7x	100bp	HiSeq2000
40kb paired	*P. formosa*	0.1x	100bp	HiSeq2000
350 bp fragment	*P. latipinna*	25x	100bp	HiSeq2000
3kb paired	*P. latipinna*	12x	100bp	HiSeq2000
350 bp fragment	*P. mexicana*	30x	100bp	HiSeq2000
3kb paired	*P. mexicana*	14x	100bp	HiSeq2000

Supplementary Table 3. A summary of sample details

Sample	Location	Geographic	Sympatric host/allopatry	GPS	SRA Exp Acc	ELC % Dupli–cation	Total Bases Kb (PF)	Fold coverag e (950Mb)	% >Q30 (R1)	% >Q30 (R2)
Pfo_Bar-1	big river, high population density	Rio Purification near Barretal	*P. mexicana*	N 24°02.848´ W 98° 22.264´	Assembly GCF_000485575.1			95		
Pfo_Tan-1	sulfur creek	Taninul, Sulfur Creek	*P. mexicana*	N 21°56.363´ W 98° 53.296´	SRX1009698	3,39	10689044	11,25	0,91	0,84
Pfo_Tan-2	sulfur creek	Taninul, Sulfur Creek	*P. mexicana*	N 21°56.363´ W 98° 53.296´	SRX1009697	2,43	8592377	9,04	0,91	0,86
Pfo_Tan-3	sulfur creek	Taninul, Sulfur Creek	*P. mexicana*	N 21°56.363´ W 98° 53.296´	SRX1009696	2,17	8640179	9,09	0,92	0,87
Pfo_Tan-4	sulfur creek	Taninul, Sulfur Creek	*P. mexicana*	N 21°56.363´ W 98° 53.296´	SRX1009699	2,43	9807571	10,32	0,91	0,85
Pfo_Vic-1	small ditch	Higway 80, km 105 north of Ciudad Mante	*P. mexicana, P. latipunctata*	N 22° 48' 44.280" W 99° 0' 45.000"	SRX1009694	2,40	9018659	9,49	0,92	0,87
Pfo_Cha	lake, still water, coastal area, polluted,medium population density	Laguna Champaxan	*P. mexicana, P. latipinna*	N 22°23.425´ W 97° 55.828´	SRX1009695	2,61	9562254	10,07	0,92	0,88
Pfo_Gua	fast flowing small river, low population density	Rio Guayalejo	*P. mexicana*	N 23°16.624´ W 98° 56.315´	SRX1009693	2,94	10640615	11,20	0,92	0,86
Pfo_Bar-2	big river, high population density	Rio Purification near Barretal	*P. mexicana*	N 24°02.848´ W 98° 22.264´	SRX1009700	2,86	10153587	10,69	0,91	0,87
Pfo_D1	small ditch	Ditch1	*P. latipinna*	N 25° 59' 11.400" W 97° 31' 52.680"	SRX1009700	4,38	7329275	7,72	0,96	0,94
Pfo_Olm	n.a.	Olmito	*P. latipinna*	N 25° 59' 9.240" W 97° 31' 53.760"	SRX1009706	3,18	6107017	6,43	0,96	0,93
Pfo_Bar-3	big river, high population density	Rio Purification near Barretal	*P. mexicana*	N 24° 02.848´ W 98° 22.264´	SRX1009708	4,88	7449198	7,84	0,96	0,92

Sample	Location	Geographic	Sympatric host/allopatry	GPS	SRA Exp Acc	ELC % Dupli–cation	Total Bases Kb (PF)	Fold coverage (950Mb)	% >Q30 (R1)	% >Q30 (R2)
Pfo_Br	big river, high population density	near Brownsville	*P. latipinna*	N 25° 53' 58.812" W 97° 28' 46.2"	SRX1009701	5,68	7370230	7,76	0,96	0,94
Pfo_SM	big river, high population density	San Marcos River, Tx	*P. latipinna*	N 29° 51' 28.800" W 97° 51' 51.120"	SRX1009711	3,78	6161389	6,49	0,96	0,93
Pfo_nBr-1	lab strain, n.a.	Lab strain Pf I with microchromosome	original host in the wild *P. latipinna*, in the lab since >30 years Black Molly	close to N 25° 53' 58.812" W 97° 28' 46.2" but different clone	SRX1009703	4,67	7330499	7,72	0,96	0,92
Pfo_Vic-2	small ditch	ditch at Mex 80 at km 105, 11 km north of Mante	*P. mexicana, P. latipunctata*	N 22° 48' 44.280" W 99° 0' 45.000"	SRX1009710	5,03	5674957	5,97	0,96	0,94
Pfo_nBr-2	lab strain,n.a.	Lab strain Pf I with microchromosome	orignal host in the wild *P. latipinna*,in the lab since >30 years Black Molly	close to N 25° 53' 58.812" W 97° 28' 46.2" but different clone	SRX1009704	3,98	6505942	6,85	0,96	0,92
Pfo_Lim	medium –size river, low population density	Mex 85 crossing the Rio Guayalejo near El Limon	*P. mexicana*	N 22° 50' 30.120" W 99° 1' 33.439"	SRX1009705	4,23	7285986	7,67	0,96	0,93
Pfo_Pad	big river, high population density	Rio Purification near Nueva Padilla	*P. mexicana*	N 24°02'35.6" W 98°54'15.9"	SRX1009702	3,49	6433006	6,77	0,95	0,91
Pfo_Man	small tributary, high population density	tributary to the Mante canal	*P. mexicana, P. latipunctata*	N 22° 43' 28.523" W 98° 52' 8.604"	SRX1009709	4,50	8289185	8,73	0,96	0,92
Pla_Tam	Small creek, polluted, industrial outflow	North of Tampico,near Altamira	allopatric	N 22°25.451´ W 97° 53.195´	Assembly GCA_001443285.1	NA		34		
Pla_Flo	Brackish lagoon, high population density	Key Largo Florida	No *P. formosa*	N 25° 7' 28.664" W 80° 24' 23.242"	SRX3445764	NA	26449924	27,8	93	85,3
Pla_Olm	n.a.	Olmito	*P. formosa*	N 25° 59' 9.240" W 97° 31' 53.760"	SRX3445762	NA	28628155	30,1	93,6	85,9

Sample	Location	Geographic	Sympatric host/allopatry	GPS	SRA Exp Acc	ELC % Dupli–cation	Total Bases Kb (PF)	Fold coverag e (950Mb)	% >Q30 (R1)	% >Q30 (R2)
Pla_SM	big river, high population density	San Marcos River, Tx	*P. formosa*	N 29° 51' 28.800" W 97° 51' 51.120"	SRX3445760	NA	27048773	28,5	93,5	86,1
Pla_Fld	small ditch, high population density	Fort Lauderdale, Florida	No *P. formosa*	n.a., approximate coordinates: N 26° 4.342' W 80° 7.497'	SRX3445761	NA	26623080	28,0	93,2	85,7
Pla_Cha	lake, still water, coastal area polluted, low population density	Laguna Champaxan	*P. formosa, P. mexicana*	N 22°23.425´ W 97° 55.828´	SRX3445763	NA	19481509	20,5	92,5	85,1
Pme_Cav	sulfur creek, cave, high population density	Cave I/14	No *P. formosa*	N17°26.321´ W 92°46.323´	SRX3445795	NA	19151005	20,2	93,7	86,3
Pme_Azu	fast flowing river with cataracts, low population density	Aquas Azul	No *P. formosa*	N 17° 15.484' W 092° 06.900'	SRX3445767	NA	24354175	25,6	91,9	84,6
Pme_Gua	fast flowing small river, low population density	Rio Guayalejo	*P. formosa*	N 23°16.624´ W 98° 56.315´	SRX3445796	NA	27755157	29,2	92	84,7
Pme_Ver	lab strain, n.a.	Near Rio Verde	No *P. formosa*	n.a., approximate coordinates: N 21°55.182' W 99° 57.550'	SRX3445765	NA	24392978	25,7	92,9	85,3
Pme_Cha	lake, still water, coastal area, polluted, high population density	Laguna Champaxan	*P. formosa*, *P. latipinna*	N 22°23.425´ W 97° 55.828´	Assembly GCA_001443 325.1	NA		30		

Supplementary Table 4. A comparative summary of the percentage of elements per genome of five poeciliid species

Percentage in genomes	*Poecilia formosa*	*Poecilia latipinna* (Pf library)	*Poecilia mexicana* (Pf library)	*Poecilia reticulata* (Pf library)	*Xiphophorus maculatus*
DNA TRANSPOSONS	**12,522**	**12,617**	**12,759**	**13,565**	**12,009**
Academ	0,008	0,008	0,009	0,007	0,021
CMC-EnSpm	0,149	0,147	0,152	0,139	0,018
CMC-Chapaev	0,005	0,005	0,005	0,006	0,017
CMC-Transib	-	-	-	-	-
Ginger	0	0	0	0	0,001
hAT-Ac	1,24	1,253	1,248	1,403	1,206
hAT-Blackjack	-	-	-	-	0,005
hAT-Charlie	1,326	1,313	1,295	1,55	0,747
hAT-Tip100	0,147	0,145	0,149	0,124	0,336
hAT-Tol2	0,253	0,255	0,255	0,248	0,163
hAT-hAT5	0,144	0,148	0,146	0,148	0,211
hAT-Other	0,624	0,607	0,604	0,666	0,618
IS4EU	0,099	0,103	0,105	0,124	0,052
Kolobok	0,002	0,003	0,003	0	-
MULE-MuDR	0,032	0,032	0,032	0,028	-
Novosib	0,008	0,008	0,008	0,009	-
P	0,041	0,042	0,042	0,037	0,04
PIF-Harbinger	0,539	0,547	0,543	0,473	0,505
PIF-ISL2EU	0,028	0,025	0,028	0,025	0,114
PiggyBac	0,339	0,342	0,34	0,324	0,249
TcMar-ISRm11	-	-	-	-	-
TcMar-Fot1	0,007	0,005	0,005	0,003	
TcMar-Tc1	5,98	6,294	6,259	6,593	6,19
TcMar-Tc2	0,105	0,101	0,104	0,196	0,078
TcMar-Tc4	0,036	0,037	0,037	0,038	-
TcMar-Tigger	0,197	0,199	0,201	0,189	0,219
RC/Helitron	0,031	0,032	0,033	0,019	0,251
Maverick/Polinton	0,052	0,051	0,054	0,037	0,025
OTHER DNA	1,13	0,915	1,102	1,179	0,943
LTR RETROTRANSPOSONS	**0,862**	**0,879**	**0,871**	**0,639**	**0,565**
BEL-Pao	0,034	0,04	0,038	0,014	0,035
Copia	0,062	0,062	0,061	0,049	0,001
DIRS1	0,026	0,028	0,027	0,017	0,02
DIRS-Ngaro	0,138	0,145	0,136	0,168	0,103
ERV1	0,036	0,037	0,038	0,014	0,01
ERV-Other	0,004	0,004	0,004	0,001	0,118
Gypsy	0,32	0,34	0,339	0,221	0,219
OTHER LTR	0,242	0,223	0,228	0,155	0,059
LINE RETROTRANSPOSONS	**2,688**	**2,76**	**2,768**	**2,652**	**2,409**
DRE	-	-	-	-	-
I-Nimb	0,012	0,022	0,022	0,014	0,007
I-Jockey	0,049	0,053	0,053	0,026	0,058
I-Other	0,143	0,155	0,151	0,128	0,045

Percentage in genomes	*Poecilia formosa*	*Poecilia latipinna* (Pf library)	*Poecilia mexicana* (Pf library)	*Poecilia reticulata* (Pf library)	*Xiphophorus maculatus*
L1-Tx1	0,019	0,019	0,019	0,014	0,02
L1-Other	0,045	0,049	0,048	0,029	0,104
L2	1,226	1,241	1,246	1,195	0,89
Proto2	0,01	0,011	0,011	0,005	-
R2	0,006	0,002	0,005	0,001	0,003
R4-Dong	0,074	0,098	0,101	0,072	0,008
RTE-BovB	0,284	0,265	0,272	0,347	0,409
RTE	0,231	0,24	0,236	0,308	0,128
Rex-Babar	0,545	0,56	0,558	0,484	0,687
Penelope	0,004	0,003	0,003	0,003	0,004
OTHER LINE	0,04	0,042	0,043	0,026	0,046
SINE RETROTRANSPOSONS	**0,387**	**0,406**	**0,4**	**0,713**	**0,376**
5S-Deu-L2	-	-	-	-	-
Hpa	-	-	-	-	0,005
MIR	0,08	0,084	0,083	0,093	0,112
tRNA-L2	-	-	-		-
tRNA-V	0,295	0,301	0,295	0,6	0,196
tRNA-Other	0,012	0,021	0,022	0,02	0,063
UNKNOWN	**6,17**	**6,222**	**6,218**	**5,727**	**5,588**
TOTAL TEs	**22,614**	**22,884**	**22,145**	**23,296**	**20,947**
TOTAL TEs (no Unknown)	**16,452**	**16,662**	**15,927**	**17,569**	**15,359**
Satellite	0,125	0,127	0,129	0,117	0,142
Simple repeats	0,974	1,183	1,071	0,992	0,765
Low complexity	0,823	0,805	0,778	0,733	0,66
TOTAL REPEATS	**24,536**	**24,999**	**24,123**	**25,138**	**22,514**

Supplementary Table 5. A summary of transposable element copy number expansion by family

	Coverage (bp)	Coverage (%)	Number of Repeat Masker copies	Percentage of copies (length >30%)	Percentage of copies (length >50%)	Percentage of copies (length >80%)
DNA TRANSPOSONS						
Other DNA	6734774	0,973	25956	45,4	30,5	14,2
Academ	58006	0,008	227	16,3	4,8	1,3
CMC-EnSpm	1031743	0,149	3195	20,1	10,4	3
Chapaev-Chap3	34604	0,005	133	67,7	40,6	14,3
Crypton	108	0	2	100	100	100
EnSpm	2519	0	4	50	50	50
Helitron	27216	0,004	34	11,8	5,9	5,9
RC/Helitron	184844	0,027	694	23,8	13,1	5,6
IS4EU	687935	0,099	2208	34,4	19,1	10,1
Kolobok	16064	0,002	104	44,2	27,9	1
MULE-MuDR	218883	0,032	914	49,6	22,4	7,3
MuDr	354	0	4	100	50	50
Novosib	51959	0,008	644	17,5	6,8	0,3
P	281258	0,041	1247	56,1	34,1	16,1
Harbinger	167876	0,024	427	34,2	17,1	7,7
PIF-Harbinger	3563101	0,515	13548	46,3	28,8	3,6
PIF-ISL2EU	192257	0,028	894	25,8	11,2	1,6
PiggyBac	2345908	0,339	8135	52,9	33,2	13,5
Maverick	243281	0,035	1443	2,2	0,6	0,1
Polinton	119684	0,017	110	4,5	2,7	0,9
TcMar	721198	0,104	3283	11,3	5,3	2,1
TcMar-Fot1	45868	0,007	254	11,8	4,3	0,4
TcMar-Tc1	41377647	5,98	237288	21,5	7,8	1,4
TcMar-Tc2	727784	0,105	3274	28,6	14,1	2,4
TcMar-Tc4	251268	0,036	932	38,7	16,3	6
TcMar-Tigger	1361698	0,197	6060	59	35,5	12,6
Zisupton	1409	0	3	66,7	66,7	33,3
hAT	4318533	0,624	20430	50,9	31,1	11
hAT-Ac	8581601	1,24	38389	39,6	19,9	8,2
hAT-Charlie	9173367	1,326	40945	21,4	8,7	3
hAT-Tip100	1013853	0,147	4364	25,1	12,6	4,7
hAT-Tol2	1751664	0,253	7032	68,4	45,3	21,4
hAT-hAT5	998666	0,144	5485	37,9	19,9	4,3
MITE	252430	0,036	1164	91,8	82,8	35,2
LINE RETROTRANSPOSONS						
LINE	279149	0,04	497	11,1	3,4	0,6
CR1	20381	0,003	93	8,6	4,3	1,1
Dong-R4	508945	0,074	2352	13,9	7,1	2,8
Hero	951	0	4	25	25	25
I	988998	0,143	3625	3,9	1,8	0,5

	Coverage (bp)	Coverage (%)	Number of Repeat Masker copies	Percentage of copies (length >30%)	Percentage of copies (length >50%)	Percentage of copies (length >80%)
I-Nimb	84816	0,012	541	3,7	0,4	0,4
Jockey	337374	0,049	1297	8,1	3,5	1,1
L1	83862	0,012	434	70	40,1	11,1
LINE1	227603	0,033	381	24,1	11,3	3,9
L1-Tx1	65301	0,009	383	6	4,7	0,3
LINE1-Tx1	67580	0,01	168	71,4	45,8	16,1
L2	8460146	1,223	41004	10,8	4,2	0,9
Nimb	63028	0,009	94	34	22,3	10,6
Penelope	25292	0,004	39	43,6	12,8	7,7
Proto2	68533	0,01	203	31	22,2	7,9
R2	39134	0,006	48	39,6	35,4	20,8
RTE	1600791	0,231	6511	5,1	1,5	0,2
RTE-BovB	1966255	0,284	7139	23,5	10,4	2,7
Rex-Babar	3714392	0,537	19120	24,9	10,6	2,5
Rex1	53918	0,008	162	28,4	18,5	5,6
LTR RETROTRANSPOSONS						
LTR	1676754	0,242	4315	16,6	6,5	0,7
BEL	236627	0,034	380	14,7	7,4	2,9
Copia	431257	0,062	2077	19	6,1	0,9
DIRS	34560	0,005	39	64,1	43,6	30,8
DIRS1	148121	0,021	327	7,3	3,7	0,9
ERV	28467	0,004	31	48,4	32,3	19,4
ERV1	246743	0,036	617	2,4	1,3	1
Gypsy	2212111	0,32	5782	9,5	4,5	2,1
Ngaro	957078	0,138	7120	11,2	4,8	0,4
SINE RETROTRANSPOSONS						
SINE	896816	0,13	5253	37,6	19,7	4,2
MIR	552654	0,08	3607	34,3	17,2	4
SINE2	65177	0,009	889	87	62,5	16,3
SINE3	958	0	13	100	100	76,9
V	2038885	0,295	11555	37,5	22,7	10,5
tRNA	82287	0,012	616	43	23,9	1,1
SINE?	244186	0,035	639	6,4	2,2	0,2
UNKNOWN	42690186	6,17	183038	50,3	29,6	10,2
Low_complexity	5694101	0,823	140266	100	99,9	99,7
Satellite	867850	0,125	3300	46,8	33,4	14
Simple_repeat	6736242	0,974	138596	99,3	98,9	98
TOTAL	171053653	24,721	1021420			

Supplementary Table 6. Transposable element coverage in the transcriptomes of *P. formosa* (left) and *X. maculatus* (right)

MOLLY	Transcriptome coverage (bp)	Transcriptome coverage (%)	Number of hits
DNA TRANSPOSONS		**0,38534**	
Others	7496	0,0105	76
Academ	314	0,00044	1
Chapaev-Chap3	406	0,00057	3
CMC-EnSpm	2220	0,00311	17
EnSpm	501	0,0007	3
Harbinger	4428	0,0062	22
Helitron	8863	0,01242	23
IS4EU	4146	0,00581	25
MULE-MuDR	757	0,00106	4
Mariner	2913	0,00408	14
Maverick	683	0,00096	3
Novosib	39	0,00005	1
P	93	0,00013	1
PIF-Harbinger	7857	0,01101	69
PIF-ISL2EU	744	0,00104	9
PiggyBac	9833	0,01378	45
Polinton	14217	0,01992	23
TcMar	3092	0,00433	17
TcMar-Fot1	487	0,00068	3
TcMar-Tc1	168617	0,23623	1193
TcMar-Tc2	1384	0,00194	15
TcMar-Tc4	193	0,00027	1
TcMar-Tigger	621	0,00087	8
Zisupton	1554	0,00218	5
hAT	10659	0,01493	68
hAT-Ac	7128	0,00999	82
hAT-Charlie	12849	0,018	121
hAT-Tip100	1312	0,00184	13
hAT-Tol2	745	0,00104	10
hAT-hAT5	899	0,00126	16
LINE RETROTRANSPOSONS		**0,25635**	
Others	15530	0,02176	22
CR1	4130	0,00579	5
Dong-R4	14224	0,01993	22
I	6964	0,00976	15
I-Nimb	5650	0,00792	16
Jockey	16153	0,02263	49
L1	141	0,0002	2
L1-Tx1	3652	0,00512	15
L2	33660	0,04716	210
LINE1	20701	0,029	35

PLATY	Transcriptome Coverage (bp)	Transcriptome coverage (%)	Number of hits
DNA TRANSPOSONS		**3,122**	
Others	82069	0,155	415
Academ	4881	0,009	18
CMC-Chapaev-3	548	0,001	3
Chapaev-Chap3	701	0,001	2
CMC-EnSpm	987	0,002	14
Harbinger	8504	0,016	21
Helitron	61955	0,117	260
IS4EU	4757	0,009	27
Kolobok	506	0,001	1
MITE	7345	0,014	50
Mariner	644	0,001	7
P	4185	0,008	17
PIF-Harbinger	33091	0,063	137
PIF-ISL2EU	4390	0,008	30
PiggyBac	19754	0,037	82
Polinton	5125	0,01	41
SET_Mariner	424	0,001	3
Tc1	15320	0,029	87
TcMar-Tc1	982620	1,859	5805
TcMar-Tc2	11682	0,022	50
TcMar-Tigger	18646	0,035	89
Tol2	1551	0,003	1
Unclassified	8416	0,016	65
hAT	78675	0,149	465
hAT-Ac	127443	0,241	633
hAT-Blackjack	576	0,001	3
hAT-Charlie	93743	0,177	587
hAT-Tip100	35574	0,067	240
hAT-Tol2	13934	0,026	71
hAT-hAT5	23422	0,044	145
LINE RETROTRANSPOSONS		**0,598**	
LINE	2271	0,004	17
CR1	704	0,001	4
Dong-R4	295	0,001	2
I	71448	0,135	174
I-Nimb	1987	0,004	14
Jockey	29566	0,056	104
L1	5427	0,01	21
L1-Tx1	3580	0,007	10
L2	86491	0,164	485
L2-Maui	204	0	3

MOLLY	Transcriptome coverage (bp)	Transcriptome coverage (%)	Number of hits
LINE1-Tx1	4565	0,0064	10
Nimb	3062	0,00429	7
Proto2	940	0,00132	3
R2	435	0,00061	1
RTE	8485	0,01189	39
RTE-BovB	13543	0,01897	67
Rex-Babar	29234	0,04096	139
Rex1	1884	0,00264	8
LTR RETROTRANSPOSONS		**0,45825**	
Others	4039	0,00566	27
BEL	38911	0,05451	116
Copia	9826	0,01377	28
DIRS	872	0,00122	3
DIRS1	8175	0,01145	30
ERV	4705	0,00659	16
ERV1	25875	0,03625	49
Gypsy	232400	0,32558	589
Ngaro	2301	0,00322	20
SINE RETROTRANSPOSONS		**0,01133**	
Others	3510	0,00492	30
MIR	304	0,00043	5
SINE2	178	0,00025	3
SINE3	46	0,00006	1
V	4005	0,00561	38
SINE?	42	0,00006	1
UNKNOWN	227462	**0,31867**	959
TOTAL	**1020654**	**1,42994**	**4471**

PLATY	Transcriptome Coverage (bp)	Transcriptome coverage (%)	Number of hits
Nimb	51	0	1
R2	41	0	1
RTE	9339	0,018	60
RTE-BovB	25756	0,049	129
Rex-Babar	77354	0,146	481
Unclassified	1228	0,002	11
PLE/Penelope	625	0,001	1
LTR RETROTRANSPOSONS		**0,103**	
LTR	4266	0,008	13
BEL	358	0,001	1
DIRS1	247	0	4
ERV	13832	0,026	67
ERV1	409	0,001	3
Gypsy	19711	0,037	77
Ngaro	10914	0,021	101
Pao	1199	0,002	4
Unclassified	3648	0,007	2
SINE RETROTRANSPOSONS		**0,118**	
SINE	19823	0,037	162
Hpa	572	0,001	8
MIR	11803	0,022	94
V	22462	0,042	132
tRNA	7784	0,015	53
SINE?	400	0,001	5
UNKNOWN	817464	1,546	3831
TOTAL	**2902727**	**5,487**	**15444**

Supplementary Table 10: Meiosis genes in *P. formosa*, manually checked to be intact

Symbol	Gene ID	Name
REC8	ENSPFOG00000009178	REC8 homolog
TEX11	ENSPFOG00000010478	Testis expressed 11
SYCP3	ENSPFOG00000013912	Synaptonemal complex protein 3
MEIOB	ENSPFOG00000014271	meiosis specific with OB domains [
MEI4	ENSPFOG00000022187	Meiosis inhibitor 2
SYCE2	ENSPFOG00000022749	Synaptonemal complex central element protein 2
SYCE3	ENSPFOG00000022749	Synaptonemal complex central element protein 3
MOS	ENSPFOG00000009336	V-mos Moloney murine sarcoma viral oncogene homolog
	ENSPFOG00000009339 [1]	
SYCP1	XM_007547092.2 (not annotated in Ensembl)	Synaptonemal complex protein 1
MEI1	ENSPFOG00000000271	Meiosis inhibitor 1
MND1	ENSPFOG00000004191	Meiotic nuclear divisions 1 homolog (S. cerevisiae)
HORMAD1	ENSPFOG00000008020	HORMA domain containing 1
SMC1B	ENSPFOG00000009938	Structural maintenance of chromosomes 1B
HSPA2	ENSPFOG00000007103	Heat shock 70kDa protein 2
DMC1	ENSPFOG00000005497	DMC1 dosage suppressor of mck1 homolog, meiosis-specific homologous recombination (yeast)
STAG3	ENSPFOG00000001981	Stromal antigen 3
SPO11	ENSPFOG00000009343	SPO11 meiotic protein covalently bound to DSB homolog (*S. cerevisiae*)
HORMAD1-like	LOC103140019 (not annotated in Ensembl)	HORMA domain containing 1-like
SPATA22	ENSPFOG00000023352	Spermatogenesis associated 22
SYCE1	ENSPFOG00000023217	Synaptonemal complex central element protein 1

[1]Remarks: duplicate in *P. formosa*

Supplementary Table 11: Spermatogenesis genes in *P.formosa*, manually checked to be intact

Gene ID	Name
ENSPFOG00000000171	ADAMTS like 1
ENSPFOG00000004159	alkB homolog 5, RNA demethylase
ENSPFOG00000016332	asunder, spermatogenesis regulator homolog (Drosphila)
ENSPFOG00000001640	breast cancer 2, early onset
ENSPFOG00000008020	HORMA domain containing 1
ENSPFOG00000019731	germ cell-less, spermatogenesis associated 1
ENSPFOG00000019052	male-enhanced antigen 1
ENSPFOG00000022187	meiosis-specific 4 homolog (S. cerevisiae)
ENSPFOG00000013317	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
ENSPFOG00000008841	ovo-like zinc finger 1
ENSPFOG00000013689	phospholipase D family, member 6
ENSPFOG00000003357	deleted in azoospermia-like
ENSPFOG00000008276	NME/NM23 family member 5
ENSPFOG00000022226	outer dense fiber of sperm tails 3-like 2
ENSPFOG00000001983 [1]	
ENSPFOG00000014029	proteasome activator subunit 4a
ENSPFOG00000006853	proteasome activator subunit 4b
ENSPFOG00000018868	ring finger protein 8, E3 ubiquitin protein ligase
ENSPFOG00000009878	rhophilin associated tail protein 1-like
ENSPFOG00000001295	SAS-6 centriolar assembly protein
ENSPFOG00000003906	smoothened, frizzled class receptor
ENSPFOG00000022167	lymphocyte antigen 6 complex locus protein G6d-like
ENSPFOG00000016944 [1]	
ENSPFOG00000006292	spermatogenesis associated 13
ENSPFOG00000015317	spermatogenesis associated 17
ENSPFOG00000017824 [2]	spermatogenesis associated 18
ENSPFOG00000000889	
ENSPFOG00000007371	spermatogenesis associated 2
ENSPFOG00000022764	spermatogenesis associated 20
ENSPFOG00000019016	spermatogenesis associated 4
ENSPFOG00000018264	spermatogenesis associated 5
ENSPFOG00000005380	spermatogenesis associated 5-like 1
ENSPFOG00000023137	spermatogenesis associated 6
ENSPFOG00000014690	spermatogenesis associated 7
ENSPFOG00000014616	serine/threonine/tyrosine interacting protein
ENSPFOG00000022749	synaptonemal complex central element protein 2
ENSPFOG00000002145	tudor domain containing 1
ENSPFOG00000002400	tudor domain containing 9
ENSPFOG00000014918	telomerase reverse transcriptase
ENSPFOG00000007470	tumor protein p53
ENSPFOG00000000149	thyroid hormone receptor interactor 13
ENSPFOG00000016882	VPS33B interacting protein, apical-basolateral polarity regulator, spe-39 homolog

[1]2 paralogs in teleost as result of WGD [2]duplicate in Poecilia formosa

Supplementary Table 12: Sexual development genes in *P. formosa*, manually checked to be intact

Symbol	Gene ID	Name
ALDH1a2	ENSPFOG00000004048	aldehyde dehydrogenase 1 family, member A2
ALDH1a3	ENSPFOG00000010138	aldehyde dehydrogenase 1 family, member A3
AMH	ENSPFOG00000001033	anti-Mullerian hormone
AMHR2	ENSPFOG00000004667	anti-Mullerian hormone receptor, type II
AR	ENSPFOG00000019378	androgen receptor
Cxcl12/SDF 1	ENSPFOG00000002793	chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
cxcl12b	ENSPFOG00000006643	chemokine (C-X-C motif) ligand 12b (stromal cell-derived factor 1)
CXCR4a	ENSPFOG00000010131	chemokine (C-X-C motif) receptor 4a
CXCR4b	ENSPFOG00000014542	chemokine (C-X-C motif), receptor 4b
	ENSPFOG00000014406	[1]
CYP11C1	ENSPFOG00000005578	cytochrome P450, family 11, subfamily C, polypeptide 1
cyp19a1a	ENSPFOG00000017864	cytochrome P450, family 19, subfamily A, polypeptide 1a
cyp19a1b	ENSPFOG00000019633	cytochrome P450, family 19, subfamily A, polypeptide 1b
cyp26a1	ENSPFOG00000009240	cytochrome P450, family 26, subfamily A, polypeptide 1
cyp26b1	ENSPFOG00000000077	cytochrome P450, family 26, subfamily b, polypeptide 1
cyp26c1	ENSPFOG00000017561	cytochrome P450, family 26, subfamily C, polypeptide 1
DAX1	ENSPFOG00000016778	orphan nuclear receptor Dax-1
DHH	ENSPFOG00000006989	desert hedgehog
DMRT1	ENSPFOG00000001335	doublesex and mab-3 related transcription factor 1
DMRT3a	ENSPFOG00000001398	doublesex and mab-3 related transcription factor 3a
ESR1	ENSPFOG00000019595	estrogen receptor 1
ESR2a	ENSPFOG00000016031	estrogen receptor 2a
FGF16	ENSPFOG00000009673	fibroblast growth factor 16
FGF20a	ENSPFOG00000007108	fibroblast growth factor 20a
FGF20b	ENSPFOG00000001230	fibroblast growth factor 20b
FGF9	ENSPFOG00000017395	fibroblast growth factor 9
FOXL2	ENSPFOG00000020308	forkhead box L2
FST	ENSPFOG00000015402	follistatin
	ENSPFOG00000015386	[1]
GATA-4	ENSPFOG00000022572	GATA-binding protein 4
GSDF	ENSPFOG00000004522	gonadal somatic cell derived factor
HHIP	ENSPFOG00000008387	hedgehog interacting protein
LRPPRC	ENSPFOG00000011227	leucine-rich pentatricopeptide repeat containing
Ptch2	ENSPFOG00000005637	patched 2
PAX2a	ENSPFOG00000013202	paired box 2a
PAX2b	ENSPFOG00000003702	paired box 2b
RSPO1	ENSPFOG00000002378	R-spondin 1
SF 1	ENSPFOG00000013683	splicing factor 1
SOX10	ENSPFOG00000012609	SRY (sex determining region Y)-box 10

Symbol	Gene ID	Name
SOX5	ENSPFOG00000004694	SRY (sex determining region Y)-box 5
SOX8a	ENSPFOG00000017781	SRY (sex determining region Y)-box 8a
Sox8b	ENSPFOG00000009378	SRY (sex determining region Y)-box 8b
SOX9a	ENSPFOG00000009155	SRY (sex determining region Y)-box 9a
SOX9b	ENSPFOG00000017723	SRY (sex determining region Y)-box 9b
SRD5A1	ENSPFOG00000004057	steroid-5-alpha-reductase, alpha polypeptide 1 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 1)
SRD5A2a	ENSPFOG00000019493	steroid-5-alpha-reductase, alpha polypeptide 2a
SRD5A2b	ENSPFOG00000018194	steroid-5-alpha-reductase, alpha polypeptide 2b
SRD5A3	ENSPFOG00000017304	steroid 5 alpha-reductase 3
WNT4a	ENSPFOG00000004485	wingless-type MMTV integration site family, member 4a
WNT4b	ENSPFOG00000014292	wingless-type MMTV integration site family, member 4b
WT1a	ENSPFOG00000010572	wilms tumor 1a
WT1b	ENSPFOG00000017089	wilms tumor 1b

[1]duplicate in Poecilia formosa

Supplementary Table 16. GO-term enrichment of biological processes in the *P. formosa* samples

GO ID	Term	Annotated	Significant	Expected	raw p-value	BH adjusted p-value
GO:0015074	DNA integration	64	12	0.26	1.7e-17	4.875600e-14
GO:0006259	DNA metabolic process	408	18	1.67	1.7e-14	2.437800e-11
GO:0006313	transposition, DNA-mediated	32	8	0.13	4.5e-13	3.226500e-10
GO:0032196	transposition	32	8	0.13	4.5e-13	3.226500e-10
GO:0006310	DNA recombination	105	8	0.43	9.5e-09	5.449200e-06
GO:0006508	proteolysis	776	16	3.18	4.7e-08	2.246600e-05
GO:0006334	nucleosome assembly	28	5	0.11	8.7e-08	2.772400e-05
GO:0031497	chromatin assembly	28	5	0.11	8.7e-08	2.772400e-05
GO:0034728	nucleosome organization	28	5	0.11	8.7e-08	2.772400e-05
GO:0006333	chromatin assembly or disassembly	30	5	0.12	1.3e-07	3.728400e-05
GO:0043170	macromolecule metabolic process	4886	39	20.00	1.9e-07	4.953818e-05
GO:0006323	DNA packaging	34	5	0.14	2.4e-07	5.736000e-05
GO:0065004	protein-DNA complex assembly	36	5	0.15	3.3e-07	6.760286e-05
GO:0071824	protein-DNA complex subunit organization	36	5	0.15	3.3e-07	6.760286e-05
GO:0071103	DNA conformation change	65	5	0.27	6.5e-06	1.242800e-03
GO:0044238	primary metabolic process	6067	41	24.84	8.8e-06	1.577400e-03
GO:0071704	organic substance metabolic process	6215	41	25.44	1.8e-05	3.036706e-03

Supplementary Table 17. GO-term enrichment of biological processes in the *P. latipinna* samples

GO ID	Term	Anno-tated	Signifi cant	Expected	raw p-value	BH adjusted p-value
GO:0015074	DNA integration	64	11	0.24	4.8e-16	1.376640e-12
GO:0006313	transposition, DNA-mediated	32	8	0.12	2.4e-13	2.294400e-10
GO:0032196	transposition	32	8	0.12	2.4e-13	2.294400e-10
GO:0006259	DNA metabolic process	408	14	1.55	2.1e-10	1.505700e-07
GO:0006310	DNA recombination	105	8	0.40	5.1e-09	2.925360e-06
GO:0007608	sensory perception of smell	70	5	0.27	6.5e-06	3.107000e-03
GO:0043170	macromolecule metabolic process	4886	34	18.55	1.0e-05	4.097143e-03
GO:0007606	sensory perception of chemical stimulus	79	5	0.30	1.2e-05	4.302000e-03
GO:0016567	protein ubiquitination	220	7	0.84	1.8e-05	5.736000e-03
GO:0032446	protein modification by small protein conjugation	233	7	0.88	2.6e-05	7.456800e-03
GO:0070647	protein modification by small protein conjugation or removal	245	7	0.93	3.6e-05	9.386182e-03

Supplementary Table 18. GO-term enrichment of biological processes in the *P. mexicana* samples

GO ID	Term	Anno-tated	Signi-ficant	Expected	raw p-value	BH adjusted p-value
GO:0015074	DNA integration	64	11	0.24	4.8e-16	1.376640e-12
GO:0006313	transposition, DNA-mediated	32	8	0.12	2.4e-13	2.294400e-10
GO:0032196	transposition	32	8	0.12	2.4e-13	2.294400e-10
GO:0006259	DNA metabolic process	408	14	1.55	2.1e-10	1.505700e-07
GO:0006310	DNA recombination	105	8	0.40	5.1e-09	2.925360e-06
GO:0007608	sensory perception of smell	70	5	0.27	6.5e-06	3.107000e-03
GO:0043170	macromolecule metabolic process	4886	34	18.55	1.0e-05	4.097143e-03
GO:0007606	sensory perception of chemical stimulus	79	5	0.30	1.2e-05	4.302000e-03
GO:0016567	protein ubiquitination	220	7	0.84	1.8e-05	5.736000e-03
GO:0032446	protein modification by small protein conjugation	233	7	0.88	2.6e-05	7.456800e-03
GO:0070647	protein modification by small protein conjugation or removal	245	7	0.93	3.6e-05	9.386182e-03

Supplementary Table 19. GO-term enrichment of molecular functions in the *P. formosa* samples

GO ID	Term	Anno-tated	Signi-ficant	Expected	raw p-value	BH adjusted p-value
GO:0004175	endopeptidase activity	410	22	4,26	2.90E-10	2.55E-07
GO:0004803	transposase activity	31	8	0,32	7.40E-10	3.25E-07
GO:0003676	nucleic acid binding	2794	60	29	7.70E-09	2.26E-06
GO:0004252	serine-type endopeptidase activity	173	13	1,8	3.10E-08	6.81E-06
GO:0008236	serine-type peptidase activity	200	13	2,08	1.70E-07	2.49E-05
GO:0017171	serine hydrolase activity	200	13	2,08	1.70E-07	2.49E-05
GO:0070011	peptidase activity, acting on L-amino acids	596	22	6,19	2.50E-07	3.14E-05
GO:0008233	peptidase activity	608	22	6,31	3.50E-07	3.85E-05
GO:0004842	ubiquitin-protein transferase activity	163	9	1,69	5.20E-05	0,005078667
GO:0019787	ubiquitin-like protein transferase activity	168	9	1,74	6.60E-05	0,0058014
GO:0032446	protein modification by small protein conjugation	233	9	1,42	1.20E-05	0,0023672
GO:0070647	protein modification by small protein conjugation or removal	245	9	1,49	1.80E-05	0,003328875
GO:0071704	organic substance metabolic process	5955	55	36,33	2.50E-05	0,004351471
GO:0050900	leukocyte migration	33	4	0,2	4.60E-05	0,007561889

Supplementary Table 20. GO-term enrichment of molecular functions in the *P. latipinna* samples

GO ID	Term	Anno-tated	Signi-ficant	Expected	raw p-value	BH adjusted p-value
GO:0004803	transposase activity	31	8	0.21	2.7e-11	2.370600e-08
GO:0004175	endopeptidase activity	410	14	2.82	8.6e-07	3.775400e-04
GO:0003676	nucleic acid binding	2794	40	19.22	2.0e-06	5.853333e-04
GO:0004252	serine-type endopeptidase activity	173	9	1.19	3.0e-06	6.585000e-04
GO:0008236	serine-type peptidase activity	200	9	1.38	9.9e-06	1.448700e-03
GO:0017171	serine hydrolase activity	200	9	1.38	9.9e-06	1.448700e-03
GO:0046982	protein heterodimerization activity	120	7	0.83	1.9e-05	2.383143e-03
GO:0070011	peptidase activity, acting on L-amino acid peptides	596	14	4.10	6.0e-05	6.585000e-03
GO:0008233	peptidase activity	608	14	4.18	7.4e-05	7.219111e-03

Supplementary Table 21. GO-term enrichment of molecular functions in the *P. mexicana* samples

GO ID	Term	Anno-tated	Signi-ficant	Expected	raw p-value	BH adjusted p-value
GO:0030259	lipid glycosylation	21	9	0.35	2.3e-11	1.568140e-07
GO:0019882	antigen processing and presentation	99	14	1.66	1.0e-09	3.409000e-06
GO:0055114	oxidation-reduction process	796	34	13.36	4.9e-07	1.113607e-03
GO:0031341	regulation of cell killing	55	8	0.92	3.5e-06	5.965750e-03
GO:0048387	negative regulation of retinoic acid receptor signaling pathway	16	5	0.27	4.8e-06	6.545280e-03
GO:0048385	regulation of retinoic acid receptor signaling pathway	18	5	0.30	9.2e-06	9.374750e-03
GO:0001916	positive regulation of T cell mediated cytotoxicity	31	6	0.52	1.1e-05	9.374750e-03
GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	31	6	0.52	1.1e-05	9.374750e-03

Supplementary Table 22: Gene copies and expression of meiosis gene showing CNV in *P. formosa*

Gene	Ensembl Gene ID	Expr. Ovary mature[1]	Expr. Ovary immat[1]
cdk1	ENSPFOG00000015459	812	1498
	ENSPFOG00000015509	404	718
	ENSPFOG00000015558	85	425
bod1l1	ENSPFOG00000002769	1	262
		2370	16301
	ENSPFOG00000002857	21	381
aurkb	ENSPFOG00000002811	1156	2692
	ENSPFOG00000002658	786	3231
securin/pttg1	ENSPFOG00000006275	1046	906
	ENSPFOG00000017444	802	1034
septin 6	ENSPFOG00000024721	380	937
	ENSPFOG00000024595	1218	2484
		57	482
pttg1	ENSPFOG00000023510	802	1034
	ENSPFOG00000006275	1046	906
mos	ENSPFOG00000009336	130	621
		94	1736
	ENSPFOG00000009339	1000	5127
		23	160
spdya/ringo	ENSPFOG00000011326	627	3111
	ENSPFOG00000011342	691	3715
cntln	ENSPFOG00000000922	23	487
		1	69

[1] numbers are raw read counts from RNAseq analyses of mature and immature ovaries of *P. formosa*.

Supplementary Table 23. A summary of allele ancestry using INTROGRESS method

	Lower estimate	H index	Upper estimate
Pfo_nBr-1	0.496606	0.500603	0.5046
Pfo_nBr-2	0.49617	0.500786	0.505403
Pfo_Br	0.497923	0.501963	0.506002
Pfo_D1	0.499897	0.5039	0.507901
Pfo_Vic-2	0.49485	0.5004	0.50595
Pfo_Man	0.494666	0.498321	0.501977
Pfo_Lim	0.48907	0.493171	0.497274
Pfo_Bar-3	0.50194	0.505908	0.509873
Pfo_Olm	0.498484	0.503459	0.508434
Pfo_SM	0.499818	0.504691	0.509562
Pfo_Pad	0.500726	0.505447	0.510166
Pfo_Cha	0.49692	0.500238	0.503556
Pfo_Vic-1	0.493504	0.496927	0.500352
Pfo_Tan-1	0.501016	0.504183	0.507346
Pfo_Tan-2	0.494273	0.497748	0.501224
Pfo_Tan-3	0.500531	0.503993	0.507453
Pfo_Tan-4	0.501509	0.504738	0.508025
Pfo_Bar-2	0.497305	0.500499	0.503692
Pfo_Gua	0.493147	0.496457	0.499768

Supplementary Table 24: Introgressed scaffolds in *P. formosa*

		Samples				
		Pfo_Tan-2	**Pfo_nBr-2**	**Pfo_nBr-1**	**Pfo_Br**	**Pfo_Vic-2**
Introgressed scaffolds	**KI519701.1** (2,222,512)	x				
	KI519768.1 (1,584,072**)**	x				
	KI519777.1 (1,448,886)	x				
	KI519845.1 (1,003,796)	x				
	KI519919.1 (742,264)	x				
	KI519939.1 (700,049)	x				
	KI520051.1 (505,600)	x				
	KI520145.1 (313,854)	x				
	KI520148.1 (281,439)	x				
	KI520217.1 (209,494)	x				
	KI520378.1 (108,081)	x				
	KI519925.1 (309,090)		x			
	KI519926.1 (273,463)		x			
	KI519991.1 (563,587)		x	x		
	KI520043.1 (519,751)		x	x		
	KI520074.1 (423,533)		x	x		
	KI520164.1 (352,441)				x	
	KI520210.1 (233,258)				x	
	KI520162.1 (324,406)					x
Total introgressed scaffold size		8,116,251	1,569,673	1,506,871	585,699	324,406

Supplementary Table 26: GO-term enrichment of molecular functions in introgressed genes

GO ID	Term	Anno-tated	Signifi-cant	Expec-ted	raw p-value	BH adjusted p-value
GO:0004298	threonine-type endopeptidase activity	24	4	0.28	0.00015	0.0651
GO:0070003	threonine-type peptidase activity	24	4	0.28	0.00015	0.0651
GO:0046872	metal ion binding	2730	48	31.57	0.00121	0.215264
GO:0043167	ion binding	2806	49	32.45	0.00123	0.215264
GO:0043169	cation binding	2733	48	31.61	0.00124	0.215264
GO:0016887	ATPase activity	193	8	2.23	0.00182	0.215915
GO:0005231	excitatory extracellular ligand-gated on channel activity	46	4	0.53	0.00193	0.215915
GO:0050661	NADP binding	22	3	0.25	0.00199	0.215915
GO:0004551	nucleotide diphosphatase activity	7	2	0.08	0.00269	0.259435555555556
GO:0004396	hexokinase activity	8	2	0.09	0.00356	0.280916363636364
GO:0005536	glucose binding	8	2	0.09	0.00356	0.280916363636364
GO:0048029	monosaccharide binding	11	2	0.13	0.00683	0.438629333333333
GO:0004970	ionotropic glutamate receptor activity	35	3	0.4	0.00758	0.438629333333333
GO:0005234	extracellular-glutamate-gated channel activity	35	3	0.4	0.00758	0.438629333333333
GO:0008066	glutamate receptor activity	35	3	0.4	0.00758	0.438629333333333
GO:0046914	transition metal ion binding	1418	26	16.4	0.01199	0.619854117647059
GO:0003707	steroid hormone receptor activity	77	4	0.89	0.01214	0.619854117647059
GO:0001871	pattern binding	16	2	0.19	0.01434	0.64604
GO:0030247	polysaccharide binding	16	2	0.19	0.01434	0.64604
GO:0030246	carbohydrate binding	125	5	1.45	0.01503	0.64604
GO:0015085	calcium ion transmembrane transporter activity	83	4	0.96	0.01563	0.64604
GO:0004721	phosphoprotein phosphatase activity	181	6	2.09	0.01872	0.738589090909091
GO:0072509	divalent inorganic cation transmembrane transporter activity	93	4	1.08	0.02274	0.858187826086957
GO:0008270	zinc ion binding	1221	22	14.12	0.02448	0.88536
GO:0005044	scavenger receptor activity	56	3	0.65	0.02693	0.9350096
GO:0005262	calcium channel activity	60	3	0.69	0.03216	0.970424

GO ID	Term	Anno-tated	Signifi-cant	Expec-ted	raw p-value	BH adjusted p-value
GO:0004527	exonuclease activity	25	2	0.29	0.03351	0.970424
GO:0017154	semaphorin receptor activity	25	2	0.29	0.03351	0.970424
GO:0019200	carbohydrate kinase activity	25	2	0.29	0.03351	0.970424
GO:0038024	cargo receptor activity	61	3	0.71	0.03354	0.970424
GO:0016787	hydrolase activity	2239	35	25.9	0.03546	0.97297375
GO:0016462	pyrophosphatase activity	710	14	8.21	0.03587	0.97297375
GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	715	14	8.27	0.03771	0.991887272727273
GO:0016817	hydrolase activity, acting on acid anhydrides	722	14	8.35	0.04039	1
GO:0042626	ATPase activity, coupled to transmembrane movement of substances	67	3	0.77	0.04249	1
GO:0022824	transmitter-gated ion channel activity	68	3	0.79	0.04408	1
GO:0022835	transmitter-gated channel activity	68	3	0.79	0.04408	1
GO:0015399	primary active transmembrane transporter activity	70	3	0.81	0.04737	1
GO:0015405	P-P-bond-hydrolysis-driven transmembrane transporter activity	70	3	0.81	0.04737	1
GO:0016820	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	70	3	0.81	0.04737	1
GO:0035586	purinergic receptor activity	31	2	0.36	0.04967	1

Supplementary Table 27. A summary of sequence coverage, polymorphism counts, maximum-likelihood estimates of theta ($4N_e\mu$) and sequencing error rate for all isolates of *P. formosa* and parental species *P. latipinna* and *P. mexicana*

Species	Sample ID	Sequence coverage	No. of sites (>4x)	Heterozygosity	Sequencing error rate
P. formosa	Pfo_nBr-1	7,72	4,13E+08	1,30E-02	6,48E-04
P. formosa	Pfo_nBr-2	6,85	3,92E+08	1,27E-02	6,47E-04
P. formosa	Pfo_Br	7,76	4,14E+08	1,30E-02	5,85E-04
P. formosa	Pfo_D1	7,72	4,15E+08	1,30E-02	5,83E-04
P. formosa	Pfo_Vic-2	5,97	3,73E+08	1,24E-02	6,80E-04
P. formosa	Pfo_Man	8,73	4,25E+08	1,31E-02	5,71E-04
P. formosa	Pfo_Lim	7,67	4,09E+08	1,27E-02	5,90E-04
P. formosa	Pfo_Bar-3	7,84	4,15E+08	1,28E-02	5,77E-04
P. formosa	Pfo_Olm	6,43	3,86E+08	1,26E-02	7,69E-04
P. formosa	Pfo_SM	6,49	3,88E+08	1,25E-02	6,78E-04
P. formosa	Pfo_Pad	6,77	3,93E+08	1,25E-02	6,42E-04
P. formosa	Pfo_Cha	10,07	4,36E+08	1,34E-02	7,06E-04
P. formosa	Pfo_Vic-1	9,49	4,30E+08	1,33E-02	7,40E-04
P. formosa	Pfo_Tan-1	11,25	4,41E+08	1,36E-02	7,51E-04
P. formosa	Pfo_Tan-2	9,04	4,25E+08	1,33E-02	6,45E-04
P. formosa	Pfo_Tan-3	9,09	4,27E+08	1,34E-02	6,73E-04
P. formosa	Pfo_Tan-4	10,32	4,36E+08	1,35E-02	7,11E-04
P. formosa	Pfo_Bar-2	10,69	4,43E+08	1,33E-02	6,45E-04
P.formosa	Pfo_Bar-1	245,9	4,93E+08	1,43E-02	1,66E-03
P. formosa	Pfo_Gua	11,2	4,41E+08	1,34E-02	7,10E-04
P. latipinna	Pla-SM	28,5	4,39E+08	2,14E-03	3,78E-03
P. latipinna	Pla-Fld	28	4,39E+08	1,83E-03	3,74E-03
P. latipinna	Pla-Olm	30,1	4,41E+08	1,05E-03	3,90E-03
P. latipinna	Pla-Flo	27,8	4,41E+08	3,09E-03	4,09E-03
P. latipinna	Pla-Cha	20,5	4,34E+08	1,25E-03	3,92E-03
P. latipinna	Pla-Tam	20,5	4,50E+08	1,10E-03	1,11E-03
P. mexicana	Pme-Ver	25,6	4,37E+08	2,02E-03	4,02E-03
P. mexicana	Pme-Azu	29,2	4,40E+08	7,93E-04	4,11E-03
P. mexicana	Pme-Gua	30,6	4,45E+08	5,32E-03	4,06E-03
P. mexicana	Pme-Cav	26,3	4,39E+08	2,28E-03	4,10E-03
P. mexicana	Pme_Cha	22,7	4,53E+08	1,43E-03	1,19E-03

Supplementary Table 28: Allelic distance of 15 immune genes within individuals

locus	Alignment length [bp]	*P. latipinna*	*P. mexicana*	*P. formosa*
C2	7302	38.5	0	99
C3	3922	1.5	0	74
CD59	4368	10	4	35
CLEC	6113	10	91	91
InPhi4a	3609	0	9	39
IRF2	1786	4	0	61
IL1a	10224	2	1	71
IL21	12552	0	0	229
TLR1	3633	0	0	30
TLR5	6733	0	0	150
TLR7	6584	0	0	94
TLR8	7691	17	17	127
TLR9	5746	2	2	69
TLR21	4273	0	0	37
TRAF6	6077	1	0	115
Median overall		1.5	0	74

Given is the median number of differences of alleles within *Poecilia formosa* individuals (N =20) and within individuals from the parental species (*P. latipinna* N = 6; *P. mexicana* N = 5). The alignment length for each gene locus is given in base pairs (bp).

Supplementary Table 29: Analysis of overall dN-dS rates for 15 immune genes

	dN-dS[1]		
immune gene	***P. mexicana***	***P. latipinna***	***P. formosa***
C2	-1,085	-0,275	-1,433
C3	0,183	-1,122	-0,366
CLEC	-0,175	-0,802	-0,089
CD59	0,136	0,391	0,363
InPhy4a	0,104	-0,014	-0,632
IL1a	-0,444	-0,337	0,108
IL21	0,409	-0,184	-0,261
IRF2	0,42	-0,381	0,183
TLR1	-1,495	-0,804	-0,997
TLR5	-0,473	0,431	-0,011
TLR7	-0,124	-0,172	-0,912
TLR8	-1,137	-0,348	-0,427
TLR9	-0,924	-0,072	-1,022
TLR21	-1,108	-1,875	-1,142
TRAF6	0,138	-0,619	-0,607

[1]Given are mean values of pairwise sequence analyses for all *Poecilia* species. (dN = dS - values of/close to zero = neutral evolution, dN < dS = values above 0 = positive selection, dN > dS = values below 0 = purifying selection).

Supplementary Table 30. Nucleotide composition of the assemblies in the different masking steps

Assembly	A	C	G	T	N	%masked sequence
pformosa.5.1.2.fa	217240831	139917828	139778204	217251047	34735552	4.64
pformosa.5.1.2.rm.trf.fa	210400740	137048710	136928116	210412892	54133004	7.23
pformosa.5.1.2.rm.trf.kmer.fa	202392947	131630920	131516985	202418966	80963644	10.81
pformosa.5.1.2.rm.trf.kmer.36bps_padding.fa	189586924	123547038	123444004	189617118	122728378	16.39

Supplementary Table 31. Genes involved in the processing of small non-coding RNAs (miRNA, piRNA)

Gene	ENSEMBL Gene ID	Remarks
Drosha	ENSPFOG00000000371	
DGCR8	ENSPFOG00000010796	
DDX17	ENSPFOG00000011861	
DDX5	ENSPFOG00000004109	2 paralogs in teleost as result of WGD
	ENSPFOG00000011690	
TARDP	ENSPFOG00000016117	2 paralogs in teleost as result of WGD
	ENSPFOG00000004365	
Xpo5	ENSPFOG00000004497	
Dicer1	ENSPFOG00000008531	
TARBP2	ENSPFOG00000007586	
Ago1	ENSPFOG00000005867	
Ago2	ENSPFOG00000005321	
Ago3b	ENSPFOG00000006684	2 paralogs in teleost as result of WGD
Ago3l	LOC103139281	NCBI ID, not annotated in Ensembl
Ago4	ENSPFOG00000001935	
Piwil1	ENSPFOG00000016960	Piwil3,4 absent in all fish
Piwil2	ENSPFOG00000012110	
Dis3L2	ENSPFOG00000007818	
Lin28a	ENSPFOG00000014054	
ZCCHC11	ENSPFOG00000002439	syn TUT4
ZCCHC6	ENSPFOG00000011606	syn TUT7

Supplementary Table 32. Putative LoF genes of *P. formosa* displaying gene ontology enrichment for the neuromuscular process[1]

Gene name	Gene description	Ensembl gene ID
CNTNAP1	contactin associated protein 1	ENSG00000108797
FABP7	fatty acid binding protein 7, brain	ENSG00000164434
STRA6	stimulated by retinoic acid gene 6 homolog (mouse)	ENSG00000137868
TNR	tenascin R	ENSG00000116147
ADCY5	adenylate cyclase 5	ENSG00000173175
IGDCC3	immunoglobulin superfamily, DCC subclass, member 3	ENSG00000174498
PCDH15	protocadherin-related 15	ENSG00000150275
NR4A3	nuclear receptor subfamily 4, group A, member 3	ENSG00000119508

[1]The number of reference genes in this category is 75, the experimentally observed were 8 while the expected was 1.08. The ratio of enrichment was 7.4 with an adjusted p value of 0.0157.

Supplementary Table 33: Summary of gene gain and loss events inferred after correcting for annotation and assembly error across all 8 species

	Expansions			**Contractions**			**No Change**	**Avg. Expansion**
	Families	**Genes**	**genes/ expansion**	**Families**	**Genes**	**genes/ contraction**	**Families**	
Cavefish	1095 (26)[1]	1738	1.59	920 (30)	1152	1.25	7897	0.05912
Tilapia	613 (98)	1646	2.69	1488 (14)	1594	1.07		
Tetraodon	1057 (28)	1571	1.49	2348 (70)	2762	1.18	6507	-0.120157
Amazon molly	756 (109)	1117	1.48	774 (22)	850	1.10	8382	0.026937
Platyfish	475 (49)	588	1.24	530 (43)	695	1.31	8907	-0.010795
Zebrafish	1192 (96)	2429	2.04	692 (4)	779	1.13	8028	0.166465
Medaka	454 (38)	770	1.70	2175 (47)	2638	1.21	7283	-0.188458
Stickleback	579 (54)	936	1.62	1617 (18)	1841	1.14	7716	-0.091303

[1]The number of rapidly evolving families is shown in parentheses for each type of change.

Supplementary Table 34. Duplications per individual of *P. formosa*, *P. latipinna* and *P. mexicana*

Sample	#Duplicated bps	%Genome duplicated	#Shared duplicated bps	%Shared duplicated bps	#Uniquely duplicated bp	%Uniquely duplicated bp
Pfo_Tan-1	22.520.571	3,01	14.763.255	65,55	275.246	1,22
Pfo_Tan-2*	28.122.731	3,76		52,5	5.637.460	20,05
Pfo_Tan-3	22.689.273	3,03		65,07	184.990	0,82
Pfo_Tan-4	23.010.402	3,07		64,16	261.022	1,13
Pfo_Gua	21.346.522	2,85		69,16	318.620	1,49
Pfo_Vic-1	22.092.369	2,95		66,83	131.109	0,59
Pfo_Cha	22.667.534	3,03		65,13	311.707	1,38
Pfo_Bar-2	24.329.074	3,25		60,68	727.509	2,99
Pfo_D1	22.154.026	2,96		66,64	170.005	0,77
Pfo_Olm	21.630.294	2,89		68,25	250.634	1,16
Pfo_Bar-3	22.165.973	2,96		66,6	236.027	1,06
Pfo_Br*	22.945.498	3,06		64,34	576.808	2,51
Pfo_SM	21.252.111	2,84		69,47	137.235	0,65
Pfo_nBr-1*	23.359.213	3,12		63,2	355.257	1,52
Pfo_Vic-2*	21.054.287	2,81		70,12	514.987	2,45
Pfo_nBr-2*	22.937.782	3,06		64,36	720.227	3,14
Pfo_Lim	22.116.685	2,95		66,75	231.693	1,05
Pfo_Pad	21.451.526	2,86		68,82	169.007	0,79
Pfo_Man	22.863.953	3,05		64,57	516.602	2,26
Pla_Flo	23.790.400	3,18	16.612.356	69,83	1.441.780	6,06
Pla_Fld	22.701.944	3,03		73,18	986.290	4,34
Pla_Cha	22.374.363	2,99		74,25	1.120.200	5,01
Pla_Olm	23.145.296	3,09		71,77	1.090.373	4,71
Pla_SM	22.858.166	3,05		72,68	948.018	4,15
Pme_Azu	21.963.694	2,93	16.491.815	75,09	1.023.124	4,66
Pme_Cav	21.865.150	2,92		75,43	1.366.910	6,25
Pme_Gua	22.764.958	3,04		72,44	1.471.647	6,46
Pme_Ver	22.802.643	3,04		72,32	1.281.704	5,62

Samples marked with an asterisk (*) carry putative introgressed scaffolds. Introgressed bps are included in the computed number of duplicated bps and thus these samples are not included in the calculation of percentage shared and uniquely duplicated bps among samples.