Population genomics reveals a possible history of backcrossing and recombination in the gynogenetic fish *Poecilia formosa*

Laura Alberici da Barbiano, Zachariah Gompert, Andrea S. Aspbury, Caitlin R. Gabor, and Chris C. Nice

Department of Biology, Texas State University, San Marcos, TX 78666

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved July 12, 2013 (received for review February 26, 2013)

Unisexual sperm-dependent vertebrates are of hybrid origins, rare, and predicted to be short-lived as a result of several challenges arising from their mode of reproduction. In particular, because of a lack of recombination, clonal species are predicted to have a low potential to respond to natural selection. However, many unisexual sperm-dependent species persist, and assessing the genetic diversity present in these species is fundamental to understanding how they avoid extinction. We used population genomic methods to assess genotypic variation within the unisexual fish *Poecilia formosa*. Measures of admixture and population differentiation, as well as clustering analyses, indicate that the genomes of individuals of *P. formosa* are admixed and intermediate between *Poecilia latipinna* and *Poecilia mexicana*, consistent with the hypothesis of their hybrid origins. Bayesian genomic cline analyses indicate that about 12% of sampled loci exhibit patterns consistent with inheritance from only one parent. The estimation of observed heterozygosity clearly suggests that *P. formosa* is not comprised of direct descendants of a single nonrecombining asexual F1 hybrid individual. Additionally, the estimation of observed heterozygosity provides support for the hypothesis that the history of this unisexual species has included backcrossing with the parent species before the onset of gynogenesis. We also document high levels of variation among asexual individuals, which is attributable to recombination (historical or ongoing) and the accumulation of mutations. The high genetic variation suggests that this unisexual vertebrate has more potential to respond to natural selection than if they were frozen F1 hybrids.

The maintenance of sex presents a conundrum for evolutionary biology because the costs of sexual reproduction (cost of producing males, energy expenditure to find a mate, exposure to diseases, and segregation of alleles) appear to be immediate and substantial, whereas its benefits (facilitation of adaptations, elimination of deleterious mutations) are postponed (reviewed in ref. 1). The long-term maintenance of unisexual organisms is of interest to evolutionary biologists as well because the advantages of asexual reproduction are all immediate (no cost of producing males and, therefore, exponential population growth), but the long-term costs are substantial (accumulation of deleterious mutations and lack of genetic recombination to respond to environmental changes). Asexual vertebrate species are, therefore, predicted to be short-lived compared with sexually reproducing species (2–4). However, recent work focused on nonvertebrate species has challenged the view that recombination is absent in asexual lineages and that, therefore, those species are doomed to extinction. Asexual aphids, fungi, and microcrustaceans have all been shown to be genetically variable [aphids (5), fungi (6), *Daphnia* (7)] and, in some cases, mitotic recombination facilitates the spread of beneficial mutations (8). Therefore, understanding how much genetic variation is present in asexual lineages, and whether the presence of this variation and the mechanisms that facilitate it are shared among taxa, is an essential step toward understanding the evolution of sexual and asexual reproduction and, perhaps, challenging existing paradigms.

Asexuality is common in many phyla (reviewed in ref. 7), but it is relatively rare in vertebrates (1). All known unisexual (all-female) vertebrates are products of hybridization events between sexually reproducing species (ref. 9 and references therein), constitute only 0.1% of extant vertebrate species (1, 9). One type of asexual reproduction found in unisexual vertebrates is gynogenesis, where females must mate with males of a closely related species (but refer to ref. 10 for exceptions), but the nonrecombinant embryos do not inherit any genetic information from the sperm donor (9). Because gynogens require sperm to initiate development of offspring, but no paternal genes are expressed, they are considered “sexual parasites” (11).

The maintenance of a gynogenetic species is paradoxical because gynogens face the costs of both sexual and asexual reproduction: the cost of finding a mate, exposure to diseases, accumulation of deleterious mutations, and lack of genetic recombination to facilitate adaptation. In addition, because male sperm donors do not gain a fitness advantage from mating with gynogens, selection should favor males that avoid mating with them.

Given the extensive and diverse list of challenges faced by gynogenetic species, they are predicted to be short-lived with a limited potential to respond to natural selection. Nevertheless, gynogenetic species persist, and some have origins in the distant past (12, 13). This suggests that gynogenetic species might be able to avoid or ameliorate some of the costs associated with their reproductive mode. One question that arises is how much genetic variation persists in gynogenetic species?

The Amazon molly (*Poecilia formosa*) is an excellent system to explore this question. *P. formosa* is the first vertebrate recognized as asexual (11) and is a gynogenetic species that uses *Poecilia mexicana* (Atlantic molly), *Poecilia latipinna* (sailfin molly), and *Poecilia latipunctata* (Tamesi molly) as sexual hosts (14). Like every other known unisexual vertebrate, *P. formosa* is thought to be a hybrid lineage (9, 13, 15–17). *P. mexicana* is recognized to be the maternal species of *P. formosa* (13, 15–17), whereas *P. latipinna* (or an extinct ancestor of *P. latipinna*) is the putative paternal species (17). *P. formosa* lives in sympatry with at least one of the two parent species throughout its range from the Tampa region in Mexico to the southeastern United States (Fig. 1A). Although recent studies suggest that *P. formosa* is a species that consists of “frozen” F1 hybrid clones (i.e., individuals with...


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the Dryad Digital Repository, www.datadryad.org (Dryad data package DOI: 10.5061/dryad.h6p35).

1To whom correspondence should be addressed. E-mail: la1122@txstate.edu.

This article contains supporting information online at www.pnas.orglookup/suppl/doi:10.1073/pnas.1303730110/-/DCSupplemental.
ancestry from *P. mexicana* and *P. latipinna* at all loci) (15–19), this result is based on limited genetic data. Additionally, it is still not clear whether *P. formosa* is the product of a single or multiple hybridization events, although a recent investigation supports the hypothesis of a single event possibly giving rise to several clonal lineages (19).

The overall objective of this research was to examine genotypic variation in *P. formosa* at a genomic scale. We generated thousands of DNA sequence markers to address three main questions. (i) Is *P. formosa* the product of hybridization between *P. latipinna* and *P. mexicana*? This is the conclusion of several previous studies (11, 13, 15–19), and, here, we attempt to confirm this result with a genome-wide survey of *P. formosa* and its putative parent species. (ii) Is *P. formosa* composed of frozen F\(_1\) hybrid clones, or is there evidence of a more complicated history of nonclonal reproduction or recombination? Previous genetic analyses (13, 15–19), and the fact that no laboratory has been able to synthesize *P. formosa* from artificial hybridization experiments (15, 19–21), suggest that the genome of this species is more complicated than that of a simple F\(_1\) hybrid. If *P. formosa* is not composed of clones of F\(_1\) hybrid lineages, then, what is the genetic contribution of each parent species? (iii) How much genotypic diversity exists within *P. formosa* and is this genotypic diversity consistent with *P. formosa* having evolved from a single or multiple independently formed hybrid individuals? Quantifying genotypic variation will provide some estimate of this species’ potential to respond to selection.

**Results and Discussion**

We used a next-generation sequencing population genetics approach to collect information on variation from across the genomes of *P. formosa*, *P. latipinna*, and *P. mexicana* and to understand the genomic composition of the gynogenetic species. The methodology used herein allowed us to obtain genotype information from thousands of variable sites dispersed across the entire genomes of these fishes, with which we achieved a higher level of resolution of patterns of genetic variation than previous studies. We generated DNA-sequence data using the Illumina GAII platform for 192 fish: 41 *P. formosa* (5 localities where *P. formosa* is sympatric with *P. mexicana* and 6 localities where it is sympatric with *P. latipinna*), 82 *P. latipinna* (from 22 localities across Louisiana, Texas, and Mexico), and 69 *P. mexicana* (from 13 localities across Mexico and Honduras; Fig. 1). We identified 26,313 single-nucleotide polymorphism (SNP) markers that were used for the analyses.

Our first goal was to confirm the hybrid origin of *P. formosa* and determine the genomic contribution from its parent species. Our results are in agreement with the findings of several previous studies: *P. formosa* is a hybrid between *P. mexicana* and *P. latipinna* (11, 14–19). A principal components analysis (PCA) of the genotype estimates of each individual at each locus (Fig. 1B–D), an admixture analysis performed in STRUCTURE with K = 2 (forcing individuals to be assigned to one of only two clusters) (Fig. 2), calculation of genome-average pairwise GST (Fig. 3), and the hybrid index for each *P. formosa* (Fig. 3) all suggest that gynogenetic individuals have a genome that is intermediate between *P. latipinna* and *P. mexicana*. Principal component axes 1 (PC1) and 2 (PC2) collectively explained 63% of the genetic variation (Fig. 1B). PC1 appears to divide the three species into three distinct clusters, with *P. formosa* in an intermediate position between the parent species, consistent with the hypothesis of a hybrid origin for *P. formosa* (Fig. 1B). PC2 separates the asexual species from the sexual species and illustrates the variation in genotypes among individuals (Fig. 1B). Interestingly, the genotypic

---

Fig. 1. Sampled populations (A) and PCA plots for the 192 individuals based on the genotype probabilities at each locus (B–D). Dark blue, north *P. latipinna*; blue, central *P. latipinna*; light blue, south *P. latipinna*; orange, north *P. formosa* sympatric with *P. latipinna*; red, south *P. formosa* sympatric with *P. mexicana*; light green, north *P. mexicana*; green, central *P. mexicana*; dark green, south *P. mexicana*. 
variation within *P. formosa* was high: individuals of *P. formosa* did not all cluster together but were differentiated along the second principal component just as much as the sexual individuals. PC3 explained 5.7% of the variation and separated the populations of *P. mexicana* into three groups, which corresponded to three geographic regions (north, central, and south) (Fig. 1C). PC4 explained 4.0% of the variation and divided *P. latipinna* into two geographic regions: north plus central (the central group is comprised of populations in central Texas that were introduced from the north group, including population in Florida and the Gulf Coast of the United States) and south (Fig. 1D).

Estimates of pairwise GST also confirmed the genomic intermediacy of *P. formosa* (Fig. S1 and Table 1). Differentiation between *P. latipinna* and *P. mexicana* was approximately twice that observed between *P. formosa* and either parent species (Table 1). Within-species (among-population) differentiation was an order of magnitude lower, and lowest in *P. formosa*, although distinctly nonzero. Estimation of hybrid indices for the asexual individuals was also consistent with the hypothesis of hybrid origin: for the 41 *P. formosa* examined, hybrid index ranged from 0.37 to 0.56 (mean, 0.49; Fig. 3). These estimates correspond to three geographic regions (north, central, and south) (Fig. 1C). PC4 explained 4.0% of the variation and divided *P. latipinna* into two geographic regions: north plus central (the central group is comprised of populations in central Texas that were introduced from the north group, including population in Florida and the Gulf Coast of the United States) and south (Fig. 1D).

Estimates of pairwise GST also confirmed the genomic intermediacy of *P. formosa* (Fig. S1 and Table 1). Differentiation between *P. latipinna* and *P. mexicana* was approximately twice that observed between *P. formosa* and either parent species (Table 1). Within-species (among-population) differentiation was an order of magnitude lower, and lowest in *P. formosa*, although distinctly nonzero. Estimation of hybrid indices for the asexual individuals was also consistent with the hypothesis of hybrid origin: for the 41 *P. formosa* examined, hybrid index ranged from 0.37 to 0.56 (mean, 0.49; Fig. 3A). These estimates are in agreement with the estimates of admixture proportion calculated using STRUCTURE (Fig. 2A).

Given the patterns and diversity captured by both the estimates of hybrid index and the PCA, we calculated the probability of ancestry for each locus in *P. formosa* given the two parent populations and the individual’s hybrid index. This estimate is calculated using the Bayesian genomic cline model (22, 23) and is summarized in the α parameter. The α parameter is a population-level parameter that specified whether, on average, individuals are less or more likely to have either parent species’ ancestry given their hybrid index (which is a genome-wide measure of ancestry). The parameter α provides information about ancestry for a locus across individuals and not allelic state; thus, we used α to examine the organization of the genome of *P. formosa* with respect to ancestry. For frozen F1 hybrids obtained from parent species that are differentiated at all loci, we expect hybrid indexes of 0.5 (consistent with our mean estimates for *P. formosa*) and values of the cline parameter α to be zero across all loci, indicating no excess ancestry. However, our results suggest that *P. formosa* is either not a frozen F1 hybrid derived from a single F1 individual or substantial evolution has occurred in *P. latipinna* and *P. mexicana* since they gave rise to *P. formosa*. As shown in Fig. 3B, most loci in *P. formosa* have parameter α estimates not different from zero, indicating no excess ancestry for either parent species at those loci. However, about 12% of the loci have excess ancestry from one or the other parent species, consistent with the conclusion that genetic recombination might have occurred in *P. formosa*. An alternative explanation for these results could be that the allele frequencies of the parent species at the time of hybridization were much more similar than they are now. Interestingly 15% of the SNPs analyzed in this study are polymorphic in all three species, whereas 12.5% are shared only by *P. formosa* and *P. latipinna* and 21.6% are shared by *P. formosa* and *P. mexicana*. These results suggest that possibly *P. latipinna* and *P. mexicana* were more similar at the time of hybridization but have since diverged because of drift, mutations, and selection. In fact, only 1.8% of SNPs are now shared exclusively between the two sexual species, indicating that they are now clearly differentiated (as also suggested from the estimates of GST).

The possibility that our results are attributable only to the fact that the parent populations were more similar at the time of hybridization and differentiated since is not a trivial one. To estimate both the hybrid index and α parameter we must assign

---

*Fig. 2.* Results from STRUCTURE clustering analyses. Admixture proportion for K = 2 (A) and mean assignment probabilities to cluster 1 ±95% credible intervals for K = 2 (B). Admixture proportions for K = 4 (C) and mean assignment probability ±95% credible intervals for K = 4 for cluster individuals were assigned to D. K = 2 and K = 4 were chosen as the appropriate number of groups after examination of the log of the marginal likelihood and the ad hoc delta(K) statistic (41) for each K (Fig. 5A).
putative parent species to *P. formosa*. We chose populations of *P. mexicana* found in the northern part of the species’ range and populations of *P. latipinna* found in the southern part of its range because previous work has suggested that the region of Tampico, Mexico, is where *P. formosa* has originated (13, 17). However, this ad hoc decision might have influenced the estimates of $\alpha$, given that the allele frequencies of the sexual individuals found in those populations now might be significantly different from the allele frequencies during the time of hybridization. To address this possibility, we also estimated $\alpha$ for all of the other parent population combinations and found that the locus-specific $\alpha$ estimates did not correlate between different parent population combinations (Table S1 and Fig. S2). The patterns found during the estimation of $\alpha$ across the genome can be explained by (i) recombination in the gynogenetic species coupled with differentiation between the parent populations or (ii) high differentiation in the parent species (and populations within each parent species) after the hybridization event. The estimation of linkage and Hardy–Weinberg (HW) disequilibria (Fig. S3) provide evidence to support the hypothesis that some recombination has occurred in *P. formosa*. In fact, calculations of a Burrow’s composite measure of linkage disequilibrium demonstrate that *P. formosa* exhibits less disequilibrium than expected for an asexual lineage of hybrid origin descending from a single F1 individual, which suggests that recombination has reduced the linkage disequilibria created by admixture (SI Results and Discussion and Fig. S3).

To rule out the hypothesis that the excess ancestry recorded was attributable to recent gene flow with either of the parent species, we compared the estimations of the $\alpha$ parameter for the northern and southern populations of *P. formosa*, where *P. formosa* is sympatric with *P. latipinna* and *P. mexicana*, respectively. This analysis revealed no difference in the pattern of parent contribution (Fig. 3C). Thus, the excess ancestry for $\sim$12% of the loci surveyed does not vary with geography across *P. formosa* and does not appear to be a function of recent gene flow from the parent species. Although previous studies have found that microchromosomes might be inherited by *P. formosa* (24), we found no evidence to support their results. However, given that we only used SNPs that had a minimum coverage of five reads per population, and microchromosome inheritance has been recorded only in a handful of populations not sampled extensively in this study, it is possible that microchromosomes were not included in our analysis at all because they are not present in every population.

Together with investigating the ancestry of *P. formosa*, we tried to shed light on the allelic state of the loci used for the analyses. The estimation of observed heterozygosity across loci revealed patterns that are inconsistent with the hypothesis that *P. formosa* is a frozen F1 hybrid descended from a single F1 individual (Fig. S4). Specifically, loci with observed heterozygosities

---

**Table 1. Summary of genome-wide $G_{ST}$ calculation**

<table>
<thead>
<tr>
<th>Pairwise comparisons</th>
<th>$n$</th>
<th>Mean $G_{ST}$</th>
<th>Maximum $G_{ST}$</th>
<th>Minimum $G_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intraspecific comparisons</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>P. latipinna</em> vs. <em>P. latipinna</em></td>
<td>3</td>
<td>0.06 ± 0.02</td>
<td>0.043</td>
<td>0.081</td>
</tr>
<tr>
<td><em>P. mexicana</em> vs. <em>P. mexicana</em></td>
<td>3</td>
<td>0.096 ± 0.02</td>
<td>0.077</td>
<td>0.125</td>
</tr>
<tr>
<td><em>P. formosa</em> vs. <em>P. formosa</em></td>
<td>1</td>
<td>0.028</td>
<td>Upper bound: 0.0278</td>
<td>Lower bound: 0.0288</td>
</tr>
<tr>
<td>Interspecific comparisons</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>P. latipinna</em> vs. <em>P. mexicana</em></td>
<td>9</td>
<td>0.361 ± 0.01</td>
<td>0.341</td>
<td>0.380</td>
</tr>
<tr>
<td><em>P. mexicana</em> vs. <em>P. formosa</em></td>
<td>6</td>
<td>0.155 ± 0.02</td>
<td>0.132</td>
<td>0.180</td>
</tr>
<tr>
<td><em>P. formosa</em> vs. <em>P. latipinna</em></td>
<td>6</td>
<td>0.163 ± 0.01</td>
<td>0.144</td>
<td>0.170</td>
</tr>
</tbody>
</table>

Mean and range of $G_{ST}$ for comparisons between populations of *P. latipinna* and *P. mexicana* and for comparison within species among populations of *P. latipinna* (North, Central, and South; Fig. 1) and *P. mexicana* (North, Central, and South; Fig. 1). Mean and upper and lower bounds provided for the comparison between the two regions (North and South) of *P. formosa*. 

---

*Fig. 3.* Posterior probability of estimates of hybrid indices and cline parameter $\alpha$ for *P. formosa*. The assigned putative parent populations were *P. mexicana* found in the northern part of its range and *P. latipinna* found in the southern part of its range. A depicts the posterior probability distribution of the hybrid index estimates ($\pm$95% credible intervals) for the 41 *P. formosa* used in this study. The vertical dotted line in A divides *P. formosa* sympatric with *P. latipinna* (Left) from *P. formosa* sympatric with *P. mexicana* (Right). Each black line in B represents the 95% credible interval for the estimate of the cline parameter $\alpha$ for each of the 26,313 SNPs, for all 41 *P. formosa*. C shows the correlation between the alpha estimates at each locus in the populations of *P. formosa* in the south vs. populations in the north.
different from 0 (for loci at which the parents are not differentiated) or 1 (for loci at which the parents are differentiated) are not expected in F₁ hybrids descended from a single F₁ individual (which would be consistent with the hypothesis of a single hybridization forming the mother of all P. formosa). The variation in heterozygosity across loci (Fig. S4) is, instead, consistent with a short history of backcrossing, perhaps before the onset of gynogenesis or with the hypothesis that P. formosa is comprised of descendants of multiple, independently formed F₁ gynogenetic individuals. However, the data gathered for the present study and work performed in other laboratories suggest that a short history of recombination is more likely than the hypothesis of independent origins of clonally reproducing F₁ hybrids. The ancestral P. formosa might have been a sexually reproducing hybrid for some time before becoming gynogenic, as hypothesized by Turner et al. (16), and later by Stöck et al. (19). Unsexual sperm-dependent organisms might be rare because certain genetic compositions (and possibly specific epistatic interactions) are necessary for the evolution of gynogenesis. Stöck et al. (19) referred to this idea as the “rare formation hypothesis.” Perhaps the specific combination of alleles required for gynogenesis only occurred after some cycles of recombination and independent assortment of alleles. This possibility could explain why no one has been able to reproduce P. formosa in the laboratory, even after extensive attempts to do so (13–16, 19–21).

A recent study also found some loci in P. formosa that were homoyzogous for one of the parent species, and the authors suggested that mitotic gene conversion (or mitotic recombination) during gamete formation might explain the pattern (18, 19). Our results are consistent with the hypothesis of mitotic gene conversion. When this particular type of recombination occurs, some loci become homozygous for one of the alleles (25), causing a loss of heterozygosity and an increase in linkage disequilibria. The probability of the occurrence and success of gene conversion varies across the genome (26). This mechanism causes genotypes to vary among individuals and causes a decay of admixture linkage disequilibria because recombination within admixed individuals and between chromosomes of different ancestry occurs. Thus, mitotic gene conversion could potentially explain the observed variation in ancestry across loci in P. formosa and the variation among individuals of P. formosa.

The study was designed to determine the amount of genotypic variation present within P. formosa. The GST calculations and the PCA suggest relatively high genotypic diversity in P. formosa. This observation is in agreement with previously published results (17, 27, 28), which all found that P. formosa was genotypically variable. Stöck et al. (19) suggest that the high genetic diversity in P. formosa is attributable to high mutation rates because the phylogenetic analyses of mtDNA variation suggested a monophyletic origin of P. formosa (19). Turner et al. (27) also suggested that high mutation rates are more probable than multiple hybrid origins based on allozyme data. However, some of the P. formosa studied by Turner et al. (27) were collected in the population where triploid individuals are present, and, therefore, the high clonal diversity found in this population in the Rio Purification might have been caused by the presence of triploids. We did not include triploids in this study. To address the possibility suggested in previous studies, that high mutation rates within P. formosa contribute to high genetic variation, we calculated the percentage of variable SNPs private to P. formosa and found that 3% of the SNPs used in this study are only found in the gynogenetic species. Thus, mutation accumulation within P. formosa has been moderate and does not fully explain the genotypic diversity in this species.

Our analyses using more than 25,000 SNPs from across the genome of P. formosa document considerable genotypic variation within this gynogenetic species. Given the complexity of the genomic patterns across loci and among individuals, it is not currently possible to make definitive inferences about the number of clonal lineages or the number of hybridization events involved in the origin of P. formosa. We can conclude, however, that our results are consistent with the hypothesis that the formation of P. formosa as a unisexual species is the result of hybridization and possible subsequent recombination. Recombination might have occurred following hybridization and the resulting genotypic variation has obscured our ability to discern distinct lineages. Results from calculation of genetic distances between individuals across loci illustrates that individual P. formosa are not identical to one another (Fig. S5). However, differentiation between P. formosa from different geographic regions appears to be less than among populations of either parent species (Fig. S5). Recombination following hybridization (i.e., recombination via sexual reproduction) could have occurred before the onset of the gynogenetic reproductive mode, resulting in numerous genotypes (as suggested from the calculation of observed heterozygosity). Alternatively, some form of asexual recombination, most likely mitotic gene conversion or automixis (29, 30, 31), might have occurred (or might still be occurring). Similar mechanisms for the production of genetic diversity have been proposed for the unisexual lizards in the genus Darevskia [formerly Lacerta, Lacertidae (32, 33)] and have been shown to possibly play a significant role in the maintenance of asexual species (8).

Regardless of the origins of genetic variation in P. formosa, it is clear that this variation could contribute to the persistence of this species. Coexistence between a unisexual sperm-dependent species, and its host can be achieved and maintained if genetic variation is present in a population because natural selection can select against the clones that overlap extensively in resource use with their host species (frozen niche variation) (34, 35). Intriguingly, some form of asexual recombination, such as mitotic gene conversion or automixis, might also facilitate a reduction in the rate of accumulation of deleterious mutations (2) and increase the longevity of P. formosa beyond what is predicted by theoretical models (36).

Materials and Methods

In-depth descriptions of the protocols, models, and calculations used are in SI Materials and Methods. This work was performed under Institutional Animal Care and Use Committee no. 0818-0325-18.

Next-generation DNA sequence data from 192 fish were generated with the Illumina GAII platform following recently developed methods (for more details see, refs. 37 and 38 and SI Materials and Methods). A total of 32,492 variable sites were identified using custom Perl scripts together with samtools and bedtools. Because of the low numbers of individuals sampled from each locality, we pooled individuals across localities into eight geographical regions to obtain adequate sample sizes to perform all of our analyses (Fig. 1A).

We trimmed data to only those SNPs with a minimum of five reads per km. By region (population grouping), which produced 26,313 SNPs, and then used Bayesian hierarchical models to estimate allele frequencies for each locus based on the observed data by using the allele frequency Bayesian model presented in Gompert and Buerkle (37). We summarized population genetic structure at the individual level using both a PCA and STRUCTURE 2.2 (40, 41). We also summarized population genetic structure at the population level by calculating pairwise ΦST statistics (42) for all combinations of regional groupings (Fig. 3).

To investigate the genomic composition of P. formosa, we used a Bayesian approach to estimate hybrid index for all P. formosa individuals and to assess ancestry (relative to the putative parental species) at all SNP loci for all individuals. To obtain a clearer picture about the robustness of the pattern obtained using the putative parent populations (P. latipinna and P. mexicana in northern Mexico), we repeated the estimation of both the hybrid index (Fig. 52) and for all combinations of possible parent populations and calculated the correction coefficient between these new estimates vs. the estimates obtained with the putative parent populations (Table S1).

Results from the genomic clines analysis (Results and Discussion) provided possible evidence of a history of recombination in P. formosa. This was unexpected given the asexual nature of P. formosa; and the predicted lack of recombination in this asexual species. Consequently, we predicted substantially higher linkage disequilibrium in this species compared with the parental species. We, therefore, calculated Burrow’s composite measure of

Alberici da Barbiano et al.
linkage disequilibrium (∆) between all pairs of variable loci (43, 44) and performed a set of simulations to untangle the effects of linkage disequilibrium and HW disequilibrium on ∆.

To further investigate the results of the linkage disequilibrium estimation and to better understand the allelic state of the loci analyzed, we calculated the observed heterozygosity for each locus in each population (Fig. 54).

To address the question of whether mutation accumulation only can explain the variation present in _P. formosa_ (as suggested by previous studies), we calculated the proportion of variable SNPs private to _P. formosa_, _P. mexicana_, and _P. latipinna_, as well as the proportion of SNPs shared by all species and by only two species.

As an alternative means of illustrating the genotypic variation observed within _P. formosa_ (Fig. 18), we calculated the “genotypic distance” between each pair of individuals at each locus as a measure of genotypic dissimilarity among individuals (Fig. 55).

ACKNOWLEDGMENTS. We thank Michael Sandel, Ryan Earley, Margaret Ptecek, Jose Rainha Zúñiga-Vega, Kristen Davis, and Celeste Espinido for collecting fish; Alex Buerkle for computational help and suggestions; Tom Parchman for help with scripts; Kate Bell, Lauren Lucas, Darrin Hulsey, and Mariana Mateos for suggestions; the Mexican government for collecting permit DGOPA.0711.210612.1750; the Texas State University computer cluster; and the University of Wyoming Seismic computer cluster. This work was funded by the National Science Foundation (Integrative Organismal Systems Grant IOS-1021873 to C.R.G., A.S.A., and C.C.N.; and Division of Environmental Biology Grant DEB-1050355 to C.C.N.). Additional funding was provided by the Raney Fund Award (American Society for Ichthyologists and Herpetologists), the Theodore Roosevelt Memorial Grant (American Museum of Natural History), the Howard McAlpine Student Research Award (Southwestern Association of Naturalists), the Vern Parish Grant (American Livebearer Association), and the Texas Academy of Science Student Research Grant (Texas Academy of Science) (to L.A.D.B.).

Supporting Information

Albericida da Barbiano et al. 10.1073/pnas.1303730110

SI Materials and Methods
Next-generation DNA sequence data were generated with the Illumina GAIIX platform following recently developed methods (for more details see 1, 2). We generated DNA-sequence data for 192 fish: 41 Poecilia formosa sampled from 5 localities where *P. formosa* is sympatric with *Poecilia mexicana* and 6 localities where *P. formosa* is sympatric with *Poecilia latipinna*; 82 *P. latipinna* from 22 localities across Louisiana, Texas, and Mexico; and 69 *P. mexicana* from 13 localities across Mexico and Honduras (Fig. 1A). We isolated and purified DNA from caudal fin clips following the Gentra Systems PURGENE DNA-isolation protocol. We fragmented the genomic DNA using restriction enzymes (EcoRI and MseI) to generate a genomic DNA library for each individual. Adapters with Illumina primer sites were ligated to the ends of these fragments. Each individual’s library of genomic fragments was labeled by the inclusion of a unique 10-bp-long identification sequence (i.e., barcode) added to the EcoRI adapter (1). Individual libraries were amplified with two rounds of PCR using the Illumina primers, after which PCR products were pooled across all individuals. The result is a pooled library for all 192 individuals, with fragments uniquely identified to individual with a 10-bp barcode. We then separated fragments on a 2% (mass/vol) agarose gel and isolated fragments between 250 and 500 bp in length by cutting the gel. We used the Qiagen Gel Extraction 15 Kit (catalog no. 28706; Qiagen) to purify these fragments. This reduced-complexity genomic-DNA library was sequenced at the National Center for Genome Research using the Illumina GAIIX platform.

The resulting sequence reads were processed using a series of quality control steps to identify variable sites following the methods of Gompert et al. (1). Briefly, a custom Perl script was used to identify sequences to individual based on the barcode sequence, remove the 10-bp barcode and 6-bp EcoRI cut site, and remove reads that contained adapter sequences or were of poor quality. We used SeqMan NGEN 3.0.4 (DNASTAR) to perform a de novo assembly using a subset of sequence reads (8 million) and concatenated the consensus sequences from the resulting contigs to create an artificial chromosome for reference-based assembly of the entire data set. This reference included 215,622 consensus sequences, each 90-bp long. We assembled the full dataset (43 million sequences) to this artificial reference using SeqMan NGEN 3.0.4 (DNASTAR). We then used custom Perl scripts (available at http://uweb.txstate.edu/~la1122/LauraAlbericidaBarbiano/Home.html), together with samtools and bcftools (3) to identify variable sites. Base quality scores were incorporated into the identification of variable sites, and SNPs were only called if at least 25% of the individuals had data for that locus. We identified 32,492 variable sites.

Because of the low numbers of individuals sampled from each locality, we pooled individuals across localities into eight geographical regions to obtain adequate sample sizes to perform all of our analyses (Fig. 1A). Regional groupings included three geographical regions for *P. latipinna*: north (Florida, Louisiana, and East Texas), central (populations in central Texas), and south (south Texas and northern Mexico). For *P. formosa*, grouping included two regions: north (localities sympatric with *P. latipinna*; these also included populations found in central Texas where individuals of *P. formosa* were introduced from Brownsville, TX) and south (localities sympatric with *P. mexicana*). Three regions were identified for *P. mexicana*: north (northern Mexico), central (central Mexico), and south (southern Mexico, Yucatan Peninsula, and Honduras; Fig. 1A).

Population Genetic Analyses. We trimmed data to only those SNPs with a minimum of five reads per marker per region (population grouping), which produced 26,313 SNPs. We used Bayesian hierarchical models to estimate allele frequencies for each locus based on the observed data by using the allele frequency Bayesian model presented in Gompert and Buerkle (4), which is similar to the models used by Pritchard et al. (5), Gillespie et al. (6), and Hedrick (7). Two assumptions of the model are that (i) the data do not contain errors (this is a simplification of the reality of our data) and (ii) sequenced is sampled stochastically and have a limited coverage for each nucleotide. The model treats the genotypes of individuals at each locus and the population allele frequencies as unknown model parameters, which are estimated from the sequence data (for more details on the model, see ref. 2). The allele frequency model was written by Z.G. and relies on the GNU Scientific Library (8). The posterior probabilities for parameter estimates (allele frequencies for each population and genotypes for each individual for each locus) were obtained using Markov Chain Monte Carlo (MCMC) of 20,000 steps, and we retained samples every 10th step. Mixing of the chains was diagnosed using the coda package in R. To determine whether the different chains converged, we visually inspected the density distribution of the posterior probabilities, as well as calculated the Gelman and Rubin diagnostic. Chains were accepted only if the diagnostic resulted in a value of 1, indicating convergence among the chains. Additionally, only runs that yielded effective sample sizes of at least 150 for a randomly selected parameter values were accepted.

We summarized population genetic structure at the individual level in two ways. First, a principal component analysis (PCA) was performed using the genotype posterior probabilities for the three genotypes for each individual for each SNP locus as variables. We used the covariance matrix to produce the PCA in R (using the prcomp function in the composition package in R) to center but not scale the genotype probabilities. Second, we used the admixture model in STRUCTURE 2.2 (5, 9). For this analysis, we sampled one sequence for each SNP locus for each individual in proportion to the frequency of reads for each individual at that locus. Individuals were, thus, assigned a 1 or a 0 depending on which SNP sequence was sampled for that individual and a -9 (missing information) for the alternative allele for each locus (script written by Tom Parchman, University of Wyoming, Laramie, WY). This was done so to have an infill similar to those used for dominant markers, where heterozygosity at a locus cannot be verified. We sampled individual reads in this way from 500 random loci. The admixture model in STRUCTURE was then used to estimate the admixture proportions of each of K groups. The model was run for K = 2–9 (number of geographical regions + 1) and each analysis of K groups was repeated 10 times. MCMC chains of 80,000 steps with a burn-in of 30,000 were used for each analysis. To estimate the appropriate number of groups (K), the log of the marginal likelihood for each run was plotted against K and the ad hoc delta(K) statistic was calculated and plotted against K. We used the assignment probabilities for *P. formosa* for K = 2 as an estimate of admixture proportion.

We also summarized population genetic structure at the population level by calculating pairwise GST statistics (10) for all combinations of regional groupings. Pairwise GST estimates were summarized using nonmetric multidimensional scaling (NMDS) to ordinate populations. NMDS was performed using the Modern Applied Statistics with S package in R, and a plot of the first three dimensions was used to display genetic structure at the population level (Fig. S1).
Genomic Clines Analyses. To investigate the genomic composition of *P. formosa*, we used a Bayesian approach to estimating hybrid index for all *P. formosa* individuals and assigning ancestry (relative to the putative parent species) at all SNP loci for all individuals. We used the Bayesian genomic cline model (11) to estimate the hybrid index of the 41 *P. formosa* given their putative parent populations as prior information. We set populations of *P. latipinna* found in the southern part of its range (*P. latipinna* south) and populations of *P. mexicana* found in the northern part of its range (*P. mexicana* north; Fig. 1A) as the putative parent populations. It is not known exactly where *P. formosa* originated, but genetic evidence points to the region of Tampico [corresponding to the southern portion of the range of *P. latipinna* and the northern range of *P. mexicana* (12)]. The cline parameter h (hybrid index) is the probability of ancestry of an admixed individual given two parent populations and is equivalent to an estimate of admixture proportion (11, 13–15). We were also specifically interested in determining whether the genomic compositions of individual *P. formosa* differ from that of *F₁* hybrids or, alternatively, more complicated hybrid classes. Cline parameter α, a locus-specific component of the Bayesian genomic cline model, denotes an increase or decrease in the probability of parent 1 ancestry (in this case *P. latipinna*) relative to a null expectation based on the hybrid index (11, 13–15). Given a hybrid index, if there is excess contribution from either parent species, then the α index will be different from 0. The calculation of cline parameter α provides an estimate of excess ancestry relative to hybrid index for each locus (13–15). We ran five chains for 80,000 steps with 20,000 burn-ins to estimate both the hybrid index for all individuals and the α index for all loci. To obtain a clearer picture about the robustness of the pattern obtained using the putative parent populations, we repeated the estimation of both the hybrid index (Fig. S2) and α for all combinations of possible parent populations. We then calculated the correlation coefficient between the estimates obtained with the new combinations vs. the estimates obtained with the putative parent populations (Table S1). Given that the correlations were performed using the point estimate of α, we then calculated how many loci with α estimates equal, lower, and greater than 0 are shared by the putative parent populations and the other possible parent populations (Table S1; R Script available at http://uweb.txstate.edu/~la1122/LauraAlbericidaBarbiano/Home.html).

Linkage and Hardy–Weinberg Disequilibria. Results from the genomic clines analysis (SI Results and Discussion) provided possible evidence of a history of recombination in *P. formosa*. This was unexpected given the hybrid origin of *P. formosa* and the presumed lack of recombination in this asexual species. Consequently, we predicted substantially higher linkage disequilibrium in this species compared with the parent species. We, therefore, calculated Burrow’s composite measure of linkage disequilibrium (Δ) between all pairs of variable loci (16, 17). We calculated Δ between each pair of loci (Δij) iteratively 75 times following the formula found in Weir (16) and Zaykin (17) by using the estimated genotype posterior probabilities. We created a matrix with genotype counts at each locus and then calculated ΔAB = (1/n)(2nAABB + nABB + nABB + 1/2nAAB) − 2pABpB, where n is the number of individuals in the sample, n denotes the genotypic counts for each pair of loci (A and B), and p denotes the allelic frequencies at each locus (16, 17). We then averaged the 75 iterations to obtain a mean linkage disequilibrium for each pair of loci and obtained a final matrix with values of Δ for each pair of loci for each one of our populations (script available in the Dryad Digital Repository). For each geographic region, we calculated the average Δ across all pairs of loci using a custom R script and then plotted the distribution of Δ (Fig. S2A). As a comparison, we then calculated Δ for a simulated population of *P. formosa* created by sampling genotypes for a population of 41 synthetic hybrids between *P. mexicana* and *P. latipinna* given the allele frequencies of *P. latipinna* and *P. mexicana* found in the area of the original hybrid event. (Script was written and is available at http://uweb.txstate.edu/~la1122/LauraAlbericidaBarbiano/Home.html; Fig. S2A)

A second set of simulations was performed to untangle the effects of linkage disequilibrium and Hardy–Weinberg (HW) disequilibrium on Δ. We created two parent populations fixed for opposite alleles at 10 loci. We then created multiple filial populations by sampling alleles according to their frequency in the parent population, so that filial populations varied in the proportion of heterozygotes; Δ was then calculated for each filial population (ref. 16 and Fig. S2B).

Observed Heterozygosity. To further investigate the results of the linkage disequilibrium estimation and to better understand the allelic state of the loci analyzed, we calculated the observed heterozygosity for each locus in each population (Fig. S4). The observed heterozygosity was calculated for each locus by averaging the posterior probability of a locus being heterozygous among all of the chains of the genotype probability model (Bayesian hierarchical model explained previously).

Private Alleles. To address the question of whether mutation accumulation only can explain the variation present in *P. formosa* (as suggested by previous studies), we calculated the proportion of variable SNPs private to *P. formosa*, *P. mexicana*, and *P. latipinna*, as well as the proportion of SNPs shared by all species and by only two species (custom Perl script is available at http://uweb.txstate.edu/~la1122/LauraAlbericidaBarbiano/Home.html).

Genotypic Distance Among Individuals. As an alternative means of illustrating the genotypic variation observed within *P. formosa* (Fig. 1B), we calculated the “genotypic distance” between each pair of individuals at each locus as a measure of genotypic dissimilarity among individuals. We first calculated the mean genotypic of each individual at each locus by multiplying the probabilities of being homozygous for one allele, heterozygous or homozygous for the alternative allele by 0, 1, or 2 respectively, and then summed the values, which provides a “mean genotype.” For each pair of individuals, we then took the difference between the genotypes at all loci and averaged across loci to obtain the overall genetic distance between the two individuals. We summarized the results in R using the image.plot function in the fields package. This process was used to calculate distances between all 192 individuals used in this study (Fig. S5).

SI Results and Discussion

Hybrid Index and α. The results obtained from the calculation of the parameter (Bayesian genomic cline analysis) suggest that 12% of the surveyed loci in *P. formosa* show an excess ancestry from the parent species, possibly indicating a history of recombination. However, much of the genome appears to remain presumed lack of recombination in this asexual species. Consequently, we predicted substantially higher linkage disequilibrium in this species compared with the parent species. We, therefore, calculated Burrow’s composite measure of linkage disequilibrium (Δ) between all pairs of variable loci (16, 17). We calculated Δ between each pair of loci (Δij) iteratively 75 times following the formula found in Weir (16) and Zaykin (17) by using the estimated genotype posterior probabilities. We created a matrix with genotype counts at each locus and then calculated ΔAB = (1/n)(2nAABB + nABB + nABB + 1/2nAAB) − 2pABpB, where n is the number of individuals in the sample, n denotes the genotypic counts for each pair of loci (A and B), and p denotes the allelic frequencies at each locus (16, 17). We then averaged the 75 iterations to obtain a mean linkage disequilibrium for each pair of loci and obtained a final matrix with values of Δ for each pair of loci for each one of our populations (script available in the Dryad Digital Repository). For each geographic region, we calculated the average Δ across all pairs of loci using a custom R script and then plotted the distribution of Δ (Fig. S2A). As a comparison, we then calculated Δ for a simulated population of *P. formosa* created by sampling genotypes for a population of
not definitive. However, we found no evidence against the hypothesis that recombination has occurred in *P. formosa*.

The estimation of hybrid indices in all of the possible combinations of parent populations suggests that the putative parent populations (*P. latipinna* south and *P. mexicana* north) are very likely the correct populations to use for the estimation of $\alpha$. In fact, any estimation of the hybrid index that included one of the two putative parent populations shifted the hybrid index more toward $1$ (when *P. mexicana* north was considered) and more toward $0$ (whenever *P. latipinna* south was considered), suggesting that these two populations are more similar to *P. formosa* than any other population of the parent species (Fig. S2).

**Linkage Disequilibrium.** Given the clonal and hybrid nature of *P. formosa*, we expected to observe substantial linkage disequilibrium resulting from admixture. When we calculated the Burrow’s composite measure of HW and linkage disequilibrium, $\Delta$, however, we found that the distribution of $\Delta$ in our asexual population was not much different from that of its sexual parents (Fig. S3). To try to interpret this result, we compared the results of the $\Delta$ calculations (Fig. S3A) to the results obtained from a simulated hybrid dataset (Fig. S3B), but we were unable to make confident inferences about how much linkage disequilibrium is present in *P. formosa*. A possible cause for the lack of clear results from the calculation of $\Delta$ is that the estimation is confounded by pooling sampling localities into geographic regional groups, which might create a Wahlund effect.

More work is necessary to properly understand the results obtained for $\Delta$ in *P. formosa*. For example, performing the same analyses with large samples from each locality will remove the confounding effects of pooling individuals into regional “populations” and provide more precise inferences about the amount of recombination that has occurred or is occurring in *P. formosa*.


![Fig. S1. Nonmetric multidimensional scaling of pairwise G_{ST} between all populations.](image-url)
Fig. S2. (A–I) Estimates of hybrid index for all *P. formosa* from all combinations of parent species populations. *G* depicts the hybrid index of individual *P. formosa* obtained from the putative parent populations (*P. latipinna* in the south of its range and *P. mexicana* in the north part of its range). For simplicity and clarity, names of parent populations were abbreviated to Pl for *P. latipinna* and Pm for *P. mexicana*; the third letter in the abbreviation refers to the geographic region of the populations (n, north; c, central; s, south).
Fig. S3. Mean genome-wide values of Burrow’s composite measure of linkage disequilibrium (Δ) for each geographic region (A) and values of Δ relative to the effects of linkage and HW disequilibria (B). Dotted line in A separates asexual (Left) from sexual (Right) populations. Values for B were obtained from a simulated dataset of parent populations fixed for opposite alleles, and filial populations that varied in the number of heterozygous offspring. The simulations show that a Δ value of ~0 (similar to the value we obtained with our dataset) is obtained when both linkage disequilibrium and HW disequilibrium are at their maxima.
Fig. S4. Distributions of observed heterozygosity across all SNPs for *P. latipinna* (A–C), *P. mexicana* (D–F), and *P. formosa* (G and H).
Fig. S5. Mean genetic distance among all individuals (A) and among individual *P. formosa* only (B). *Poecilia formosa* appears to be genetically intermediate between *P. latipinna* and *P. mexicana* (A). A high amount of genotypic variation can be detected within *P. formosa* (B). Note that a different scale of values is used in the legends of A and B.
Fig. S6. Summary of the likelihood of $K$ ($L(K)$) given the number of clusters (A) and calculation of delta($K$) (14) (B). $K = 2$ and $K = 4$ appear to be the best to describe the data. See Materials and Methods and Results and Discussion for information about these results.

Table S1. Correlations between the estimation of the $\alpha$ parameter among all of the possible parent species population combinations vs. the estimates obtained from the two putative parent populations (P. latipinna south and P. mexicana north)

<table>
<thead>
<tr>
<th>Combination of parent population vs. putative parent populations</th>
<th>Correlation coefficient of $\alpha$ estimates</th>
<th>No. of shared loci with $\alpha &gt; 0$</th>
<th>No. of shared loci with $\alpha = 0$</th>
<th>No. of shared loci with $\alpha &lt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. latipinna north + P. mexicana north</td>
<td>0.51</td>
<td>1,871</td>
<td>24,370</td>
<td>72</td>
</tr>
<tr>
<td>P. latipinna north + P. mexicana central</td>
<td>0.39</td>
<td>934</td>
<td>23,943</td>
<td>1,436</td>
</tr>
<tr>
<td>P. latipinna north + P. mexicana south</td>
<td>0.30</td>
<td>873</td>
<td>24,012</td>
<td>1,428</td>
</tr>
<tr>
<td>P. latipinna central + P. mexicana north</td>
<td>0.50</td>
<td>1,875</td>
<td>24,369</td>
<td>69</td>
</tr>
<tr>
<td>P. latipinna central + P. mexicana central</td>
<td>0.24</td>
<td>921</td>
<td>25,321</td>
<td>71</td>
</tr>
<tr>
<td>P. latipinna central + P. mexicana south</td>
<td>0.22</td>
<td>837</td>
<td>24,312</td>
<td>1,164</td>
</tr>
<tr>
<td>P. latipinna south + P. mexicana central</td>
<td>0.60</td>
<td>765</td>
<td>23,583</td>
<td>1,965</td>
</tr>
<tr>
<td>P. latipinna south + P. mexicana south</td>
<td>0.47</td>
<td>696</td>
<td>23,728</td>
<td>18,889</td>
</tr>
</tbody>
</table>

The last three columns provide the number of loci shared by the putative parent populations and the potential parent populations with $\alpha > 0$, $\alpha = 0$, or $\alpha < 0$. 