Supporting Information

Alberici da Barbiano et al. 10.1073/pnas.1303730110

SI Materials and Methods

Next-generation DNA sequence data were generated with the Illumina GAII 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. 1*A*). 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 (EcoR1 and Mse1) 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 EcoR1 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 Qiaquick 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 GAII 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 EcoR1 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. 1*A*). 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. 1*A*).

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*) sequences are 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 infile 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 G_{ST} statistics (10) for all combinations of regional groupings. Pairwise G_{ST} 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 assessing 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. 1*A*) 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 F_1 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 (Δ_{ij}) 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 $\Delta_{AB} = (1/n)(2n_{AABB} + n_{AABb} + n_{AaBB} + 1/2n_{AaBb}) - 2p_Ap_B$, where n is the number of individuals in the sample, n denotes the genotype counts for each pair of loci (A and B), and p denotes the allele frequencies at each locus (16, 17). We then averaged the 75 iterations to obtain a mean linkage disequilibria 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. S2*A*). 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. S2*A*.)

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. S2*B*).

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. 1*B*), 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 genotype 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 α 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 admixed. To determine the robustness of these patterns, we estimated the hybrid indices and α for all of the possible parent species population combinations. Although some loci in every combination exhibited excess ancestry for one or the other parent species, the loci were not always the same among the different combinations vs. the putative parent populations (Table S1). The correlations between the estimates of α for each of the 26,313 loci were low among all combinations. These results make our interpretation of α hard because if *P. formosa* were a "frozen" F_1 descending from a single individual, we would expect the estimation of α for each locus to be the same regardless of the parent populations examined, unless significant diversification has occurred among population of one or both parent species. Given that we cannot tear these two possibilities apart with the present dataset, our conclusions about the estimation of α are

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 α. 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 disequilibria, Δ, however, we found that the distribution of Δ 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 Δ calculations (Fig. S3*A*) to the results obtained from a simulated hybrid dataset (Fig. S3*B*), 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 Δ 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 Δ 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*.

1. Gompert Z, et al. (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution* 66(7):2167–2181.
2. Parchman TL, et al. (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Mol Ecol* 21(12):2991–3005.
3. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
4. Gompert Z, Buerkle CA (2011) A hierarchical Bayesian model for next-generation population genomics. *Genetics* 187(3):903–917.
5. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155(2):945–959.
6. Gillespie J (2004) *Populations Genetics: A Concise Guide* (Johns Hopkins Univ Press, Baltimore), 2nd Ed.
7. Hedrick P (2005) *Genetics of Populations* (Jones and Bartlett Publishers, Burlington, MA), 3rd Ed.
8. Galassi M, et al. (2009) *GNU Scientific Library: Reference Manual* (Network Theory, Godalming, UK.).
9. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164(4):1567–1587.
10. Nei M (1973) Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA* 70(12):3321–3323.
11. Gompert Z, Buerkle CA (2011) Bayesian estimation of genomic clines. *Mol Ecol* 20(10): 2111–2127.
12. Turner BJ, Brett BH, Miller RR (1980) Interspecific hybridization and the evolutionary origin of a gynogenetic fish, *Poecilia formosa*. *Evolution* 34(5):917–922.
13. Gompert Z, Buerkle CA (2009) A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Mol Ecol* 18(6):1207–1224.
14. Buerkle CA (2005) Maximum-likelihood estimation of a hybrid index based on molecular markers. *Mol Ecol Notes* 5(3):684–687.
15. Gompert Z, Parchman TL, Buerkle CA (2012) Genomics of isolation in hybrids. *Philos Trans R Soc Lond B Biol Sci* 367(1587):439–450.
16. Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics* 35(1):235–254.
17. Zaykin DV (2004) Bounds and normalization of the composite linkage disequilibrium coefficient. *Genet Epidemiol* 27(3):252–257.

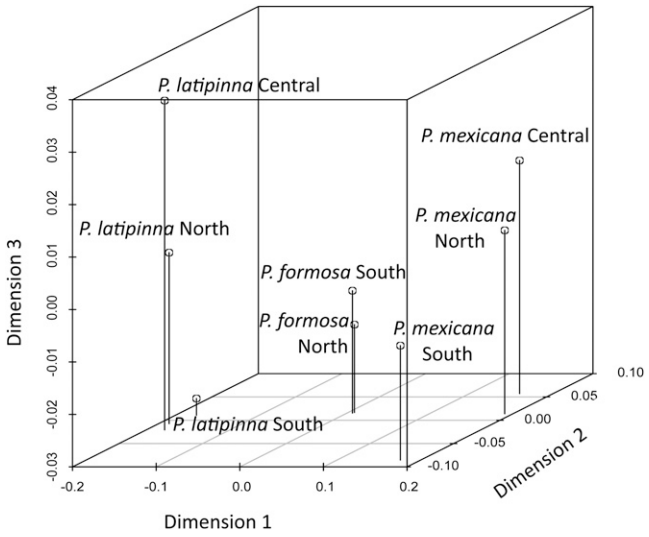

Fig. S1. Nonmetric multidimensional scaling of pairwise G_{ST} between all populations.

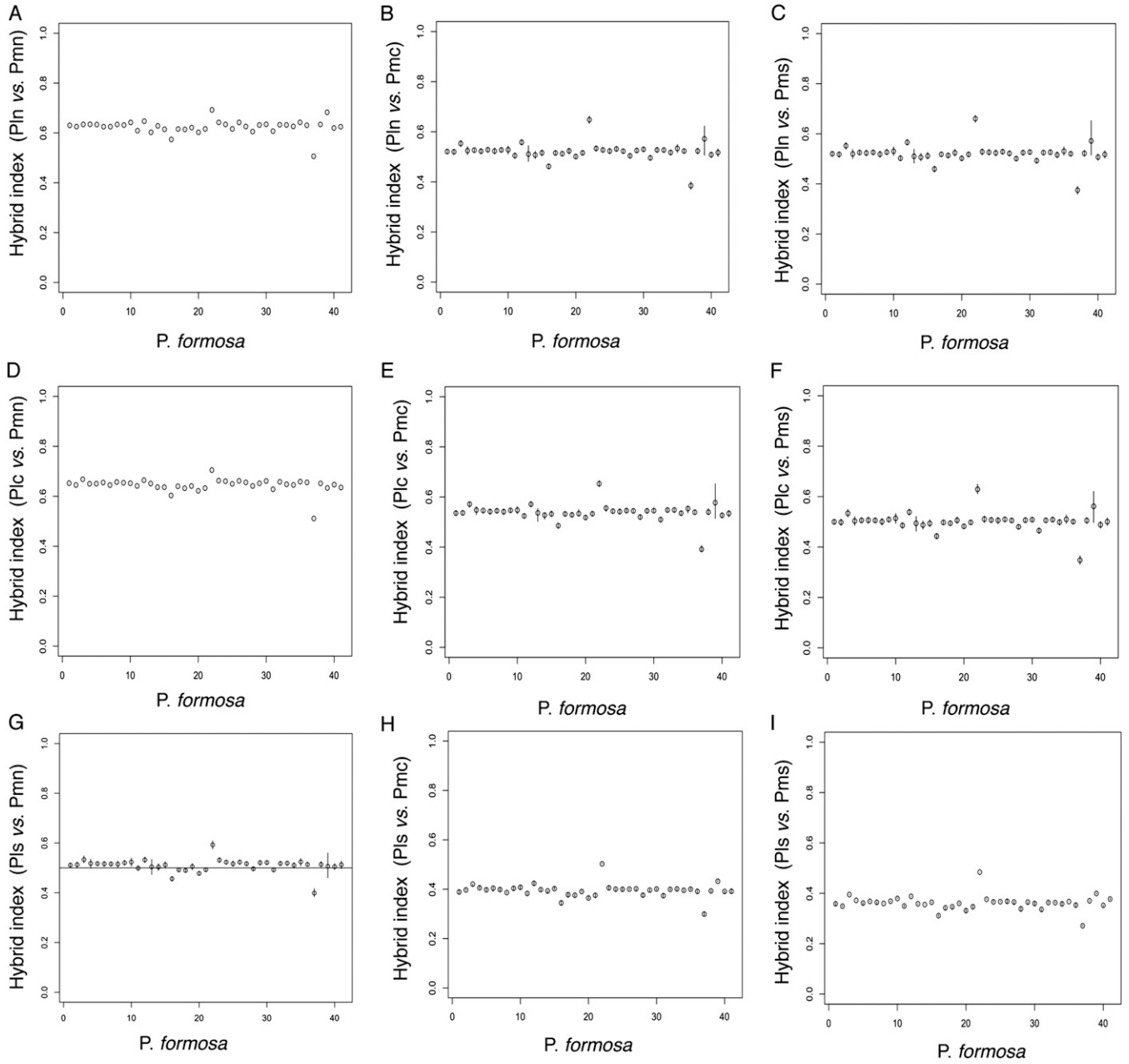

Fig. S2. (*A–I*) Estimates of hybrid index for all *P. formosa* from all combinations of parent species populations. *G* depicts the hybrid index of individual *P. formosa* obtained from the putative parent populations (*P. latipinna* in the south of its range and *P. mexicana* in the north part of its 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).

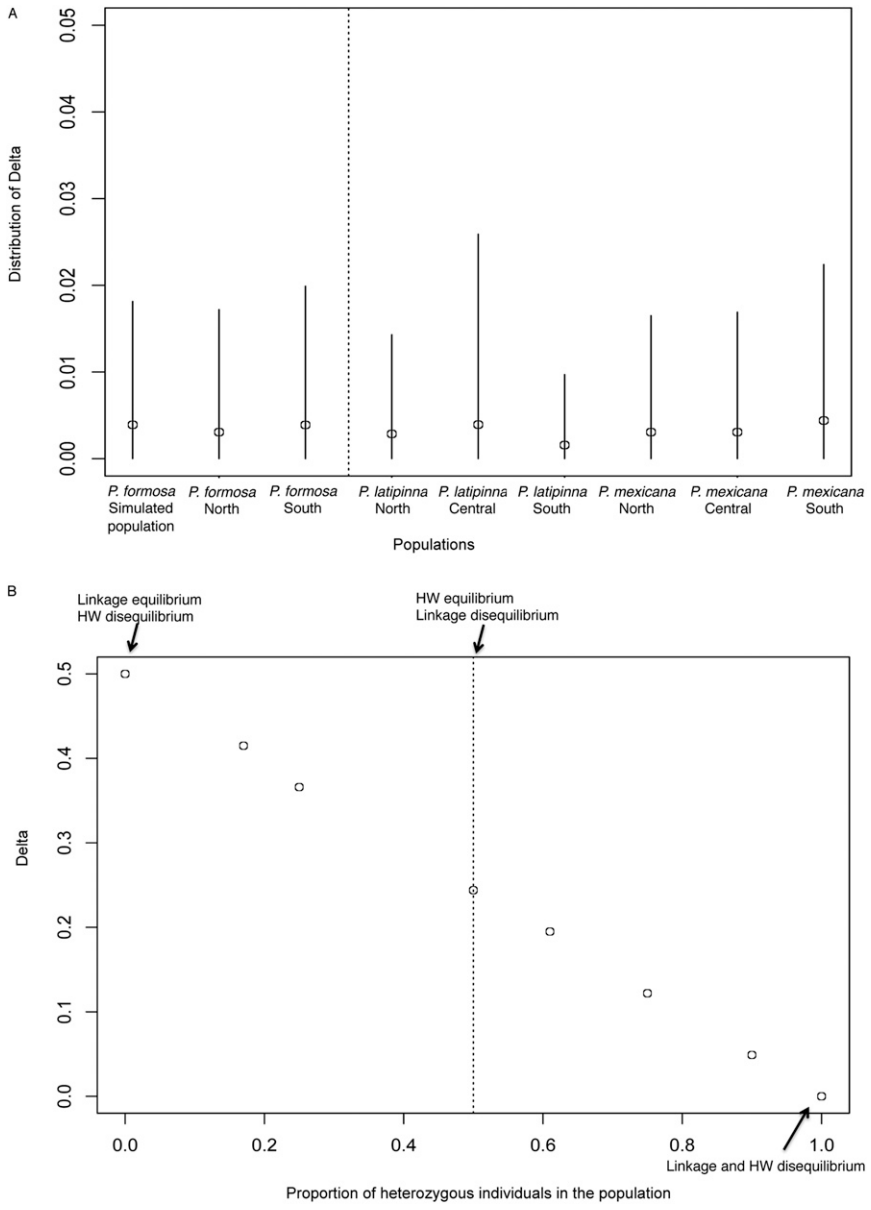

Fig. S3. Mean genome-wide values of Burrow's composite measure of linkage disequilibrium (Δ) for each geographic region (*A*) and values of Δ relative to the effects of linkage and HW disequilibria (*B*). Dotted line in *A* separates asexual (*Left*) from sexual (*Right*) populations. Values for *B* 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.

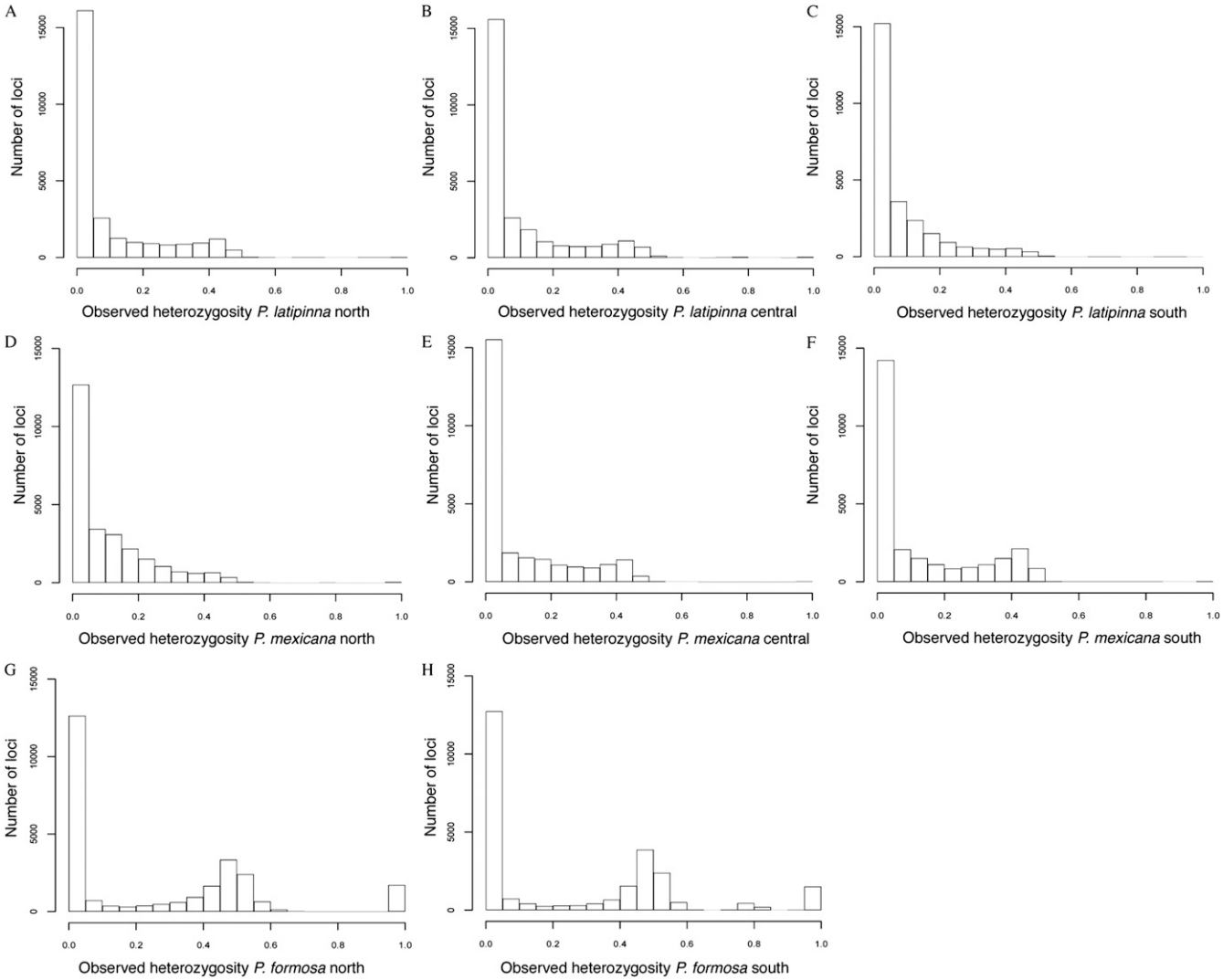

Fig. S4. Distributions of observed heterozygosity across all SNPs for *P. latipinna* (*A*–*C*), *P. mexicana* (*D*–*F*), and *P. formosa* (*G* and *H*).

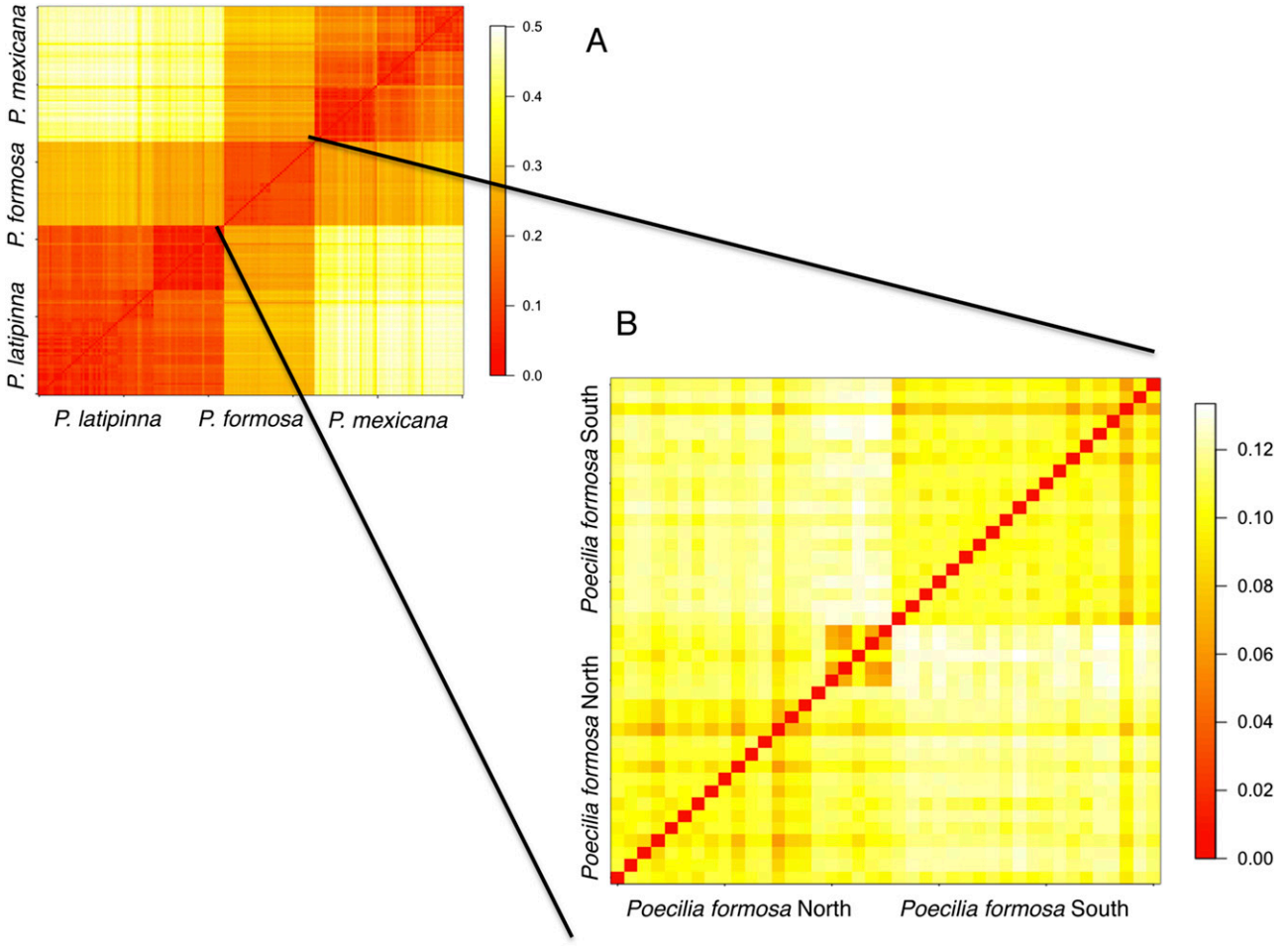

Fig. S5. Mean genetic distance among all individuals (*A*) and among individual *P. formosa* only (*B*). *Poecilia formosa* appears to be genetically intermediate between *P. latipinna* and *P. mexicana* (*A*). A high amount of genotypic variation can be detected within *P. formosa* (*B*). Note that a different scale of values is used in the legends of *A* and *B*.

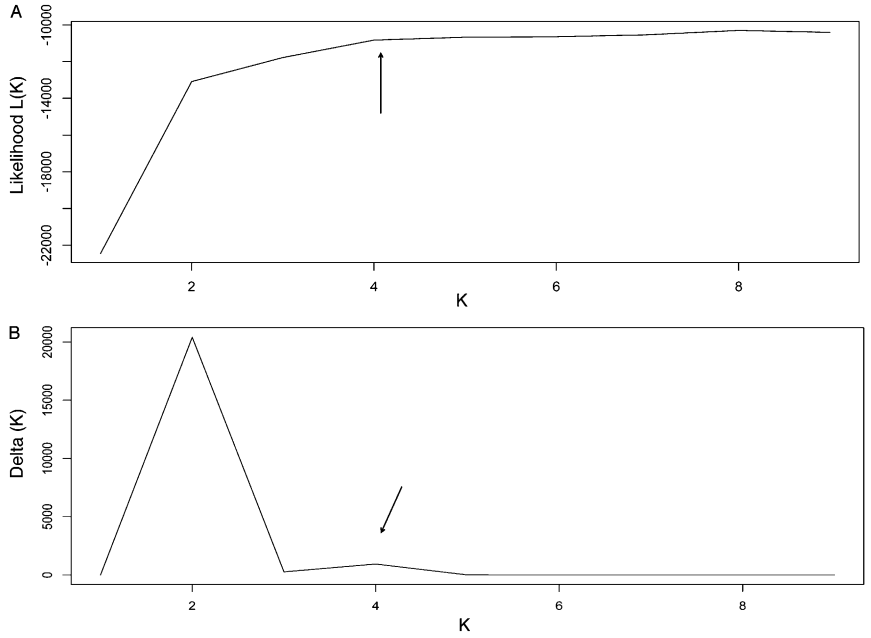

Fig. S6. Summary of the likelihood of K (L(K)) given the number of clusters (*A*) and calculation of delta(K) (14) (*B*). K = 2 and K = 4 appear to be the best to describe the data. See *Materials and Methods* and *Results and Discussion* for information about these results.

Table S1. Correlations between the estimation of the α 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)

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

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 $\alpha < 0$.