doi:10.1111/evo.12009

AN EVALUATION OF THE HYBRID SPECIATION HYPOTHESIS FOR *XIPHOPHORUS CLEMENCIAE* BASED ON WHOLE GENOME SEQUENCES

Molly Schumer,[1,2] Rongfeng Cui,[3,4] Bastien Boussau,[5,6] Ronald Walter,[7] Gil Rosenthal,[3,4] and Peter Andolfatto[1,8]

[1]*Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544*

[2]*E-mail: schumer@princeton.edu*

[3]*Department of Biology, Texas A&M University, College Station, Texas*

[4]*Centro de Investigaciones Científicas de las Huastecas Aguazarca, Calnali, Hidalgo, Mexico*

[5]*Department of Integrative Biology, University of California Berkeley, Berkeley, California*

[6]*Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Lyon, France*

[7]*Department of Chemistry and Biochemistry, Texas State University, San Marcos, Texas 78666*

[8]*Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544*

Received August 28, 2012
Accepted October 24, 2012
Data Archived: Dryad doi:10.5061/dryad.6k7gh

Once thought rare in animal taxa, hybridization has been increasingly recognized as an important and common force in animal evolution. In the past decade, a number of studies have suggested that hybridization has driven speciation in some animal groups. We investigate the signature of hybridization in the genome of a putative hybrid species, *Xiphophorus clemenciae*, through whole genome sequencing of this species and its hypothesized progenitors. Based on analysis of this data, we find that *X. clemenciae* is unlikely to have been derived from admixture between its proposed parental species. However, we find significant evidence for recent gene flow between *Xiphophorus* species. Although we detect genetic exchange in two pairs of species analyzed, the proportion of genomic regions that can be attributed to hybrid origin is small, suggesting that strong behavioral premating isolation prevents frequent hybridization in *Xiphophorus*. The direction of gene flow between species is potentially consistent with a role for sexual selection in mediating hybridization.

KEY WORDS: **Hybridization, incomplete lineage sorting, introgression, next-generation sequencing, sexual selection.**

Gene flow is a potentially important force in adaptation and speciation, and can play a creative role in the evolutionary process (Arnold 2006). Research has shown that hybridization between animal species is remarkably common (Dowling and Secor 1997; Mallet 2005), and that reproductive isolation can be maintained even in the presence of frequent hybridization. Despite predictions that gene flow may prevent speciation, some of the fastest radiating groups frequently hybridize (Salzburger et al. 2002; Mallet 2008), and recent studies have suggested that hybridization can play an important role in the spread of adaptive alleles between species (Castric et al. 2008; Mullen et al. 2008; Whitney et al. 2010; Song et al. 2011; Hovick et al. 2012). Whole genome sequencing projects have resulted in a much more detailed picture of the adaptive significance of hybridization and patterns of introgression (e.g., Heliconius Genome Consortium 2012).

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

In the last decade, a number of studies have proposed cases of speciation mediated by hybridization. Although hybrid speciation via polyploidy is common in plants (Soltis et al. 2004), hybrid speciation without chromosome doubling, or homoploid hybrid speciation, is considered rare. Unlike alloploid hybrids, homoploid hybrids are likely to experience continued gene flow with parental species, preventing speciation. Theoretically, hybrid speciation can occur when recombinant phenotypes make hybrids better adapted to particular environments than parental species, allowing them to colonize a niche unavailable to either parental species. Proposed cases include adaptations to high elevation habitats in the butterfly genera *Papilio* and *Lycaeides* (Gompert et al. 2006; Kunte et al. 2011), utilization of new host plants in host-specific fruitflies (Schwarz et al. 2005), adaptation to extreme habitats in *Helianthus* sunflowers (Rieseberg et al. 1995, 2003), and tolerance of warmer waters in an invasive hybrid group of sculpins (Nolte et al. 2005). Another proposed mechanism of hybrid speciation is strong sexual isolation between parental species and their hybrid offspring. This mechanism is likely to be even more rare because most hybrids show weak reproductive isolation from parental species (Christophe and Baudoin 1998; Velthuis et al. 2005; Ganem et al. 2008) but see (Mavarez et al. 2006; Salazar et al. 2010). Putative cases of hybrid speciation via natural or sexual selection have remained controversial (Brower 2011), in part due to the particular ecological conditions that favor this mode of speciation and in part due to the evidence that has been available to support these claims. Much past genetic research on hybrid speciation has relied on a small number of markers to infer genome-wide patterns of mosaicism (e.g., Nolte et al. 2005; Schwarz et al. 2005; Meyer et al. 2006). Whole-genome sequencing now gives us the tools to investigate potential cases of hybrid speciation and distinguish hybrid speciation from other processes that can produce similar patterns.

Although a potentially important evolutionary phenomenon, distinguishing hybridization and hybrid speciation from other processes can be challenging. Incomplete lineage sorting (ILS) can appear similar to hybridization especially in cases where divergence time is short. In the *Drosophila melanogaster* species group, Pollard et al. (2006) found that only approximately 58% of gene trees supported the species tree, and incomplete lineage sorting resulted in support for two other topologies in the remaining gene trees. Such findings highlight the importance of genome-wide data because sampling few loci can generate misleading results. One promising method that could be used to distinguish between ILS and hybridization is evaluating the size distribution of phylogenetically discordant segments. Recent gene flow will produce large discordant regions (Hanson 1959; Martinsen et al. 2001), whereas the size of segments supporting discordant trees in the case of incomplete lineage sorting is expected to be small (Hobolth et al. 2011; Pruefer et al. 2012). However, in the case of ancient gene flow, recombination will break apart introgressed regions, causing ILS and introgression to appear more similar. In addition, secondary introgression may appear similar to hybrid speciation in certain contexts. Use of larger datasets can help distinguish between introgression, hybrid speciation, and ILS by considering genome-wide patterns.

Another difficulty of studying hybrid speciation stems from disagreement over the nature and extent of genetic contribution required to link hybridization to the speciation process. In general, researchers have identified hybrid species based on genomic mosaicism resulting from recombination between the two parental genomes (Mallet 2005, 2008). This type of hybrid speciation has been documented in *Helianthus* sunflowers; the genome of *H. anomalus* is a combination of the two parental genomes, even preserving patterns of linkage disequilibirum from its progenitors (Rieseberg et al. 1995). Similarly, genome wide divergence patterns in the pathogenic fungus *Zymospetoria pseudotritici* suggest that this species arose from a single hybridization event between two haploid heterospecifics (Stukenbrock et al. 2012). In contrast, other researchers have suggested a mechanism of hybrid speciation called hybrid trait speciation, in which a trait combination acquired by hybridization is important in driving the speciation process (Jiggins et al. 2008; Salazar et al. 2010). Hybrid trait speciation has been proposed for *Heliconius* butterflies in which the spread of the mimicry locus through hybridization may have allowed for rapid diversification (Heliconius Genome Consortium 2012). In this scenario, new traits arising from recombination between parental genomes are key to the speciation process, but in most cases it is difficult to distinguish between introgression and introgression that facilitates speciation, especially if the loci underlying ecologically important traits are unknown. Although distinguishing between introgression and hybrid trait speciation remains challenging, the first step in investigating the role of past hybridization in speciation is determining how much of the genome can be attributed to hybrid origin.

The genus *Xiphophorus* is composed of 26 species of freshwater fish remarkable both for their striking morphological diversity driven by sexual selection and for the broad interfertility between reproductively isolated species in this genus (Kazianis et al. 1996; Kallman and Kazianis 2006). *Xiphophorus* is made up of three monophyletic groups: the southern swordtails, the northern swordtails, and the platyfish (Fig. 1A). Despite sympatric distribution of interfertile species, most species do not naturally hybridize (but see Culumber et al. 2011; Rosenthal and Garcia de Leon 2011), because there are strong behavioral, prezygotic barriers to mating (Clark et al. 1954; McLennan and Ryan 1999). However, these barriers are prone to disruption due to disturbance of communication channels (Fisher et al. 2006), low

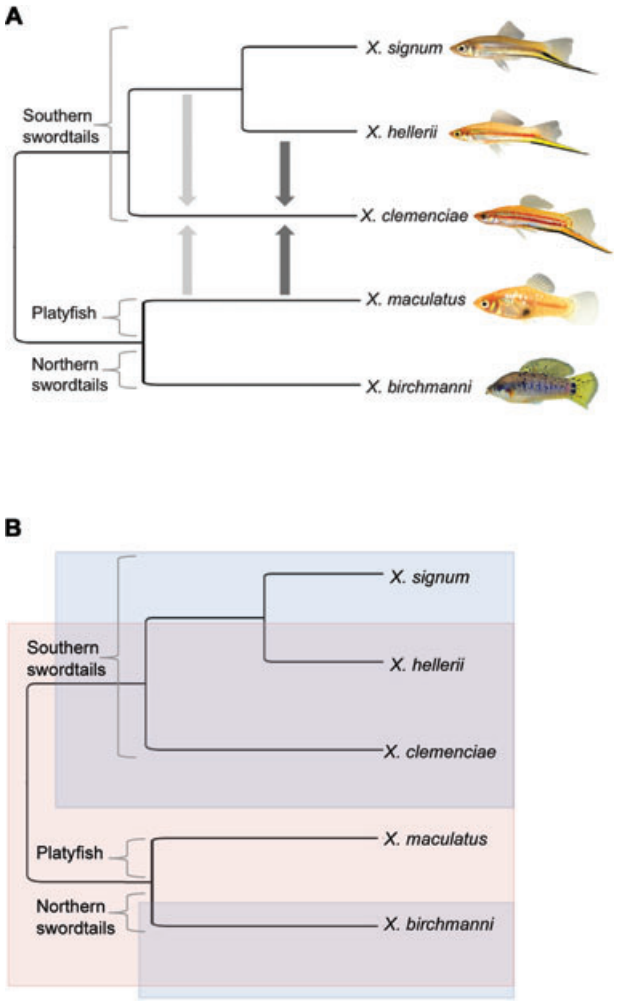

Figure 1. (A) Unrooted nuclear phylogeny of the species used in this study (reproduced from Meyer et al. 2006). Light gray and dark gray arrows show the two hybridization events proposed to have led to *X. clemenciae*; Meyer et al. (2006) proposed that either a hybridization event between extant swordtails and platyfish lead to the formation of *X. clemenciae* (dark gray arrows), or an ancient hybridization event was involved (light gray arrows). Both hypotheses can be addressed with our approach. (B) Experimental design used in the present study. Most analyses focus on species highlighted in red: *X. maculatus*, *X. clemenciae, X. hellerii*, and outgroup *X. birchmanni*. Species highlighted in blue (*X. clemenciae*, *X. hellerii*, *X. signum*, and *X. birchmanni*) were used for an additional set of comparisons to investigate whether gene flow has occurred between *X. clemenciae* and *X. hellerii*.

densities of conspecifics (Willis et al. 2011) or heightened predation risk (Willis et al. 2012). Past hybridization has been proposed based on discordance between mitochondrial and nuclear phylogenetic trees (Meyer et al. 1994, 2006). *X. clemenciae*, a southern swordtail, is closely related to *X. hellerii* based on nuclear sequences, and is also morphologically and behaviorally similar to other swordtails. However, mitochondrial sequence analysis grouped *X. clemenciae* with the swordless platyfish *X. maculatus.* Although most *Xiphophorus* species have strong premating isolation, *X. maculatus* females show preference for the elongated caudal fin typical of swordtails (Basolo 1990, 1995; Meyer et al. 2006). In addition to sequence data, this led to the proposal that *X. clemenciae* is derived from ancient hybridization between *X. maculatus*-like and *X. hellerii*-like ancestors (Fig. 1A, Meyer et al. 2006; Jones et al. 2012). *X. clemenciae* is one of the first vertebrates proposed to be a hybrid species, and one of the few cases in which hybrid speciation was thought to be mediated via sexual selection.

A hybrid origin of *X. clemenciae* is supported by hybrid fertility and by behavioral and morphological data. However, due to the limited number of regions sampled, disagreement between gene trees could also be explained by secondary introgression or incomplete lineage sorting. We examine the evolutionary history of *X. clemenciae* and potential evidence for hybrid ancestry through whole-genome sequencing of *X. clemenciae* and its putative parent species, *X. hellerii* and *X. maculatus*.

A number of methods have been developed to distinguish between incomplete lineage sorting and introgression, but few methods have been developed to distinguish between introgression and hybrid speciation (but see Kubatko, 2009). We therefore take a two-part approach to investigate hybrid ancestry in *X. clemenciae* (Fig. 1B). We use genome-wide phylogenetic analysis to ask whether the genome of *X. clemenciae* is a mosaic of regions more closely related to *X. maculatus* and regions more closely related to *X. hellerii.* If *X. clemenciae* has a mosaic genome, we predict that divergence time and phylogenetic relationships between *X. clemenciae, X. maculatus*, and *X. hellerii* will vary strongly depending on the focal genomic region. Next, we investigate whether secondary introgression has occurred between *X. clemenciae*, *X. hellerii*, and *X. maculatus*, by analyzing the size distribution of regions supporting discordant topologies and using explicit tests for hybridization. We supplement these direct tests for gene flow with simulations of different models of speciation. Finally, to investigate whether gene flow has occurred between *X. hellerii* and *X. clemenciae* since speciation, we include genomic data from another southern swordtail species to clarify phylogenetic relationships between *X. hellerii* and *X. clemenciae* (Fig. 1B). Together, these techniques allow us to distinguish between introgression, ILS, and mosaic genome hybrid speciation. Our results highlight the importance of genome-wide datasets in understanding the role of genetic exchange in speciation.

Methods

GENOME SEQUENCING

One individual of *X. hellerii* (Río Sarabia near Oaxaca) and *X. clemenciae* (Río Grande, Oaxaca) were obtained from the

Xiphophorus Genetic Stock Center (Texas State University, San Marcos, TX). One *X. birchmanni* individual was obtained from a wild population in Río Coacuilco at Coacuilco, Mexico. Genomic DNA was extracted from fin clips using the Agencourt bead–based DNA purification kit (Beckman Coulter Inc., Brea, CA) following manufacturer's protocol with slight modifications. Fin clips were incubated in a 55°C shaking incubator (100 rpm) overnight in 94 μL of lysis buffer with 3.5 μL 40 mg/mL proteinase K and 2.5 DTT, followed by bead binding and purification. Genomic DNA was quantified and evaluated for purity using a Nanodrop 1000 (Thermo Scientific, Wilmington, DE). One microgram was then sheared with a Covaris sonicator (Covaris, Woburn, MA) to approximately 500 bp. The sheared DNA was prepared for sequencing following the protocol outlined in Quail et al. (2009). Briefly, the sheared DNA was end-repaired, and an A-tail was added to facilitate adapter ligation. After adapters were ligated, the product was run on a 2% Agarose gel and fragments between 350 and 500 bp were selected, purified, and PCR amplified for 14 and 16 cycles. Purified samples were analyzed for quality and size distribution on a Bioanalyzer 2100 (Agilent, Santa Clara, CA) and sequenced on an Illumina HiSeq 2000 sequencer at the Lewis-Sigler Institute Sequencing Facility (Princeton University, Princeton, NJ).

Raw 101 bp reads were trimmed to remove low-quality bases (Phred quality score < 20) and reads with fewer than 30 bp of contiguous high-quality bases were removed using the script TQSfastq.py (http://code.google.com/p/ngopt/source/browse/trunk/SSPACE/tools/TQSfastq.py). The number of reads per species and alignment statistics are summarized in Table S1. Trimmed reads were aligned to the *X. maculatus* reference genome (GenBank Assembly ID: GCA_000241075.1, Ensembl annotation: http://pre.ensembl.org/Xiphophorus_maculatus) using STAMPY v1.0.17 (Lunter and Goodson 2011) and the *X. maculatus* mitochondrial genome (GenBank accession no.: AP005982.1) using bwa (Li and Durbin 2009). Mapped reads were analyzed for variant sites using the samtools/bcftools pipeline (Li et al. 2009). Using a custom python script, we used the *X. maculatus* reference genome and mitochondrial genome as a scaffold, and for each species created a new version that incorporated variant sites detected by samtools and masked any sites that had coverage lower than 10 in the mpileup. Because *X. birchmanni* is the only species in which individuals were not lab bred for multiple generations, polymorphism from *X. birchmanni* was used to estimate θ, the population mutation rate, for coalescent simulations. Because the *X. maculatus* reference sequence is currently made up of 20,640 supercontigs, we used the largest 150 scaffolds in our analysis that comprises 56% of the length of the assembled genome. Raw sequences are available through NCBI Sequence Read Archive (Acc no. SRA060275).

TRANSCRIPTOME SEQUENCING OF *X. SIGNUM*

To test for gene flow between *X. hellerii* and *X. clemenciae*, we supplemented genomic data with RNAseq data for *X. signum* (see Fig. 1B). These data were generated from an mRNAseq library derived from the brain tissue of *X. signum* (obtained from the Xiphophorus Genetic Stock Center) prepared with Illumina's TruSeq mRNA Sample Prep Kit (Illumina Inc., San Diego, CA) following manufacturer's instructions. This library was assessed for quality as described above and sequenced with other samples in an Illumina Paired-End lane. Briefly, 18,926,812 reads from *X. signum* were trimmed to QV > 20 and mapped to the *X. maculatus* reference genome using STAMPY. Reads with poor mapping scores (< 20) and regions of high divergence (greater than 6 differences in a sliding window of 21 bp) were excluded due to potential misalignment at splice junctions. Bases with less than $10\times$ coverage or with polymorphism were masked. Alignments were analyzed for variant sites as described above. If alignments were separated by less than 100 bp they were concatenated for analysis. Only alignments of 800 bp or greater (maximum alignment size: 29,857 bp, median size: 2,519 bp) were used in subsequent analysis. This resulted in 3,095 alignments totaling 10,338,221 bp for analysis of phylogenetic relationships between the southern swordtails. Raw reads are available through NCBI Sequence Read Archive (Acc no. SRA060275).

ANALYSIS OF POTENTIAL GENE FLOW

i Detecting genomic mosaicism using the AU test and PhyML_multi

To investigate whether the genome of *X. clemenciae* showed evidence of mosaicism, we examined phylogenetic relationships throughout the genome using both the approximately unbiased test (AU test; Shimodaira 2002) and PhyML_multi (Boussau et al. 2009). Both of these methods can be used to determine phylogenetic relationships over but the AU test is window-based, whereas PhyML_multi uses an HMM-based approach to determine breakpoints between alternate topologies (for proof of concept see Supplementary Materials i and ii). We found that although these two tests result in the same genome-wide pattern, many of the specific regions they identify as supporting discordant topologies are distinct (Tables S2, S3), likely due to different levels of sensitivity (Supplementary Materials i and ii, Figs. S1 to S4), and as a result we used both methods in our analysis to gain a more complete picture of the history of hybridization in *X. clemenciae*. PhyML_multi cannot be used on the analysis including RNAseq data from *X. signum* because the median alignment size is smaller than the range for which PhyML_multi is accurate (Supplementary Materials i).

For the AU test between *X. maculatus*, *X. hellerii*, *X. clemenciae*, and *X. birchmanni* a custom script was used to extract

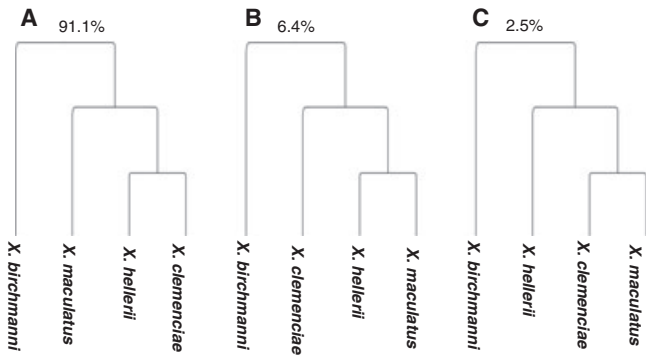

Figure 2. With potential gene flow between *X. hellerii*, *X. clemenciae*, and *X. maculatus*, three unrooted topologies are likely relative to *X. birchmanni*. Topology (A) is the genome-wide consensus tree, (B) suggests incomplete lineage sorting or introgression between *X. hellerii* and *X. maculatus*, and (C) suggests incomplete lineage sorting or introgression between *X. clemenciae* and *X. maculatus*. The percentage of informative 10 kb regions that support each of the three topologies based on AU tests of Scaffolds 0 to 148 is listed above each topology.

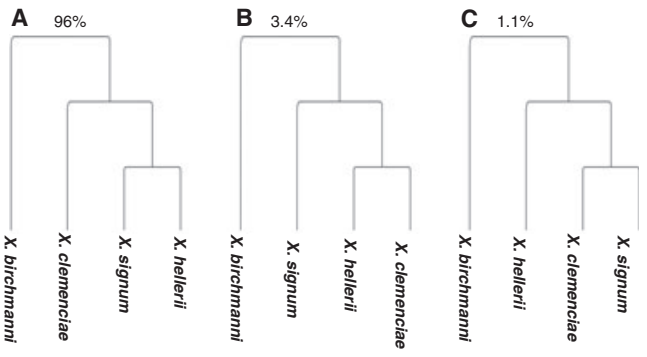

Figure 3. To investigate gene flow between *X. hellerii* and *X. clemenciae*, AU tests were performed including *X. hellerii*, *X. clemenciae*, *X. signum*, and *X. birchmanni*. Three unrooted topologies are likely relative to *X. birchmanni*. Topology (A) is the consensus tree from RNAseq data and agrees with previous nuclear phylogenies (i.e., Meyer et al. 2006), (B) suggests incomplete lineage sorting or introgression between *X. clemenciae* and *X. hellerii*, and (C) suggests incomplete lineage sorting or introgression between *X. clemenciae* and *X. signum*. The percent of informative regions that support each of the three topologies genome-wide is listed above each topology.

alignments of scaffolds 0 to 149 in nonoverlapping 10 kb windows. Windows in which the reference contained Ns, low coverage regions in any species ($< 10\times$), unreliable SNP calls (variant quality score < 20) and sites that were polymorphic or contained indels were excluded. This resulted in 33,564 alignments of 219,729,172 bp (median alignment length 9810 bp). We tested each window's support for all possible unrooted 4-taxa tree topologies. Figure 2 shows the three topologies of interest: A ([*X. hellerii*, *X. clemenciae*], *X. maculatus*, *X. birchmanni*); B ([*X. hellerii*, *X. maculatus*], *X. clemenciae*, *X. birchmanni*), and C ([*X. maculatus*, *X. clemenciae*], *X. hellerii*, *X. birchmanni*]. Topology A is the likely species tree, topology B supports introgression from platyfish into *X. hellerii* and topology C supports introgression from platyfish into *X. clemenciae*. RAxML 7.2.8 (Stamatakis 2006) was used to calculate site-wise likelihoods for each window optimized under the General Time Reversible model with a gamma distribution of mutation rates (GTR + GAMMA). The program Consel 0.20 (Shimodaira and Hasegawa 2001) uses these likelihoods as input for AU tests. If a particular topology had an AU *P*-value greater than 0.95 in a window, we assigned the window that topology. The AU *P*-value is the probability that the alternate topologies have been correctly rejected. To obtain a size distribution of discordant genomic blocks, we repeated this analysis with 5 kb windows and counted the number of contiguous windows that support one topology with AU *P*-values greater than 0.95.

As a secondary approach, to identify genomic regions supporting discordant topologies we used PhyML_multi (Boussau et al. 2009, Supplementary Materials i). Using a custom perl script, we input alignments (quality controlled as above) of the maximum size supported by PhyML_multi (100 kb), specifying as input trees the three topologies of interest (Fig. 2). PhyML_multi evaluates the input trees for each window, calculates site likelihoods for each likely input tree, and uses a Hidden Markov Model to identify breakpoints between topologies (Boussau et al. 2009). For each region, we recorded whether a discordant segment was detected using the Viterbi algorithm (instead of the Forward-Backward algorithm, see Supplementary Materials i), the size of this segment, and which alternative topology this segment supported. We discarded discordant segments less than 5.5 kb in length because quality tests (Figs. S1, S2, Supplementary Materials i) suggested that PhyML_multi does not reliably identify segments of this size.

We repeated the AU test for another group of species (highlighted in blue in Fig. 1B) to test whether *X. clemenciae* had extensive portions of the nuclear genome derived from *X. hellerii*. Figure 3 shows the three topologies of interest in this analysis: A ([*X. hellerii*, *X. signum*], *X. clemenciae*, *X. birchmanni*), B ([*X. hellerii*, *X. clemenciae*], *X. signum*, *X. birchmanni*), and C ([*X. clemenciae*, *X. signum*], *X. hellerii*, *X. birchmanni*). Topology A is supported by previous molecular phylogenies (Meyer et al. 1994, 2006), topology B is consistent with hybridization between *X. hellerii* and *X. clemenciae*, and topology C is consistent with hybridization between *X. clemenciae* and *X. signum*.

ii Detecting introgressive hybridization

Even in the absence of a mosaic genome, introgression from *X. maculatus* may have been important in the evolution of

X. clemenciae. Large regions that support alternative topologies are candidates for recent introgression. We identify large regions (>10 kb in the AU test or >5.5 in PhyML_multi) that support the relationship *X*, *hellerii-X*, *maculatus* and *X. clemenciae-X*. Our simulations suggest that regions of this size are unlikely to be caused by ILS, although a combination of ILS and very low recombination rates could result in discordant regions of this size (Supplementary Materials iii).

As an additional test for introgression, we estimated Patterson's *D*-statistic. The *D*-statistic tests whether shared sites between species are in excess of that expected by ILS alone (Green et al. 2010). We calculated the *D*-statistic for individual scaffolds and for the entire dataset (scaffolds 0–149). We calculated ABBA as the number of shared derived substitutions between *X. hellerii* and *X. maculatus*, and BABA as the number of shared derived substitutions between *X. clemenciae* and *X. maculatus*. To test for statistical significance of the *D*-statistic in this dataset, we used a block jackknife method (Heliconius Genome Consortium 2012; Green et al. 2010) to determine standard error for each scaffold and for the combined dataset (scaffolds 0–149), and performed a two-sided *z*-test. For all analyses of the *D*-statistic we excluded windows in which greater than 90% of the sites were uninformative (*N* in one or more species), and used a coverage cutoff of 20 for SNPs to include only high confidence sites. A block size of 1 Mb was used for the genome-wide *D*-statistic and a block size of 100 kb was used within scaffolds (see Supplementary Materials iv); jackknife bootstrapping was performed using the bootstrap package in R (http://cran.r-project.org/web/packages/bootstrap/index.html, R Development Core Team 2008). A *D*-statistic that diverges significantly from zero can also indicate the presence of population structure (Green et al. 2010).

To test for introgression between *X. hellerii* and *X. clemenciae* using the *D*-statistic, we similarly calculated ABBA as the number of shared derived substitutions between *X. clemenciae* and *X. hellerii*, and BABA as the number of shared derived substitutions between *X. clemenciae* and *X. signum*. To test for statistical significance of the *D*-statistic, we performed a two sample *z*-test. Because the length of alignments and space between the alignments is variable with RNAseq data, we calculated standard error by jackknifing the *D*-statistic for each scaffold (0–149).

iii Examination of potential mitochondrial introgression

Previous research based on two mitochondrial sequences suggested that a platyfish mitochondrial sequence introgressed into *X. clemenciae*. We repeated the analysis of mitochondrial relationships using the whole mitochondrial sequence. Multispecies mitochondrial alignments were produced as above, using a *X. maculatus* mitochondrial genome (GenBank accession no. AP005982.1, unknown strain; Setiamarga et al. 2008) as reference. Protein coding genes were aligned to their correct reading frames using Muscle (Edgar 2004) in MEGA 5 (Kumar et al. 2008). We built a maximum likelihood mitochondrial tree based on the entire mitochondrial sequence, coding sequences only, first and second positions of protein coding sequences only, and four-fold degenerate sites. Due to potential long branch attraction using the entire sequence and four-fold degenerate sites, we included only coding sequences in our final analysis. All analyses were performed in RAxML 7.2.8 using general time reversible (GTR) + GAMMA for each partition with nodal support using 500 rapid bootstraps with the GTR substitution model with the CAT approximation of rate heterogeneity (GTR + CAT).

COALESCENT SIMULATION OF SPECIATION MODELS

X. clemenciae may have originated through speciation with no gene flow, hybrid speciation, or speciation with limited gene flow. To investigate these potential models of speciation, we simulate the neutral coalescent with recombination (Hudson 1990) in a simple allopatric speciation model, speciation of *X. clemenciae* via admixture with different proportions of contributions from each parental genome, and speciation with limited gene flow (Fig. S5). To estimate the parameters for our simulations, we use the average proportion of polymorphic sites in *X. birchmanni* as an estimate of θ. Because no genome-wide mutation rate is available for fish species (but see initial estimates for *Xiphophorus* species: Shen et al. 2012), we used the mutation rate (μ) of 3.8×10^{-8} per base pair per generation derived from *Mus musculus* (Lynch 2010) because it has a similar genome size to *Xiphophorus* (approximately two times larger). We assumed an average genome-wide recombination rate (r) for *Xiphophorus* of 1 cM/3.78 Mb (Walter et al. 2004). We estimated the population recombination rate, $\rho = 4N_e r$, using $\theta/4\mu$ as an estimate of N_e. Based on these values, we found that the effective population size is approximately 10,500 and ρ is approximately 0.0016, suggesting that ρ/θ is approximately 1.

Due to limitations in processing speed, we simulated 100 kb of sequence in 10 kb regions scattered throughout a 10 Mbp region (see Fig. S6) using msHOT (Hellenthal and Stephens 2007). For each model (allopatric, admixture, and limited hybridization), we varied parameters to achieve the best match between the simulated and observed data. Five hundred simulations were completed for each model. In all simulations, time of divergence was a free parameter. Simulations of allopatric speciation with per site θ of 0.0016 matched the mean but not the variance of the observed data (data not shown). Because the ancestral population size may have been larger (and thus result in a larger ancestral θ, θ_A), we repeated our simulations over a range of θ_A (1× to 9× the current θ) and varied splitting time concordantly so that mean divergence of the simulated data matched the real data. For the admixture model we set ancestral population size equal to current population size ($\theta_A = \theta$) and varied extent of admixture and time of

Table 1. Summary of divergence, polymorphism, coverage, and alignment length (covered at 1× or greater) in the three species for which whole genome data was collected.

Species	Average Coverage	Length of Alignment (Percent of Reference)	Percent of Genome Covered ≥ 10×	Divergence From *X. maculatus*	Percent Polymorphic Sites
X. hellerii	24	645,260,495 bp (99%)	70%	1.8%	0.22%
X. clemenciae	21	642,871,178 bp (98%)	64%	1.4%	0.19%
X. birchmanni	45	648,384,058 bp (99%)	79%	1.6%	0.16%

admixture so that mean divergence matched the observed data. As an additional model we included allopatric speciation with limited hybridization. For this model we set the ancestral population size equal to current population size ($\theta_A = \theta$) and varied rate of migration from 0 to 1 ($4m \times N_e$, where m is the proportion of each subpopulation made up of migrants each generation) and splitting time so that mean divergence matched the observed data. Scripts used for coalescent simulations are available through the DRYAD repository: doi:10.5061/dryad.6k7gh.

Results

GENOME SEQUENCING

Average genome wide coverage ranged from 21 to 45× and the total length of the alignment for each species ranged from 98% to 99% of the assembled *X. maculatus* genome (Table 1). Divergence between species, considering only sites with coverage greater than or equal to 20 was as follows: 1.4% *X. hellerii-X. clemenciae*, 1.44% *X. maculatus-X. clemenciae*, 1.8% *X. maculatus-X. hellerii*, 1.7% *X. birchmanni-X. clemenciae*, 1.6% *X. maculatus-X. birchmanni*, and 2.0% *X. hellerii-X. birchmanni* (see Fig. S7 for coverage cutoff validation), and genome wide polymorphism ranged from 0.16% for *X. birchmanni* to 0.22% for *X. hellerii* (Table 1). More details on the sequencing and alignment statistics can be found in Table S1.

ANALYSIS OF POTENTIAL GENE FLOW

i No evidence for genomic mosaicism in *X. clemenciae*

We predicted that if *X. clemenciae* was a hybrid species derived from *X. maculatus* and *X. hellerii*, we would observe larger than expected discordance between gene trees and an excess of topologies supporting a close grouping between *X. clemenciae* and *X. maculatus*. The majority of maximum likelihood topologies supported the relationship shown in Figure 2A: ([*X. hellerii*, *X. clemenciae*], *X. maculatus*, *X. birchmanni*). At an AU *P*-value of > 0.95, only 9% of trees based on 10 kb windows supported alternate topologies (Fig. 2). Of these, 6.4% supported a closer grouping of *X. maculatus* and *X. hellerii*, and 2.5% supported a closer grouping of *X. maculatus* and *X. clemenciae*. The number of discordant regions supporting the grouping of *X. maculatus-X. hellerii* is significantly greater than the number of regions supporting *X. maculatus-X. clemenciae* (95% CI of 1000 nonparametric bootstraps *X. maculatus-X. hellerii*: 6.0%,7%; *X. maculatus-X. clemenciae*: 2.2%, 2.7%). These results are inconsistent with hybrid speciation, but also suggest that ILS alone cannot account for the genome-wide pattern. Analysis of phylogenetic relationships with PhyML_multi similarly found that > 99% of windows grouped *X. hellerii-X. clemenciae*, whereas all remaining windows found support for the *X. hellerii-X. maculatus* topology. Interestingly, PhyML_multi found no regions > 5.5 kb that supported a close relationship between *X. clemenciae-X. maculatus*; this was not unexpected given its limited sensitivity to detect small discordant regions.

To determine whether *X. hellerii* made a large genomic contribution to *X. clemenciae*, we repeated the AU analysis using an additional southern swordtail species, *X. signum* (species highlighted in blue in Fig. 1B), which is more closely related to *X. hellerii* than *X. clemenciae*. A total of 94.4% of the maximum likelihood topologies support the relationship ([*X. hellerii*, *X. signum*], *X. clemenciae*, *X. birchmanni*) at an AU *P*-value of > 0.95 (Fig. 3A). More regions (3.4%) supported the *X. clemenciae-X. hellerii* topology (Fig. 3B) than the *X. clemenciae-X. signum* (Fig. 3C) topology (1.1%), suggesting potential introgression between *X. clemenciae* and *X. hellerii* (95% CI of 1000 nonparametric bootstraps *X. hellerii-X. clemenciae*: 2.8%, 3.6%; *X. signum-X. clemenciae*: 1.1%, 1.6%) but ruling out the hypothesis that much of the *X. clemenciae* genome is derived from backcrossing with *X. hellerii* (Meyer et al. 2006).

Although these results confirm that the genome of *X. clemenciae* is not a mosaic of *X. hellerii* and *X. maculatus*, they do not rule out instances of limited introgression between any of these three species. To investigate this possibility, we considered the size distribution of discordant regions, explicit tests for introgression, and simulations of the coalescent process.

ii Evidence for historical gene flow between *Xiphophorus* species

Discordant regions identified by the AU test and PhyML_multi could be the result of introgression or ILS. To further investigate whether discordant regions are likely to be due

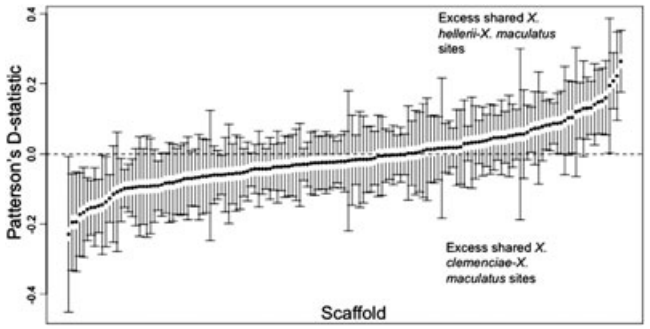

Figure 4. Patterson's D-statistic calculated for each of scaffolds 0 to 149 plotted with 2∗Standard Error derived from block jackknife bootstrapping (see text for details). A positive D-statistic indicates potential introgression between *X. hellerii* and *X. maculatus*; a negative D-statistic indicates potential introgression between *X. clemenciae* and *X. maculatus*. The genome wide D-statistic of −0.01 is not significantly different from zero (dashed line). Scaffolds are not listed in size order but sorted by D-statistic value; scaffolds and corresponding D-statistics are listed in Table S4.

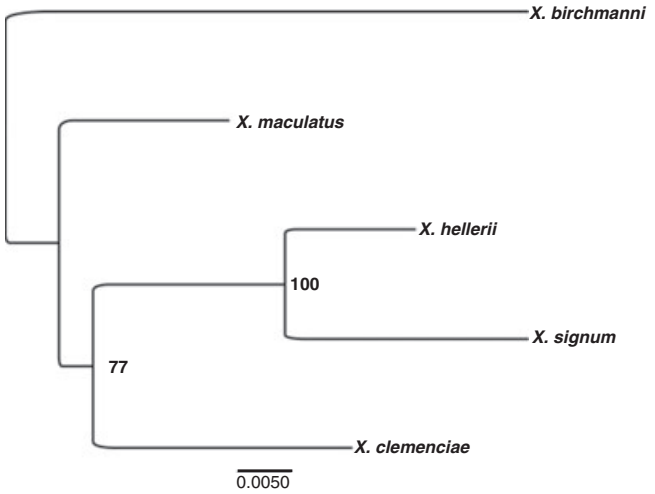

Figure 5. Maximum likelihood mitochondrial phylogeny based on the whole mitochondrial coding sequence for *Xiphophorus* species included in this study with rapid bootstrap values shown at nodes. This phylogeny shows weak resolution of the position of *X. clemenciae* and does not group *X. clemenciae* with platyfish *X. maculatus*.

to ILS or introgression, we determined the size distribution of discordant segments. Based on AU tests in 5 kb windows, 12.5% of regions supporting the (*X. clemenciae*, *X. maculatus*) topology and 22% of regions supporting the (*X. hellerii*, *X. maculatus*) topology are adjacent to a neighboring region supporting the same topology (Fig. S8). This suggests that approximately 80% to 90% of the regions supporting discordant topologies are 5 kb or smaller, consistent with ILS (Fig. S6). These results also show that approximately 20% of regions supporting a close grouping of *X. hellerii-X. maculatus* are larger than 5 kb. Using PhyML_multi, we detected 226 regions supporting a close grouping of *X. hellerii* and *X. maculatus* with a median size of 9.9 kb (Table S3). Many regions supporting recent gene flow between *X. hellerii* and *X. maculatus* identified by the AU test and Phyml_multi (Fig. S9) are larger than expected based on ILS (Fig. S10). This suggests that there may have been significant recent gene flow between *X. hellerii* and *X. maculatus* but not between *X. maculatus* and *X. clemenciae*. Analysis of size distribution is not possible for the dataset including *X. signum* (species highlighted in blue in Fig. 1B) because of the discontinuity of the RNAseq data.

Patterson's *D*-statistic can also be used to reject the null hypothesis of ILS. We first calculated the *D*-statistic for introgression between *X. maculatus*, *X. hellerii*, and *X. clemenciae*. The genome wide *D*-statistic was not significantly different from zero ($D = -0.01$, $P = 0.10$), but we detected 16 scaffolds with a significantly positive *D*-statistic, and 18 scaffolds with a significantly negative *D*-statistic (Fig. 4). The regions with a significantly positive or negative *D*-statistic are candidates for introgression between *X. maculatus* and *X. hellerii*, and *X. maculatus* and *X. clemenciae*, respectively. Although these results differ from our findings from the AU test and PhyML_multi, performance tests applying the *D*-statistic to simulations of allopatric speciation suggest that processes other than gene flow can produce *D*-statistics significantly different from zero (Supplementary Materials iv). In addition, we find that the *D*-statistic is highly consistent with the AU test in regions with strong support for discordant topologies (Supplementary Materials iv).

We also used Patterson's *D*-statistic to investigate introgression between *X. hellerii* and *X. clemenciae*. The *D*-statistic ($D = 0.17$, $P = 9.6e{-}13$), based on RNAseq reads mapped to scaffolds 0 to 149, shows a significant excess of shared sites between *X. hellerii* and *X. clemenciae*, supporting the findings of the AU test. This suggests that gene flow has occurred between *X. hellerii* and *X. clemenciae* since the divergence of *X. signum* and *X. hellerii*.

iii Re-evaluating the evidence for mitochondrial introgression

To test for mitochondrial introgression, we built a maximum likelihood mitochondrial tree based on the coding regions. This phylogeny groups *X. clemenciae* with the southern swordtails but has weak support at this node (Fig. 5). Different partitioning of the data (using 1st and 2nd positions of protein coding regions) did not improve support at this node or group *X. clemenciae* with platyfish as found previously (Meyer et al. 1994, 2006). Although we include more sites, previous analyses were based on more species (Meyer et al., 1994, 2006) and it is unclear which analysis is more informative. Including full mitochondrial sequences of additional platyfish and southern swordtails would likely improve resolution of *X. clemenciae*'s relatedness to other *Xiphophorus* species. However, our simulations of mitochondrial

Table 2. Summary of input parameters and 500 simulations of 100 kb (of an 10 Mb segment) under three speciation models (simulated in msHOT) compared to observed values. Variance is calculated for the number of divergent sites between 10 kb windows. Divergence was calculated by the number of fixed differences between simulated sequence pairs, and ranged from 1.43% to 1.45% between *X. clemenciae* and *X. maculatus*, and 1.39% to 1.42% between *X. clemenciae* and *X. hellerii*.

Speciation Model	Input Parameters	Variance in Divergence (95% Confidence Intervals) *X. clemenciae–X. maculatus*	Variance in Divergence (95% Confidence Intervals) *X. clemenciae–X. hellerii*
Observed values	$T_{\mathrm{div(hel,clem)}} = 3.9$[1]	1008	696.0
	$T_{\mathrm{div(hel,mac)}} = 5$		
	$\theta = 0.0016$		
	$\rho = 0.0016$		
Allopatric speciation	$T_{\mathrm{div(hel,clem)}} = 0.001$	778.9	780.6
	$T_{\mathrm{div(hel,mac)}} = 0.008$	(2.5%: 170.0, 97.5%: 1799)	(2.5%: 150.0, 97.5%: 2084)
	$\theta_A = \rho = 0.014$		
Allopatric speciation with migration	$T_{\mathrm{div(hel,clem)}} = 4.8$	937.6	959.2
	$T_{\mathrm{div(hel,mac)}} = 5$	(2.5%: 214.2, 97.5%: 2753)	(2.5%: 171.7; 97.5%: 2763)
	$4N_e \times m = 0.6$		
	$\theta_A = \rho = 0.0016$		
Admixture (hybrid Speciation)	$T_{\mathrm{div(hel,mac)}} = 4$	67.10	75.05
	$T_{\mathrm{admix(hel,mac)}} = 3.85$	(2.5%: 60.80, 97.5%: 70.42)	(2.5%:65.50, 97.5%: 78.85)
	$\theta_A = \rho = 0.0016$		

[1]Time of divergence calculated as $T_{\mathrm{div}} = 1/2\,[(D_{xy}/\theta) - 1]$.

sequences using *ms* (Hudson 2002) and Seq-gen (Rambaut and Grassly 1997) demonstrate that discordant topologies based on mitochondrial sequences may be common due to incomplete lineage sorting (Supplementary Materials v; Ballard and Rand 2005). In an allopatric model of coalescence of mitochondrial sequences, lineage sorting resulted in mitonuclear discordance in 22% of simulations; with hybridization discordance was found in 61% of simulations. Frequent exchange of mitochondrial haplotypes has been observed in other fish species (e.g., Keck and Near 2010), and mito-nuclear discordance may not be a reliable indicator of a history of hybridization in these groups.

EVALUATION OF DIFFERENT SPECIATION MODELS

We simulated divergence among *X. maculatus*, *X. hellerii*, and *X. clemenciae* using a simple allopatric speciation model, a model that incorporates admixture in the origin of *X. clemenciae*, and a model including limited hybridization between all three species (Fig. S5). We find that a simple allopatric model ($\theta_A = \theta$) does not describe patterns of variation in the observed data. Similarly, the admixture model describes the mean but not the variance observed in the real data, even when different admixture proportions are included in the model. Increasing the ancestral population size ($\theta_A = 8.5 \times \theta$) in the allopatric model results in simulated data which describes the mean divergence and variance observed in the actual data (Table 2), but a larger ancestral population size may not be a realistic assumption. A model including limited gene flow and assuming no change in population size ($\theta_A = \theta$) can also describe both the mean and variance in the observed data (Table 2), and fits well with our genetic data. These basic simulations demonstrate that, although admixture is possible in the evolutionary history of *X. clemenciae*, the observed data can be described adequately with a simple allopatric speciation model or an allopatric speciation model with limited gene flow.

Discussion

Distinguishing hybrid speciation from other processes can be difficult. In species with high levels of hybridization, introgression can closely resemble the mosaic pattern expected in cases of hybrid speciation, especially when a limited number of markers are used. Similarly, if the branch length between speciation events is short, ILS can result in large portions of the genome that do not support the species tree even when genome-wide datasets are used (e.g., 42% of gene trees in Pollard et al. 2006). Phylogenetic discordance due to ILS can persist even when the speciation events that produced this discordance are relatively ancient (Degnan and Rosenberg 2009). In the case of *X. clemenciae*, including genome-wide data gives us a more detailed picture of the role of hybridization in the evolutionary history of this species.

We expected that if *X. clemenciae* originated from hybridization of a platyfish and a swordtail, its phylogenetic relationship to *X. hellerii* and *X. maculatus* would vary strongly depending on

the genomic region being examined. This pattern was not evident in genome-wide data, and we saw very few regions in which *X. clemenciae* was more closely related to *X. maculatus* than to *X. hellerii*. We find that based on the whole mitochondrial protein coding sequence, *X. clemenciae* is not grouped closely with *X. maculatus*, but the position of *X. clemenciae* is not well resolved (Fig. 5). In sum, this evidence suggests that hybridization with a platyfish species was not involved in the origin of *X. clemenciae*.

Cytonuclear signatures, the relationship between mitochondrial and nuclear genotypes, are frequently used to infer the presence and direction of hybridization (Arnold 1993; Scribner and Avise 1994; Avise 2000). Our results demonstrate that cytonuclear signatures can be misleading. In particular, our simulations of mitochondrial coalescence suggested that we can expect to observe a paraphyletic tree 22% of the time even under a simple allopatric model. A meta-analysis of phylogenetic studies incorporating mitochondrial markers found that monophyly is not supported in upward of 20% of species surveyed (Funk and Omland 2003). Although this could reflect ubiquitous hybridization, studies have also suggested that such paraphyly can be achieved by selection or by lineage sorting (Ballard and Rand 2005). Thus, evidence for hybridization based on mitochondrial data alone should be interpreted cautiously.

An alternative hypothesis proposed for the origin of *X. clemenciae* based on behavioral data is initial hybridization between *X. maculatus* and *X. hellerii* followed by repeated backcrossing between *X. clemenciae* and an isolated *X. hellerii* population (Meyer et al. 2006). Such patterns of backcrossing could reduce genomic contribution from *X. maculatus* to the levels we observed. Although this hypothesis is difficult to distinguish from normal allopatric speciation between *X. hellerii* and *X. clemenciae*, it generates the prediction that *X. clemenciae* should be more closely related to *X. hellerii* than to other southern swordtails through much of its genome. We find that levels of discordance are likely too low to be consistent with extensive admixture (*X. hellerii* and *X. clemenciae* are most closely related in ∼3.4% of alignments), but may reflect past hybridization.

Previous research suggested that *X. clemenciae* may be a hybrid of platyfish and southern swordtail lineages based on both nuclear and mitochondrial data (Meyer et al. 1994, 2006). We do not find evidence that *X. clemenciae* has genomic contributions from platyfish, or that *X. clemenciae* has extensively admixed with *X. hellerii*. However, similarly to Meyer et al. (1994, 2006) our results suggest that hybridization has occurred between *X. clemenciae, X. hellerii*, and *X. maculatus*.

Our analysis reveals that there has been gene flow both between *X. hellerii* and *X. maculatus*, and between *X. clemenciae* and *X. hellerii*. Although *X. clemenciae* is grouped with *X. maculatus* in 2.5% of windows, *X. hellerii* is grouped with *X. maculatus* in 6.4% of windows. Similarly, in analysis of the southern swordtails, *X. clemenciae* is grouped with *X. signum* in 1.1% of windows, whereas it is grouped with *X. hellerii* in 3.4% of windows. With no gene flow between species, the support for discordant topologies due to lineage sorting should be equal (but see Slatkin and Pollack 2008). The discordant regions we detect are also much larger than the expected size for discordant regions caused by incomplete lineage sorting (Figs. S8 to S10), although regions of low recombination could also produce discordant segments of comparable size (Supplementary Materials iii). The excess of closely related regions between *X. maculatus-X. hellerii* and *X. hellerii-X. clemenciae* suggests that gene flow has occurred, but in both cases, genomic contributions that can be attributed to hybridization are relatively small. Although we cannot evaluate the expected time of admixture in *X. hellerii-X. clemenciae*, the median discordant segment size of ∼10 kb in regions that group *X. hellerii-X. maculatus* suggests relatively ancient hybridization (as suggested in Meyer et al. 2006), but also that hybridization occurred well after speciation. A recent study using microsatellite markers found little evidence for current gene flow between *X. hellerii*, *X. clemenciae*, and *X. maculatus* throughout their ranges (Jones et al. 2012). The strains of *X. hellerii*, *X. clemenciae*, and *X. maculatus* we sampled are from adjacent river systems, raising the possibility that gene flow between these species is ongoing. Comparing levels of introgression and genomic regions that have introgressed between different populations of these *Xiphophorus* species will allow us to determine whether introgression has occurred repeatedly between species. A population level approach to gene flow between *X. hellerii*, *X. clemenciae*, and *X. maculatus* using techniques such as reduced representation sequencing is an important next step in confirming the patterns of introgression found in this study more broadly.

Sexual selection has received significant attention as a barrier to hybridization, but our results are consistent with a role for female preference in mediating gene flow between species (Basolo et al. 1990; Meyer et al. 2006). Previous research has shown that *X. maculatus* females prefer males with a sword ornament on the caudal fin, even though conspecific males are unornamented (Basolo 1990). Such preferences for heterospecific traits could drive hybridization. Given that *X. hellerii* and *X. maculatus* are sympatric through much of their ranges, hybridization has been suspected but not documented between the two species. Our finding of substantial gene flow between *X. hellerii* and *X. maculatus* suggests that hybridization has occurred between these two species historically. The strength of premating isolation between *X. hellerii* and *X. clemenciae* has not been studied, but the two species are similar in a number of secondary sexual characteristics. Given evidence of past hybridization between *X. hellerii* and *X. clemenciae*, investigating the strength of isolating mechanisms between these species is an interesting future direction. Even with strong isolating behavioral barriers, ecological changes such as

habitat disturbance or skewed sex ratios have been shown to alter preferences in ways that promote hybridization (Fisher et al. 2006; Willis et al. 2011, 2012).

Since the recognition that hybridization is relatively ubiquitous in animal species, one of the major questions has been what types of genomic regions are likely to introgress. A number of studies have shown widespread introgression of regions underlying ecologically important traits such as mimicry (Mullen et al. 2008; Heliconius Genome 2012), toxin resistance (Song et al. 2011), and abiotic tolerance (Whitney et al. 2010). Given the rapid increase in the number of animal taxa in which hybridization has been documented, both investigating the functional identity of introgressed regions, and distinguishing between adaptive and neutral introgression, will be crucial areas of future research.

How important is hybridization in speciation and how can we distinguish hybrid speciation from other processes? Although hybrid speciation has been proposed in a range of taxa, verified cases of genomic mosaicism remain scarce. Advances in sequencing will result in a clearer picture of whether hybrid speciation in animals is as rare as historically thought, or an important force in speciation. What has become clear is that hybridization has great potential to spread adaptive alleles between species. Recently, researchers have proposed that such instances of adaptive introgression can lead to speciation. In this scenario, introgression of a trait allows for colonization of a new niche or immediate sexual isolation from parental species, ultimately resulting in speciation (Jiggins et al. 2008). This hypothesis, although intriguing, has not been widely tested, and requires detailed knowledge of the genetic basis of ecologically important traits and the evolutionary history of the focal species. Because hybrid trait speciation predicts hybrid ancestry at only a few loci, and hybridization is much more widespread than previously thought, many species are likely to resemble this genomic pattern. Thus, claims of hybrid trait speciation should be based on strong evidence that introgression drove speciation. The first step in determining whether adaptive introgression may have led to speciation is the genetic mapping of ecologically relevant traits.

Conclusions

We do not find evidence of a mosaic genome in *X. clemenciae*, or extensive backcrossing of *X. clemenciae* with *X. hellerii*. Despite this, hybridization has been an important part of the evolutionary history of these *Xiphophorus* species. Gene flow from heterospecifics may be responsible for a number of the genomic regions analyzed in *X. hellerii* and *X. clemenciae*. The patterns of gene flow among *X. maculatus*, *X. hellerii*, and *X. clemenciae* are consistent with current species distribution and suggest that female preferences may play a role in hybridization. Female preference likely plays a major role in hybridization in *Xiphophorus* because of weak postzygotic isolation between most species. Investigating the role of female preference in patterns of hybridization and introgression, in *Xiphophorus* and other species groups, is a rich area for future research.

ACKNOWLEDGMENTS

The authors thank the *Xiphophorus* Genetic Stock Center for providing samples for sequencing. Funding for this project was provided in part by the Society for the Study of Evolution's Rosemary Grant Award to M. Schumer and R. Cui. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE0646086 and was partially supported by National Institutes of Health, National Center for Research Resources, Division of Comparative Medicine grant R24OD011120 (R. Walter) including an ARRA supplement to this award. The authors are grateful to Heidi Fisher, two anonymous reviewers, and members of the Rosenthal and Andolfatto labs for providing helpful comments on earlier versions of this article. The authors also thank staff of the Lewis Sigler Center for Integrative Genomics and the Texas A&M University Brazos HPC cluster.

LITERATURE CITED

Arnold, J. 1993. Cytonuclear disequilibria in hybrid zones. Annu. Rev. Ecol. Syst. 24:521–554.

Arnold, M. L. 2006. Evolution through genetic exchange. Oxford Univ. Press, Oxford, U.K.

Avise, J. C. 2000. Cytonuclear genetic signatures of hybridization phenomena: rationale, utility, and empirical examples from fishes and other aquatic animals. Rev. Fish Biol. Fisher. 10:253–263.

Ballard, J. W. O., and D. M. Rand. 2005. The population biology of mitochondrial DNA and its phylogenetic implications. Annu. Rev. Ecol. Syst. 36:621–642.

Basolo, A. L. 1990. Female preference predates the evolution of the sword in swordtail fish. Science 250:808–810.

———. 1995. A further examination of preexisting bias favoring a sword in the genus Xiphophorus. Anim. Behav. 50:365–375.

Boussau, B., L. Gueguen, and M. Gouy. 2009. A mixture model and a hidden markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. Evol. Bioinform. 5:67–79.

Brower, A. V. Z. 2011. Hybrid speciation in Heliconius butterflies? A review and critique of the evidence. Genetica 139:589–609.

Castric, V., J. Bechsgaard, M. H. Schierup, and X. Vekemans. 2008. Repeated adaptive introgression at a gene under multiallelic balancing selection. Plos Genetics 4:1–9.

Christophe, N., and C. Baudoin. 1998. Olfactory preferences in two strains of wild mice, Mus musculus musculus and Mus musculus domesticus, and their hybrids. Anim. Behav. 56:365–369.

Clark, E., L. R. Aronson, and M. Gordon. 1954. Mating behavior patterns in two sympatric species of Xiphophorin fishes; their inheritance and significance in sexual isolation. Bull. Amer. Mus. Nat. Hist. 103:135–226.

Culumber, Z. W., H. S. Fisher, M. Tobler, M. Mateos, P. H. Barber, M. D. Sorenson, and G. G. Rosenthal. 2011. Replicated hybrid zones of Xiphophorus swordtails along an elevational gradient. Mol. Ecol. 20:342–356.

Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Dowling, T. E., and C. L. Secor. 1997. The role of hybridization and introgression in the diversification of animals. Annu. Rev. Ecol. Syst. 28:593–619.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform. 5:1–19.

Fisher, H. S., B. B. M. Wong, and G. G. Rosenthal. 2006. Alteration of the chemical environment disrupts communication in a freshwater fish. P. Roy. Soc. B. 273:1187–1193.

Funk, D. J., and K. E. Omland. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397–423.

Ganem, G., C. Litel, and T. Lenormand. 2008. Variation in mate preference across a house mouse hybrid zone. Heredity 100:594–601.

Gompert, Z., J. A. Fordyce, M. L. Forister, A. M. Shapiro, and C. C. Nice. 2006. Homoploid hybrid speciation in an extreme habitat. Science 314:1923–1925.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, *et al*. 2010. A draft sequence of the neandertal genome. Science 328:710–722.

Hanson, W. D. 1959. The breakup of initial linkage blocks under selected mating systems. Genetics 44:857–868.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94–98.

Hellenthal, G., and M. Stephens. 2007. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics 23:520–521.

Hobolth, A., J. Y. Dutheil, J. Hawks, M. H. Schierup, and T. Mailund. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21:349–356.

Hovick, S. M., L. G. Campbell, A. A. Snow, and K. D. Whitney. 2012. Hybridization alters early life-history traits and increases plant colonization success in a novel region. Amer. Nat. 179:192–203.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7:1–44.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Jiggins, C. D., C. Salazar, M. Linares, and J. Mavarez. 2008. Hybrid trait speciation and Heliconius butterflies. Philos. T. Roy. Soc. B. 363:3047–3054.

Jones, J. C., J.-A. Perez-Sato, and A. Meyer. 2012. A phylogeographic investigation of the hybrid origin of a species of swordtail fish from Mexico. Mol. Ecol. 21:2692–2712.

Kallman, K. D., and S. Kazianis. 2006. The genus xiphophorus in Mexico and Central America. Zebrafish 3:271–285.

Kazianis, S., D. C. Morizot, B. B. McEntire, R. S. Nairn, and R. L. Borowsky. 1996. Genetic mapping in Xiphophorus hybrid fish: assignment of 43 AP-PCR/RAPD and isozyme markers to multipoint linkage groups. Genome Res. 6:280–289.

Keck, B. P., and T. J. Near. 2010. Geographic and temoporal aspects of mitochondrial replacement in nothonotus darters (Teleostei: Percidae: Etheostomatinae). Evolution 64:1410–1428.

Kumar, S., M. Nei, J. Dudley, and K. Tamura. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. 9:299–306.

Kunte, K., C. Shea, M. L. Aardema, J. M. Scriber, T. E. Juenger, L. E. Gilbert, and M. R. Kronforst. 2011. Sex chromosome Mosaicism and Hybrid Speciation among Tiger Swallowtail Butterflies. PLoS Genetics 7: 1–14.

Kubatko, L. S. 2009. Hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478–488.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and P. Genome Project Data. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

Lunter, G., and M. Goodson. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21:1–4.

Lynch, M. 2010. Evolution of the mutation rate. Trends Genet. 26:345–352.

Mallet, J. 2005. Hybridization as an invasion of the genorne. Trends. Ecol. Evolut. 20:229–237.

———. 2008. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philos. T. Roy. Soc. B. 363:2971–2986.

Martinsen, G. D., T. G. Whitham, R. J. Turek, and P. Keim. 2001. Hybrid populations selectively filter gene introgression between species. Evolution 55:1325–1335.

Mavarez, J., C. A. Salazar, E. Bermingham, C. Salcedo, C. D. Jiggins, and M. Linares. 2006. Speciation by hybridization in Heliconius butterflies. Nature 441:868–871.

McLennan, D. A., and M. J. Ryan. 1999. Interspecific recognition and discrimination based upon olfactory cues in northern swordtails. Evolution 53:880–888.

Meyer, A., J. M. Morrissey, and M. Schartl. 1994. Recurrent origin of a sexually selected trait in Xiphophorus fishes inferred from a molecular phylogeny. Nature 368:539–542.

Meyer, A., W. Salzburger, and M. Schartl. 2006. Hybrid origin of a swordtail species (Teleostei : Xiphophorus clemenciae) driven by sexual selection. Mol. Ecol. 15:721–730.

Mullen, S. P., E. B. Dopman, and R. G. Harrison. 2008. Hybrid zone origins, species boundaries, and the evolution of wing-pattern diversity in a polytypic species complex of North American admiral butterflies (Nymphalidae : Limenitis). Evolution 62:1400–1417.

Nolte, A. W., J. Freyhof, K. C. Stemshorn, and D. Tautz. 2005. An invasive lineage of sculpins, Cottus sp (Pisces, Teleostei) in the Rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. Philos. T. Roy. Soc. B. 272:2379–2387.

Nolte, A. W. and D. Tautz. 2010. Understanding the onset of hybrid speciation. Trends Genet. 26:54–58.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2:1634–1647.

Pruefer, K., K. Munch, I. Hellmann, K. Akagi, J. R. Miller, B. Walenz, S. Koren, G. Sutton, C. Kodira, R. Winer, *et al*. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486:527–531.

Quail, M. A., H. Swerdlow, and D. Turner. 2009. Improved protocols for the Illumina Genome Analyzer Sequencing Systerm. Curr. Protoc. Hum. Genet. 18:18.2.1–18.2.27.

R Development Core Team. 2008. R: A language and environment for statistical computing. Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic frees. Comput. Appl. Biosci. 13:235–238.

Rieseberg, L. H., O. Raymond, D. M. Rosenthal, Z. Lai, K. Livingstone, T. Nakazato, J. L. Durphy, A. E. Schwarzbach, L. A. Donovan, and C. Lexer. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. Science 301:1211–1216.

Rieseberg, L. H., C. Vanfossen, and A. M. Desrochers. 1995. Hybrid speciation accompanied by genomic reorganization in wild sunflowers. Nature 375:313–316.

Rosenthal, G. G., and F. J. Garcia de Leon. 2011. Speciation and hybridization. Pp. 109–119 *in* I. Schlupp, A. Pilastro, J. Evans, eds., Ecology and evolution of poeciliid fishes. Univ. of Chicago Press, Chicago.

Salazar, C., S. W. Baxter, C. Pardo-Diaz, G. Wu, A. Surridge, M. Linares, E. Bermingham, and C. D. Jiggins. 2010. Genetic evidence for hybrid trait speciation in heliconius butterflies. PLoS Genet. 6:1–12.

Salzburger, W., S. Baric, and C. Sturmbauer. 2002. Speciation via introgressive hybridization in East African cichlids? Mol. Ecol. 11:619–625.

Schwarz, D., B. M. Matta, N. L. Shakir-Botteri, and B. A. McPheron. 2005. Host shift to an invasive plant triggers rapid animal hybrid speciation. Nature 436:546–549.

Scribner, K. T., and J. C. Avise. 1994. Cytonuclear genetics of experimental fish hybrid zones inside Biosphere-2. Proc. Natl. Acad. Sci. U.S.A. 91:5066–5069.

Setiamarga, D. H. E., M. Miya, Y. Yamanoue, K. Mabuchi, T. P. Satoh, J. G. Inoue, and M. Nishida. 2008. Interrelationships of Atherinomorpha (medakas, flyingfishes, killifishes, silversides, and their relatives): the first evidence based on whole mitogenome sequences. Mol. Phylogenet. Evol. 49:598–605.

Shen, Y., J. Catchen, T. Garcia, A. Amores, I. Beldorth, J. Wagner, Z. Zhang, J. Postlethwait, W. Warren, M. Schartl *et al*. 2012. Identification of transcriptome SNPs between Xiphophorus lines and species for assessing allele specific gene expression within F1 interspecies hybrids. Comp. Biochem. Physiol. C. Toxicol. Pharmacol. 155:102–108.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Slatkin, M., and J. L. Pollack. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. Mol. Biol. Evol. 25:2241–2246.

Soltis, D. E., P. S. Soltis, and J. A. Tate. 2004. Advances in the study of polyploidy since Plant speciation. New Phytol. 161:173–191.

Song, Y., S. Endepols, N. Klemann, D. Richter, F.-R. Matuschka, C.-H. Shih, M. W. Nachman, and M. H. Kohn. 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. Curr. Biol. 21:1296–1301.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stukenbrock, E. H., F. B. Christiansen, T. T. Hansen, J. Y. Dutheil, and M. H. Schierup. 2012. Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. Proc. Natl. Acad. Sci. U.S.A. 109:10954–10959.

Velthuis, B. J., W. C. Yang, T. van Opijnen, and J. H. Werren. 2005. Genetics of female mate discrimination of heterospecific males in Nasonia (Hymenoptera, Pteromalidae). Anim. Behav. 69:1107–1120.

Walter, R. B., J. D. Rains, J. E. Russell, T. M. Guerra, C. Daniels, D. A. Johnston, J. Kumar, A. Wheeler, K. Kelnar, V. A. Khanolkar, *et al*. 2004. A microsatellite genetic linkage map for xiphophorus. Genetics 168:363–372.

Whitney, K. D., R. A. Randell, and L. H. Rieseberg. 2010. Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. New Phytol. 187:230–239.

Willis, P. M., G. G. Rosenthal, and M. J. Ryan. 2012. An indirect cue of predation risk counteracts female preference for conspecifics in a naturally hybridizing fish *Xiphophorus birchmanni*. PloS one 7: e34802.

Willis, P. M., M. J. Ryan, and G. G. Rosenthal. 2011. Encounter rates with conspecific males influence female mate choice in a naturally hybridizing fish. Behav. Ecol. 22:1234–1240.

Associate Editor: Artyom Kopp

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Table S1. Information on alignments of each species to the *X. maculatus* reference genome.

Table S2. Large discordant regions identified by the AU test.

Table S3. Large discordant regions identified by PhyML_multi.

Table S4. D-statistic and jackknife standard error for each scaffold (0–149).

Table S5. Values of Patterson's D-statistic for regions supporting a discordant topology (as identified by an AU *P*-value > 0.9 for an alternate topology) suggest that the D-statistic and AU test give consistent results.

Figure S1. Performance of PhyML_multi on test sequences with known breakpoints between topologies with discordant segments ranging from 2 kb to 10 kb in size.

Figure S2. (A) Size distribution of erroneous discordant regions detected by PhyML_multi in 900 simulated sequences, excluding breakpoints detected in the last 10 kb of the simulated alignment. This suggests that most erroneous discordant regions detected by PhyML_multi are <5 kb. (B) Number of incorrectly identified breakpoints in 100 simulations of each size class of discordant segments. Sequences in which few discordant segments are detected also have few instances of detection of erroneous breakpoints.

Figure S3. AU *P*-value for support of the discordant topology in 1000 simulations of (A) windows containing support only for the concordant topology, (B) windows contain support only for the discordant topology, and (C) windows containing some support for both topologies.

Figure S4. AU *P*-value support for the discordant topology in 1000 simulations with varying numbers of contiguous base pairs supporting that topology in 5 kb (A) and 10 kb (B) windows.

Figure S5. Three speciation models simulated with msHOT: (A) allopatric speciation, (B) speciation via admixture, and (C) allopatric speciation with limited gene flow.

Figure S6. Schematic of simulation strategy with msHOT.

Figure S7. Percent divergence (from *X. maculatus*) and polymorphism in *X. birchmanni* at different coverage thresholds suggests that polymorphism and divergence estimates are not coverage dependent.

Figure S8. Based on AU tests, the three topologies (*X. clemenciae, X. hellerii*), (*X. maculatus*, *X. hellerii*), and (*X. maculatus*, *X. clemenciae*) are all supported in some regions in Scaffolds 0-149, but the regions that support them vary in size.

Figure S9. Size distribution of regions supporting a close grouping of *X. hellerii* and *X. maculatus* (sampled from Scaffolds 0-149) based on analysis with PhyML_multi.

Figure S10. Expected size of discordant regions due to ILS based on 100 simulations of 100 kb regions using *ms*.

Figure S11. D-statistic and 97.5% confidence intervals (calculated from 1000 nonparametric bootstraps of the simulated ABBA and BABA sites) of 1 Mb sequences generated through 100 simulations of allopatric speciation.