MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 2661–2679 doi: 10.1111/mec.13602

GENOMICS OF HYBRIDIZATION

Ancient hybridization and genomic stabilization in a swordtail fish

MOLLY SCHUMER,*† RONGFENG CUI,†‡§ DANIEL L. POWELL,†‡ GIL G. ROSENTHAL†‡ and PETER ANDOLFATTO*¶

*Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA, †Centro de Investigaciones Científicas de las Huastecas "Aguazarca", 16 de Septiembre 392, Calnali Hidalgo 43230, Mexico, ‡Department of Biology, Texas A&M University, TAMU, College Station, TX 77843, USA, §Max Planck Institute for the Biology of Aging, D-50931, Cologne, Germany, ¶Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

Abstract

A rapidly increasing body of work is revealing that the genomes of distinct species often exhibit hybrid ancestry, presumably due to postspeciation hybridization between closely related species. Despite the growing number of documented cases, we still know relatively little about how genomes evolve and stabilize following hybridization, and to what extent hybridization is functionally relevant. Here, we examine the case of *Xiphophorus nezahualcoyotl*, a teleost fish whose genome exhibits significant hybrid ancestry. We show that hybridization was relatively ancient and is unlikely to be ongoing. Strikingly, the genome of *X. nezahualcoyotl* has largely stabilized following hybridization, distinguishing it from examples such as human–Neanderthal hybridization. Hybridization-derived regions are remarkably distinct from other regions of the genome, tending to be enriched in genomic regions with reduced constraint. These results suggest that selection has played a role in removing hybrid ancestry from certain functionally important regions. Combined with findings in other systems, our results raise many questions about the process of genomic stabilization and the role of selection in shaping patterns of hybrid ancestry in the genome.

Keywords: divergence, genomic stabilization, hybridization, whole-genome sequencing

Received 1 October 2015; revision accepted 6 January 2016

Introduction

Hybridization between different species is now known to be a common evolutionary process. As a result, the genomes of many species include regions derived from hybridization. Although hybridization and hybrid ancestry in the genome are now recognized as common, we are just beginning to develop a conceptual framework for understanding how genomes evolve following hybridization and what this implies about the role of hybridization in adaptation and speciation. Many core questions about the dynamics of genome evolution after hybridization remain unanswered. Is the fixation and loss of hybridization-derived regions in the genome primarily governed by neutral processes? What role do hybrid incompatibilities play in purging hybridization-derived regions from the genome? How many hybridization-derived regions were driven to fixation by positive selection (i.e. adaptive introgression)? Beginning to address these questions is the first step in understanding the functional consequences of hybridization and its importance in speciation.

Although hybrid ancestry has been documented in the genomes of many species, less is known about the variation within and between populations of the same species in hybrid ancestry. Different populations may experience distinct histories of hybridization, resulting in different admixture proportions (Cahill *et al.* 2015). Differences between populations in effective population sizes can also contribute to variation in admixture proportions if selection is less effective at purging hybridization-derived regions from the genome in smaller populations (Sankararaman *et al.* 2014).

Correspondence: Molly Schumer, Fax: (609) 258 1712;
E-mail: schumer@princeton.edu

Unless there is ongoing gene flow between species, we expect that over time hybrid ancestry in the genome will eventually stabilize, either due to genetic drift (the strength of which depends on population sizes) or selection. Although there has been little empirical work on the process of genomic stabilization, existing results are mixed. For example, few Neanderthal-derived regions in the human genome have fixed in the 2000 generations since hybridization occurred (Sankararaman *et al.* 2012, 2014). By contrast, in a hybrid lineage of *Zymoseptoria* fungi, ~30% of the genome has fixed for one or the other parental haplotype in the last ~400 generations (Stukenbrock *et al.* 2012). Similarly, in hybrid sunflower species, genomic stabilization is believed to have occurred rapidly, in fewer than 1000 generations (Buerkle & Rieseberg 2008). Demographic differences likely have major effects on these patterns (e.g. rapid expansion of human populations and the haploid life cycle in fungi), but rapid rates of stabilization in some species suggest either an important role for selection in shaping the hybrid genome or strong genetic drift (Buerkle & Rieseberg 2008). More research is needed to understand whether there are general rules for how quickly genomes stabilize after hybridization and which processes most commonly drive this stabilization.

To understand what role selection may play in determining hybrid ancestry in the genome, we can also ask where hybrid ancestry occurs in the genome. The human–Neanderthal hybridization event has been one of the best-studied cases of genome evolution postadmixture. Rapidly evolving regions have lower-than-expected Neanderthal ancestry in modern humans (Vernot & Akey 2014), as do regions of the genome subject to purifying selection in humans (Sankararaman *et al.* 2014) but see also (Juric *et al.* 2015). These results suggest that selection has acted to shape patterns of Neanderthal-derived ancestry among human genomes.

Evidence from other species hints at similar dynamics between hybridization and selection. In the *Helianthus* sunflower hybrid species complex, experimentally synthesized hybrid lineages recapitulate patterns of ancestry observed in hybrid species, suggesting that selection had determined the architecture of ancestry in the hybrid species (Rieseberg *et al.* 1996). Similarly, experimentally synthesized autopolyploid (hybridization-derived polyploid) plants undergo comparable genomic changes to those observed in ancient autopolyploids (Song *et al.* 1995; Soltis & Soltis 1999; Tayale & Parisod 2013). In contrast, in a recent study of hybrid agricultural corn, independently selected lineages did not show consistent patterns of genetic changes (Gerke *et al.* 2015).

Studies of differential introgression in hybrid zones have also informed our understanding of the interactions between gene flow and selection. For example, male hybrids between two house mouse subspecies, *Mus musculus musculus* and *M. m. domesticus*, exhibit a range of reduced fertility phenotypes due to epistatic genetic incompatibilities (Good *et al.* 2010; White *et al.* 2011; Turner & Harr 2014; Turner *et al.* 2014). These regions are particularly resistant to introgression across hybrid zones between the two subspecies, suggesting that selection is effectively filtering gene flow throughout the genome (Payseur *et al.* 2004; Payseur & Nachman 2005). Similarly, in the hybridizing swordtail fish *Xiphophorus birchmanni* and *X. malinche*, regions of the genome implicated in hybrid incompatibilities are more divergent between species (but not between other closely related species), suggesting that these regions have resisted homogenization from historical gene flow (Schumer *et al.* 2014). A large number of studies in a range of species have found patterns of differential introgression across the genome (e.g. Martinsen *et al.* 2001; Geraldes *et al.* 2006; Carling & Brumfield 2008; Teeter *et al.* 2008; Hamilton *et al.* 2013; Larson *et al.* 2013), consistent with selection filtering gene flow, although in most cases a direct link between selection and differential introgression has not been established.

Examples of adaptive introgression, such as the evolution of warning wing colour patterns in butterflies, poison resistance in mice and drought tolerance in sunflowers (Whitney *et al.* 2010; Song *et al.* 2011; Heliconius Genome 2012), underscore the fact that hybridization-derived regions are not always fixed by passive processes. However, it is notable that all of these cases involve hybrid ancestry in just a few genomic regions and how likely adaptive introgression is on a genome-wide scale, especially in cases where a large proportion of the genome has been derived from hybridization, is an open question. Research on the functional consequences of hybridization have hinted at an excess of hybrid ancestry in some gene families and functional categories, implying that hybridization can be a mechanism for broadly co-opting functional elements from another genome. For example, modern humans harbour Neanderthal ancestry at genes involved in skin and hair phenotypes (Sankararaman *et al.* 2014; Vernot & Akey 2014), suggesting that humans migrating out of Africa acquired phenotypes may have adapted to a non-African environment via hybridization rather than *de novo* mutations.

To address these many questions about the processes shaping hybrid ancestry in the genome, we need to investigate patterns of hybrid ancestry and genomic stabilization in a broader range of species. Swordtail fish (*Xiphophorus*) have been a model system for the evolutionary genetics of hybridization for the better part of a century (Gordon 1937; Atz 1962), with a renewed surge

of interest since the advent of modern molecular methods (Meyer *et al.* 2006; Culumber *et al.* 2011; Jones *et al.* 2012; Kang *et al.* 2013). We previously reported that a species of northern swordtail, *X. nezahualcoyotl*, had significant hybrid ancestry in its transcriptome (Cui *et al.* 2013). While the majority of the transcriptome of *X. nezahualcoyotl* indicates it is sister to *X. montezumae*, a large fraction appears to be derived from a parapatric swordtail species, *X. cortezi* (Fig. 1). Here, we quantitatively evaluate this pattern using whole-genome sequences for all three species. In addition, we evaluate the extent, and timing, of genome stabilization and ask whether hybridization-derived regions exhibit specific properties in relation to gene content and constraint.

Methods

Sample collection

Sample collection procedures for this study were approved by the Texas A&M Institutional Animal Care and Use Committee (Protocol # 2013-0168) and the Mexican federal government (see Acknowledgments). Individuals were collected using baited minnow traps and seine nets in the states of San Louis Potosí and Tamaulipas, Mexico, between 2013 and 2015. *X. cortezi* and *X. montezumae* were collected from southern San Louis Postosí from the Río Huichihuayan in May of 2013 and the Río Tamasopo in March of 2015, respectively. *X. nezahualcoyotl* were sampled from two distinct populations: one individual from Arroyo Los Gallitos (hereafter 'Gallitos') in southern Tamaulipas, Mexico, in 2014 and one from a stock collected at Arroyo Las Crucitas (hereafter 'Crucitas') in southern San Louis Potosi, Mexico, in 2011. Lateral photographs of each fish were taken prior to tissue sampling. For sequencing, one male of each species was anaesthetized using tricaine methanesulfonate (MS-222). A fin clip was then taken from the upper portion of the caudal fin and preserved in 95% ethanol.

Genomic DNA extraction and library preparation

Genomic DNA was extracted from fin clips using the DNeasy kit (Qiagen, Valencia, CA, USA) and quantified and assessed for purity using a Nanodrop 1000 (Thermo Scientific, Wilmington, DE, USA). One microgram of DNA was sheared into 500-bp fragments using a Covaris LE220 sonicator (Covaris, Woburn, MA, USA) and prepared for sequencing following the protocol of Quail *et al.* (2009). Briefly, sheared DNA was end-repaired and an A-tail was added to facilitate adaptor ligation. Following adaptor ligation, libraries were run on a 2% agarose gel to select products between 400 and 600 bps and purified using a Qiagen gel purification kit (Qiagen). Following purification, samples were PCR-amplified with custom Illumina-style indexed primers for 12–14 cycles using the Phusion high-fidelity polymerase system (NEB, Ipswich, MA, USA). PCR products were purified with Agencourt AMPure XP beads

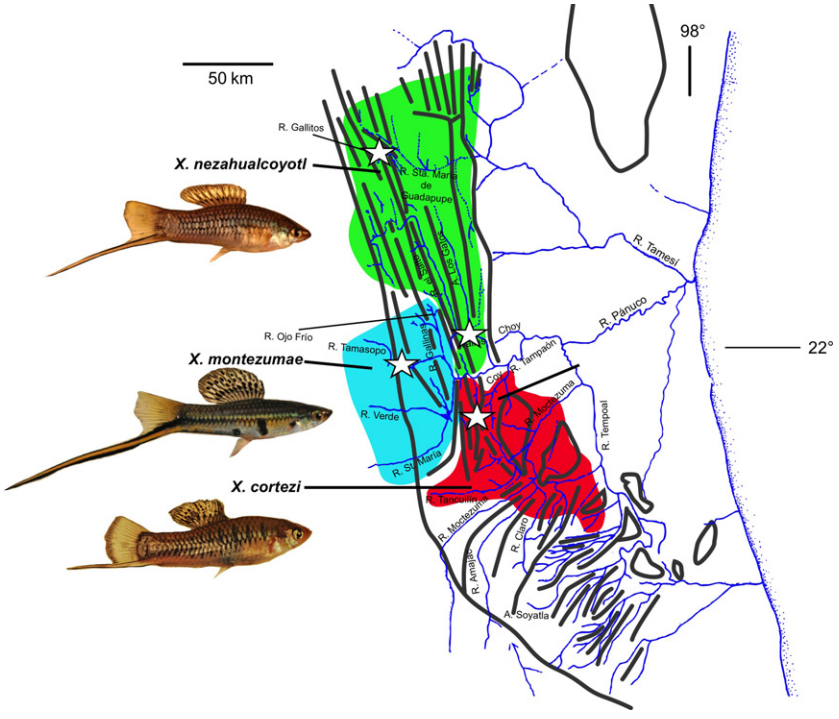

Fig. 1 Range of focal species and sampling locations. Map of the geographical ranges and major river systems of the *X. nezahualcoyotl* (green), *X. montezumae* (blue) and *X. cortezi* (red). Stars indicate exact sampling locations of the individuals used for whole-genome sequencing in this study. Inset photographs on the left show individuals from sampled populations (*X. nezahualcoyotl* – Los Gallitos population shown).

(Beckman Coulter, Brea, CA, USA), and the size distribution and purity was evaluated on a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). Libraries were sequenced on a total of nine lanes of an ILLUMINA HISEQ 2000 machine with v3 chemistry, yielding 101-bp single-end reads. Additional 141-bp single-end reads obtained on the Illumina HiSeq 2500 were collected for the *X. montezumae* individual to increase coverage. All sequence data are available through the NCBI sequence read archive (SRX1518311, SRX1518380, SRX1518263, SRX1518028).

Sequence processing and variant calling

Raw sequences were parsed by index using a custom python script (https://gist.github.com/dgrtwo/3725741). The number of reads per sample (ranging from 196 to 398 million) and mapping statistics are summarized in Table S1 (Supporting information). We refrained from trimming reads to remove low-quality bases, as recommended by the GATK pipeline (McKenna *et al.* 2010). However, we did trim reads to remove potential adapter sequences using the program CUTADAPT v1.9 (Li *et al.* 2012). Reads were mapped to the *X. maculatus* reference genome version 4.4.2 using the bwa-mem algorithm (Li & Durbin 2009). Unplaced contigs were included for read mapping but not included in downstream analyses. Variant calling was performed using the GATK version 3.4 pipeline (McKenna *et al.* 2010). Briefly, sam files were converted to bam files, duplicates were marked and removed, and bam files indexed using PICARD tools v1.118. INDEL realignment was performed with GATK, and genotypes were called with HaplotypeCaller in the gVCF format. Because there is no library of verified SNPs for the *Xiphophorus* species in this study, we did not perform base recalibration or variant recalibration. Instead, we filtered the final gVCFs for each species to identify sites for low coverage (DP < 10), low quality score (variant quality – GQ or invariant quality – RGQ < 20), low mapping quality (MQ < 40), low quality by depth score (QD < 2), high fisher strand score (FS > 60), high strand odds ratio (SOR > 4), low read position rank sum score (ReadPosRankSum < −8) and low mapping quality rank sum score (MQRankSum < −12.5). We also masked INDELs and all sites within 5 bp of INDELs. These thresholds were determined based on applying GATK's recommendations to simulated data and adjusting thresholds to improve accuracy (see Supporting information 1). Note that for invariant sites only DP and RQG filters could be used. To generate reference-based genomes for analysis, we used the *X. maculatus* genome as a scaffold and updated variant sites for each species using the program seqtk (https://github.com/lh3/seqtk), masking variant and invariant sites identified as with low quality based on the above filters. Using simulated data, we confirm that this align-to-reference and variant-calling pipeline has high sensitivity to real variants (Fig. S1, Supporting information), low error rates (Fig. S2, Supporting information) and is unlikely to result in artefacts that could be misinterpreted as hybrid ancestry using our approach (Fig. S3, see Supporting informations 2 and 3).

Whole-genome phylogenetic analysis and phylogenetic discordance

To infer the species tree for the five reference-based genomes (*X. nezahualcoyotl* (Gallitos), *X. nezahualcoyotl* (Crucitas), *X. montezumae*, *X. cortezi* and the *X. maculatus* genome reference as an outgroup), we estimated a maximum-likelihood tree with 100 rapid bootstraps using a General Time Reversible model with a gamma distribution of substitution rates (GTR + GAMMA) as implemented in RAxML 7.2.8 (Stamatakis 2006). For this initial analysis, the data was analysed as a single partition. The final length of the RAxML matrix after excluding 16% gaps and undetermined characters was 662 Mb. Genomewide, there were 272–282 000 phylogenetically informative sites (for the Crucitas and Gallitos *X. nezahualcoyotl* samples, respectively). The distribution of missing data for each sample is shown in Fig. S4.

To estimate the proportion of the genome supporting admixture between the *X. nezahualcoyotl* and *X. cortezi* lineages, we divided the whole-genome alignments into 10 kb segments and performed an approximately unbiased (AU) test on these alignments (Shimodaira 2002). We performed this analysis with the two *X. nezahualcoyotl* sequences separately, testing support for the three possible topologies of an unrooted four-taxon tree. We calculated site likelihoods using RAxML (as above) and input these likelihoods into Consel 0.2 (Shimodaira & Hasegawa 2001). For each topology, this program gives the probability that the topology is as likely as the maximum-likelihood tree. If a particular alignment had a probability >0.95 for a topology, we treated this as support for that topology. We excluded alignments that had a likelihood difference of 0 between topologies because this can be caused by a low number of informative sites (Schmidt 2009). Because this approach tends to result in ambiguous probabilities for windows of mixed ancestry (see Schumer *et al.* 2012), it will likely underestimate the proportion of the genome derived from hybridization when hybridization-derived regions are small.

Test for gene tree asymmetry

There are two major reasons why the phylogenetic relationships in a particular region of the genome will differ from the true phylogenetic relationships between species. Incomplete lineage sorting (ILS) can result in support for alternative phylogenetic relationships in particular genomic regions (Maddison 1997). A second common cause of this discordance is hybridization between species (Green *et al.* 2010). One signature of hybridization that can differ from ILS is the relative representation of different phylogenetic relationships. If there has been biased gene flow, more regions of the genome will support a particular minor topology than expected. The expectation under ILS is that an equal number of discordant trees will support each of the two possible minor topologies in a four-taxon tree. However, it is important to keep in mind that population structure during speciation can also generate gene tree asymmetry (Green *et al.* 2010). To test for significant gene tree asymmetry, we performed 1000 bootstrap resamplings with replacement of the AU test results. This allowed us to calculate confidence intervals for the proportion of the genome supporting each topology and assess whether confidence intervals for the two minor topologies overlap. As noted above, significant differences in the proportion of the genome supporting each minor topology is evidence for gene flow (or population structure during speciation).

Identifying and delineating phylogenetically discordant regions

The approaches described above have two limitations. First, while both methods allow us to estimate the proportion of the genome that supports each possible topology, they do not allow us to infer the boundaries of these regions. Second, these analyses do not allow us to distinguish between ILS and hybridization for particular genomic regions.

To address both of these limitations, we use the program PhyloNet-HMM (Liu *et al.* 2014) to delineate between regions supporting different hybridization scenarios. PhyloNet-HMM uses a hidden Markov model to detect breakpoints between regions supporting different phylogenetic relationships. PhyloNet-HMM uses the concept of parent trees that correspond to different hybridization scenarios (e.g. hybridization tree and species tree) and gene trees, which correspond to different gene tree relationships within each parent tree. Thus, this program theoretically allows for ILS of gene trees within a parent tree without switching parent tree topologies. As a result, switching from the parent tree to the hybridization tree should only occur in response to a change in hybridization signal. Our performance tests suggest that this approach accurately distinguishes between ILS and hybridization. PhyloNet-HMM rarely misassigns regions with only ILS to the hybridization tree (Fig. S5, Supporting information 4), suggesting that the majority of regions it identifies will be truly hybridization-derived.

We ran PhyloNet-HMM separately on alignments of *X. montezumae*, *X. maculatus*, *X. cortezi* and each *X. nezahualcoyotl* sample. We divided the genome into 2 Mb alignments to increase computational speed. We specified two possible parental trees, the species tree with *X. montezumae* most closely related to *X. nezahualcoyotl*, and a hybridization tree with *X. cortezi* most closely related to *X. nezahualcoyotl*. An example of the parameters file used is available in Appendix S1. Briefly, we specified substitution rates and base frequencies based on RAxML results but allowed the program to optimize all other parameters including branch lengths. We treated posterior probabilities >0.95 as support for a particular parent tree, and delineated regions of support as stretches of sites with ≥0.95 posterior support for a particular parent tree (Fig. S6, Supporting information). Note that this approach systematically underestimates the length of hybridization-derived regions (Fig. S7, Supporting information).

As recombination rates vary throughout the genome, the sizes of discordant regions will depend on the local recombination rate. Based on our performance analyses (Supporting information 4), we analyse two data sets with different expected false discovery rates, a full data set (false-positive rate 2%) and a stringent data set filtered by region size to reduce the false-positive rate (false-positive rate <0.5%). This more conservative data set only included regions >0.02 cM in length (see Supporting information 5).

Determining the direction of introgression

Phylogenetic approaches such as PhyloNet-HMM can identify potentially introgressed regions, but are not informative about the direction of introgression, or in our case, whether the region introgressed from the *X. cortezi* lineage into *X. nezahualcoyotl* or vice versa. To investigate the direction of gene flow, we compared divergence between species at regions inferred to be hybridization-derived. As an additional approach, we applied the D_{FOIL} approach (Pease & Hahn 2015; https://github.com/jbpease/dfoil). This approach, an extension of the widely used D-statistic (Green *et al.* 2010), uses site patterns in a 5-taxon tree to polarize the direction of introgression. In our case, we used the genome sequence of *X. malinche* which together with *X. birchmanni* forms the sister clade to *X. cortezi* (Cui

et al. 2013). As genomewide patterns were suggestive of a complex history of introgression (see Supporting information 6), we also applied D_{FOIL} to individual regions identified by PhyloNetHMM. Specifically, we ask what proportion of regions supported each of the two directions for introgression (i.e. into *X. nezahualcoyotl* versus into *X. cortezi*) at a *P*-value cut-off of 0.05. We note that applying D_{FOIL} to small regions can result in a higher than expected false-positive rate (see Pease & Hahn 2015), but we only apply D_{FOIL} to regions already identified as hybridization-derived by PhyloNet-HMM.

Evaluating stabilization of hybridization-derived regions

Based on our simulations, PhyloNet-HMM does not effectively infer hybridization at regions that are heterozygous for ancestry (see Supporting information 7). When PhyloNet-HMM encounters such regions, it almost always assigns them ambiguous posterior probabilities (Supporting information 7). Other HMM-based chromosome painting methods, such as Multiplexed Shotgun Genotyping, allow for inference of heterozygous hybrid ancestry, but have low sensitivity to short ancestry tracts and as a result cannot be effectively used here (Andolfatto *et al.* 2011; Schumer *et al.* 2015a). Only ~15% of the genome was assigned ambiguous posterior probabilities by PhyloNet-HMM in our analysis (see Results). These ambiguous signals could be produced by ancestry heterozygosity but also by many other factors. As a first pass for identifying heterozygous hybrid ancestry regions, we asked whether nucleotide heterozygosity was higher than expected at sites inferred to be ancestry informative between *X. montezumae* and *X. cortezi*. Using the approaches described in Supporting information 4 (incorporating population specific estimates of θ; Table S1, Supporting information), we perform coalescent simulations to estimate the number of sites that are expected to be polymorphic in *X. nezahualcoyotl* for alleles that are ancestry informative between *X. montezumae* and *X. cortezi* (e.g. *X. nezahualcoyotl* genotype AC, *X. montezumae* AA, *X. cortezi* CC). We perform 1000 simulations of 1 Mb regions to generate a null distribution.

As an additional approach, we asked whether ambiguous regions showed evidence of *cortezi–nezahualcoyotl* hybridization using the D-statistic (Green *et al.* 2010), and how the signal in these regions compared to the rest of the genome and regions identified as hybridization-derived by PhyloNet-HMM. To capture introgression at polymorphic sites with this approach, one allele was randomly sampled at each polymorphic site when calculating D.

Estimating the timing of hybridization and genome stabilization

Due to the complex history of introgression (see Results), it is not straightforward to determine either the absolute time since hybridization occurred between the *X. nezahualcoyotl* and *X. cortezi* lineages or how long it took for hybrid ancestry to stabilize in the genome. Despite these difficulties, we apply two approaches to begin to address these questions. First, we compare divergence between hybridization-derived regions and the rest of the genome (see next section). Second, we use the length of hybridization-derived haplotypes to investigate the time between hybridization and stabilization of hybrid ancestry in the genome (Supporting informations 8 and 9). This approach has several limitations when applied to our data (see full discussion in Supporting information 8), and thus, these results should be viewed as a first approximation.

Properties of hybridization-derived regions

Our aim is to determine whether hybridization-derived regions are atypical compared with the genomic background for a number of features (see sections below). One challenge in performing these comparisons is ensuring that the genomic background does not differ from hybridization-derived regions simply due to biases in detection. We use different approaches to control for this depending on the analysis, but for all comparisons, we draw regions for null data sets only from regions of the genome that were confidently called for the species tree by PhyloNet-HMM. In addition, only regions identified as hybridization-derived using both *X. nezahualcoyotl* populations were used in these analyses. As discussed above, we analysed two data sets, one restricted to regions longer than 0.02 cM ('stringent data set', estimated <0.5% FDR) and one including all segments identified as hybridization-derived by PhyloNet-HMM ('full data set', estimated 2% FDR). All null data sets were matched in region size to the corresponding focal data set.

Although the majority of introgression we detect is from the *X. cortezi* lineage into the *X. nezahualcoyotl* lineage (see Results), we also detect introgression in the opposite direction. This means that the regions we analyse are actually a combination of two distinct evolutionary histories, which could generate biases in our results. To address this concern, we repeated the major analyses described below using only sequences with significant evidence for directional introgression from *X. cortezi* (based on D_{FOIL} analysis, see above).

Gene density. To compare gene density between hybridization-derived regions and other regions of the genome, we downloaded the gene annotation file (GTF) for the *X. maculatus* assembly version 4.4.2, linkage group version 1.0 from http://genome.uoregon.edu/xma/ (Amores *et al.* 2014). Initial analyses with this GTF file uncovered a number of problems in the coordinates of genes (e.g. many premature stop codons, extracted gene region did not align to the listed gene identity). As a result, we regenerated the GTF file using the linkage group apg file from Amores *et al.* (2014) and corresponding version of the genome, the previous version of the genome in scaffold form (ftp://ftp.ensembl.org/pub/release-81/fasta/xiphophorus_maculatus/dna/), and the GTF file generated by Ensembl for the scaffolded version of the genome (ftp://ftp.ensembl.org/pub/release-81/gtf/xiphophorus_maculatus/). We used the programs chain.py (https://github.com/tanghaibao/jcvi/blob/master/formats/chain.py) and crossmap (http://crossmap.sourceforge.net/) to convert the agp file into a chain file and then lift the Ensembl GTF file onto the linkage group version of the genome. Using these, GTF coordinates resulted in only 1.5% of genes with premature stop codons (which were subsequently excluded in PAML analyses; see below). We deposited this GTF file on Dryad (doi:10.5061/dryad.tm47d).

We generated bed files containing the coordinates of discordant regions (both >0.02 cM filtered and full data sets) and used bedtools2 (Quinlan & Hall 2010) to intersect discordant regions with entries in the GTF file. We then counted the number of unique protein coding genes that overlapped with regions derived from hybridization. We bootstrapped the data with replacement 1000 times to generate a distribution of gene counts in these discordant regions. We compared this distribution to a null distribution with the same number of regions as described above and calculated the proportion of base pairs that were coding in each of the focal and null data sets.

Conserved genomic regions. Regions with low levels of divergence over large evolutionary timescales ('conserved regions') are more likely to be functionally important and as a result may be less likely to introgress. To identify these regions, we first generated alignments between the zebrafish and *X. maculatus* reference genomes. We used the repeat-masked version of danRer7 from the UCSC genome browser source because this genome had already been aligned to several other fish species: Stickleback (UCSC gasAcu1), Medaka (oryLat2), Fugu (fr3) and Pufferfish (tetNig2) genomes. We followed the pipeline for whole-genome alignment outlined by the UCSC genome browser wiki site (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto) and used the program tba (Blanchette *et al.* 2004) to combine our zebrafish–*X. maculatus* alignment with the available alignments into a single maf file containing five fish species aligned to zebrafish. We used the maf_project command of tba to convert *X. maculatus* to the reference sequence in the merged maf file and exclude alignments lacking *X. maculatus*. maf files were sorted by chromosome using a custom perl script and by coordinate using the mafSorter command from the package mafTools (Earl *et al.* 2014). We then used the program phyloFit to generate phylogenetic models for conserved and nonconserved sites (Siepel & Haussler 2004). These were used as input to the program phastCons; we ran phastCons on alignments split by chromosome, setting the target-coverage parameter to 0.25, the expected length parameter to 12, and rho to 0.4 (Siepel *et al.* 2005).

phastCons outputs likelihoods of conservation at individual bases as well as a set of coordinates and likelihoods for the most conserved elements in the genome. We overlapped the most conserved elements output by phastCons with PhyloNet-HMM results, arbitrarily focusing on the top 25th percentile of conserved elements (conservation log-likelihood score of 56), to ask about the distribution of conserved elements relative to hybridization-derived regions. In our analysis, the conservation log-likelihood score is strongly associated with length of the detected conserved element (Pearson's correlation: 0.94). We compared the proportion of highly conserved elements found in regions called for the hybridization tree versus the species tree to null data sets with randomly selected regions of the same length. We repeated this analysis for the proportion of conserved base pairs. Note that for this analysis we did not limit randomly selected sites to regions where the species or hybridization tree was confidently inferred by PhyloNet-HMM as in this case the focal data set of conserved elements was distributed throughout the genome.

Genomic and coding region divergence. We examine divergence in parts of the genome derived from hybridization both at the gene level and region level. For analysis of raw divergence in regions derived from hybridization, we calculated D_{xy} between the *X. montezumae* and *X. cortezi* genomes in these regions and compared these values to regions randomly selected from parts of the genome supporting the species tree (see details below).

To analyse coding divergence within the hybridization-derived regions, we extracted exons for each gene in these regions and calculated dN, dS and dN/dS using codeml in paml v4.8a with the F3X4 codon model

(Yang 1997). All genes with premature stop codons were excluded from the analyses. We primarily focus the PAML analysis on *X. montezumae* and *X. cortezi* but also repeat the analysis including *X. maculatus* and *X. hellerii* (data from Schumer *et al.* 2012). See Supporting information 10 for details and results of this additional analysis.

As our power to call regions as hybridization-derived is dependent on divergence between species (Supporting information 4), this analysis requires additional controls when generating null data sets. In addition to drawing null regions from those confidently called for the species tree, to detect when we are likely to have low power (defined as ≤0.9 power, see Supporting information 4), we flagged regions that fell below the 10th percentile of divergence between *X. montezumae* and *X. cortezi*. For each null data set containing low-power regions (average 7.4 of 452 for the stringent data set and 50 of 2282 for the full data set), we calculated summary statistics with and without them.

Functional analysis of regions derived from hybridization. To determine whether particular functional categories or pathways are over-represented in discordant genomic regions, we performed gene ontology (GO) analysis using a previously described custom pipeline with the GOstats package in R (Falcon & Gentleman 2007; Schumer *et al.* 2014) and KEGG pathway analysis using DAVID v6.7 (Huang *et al.* 2009). For both GO and KEGG analyses, annotated genes in the *X. maculatus* reference genome were matched with HUGO gene names and only these genes were used in the analysis. For GO analysis, we analysed biological process, molecular function and cellular component annotations and tested for significance using a hypergeometric function with a *P*-value threshold of 0.05 (Falcon & Gentleman 2007).

Because hybridization-derived regions are clustered, which could result in multiple genes with similar annotations or pathways being chosen at random, we also performed the above analyses on 10 null data sets. As GO and KEGG analyses may be sensitive to both the number of genes in a data set and the clustering of these genes in regions, instead of generating null data sets with the exact number of regions as the focal data set, we randomly sampled region sizes from the focal data set and continued to sample null regions until we reached the number of genes observed in the focal data set. We repeated the analyses described above on these data sets and asked whether null data sets generated fewer or less significantly enriched GO terms than the real data.

Results

Genome sequencing and species tree

Average genomewide depth coverage of the four sequenced swordtail genomes ranged from 21 to 41× (Table S1, Supporting information). All northern swordtail genomes were ~1.5% diverged from the *X. maculatus* reference (see Supporting information 1 for proof of principle in mapping to a divergent reference). Average pairwise sequence divergence (D_{xy}) between the sampled individuals ranged from 0.1% between the two *X. nezahualcoyotl* individuals to 0.65% for *X. montezumae–X. cortezi* (see Table S2, Supporting information for pairwise comparisons). The two *X. nezahualcoyotl* individuals, sampled from different populations, differed considerably in levels of per site nucleotide heterozygosity, 0.025–0.08%. Notably, *X. montezumae* and the *X. nezahualcoyotl* (Gallitos) exhibit remarkably low levels of polymorphism (Table 1).

Analysis of whole-genome concatenated alignments with RAxML resulted in a high confidence species tree with 100% bootstrap support for all internal nodes (Fig. 2). This species tree places the two *X. nezahualcoyotl* samples sister to *X. montezumae* (Fig. 2), as previously reported (Cui *et al.* 2013; Jones *et al.* 2013).

Gene tree asymmetry

We performed AU test analyses of the two *X. nezahualcoyotl* individuals separately in case the two populations had different histories of hybridization. Despite being sampled from distinct populations, the genome-wide patterns of hybrid ancestry were nearly identical in the

Table 1 Summary of divergence, polymorphism and alignment statistics for the four individuals sequenced for this study. See Table S1, Supporting information for more details

Sample	Divergence from *X. maculatus* per site	Nucleotide diversity (θ_π) per site	Average depth of coverage	Fraction of genome with depth coverage ≥10
X. nezahualcoyotl Gallitos	0.016	0.00025	41	0.96
X. nezahualcoyotl Las Crucitas	0.015	0.00082	26	0.94
X. cortezi	0.015	0.0011	26	0.95
X. montezumae	0.016	0.00030	21	0.94

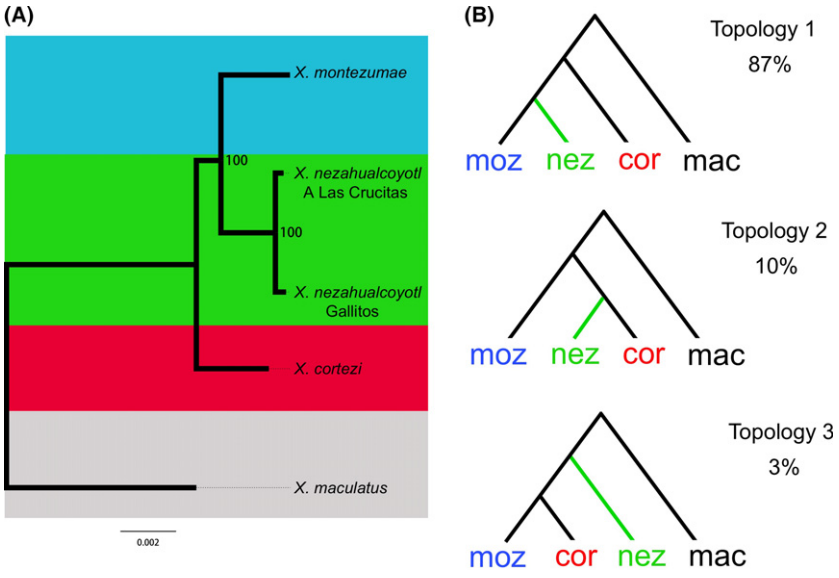

Fig. 2 Phylogenetic analysis of *X. montezumae*, *X. cortezi* and *X. nezahualcoyotl*. (A) Species tree based on whole-genome alignments and RAxML analysis. (B) Results of phylogenetic analysis with the AU test in 10 kb windows show the same major phylogenetic pattern as in A but major asymmetry in the two minor topologies (topology 2 and topology 3), suggestive of hybridization between the *X. cortezi* and *X. nezahualcoyotl* lineages. nez – *X. nezahualcoyotl*, moz – *X. montezumae*, cor – *X. cortezi* and mac – *X. maculatus*.

two samples. When ambiguous topologies were excluded in the analysis, 86% of 10 kb alignments supported the species tree relationship, while 10% of regions supported the sister relationship of *X. nezahualcoyotl* and *X. cortezi* in both samples (Crucitas – 10 ± 0.3%, Gallitos – 10.1 ± 0.4%; Fig. 2B). In contrast, only 3% of regions supported the sister relationship of *X. cortezi* and *X. montezumae*. This proportion of regions supporting the *cortezi–montezumae* topology is consistent with results from our coalescent simulations which predict 3.4 ± 0.4% of alignments supporting this topology due to ILS alone (confidence intervals from 1000 bootstrap resamplings of alignments). The *P*-value for asymmetry of the two minor topologies is <0.001 by bootstrapping. This excess in trees supporting a sister relationship between *X. nezahualcoyotl* and *X. cortezi* results in an estimate of 7.5–8% of the genome being derived from hybridization using the estimator proposed by Yu *et al.* (2012). This estimate is similar to the proportion of the genome found to support introgression based on PhyloNet-HMM results (8%, Fig. 3, Table S3, Supporting information).

Determining the direction of introgression

The phylogenetic approaches we have used above identify regions likely to represent introgression between the *X. nezahualcoyotl* and *X. cortezi* lineages but are not informative about its direction. Patterns of divergence in regions identified as hybridization-derived strongly suggest that the majority of gene flow occurred from the *X. cortezi* lineage into *X. nezahualcoyotl* (see Results). To investigate this pattern in more detail, we used the program D_{FOIL}, which leverages data from additional species to polarize the direction of introgression (see Methods). This analysis suggests a complex history of introgression, potentially consistent with bidirectional introgression between the *X. cortezi* and *X. nezahualcoyotl* lineages (J. Pease, personal communication; see also simulations in Supporting information 11). Analysis of directional patterns with D_{FOIL} in individual introgressed regions (identified by PhyloNet-HMM) demonstrates that many regions have too few informative sites to confidently assign the direction of gene flow. However, of the regions where significant directional introgression was detected at $P < 0.05$ ($N = 250$ using *X. nezahualcoyotl* Gallitos and $N = 245$ using *X. nezahualcoyotl* Crucitas), 76% supported introgression from *X. cortezi* into *X. nezahualcoyotl*. This finding is consistent with expectations based on patterns of divergence (Fig. 4) and suggests that the major direction of gene flow has been from the *X. cortezi* lineage. Regions that could be assigned an introgression direction were longer on average than other hybridization-derived regions (29 vs. 19 kb), but importantly, region length was similar in both introgression directions (95% CI into *X. nezahualcoyotl:* 3–115 kb, into *X. cortezi*: 3–109 kb).

Two distinct models of hybridization could explain the general patterns observed in our data. In the first model (Fig. S8, Supporting information), introgression primarily occurred from the *X. cortezi* lineage into the *X. nezahualcoyotl* lineage after the separation of the *X. nezahualcoyotl* and *X. montezumae* lineages. In the second model (Fig. S8, Supporting information), the *X. nezahualcoyotol* lineage was the product of hybridization between the *X. cortezi* and *X. montezumae* lineages, with the *X. montezumae* lineage contributing the majority of the genome. Distinguishing between these models is difficult given that hybridization was ancient

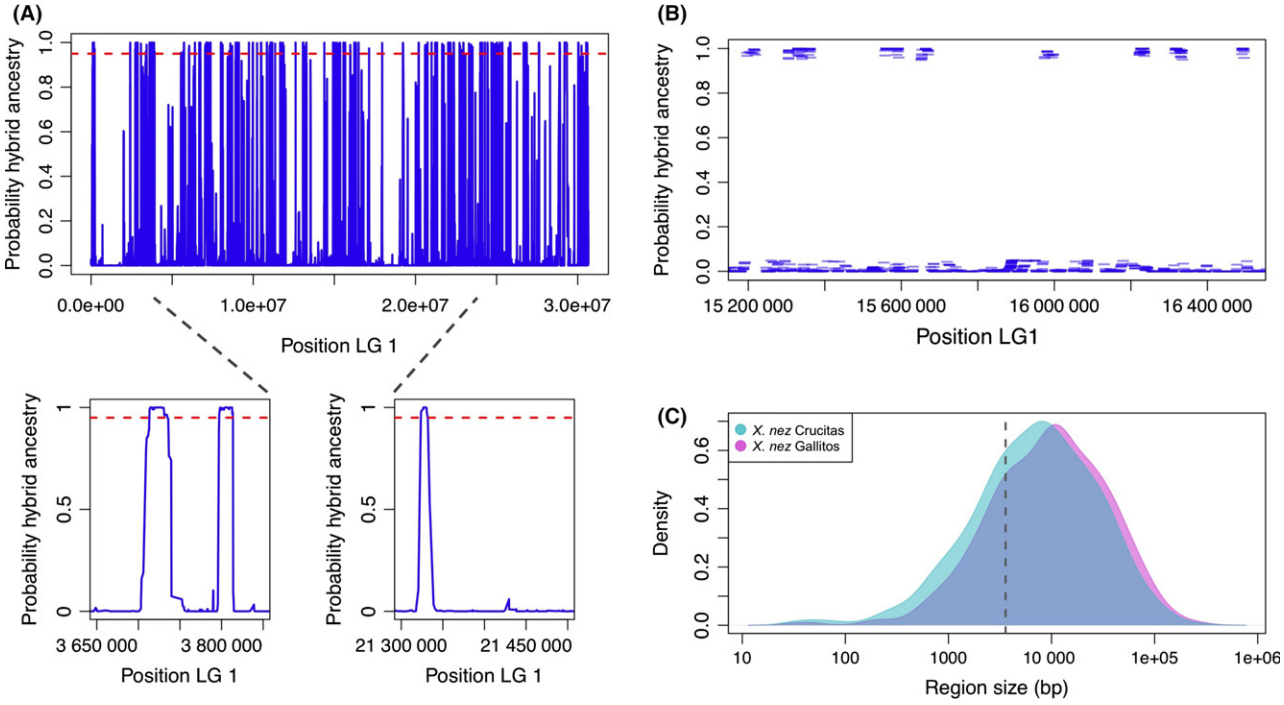

Fig. 3 Detection and delimitation of hybridization-derived regions on Linkage Group 1 using PhyloNet-HMM. (A) Posterior probabilities for the hybridization topology by position across Linkage Group 1, with insets showing close-ups of representative regions. (B) Example of a subset of Linkage Group 1 data after processing to remove sites that lack posterior support for any gene tree at ≥0.95. (C) Size distribution of regions inferred to be hybridization-derived by PhyloNet-HMM plotted on a log scale.

and we are uncertain about ancestral population sizes. However, we can begin to distinguish between these scenarios by looking at relative divergence between regions in *X. nezahualcoyotl* most closely related to *X. montezumae* and those most closely related to *X. cortezi* (see next section). When normalizing by divergence between *X. montezumae*–*X. cortezi*, we see that regions most closely related to the *cortezi* lineage in *X. nezahualcoyotl* are significantly less divergent from *X. cortezi* than the equivalent comparison for regions most closely related to *X. montezumae* (95% CI of normalized divergence of *cortezi* regions: 0.43–0.46, *montezumae* regions: 0.51–0.54). Because normalized divergence to *montezumae* regions is greater than normalized divergence to *cortezi* regions, we frame our discussion of the hybridization scenario in the context of gene flow from the *X. cortezi* lineage into the *X. nezahualcoyotl* lineage, but note that we cannot presently quantitatively distinguish the two models outlined above.

Delimiting regions derived from hybridization

Using PhyloNet-HMM, we were able to assign 84% of the genome to the species tree or hybridization tree with high confidence (Figs 3 and 5A). The 95% confidence intervals of the lengths of segments inferred to be hybridization-derived ranged from 1 to 100 kb (Fig. 3). The total number of regions supporting hybridization between *X. nezahualcoyotl* – *X. cortezi* lineages was 2282, and the average region length was 20 kb. Regions identified as discordant by the AU approach largely overlapped with regions identified as hybridization-derived by PhyloNet-HMM (Fig. S9, Supporting information). Strikingly, these regions were remarkably repeatable between the two *X. nezahualcoyotl* individuals sampled. About 99.9% of sites that were associated with introgression from the *X. cortezi* lineage in the *X. nezahualcoyotl* Crucitas sample were also associated with introgression in the Gallitos sample, and 99.7% vice versa (Table S4, Supporting information). Including ambiguous regions in this analysis, only 3% of sites fell into different categories in the two samples (species tree, hybridization tree or ambiguous). This suggests that most of the *X. nezahualcoyotl* genome has stabilized for hybrid ancestry.

Although few sites show distinct patterns between the two *X. nezahualcoyotl* populations, we perform several analyses to further investigate the possibility that ambiguous regions might be segregating for hybrid ancestry. If hybridization-derived regions were still segregating in *X. nezahualcoyotl*, we would expect to

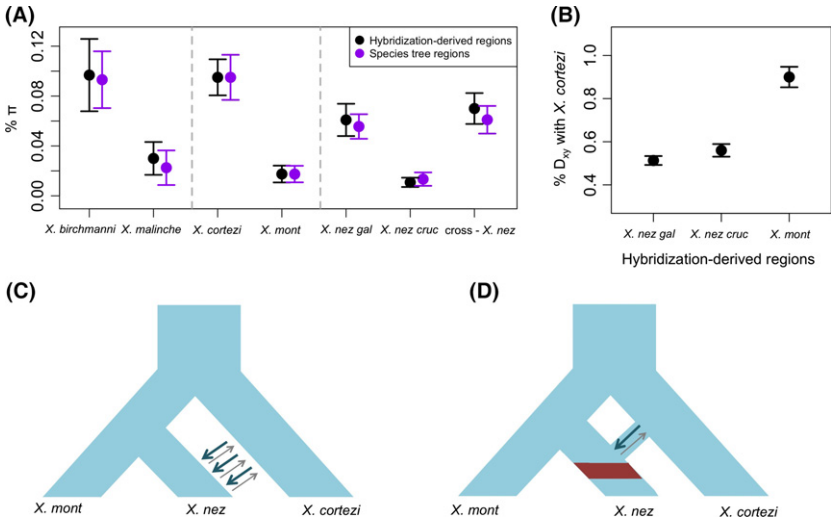

Fig. 4 Summary of nucleotide diversity, divergence and potential evolutionary scenarios. (A) Mean percentage per site heterozygosity (θ_π) (with error bars indicating two standard deviations of the mean) in regions identified as hybridization-derived or following the species tree in this study in a range of *Xiphophorus* species. Data for *X. birchmanni* and *X. malinche* were re-analyzed from Schumer *et al.* 2014. (B) D_{xy} (percentage average pairwise sequence divergence per site) between both *X. nezahualcoyotl* samples and *X. cortezi* and *X. montezumae* and *X. cortezi* in regions that are derived from hybridization (see also Fig. S12). Error bars indicate two standard errors. (C and D) show two possible evolutionary scenarios for the origin of hybrid ancestry in *X. nezahualcoyotl*. In C, ongoing migration introduces hybrid ancestry into *X. nezahualcoyotl*, resulting in segregating hybrid ancestry. In D, a pulse of hybridization followed by stabilization (red shading), driven by a bottleneck or selection, results in stabilized hybrid ancestry in *X. nezahualcoyotl*. The latter scenario is most consistent with results comparing the two *X. nezahualcoyotl* samples.

see regions of high polymorphism for *X. montezumae* and *X. cortezi* alleles. In *X. nezahualcoyotl* Crucitas, 14 440 such sites were observed, while 3938 were observed in *X. nezahualcoyotl* Gallitos. The upper 5% quantile predicted from neutral simulations was 17 119 and 4397, respectively. Thus, we do not observe an excess of polymorphism at *X. montezumae* and *X. cortezi* ancestry informative sites in either *X. nezahualcoyotl* sample, suggesting that few *X. cortezi* alleles are still segregating in either of the *X. nezahualcoyotl* individuals.

As a more sensitive approach, we calculated the D-statistic for regions ambiguously assigned by PhyloNet-HMM. Although the overall signature of introgression from ambiguous regions is weaker than the genome-wide signature (Fig. S10, Supporting information), we do see significant evidence for *cortezi–nezahualcoyotl* introgression in these regions. Interestingly, the regions with the strongest signal of introgression (the shortest 25% quantile of ambiguous regions) do have an excess of ancestry polymorphic sites (i.e. twofold–threefold greater than the genomic background; $P < 0.001$ in the Crucitas sample and $P = 0.1$ in the Gallitos sample). This suggests that there could still be regions segregating for hybrid ancestry in *X. nezahualcoyotl*, but the subset of ambiguous regions driving this signal make up only a small proportion of the genome (Fig. S10, Supporting information).

Estimates of the time to genomic stabilization and the time since admixture

By examining divergence between hybridization-derived haplotypes in *X. nezahualcoyotl* and these same regions in *X. cortezi* (Fig. S11, Supporting information), it is apparent that there has been substantial divergence between these regions, ~4× more than would be expected based solely on levels of polymorphism in *X. cortezi*. This suggests that a significant amount of time has passed since introgression (~50 000 generations assuming a per base pair mutation rate of 3.8×10^{-8}; Fig. 4, Fig. S11, Supporting information). However, analysis of tract lengths (Gravel 2012; Jin *et al.* 2014; Fig. S12; see details in Supporting information 8) suggests that hybridization-derived tracts stabilized quickly in the *X. nezahualcoyotl* genome, after ~2500 generations (or ~1250 years assuming two generations/year). Although this estimate should be interpreted with caution (see Supporting information 8), it implies that hybrid ancestry may have stabilized relatively quickly. Such rapid stabilization is inconsistent with genetic drift in large- or moderate-sized populations, but simulations suggest that a strong bottleneck could explain this rapid stabilization (Supporting information 9). Similar diversity levels in hybridization-derived regions and other genomic regions (Fig. 4) are also consistent with fixation driven by a strong bottleneck; however, fixation likely

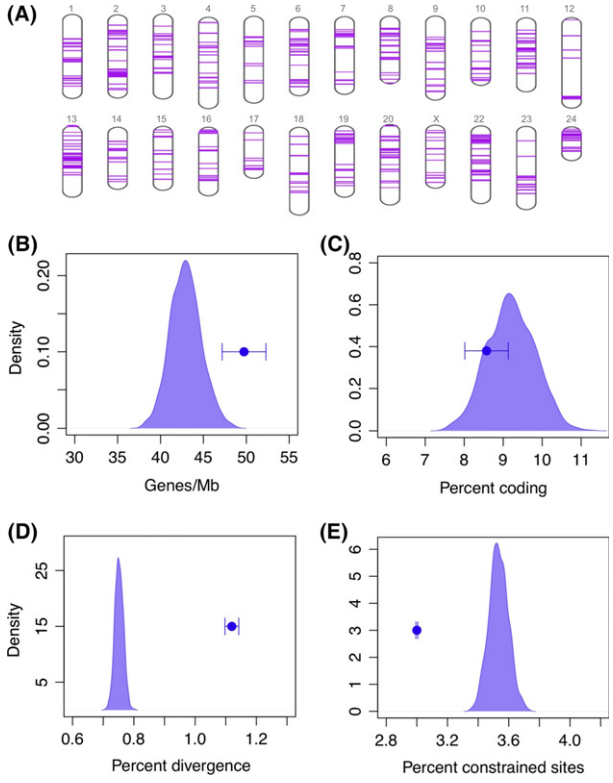

Fig. 5 Differences between hybridization-derived genomic regions and the genomic background. (A) Locations of hybridization-derived regions in the genome (stringent data set). (B) Gene density is significantly higher in hybridization-derived regions than the genomic background, but the proportion of coding sites is not (C). (D) Hybridization-derived regions are significantly more diverged between *X. montezumae* and *X. cortezi* than the genomic background and (E) have fewer evolutionarily constrained sites, as determined by PHASTCONS analysis. Error bars indicate two standard deviations of the mean. Results shown here are for the stringent data set; the same plots for the full data set can be seen in Fig. S14.

occurred sufficiently long ago for recovery of polymorphism in hybridization-derived regions.

Nonrandom patterns in hybridization-derived regions

We asked whether hybridization-derived regions show any unusual patterns with respect to gene density, divergence or rates of evolution compared with the genomic background. To do this, we considered two data sets: (i) all 2282 regions identified above 'full data set' and (ii) a data set of 452 regions filtered to exclude regions with a length of <0.02 cM 'stringent data set'. The second data set is expected to have a lower false discovery rate based on our simulations (<0.5%, see Methods; Supporting information 4). For simplicity, we report results for the more stringent data set, but note the few cases in which analysis of the two data sets yield different results.

We see several surprising patterns when comparing regions derived from hybridization to the genomic background. First, these regions are on average 1.5× more divergent between *X. montezumae* and *X. cortezi* than other genomic regions ($P < 0.001$ by simulation; Fig. 5). We observe this pattern even when regions for which we have lower power (i.e. expected power ≤0.9; see Supporting information 4) are excluded from the null data sets (Fig. S13, Supporting information), suggesting that this difference is not driven by our power to detect hybridization signal in low-divergence regions.

The high divergence of introgressed regions might suggest that only neutrally evolving hybridization-derived regions such as noncoding and nonregulatory sequences have persisted in the *X. nezahualcoyotl* genome. However, gene density is also significantly higher in hybridization-derived regions than in null data sets ($P = 0.002$ by simulation, Fig. 5; Fig. S14, Supporting information). Although gene density is higher, the proportion of each region that is coding does not differ between null and hybridization-derived regions (Fig. 5, Fig. S14, Supporting information).

Hybridization-derived regions have a significantly higher number of synonymous and nonsynonymous substitutions than the genomic background ($P < 0.001$ by simulation for dN, dS, N and S; Table 2). However, the stringent data set does not show evidence that these regions are less constrained (or positively selected) compared with the genomic background based on dN/dS ($P = 0.34$ by simulation). In contrast, the full data set has a significantly higher dN/dS ratio than the genomic background ($P = 0.008$ by simulation, Table 2). By subsampling the full data set, we conclude that this difference is likely due to a lack of power (i.e. ~30%) and that the overall pattern does suggest lower levels of constraint on proteins in hybridization regions. Further, the same patterns are observed in an independent comparison of the same regions in *X. maculatus* and *X. hellerii* (Supporting information 10), suggesting that reduced constraint is a general property of these proteins.

To evaluate patterns of hybridization in relation to constrains on both coding and noncoding DNA, we used the program PHASTCONS to identify the most conserved elements in the *Xiphophorus* genome based on comparisons to five other fish genomes. This analysis identified 472 163 conserved elements distributed across all 24 linkage groups, on average 55 base pairs in length (median 36). We find that highly conserved elements (conservation log-likelihood score >56; $N = 118\,922$ see Methods) are significantly under-represented in

Table 2 Summary of patterns of divergence between *X. montezumae* and *X. cortezi* in regions derived from hybridization versus the genomic background. See Methods for details on analyses. Given the phylogenetic placement of *X. nezahualcoyotl*, divergence between *X. montezumae* and *X. cortezi* can be used as a proxy for divergence between the *X. cortezi* and *X. nezahualcoyotl* lineages at the time of hybridization for regions that introgressed from *X. cortezi* into *X. nezahualcoyotl*. For the null data sets, the 95% confidence intervals are shown

Data set	Mean D_{xy}	Mean dN	Mean dS	Mean dN/dS
All hybridization-derived regions (N = 2282)	0.011	0.0027	0.011	0.088
Null (N = 2282)	0.00860–0.00864	0.0021–0.0022	0.0092–0.0099	0.080–0.087
Length filtered hybridization-derived regions (N = 452)	0.010	0.0026	0.011	0.085
Null (N = 452)	0.0079–0.0080	0.0020–0.0023	0.0090–0.010	0.076–0.090

D_{xy} – pairwise divergence, dN – average rate of nonsynonymous substitutions, dS – average rate of synonymous substitutions, dN/dS – sum of nonsynonymous substitution rates divided by the sum of synonymous substitution rates.

hybridization-derived regions ($P < 0.001$ by simulation; Fig. 5, Fig. S14, Supporting information). Consistent with this result, a significantly lower proportion of highly conserved base pairs are found in hybridization-derived regions ($P < 0.001$ by simulation). On average, species tree regions have 15% more conserved bases than hybridization-derived regions (Fig. 5). This effect can also be seen by comparing the two tails of the distribution of log-likelihood scores of the conserved elements identified by phastCons. The most conserved elements (based on log-likelihood score) are the least likely to be hybridization-derived, while the least conserved are the most likely to be hybridization-derived (Fig. S15, Supporting information).

Verification of results for regions with high confidence in the direction of introgression

Our analyses suggest that the majority of hybridization-derived regions we analyse are regions that introgressed from the *X. cortezi* lineage into the *X. nezahualcoyotl* lineage (76%). However, because our focal data sets likely represent a mixture of regions from both introgression directions, we repeat the analyses of divergence, coding density and conservation detailed above on 222 regions identified by D_{FOIL} as introgressed from *X. cortezi* into *X. nezahualcoyotl*. Results based on this smaller data set are identical to those observed in the stringent data set (see Supporting information 11).

Gene ontology and pathway analysis results

We investigated whether genes derived from hybridization were more likely to belong to particular functional categories. GO analysis identified 132 significantly enriched biological process GO terms in the stringent data set and 213 significantly enriched biological process GO terms in the full data set at $P < 0.05$ (Tables S5 and S6, Supporting information). These included categories related to immune response, which have been previously implicated in adaptive introgression (e.g. Abi-Rached *et al.* 2011; Mendez *et al.* 2013; Racimo *et al.* 2015). However, analysis of 10 null data sets matching the stringent and full data sets in region sizes and gene number (see Methods) shows that this number of enriched categories is expected by chance (stringent: 76–262 terms, 30% >132 terms; full: 93–261 terms, 30% >213 terms). This same result is observed whether lower P-value thresholds are used, and in no case was the most significant GO term in the focal data sets more significant than the most significant GO term in the 10 null data sets. Qualitatively similar results were observed for cellular component and molecular function GO enrichment analyses. These results suggest that although there is enrichment in particular GO categories in hybridization-derived regions, these patterns cannot be distinguished from expectations by chance. Interestingly, one of the most significant GO terms in analysis of cellular component terms, the cell periphery (Table S6, Supporting information), was repeatedly detected in null data sets as a significantly enriched term. This GO term has been previously reported as enriched in hybridization-derived regions in mice (Janousek *et al.* 2015; we note that these authors also evaluated significance using null data sets).

Results of pathway analyses in DAVID mirrored GO results. One and 10 significantly overrepresented pathways were identified in the stringent and full data sets, respectively; none of which exceeded the number of enriched pathways found in the null data sets (stringent: 3–15 pathways, 100% >1 pathway; full: 9–21 pathways, 80% >10 pathways), or were more significantly enriched than the most significant pathways identified in the null data sets. Overall, our results suggest that caution should be used when applying these types of analyses as the clustering of functionally similar genes in the genome can cause a higher than expected false-positive rate.

Discussion

As many species have genomes with hybrid ancestry, it is important to understand the different factors shaping patterns of hybrid ancestry in the genome and its evolutionary implications. Despite a few well-studied cases, we know little in general about how selection and drift act after hybridization to influence where hybrid ancestry persists in a genome. Studies in several taxa have begun to shed light on these underlying processes, suggesting an important role for epistasis and constraint, but a much broader survey is needed to begin to illuminate the rules governing patterns of hybrid ancestry in eukaryotic genomes. Our investigation of hybrid ancestry in the swordtail fish *X. nezahualcoyotl* shows that hybridization-derived regions have largely fixed, raising questions about whether this pattern is likely to occur in the absence of selection. Furthermore, we find that hybridization-derived regions are less constrained on average than the rest of the genome, supporting a role for selection in influencing what regions of the genome are able to move between species.

Hybrid ancestry and stabilization of the X. nezahualcoyotl genome

Our phylogenetic analyses demonstrate that *X. nezahualcoyotl* is most closely related to *X. montezumae*, but 8% of the genome shows evidence of admixture. Based on our finding that 76% of hybridization-derived regions introgressed from the *X. cortezi* lineage into *X. nezahualcoyotl*, we estimate that 6% of the *X. nezahualcoyotl* genome is derived from admixture with *X. cortezi*. This estimate is notably smaller than our previous estimate using RNAseq data (see analyses and discussion in Supporting information 12). Among the regions derived from hybridization are over 2000 protein coding genes, demonstrating that the functional importance of hybridization is potentially substantial. Based on divergence between hybridization-derived haplotypes in *X. nezahualcoyotol* and these same regions in *X. cortezi*, hybridization is likely to have been relatively ancient (Fig. 4; Fig. S11, Supporting information).

Several lines of evidence suggest that the genome of *X. nezahualcoyotl* has stabilized following admixture with the *X. cortezi* lineage and has likely been stable for many generations. We observe nearly complete concordance (>99%) between those regions unambiguously inferred to be hybridization-derived in the two *X. nezahualcoyotl* samples, which is inconsistent with a large number of *X. cortezi* regions segregating in either population. Although heterozygous hybridization-derived regions are not detected by PhyloNet-HMM, if many regions were heterozygous, this would result in lower overlap in hybridization-derived regions between the two samples. In addition, heterozygosity is very low overall in the northern *X. nezahualcoyotl* sample (Gallitos, Table 1), implying that few hybridization-derived sites could be segregating in this population. Furthermore, only a small fraction of ambiguous base pairs show strong evidence for hybrid ancestry (Fig. S10, Supporting information).

These patterns, in the light of the geographical location of the sampled populations (Fig. 1), suggest that *X. nezahualcoyotl* may have hybridized with the *X. cortezi* lineage in their southern range of overlap but that the genome largely stabilized for hybrid ancestry before *X. nezahualcoyotl* established its current range (Fig. 4). Sampling more individuals from a wider geographical distribution may strengthen this claim. Present day gene flow between *X. nezahualcoyotl* and *X. cortezi* is limited by deep and rapid water flow in the Río Tampaón and it is unknown how long this geographical structure has been in place. The length distribution of hybridization-derived haplotypes is consistent with ~2500 generations of recombination preceding stabilization, but divergence between hybridization-derived haplotypes in *X. nezahualcoyotl* and *X. cortezi* suggests a more ancient hybridization event (Fig. 4). However, taken together these data demonstrate that hybridization did not occur in the recent past and that gene flow between *X. nezahualcoyotl* and *X. cortezi* is unlikely to be ongoing (Fig. 4; Fig. S11, Supporting information).

These results, in the context of previous studies, raise many questions about the processes driving genomic stabilization after hybridization. The stabilization of hybridization-derived regions requires either long time periods, small population sizes or selection on hybridization-derived regions. Results in hybrid fungal lineages and hybrid *Helianthus* sunflowers have supported rapid genomic stabilization following hybridization (Buerkle & Rieseberg 2008; Stukenbrock *et al.* 2012), and studies in *Helianthus* have linked selection to genomic reorganization during this stabilization process (Rieseberg *et al.* 1996). In contrast, to our knowledge, there are no documented cases of rapid genomic stabilization in animals. For example, in the case of human–Neanderthal admixture, which occurred ~2000 generations ago (Sankararaman *et al.* 2012), much of the genome has not stabilized and few regions are at high frequency for Neanderthal ancestry (Sankararaman *et al.* 2014).

What processes have driven genomic stabilization in *X. nezahualcoyotl*? Although significant time has passed since hybridization given levels of divergence between *X. nezahualcoyotl* and *X. cortezi* at hybridization-derived

regions, stabilization of hybridization-derived regions in *X. nezahualcoyotl* may have occurred over a much shorter time period (~2500 generations, see Supporting information 8). Our recent work in two other swordtail species with slightly lower sequence divergence than *X. montezumae–X. cortezi* identified hundreds of genetic incompatibilities separating parent species (Schumer *et al.* 2014), suggesting that initial selection on hybrids in *Xiphophorus* can be strong. Selection against incompatibilities could be one mechanism for driving rapid fixation of hybridization-derived regions (Schumer *et al.* 2015b). Given current levels of polymorphism in one of the *X. nezahualcoyotl* samples, we can rule out a recent bottleneck as the cause of substantial fixation of *X. cortezi* ancestry (Fig. 4). However, simulations reveal that we cannot rule out stabilization caused by a strong bottleneck associated with hybridization and recovery (Fig. S16, Supporting information 9). Similar diversity levels in hybridization-derived regions and the rest of the genome could also support a demographic explanation for stabilization (Fig. 4). Definitively addressing this question will require a detailed investigation of the demographic history of *X. nezahualcoyotl*.

Functional significance of hybridization-derived regions

Another approach to investigating the possible role of selection in shaping hybrid ancestry in *X. nezahualcoyotl* is examining patterns in regions fixed for *X. cortezi*-derived haplotypes. We consider two major hypotheses in our functional analysis of regions derived from hybridization. The first is that regions of the genome with high divergence between species, particularly nonsynonymous divergence, will be less likely to introgress, because they might be more likely to harbour substitutions that will either be universally detrimental (and fixed by chance in one species) or detrimental in only one species' genomic background (i.e. subject to negative epistatic interactions). The second hypothesis is that regions with high substitution rates are less likely to be functionally important and thus will be enriched for hybrid ancestry compared with the rest of the genome after selection. As these hypotheses make nearly opposite predictions about divergence and constraint in hybridization-derived regions, we can ask whether either is more consistent with the major patterns in our data.

We establish that regions derived from hybridization have greater sequence divergence between *X. montezumae* and *X. cortezi* than other genomic regions, suggesting greater divergence between *X. cortezi* and *X. nezahualcoyotl* at these regions at the time of hybridization. One possible interpretation of these results is that divergence in these regions is driven by an excess of nonfunctional DNA that is evolving neutrally and thus reflects substitutions that are unlikely to result in negative epistasis. We may expect that these regions will have fewer functional elements than randomly sampled genomic regions. Instead, we see an excess of protein coding genes in these regions. However, we observe no difference in the proportion of coding base pairs between regions in the hybridization-derived and null data sets (Fig. 5). We also find that genes in hybridization-derived regions show an excess of both synonymous and nonsynonymous substitutions, demonstrating that the divergence signal in these regions is not being driven solely by noncoding substitutions.

This result contrasts sharply with findings in other studies which have suggested that divergent regions are less likely to introgress between species (Vernot & Akey 2014; Janousek *et al.* 2015). Previous studies have suggested that genomic regions with high substitution rates, particularly at nonsynonymous sites, may be more likely to be associated with negative epistasis and incompatibilities between species (Orr 1995; Orr & Turelli 2001). This idea has been invoked to explain low divergence in introgressed regions in previous studies (Janousek *et al.* 2015). However, it is also possible that high levels of observed nonsynonymous divergence reflect weaker purifying selection. This would suggest that these genes are less functionally essential and thus more likely to introgress between species than highly conserved regions, consistent with the pattern that we observe. We identified conserved elements in the *Xiphophorus* genome by determining which genomic regions have been highly conserved over ~250 million years of fish evolution. We see strong evidence that the most constrained regions of the genome are less likely to be derived from hybridization (Fig. 5; Fig. S14, Supporting information). Furthermore, this pattern is amplified in the most strongly conserved regions of the genome (Fig. S15, Supporting information). However, we note that although this pattern is highly significant, hybridization-derived regions have only ~10–15% fewer conserved bases than other regions (Fig. 5, Fig. S14, Supporting information), suggesting that other mechanisms (such as drift) could be playing a large role in fixation.

Theoretical work suggests that introgression is more likely to occur at more neutral genomic regions (Barton 1979), consistent with our results. Given this pattern, we can speculate that selection has on average acted to reduce hybrid ancestry in the most functionally important regions of the genome. This result is broadly consistent with findings in other systems that suggest that selection often constrains hybrid ancestry at functionally important genomic regions (Payseur *et al.* 2004;

Sankararaman *et al.* 2014; Schumer *et al.* 2014). We also note higher average dN/dS for proteins in hybridization-derived regions (full data set). This observation is consistent with the results of our conservation analysis demonstrating an overall pattern of lower constraint on regions derived from hybridization.

Patterns on the putative X chromosome

The X chromosome has been shown to play a disproportionate role in reproductive isolation in many species, likely as a result of more rapid evolution ('faster-X'), Haldane's rule or meiotic drive (Presgraves 2008). We find that the putative X chromosome (Schartl *et al.* 2013) has both lower levels of introgression than average and lower levels of coding introgression than average in *X. nezahualcoyotl*, but is not a clear outlier compared with other chromosomes (Fig. S17, Supporting information). This finding is consistent with previous results on the genomic distribution of hybrid incompatibility regions in other *Xiphophorus* species (Schumer *et al.* 2014). The lack of a clear role of the X chromosome in reproductive isolation in *Xiphophorus* may reflect the young age of the sex chromosome (Schultheis 2004), diversity in sex chromosomes in the genus (Schartl *et al.* 2009) or the influence of autosomal factors on sex determination (Kallman 1984).

Conclusions

Our results demonstrate that the genome of *X. nezahualcoyotl* exhibits substantial hybrid ancestry that has largely stabilized following an ancient hybridization event. Our analysis suggests that selection has played a role in shaping what regions of the genome are derived from hybridization. Notably, we see the opposite results of several previous studies that observed low divergence at introgressed regions, highlighting the fact that there is still much that we do not know about how genomes stabilize following pulses of hybridization. In particular, we know little about whether genomic stabilization after hybridization is typically driven by passive processes such as genetic drift or active processes such as selection. If fixation of hybrid ancestry is normally non-neutral, how important are different mechanisms such as selection on genetic incompatibilities, adaptive introgression, or purging of hybrid ancestry in functionally important regions of the genome? Contrasting our results to others highlights the potentially diverse genomic outcomes of hybridization and emphasizes the need for much more research to characterize the 'rules' governing posthybridization genome evolution.

Acknowledgements

We thank the federal government of Mexico for permission to collect fish under Mexican federal collector's permit to Scott Monks (FAUT-217) and a scientific collecting permit to Guillermina Alcaraz (PPF/DGOPA-173/14). We thank Liz Marchio and Kyle Piller for providing individuals collected from Las Crucitas. We thank Kevin Liu for additional information on running PhyloNet-HMM and James Pease for help running and interpreting D_{FOIL}. This project was supported by an NSF GRFP DGE0646086, NSF DDIG DEB-1405232 and grant from the American Livebearer's Association to M.S.

References

Abi-Rached L, Jobin MJ, Kulkarni S *et al.* (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, **334**, 89–94.

Amores A, Catchen J, Nanda I *et al.* (2014) A RAD-tag genetic map for the platyfish (*Xiphophorus maculatus*) reveals mechanisms of karyotype evolution among Teleost fish. *Genetics*, **197**, 625–641.

Andolfatto P, Davison D, Erezyilmaz D *et al.* (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Atz JW (1962) Effects of hybridization on pigmentation in fishes of the genus *Xiphophorus*. *Zoologica*, **47**, 153–181.

Barton NH (1979) Gene flow past a cline. *Heredity*, **43**, 333–339.

Blanchette M, Kent WJ, Riemer C *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, **14**, 708–715.

Buerkle CA, Rieseberg LH (2008) The rate of genome stabilization in homoploid hybrid species. *Evolution*, **62**, 266–275.

Cahill JA, Stirling I, Kistler L *et al.* (2015) Genomic evidence of geographically widespread effect of gene flow from polar bears into brown bears. *Molecular Ecology*, **24**, 1205–1217.

Carling MD, Brumfield RT (2008) Haldane's rule in an avian system: using cline theory and divergence population genetics to test for differential introgression of mitochondrial, autosomal, and sex-linked loci across the Passerina bunting hybrid zone. *Evolution*, **62**, 2600–2615.

Cui R, Schumer M, Kruesi K *et al.* (2013) Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes. *Evolution*, **67**, 2166–2179.

Culumber ZW, Fisher HS, Tobler M *et al.* (2011) Replicated hybrid zones of *Xiphophorus* swordtails along an elevational gradient. *Molecular Ecology*, **20**, 342–356.

Earl D, Nguyen N, Hickey G *et al.* (2014) Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Research*, **24**, 2077–2089.

Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.

Gavrilets S (1997) Hybrid zones with Dobzhansky-type epistatic selection. *Evolution*, **51**, 1027–1035.

Geraldes A, Ferrand N, Nachman MW (2006) Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics*, **173**, 919–933.

Gerke J, Edwards J, Guill K, Ross-Ibarra J, McMullen M (2015) The genomic impacts of drift and selection for hybrid performance in maize. *Genetics*, **202**, 1–11.

Good JM, Giger T, Dean MD, Nachman MW (2010) Widespread over-expression of the X chromosome in sterile F-1 hybrid mice. *Plos Genetics*, **6**, e1001148.

Gordon M (1937) Heritable color variations in the Mexican swordtail-fish – Aquarium species as the drosophila of fish genetics. *Journal of Heredity*, **28**, 223–230.

Gravel S (2012) Population genetics models of local ancestry. *Genetics*, **191**, 607–619.

Green RE, Krause J, Briggs AW *et al.* (2010) A draft sequence of the neandertal genome. *Science*, **328**, 710–722.

Hamilton JA, Lexer C, Aitken SN (2013) Differential introgression reveals candidate genes for selection across a spruce (*Picea sitchensis x P. glauca*) hybrid zone. *New Phytologist*, **197**, 927–938.

Heliconius Genome C (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**, 44–57.

Janousek V, Munclinger P, Wang L, Teeter KC, Tucker PK (2015) Functional organization of the genome may shape the species boundary in the house mouse. *Molecular Biology and Evolution*, **32**, 1208–1220.

Jin W, Li R, Zhou Y, Xu S (2014) Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping. *European Journal of Human Genetics*, **22**, 930–937.

Jones JC, Perez-Sato J-A, Meyer A (2012) A phylogeographic investigation of the hybrid origin of a species of swordtail fish from Mexico. *Molecular Ecology*, **21**, 2692–2712.

Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013) The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology*, **22**, 2986–3001.

Juric I, Aeschbacher S, Coop G (2015) The strength of selection against neanderthal introgression. *bioRxiv*, 1–24, doi:10.1101/030148.

Kallman KD (1984) A new look at sex determination in poeciliid fishes. In: *Evolutionary Genetics of Fishes* (ed. Turner BJ), pp. 95–171. Plenum Press, New York.

Kang J, Schartl M, Walter R, Meyer A (2013) Comprehensive phylogenetic analysis of all species of swordtails and platies (Pisces: Genus Xiphophorus) uncovers a hybrid origin of a swordtail fish, *Xiphophorus monticolus*, and demonstrates that the sexually selected sword originated in the ancestral lineage of the genus, but was lost again secondarily. *BMC Evolutionary Biology*, **13**, 1–19.

Larson EL, Andres JA, Bogdanowicz SM, Harrison RG (2013) Differential introgression in a mosaic hybrid zone reveals candidate barrier genes. *Evolution*, **67**, 3653–3661.

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

Li J-W, Robison K, Martin M *et al.* (2012) The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. *Nucleic Acids Research*, **40**, D1313–D1317.

Liu KJ, Dai J, Truong K *et al.* (2014) An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *Plos Computational Biology*, **10**, e1003649.

Maddison WP (1997) Gene trees in species trees. *Systematic Biology*, **46**, 523–536.

Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter gene introgression between species. *Evolution*, **55**, 1325–1335.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Mendez FL, Watkins JC, Hammer MF (2013) Neandertal origin of genetic variation at the cluster of OAS immunity genes. *Molecular Biology and Evolution*, **30**, 798–801.

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

Orr HA (1995) The population genetics of speciation – the evolution of hybrid incompatibilities. *Genetics*, **139**, 1805–1813.

Orr HA, Turelli M (2001) The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. *Evolution*, **55**, 1085–1094.

Payseur BA, Nachman MW (2005) The genomics of speciation: investigating the molecular correlates of X chromosome introgression across the hybrid zone between *Mus domesticus* and *Mus musculus*. *Biological Journal of the Linnean Society*, **84**, 523–534.

Payseur BA, Krenz JG, Nachman MW (2004) Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. *Evolution*, **58**, 2064–2078.

Pease JB, Hahn MW (2015) Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology*, **64**, 651–662.

Presgraves DC (2008) Sex chromosomes and speciation in *Drosophila*. *Trends in Genetics*, **24**, 336–343.

Quail MA, Swerdlow H, Turner DJ (2009) Improved protocols for the Illumina genome analyzer sequencing system. *Current Protocols in Human Genetics*, **62**, 1–27 Chapter 18.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E (2015) Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, **16**, 359–371.

Rieseberg LH, Sinervo B, Linder CR, Ungerer MC, Arias DM (1996) Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. *Science*, **272**, 741–745.

Sankararaman S, Patterson N, Li H, Paeaebo S, Reich D (2012) The date of interbreeding between neandertals and modern humans. *Plos Genetics*, **8**, e1002947.

Sankararaman S, Mallick S, Dannemann M *et al.* (2014) The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, **507**, 354–357.

Schartl M (2004) Sex chromosome evolution in non-mammalian vertebrates. *Current Opinion in Genetics and Development*, **14**, 634–641.

Schultheis C, Bohne A, Schartl M, Volff JN, Galiana-Arnoux D (2009) Sex determination diversity and sex chromosome evolution in Poeciliid fish. *Sexual Development*, **3**, 68–77.

Schartl M, Walter RB, Shen Y *et al.* (2013) The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nature Genetics*, **45**, 567–572.

Schmidt HA (2009) Testing tree topologies. *Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and*

Hypothesis Testing, 2nd edn, pp. 381–404. Cambridge University Press, Cambridge, UK.
Schumer M, Cui R, Boussau B *et al.* (2012) An evaluation of the hybrid speciation hypothesis for *Xiphophorus clemenciae* based on whole genome sequences. *Evolution*, **64**, 1155–1168.
Schumer M, Cui R, Powell D *et al.* (2014) High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species.*eLife*, **3**, 1–21.
Schumer M, Cui R, Rosenthal G, Andolfatto P (2015a) simMSG: an experimental design tool for high-throughput genotyping of hybrids. *Molecular Ecology Resources*, **16**, 183–192.
Schumer M, Cui R, Rosenthal GG, Andolfatto P (2015b) Reproductive isolation of hybrid populations driven by genetic incompatibilities. *Plos Genetics*, **11**, 5041.
Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, **51**, 492–508.
Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, **17**, 1246–1247.
Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, **21**, 468–488.
Siepel A, Bejerano G, Pedersen JS *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **15**, 1034–1050.
Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. *Trends in Ecology & Evolution*, **14**, 348–352.
Song KM, Lu P, Tang KL, Osborn TC (1995) Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 7719–7723.
Song Y, Endepols S, Klemann N *et al.* (2011) Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, **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 EH, Christiansen FB, Hansen TT, Dutheil JY, Schierup MH (2012) Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 10954–10959.
Tayale A, Parisod C (2013) Natural pathways to polyploidy in plants and consequences for genome reorganization. *Cytogenetic and Genome Research*, **140**, 79–96.
Teeter KC, Payseur BA, Harris LW *et al.* (2008) Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Research*, **18**, 67–76.
Turner LM, Harr B (2014) Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. *eLife*, **3**, 1–25.
Turner LM, White MA, Tautz D, Payseur BA (2014) Genomic networks of hybrid sterility. *PLoS Genetics*, **10**, e1004162.
Vernot B, Akey JM (2014) Resurrecting surviving neandertal lineages from modern human genomes. *Science*, **343**, 1017–1021.
White MA, Steffy B, Wiltshire T, Payseur BA (2011) Genetic dissection of a key reproductive barrier between nascent species of house mice. *Genetics*, **189**, U289–U988.
Whitney KD, Randell RA, Rieseberg LH (2010) Adaptive introgression of abiotic tolerance traits in the sunflower Helianthus annuus. *New Phytologist*, **187**, 230–239.
Yang ZH (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computational Applied Bioscience*, **13**, 555–556.
Yu Y, Degnan JH, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *Plos Genetics*, **8**, 456–465.

M.S. and R.C. conceived of the experiments. All authors designed experiments. M.S., R.C. and D.P. performed data analysis. D.P. collected fish samples. P.A. and G.G.R. supervised the research. All authors wrote the manuscript.

Data accessibility

Sequence data has been deposited on the NCBI SRA (SRX1518311, SRX1518380, SRX1518263, SRX1518028). Genome alignments and scripts are available on Dryad (doi:10.5061/dryad.tm47d). All steps required to reproduce major analyses are included in the Appendix S2 (Supporting information).

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Supporting text including additional analyses and simulations.

Appendix S2 Example command line used for mapping, analysis and simulations.

Fig. S1 Sensitivity to real variants in align-to-reference simulations.

Fig. S2 Error rates for align-to-reference approach.

Fig. S3 No effect of errors in genotype calling on hybridization inference.

Fig. S4 Distribution of missing data in each sample.

Fig. S5 Schematic showing sequence generation for testing of PhyloNet-HMM.

Fig. S6 Example of hybridization simulation results from PhyloNet-HMM.

Fig. S7 PhyloNet-HMM tends to underestimate the size of hybridization-derived regions.

Fig. S8 Different evolutionary scenarios potentially leading to present day *X. nezahualcoyotl* genome.

Fig. S9 Concordance between AU results and PhyloNet-HMM results for linkage group 1.

Fig. S10 Signature of introgression based on the D-statistic in the whole genome and hybridization-derived regions (left) and in regions not called confidently by PhyloNet-HMM (right).

Fig. S11 Estimates of nucleotide diversity (π) and divergence (D_{xy}) in regions with different evolutionary histories.

Fig. S12 Comparison of the length of hybridization-derived ancestry tracts to a random exponential distribution.

Fig. S13 Divergence in hybridization-derived regions when low power windows are excluded.

Fig. S14 Differences between hybridization derived genomic regions and the genomic background in the unfiltered dataset.

Fig. S15 The most strongly conserved genomic regions are the least likely to be hybridization derived.

Fig. S16 Proportion of simulated regions fixed as a function of population size after 2500 generations.

Fig. S17 Distribution of hybridization-derived regions by chromosome.

Fig. S18 Assessment of potential biases associated with our approach to estimating the time to genomic stabilization.

Table S1 Mapping and coverage statistics for focal species.

Table S2 Percent average pairwise sequence divergence between samples.

Table S3 Proportion of each linkage group inferred to be hybridization-derived.

Table S4 Regions with distinct patterns in the two X. nezahualcoyotl samples.

Table S5 Significantly enriched GO categories in the size-filtered dataset for biological processes, molecular function, and cellular component annotations.

Table S6 Significantly enriched GO categories in the size-filtered dataset for biological processes, molecular function, and cellular component annotations.