Title: Genome evolution is surprisingly predictable after initial hybridization

Authors: Quinn K. Langdon[1,2,3*], Jeffrey S. Groh[4], Stepfanie M. Aguillon[1,2,5], Daniel L. Powell[1,2], Theresa Gunn[1,2], Cheyenne Payne[1,2], John J. Baczenas[1], Alex Donny[1,2], Tristram O. Dodge[1,2], Kang Du[6], Manfred Schartl[6,7], Oscar Ríos-Cárdenas[8], Carla Gutierrez-Rodríguez[8], Molly Morris[9], Molly Schumer[1,2,10*]

[1]Department of Biology, Stanford University
[2]Centro de Investigaciones Científicas de las Huastecas "Aguazarca", A.C.
[3]Gladstone Institute of Virology, Gladstone Institutes, San Francisco, California
[4]Center for Population Biology and Department of Evolution and Ecology, University of California, Davis
[5]Department of Ecology and Evolutionary Biology, University of California, Los Angeles
[6]Xiphophorus Genetic Stock Center, Texas State University San Marcos
[7]Developmental Biochemistry, Biocenter, University of Würzburg
[8]Red de Biología Evolutiva, Instituto de Ecología, A.C.
[9]Department of Biology, Ohio University
[10]Freeman Hrabowski Fellow, Howard Hughes Medical Institute

*Correspondence to: qlangdon@gmail.com and schumer@stanford.edu

Abstract

Over the past two decades, evolutionary biologists have come to appreciate that hybridization, or genetic exchange between distinct lineages, is remarkably common – not just in particular lineages but in taxonomic groups across the tree of life. As a result, the genomes of many modern species harbor regions inherited from related species. This observation has raised fundamental questions about the degree to which the genomic outcomes of hybridization are repeatable and the degree to which natural selection drives such repeatability. However, a lack of appropriate systems to answer these questions has limited empirical progress in this area. Here, we leverage independently formed hybrid populations between the swordtail fish *Xiphophorus birchmanni* and *X. cortezi* to address this fundamental question. We find that local ancestry in one hybrid population is remarkably predictive of local ancestry in another, demographically independent hybrid population. Applying newly developed methods, we can attribute much of this repeatability to strong selection in the earliest generations after initial hybridization. We complement these analyses with time-series data that demonstrates that ancestry at regions under selection has remained stable over the past ~40 generations of evolution. Finally, we compare our results to the well-studied *X. birchmanni*×*X. malinche* hybrid populations and conclude that deeper evolutionary divergence has resulted in stronger selection and higher repeatability in patterns of local ancestry in hybrids between *X. birchmanni* and *X. cortezi*.

Introduction

Hybridization has made substantial contributions to the genomes of species across the tree of life. Dozens of studies over the past two decades have documented pervasive genetic exchange between closely related species within all major eukaryotic groups [1–8]. Hybridization has even played an important role in the evolutionary history of our own species [9–11] and that of our close relatives [12,13]. Because we now know that genetic exchange between species is pervasive, unraveling the genetic and evolutionary impacts of hybridization is a fundamental part of understanding the genomes of modern species. Moreover, characterizing the genomic consequences of hybridization promises to directly inform our understanding of the genetic changes that lead to divergence between species.

Modern genomic approaches to studying hybridization are often based on inference of local ancestry, or the ancestral source population from which a haplotype was derived, using genomic similarity to contemporary reference populations. With these approaches, researchers have moved from documenting evidence of hybridization in the genome as a whole to characterizing patterns of local variation in ancestry along the genome. Research into the history of genetic exchange between modern humans and our extinct relatives, the Neanderthals and Denisovans, was among the first to rigorously evaluate where in the genome ancestry from other lineages has been retained and where it has been lost [10,14–17]. This question has since been tackled in several species groups, including swordtail fish [18,19], *Saccharomyces* yeast [20], monkeyflowers [3,21,22], *Drosophila* [23], *Formica* ants [24], honey bees [5], *Heliconius* butterflies [25,26], and baboons [27]. Although the organisms in which these questions have been studied are diverse, some unifying observations have emerged from this work, hinting at shared principles that impact the predictability of genome evolution after hybridization. First, in most species studied to date, haplotypes that originate from the ‘minor’ parent species, or the species from which hybrids derive less of their genome, are inferred to be on average deleterious (due to a number of possible mechanisms of selection; see below, [15,28,29]). Second, genome architecture seems to play a repeatable role in the purging of minor parent ancestry following hybridization. Researchers have consistently found that regions of the genome with low rates of recombination have lower levels of minor parent ancestry, presumably because long introgressed haplotypes are more likely to contain multiple linked deleterious variants and thus be purged by selection more rapidly [3,24,25,27,30,31]. Theoretical studies have demonstrated that these dynamics are expected from first principles [32]. Similarly, researchers have found that regions of the genome especially dense in functional basepairs, including coding, conserved, and enhancer regions, are often depleted in minor parent ancestry [10,15,17,28,30] (but see [33] for discussion of the challenges of these analyses). Together, these observations point to a shared role of genome organization in the patterning of ancestry in the genome after hybridization.

These patterns highlight shared factors that drive genome evolution after hybridization across diverse taxa. However, it is still unclear whether selection drives repeatable patterns of local ancestry in replicated hybridization events between the same species, after accounting for these factors. From first principles, we might expect more repeatability in local ancestry across replicated hybrid populations in scenarios when more loci are under selection in hybrids ([15]; and the sites under selection are shared) and when selection is strong relative to genetic drift [30,32]. The specific mechanisms of selection on hybrids are also likely to play an important role in the degree to which we expect repeatability in local ancestry in replicate hybrid populations. In cases where selection on hybrids is largely driven by selection on negative epistatic interactions between substitutions that have arisen in the parental species' genomes (so-called "Dobzhansky-Muller" hybrid incompatibilities; but see [34]) or directional selection against one ancestry state (e.g. due to an excess of deleterious mutations that have accumulated along that lineage; [15,24,29,32]), we might predict that selection will drive repeatable ancestry patterns around selected sites. Moreover, theory and available empirical data predicts that the number of hybrid incompatibilities will increase non-linearly with divergence between lineages [35–40], such that hybrid incompatibilities may play a larger role in the genome evolution of hybrids formed between distant relatives. By contrast, in species where selection against hybrids is largely dependent on the ecological environment [41,42], we might predict that selection will drive distinct patterns of local ancestry in distinct environments. The demographic history of the hybrid population itself is also crucial for interpreting signals of repeatability, since variables such as the time since admixture determine the scale of ancestry variation along a chromosome and the accumulated effects of genetic drift. Importantly however, temporally-localized effects of selection can leave lasting impacts on ancestry variation, suggesting that ancestry patterns studied even long after admixture can be informative about the early stages of selection on hybrids [33,43].

Beyond the diverse biological factors at play, progress in understanding the repeatability of replicate hybridization events has been limited by the fact that only a handful of empirical studies have tackled this question. This is in part due to a lack of appropriate systems to test these questions (e.g. those with truly independent hybridization events) and in part due to the difficulty of excluding technical factors impacting the accuracy of ancestry inference that could be misinterpreted as biological signal. We focus our discussion here on studies that directly infer local ancestry states along the genome because of their precision and improved ability to distinguish hybridization from other biological processes (e.g. incomplete lineage sorting, background selection; [44–46]). However, we note that other approaches have provided important insights into the repeatability of genetic and phenotypic evolution after hybridization [3,26,39,47–50].

Some of the earliest studies to address questions about repeatability of local ancestry patterns asked whether there were shared deserts of archaic ancestry (i.e. Neanderthal and Denisovan ancestry) in the human genome [10,14]. These studies identified concordant patterns in the locations of deserts of archaic ancestry and the types of regions that harbor higher levels of archaic ancestry [10,14]. However, interpretation of these results is complicated by the challenges of distinguishing between Neanderthal and Denisovan ancestry [51], and other technical considerations [16]. Outside of hominins, three studies have explicitly inferred local ancestry and used it to evaluate the repeatability of genome evolution in replicated hybridization events. In *Drosophila*, Matute et al. (2019) showed that experimental hybrid populations generated between *Drosophila* species showed repeatable patterns of purging of minor parent ancestry [52]. In hybrid swarms generated between these species, ancestry from one parental species was consistently purged, and the regions where minor parent ancestry tracts were retained showed some level of repeatability in replicate populations. In replicate natural populations of hybrid ants that have evolved independently for tens of generations, researchers found remarkably high repeatability in local ancestry patterns across three hybrid populations, driven in part by selection against deleterious load inherited from one of the parental species [24]. Past work from our group asked about repeatability in patterns of minor parent ancestry in naturally occurring *Xiphophorus birchmanni* × *X. malinche* populations that formed independently in different river systems [53]. We found moderate predictability in local ancestry patterns between replicate *X. birchmanni* × *X. malinche* populations [54,55]. We also compared patterns of local ancestry between *X. birchmanni* × *X. malinche* hybrid populations to a hybrid population of a different type, formed between *X. birchmanni* and its more distant relative, *X. cortezi* [53], and identified weak but significant correlations in local ancestry between hybrid population types.

Here, we identify a new independently formed hybrid population between *X. birchmanni* and *X. cortezi* (Fig. 1), allowing us to ask questions about how repeatability of genome evolution scales with increasing genetic divergence between hybridizing species. We observe an extraordinary level of repeatability in local ancestry patterns across independently formed *X. birchmanni* × *X. cortezi* hybrid populations, consistent with remarkably strong selection on hybrids. We find that some of this repeatability in local ancestry is linked to large minor parent ancestry "deserts" that coincide with known hybrid incompatibilities. Using wavelet analysis [32], we find the overall correlation in ancestry between *X. birchmanni* × *X. cortezi* hybrid populations is dominated by broad genomic scales, consistent with strong selection shortly after hybridization, and that there is likely a high density of selected sites. Moreover, repeatability in *X. birchmanni* × *X. cortezi* hybrid populations greatly exceeds what is observed in hybrid populations between the more closely related species *X. birchmanni* × *X. malinche*, pointing to pronounced changes in reproductive isolation with modest increases in genetic divergence (Fig. 1). This unique system with replicated hybridizing populations in two closely related species pairs gives us unprecedented power to unravel the dynamics of selection after hybridization and its impacts on repeatability in genome evolution.

Results

Chromosome scale genome assembly for X. cortezi

We generated a nearly chromosome-scale *de novo* assembly for *X. cortezi* using PacBio HiFi long-read sequencing at ~100x coverage. The genome was highly contiguous, with a contig N50 of 28,997,520 bp. Following reference-guided scaffolding to previously generated chromosome-level *X. birchmanni* and *X. malinche* assemblies (NCBI submission ID: JAXBVF000000000)*,* the final *X. cortezi* assembly was chromosome-level with a scaffold N50 of 32,220,398 bp, and >99.4% of all sequence contained in the largest 24 scaffolds (corresponding to the 24 *Xiphophorus* chromosomes). The total assembled sequence length of 723 Mb is similar to other *Xiphophorus* assemblies and close to the expected length for this species based on previously collected flow cytometry estimates [56]. The *X. cortezi* genome was also highly complete, with 98.6% of actinopterygii BUSCOs present and in single copy (C:98.6%[S:97.0%,D:1.6%],F:0.4%,M:1.0%,n:3640), and the annotation process recovered a total of 25,032 protein coding genes (see Methods, Supporting Information 1). While the two genomes are largely syntenic, we also identified putative structural rearrangements between *X. birchmanni* and *X. cortezi* (Supporting Information 1; Table S1-S2).

Genome-wide ancestry in the Chapulhuacanito and Santa Cruz populations

Past work from our group has focused on hybridization between *X. birchmanni* and *X. cortezi* in the Santa Cruz river drainage [53]. While we collected samples from multiple sites in the Santa Cruz drainage in our previous work, our analyses suggested that hybrids at different sampling sites originated from the same hybridization event [53]. For simplicity, throughout the manuscript refer to samples collected at the Santa Cruz site as "Santa Cruz" and samples collected nearby (e.g. in historical collections) as samples from the "Río Santa Cruz." Here, we report a previously undescribed hybridization event between *X. birchmanni* and *X. cortezi* at the Chapulhuacanito population (21°12'10.58"N, 98°40'28.27"W) in the Río San Pedro drainage, 17 km away by land and 130 km away in river distance from the Santa Cruz population (Fig. 1A). While on average populations from the Santa Cruz drainage derive 85-89% of their genomes from the *X. cortezi* parental species, the Chapulhuacanito population is more admixed, with 76%

of the genome derived from *X. cortezi* on average (Fig. 1C). In both populations, *X. birchmanni* is the minor parent species.

We also sequenced historical samples from the Chapulhuacanito and Río Santa Cruz populations from 2003, 2006, and 2017. This sampling period spans ~40 generations based on reported generation times for this species group [57]. Since hybridization began in these populations more than a hundred generations before the present (see below), our earliest sampling points only survey the latest chapter in the history of *X. birchmanni* × *X. cortezi* hybrid populations. Theory predicts that in the first several generations following hybridization, admixture proportions can change dramatically due to selection [15,29,32], but after this initial period, change in genome-wide average ancestry is expected to slow dramatically [32,33]. The observed patterns in our datasets are concordant with these predictions. Genome-wide average ancestry was essentially unchanged from 2003 to recent sampling from 2019-2021 (Chapulhuacanito: 78 ± 1.2 % *X. cortezi* in 2003 and 76 ± 2 % *X. cortezi* in 2021; Río Santa Cruz: 87 ± 5% *X. cortezi* in 2003 and 88 ± 1% *X. cortezi* in 2019-2020).

Demographic history of the hybrid populations

The demographic history of each hybrid population is also expected to impact how repeatable the outcomes of selection are and should be explicitly incorporated into analyses. For hybrid populations that formed on different timescales, both the amount of time for selection to shift ancestry at target loci and for genetic drift to shift ancestry at neutral loci would be expected to impact the repeatability of local ancestry across populations. To incorporate demographic history into our analyses, we used an approximate Bayesian computation approach to explore the likely demographic histories of both the Santa Cruz and Chapulhuacanito populations (see Methods; [18]). We performed simulations drawing from uniform distributions of time since admixture, admixture proportion, and hybrid population size and log uniform distributions for migration rates from each parental species, using SLiM ([58], see Methods). We used the observed genome-wide admixture proportion, coefficient of variance in genome wide and local ancestry, and median ancestry tract length as summary statistics, and used ABCreg ([59]; see Methods) to infer posterior distributions for the time since admixture, admixture proportion, hybrid population size, and migration rates from each parental species, in both hybrid populations. While we did not recover well-resolved posterior distributions for hybrid population size for either population, we do recover well-resolved posterior distributions for other demographic parameters. Based on the maximum a posteriori (or MAP) estimate of these distributions, we find that hybridization began over a hundred generations ago in both drainages (Fig. 1D; Chapulhuacanito =137; Santa Cruz = 263; see Table S3 for 95% confidence intervals), and that the migration rate from the parental populations has been very low (MAP estimate for Santa Cruz: $m_{cortezi} = 4 \times 10^{-5}$, $m_{birchmanni} = 0.00028$; Chapulhuacanito: $m_{cortezi} = 4 \times 10^{-5}$, $m_{birchmanni} = 0.00017$; Fig. S1; see Table S3 for 95% confidence intervals). In subsequent simulations, we explicitly incorporate this inferred demographic history to build our expectations of cross-population correlations under neutrality or under different models of selection.

Confirming the independent origin of the two hybrid populations

Given geographical isolation between the *X. birchmanni* × *X. cortezi* hybrid populations (Fig. 1A), we had good reason to believe that the two populations originated independently. However, given the extraordinarily high correlations in local ancestry we observed across the two populations (see below), we sought additional evidence that they were independent in origin.

Since we inferred local ancestry for individuals from both populations, we have access to information about historical recombination events in these populations. Specifically, a subset of recombination events that occurred in the hybrid ancestors of present-day individuals will be detectable as ancestry transitions in present-day individuals. Since independently formed hybrid populations have distinct histories of recombination, we tested for potential overlap in the locations of ancestry transitions. We generated a matrix containing the locations of ancestry transitions in each hybrid individual in our dataset (see Methods) and performed a principal component analysis. We see that the Santa Cruz and Chapulhuacanito populations separate out in PC space in this analysis (Fig. 2A). This suggests that the two populations have distinct historical recombination events. We also find that the frequency at which the locations of ancestry transitions are shared between individuals in the Santa Cruz and Chapulhuacanito populations is similar to the frequency expected by chance (Fig. 2B), again pointing to independent population histories.

We further explored these patterns using high-coverage whole genome sequencing data of three individuals from sympatric *X. birchmanni* populations at both Santa Cruz and Chapulhuacanito, and three naturally occurring hybrids at the two sites. We called variants (see

Methods) and performed principal component analysis on sympatric *X. birchmanni* individuals, hybrid samples, and pure *X. birchmanni* and *X. cortezi* collected from allopatric populations (Fig. S2-S4). Moreover, we performed local ancestry inference on the natural hybrids for which we had generated deep-sequencing data and identified homozygous *X. birchmanni* and homozygous *X. cortezi* ancestry tracts within these individuals (limiting our analysis to tracts that were the same ancestry state in all six deep-sequenced hybrids). We extracted these regions from the natural hybrids and from the parental genomes and performed principal component analysis on regions of *X. birchmanni* and *X. cortezi* ancestry separately. We found that ancestry tracts derived from the two hybrid populations formed separate clusters and individuals from the two populations differ in their degree of sequence mismatch (Fig. 2C-E; Methods). Moreover, when we used the variants in these ancestry tracts to calculate a genetic relatedness matrix using GCTA [60], we see evidence of related individuals within but not between populations (see Methods). Together, genetic and ancestry transitions patterns in the two populations corroborate our expectations from geographic distance and demographic analyses, indicating that hybrid populations in the Santa Cruz and Chapulhuacanito rivers originated independently. See Supporting Information 2 for a more thorough discussion of the implications of analyses of relatedness and genetic variation within and between populations.

Correlations between minor parent ancestry, the local recombination rate, and the density of coding and conserved basepairs

Past work on hybrid populations of *Xiphophorus* and in other systems [3,14,25,27,30,53] has found that the frequency of minor parent ancestry in the genome often correlates with factors such as the local recombination rate and the density of functional basepairs (e.g. coding regions). In the presence of selection against minor parent ancestry (due to hybrid incompatibilities or other mechanisms; [15,29]), both theory and simulations [32] predict that the level of minor parent ancestry will be positively correlated with the local recombination rate. Similarly, if selected sites fall more frequently in coding (or conserved) regions of the genome, and selection is sufficiently polygenic, we might expect to see a depletion of minor parent ancestry in these regions.

We tested for correlations between the local recombination rate estimated in *X. birchmanni* and local ancestry along the genome in a range of window sizes in both

Chapulhuacanito and Santa Cruz (Fig. 3A; Table S4). Although we have developed recombination maps for both species (see Methods), we chose to use the *X. birchmanni* map because it is likely to be more accurate (see [30]; Supporting Information 3) and our analyses suggest that it is extremely similar to the *X. cortezi* map (Fig. S5-S6; Supporting Information 3). Regardless of window size, we observe strong positive correlations between the local recombination rate and average minor parent ancestry in both populations (Fig. 3A; Table S4). After controlling for the strong effects of local recombination rate, we find that the density of coding (and conserved) basepairs also correlates with the distribution of minor parent ancestry in Chapulhuacanito and Santa Cruz (Table S5-S6; see also [53]). In particular, regions of the genome with especially high density of coding (or conserved) basepairs appear to be depleted in minor parent ancestry (Fig. 3B).

Repeatability in local ancestry between replicate hybrid populations

We found that local ancestry along the genome was surprisingly repeatable across the two *X. birchmanni* × *X. cortezi* hybrid populations (Fig. 3C). That is, the observed minor parent ancestry in a given 100 kb region of the genome in one population was highly predictive of the observed minor parent ancestry in that same region in the other population (Spearman's $\rho = 0.79$; $p = 2 \times 10^{-171}$). We note that because adjacent windows are not independent, for all analyses we report p-values after thinning data to include only one window per Mb (admixture LD in both populations decays to background levels over this distance; Fig. S7). The observed correlations in ancestry across populations exceed what we have previously detected in replicate *X. birchmanni* × *X. malinche* hybrid populations (Fig. 3D). While we detected these patterns across window sizes, they generally increased with larger window sizes (Table S7). We find that these correlations are robust to controlling for shared features of genome architecture like the local recombination rate and the locations of coding and conserved basepairs using a partial correlation approach (Table S8; see Methods).

This cross-predictability is not expected under neutrality but can be produced in simulations of hybridization followed by strong selection on many loci (Supporting Information 4). This suggests that repeatability in minor parent ancestry across *X. birchmanni* × *X. cortezi* hybrid populations is driven by a shared architecture of selection on hybrids. For comparison, we evaluated correlations in local ancestry observed when subsampling individuals from the same population and sampling year, samples from the same populations but different sampling years, and populations sampled from different sites on the same river. We reasoned that for each of these comparisons, samples are expected to largely share the same demographic history and history of selection. Reassuringly, we found that correlations in these analyses greatly exceeded those observed between Chapulhuacanito and Santa Cruz (Fig. S8; Table S7).

Our simulations indicate that the striking correlations we see in local ancestry across Chapulhuacanito and Santa Cruz (Fig. 3C) could be driven by a shared architecture of selection on hybrids in these populations (see below, Supporting Information 4). However, we wanted to thoroughly rule out other possible explanations, namely that technical factors might contribute to this signal. These approaches are described in detail in the Methods and Supporting Information, but we discuss them briefly here. We used simulations and analyses of lab generated crosses to confirm that our local ancestry inference approach is highly accurate (Fig. S9-Fig. S11; Methods and Supporting Information 5). We used simulations to artificially induce high error rates in local ancestry inference and found that it could not generate the patterns observed in our data (Supporting Information 5). We repeated analyses removing regions that are prone to error in local ancestry inference (Table S9; Supplementary Information 6) and controlling for our power to infer local ancestry along the genome (see Methods; Table S9), among other analyses (see Table S9; Methods; Supporting Information 6). None of these analyses qualitatively changed our results (Supplementary Information 4-6).

Simulations indicate that it is possible for selection alone to drive cross-population correlations at the magnitude we infer in Santa Cruz and Chapulhuacanito in scenarios where selection acts on many loci and is exceptionally strong (average *s* drawn from an exponential distribution of 0.4-0.6; Supplementary Information 2). Our results from lab-generated *X. birchmanni* × *X. cortezi* hybrids indicate that hybrids in this cross suffer immense fitness consequences, suggesting that such strong selection is plausible (see Discussion; [61]). Indeed, evaluating patterns of local ancestry across the two independently formed populations, we can see evidence for large, shared deserts of minor parent ancestry (Fig. 4A). This hints that the correlations we observe in our data may be largely driven by strong selection acting shortly after hybridization, resulting in shared patterning of minor ancestry over broad spatial scales along the genome. To evaluate this question in more depth and across spatial scales in the genome, we next used a wavelet-based analysis of cross population ancestry correlations [33].

Wavelet transform approach to infer the spatial scale of correlations in ancestry

In our windowed analyses, the correlations in ancestry between the Santa Cruz and Chapulhuacanito populations increase as we consider larger window sizes, suggesting that the observed correlations are driven by covariation in ancestry at large genomic scales (Table S7). Similarly, we find that the correlations between recombination rate and minor parent ancestry become stronger in larger genomic windows (Table S4).

Theory predicts that the strength of selection on hybrids will vary dramatically over time, since the removal of ancestry tracts harboring alleles that are deleterious in hybrids will be most rapid in the earliest generations following hybridization when ancestry tracts are long [29,31,32]. Furthermore, these dynamics can establish spatial ancestry patterns along the genome that persist over time and constrain subsequent evolution. This leads to the prediction that the genomic scale of autocorrelation in ancestry will be informative about the timing and strength of selection (relative to the onset of hybridization) [33]. To better understand the role of selection in shaping genomic ancestry patterns across replicate hybrid populations, we applied recently developed methods based on the Discrete Wavelet Transform [33] to our data (see Methods; Supporting Information 7). The intuition behind this analysis is as follows: moving along a chromosome, the ancestry proportion deviates around its chromosome-wide average, and this variation occurs over a range of different spatial scales (genomic window sizes, roughly speaking). The wavelet transform can be used to summarize the scales of variance in ancestry along a chromosome, as well as the contributions of each scale to the overall correlation between two signals measured along the chromosome (e.g. ancestry and recombination), where each component carries independent information about the overall correlation (see Methods). Because the scale of variation is ultimately determined by the lengths of admixture tracts, these signals contain information about the timing of selection and drift relative to the onset of hybridization [33].

Using this approach, we found that the overall correlation between minor parent ancestry and recombination in both replicate populations is predominantly attributable to broad genomic scales (Fig. 4C). Furthermore, wavelet correlations between minor parent ancestry and recombination were strongly positive in replicate populations, with the strongest correlations observed at the broadest genomic scales (Fig. S12). As discussed in Groh and Coop (2023), the squared correlation coefficients for ancestry vs. recombination can be interpreted as the percent of variance in ancestry at each scale attributable to selection, since these correlations are only generated by selection and not by drift (barring errors in ancestry inference; see Supporting Information 7). Applying this logic, we find that correlations with recombination indicate roughly 80% of the variance in ancestry at the broadest genomic scales (e.g. 16 Mb) in the Santa Cruz and Chapulhuacanito populations can be attributed to selection against minor parent ancestry. By contrast, comparatively little of the variance in ancestry at fine genomic scales is attributable to selection against minor parent ancestry (e.g. 0.2% at a scale of 32 kb).

We next applied this approach to the correlation between minor parent ancestry across the two replicate *X. birchmanni* and *X. cortezi* populations. We found that across scales, cross-population ancestry correlations between *X. birchmanni* × *X. cortezi* hybrid populations were stronger than the correlations observed with recombination rate, especially at finer spatial scales. Thus, ancestry in a replicate hybrid population is a better predictor of fine-scale genetic ancestry patterns than recombination rate. This implies that recombination alone only captures a portion of the total effects of selection on ancestry patterns, and that its effects in mediating parallel genomic outcomes of hybridization manifest predominantly over broad genomic scales. From cross-population ancestry wavelet correlations, estimates of the proportion of ancestry variance attributable to selection on minor parent ancestry range from ~25% at a scale of 32 kb to as high as 93% at a scale of 8 Mb (Fig. 4D). Surprisingly, we found that significant positive correlations persisted even at very small spatial scales (Fig. S13). This pattern is consistent with convergent selection shaping very fine scale ancestry patterns, although we discuss important caveats to this interpretation in in Supporting Information 7. Nonetheless, the magnitude and scales of ancestry correlations across populations suggest that predictability is driven by both early and continued selection on hybrids.

For comparison, we repeated these analyses in two *X. birchmanni* × *X. malinche* hybrid populations, the Acuapa population and the Aguazarca population [30]. *X. birchmanni* and *X. malinche* are more closely related than *X. birchmanni* and *X. cortezi* (Fig. 1B), and hybridization began more recently (within the last 50-100 generations; [18]). We again find strong positive correlations between minor parent ancestry in the two populations at broad genomic scales, but these are noticeably reduced compared to the cross-population comparison between the two *X. birchmanni* × *X. cortezi* populations (Fig. 4D, Fig. S13, Supporting Information 7). These results are consistent with weaker selection overall against minor parent ancestry in *X. birchmanni* × *X. malinche* hybrid populations, and/or fewer loci under selection, both of which may be expected given that these species diverged more recently (Fig. 1B; [30,62]). Moreover, previous work analyzing wavelet correlations between minor parent ancestry and recombination rate in *X. birchmanni* × *X. malinche* populations found that only ~20% of the variation in minor parent ancestry at large spatial scales was attributable to selection [33]. Overall, these results suggest that genome evolution after hybridization is substantially more predictable for *X. birchmanni* × *X. cortezi* hybrids.

Finally, we examined the genomic scale of shared ancestry patterns between a *X. birchmanni* × *X. cortezi* hybrid population and a *X. birchmanni* × *X. malinche* hybrid population (the Chapulhuacanito and Acuapa populations respectively). We observed positive correlations in minor parent ancestry at broad scales but find that these correlations are dramatically reduced at fine scales, especially compared to analyses of the populations of the same hybridizing pair (Fig. 4D). This would be expected if replicate populations of the same hybridizing pair show greater overlap in the fine-scale targets of selection than populations of different hybridizing pairs. The positive fine-scale ancestry correlations within replicate hybrid populations (Fig. 4D, Fig. S13) are consistent with this interpretation (see also Supporting Information 7). We thus suggest that broad-scale predictability among different hybridizing pairs may be driven primarily by effects of shared genome architecture rather than shared identity of selected loci.

Repeatability in minor parent deserts and islands between replicate X. birchmanni × X. cortezi populations

Given that the results of wavelet-based analyses point to shared targets of selection across *X. birchmanni* × *X. cortezi* hybrid populations, we were interested in whether we could identify individual loci that are likely to be under selection. Loci that are shared targets of selection could be alleles that are globally deleterious (or beneficial), or those that are involved in hybrid incompatibilities between *X. birchmanni* × *X. cortezi*. Using our large recent population samples, we identified contiguous regions of low minor parent ancestry, or minor parent ancestry "deserts", in each *X. birchmanni* × *X. cortezi* hybrid population and asked how frequently they overlapped across populations (see Methods). Simulations suggest that our approach has high sensitivity and low false positive rates (~70% power at s=0.05; average of 2-4 shared deserts detected genome-wide in neutral regions; see Supporting Information 8). We identified 115 "deserts" of low minor parent ancestry in Santa Cruz and 152 deserts in Chapulhuacanito.

Strikingly, 38 of these regions overlapped, exceeding expectations by chance (Fig. 5A; see Methods). The average length of these regions was 1.8 Mb with a total of ~40 Mb of the 723 Mb genome falling into shared deserts. Since the typical ancestry tract length for *X. cortezi* (i.e. the major parent) in these populations is much smaller (~150 kb), this hints that these regions may have changed in ancestry shortly after initial hybridization. These shared minor parent ancestry deserts are excellent candidates for shared regions under selection in the two hybrid populations.

Similarly, we identified regions of especially high minor parent ancestry in each *X. birchmanni* × *X. cortezi* hybrid population and asked how frequently they overlapped across populations compared to expectations by chance (see Methods). In doing so, we found evidence for 89 shared minor parent "islands" out of 238 islands in Santa Cruz and 147 in Chapulhuacanito, again exceeding the level of sharing expected by chance (Fig. 5A; Methods). The typical length of shared islands was 190 kb, much smaller than that observed for shared deserts, but together these regions still covered a substantial portion of the genome (~29 Mb). We report the genes observed in these regions (Table S10) and analysis of functional enrichment in the supplementary materials (Supporting Information 9).

We compared minor parent deserts and islands identified in the *X. birchmanni* × *X. cortezi* hybrid populations to those detected in the *X. birchmanni* × *X. malinche* hybrid populations. As expected, we found many fewer shared deserts and islands across hybrid population types (Fig. S14), with shared deserts and islands only slightly exceeding expectations by chance in most comparisons.

Since we had access to time-series data for both the Santa Cruz and Chapulhuacanito populations, we were interested in evaluating how ancestry at minor parent deserts and islands has changed over the last 40 generations. Given that both hybrid populations are estimated to be over 100 generations old, we would expect that loci under strong or moderate selection would be fixed even at the earliest time points in our dataset. Indeed, we find that regions that fall into shared ancestry deserts tend to have low minor parent ancestry in 2003 and maintain low ancestry through time (Fig. 5B). The same is generally true for regions of high minor parent ancestry, although we do identify six minor parent islands where minor parent ancestry significantly increases between 2003 and 2020-2021 (Table S10).

Finally, we evaluated ancestry at rearrangements identified between *X. birchmanni* and *X. cortezi* based on our new PacBio HiFi based assemblies. We identified nine inversions greater than 100 kb, ranging in size from 218 kb to 6.7 Mb (Table S2; Fig. S15). These inversions were concentrated on chromosomes 8 and 17 (six out of nine of the inversions). As chromosomal inversions tend to suppress recombination in heterozygotes we predicted that these regions would be especially depleted in minor parent ancestry. Notably, we found that on average these regions were depleted in minor parent ancestry compared to expectations by chance (Fig. 5E), but not when compared to non-inverted regions of the genome that had exceptionally low recombination rates (Fig. 5E).

Ancestry at known incompatibilities identified between X. birchmanni and X. cortezi

We were also interested in evaluating patterns of minor parent ancestry locally at regions that are known to be under selection in hybrids between *X. birchmanni* and *X. cortezi*. Other work from our lab has identified a mitonuclear hybrid incompatibility between individuals with the *X. cortezi* mitochondria and homozygous *X. birchmanni* ancestry at *ndufs5* and *ndufa13* [53,55,61]. F_2 hybrids that inherit the *X. cortezi* mitochondrial haplotype and two copies of the *X. birchmanni* allele at *ndufs5* experience mortality during embryonic development [55,61]. Inheriting the *X. cortezi* mitochondrial haplotype and two copies of the *X. birchmanni* allele at *ndufa13* causes higher rates of post-natal mortality. Because hybrid populations at both Santa Cruz and Chapulhuacanito have fixed the *X. cortezi* mitochondrial haplotype (Table S11; [53,55]), this leads to the strong expectation that they will largely have purged *X. birchmanni* ancestry at *ndufs5* and *ndufa13*.

We evaluated ancestry in these regions of the genome in our large sample of hybrid individuals from both Santa Cruz and Chapulhuacanito. We identified a large, shared ancestry desert surrounding *ndufa13* on chromosome 6 (Fig. 5C). For *ndufs5*, the region surrounding the gene on chromosome 13 was identified as an ancestry desert in Chapulhuacanito, but not in Santa Cruz. Closer examination of this region (Fig. S16) indicates that *X. birchmanni* ancestry at *ndufs5* is depleted in Santa Cruz but falls just above the 5% quantile of minor parent ancestry used to identify deserts genome-wide in Santa Cruz (the 5% quantile was 2.2% *X. birchmanni* ancestry while an average of 2.3% *X. birchmanni* ancestry was observed at *ndufs5;* see Methods). Moreover, both regions were consistently low in *X. birchmanni* ancestry through time in our samples from Chapulhuacanito (Fig. 5D) and no individuals homozygous for *X. birchmanni* ancestry at either region were observed across the two populations. Based on predictions from Hardy-Weinberg equilibrium <0.05% of mating events would be expected to produce embryos incompatible at *ndufs5* or *ndufa13* in either population. Since these two genes form part of mitochondrial protein complex I, we also analyzed ancestry at genes that are involved in protein complexes genome-wide (Fig. S17; Supporting Information 10).

Discussion

The extent to which genome evolution after hybridization is predictable is an open question in evolutionary biology. Given the large number of species that have exchanged genes with their close relatives, the answer to this question has wide ranging implications for species across the tree of life. Few studies to date have been able to tackle this question because addressing it requires access to multiple, independently formed hybrid populations and accurate local ancestry inference approaches where technical factors such as variation in error rates or power to infer ancestry along the genome can be excluded as drivers of the observed patterns. Even well-studied cases with excellent genomic resources such as the human-Neanderthal and human-Denisovan admixture events present a challenge in appropriately accounting for such technical factors.

Here, we further developed *Xiphophorus* as a natural biological system in which to address these fundamental questions. We describe two hybrid populations between *X. cortezi* and *X. birchmanni* that formed in different river drainages in the last ~150 to 300 generations. Multiple lines of evidence—from geography to genetic variation to recombination history—confirm that the two hybrid populations formed independently. *X. cortezi* and *X. birchmanni* diverged an approximately 450k generations ago [62] and we estimate pairwise sequence divergence at 0.6%. Since levels of within-species polymorphism are relatively low, this results in a high density of fixed ancestry informative sites – approximately 4 per kb – with which to precisely infer ancestry along the genome and compare ancestry variation across the two populations.

Shortly after hybridization, hybrid genomes may contain large numbers of selected alleles that are linked on the same haplotype. Accordingly, both theory and empirical results have indicated that selection interacts with the global and local recombination rate to reshape minor parent ancestry in the genome (assuming that minor parent ancestry is on average deleterious; [30–32]). As in previous studies of *Xiphophorus* hybrids [30,53–55], we find a strong depletion of ancestry from the minor parent species (*X. birchmanni* in both populations) in regions of the genome with low recombination rates (Fig. 3A), as well as a more subtle depletion of minor parent ancestry in regions of the genome of high coding (or conserved) basepair density (Fig. 3B). Moreover, wavelet analyses indicate that correlations between minor parent ancestry and recombination rate are primarily driven by the broadest spatial scales (i.e. >4 Mb; Fig. 4B, S12), suggesting that selection on early generation hybrids is driving patterning of minor parent ancestry at a genome-wide scale in both populations [33]. These analyses suggest that a striking amount of local ancestry variation at broad spatial scales is attributable to the action of natural selection (~80%).

Perhaps the most surprising result of our study is the extraordinarily high correlations in local ancestry across the two *X. cortezi* and *X. birchmanni* hybrid populations (Fig. 3C). The results of wavelet analyses indicate that broad-scale changes in ancestry along the genome in one hybrid population (at the scale of >8 Mb) predict a remarkable ~90% of the variance in the other hybrid population. We found that this cross-population repeatability was robust to iterations of the analysis controlling for potential technical confounders (see Methods; Table S9). Since shared patterns of ancestry deviations are not predicted under neutrality, these results demonstrate that the correlations we observe are attributable to natural selection driving parallel changes in minor parent ancestry in the two hybrid populations, presumably due to selection on the same loci. Since these correlations are strongest at the broadest spatial scales in the genome, this indicates that natural selection acting shortly after hybridization was important in establishing them. The degree of cross-population repeatability we observe here exceeds that reported in other studies that have found evidence for such patterns [24,33,52,53].

What mechanisms could drive such high repeatability in minor parent ancestry across independently formed hybrid populations? Given the frequency of hybrid incompatibilities in *Xiphophorus* [30,54,55] and the fact that neither *X. birchmanni* or *X. cortezi* have experienced sustained bottlenecks like those observed in other *Xiphophorus* species ([30,62]; Fig. S4), we predicted that selection on hybrid incompatibilities may be an important driver of this signal. In simulations, we confirmed that strong selection on the same hybrid incompatibilities can, in principle, generate exceptionally high correlations in local ancestry across populations, similar to those observed in our data (Supporting Information 4; Fig. S18-S19). Results from artificial crosses between *X. cortezi* and *X. birchmanni* support the conclusion that selection is extremely strong on early-generation hybrids. One F_1 cross direction fails to develop (with *X. birchmanni* mothers) and the other produces offspring with a 6:1 male sex-bias (with *X. cortezi* mothers; [61]).

In the case of strong selection against intrinsic hybrid incompatibilities, we expect to see large 'deserts' of minor parent ancestry that are shared across independently formed hybrid populations. Genome-wide we observe over a hundred such deserts in *X. birchmanni* × *X. cortezi* populations and find that more than 25% of these minor parent ancestry deserts are repeated across the two populations (Fig. 5A). Moreover, in cases where deserts are not replicated across populations, minor parent ancestry still tends to be low in the second population (on average falling in the lowest quartile of minor parent ancestry; Fig. 4A). Consistent with our findings that selection acted early after hybridization, we find that minor parent deserts are typically large (on average 1.8 Mb). These regions are exciting candidates to pursue as we begin to map hybrid incompatibilities between *X. birchmanni* and *X. cortezi* in natural populations and in the laboratory.

Beyond these genome-wide patterns, we know the precise locations of two loci that cause a lethal mitonuclear incompatibility in *X. birchmanni* × *X. cortezi* hybrids when they are mismatched with mitochondrial ancestry [55,61]. If selection on hybrid incompatibilities is responsible for local deviations in ancestry in *X. birchmanni* × *X. cortezi* hybrid populations, we should see biased ancestry in these specific regions of the genome in both hybrid populations. Indeed, we identify large regions depleted of minor parent ancestry surrounding the genes involved in lethal mitonuclear incompatibilities on chromosome 6 (Fig. 5C) and chromosome 13 (Fig. S16). Based on these results at known incompatibilities, we infer that shared local ancestry patterns in *X. birchmanni* × *X. cortezi* hybrid populations are at least in part driven by strong selection against hybrid incompatibilities.

We also observed unexpectedly large overlap in regions of the genome where minor parent ancestry is elevated across the two populations. Eighty-nine of the 147 regions with elevated minor parent ancestry in Chapulhuacanito were also elevated in the Santa Cruz population (~60%). This enrichment may indicate that *X. birchmanni* ancestry in these regions is beneficial to hybrids, although we found no patterns of gene enrichment within islands that exceeded expectations by chance (Supporting Information 9), nor overlap with previously mapped QTL for sexually selected traits or ecological adaptations in *Xiphophorus* species [63,64]. The combined dynamics of genome-wide selection against deleterious and adaptive variation in hybrids are poorly understood in most cases (but see [11,15,24]), pointing to exciting directions for future work.

The variety of hybrid populations within *Xiphophorus* allowed us to ask how predictability of genome evolution after hybridization varies with genetic divergence. We analyzed replicate hybrid populations formed between both *X. birchmanni* and *X. malinche* and *X. birchmanni* and *X. cortezi*. Since *X. birchmanni* and *X. malinche* are more closely related than *X. birchmanni* and *X. cortezi*, theory predicts that the total strength of selection on *X. birchmanni* × *X. malinche* hybrids across the genome should be weaker [35]. Notably, the correlations in local ancestry we observed in the *X. cortezi* × *X. birchmanni* hybrid populations greatly exceed those observed in *X. birchmanni* × *X. malinche* hybrid populations. Comparisons across hybrid population types (i.e. comparing *X. cortezi* × *X. birchmanni* hybrid populations to *X. birchmanni* × *X. malinche* hybrid populations) yield the lowest predictability in minor parent ancestry (Table S7). Our wavelet analyses suggests that repeatability across hybrid population types is limited to the broadest genomic scales, potentially reflecting the effects of shared genomic architecture rather than shared targets of selection. This result is consistent with the idea that loci involved in hybrid incompatibilities may arise idiosyncratically between lineages, as different sets of mutations fix along different evolutionary branches. We note that while *X. cortezi* × *X. birchmanni* populations tend to be older than *X. birchmanni* × *X. malinche* populations (Fig. 1; [30]), wavelet analyses suggest that in both cases much of the observed variation in minor parent ancestry along the genome is established in the earliest generations following hybridization (Fig. 4; [33]).

Hybridization is a common evolutionary process that profoundly shapes genome evolution. Our accurate local ancestry inference approaches allowed us to uncover striking repeatability in local ancestry across independently formed *X. birchmanni* × *X. cortezi* hybrid populations and begin to unravel the fundamental question of how these patterns scale with evolutionary divergence between species [65]. We find that both local factors like the locations of hybrid incompatibilities and global factors such as the recombination landscape in the genome shape this process. The extent to which the patterns observed in *Xiphophorus* hybrids are generalizable to other hybridizing species is an exciting question that awaits results from other taxonomic groups.

Methods

Sample collection

Samples for low-coverage whole genome sequencing were collected from two different geographical regions (Fig. 1). Wild fish were collected using baited minnow traps in Hidalgo and San Luis Potosí, Mexico. We previously identified hybrids between *X. birchmanni* and *X. cortezi* at multiple sites on the Río Santa Cruz in northern Hidalgo [18,62]. We continued to sample from these sites for the present analysis (Huextetitla - 21°9’43.82"N 98°33’27.19"W and Santa Cruz - 21°9’27.63"N 98°31’13.79"W). We also added a new site in a different drainage (Fig. 1), near the town of Chapulhuacanito (21°12'10.58"N 98°40'28.27"W). This site also contained *X. birchmanni* × *X. cortezi* hybrids (see Results), but this hybridization event is clearly independent given the geographical distance and lack of river connectivity between these locations. At both collection sites, nearly pure *X. birchmanni* individuals were also sampled. These individuals were identified based on their genome-wide ancestry and excluded from further analysis.

We combined previously collected datasets from the Río Santa Cruz (N=254; [18,62]) with 216 new samples collected from Chapulhuacanito in June of 2021. Collected fish were anesthetized in 100 mg/mL buffered MS-222 and water, following Stanford APLAC protocol #33071. A small fin clip was taken from the caudal fin of each individual and preserved in 95% ethanol for later DNA extraction.

For this study, we also took advantage of historical collections from 2003, 2006, and 2017 in the same regions. These samples were matched to present-day collection sites using GPS coordinates and represented a mix of fin clips preserved in DMSO and whole fish preserved in 95% ethanol. We prepared libraries, sequenced all samples, and identified 76 hybrids from historical samples from Chapulhuacanito and 23 from the Río Santa Cruz.

Chromosome scale assembly for X. cortezi

We generated a new reference genome for *X. cortezi* for this project from a lab-raised male descended from an allopatric population sampled on the Río Huichihuyan. Previous work involving *X. cortezi* used a draft genome assembled with 10X chromium linked read technology [18,62]. We assembled the new reference using PacBio HiFi data.

Genomic DNA was isolated from tissue using QIAGEN's Genomic-Tip 500/G columns following the manufacturer's recommendations with some adaptations. ~400 mg of body tissue was digested in 1.5 mL of Proteinase K and 19 mL Buffer G2 at 50°C for 2 hours, inverting the sample every half hour. Following the incubation, the column was equilibrated using 10 mL of Buffer QBT. The sample was vortexed for 10s at maximum speed, then immediately applied to the column. Two washes were performed with a total of 30 mL of Buffer QC. The column was then transferred to a clean 50 mL tube and genomic DNA was eluted from the column with 15 mL of Buffer QF that was prewarmed to 50°C. The DNA was precipitated using 10.5 mL of isopropanol, mixed gently, then centrifuged immediately at a speed of 5000 x g for 15 minutes at 4°C. The DNA pellet was then washed with 4 mL of cold 70% ethanol and re-pelleted via centrifugation. Then the pellet was air-dried for 10 min and resuspended in 1.5 mL of Buffer EB. Genomic DNA was quantified and assessed for quality using a Qubit fluorometer, Nanodrop, and Agilent 4150 TapeStation. Extracted DNA was sent to Admera Health Services, South Plainfield, NJ for PacBio library prep and sequencing on SMRT cells. Raw sequence data is available on NCBI's Sequence Read Archive (SRAXXXXX).

To remove residual adapter contamination from the HiFi reads, we used HiFiAdapterFilt [66] with the default match parameter of 97% and a length parameter of 30bp. We then generated a phased genome assembly with hifiasm (v0.16.1; [67]). The resulting primary assembly was 144 contigs with a N50 of 28,997,520 bp. To achieve a chromosome-level assembly, we scaffolded the *X. cortezi* genome to the chromosome-level genomes of species in its sister clade: *X. birchmanni* and *X. malinche* (NCBI submission ID: JAXBVF000000000) using RagTag (v2.1.0; [68]). Where these scaffolded genomes differed in synteny, we used the chromosome-level assemblies of *X. hellerii*, *X. maculatus*, and *X. couchianus* as outgroups to select the ancestral orientation for *X. cortezi*. This scaffolded *X. cortezi* genome had a scaffold N50 of 32,220,398 bp and length of 723,632,656 bp. These putative *X. cortezi* chromosomes were aligned to the *X. maculatus* genome assembly using minimap2 (v2.24; [69]) and oriented and numbered according to identity with *X. maculatus*.

Chromosome 21 is known to contain the major sex determination locus in many *Xiphophorus* species [70]. To resolve potential structural variation at this locus and include both X and Y linked sequence in the *X. cortezi* reference genome, we generated an alignment between the two inferred haplotypes for chromosome 21. We found that one chromosome 21 haplotype was syntenic to chromosome 21 in *X. birchmanni*, while the other contained a 7 Mb chromosomal inversion relative to *X. birchmanni*, which is syntenic to all other *Xiphophorus* species and likely represents the ancestral *Xiphophorus* arrangement of the Y-chromosome [71].

The mitochondrial genome was assembled from the adapter-filtered hifi reads using MitoHiFi (v3.2; [72]) with default parameters and using the *X. maculatus* mitochondrial genome as a reference. We used BLASTn [73] searches to identify and subsequently remove mitochondrial contaminant sequences present in the nuclear genome, which were present on only 6 contigs that were all less than 40 kb in length. Following contaminant removal, the mitochondrial genome assembled with MitoHiFi was added to the *X. cortezi* assembly. The final assembly is available on Dryad (Accession pending).

Annotation of the X. cortezi assembly

The *X. cortezi* genome was annotated using a pipeline adapted from a previous study [74]. Transposable elements (TE) in the assembly were identified using RepeatModeler and RepeatMasker [75]. RepeatModeler was first used for an automated genomic discovery of transposable element families in the assembly. This result, together with Repbase and FishTEDB [76,77], was input into RepeatMasker for an additional retrieval of TEs based on sequence similarity. For protein coding gene annotation, TEs from known-families were hard-masked and simple repeats were soft-masked from the assembly. We used a tool designed to parse RepeatMasker output files [78] to compute quantitative information on representation of different TE families. We repeated this approach for the *X. birchmanni* PacBio reference assembly generated using the same approach. Analysis of differences between the two species in repeat content is available in Supporting Information 1.

Protein coding genes were annotated by collecting and synthesizing gene evidence from homologous alignment, transcriptome mapping and *ab initio* prediction. For homologous alignment, 455,817 protein sequences were collected from the vertebrate database of Swiss-Prot (https://www.uniprot.org/statistics/Swiss-Prot), RefSeq database (proteins with ID starting with “NP” from “vertebrate_other”) and the NCBI genome annotation of human (GCF_000001405.39_GRCh38), zebrafish (GCF_000002035.6), platyfish (GCF_002775205.1), medaka (GCF_002234675.1), mummichog (GCF_011125445.2), turquoise killifish (GCF_001465895.1) and guppy (GCF_000633615.1). We then aligned those protein sequences onto the assembly using both GeneWise and Exonerate (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate) to collect homologous gene models. In order to speed up GeneWise, GenblastA was used to retrieve the rough alignment region of the assembly for each protein [79].

For transcriptome mapping, we used previously collected RNA-seq reads from multiple tissues [54], cleaned them using fastp [80], and mapped them to the assembly using HISAT [81]. StringTie was then used to interpret gene models from the mapping results [81]. In parallel, we used Trinity to assemble RNA-seq reads into transcript sequences and aligned them to assembly for gene modeling using Splign [82,83].

We used AUGUSTUS for the *ab intio* gene prediction [84]. AUGUSTUS was trained for the first round using BUSCO genes. Genes that were predicted repeatedly by Exonerate, Genewise, StringTie and Splign were considered to be high quality genes and were used to train AUGUSTUS for the second round. All collected homologous and transcriptome gene evidence were used as hints for AUGUSTUS for the ab-initio gene prediction.

To generate the final consensus annotation, we screened homology gene models locus by locus. When two gene models competed for a splice side, we kept the one better supported by transcriptome evidence (using transcriptome data from [54]). When a terminal exon (with a start/stop codon) from an ab-initio or homology gene model was better supported by transcriptome data than that of the previously selected gene model, the exons in question were replaced by the predictions of the gene model best supported in the transcriptome data. We also kept an ab-initio prediction when its transcriptome support was 100% and it had no homology prediction competing for splice sites.

Low coverage whole genome sequencing

We extracted DNA from fin clips collected from wild-caught fish using the Agencourt DNAdvance kit (Beckman Coulter, Brea, California). We used half-reactions but otherwise followed the manufacturer's instructions for DNA extraction. We used a BioTek Synergy H1 (Agilent, Santa Clara, CA) microplate reader to quantify extracted DNA. We diluted DNA to a concentration of 10 ng/ul and then prepared tagmentation-based libraries from this genomic DNA for low coverage whole genome sequencing. The approach used for generating libraries is described in Langdon et al. 2022 [18]. Dual-indexed libraries were bead purified with 18% SPRI magnetic beads, quantified on a qubit fluorometer (Thermo Scientific, Wilmington, DE), and visualized on an Agilent 4200 Tapestation (Agilent, Santa Clara, CA). Purified libraries were sequenced by Admera Health Services (South Plainfield, NJ) on an Illumina HiSeq 4000 instrument.

Whole genome resequencing

To evaluate patterns of genetic variation within ancestry tracts, we sequenced a subset of individuals (N=3 per genotype per population) at high coverage. For these individuals, we prepared libraries following the approach of Quail et al. 2009 [85]. We used 500 ng – 1 ug of DNA per sample and sheared this input DNA to approximately 400 bp using a QSonica sonicator. The fragmented DNA underwent an end-repair reaction with dNTPs, T4 DNA polymerase, Klenow DNA polymerase and T4 PNK for 30 minutes at room temperature. An A-tail was added to the end-repaired DNA using a mix of Klenow exonuclease and dATP, incubated for 30 minutes at 37 C. The A-tail facilitated ligation of adapters with DNA ligase in a 15 minute reaction performed at room temperature. The resulting sample was purified using the Qiagen QIAquick PCR purification kit. Barcodes were added during a final PCR amplification step using the Phusion PCR kit, which was run for 12 cycles. This reaction was purified with 18% SPRI beads and libraries were visualized on the Agilent 4200 Tapestation and quantified using a Qubit fluorometer. These libraries were also sent to Admera Health Services for sequencing on an Illumina HiSeq 4000 machine.

Inferring recombination maps for X. birchmanni and X. cortezi

In past work, we used population genetic methods to infer a linkage disequilibrium (LD) based recombination map for an earlier version of the genome assembly for *X. birchmanni* [30]. We repeated the same approaches with the new *X. birchmanni* reference genome to generate a new LD-based map. Briefly, we used the previously published resequencing data for 22 adult *X. birchmanni* individuals and a pedigreed family with five offspring [30], for a total of 24 unrelated adults. We mapped reads to the genome with *bwa* mem, realigned indels with PicardTools, and called variants with GATK (v3.4; [86]). We filtered variant and invariant sites based on quality thresholds as we had with the original recombination map (DP<10; RGQ <20; QD<10; MQ < 40; FS>10; SOR > 4; ReadPosRankSum< -8; MQRankSum < -12.5). We excluded sites that overlapped with annotated repetitive regions or had <0.5X or >2X the average genome-wide coverage for that individual. For invariant sites, only RGQ and DP filters could be used. Using this filtered list of sites, we inferred the expected error rate with plink [87] using expectations of mendelian segregation in the pedigree. Finding evidence of a low error-rate (~0.45% per SNP across 5 offspring), we first removed these errors and then proceeded to phasing and inferring the LD map. We performed phasing using the program shapeit2 with the duohmm flag for inclusion of family data [88]. Past simulations matching parameters observed in *X. birchmanni* have suggested that although phasing likely introduces errors, improvements in map resolution outweigh errors introduced by phasing [30].

We inferred the LD map using LDhelmet. LDhelmet relies on a mutation transition matrix for recombination map inference [89] and also can take advantage of distributions of ancestral alleles when computing likelihoods. To infer ancestral alleles for both purposes, we used phylofit [90]. Previous simulations matching parameters observed in *X. birchmanni* have suggested that this approach results in accurate inference of ancestral sequences [30]. We used previously collected whole genome sequence data from 11 species of *Xiphophorus* (Table S12) to infer the likely ancestral basepair at variable sites as described previously [30] using the prequel command [90]. To run phylofit, we provided the aligned sequences and the inferred species tree for this groups of species [91]. For mutation matrix inference based on phylofit output, we used a threshold of 0.99 to convert posterior probabilities for the ancestral basepair to hard calls.

We then used phased haplotypes from all unrelated *X. birchmanni* individuals (48 haplotypes in total) and the mutation transition matrix to infer an LD-based recombination map with LDhelmet [89]. The total number of SNPs input into LDhelmet was 2,565,331. We first computed a likelihood lookup table for ρ values using a grid table ranging from 0 – 10 (sampling in intervals of 0.01 from 0-1 and 1 from 1-10). We next inferred recombination rates using LDhelmet's rjMCMC procedure with a block penalty of 50, a burn-in of 100,000, and ran the Markov chain for 1,000,000 iterations. Past work has suggested that a block penalty of 50 improves accuracy for inference of broad scale recombination rates in *Xiphophorus* [30]. Following map inference, we excluded SNP intervals with implausible high recombination rates (ρ/bp ≥ 0.4) and summarized recombination rates in windows of physical distance ranging from 5 kb – 5 Mb. We also used the local recombination rate estimates and the inferred lengths of each chromosome in cMs to divide the chromosome into windows of genetic distance for certain analyses (Supporting Information 11).

Because we had access to whole-genome resequencing data for 9 unrelated *X. cortezi* individuals from the Huichihuyan river (near the Nacimiento) from previous work [18,54], we decided to supplement this data to build an LD-based map for this species as well. To generate a comparable sample size for this inference, we sequenced an additional 8 individuals following the whole genome resequencing protocol described above. The average coverage for the *X. cortezi* individuals was ~65X, and the range was 19-113X. We inferred an LD-based map for *X. cortezi* as described above, except that we lacked access to pedigree data for mendelian error correction and phasing.

With the lower sample size and lack of pedigree data, we expected the *X. cortezi* map to be less accurate than the *X. birchmanni* map but used it to test general hypotheses. Swordtails have deleted the N-terminal domain of PRDM9 [92], and have a conserved PRDM9 zinc-finger binding domain across the clade. Past work has indicated that swordtails behave as PRDM9 knock-outs with a higher frequency of recombination events near the TSS, CpG islands, and H3K4me3 marks [30,92]. Using the inferred LD map for *X. cortezi*, we confirmed that we observe elevated recombination rates close to the TSS and H3K4me3 peaks, similar to patterns observed in *X. birchmanni* (Fig. S6; Supporting Information 3). We note that the median inferred ρ/bp in *X. cortezi* was substantially higher than in *X. birchmanni* (0.0027 versus 0.00076). Based on the results of our analyses of historical population sizes (see Supporting Information 3), we expect to see elevated ρ/bp in *X. cortezi* since ρ reflects *4Ne*r* and we infer that *X. cortezi* has had approximately 2X the effective population size of *X. birchmanni* over the past 100k generations. However, ρ/bp may also be impacted by a higher error rate in the *X. cortezi* recombination map given the lack of access to pedigree data.

Changes to the local ancestry inference pipeline

We previously developed approaches for local ancestry inference for hybrids between *X. birchmanni* and *X. cortezi*, but we made several improvements upon previous implementations for this project. First, we used a new chromosome scale assembly for *X. cortezi* generated by PacBio HiFi technology (see above). To more accurately quantify allele frequencies in the parental species, we sampled additional allopatric populations of *X. cortezi*, which has been less intensively sampled from a genomic perspective than *X. birchmanni*, and sequenced F_1 hybrids between the two species for error correction. We also identified and corrected an error in the *ancestryinfer* code (https://github.com/Schumerlab/ancestryinfer) that had resulted in a number of ancestry informative sites being erroneously excluded in previous versions of the pipeline.

Using the new assemblies, we identified candidate ancestry informative sites by aligning resequencing data from a high coverage *X. cortezi* individual to the *X. birchmanni* PacBio assembly (as described previously; [53,54,93]), and identifying all sites that were homozygous for different states in this data. We then treated these sites (2.64 million) as potential ancestry informative sites and evaluated their frequency in allopatric *X. cortezi* and *X. birchmanni* populations using 1X whole genome sequence data from 90 individuals of each species from three source populations each (*X. birchmanni*: Coacuilco, Talol, Xaltipa; *X. cortezi*: Puente de Huichihuyan, Octzen, Calle Texacal). We identified sites that had a 98% or greater frequency difference between the two species as our filtered set of ancestry informative sites (1,001,684 sites). We used these sites and their observed frequencies in the parental species as input for our ancestry HMM pipeline (*ancestryinfer*; [93]).

We next took advantage of our lab-generated F_1 hybrids to further filter these ancestry informative sites. We collected ~1X whole genome sequence data for 42 F_1 hybrids we generated between *X. birchmanni* and *X. cortezi* and analyzed these individuals using the *ancestryinfer* pipeline, specifying the *X. birchmanni* reference as genome 1 and the *X. cortezi* reference as genome 2. We set the error rate to 0.02 for this initial analysis. After running the pipeline, we converted posterior probabilities for each ancestry state into hard-calls using a posterior probability threshold of 0.9. Because F_1 hybrids should be heterozygous for all ancestry informative sites across the genome, we identified ancestry informative sites that were called with high confidence as homozygous *X. birchmanni* or homozygous *X. cortezi* and excluded these sites. This resulted in a final set of 995,825 ancestry informative sites which we used for downstream analyses, or a median of one marker every 240 basepairs across the 24 major chromosomes.

We tested the performance of this approach on 30 *X. cortezi* individuals we had not used in our initial filtering, 12 *X. birchmanni* individuals, 13 F_1 hybrids, 26 F_2 hybrids, and 5 BC_1 hybrids (backcrossed to *X. cortezi*) where we have clear expectations for true ancestry. Based on this analysis, we found that performance of the HMM approach was excellent (Fig. S9-S10).

Local ancestry inference and processing for downstream analysis

Using the ancestry informative sites described above, we next proceeded to analyze hybrid individuals from Chapulhuacanito and the Río Santa Cruz using the *ancestryinfer* pipeline. We inferred local ancestry for 291 individuals from Chapulhuacanito and 277 individuals from the Río Santa Cruz. Because previous analyses have indicated that *ancestryinfer* is not sensitive to priors for initial admixture time and admixture proportions [93], we set the prior for the genome-wide admixture proportion to 0.5 and the prior for the number of generations since initial admixture to 50. However, we repeated local ancestry inference for both populations following demographic inference using ABCreg (see next section) using priors inferred from this analysis for initial admixture time and admixture proportion. We found that our results were qualitatively unchanged (Table S9). For all analyses, we used a uniform recombination prior, set to the median per-basepair recombination rate in Morgans inferred for *X. birchmanni*.

For a number of downstream analyses, it was useful to convert posterior probabilities for different ancestry states into hard-calls. As we have previously, we used a posterior probability threshold of 0.9 or greater to assign an ancestry informative site to a given ancestry state (e.g. homozygous *X. birchmanni*, heterozygous for ancestry, or homozygous *X. cortezi*). Ancestry informative sites with lower than a 0.9 probability for any ancestry state were masked. We also filtered out sites that were covered in fewer than 25% of individuals. This resulted in 994,891 sites across the genome in Santa Cruz and 994,906 sites across the genome in Chapulhuacanito for downstream analysis. All local ancestry results are available on Dryad (Accession pending).

Consistent with previous work [53,62], a subset of the individuals we sequenced were nearly pure *X. birchmanni* (>98% of the genome derived from the *X. birchmanni* parent species). We identified and excluded these individuals from our dataset before examining patterns of local ancestry within the two hybrid populations, resulting in a dataset of 114 hybrid individuals from Chapulhuacanito and 276 from the Río Santa Cruz populations. We summarized minor parent ancestry across individuals by average ancestry hard-calls in non-overlapping windows of a range of sizes (e.g. 100 kb – 500 kb and 0.1 – 0.5 cM).

Demographic inference in the Chapulhuacanito and Santa Cruz populations

To inform our understanding of patterns of local ancestry along the genome in Chapulhuacanito and Santa Cruz, we wanted to better understand the likely demographic history of these populations. To do so, we used a regression-based Approximate Bayesian Computation or ABC approach with the software ABCreg [59]. We previously applied a similar approach to infer the likely demographic history of the Santa Cruz population [18] but repeat it for both populations here taking advantage of our larger empirical datasets and updated local ancestry inference pipeline. All simulations were performed in SLiM [58].

For each simulation, we drew each population demographic parameter from a uniform or log-uniform prior distribution, performed simulations, and calculated summary statistics for the simulation. We recorded the summary statistics and simulated parameters and compared them to the same statistics calculated from the real data. We modeled one chromosome 25 Mb in length with local recombination rates matching those observed on *X. birchmanni* chromosome 2. We used both global and local metrics as summary statistics (Table S13). We used the tree sequence recording functionality of SLiM to determine local ancestry of each individual in the hybrid population [94]. To perform each simulation, we used the following steps:

1. We initialized parental populations and formed a hybrid population between them. We determined the admixture proportion by drawing from a uniform prior distribution for the proportion of the genome derived from parent 1 (0.5-1). Similarly, we determined the hybrid population size for the simulation by drawing from a uniform prior ranging from 2-10,000 individuals.
2. We drew a migration rate from each parental species from a log uniform prior distribution (ranging from m=0-3% per generation based on previous results [94]).
3. We drew a time since initial admixture parameter from a uniform distribution ranging from 10-400 generations.
4. We performed the simulation for the number of generations drawn in step 3 above, implementing migration from the parental species each generation at the rate drawn in step 2.
5. We randomly sampled 69 and 242 individuals from the population to match the number of hybrid individuals sampled in Chapulhuacanito and Santa Cruz in 2021 and 2020 respectively and calculated summary statistics.

6. We generated summary statistics to compare to summary statistics calculated based on the real data.
7. We repeated this procedure until 150,000 simulations had been generated.
8. We ran the program ABCreg with the tolerance parameter set to 0.005.

To evaluate the validity of our approach, we tested how well this procedure worked to infer parameters for a simulated population with known history. We randomly sampled 100 simulations generated as described above and treated these simulations as if they were the real data. We calculated summary statistics and ran ABCreg as described above (excluding these simulations from the full ABCreg dataset). We then calculated the 95% quantile of the posterior distribution for each demographic parameter and asked how well this distribution captured the known parameters for the focal simulation. In general, the 95% quantile of the posterior distributions for each test set overlapped with the true value (Fig. S20). However, performance was poorer when we asked how often the true value fell in the 50% quantile of the posterior distribution produced by ABCreg, indicating that we should view MAP estimates as approximate estimates of the likely demographic history of each population (Fig. S20).

Repeatable patterns of minor parent ancestry as a function of genomic architecture

Using the LD-based recombination map described above, we evaluated evidence for a correlation between the local recombination rate in windows and minor parent ancestry in those same windows. As previously reported [53], we found strong correlations between local recombination rate and average minor parent ancestry in the Santa Cruz population, with minor parent ancestry being more common in regions of the genome with the highest local recombination rates. We repeated this analysis for the Chapulhuacanito population and replicated this pattern.

Because recombination events in *Xiphophorus* species appear to disproportionately localize to functionally dense regions of the genome (e.g. transcriptional start sites, CpG islands, and H3K4me3 peaks; [19,92]), we wanted to control for proximity to some of these elements in our analyses. We calculated the number of coding and conserved basepairs in each window and incorporated this into our analysis. We calculated the Spearman's partial correlation between recombination rate, minor parent ancestry, and coding (or conserved basepairs) across a range of window sizes (Table S5). We also repeated this analysis calculating average ancestry and the number of coding (or conserved) basepairs in windows of a particular genetic length (0.1-0.5 cM; Table S6; Supporting Information 11).

Cross-population repeatability in ancestry

The Santa Cruz and Chapulhuacanito populations occur in separate river systems and thus originated from independent hybridization events between *X. birchmanni* and *X. cortezi*. We wanted to understand the extent to which local ancestry between these two populations was correlated. Presumably, correlations that are observed (barring those due to technical artefacts) should be driven by shared sources of selection, either due to shared loci under selection or shared genomic architecture (e.g. similar recombination maps and locations of coding and conserved basepairs between species).

To evaluate this, we used a Spearman's correlation test implemented in R. We calculated these correlations in windows of a range of physical sizes (100 kb – 500 kb) and genetic sizes (0.1-0.5 cM). We performed each of these calculations thinning the data to retain only a single window every Mb or a single window every 1.5 cM (typically ~600 kb in *Xiphophorus*). This analysis should be conservative since admixture linkage disequilibrium decays to background levels over ~500 kb in Santa Cruz and Chapulhuacanito (Fig. S7). We found that the cross-population correlations in local ancestry in these analyses were surprisingly high (Table S7). Given this observation we sought to exclude several technical factors that might be artificially inflating this correlation.

Because power to infer ancestry will vary along the genome, we wanted to evaluate whether accounting for this power variation impacted the signal we observed. Certain regions of the genome have a higher density of ancestry informative sites between *X. birchmanni* and *X. cortezi*. We determined the median distance between ancestry informative sites (240 bp), and thinned markers such that in regions with higher marker frequency, we retained at most one marker per 240 bp. We also identified and excluded windows in which we have especially low power to infer ancestry (the number of ancestry informative sites fell in the lower 5% quantile of the genome-wide distribution).

We investigated the impact of removing other regions of the genome where we might expect to have a higher error rate in local ancestry inference. Analysis of our assemblies using seqtk telo [95] suggest that some of our chromosomes include assembled telomeric regions (Fig. S15). Since these regions may be especially challenging to analyze, we recalculated cross-population correlations excluding any region within 1 Mb of the end of a chromosome. We also generated a version of the ancestry informative sites excluding markers that overlapped with repetitive regions and recalculated cross-population correlations. We performed a number of other complementary analyses that are described in detail in Supporting Information 4-6.

Overall, our qualitative results were unchanged in each of the modifications described above and in the series of additional analyses described in Supporting Information 4-6 (Table S9). As a sanity check, we also performed analyses where we generated sub-populations from either Santa Cruz or Chapulhuacanito and asked about the observed correlations in ancestry when individuals truly originate from the same population. We also compared correlations in ancestry between samples from the same population over time, and from different sites in the same river. Reassuringly, all of these comparisons yielded correlations in ancestry that exceeded what we observed between independently formed populations at Santa Cruz and Chapulhuacanito (Table S7; Fig. S8).

Evidence of independent formation of Santa Cruz and Chapulhuacanito

While the Río Santa Cruz and Río San Pedro are separated by over 130 km of river miles, we wanted to perform additional analyses to confirm that they were independent in origin given the strong correlations in local ancestry that we observe across the two populations. To do so, we used a combination of approaches. First, we performed principal component analysis of the locations of observed ancestry transitions in Santa Cruz and Chapulhuacanito. If these populations formed and evolved independently, we would expect that observed ancestry transitions (which reflect recombination events in the ancestors of each hybrid individual) would occur largely in different locations across the two populations.

To generate a dataset for principal component analysis, we first identified the approximate locations of ancestry transitions for each hybrid individual in our datasets from Santa Cruz and Chapulhuacanito, defined as the interval over which the posterior probability changes from 0.9 posterior probability for one ancestry state to 0.9 for another ancestry state. Ancestry transitions that were supported by flanking segments of <5 kb were removed. Qualitatively, this removed transitions where the ancestry state switched and then immediately reverted, which we hypothesized might be more likely to be errors. Once ancestry transitions have been identified, we binned the genome into windows of 0.5 cM, and for a given individual recorded a 1 if there was a transition observed in that window and a zero if there was not. While most ancestry transitions were well resolved (mean 13.6 kb), some spanned multiple windows, so we used the midpoint as the location of the ancestry transition for these purposes. We ran a principal component analysis of this matrix in R.

While the variation in ancestry transition locations suggested broadly different histories of recombination in the Santa Cruz and Chapulhuacanito populations, this pattern could also be consistent with an initial period of shared history followed by vicariance. Thus, as a complementary approach, we quantified how frequently the locations of ancestry transitions were shared between pairs of individuals in the Santa Cruz and Chapulhuacanito populations, compared to expectations by chance. Shared ancestry transitions, reflecting ancestral recombination events, could occur by chance due to recombination hotspots or poor resolution of the precise locations of recombination events, but an excess of shared transitions is likely to reflect shared ancestors and thus a shared population history. For both the real and simulated data, we excluded ancestry transitions that were poorly resolved (>250 kb in length) from our analysis. For the real data, we quantified how frequently the intervals of ancestry transitions overlapped in pairs of individuals from the two populations using bedtools [96], treating ≥1 basepair shared as an overlap. To generate simulated data, we performed a series of steps. We first excluded windows where ancestry for one parental species had fixed in the real data, as these regions of the genome cannot contain ancestry transitions in the real data. For each individual, we iterated through all of the ancestry transitions observed, randomly sampling a new location for each ancestry transition, weighted by the *X. birchmanni* recombination map (summarized in 100 kb windows). Within that randomly selected window, we used R's runif function to identify a start position of the recombination interval and set the stop position based on the interval length. We repeated this until all ancestry transitions for an individual had been assigned a random position weighted by the recombination map and repeated this process for all individuals. Next, we quantified the overlap of ancestry transitions in pairs of simulated Santa Cruz and Chapulhuacanito individuals relative to the real data. These results are shown in Fig. 2A.

We also collected high-coverage whole genome sequencing data (>20X) for 3 hybrid individuals from each population and 3 pure *X. birchmanni* individuals found in the same populations (i.e. individuals inferred to derive >98% of their genome from *X. birchmanni*). For these individuals, we called variants throughout the genome as described above for inference of LD-based recombination maps (mapping and variant calling were performed using both references independently, and results were qualitatively similar). Using this variant information, we performed a principal component analysis on the observed variants genome-wide in the individuals collected from Santa Cruz and Chapulhuacanito as well as previously collected high coverage data from source populations of *X. birchmanni* (Coacuilco; [30]) and *X. cortezi* (Huichihuyan and Puente de Huichihuyan; [53,62]).

Because hybrids combine the genomes of the *X. birchmanni* and *X. cortezi* individuals that contributed to the hybridization event, we were also interested in subsetting these regions of the genome and analyzing them separately. To do so, we conducted local ancestry inference on the six hybrid individuals as described above, and identified regions where we had high confidence that all six hybrid individuals in our dataset were homozygous *X. cortezi* or homozygous *X. birchmanni*. We extracted the variants in these segments from hybrids and from the corresponding parental species plink files (i.e. from *X. cortezi* individuals for analysis of homozygous *X. cortezi* ancestry tracts). We performed a separate PCA on the *X. cortezi* and *X. birchmanni* derived regions in hybrids. If Santa Cruz and Chapulhuacanito somehow shared a population history, we would expect these regions to cluster closely together and potentially overlap in a principal component analysis. See Supporting Information 2 for a more detailed discussion of these results and their implications.

To more closely evaluate possible relatedness in these ancestry tracts, we used the program GCTA to generate a genetic relatedness matrix for these six hybrid individuals [60]. We performed separate analyses for the *X. cortezi* and *X. birchmanni* ancestry tracts. As above, we only analyzed regions where all six hybrid individuals were homozygous *X. cortezi* or *X. birchmanni* in that region respectively. We included data from the relevant parental populations for comparison. We found that all hybrid individuals from the same population were inferred to have some degree of relatedness based on analysis of both *X. cortezi* and *X. birchmanni* derived ancestry tracts, but all values for cross-population comparisons were negative (Fig. S21).

Since most of the genome of individuals in the two hybrid populations is derived from *X. cortezi*, we performed an additional analysis focusing on *X. cortezi* ancestry tracts in the high coverage individuals. Reasoning that populations derived from distinct source populations and with independent demographic histories should harbor distinct frequencies of genetic variants, we performed a "mismatch" analysis. We subset our data to focus only on regions that were homozygous *X. cortezi* across all six high coverage hybrid individuals. For each pair of individuals in our dataset, we counted each site where individual 1 was homozygous for one allele and individual 2 was homozygous for another (within *X. cortezi* ancestry tracts). We counted the total number of these sites along the genome, divided by the total number of sites that passed quality thresholds in both individuals (within *X. cortezi* ancestry tracts), and treated this as our mismatch statistic. We compared this mismatch statistic within-populations versus between-populations (Fig. 2E).

Spatial scale of cross-population correlations in ancestry

In the absence of selection or genetic drift, the ancestry proportion in a hybrid population would remain uniform along a chromosome in a hybrid population. In real populations, ancestry varies along the genome due to the combined effects of recombination with genetic drift, selection, and repeated admixture events. The spatial scale of ancestry variation along the genome holds important information about the timing of demographic and selective events, since recombination progressively shortens ancestry tracts across generations. We took advantage of the recent application of the Discrete Wavelet Transform to decompose correlations between genomic signals into independent components associated with different spatial genomic scales [33]. Briefly, the method transforms a signal measured along the genome (e.g. ancestry) into a set of coefficients that measure *changes* in the signal between adjacent windows at different locations and with windows of different sizes. The wavelet transform is performed on two signals separately, and the correlation between the coefficients at a given scale for the two signals are weighted by the variance at that scale (also determined from the wavelet coefficients) to give the contribution of each scale to the overall correlation. This approach offers the advantage that correlations across scales carry independent information, in contrast to traditional window-based analyses used elsewhere in the manuscript where results across different window sizes are confounded due to the nestedness of windows of different sizes.

As this analysis requires evenly spaced measurements along a chromosome, we first interpolated admixture proportions within diploid individuals to a 1 kb grid for each chromosome, then averaged across individuals to obtain the interpolated sample admixture proportion. We used the inferred recombination maps described above to obtain estimates of recombination in windows centered on the interpolated ancestry measurements. We applied a threshold to recombination values of $\rho \geq 0.005$ (corresponding to 4% of the genome) which we found improved the strength of correlation between genetic lengths of chromosome inferred from the LD map vs. from an F_2 linkage map.

We used the R package *gnomwav* [33] to estimate wavelet correlations between signals (minor parent ancestry between populations, recombination vs. minor parent ancestry) at a series of genomic scales for each, with the smallest scale being the resolution of interpolation, and the largest scale corresponding to variation in signals occurring over roughly half of a chromosome. Wavelet correlations were averaged across chromosomes and error bars were obtained from a weighted jackknife procedure following [33]. To obtain the contribution of each scale to the overall correlation, we weighted correlations by the wavelet variances as described in [33]. We ran these analyses for interpolation distances of 1 kb and 32 kb.

Shared minor parent deserts and islands

We were interested in identifying regions that were likely under selection in both *X. cortezi* × *X. birchmanni* hybrid populations. Guided by the results of simulations, we used an ad-hoc approach to identify regions with shared patterns of unusual ancestry across the two populations (see Supporting Information 8). We first identified ancestry informative sites where the minor or major parent ancestry fell in the lower 5% tail of genome-wide ancestry. We then selected the 0.05 cM window that overlapped this ancestry informative site and confirmed that the broader region fell within the lower 10% tail of genome-wide ancestry. We expanded out in the 5' and 3' directions in windows of 0.05 cM from this focal window until we reached a window on each edge that exceeded the 10% ancestry quantile. We treated this interval as an estimate of the boundary of the minor parent ancestry desert or island.

Because this approach may prematurely truncate ancestry deserts and islands (particularly in scenarios with error), in a separate analysis we merged any of deserts (or islands) that fell within 50 kb of each other. We filtered these merged regions to remove any regions with fewer than 10 ancestry informative sites, with fewer than 10 single nucleotide polymorphisms present in the recombination map, or that were less that 10 kb in length.

By defining deserts and islands in 0.05 cM windows we could easily overlap these regions between different populations and determine how many are shared between sampling sites. This allowed us to define regions that have shared ancestry patterns between Chapulhuacanito and Santa Cruz despite their independent origin. To compare the observed number of shared ancestry deserts and islands to what we would expect by chance, given the overall patterns of ancestry variation along the genome in the two populations, we permuted the data in 0.05 cM windows and asked how frequently ancestry deserts and islands were identified as being shared in *X. birchmanni* × *X. cortezi* populations, as we had with the real data. We repeated this procedure 1000 times. Based on these permutations, we found that few shared minor parent ancestry deserts or islands were expected by chance (Fig. 4A).

Since ancestry in a given window is strongly correlated with ancestry in the neighboring windows, especially at smaller spatial scales, we also wanted to performed permutations that preserved this ancestry structure. Specifically, for Chapulhuacanito, we shifted the window labels of ancestry summarized in 0.05 cM windows by 12.5 cMs, and we asked whether any windows that were major (or minor) parent ancestry outliers in the shifted data overlapped with the ancestry deserts (or islands) identified in Santa Cruz (using the same criteria as in the real data). We repeated this procedure 132 times to fully tile the whole genome. Consistent with the naïve permutation approach, we found that few minor parent ancestry outliers in *X. birchmanni* × *X. cortezi* hybrid populations overlapped minor parent deserts identified in the Santa Cruz hybrid population by chance (Fig. S22).

We were interested in whether any of the shared deserts or islands between Chapulhuacanito and Santa Cruz were also ancestry outliers in *X. birchmanni* × *X. malinche* populations. Given the complexity of simulations preserved LD structure across the five hybrid populations we wanted to evaluate, we simply performed the naïve simulations in 0.05 cM windows described above. Based on these permutations, we found that few minor parent ancestry outliers in *X. birchmanni* × *X. malinche* hybrid populations are expected to overlap minor parent deserts (or islands) identified in *X. birchmanni* × *X. cortezi* hybrid populations by chance.

Time series analysis

We were interested in understanding how ancestry at minor parent deserts and islands has changed over time. We focus this analysis on Chapulhuacanito due to insufficient sampling over time from the Río Santa Cruz (both in terms of numbers of hybrids sampled and number of sampling years available). The first samples we have access to from Chapulhuacanito are from 2003, approximately 40 generations ago. However, based on our demographic inference, this population likely underwent ~100 generations of evolution between initial hybridization and our first sampling year, meaning that even in our earliest samples we are evaluating ancestry in a late-stage hybrid population.

We focused our analysis on deserts and islands identified in 2021, but our results were qualitatively similar when ascertainment was performed in other years. Using the coordinates determined in 2021, we calculated average minor parent ancestry in the same region in each of the other years sampled (2002, 2006, and 2017). For minor parent islands, which showed greater levels of fluctuation in ancestry over time (see Results), we used a linear model implemented in R to test for a significant relationship between year and minor parent ancestry.

Figures

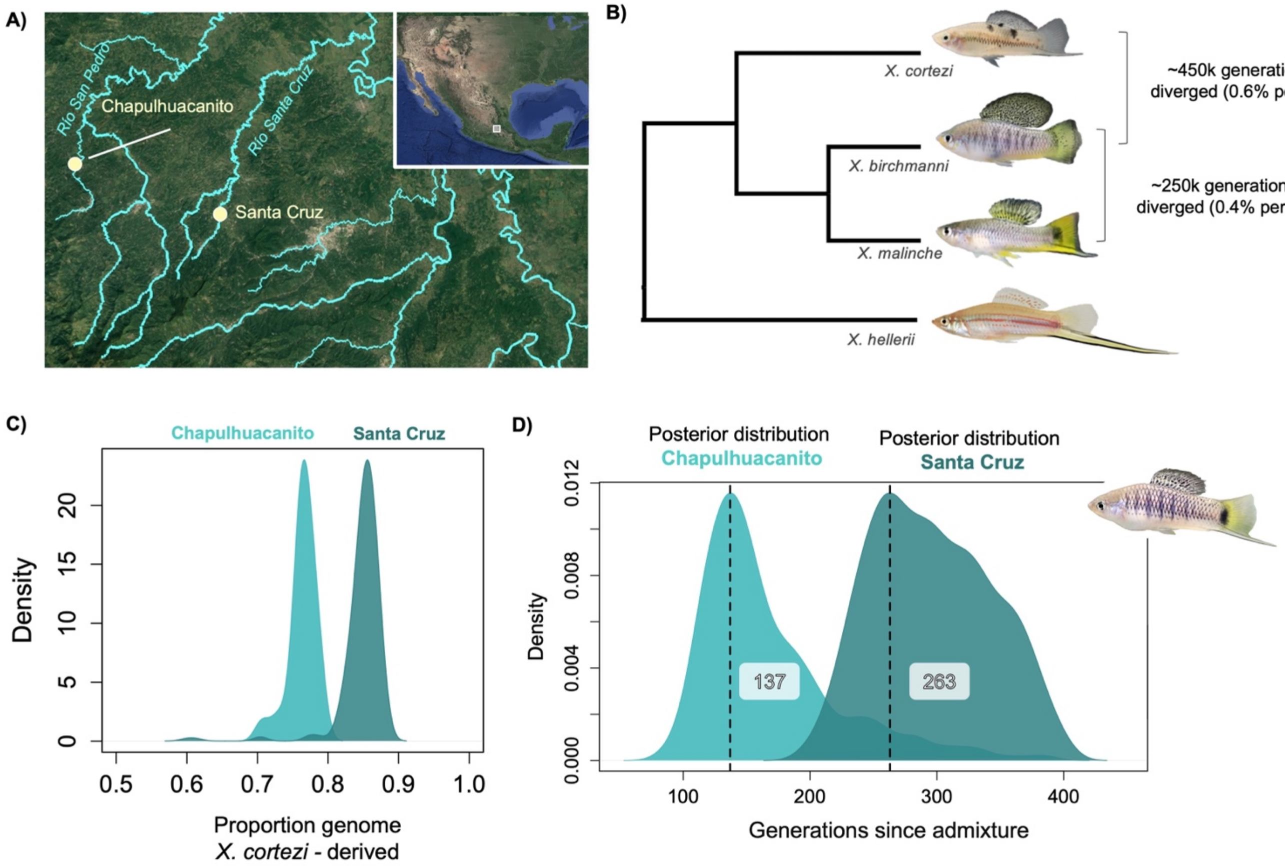

Fig. 1. A) Map of collection sites of *X. birchmanni* × *X. cortezi* hybrids in two different river drainages. **B)** Phylogenetic relationships between *X. birchmanni, X. malinche,* and *X. cortezi* and estimated divergence times from previous work. **C)** Distributions of inferred admixture proportions from samples from Chapulhuacanito in 2021 and Santa Cruz in 2020. Both populations derive the majority of their genomes from the *X. cortezi* parental species, but Chapulhuacanito has substantially more ancestry derived from the *X. birchmanni* parental species. **D)** Results of approximate Bayesian computation approaches inferring the population history of Chapulhuacanito and Santa Cruz indicate that admixture likely began at different times in the two populations. The dashed line and numbers indicate the maximum a posteriori estimate of the time since initial admixture in both populations. Inset show male hybrid collected from the Santa Cruz population. Other results from ABC analyses can be found in Fig. S1 and Table S3.

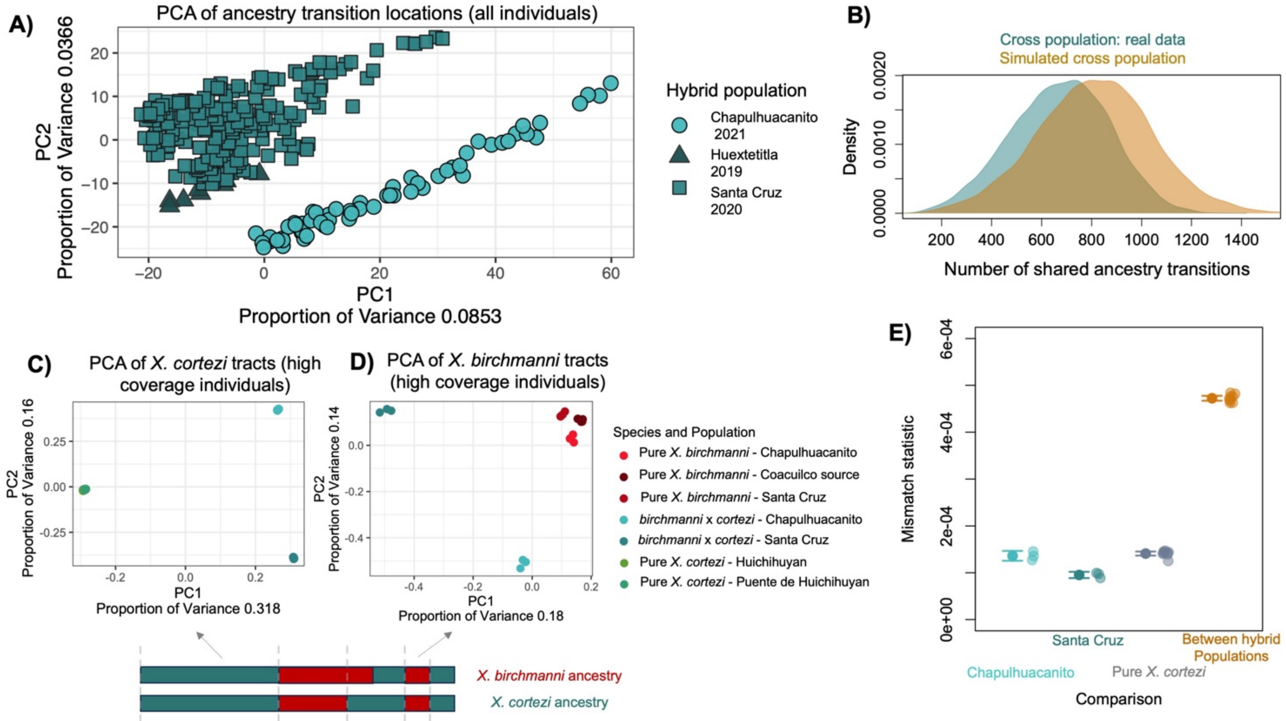

Fig. 2. A) PCA analysis of the locations of ancestry transitions indicates that the Santa Cruz and Chapulhuacanito populations have distinct recombination histories, while other individuals from the Santa Cruz drainage (the "Huextetitla" population) cluster with Santa Cruz. **B**) Using simulations, we also find that the number of shared ancestry transitions across populations (i.e. cases where ancestry transitions occur in the same physical location along the genome) is comparable to that expected by chance. Blue distribution shows the number of overlapping ancestry transitions across all pairs of individuals in Santa Cruz and Chapulhuacanito, and orange distribution shows the results of simulations using the *X. birchmanni* recombination map (see Methods). Importantly, the shared ancestry transitions in the two populations do not exceed the number expected by chance. **C) & D)** We also evaluated patterns of genetic variation using SNPs in high coverage individuals, subsetting the data to analyze tracts that are homozygous for *X. cortezi* **(C)** or *X. birchmanni* **(D)** in hybrid individuals**.** Schematic of diploid hybrid individual below the plots shows our approach for selecting regions for PCA analysis based on local ancestry in the six hybrid individuals. Tracts from individuals in different hybrid populations separate from each other and the parental populations in PCA space (**C, D**). The sympatric *X. birchmanni* populations (**D**) found in both sites are genetically distinct from each other and the Coacuilco reference population but modestly so. See Supporting Information 2 for a more in-depth discussion of these results. **E)** Results of a "mismatch" analysis for comparisons of *X. cortezi* ancestry tracts within the six high coverage hybrid individuals and in pure *X. cortezi* source populations. We counted the number of sites where pairs of individuals from Santa Cruz and Chapulhuacanito were homozygous for different SNPs over the total number of sites that passed our quality thresholds in each comparison (see Methods). We found striking differences for within population versus between population pairs. We repeated the same analysis for two *X. cortezi* populations on the Río Huicihuyan for comparison. Semi-transparent points show the results of each comparison, bars and whiskers show the mean ± 2 standard errors.

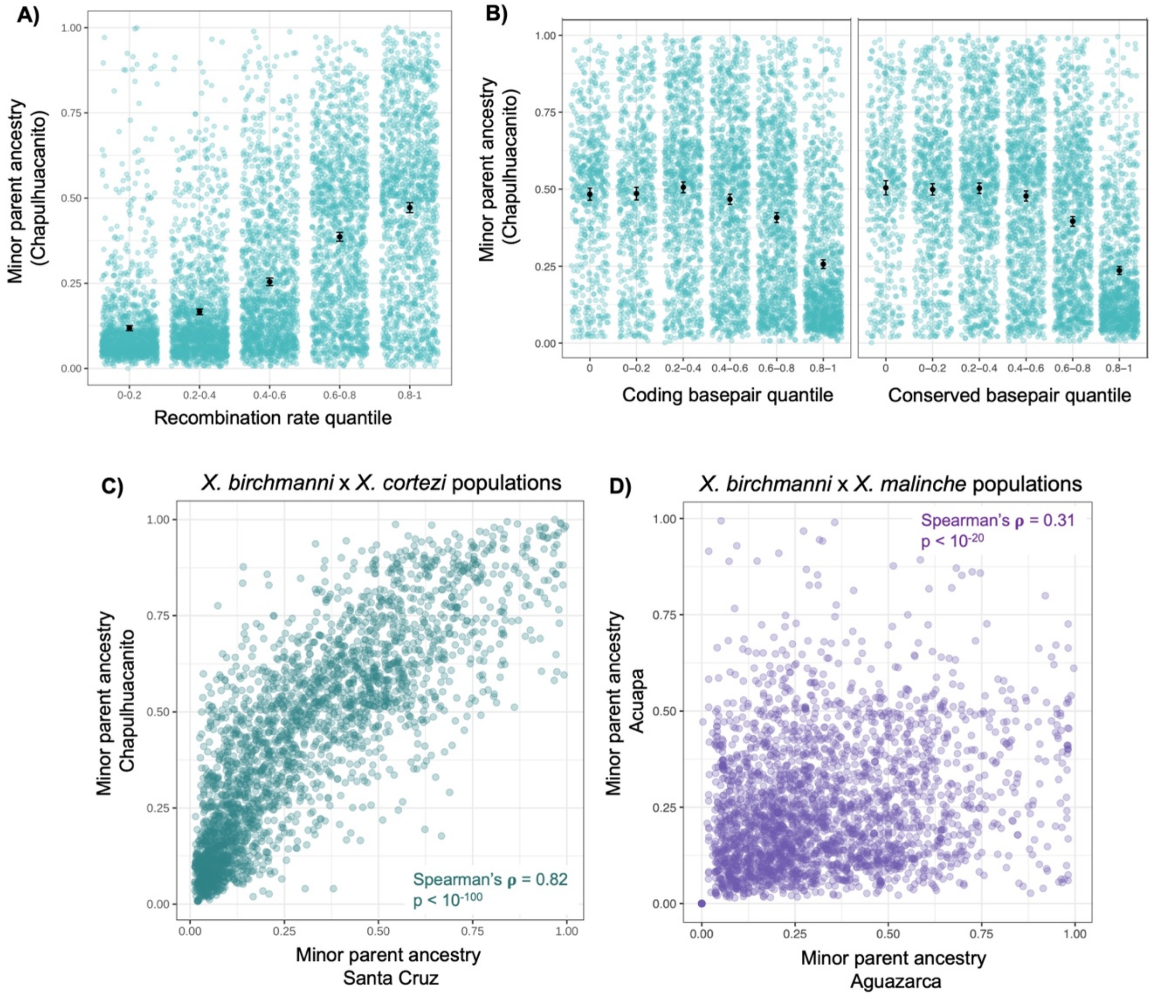

Fig. 3. A) Minor parent ancestry in the Chapulhuacanito population is strongly correlated with the local recombination rate. Here, ancestry and recombination are summarized in 250 kb windows (see also Fig. 4C for wavelet-based analysis). **B)** After accounting for the strong effect of recombination rate by summarizing ancestry in 0.25 cM windows, we also find that minor parent ancestry is depleted in regions of the genome linked to large numbers of coding or conserved basepairs. We previously reported similar results for the Santa Cruz population for both recombination rate and functional basepair density [53], and for *X. birchmanni* × *X. malinche* hybrid populations [30]. **C)** Average minor parent ancestry is strikingly correlated across the Santa Cruz and Chapulhuacanito populations. Shown here are analyses of 0.5 cM windows (Spearman's ρ = 0.82, $p < 10^{-100}$); these results are observed across all spatial scales tested in both physical and genetic distance (see Fig. S8, Table S7-S8). **D)** By contrast, minor parent ancestry is substantially less correlated between two *X. birchmanni* × *X. malinche* hybrid populations. Shown here are analyses of 0.5 cM windows (Spearman's ρ = 0.31, $p < 10^{-20}$). For additional comparisons of ancestry in *X. birchmanni* × *X. malinche* hybrid populations, see [30,53].

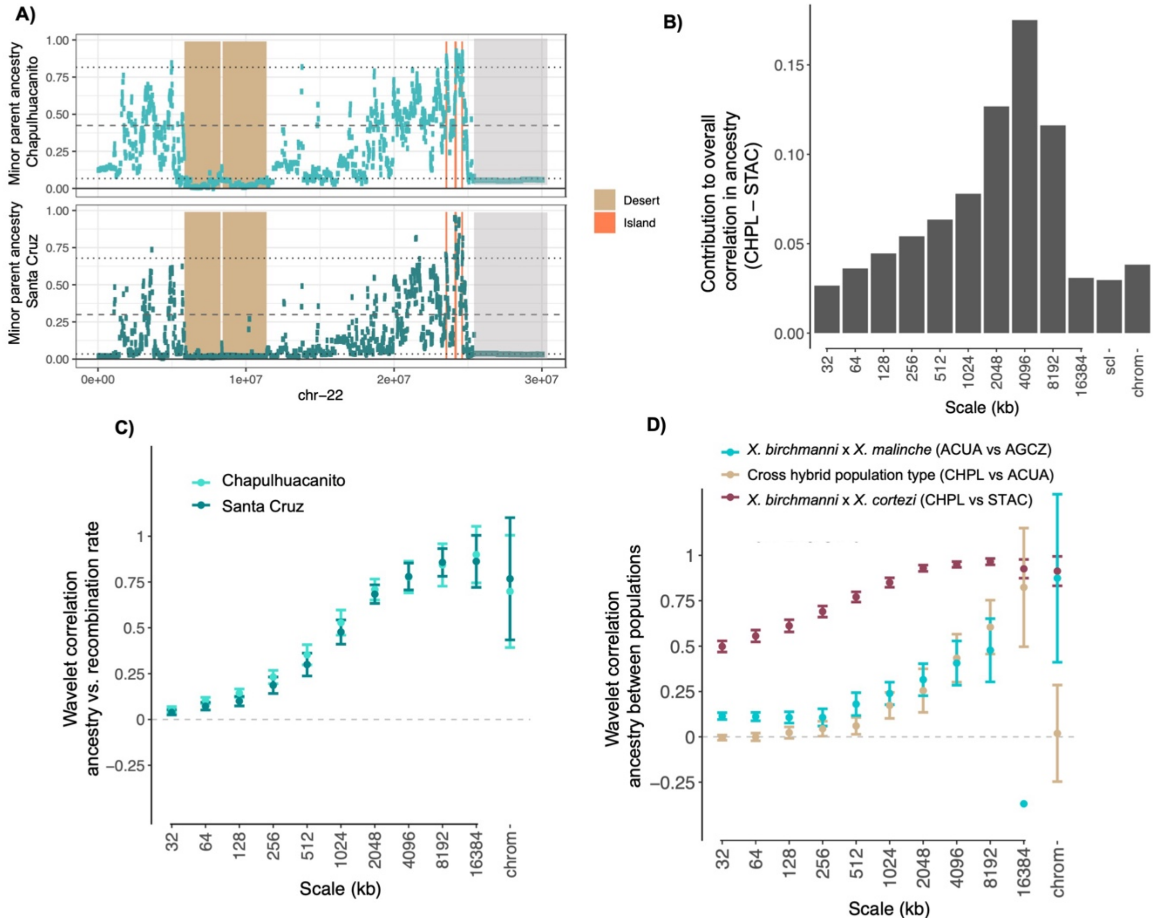

Fig. 4. A) Example of large shared minor parent ancestry deserts identified on chromosome 22 (tan) as well as shared minor parent ancestry islands (peach) in Chapulhuacanito and Santa Cruz. Note also large regions of low minor parent ancestry at ~25-30 Mb found across both populations that do not pass the threshold for being designated as shared ancestry deserts (light gray; in this case the region exceeds the 5% threshold for Santa Cruz). Dashed lines indicate the average ancestry genome-wide and dotted lines represent lower and upper 10% quantiles of minor parent ancestry. **B**) Spatial wavelet decomposition of the overall Pearson correlation between inferred minor parent ancestry in Chapulhuacanito vs. Santa Cruz (CHPL vs STAC) measured at a resolution of 1 kb. The contribution of a given spatial scale is a weighted correlation of wavelet coefficients for the two signals at that scale, weighted by the portion of the total variance attributable to that scale (see Methods). Correlations among chromosome means also contribute (chrom), as well as a leftover component (scl) due to irregularity of chromosome lengths. **C**) Wavelet correlations between inferred minor parent ancestry and recombination rate for both Chapulhuacanito and Santa Cruz populations. Note that here correlations at each scale are not weighted by variances at the corresponding scales. Points are weighted averages across chromosomes with error bars representing 95% jackknife confidence intervals. **D**) Wavelet correlations between inferred minor parent ancestry proportion in cross population comparisons between hybrids derived from the same hybridizing pair (CHPL vs. STAC - *X. birchmanni* × *X. cortezi*; ACUA vs. AGCZ – *X. birchmanni* × *X. malinche*) and from different hybridizing pairs (CHPL - *X. birchmanni* × *X. cortezi* vs. ACUA - *X. birchmanni* × *X. malinche* hybrids at the Acuapa site). Points are weighted averages across chromosomes with error bars representing

95% jackknife confidence intervals. For visualization, we omit the confidence interval for the wavelet correlation of ancestry in the two *X. birchmanni* × *X. malinche* populations (ACUA vs. AGZC) at the largest scale, since it is large and overlaps with zero. Note that the identity of the minor parent species differs across hybrid population types (*X. birchmanni* in Chapulhuacanito and Santa Cruz and *X. malinche* in Acuapa and Aguazarca).

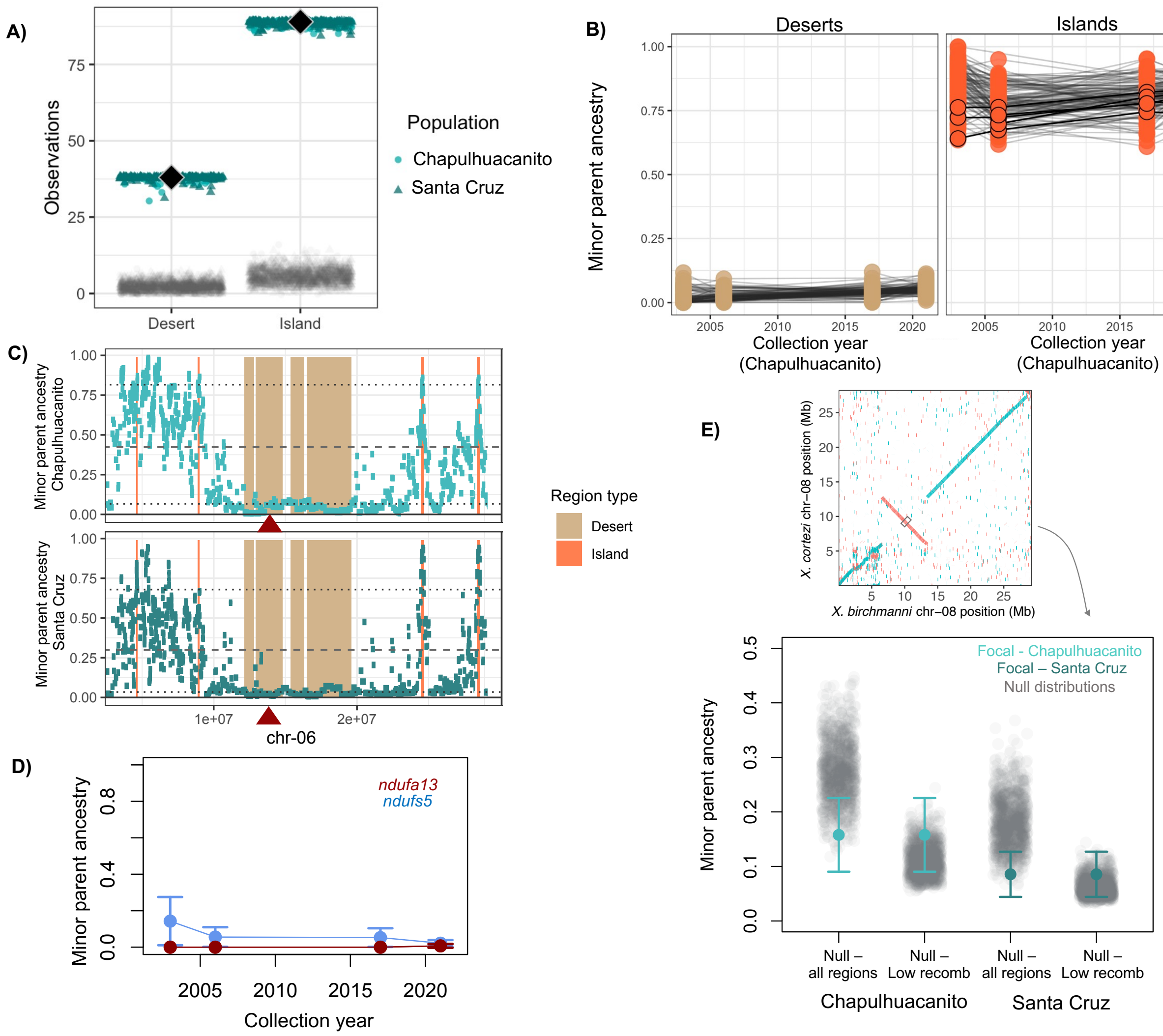

Fig. 5. A) Shared minor parent ancestry deserts and islands in *X. birchmanni* × *X. cortezi* populations (Chapulhuacanito and Santa Cruz - colored points) show a much greater overlap than expected by chance (gray points, see methods). Black diamonds show the observed number of shared minor parent deserts or islands across the two populations and colored points show the results of jack-knife bootstrapping the data from each population in 10 cM blocks (circles – bootstrap results from Chapulhuacanito, triangles – bootstrap results from Santa Cruz). **B)** Ancestry of shared minor parent deserts (left) and islands (right) through time in the Chapulhuacanito dataset. Points show ancestry at individual deserts or islands for each sampling year, and lines connect results for a given desert or island across years. Shared minor parent ancestry deserts were largely fixed by the onset of genetic monitoring of these populations approximately 40 generations ago. Islands also tended to have high minor parent frequency at the onset of sampling, but several islands do change significantly in minor parent ancestry over the sampling period (Table S10). Islands that increase significantly in minor parent ancestry through time are outlined in black. **C)** One shared minor parent ancestry desert on chromosome 6 overlaps with a known mitonuclear incompatibility generated by combining the *X. cortezi* mitochondria with homozygous *X. birchmanni* ancestry at *ndufa13* [55,61]. Local ancestry along chromosome 6 in Chapulhuacanito is shown in the top plot and local ancestry along chromosome

6 in Santa Cruz is shown in the bottom plot. The locations of shared deserts and islands are highlighted in tan and peach respectively. The location of *ndufa13* is indicated by the red triangle. Dashed lines indicate the average ancestry genome-wide and dotted lines represent lower and upper 10% quantiles of minor parent ancestry. **D**) Both *ndufa13* and another gene involved in mitonuclear incompatibility between *X. cortezi* and *X. birchmanni*, *ndufs5*, are nearly fixed for major parent ancestry at the onset of our time-series sampling. **E)** With our new long-read reference assemblies, we evaluated minor parent ancestry at the center of inversions (focusing on inversions >100 kb) that differentiated *X. birchmanni* and *X. cortezi* in the two hybrid populations. Example alignment of a large inversion identified on chromosome 8 is shown in the inset. For each inversion, we sampled ancestry in a 50 kb window that overlapped with the center of the inversion (schematically shown by the gray rectangle in the inset). We found that minor parent ancestry was modestly depleted in the two hybrid populations (colored points and whiskers) at inversions compared to the randomly sampled regions of the genome (gray points – “all regions”). However, when we generated null datasets only from regions of the genome with low recombination rates (lowest 5% quantile of recombination rate) we found that inversions did not show unusually high depletion of minor parent ancestry. This suggests that depletion of minor parent ancestry at inversions may be driven by reduced recombination in these regions in hybrids.

Acknowledgements

We thank Yaniv Brandvain, Kelley Harris, and members of the Schumer lab and Coop lab for helpful comments on earlier versions of this manuscript. We are grateful to the Federal Government of Mexico for permission to collect fish. Stanford University and the Stanford Research Computing Center provided computational support for this project. This work was supported by a Knight-Hennessy Scholars fellowship and NSF GRFP to B. Moran, NSF GRFP to T. Dodge, a CEHG fellowship and NSF PRFB (2010950) to Q. Langdon, and Stanford Science Fellowship to S. M. Aguillon. This work was supported by a Hanna H. Gray Fellowship, Freeman-Hrabowski Fellowship, Sloan Fellowship, and NIH grant 1R35GM133774 to M. Schumer, and a Stanford Tinker award to M. Schumer and C. Gutierrez Rodríguez. This research was also supported by funding from the National Science Foundation (IBN 9983561) and Ohio University (Research Incentive) to MRM.

References

1. Taylor SA, Larson EL. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. Nature Ecology & Evolution. 2019;3: 170–177. doi:10.1038/s41559-018-0777-y

2. Langdon QK, Peris D, Eizaguirre JI, Opulente DA, Buh KV, Sylvester K, et al. Postglacial migration shaped the genomic diversity and global distribution of the wild ancestor of lager-brewing hybrids. PLOS Genetics. 2020;16: e1008680. doi:10.1371/journal.pgen.1008680

3. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and Introgression between Mimulus nasutus and Mimulus guttatus. PLOS Genetics. 2014;10: e1004410. doi:10.1371/journal.pgen.1004410

4. Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, et al. Widespread introgression across a phylogeny of 155 Drosophila genomes. Current Biology. 2022;32: 111-123.e5. doi:10.1016/j.cub.2021.10.052

5. Calfee E, Agra MN, Palacio MA, Ramírez SR, Coop G. Selection and hybridization shaped the rapid spread of African honey bee ancestry in the Americas. PLOS Genetics. 2020;16: e1009038. doi:10.1371/journal.pgen.1009038

6. Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, O'Brien JE, et al. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res. 2008;18: 67–76. doi:10.1101/gr.6757907

7. Taylor SA, White TA, Hochachka WM, Ferretti V, Curry RL, Lovette I. Climate-Mediated Movement of an Avian Hybrid Zone. Current Biology. 2014;24: 671–676. doi:10.1016/j.cub.2014.01.069

8. Rosenthal GG, de la Rosa Reyna XF, Kazianis S, Stephens MJ, Morizot DC, Ryan MJ, et al. Dissolution of sexual signal complexes in a hybrid zone between the swordtails Xiphophorus birchmanni and Xiphophorus malinche (Poeciliidae). Copeia. 2003;2003: 299–307. doi:10.1643/0045-8511(2003)003[0299:dossci]2.0.co;2

9. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328: 710–722. doi:10.1126/science.1188021

10. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Current Biology. 2016;26: 1241–1247. doi:10.1016/j.cub.2016.03.037

11. Vernot B, Akey JM. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science. 2014;343: 1017–1021. doi:10.1126/science.1245938

12. de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science. 2016;354: 477–481. doi:10.1126/science.aag2602

13. Tung J, Barreiro LB. The contribution of admixture to primate evolution. Current Opinion in Genetics & Development. 2017;47: 61–68. doi:10.1016/j.gde.2017.08.010

14. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507: 354–357. doi:10.1038/nature12961

15. Juric I, Aeschbacher S, Coop G. The Strength of Selection against Neanderthal Introgression. PLOS Genetics. 2016;12: e1006340. doi:10.1371/journal.pgen.1006340

16. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple Deeply Divergent Denisovan Ancestries in Papuans. Cell. 2019;177: 1010-1021.e32. doi:10.1016/j.cell.2019.02.035

17. Telis N, Aguilar R, Harris K. Selection against archaic hominin genetic variation in regulatory regions. Nat Ecol Evol. 2020;4: 1558–1566. doi:10.1038/s41559-020-01284-0

18. Langdon QK, Powell DL, Kim B, Banerjee SM, Payne C, Dodge TO, et al. Predictability and parallelism in the contemporary evolution of hybrid genomes. PLOS Genetics. 2022;18: e1009914. doi:10.1371/journal.pgen.1009914

19. Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science. 2018;360: 656. doi:10.1126/science.aar3684

20. Clark A, Dunham MJ, Akey JM. The genomic landscape of Saccharomyces paradoxus introgression in geographically diverse Saccharomyces cerevisiae strains. bioRxiv; 2022. p. 2022.08.01.502362. doi:10.1101/2022.08.01.502362

21. Aeschbacher S, Selby JP, Willis JH, Coop G. Population-genomic inference of the strength and timing of selection against gene flow. Proceedings of the National Academy of Sciences. 2017;114: 7061–7066. doi:10.1073/pnas.1616755114

22. Kenney AM, Sweigart AL. Reproductive isolation and introgression between sympatric Mimulus species. Mol Ecol. 2016;25: 2499–2517. doi:10.1111/mec.13630

23. Corbett-Detig R, Nielsen R. A Hidden Markov Model Approach for Simultaneously Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data in Samples of Arbitrary Ploidy. PLOS Genetics. 2017;13: e1006529. doi:10.1371/journal.pgen.1006529

24. Nouhaud P, Martin SH, Portinha B, Sousa VC, Kulmuni J. Rapid and predictable genome evolution across three hybrid ant populations. PLOS Biology. 2022;20: e3001914. doi:10.1371/journal.pbio.3001914

25. Martin SH, Davey JW, Salazar C, Jiggins CD. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLOS Biology. 2019;17: e2006288. doi:10.1371/journal.pbio.2006288

26. Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, et al. Genomic architecture and introgression shape a butterfly radiation. Science. 2019;366: 594–599. doi:10.1126/science.aaw2090

27. Vilgalys TP, Fogel AS, Anderson JA, Mututua RS, Warutere JK, Siodi IL, et al. Selection against admixture and gene regulatory divergence in a long-term primate field study. Science. 2022;377: 635–641. doi:10.1126/science.abm4917

28. Moran BM, Payne C, Langdon Q, Powell DL, Brandvain Y, Schumer M. The genomic consequences of hybridization. Wittkopp PJ, editor. eLife. 2021;10: e69016. doi:10.7554/eLife.69016

29. Harris K, Nielsen R. The Genetic Cost of Neanderthal Introgression. Genetics. 2016;203: 881–891. doi:10.1534/genetics.116.186890

30. Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science. 2018;360: 656. doi:10.1126/science.aar3684

31. Nachman MW, Payseur BA. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos Trans R Soc Lond B Biol Sci. 2012;367: 409–421. doi:10.1098/rstb.2011.0249

32. Veller C, Edelman NB, Muralidhar P, Nowak MA. Recombination and selection against introgressed DNA. Evolution. 2023;77: 1131–1144. doi:10.1093/evolut/qpad021

33. Groh JS, Coop G. The temporal and genomic scale of selection following hybridization. bioRxiv; 2023. p. 2023.05.25.542345. doi:10.1101/2023.05.25.542345

34. Thompson KA, Brandvain Y, Coughlan J, Delmore KE, Justen H, Linnen CR, et al. The ecology of hybrid incompatibilities. Cold Spring Harbor Perspectives on Speciation. 2024.

35. Orr HA. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics. 1995;139: 1805–1813.

36. Orr HA, Turelli M. The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. Evolution. 2001;55: 1085–1094. doi:10.1111/j.0014-3820.2001.tb00628.x

37. Moyle LC, Nakazato T. Hybrid Incompatibility “Snowballs” Between Solanum Species. Science. 2010;329: 1521–1523. doi:10.1126/science.1193063

38. Moyle LC, Payseur BA. Reproductive isolation grows on trees. Trends in Ecology & Evolution. 2009;24: 591–598. doi:10.1016/j.tree.2009.05.010

39. Wang RJ, White MA, Payseur BA. The Pace of Hybrid Incompatibility Evolution in House Mice. Genetics. 2015;201: 229–242. doi:10.1534/genetics.115.179499

40. Matute DR, Butler IA, Turissini DA, Coyne JA. A Test of the Snowball Theory for the Rate of Evolution of Hybrid Incompatibilities. Science. 2010;329: 1518–1521. doi:10.1126/science.1193440

41. Thompson KA, Peichel CL, Rennison DJ, McGee MD, Albert AYK, Vines TH, et al. Analysis of ancestry heterozygosity suggests that hybrid incompatibilities in threespine stickleback are environment dependent. PLOS Biology. 2022;20: e3001469. doi:10.1371/journal.pbio.3001469

42. Arnegard ME, McGee MD, Matthews B, Marchinko KB, Conte GL, Kabir S, et al. Genetics of ecological divergence during speciation. Nature. 2014;511: 307–311. doi:10.1038/nature13301

43. Hajdinjak M, Fu Q, Hübner A, Petr M, Mafessoni F, Grote S, et al. Reconstructing the genetic history of late Neanderthals. Nature. 2018;555: 652–656. doi:10.1038/nature26151

44. Geza E, Mugo J, Mulder NJ, Wonkam A, Chimusa ER, Mazandu GK. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Brief Bioinform. 2018;20: 1709–1724. doi:10.1093/bib/bby044

45. Martin SH, Davey JW, Jiggins CD. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Molecular Biology and Evolution. 2015;32: 244–257. doi:10.1093/molbev/msu269

46. Racimo F, Marnetto D, Huerta-Sánchez E. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. Molecular Biology and Evolution. 2017;34: 296–317. doi:10.1093/molbev/msw216

47. Runemark A, Trier CN, Eroukhmanoff F, Hermansen JS, Matschiner M, Ravinet M, et al. Variation and constraints in hybrid genome formation. Nat Ecol Evol. 2018;2: 549–556. doi:10.1038/s41559-017-0437-7

48. Chaturvedi S, Lucas LK, Buerkle CA, Fordyce JA, Forister ML, Nice CC, et al. Recent hybrids recapitulate ancient hybrid outcomes. Nature Communications. 2020;11: 2179. doi:10.1038/s41467-020-15641-x

49. Westram AM, Faria R, Johannesson K, Butlin R. Using replicate hybrid zones to understand the genomic basis of adaptive divergence. Molecular Ecology. 2021;n/a. doi:https://doi.org/10.1111/mec.15861

50. Mitchell N, Luu H, Owens GL, Rieseberg LH, Whitney KD. Hybrid evolution repeats itself across environmental contexts in Texas sunflowers (Helianthus). Evolution. 2022;76: 1512–1528. doi:10.1111/evo.14536

51. Yuan K, Ni X, Liu C, Pan Y, Deng L, Zhang R, et al. Refining models of archaic admixture in Eurasia with ArchaicSeeker 2.0. Nat Commun. 2021;12: 6232. doi:10.1038/s41467-021-26503-5

52. Matute DR, Comeault AA, Earley E, Serrato-Capuchina A, Peede D, Monroy-Eklund A, et al. Rapid and Predictable Evolution of Admixed Populations Between Two Drosophila Species Pairs. Genetics. 2019 [cited 20 Apr 2020]. doi:10.1534/genetics.119.302685

53. Langdon QK, Powell DL, Kim B, Banerjee SM, Payne C, Dodge TO, et al. Predictability and parallelism in the contemporary evolution of hybrid genomes. PLOS Genetics. 2022;18: e1009914. doi:10.1371/journal.pgen.1009914

54. Powell DL, García-Olazábal M, Keegan M, Reilly P, Du K, Díaz-Loyo AP, et al. Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. Science. 2020;368: 731–736. doi:10.1126/science.aba5216

55. Moran BM, Payne CY, Powell DL, Iverson ENK, Banerjee SM, Langdon QK, et al. A Lethal Genetic Incompatibility between Naturally Hybridizing Species in Mitochondrial Complex I. 2021 Jul p. 2021.07.13.452279. doi:10.1101/2021.07.13.452279

56. Tiersch TR, Chandler RW, Kallman KD, Wachtel SS. Estimation of nuclear DNA content by flow cytometry in fishes of the genus Xiphophorus. Comparative Biochemistry and Physiology Part B: Comparative Biochemistry. 1989;94: 465–468. doi:10.1016/0305-0491(89)90182-X

57. Rosenthal GG. Swordtails and Platyfishes. In: Breed MD, Moore J, editors. Encyclopedia of Animal Behavior. Oxford: Academic Press; 2010. pp. 363–367. doi:10.1016/B978-0-08-045337-8.00273-4

58. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Hernandez R, editor. Molecular Biology and Evolution. 2019;36: 632–637. doi:10.1093/molbev/msy228

59. Thornton KR. Automating approximate Bayesian computation by local linear regression. BMC Genet. 2009;10: 35. doi:10.1186/1471-2156-10-35

60. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011;88: 76–82. doi:10.1016/j.ajhg.2010.11.011

61. Aguillon SM, Haase-Cox S, Langdon QK, Gunn TR, Banerjee SM, Guiterrez-Rodriguez C, et al. Multiple mechanisms maintain species barriers in hybridizing fish. In preparation.

62. Powell DL, Moran BM, Kim BY, Banerjee SM, Aguillon SM, Fascinetto-Zago P, et al. Two new hybrid populations expand the swordtail hybridization model system. Evolution. 2021;75: 2524–2539. doi:10.1111/evo.14337

63. Powell DL, Payne C, Banerjee SM, Keegan M, Bashkirova E, Cui R, et al. The Genetic Architecture of Variation in the Sexually Selected Sword Ornament and Its Evolution in

Hybrid Populations. Current Biology. 2021 [cited 28 Jan 2021]. doi:10.1016/j.cub.2020.12.049

64. Payne C, Bovio R, Powell DL, Gunn TR, Banerjee SM, Grant V, et al. Genomic insights into variation in thermotolerance between hybridizing swordtail fishes. Molecular Ecology. 2022. doi:10.1111/mec.16489

65. Moyle LC, Payseur BA. Reproductive isolation grows on trees. Trends Ecol Evol. 2009;24: 591–598. doi:10.1016/j.tree.2009.05.010

66. Sim SB, Corpuz RL, Simmonds TJ, Geib SM. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics. 2022;23: 157. doi:10.1186/s12864-022-08375-1

67. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18: 170–175. doi:10.1038/s41592-020-01056-5

68. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biology. 2022;23: 258. doi:10.1186/s13059-022-02823-7

69. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100. doi:10.1093/bioinformatics/bty191

70. Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, et al. The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. Nature Genetics. 2013;45: 567.

71. Powell D. Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. Dryad; 2020. p. 2648989706 bytes. doi:10.5061/DRYAD.Z8W9GHX82

72. Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, Blaxter M, Mieszkowska N, Hall N, et al. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. BMC Bioinformatics. 2023;24: 288. doi:10.1186/s12859-023-05385-y

73. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421

74. Du K, Pippel M, Kneitz S, Feron R, da Cruz I, Winkler S, et al. Genome biology of the darkedged splitfin, Girardinichthys multiradiatus, and the evolution of sex chromosomes and placentation. Genome Res. 2022;32: 583–594. doi:10.1101/gr.275826.121

75. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences. 2020;117: 9451–9457. doi:10.1073/pnas.1921046117

76. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015;6: 11. doi:10.1186/s13100-015-0041-9

77. Shao F, Wang J, Xu H, Peng Z. FishTEDB: a collective database of transposable elements identified in the complete genomes of fish. Database (Oxford). 2018;2018. doi:10.1093/database/bax106

78. Bailly-Bechet M, Haudry A, Lerat E. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mobile DNA. 2014;5: 13. doi:10.1186/1759-8753-5-13

79. She R, Chu JS-C, Wang K, Pei J, Chen N. GenBlastA: enabling BLAST to identify homologous gene sequences. Genome Res. 2009;19: 143–149. doi:10.1101/gr.082081.108

80. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34: i884–i890. doi:10.1093/bioinformatics/bty560

81. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12: 357–360. doi:10.1038/nmeth.3317

82. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8: 1494–1512. doi:10.1038/nprot.2013.084

83. Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. Biology Direct. 2008;3: 20. doi:10.1186/1745-6150-3-20

84. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34: W435–W439. doi:10.1093/nar/gkl200

85. Quail MA, Swerdlow H, Turner DJ. Improved Protocols for the Illumina Genome Analyzer Sequencing System. Current Protocols in Human Genetics. 2009;62: 18.2.1-18.2.27. doi:10.1002/0471142905.hg1802s62

86. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

87. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81: 559–575. doi:10.1086/519795

88. Delaneau O, Coulonges C, Zagury J-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics. 2008;9: 540. doi:10.1186/1471-2105-9-540

89. Chan AH, Jenkins PA, Song YS. Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. PLOS Genetics. 2012;8: e1003090. doi:10.1371/journal.pgen.1003090

90. Siepel A, Haussler D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol. 2004;21: 468–488. doi:10.1093/molbev/msh039

91. Preising GA, Gunn T, Baczenas JJ, Pollock A, Powell DL, Dodge TO, et al. Recurrent evolution of small body size and loss of the sword ornament in Northern Swordtail fish. bioRxiv; 2022. p. 2022.12.24.521833. doi:10.1101/2022.12.24.521833

92. Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, et al. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. In: eLife [Internet]. 6 Jun 2017 [cited 23 Jul 2019]. doi:10.7554/eLife.24133

93. Schumer M, Powell DL, Corbett-Detig R. Versatile simulations of admixture and accurate local ancestry inference with mixnmatch and ancestryinfer. Mol Ecol Resour. 2020;20: 1141–1151. doi:10.1111/1755-0998.13175

94. Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Molecular Ecology Resources. 2019;19: 552–566. doi:10.1111/1755-0998.12968

95. Li H. lh3/seqtk. 2023. Available: https://github.com/lh3/seqtk

96. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033