ACCEPTED MANUSCRIPT

High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species

Molly Schumer, Rongfeng Cui, Daniel Powell, Rebecca Dresner, Gil G Rosenthal, Peter Andolfatto

DOI: http://dx.doi.org/10.7554/eLife.02535

Cite as: eLife 2014;10.7554/eLife.02535

Received: 14 February 2014
Accepted: 2 June 2014
Published: 4 June 2014

This PDF is the version of the article that was accepted for publication after peer review. Fully formatted HTML, PDF, and XML versions will be made available after technical processing, editing, and proofing.

This article is distributed under the terms of the Creative Commons Attribution License permitting unrestricted use and redistribution provided that the original author and source are credited.

Stay current on the latest in life science and biomedical research from eLife.
Sign up for alerts at elife.elifesciences.org

Title: High-resolution Mapping Reveals Hundreds of Genetic Incompatibilities in Hybridizing Fish Species

Authors: Molly Schumer[1], Rongfeng Cui[3,4], Daniel Powell[3,4], Rebecca Dresner[1], Gil Rosenthal[3,4] and Peter Andolfatto[1,2]

Affiliations:

[1]Department of Ecology and Evolutionary Biology and the [2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA.

[3]Department of Biology, Texas A&M University, TAMU, College Station, TX, USA

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

Corresponding Author: Molly Schumer, schumer@princeton.edu,106A Guyot Hall, Princeton University, Princeton NJ, 08544. Tel: 609-258-3160.

Running title: Mapping genetic incompatibilities in hybridizing species

Key words: Hybridization, speciation, hybrid fitness, genetic incompatibilities

Data availability:

Sequence data: SRA Accession # SRX544941

Scripts used in analysis: Dryad (Provisional DOI: doi:10.5061/dryad.q6qn0)

Abstract

Hybridization is increasingly being recognized as a common process in both animal and plant species. Negative epistatic interactions between genes from different parental genomes decrease the fitness of hybrids and can limit gene flow between species. However, little is known about the number and genome-wide distribution of genetic incompatibilities separating species. To detect interacting genes, we perform a high-resolution genome scan for linkage disequilibrium between unlinked genomic regions in naturally occurring hybrid populations of swordtail fish. We estimate that hundreds of pairs of genomic regions contribute to reproductive isolation between these species, despite them being recently diverged. Many of these incompatibilities are likely the result of natural or sexual selection on hybrids, since intrinsic isolation is known to be weak. Patterns of genomic divergence at these regions imply that genetic incompatibilities play a significant role in limiting gene flow even in young species.

Introduction

Hybridization between closely related species is remarkably common (Mallet 2005). Many hybridizing populations and species remain genetically and ecologically distinct despite bouts of past admixture (e.g. Scascitelli et al. 2010; Vonholdt et al. 2010). This has led to a surge of interest in identifying which and how many loci are important in maintaining species barriers. Recent work has focused on identifying so-called "genomic islands" of high divergence between closely related species (e.g. Turner et al. 2005; Nadeau et al. 2012). This approach assumes that the most diverged regions between species are most likely to be under divergent selection between species or

47 important in reproductive isolation. However, divergence-based measures need to be
48 interpreted with caution because they are susceptible to artifacts as a result of linked
49 selection events (including background selection and hitchhiking) such that outlier
50 regions might reflect low within-population polymorphism rather than unusually high
51 divergence (discussed in Charlesworth 1998; Noor and Bennett 2009; Renaut et al. 2013),
52 and there are many possible causes of elevated divergence that are not linked to isolation
53 between species.

54 Investigating genome-wide patterns in naturally occurring or laboratory generated
55 hybrid populations is another approach to characterize the genetic architecture of
56 reproductive isolation (Payseur 2010). Hybridization leads to recombination between
57 parental genomes that can uncover genetic incompatibilities between interacting genes.
58 When genomes diverge in allopatry, substitutions that accumulate along a lineage can
59 lead to reduced fitness when hybridization decouples them from the genomic background
60 on which they arose. The best understood of these epistatic interactions, called "Bateson-
61 Dobzhansky-Muller" (BDM) incompatibilities (Coyne and Orr 2004), can occur as a
62 result of neutral substitution or adaptive evolution, and are thought to be common based
63 on theoretical (Orr 1995; Turelli et al. 2001) and empirical studies (Presgraves 2003;
64 Presgraves et al. 2003; Payseur and Hoekstra 2005; Brideau et al. 2006; Sweigart et al.
65 2006). One defining feature of BDM incompatibilities is that they are predicted to have
66 asymmetric fitness effects in different parental backgrounds, such that only a subset of
67 hybrid genotypes are under selection. Though the BDMI model is an important
68 mechanism of selection against hybrids, other evolutionary mechanisms can contribute to
69 hybrid incompatibility. For example, co-evolution between genes can result in selection

on all hybrid genotype combinations due to the accumulation of multiple substitutions (Seehausen et al. 2014). Similarly, natural or sexual selection against hybrid phenotypes can be considered a form of hybrid incompatibility; in this case the genotypes under selection will depend on their phenotypic effects.

How common are hybrid incompatibilities and what is their genomic distribution? Most studies to date have addressed this question by mapping hybrid incompatibilities that contribute to inviability or sterility (Orr 1989; Orr and Coyne 1989; Barbash et al. 2003; Presgraves 2003; Presgraves et al. 2003), in part because these incompatibilities affect hybrids even in a lab environment. Initial genome-wide studies in *Drosophila* and other organisms suggest that the number of incompatibilities contributing to hybrid viability and sterility can range from a handful to hundreds accumulating between deeply diverged species (Harushima et al. 2001; Presgraves 2003; Moyle and Graham 2006; Masly and Presgraves 2007; Ross et al. 2011); recent work has also suggested that substantial numbers of incompatibilities segregate within species (Cutter 2012; Corbett-Detig et al. 2013). However, because research has focused primarily on postzygotic isolation, little is known about the total number of loci contributing to reproductive isolation between species. For example, research shows that selection against hybrid genotypes can be strong even in the absence of postzygotic isolation (Fang et al. 2012). Thus, the focus on hybrid sterility and inviability is likely to substantially underestimate the true number of genetic incompatibilities distinguishing species.

If negative epistatic interactions are important in maintaining reproductive isolation, specific patterns of genetic variation are predicted in hybrid genomes. In particular, selection against hybrid individuals that harbor unfavorable allele

93 combinations in their genomes will lead to under-representation of these allelic
94 combinations in a hybrid population. Thus, selection can generate non-random
95 associations, or linkage disequilibrium (LD), among unlinked loci in hybrid genomes
96 (Karlin 1975; Hastings 1981). Patterns of LD in hybrid populations can therefore be used
97 to identify genomic regions that are important in establishing and maintaining
98 reproductive isolation between species.

99 Only a handful of studies have investigated genome-wide patterns of LD in hybrid
100 populations. Gardener et al. (2000) evaluated patterns of LD at 85 widely dispersed
101 markers (~0.03 markers/Mb) in sunflowers and found significant associations among
102 markers known to be related to infertility in hybrids. Similarly, Payseur and Hoekstra
103 (2005) evaluated patterns of LD among 332 unlinked SNPs (~0.12 markers/Mb) in inbred
104 lines of hybrid mice and identified a set of candidate loci with strong conspecific
105 associations. More recently, Hohenlohe et al. (2012) investigated genome-wide patterns
106 of LD at ~2,000 sites (~4.5 markers/Mb) in oceanic and freshwater sticklebacks and
107 found two unlinked regions in strong LD that are highly differentiated between
108 populations. These studies suggest hybridization can expose the genome to strong
109 selection that leaves detectable signatures of LD in hybrids.

110 Here, we evaluate genome-wide patterns of LD in replicate hybrid populations of
111 two species of swordtail fish, *Xiphophorus birchmanni* and *X. malinche*. These species
112 are recently diverged (0.5% genomic divergence per site; 0.4% genomic divergence
113 following polymorphism masking) and form multiple independent hybrid zones in river
114 systems in the Sierra Madre Oriental of Mexico (Culumber et al. 2011). *X. malinche* is
115 found at high elevations while *X. birchmanni* is common at low elevations; hybrids occur

where the ranges of these two species overlap. The strength of selection on hybrids between these two species is unknown, but several lines of evidence have suggested that selection may be weak. Hybrids are abundant in hybrid zones, often greatly outnumbering parental individuals. Hybrids are tolerant of the thermal environments at the elevations in which they are found (Culumber et al. 2012). Though there is some evidence of BDMIs between the species that cause lethal melanomas, these melanomas typically affect hybrids post-reproduction (Schartl 2008) and may constitute a weak or even favorable selective force (Fernandez and Morris 2008; Fernandez and Bowser 2010). Finally, recent behavioral studies show that once hybrids are formed, hybrid males actually have an advantage due to sexual selection compared to parental individuals (Figure 1; and see Fisher et al. 2009; Culumber et al. in press). However, the genomes of adult hybrids sampled from hybrid populations have already been subject to multiple generations of selection (at least 30; Rosenthal et al. 2003), making it difficult to evaluate the extent of selection on hybrids without genetic information. By surveying the genomes of hybrids from natural populations we are able to identify interacting genomic regions under selection in hybrids, giving a powerful picture of the number of regions involved in reproductive isolation between this recently diverged species pair.

To evaluate genome-wide patterns of LD in hybrid populations, we further develop the multiplexed shotgun genotyping (MSG) protocol of Andolfatto et al. (2011). Originally developed for QTL mapping in controlled genetic crosses, we describe modifications that make the technique applicable to population genetic samples from hybrid populations. With this approach, we assign ancestry to nearly 500,000 ancestry informative markers throughout the genome, allowing us to evaluate genome wide

patterns of LD at unprecedented resolution (~820 markers/Mb). The joint analysis of two independent hybrid zones allows us to distinguish the effects of selection against genetic incompatibilities from confounding effects due to population history (Gardner et al. 2000). We further evaluate loci that are in significant LD to investigate the mechanisms of selection on these pairs. Our results support the conclusion that a large number of loci contribute to reproductive isolation between these two species, that these loci have higher divergence than the genomic background, and that selection against genetic incompatibilities maintains associations between loci derived from the same parental genome.

Results

Hybrid zones have distinct demographic histories

We used a modified version of the MSG analysis pipeline, optimized for genotyping in natural hybrids (Materials and Methods, Figure 1 – figure supplement 1) to genotype individual fish collected from two independently formed hybrid zones, Calnali and Tlatemaco (Figure 1, Culumber et al. 2011). We genotyped 143 hybrid individuals from Calnali, 170 hybrids from Tlatemaco, and 60 parents of each species, determining ancestry at 469,400 markers distinguishing *X. birchmanni* and *X. malinche* at a median density of 1 marker per 234 bp (Materials and Methods). On average, hybrids from the Calnali hybrid zone derived only 20% of their genome from *X. malinche*, while hybrids from the Tlatemaco hybrid zone had 72% of their genomes originating from *X. malinche*. Most hybrid individuals were close to the average hybrid index in each group (Tlatemaco: 95% of individuals 66-80% *malinche* ancestry; Calnali: 95% of individuals

14-35% *malinche* ancestry). We determined the time since hybridization based on the decay in linkage disequilibrium (see Materials and Methods), assuming an average genome-wide recombination rate of 1 cM/378 kb (Walter et al. 2004). Estimates of hybrid zone age were similar for the two hybrid zones (Tlatemaco 56 generations, CI: 55-57, Calnali 35 generations CI: 34.5-35.6). Interestingly, these estimates are remarkably consistent with available historical estimates, which suggest that hybridization began within the last ~40 generations due to disruption of chemical cues by pollution (Fisher et al. 2006).

Significant LD between unlinked genomic regions

Ancestry calls at 469,400 markers genome-wide were thinned to 12,229 markers that are sufficient to describe all changes in ancestry (+/- 10%) across individuals in both populations (see Materials and Methods). Using this thinned dataset, we analyzed patterns of LD among all pairs of sites. The physical distance over which R^2 decayed to <0.5 was approximately 300 kb in both populations (Figure 2 – figure supplement 1). Average genome-wide R^2 between physically unlinked loci was 0.003 in Tlatemaco and 0.006 in Calnali (Figure 2A), and did not significantly differ from null expectations (1/2n, where n is the sample size). A p-value for the R^2 value for each pair of effectively unlinked sites was estimated using a Bayesian Ordered Logistic Regression t-test (Figure 2B, Materials and Methods).

The expected false discovery rates (FDRs) associated with given p-value thresholds applied to both populations were determined by simulation (Materials and Methods, Figure 3, Figure 3 – figure supplement 1). Using data from two hybrid

185 populations to examine LD circumvents two problems: 1) within a single population
186 cases of LD between unlinked sites are unlikely to be strong enough to survive false
187 discovery corrections and 2) even interactions that are highly significant could be caused
188 by population demographic history. The joint distribution of p-values in the two
189 populations implies several hundreds of unlinked locus pairs are in significant LD in both
190 populations (Figure 3 – figure supplement 1); 327 pairs of regions were significant at
191 FDR=5% and 150 pairs were significant at a more stringent FDR=2% (Figure 3). The
192 genomic distribution of loci in significant LD at FDR=5% is shown in Figure 4. For
193 simplicity we focus on statistics for the less stringent dataset (Supplementary file 1A), but
194 nearly identical results were found for the more stringent dataset (Tables 1 & 2).
195 Average R^2 for unlinked regions in LD is 0.08 in Tlatemaco and 0.12 in Calnali,
196 both significantly higher than the genomic background ($p<0.001$, by bootstrapping). LD
197 regions were non-randomly distributed in the genome ($\chi^2=87$, df=23, $p<3e-9$). Contrary
198 to findings of a large-X effect in other taxa (see Discussion), we do not find a significant
199 excess of LD pairs involving the X chromosome ($p=0.11$, Binomial test). Focusing on the
200 overlap of significant regions in both populations allowed us to narrow candidate regions
201 (Figure 4 – figure supplement 3). Regions in significant LD in both populations have a
202 median size of 45 kb (Figure 4 – figure supplement 1; Figure 4 – figure supplement 2);
203 67% of regions contain 10 genes or fewer, and 13 pairs of regions are at single gene
204 resolution (Supplementary File 1B). Approximately 10% of the genomic regions
205 identified are very large and contain hundreds of genes (>5 Mb, Supplementary File 1A),
206 potentially as a result of reduced recombination or selection. Unlinked regions in
207 significant LD were more strongly linked to neighboring loci (Figure 4 – figure

supplement 4), ruling out the possibility that mis-assemblies underlie the patterns we observe.

Evidence for hybrid incompatibilities

Models of selection against hybrid genotypes (Figure 5, Figure 5 – figure supplement 1) predict that certain genotype combinations will be less common (Karlin 1975). In particular, selection against hybrid incompatibilities is expected to generate positive R, or conspecific associations between loci. Among loci in significant LD, we found an excess of conspecific associations in both hybrid populations (94% of pairs in Calnali and 67% of pairs in Tlatemaco, $p<0.001$ for both populations relative to the genomic background by bootstrapping).

We focus all subsequent analyses on the subset of significant LD pairs (FDR=0.05) with conspecific associations in both populations (207 locus pairs) because these sites will be enriched for hybrid incompatibilities, but similar results are observed for the whole dataset (Supplementary File 1C). To estimate parameters under a classic BDM incompatibility model (Figure 5- figure supplement 1), we use an approximate Bayesian approach to simulate selection on two-locus interactions (see Materials and Methods). Though it is likely that other types of hybrid incompatibilities were identified in our analysis, estimates are similar using other models (see Materials and Methods; Figure 5-figure supplement 1C). These simulations demonstrate that our results are well described by a model of selection against hybrid incompatibilities (Figure 5, Figure 5 – figure supplement 1) and that sites are on average under weak to moderate selection

(Figure 5, maximum a posteriori estimates for Tlatemaco s^=0.027, and Calnali s^=0.074).

Elevated divergence at loci linked to incompatibilities

Strikingly, the median divergence between *X. birchmanni* and *X. malinche* at loci in significant conspecific LD (FDR=0.05) was much higher than the genomic background ($p<0.001$ by bootstrapping, Figure 6A). Elevated divergence could be caused by differences in selection on individual loci, differences in mutation rate, or reduced susceptibility of these genomic regions to homogenization by ongoing gene flow. To distinguish among these causes, we examined rates of synonymous substitution (dS). Elevated divergence compared to the genomic background is also observed at synonymous sites ($p<0.01$ by bootstrapping, Table 2). We also examined the same regions in two swordtail species in an independent *Xiphophorus* lineage, *X. hellerii* and *X. clemenciae* and found that the level of genomic divergence in this species pair was not significantly different from the genomic background (Figure 6B). Together, these results imply that variation in selective constraint or mutation rate do not explain elevated divergence between *X. birchmanni* and *X. malinche* at loci in conspecific LD.

Gene Ontology Analysis

We performed gene ontology (GO) analysis on unlinked regions in significant conspecific LD and, surprisingly, found no significantly enriched GO categories (see Materials and Methods). This result holds when restricting the analysis to regions

252 resolved to only a few genes (≤ 10 genes, 242 regions). This suggests that regions in
253 significant LD contain genes with a broad range of functional roles.
254

255 ## Discussion

256 In this study we assign ancestry to nearly 500,000 markers genome-wide in
257 samples from two hybrid fish populations, providing a portrait of the genetic architecture
258 of hybrid incompatibilities between two closely related species at unprecedented
259 resolution. We discover significant LD between hundreds of pairs of unlinked genomic
260 regions and show that a model of selection against hybrid incompatibilities describes the
261 observed conspecific LD patterns. This implies that many negative epistatic interactions
262 segregate in these hybrid fish populations despite the fact that intrinsic post-zygotic
263 isolation is thought to be weak (Rosenthal et al. 2003).
264

265 ### *How many regions are involved in hybrid incompatibility?*

266 Our analysis focuses on a high confidence set of unlinked loci in significant LD.
267 Despite these findings, the true number of loci involved in incompatibilities may involve
268 hundreds more interactions for several reasons. First, we conservatively exclude
269 interactions within chromosomes due to the lack of detailed genetic map information.
270 Second, we only have only moderate power to detect incompatibilities (for example,
271 ~30% using parameter estimates for Calnali). Third, relaxing the p-value threshold
272 suggests the true number of pairs in significant LD could be much larger (Figure 3 –
273 figure supplement 1). Finally, our experimental design incorporates two populations with
274 opposite trends in genome-wide ancestry and independent histories of hybridization. This

conservative approach allows us to exclude effects of population history, but false negatives in our joint analysis may result from effects that are population-specific (for example, different effects of extrinsic selection in the two populations). Investigating the role of such a large number of hybrid incompatibilities in determining the structure of hybrid genomes is an exciting area for future empirical and theoretical research.

How strong is selection on hybrid incompatibilities?

Using an approximate Bayesian approach, we estimate that the average selection coefficient on negative epistatic interactions is in the vicinity of s=0.03 (Tlatemaco) to s=0.07 (Calnali); scaling these by estimates of the effective population size (2Ns ~10 and ~45, respectively) suggests that selection is strong enough to be deterministic but of moderate strength relative to drift. Given the large number of putative incompatibilities, even weak selection would introduce substantial genetic load in hybrids. Depending on dominance effects, this could explain why hybrid populations are skewed in genome-wide ancestry. However, we may overestimate the potential for genetic load if epistatic interactions are complex. Though this study focuses on pairwise comparisons because statistically evaluating high-order interactions in a dataset of this size is intractable, more complex interactions are predicted to be likely (Orr 1995). The fact that many locus pairs (~40% of regions localized to 1 Mb or less) identified in this study interact with multiple partners provides indirect support for this prediction.

Insights from a genome-wide approach in natural hybrids

297 Most work on hybrid incompatibilities has focused on characterizing specific
298 incompatibilities at candidate genes, such as those associated with mapped QTL
299 distinguishing species (Ting et al. 1998; Presgraves et al. 2003; Lee et al. 2008; Tang and
300 Presgraves 2009; Moyle and Nakazato 2010). The idea that negative epistatic interactions
301 may be pervasive in closely related species or populations has come from multiple
302 studies of potential candidate genes (e.g. *Arabidopsis thaliana:* Bomblies et al. 2007;
303 Smith et al. 2011, *Xiphophorus:* Nairn et al. 1996*, Drosophila:* Barbash et al. 2000;
304 Bayes and Malik 2009; Tang and Presgraves 2009, *Caenorhabditis elegans:* Seidel et al.
305 2008) and several genome-wide studies (Harushima et al. 2001; Presgraves 2003;
306 Payseur and Hoekstra 2005; Moyle and Graham 2006; Masly and Presgraves 2007;
307 Matute et al. 2010). These studies suggest that on the order of 100 incompatibilities
308 explain isolation between closely related species. In contrast, though *X. malinche* and *X.*
309 *birchmanni* have a similar divergence time (~2 N_e generations) to a previously studied
310 *Drosophila* species pair (Masly and Presgraves 2007), we identify approximately 4-fold
311 more genetic incompatibilities at FDR=0.05.
312 What accounts for this difference? Previous studies on the genomic distribution of
313 hybrid incompatibilities have focused almost entirely on incompatibilities involved in
314 post-zygotic isolation (in some studies, exclusively in males; Presgraves 2003).
315 However, such studies can only provide a lower limit on the number of loci involved in
316 reproductive isolation between lineages. A recent study in *Drosophila simulans* and *D.*
317 *sechellia* investigated the effects of interspecific competition on the fitness of
318 introgressed lines (Fang et al. 2012). Though introgressed lines had no detectable
319 differences in fertility or viability, the authors nonetheless detected strong selection on

hybrids in competition experiments (Fang et al. 2012), implying that the number of loci involved in reproductive isolation has been vastly underestimated in most studies.

Our results lend support to this point of view. The hybrid genomes we analyze in this study have been exposed to over 30 generations of intrinsic and extrinsic selection. Given that post-zygotic isolation between *X. birchmanni* and *X. malinche* is weak (Rosenthal et al. 2003), we propose that extrinsic selection is more likely the cause of the majority of hybrid incompatibilities detected in this study. Future research comparing hybrid incompatibilities in lab hybrids to natural hybrids will begin to elucidate how many incompatibilities are targets of extrinsic versus intrinsic selection.

Many studies of hybrid incompatibilities have focused on organisms with clear species boundaries—those that do not frequently hybridize in nature. In species that do frequently hybridize, early research supported the conclusion that these species retain their identity through a few highly differentiated genomic regions, inferred to contain genes responsible for reproductive isolation (Turner et al. 2005; Ellegren et al. 2012). More recent studies have suggested that even hybridizing species remain genetically differentiated through much of the genome (Lawniczak et al. 2010; Michel et al. 2010). Our findings in *X. birchmanni* and *X. malinche* support the latter conclusion and suggest that hybrid incompatibilities may be a common mechanism of restricting gene flow genome-wide even in species with incomplete reproductive isolation.

Evidence for reduced gene flow associated with putative incompatibilities.

Theory predicts that more rapidly evolving genomic regions will be more likely to result in BDM incompatibilities (Orr 1995; Orr and Turelli 2001). However, an issue that

arises in testing this prediction is that functionally diverged regions associated with hybrid incompatibilities may also resist homogenization due to gene flow, conflating the cause and the effect of elevated divergence (but see below). While our study confirms that unlinked genomic regions in LD pairs are significantly more diverged between *X. birchmanni* and *X. malinche* compared to the genomic background (Figure 6), we also find that these regions do not show elevated divergence in an independent comparison of *Xiphophorus* fish (Figure 6B). This supports the hypothesis that these regions are resistant to homogenization due to ongoing gene flow between *X. birchmanni* and *X. malinche.* This finding is interesting because theoretical work suggests BDM incompatibilities are ineffective barriers to gene flow, especially when migration rates are high (Gavrilets 1997; Gompert et al. 2012), but incompatibilities in which all hybrid genotypes are under selection more effectively limit gene flow (Gavrilets 1997).

Functional evaluation of loci associated with putative incompatibilities

Remarkably, we found not a single significantly enriched GO category in well-resolved pairs in significant LD. This is in stark contrast to previous studies, such as that of Payseur and Hoekstra (2005), who found 17 over-represented GO categories in a dataset of ~180 pairs of loci (at 2 Mb resolution) detected in *Mus*. Our study suggests a much more equal representation of functional categories among genes involved in incompatibilities.

One of the first putative BDM incompatibilities identified at the genetic level involves the *Xmrk-2* gene in *Xiphophorus* hybrids. Hybrids between many *Xiphophorus* species develop lethal melanomas which have been hypothesized to reinforce species

366 boundaries (Orr and Presgraves 2000). Decoupling of *Xmrk-2* from its repressor (thought
367 to be the gene *cdkn2x*) through hybridization triggers melanoma development (Nairn et
368 al. 1996). Though melanomas can develop in *X. malinche* - *X. birchmanni* hybrids, we
369 found no evidence of LD between *Xmrk-2* and *cdkn2x* in either population. This may
370 support previous hypotheses that melanoma is not under strong selection in hybrids
371 because it affects older individuals (Schartl 2008) or provides an advantage in mate
372 choice (Fernandez and Morris 2008; Fernandez and Bowser 2010). Alternatively, *cdkn2x*
373 may not in fact be the repressor of *Xmrk-2*.
374

375 ## *No evidence for a large X effect*

376 Theory predicts that the X-chromosome will play a major role in the
377 establishment of reproductive isolation due to Haldane's rule, faster-X evolution or
378 meiotic drive (Presgraves 2008). Intriguingly, we do not see an excess of interactions
379 involving group 21, the putative X chromosome (Schartl et al. 2013). This is in contrast
380 to results in a large number of species that demonstrate that sex chromosomes play a
381 disproportionate role in the evolution of reproductive isolation (Sperling 1994;
382 Presgraves 2002; Payseur et al. 2004; Turner et al. 2005; Pryke 2010), including studies
383 on LD (Payseur and Hoekstra 2005). However, the X chromosome in *Xiphophorus* is
384 very young (Schartl 2004) and sex determination may also be influenced by autosomal
385 factors (Kallman 1984). Since the non-recombining portion of the Y chromosome is
386 small in *Xiphophorus*, this will reduce the effects of recessive X-chromosome
387 incompatibilities on male fitness.
388

389 *What explains significant heterospecific associations?*

390 We detect an excess of conspecific associations among significant LD pairs,
391 which we evaluate (above) in the context of selection on hybrid incompatibilities.
392 However, the proportion of locus pairs in significant LD that are in conspecific
393 association differs dramatically in the two populations. Six percent of locus pairs in
394 Calnali and 33% in Tlatemaco have significant heterospecific associations (i.e.
395 significantly negative R) at FDR=0.05. Heterospecific associations may be the result of
396 beneficial epistatic interactions that result in hybrid vigor. For example, hybrid males are
397 better at buffering the locomotor costs of sexual ornamentation (Johnson 2013) and have
398 an advantage compared to parental males from sexual selection (Culumber et al. in
399 press), which could in part counteract the negative fitness effects of genetic
400 incompatibilities. However, the fact that few loci are heterospecific in association in both
401 populations (2%, Figure 3 – figure supplement 1B) suggests that these patterns are not
402 repeatable across populations. One possible explanation for this is divergent effects of
403 mate preferences in the two populations. Behavioral studies have shown that *X. malinche*
404 females prefer unfamiliar male phenotypes (Verzijden et al. 2012) while *X. birchmanni*
405 females prefer familiar male phenotypes (Verzijden and Rosenthal 2011; Verzijden et al.
406 2012). Given that Tlatemaco hybrids are primarily *malinche* and Calnali hybrids are
407 primarily *birchmanni*, divergent effects of male phenotypes on mating preferences could
408 produce the observed patterns.

409

410

411

Conclusions

We find hundreds of pairs of unlinked regions in significant LD across the genomes of *X. birchmanni*-*X. malinche* hybrids in two independent hybrid populations. These associations are largely well described by a model of selection against hybrid incompatibilities, implying that reproductive isolation in these recently diverged species involves many loci. These regions were also more divergent between species than the genomic background, likely as a result of reduced permeability to ongoing gene flow between the species. By using samples from two populations with independent histories of hybridization, we are able to exclude population structure and drift as potential causes of these patterns. Our results suggest that past research has vastly underestimated the number of regions responsible for reproductive isolation between species by focusing on intrinsic postzygotic reproductive isolation. In addition, our results demonstrate that even in species without strong intrinsic post-zygotic isolation, hybrid incompatibilities are pervasive and play a major role in shaping the structure of hybrid genomes.

Materials and Methods

Genome sequencing and pseudogenome assembly

We created "pseudogenomes" of *X. malinche* and *X. birchmanni* based on the *X. maculatus* genome reference sequence. As raw materials, we used previously collected Illumina sequence data (Acc # SRX201248; SRX246515) derived from a single wild-caught male for each species (Cui et al. 2013; Schumer et al. 2013) and the current genome assembly for *X. maculatus* (Schartl et al. 2013). We used a custom python script to trim reads to remove low quality bases (Phred quality score<20) and reads with fewer

435 than 30 nucleotides of contiguous high quality bases and aligned these reads to the *X.*
436 *maculatus* reference using STAMPY v1.0.17 (Lunter and Goodson 2011) with default
437 parameters except for expected divergence set to 0.03. Between 98%-99% of reads from
438 both species mapped to the *X. maculatus* reference. Mapped reads were analyzed for
439 variant sites using the samtools/bcftools pipeline (Li and Durbin 2009). We used the VCF
440 files and a custom python script to create a new version of the *X. maculatus* reference
441 sequence for each species that incorporated variant sites and masked any sites that had
442 depth <10 reads or were called as heterozygous.
443 As an additional step to mask polymorphisms, we prepared multiplexed shotgun
444 genotyping (MSG) libraries (Andolfatto et al. 2011) for 60 parental individuals of each
445 species (Figure 1 – figure supplement 1), generating 78,881,136 single end 101
446 nucleotide reads at MseI sites for *X. malinche* and 80,189,844 single end 101 nucleotide
447 reads for *X. birchmanni*. We trimmed these reads as described above and mapped them to
448 the *X. malinche* and *X. birchmanni* reference pseudogenomes, respectively, using bwa (Li
449 and Durbin 2009). We analyzed variant sites as described above and excluded all sites
450 that were either polymorphic in the sampled parentals or had fixed differences between
451 the sampled parentals and the reference (excluding indels). After masking, 0.4% of sites
452 genome-wide were ancestry informative markers (AIMs) between *X. birchmanni* and *X.*
453 *malinche*. The total number of AIMs in the assembled 24 linkage groups was 2,189,807.
454 For the same 60 parental individuals of each species, we evaluated MSG output to
455 determine any markers that did not perform well in genotyping the parental individuals
456 (average probability of matching same-parent <0.9). We found that 1.7% of covered
457 markers performed poorly in *X. malinche* and 0.3% of markers performed poorly in *X.*

458 *birchmanni*; we excluded these 10,877 markers in downstream analysis, leaving
459 2,178,930 AIMs.
460
461 ## *Sample collection*
462 The procedures used in this study were approved by the Institutional Animal Care
463 and Use Committee at Texas A&M University (Protocols # 2010-111 and 2012-164).
464 Individuals were collected from two independent hybrid zones (Calnali-mid and
465 Tlatemaco, Culumber et al. 2011) in 2009, and between 2012-2013. Individuals were
466 caught in the wild using baited minnow traps, and lightly anesthetized with tricaine
467 methanesulfate. Fin clips were stored in 95% ethanol until extraction. Population
468 turnover rate is high between years, and sites were sampled only once per year. We also
469 performed relatedness analyses to ensure that individuals had not been resampled (data
470 not shown).
471
472 ## *MSG library preparation and sequencing*
473 DNA was extracted from fin clips using the Agencourt bead-based purification
474 method (Beckman Coulter Inc., Brea, CA) following manufacturer's instructions with
475 slight modifications. Fin clips were incubated in a 55 °C shaking incubator (100 rpm)
476 overnight in 94 µl of lysis buffer with 3.5 µl 40 mg/mL proteinase K and 2.5 DTT,
477 followed by bead binding and purification. Genomic DNA was quantified using a
478 Typhoon 9380 (Amersham Biosciences, Pittsburgh, PA) and evaluated for purity using a
479 Nanodrop 1000 (Thermo Scientific, Wilmington, DE); samples were diluted to 10 ng/µl.

MSG libraries were made as previously described (Andolfatto et al. 2011). Briefly, 50 ng of DNA was digested with MseI; following digestion custom barcodes were ligated to each sample. Five µl of sodium acetate and 50 µl of isopropanol were added to each sample and samples were pooled (in groups of 48) and precipitated overnight at -20 °C. Following overnight precipitation, samples were extracted and resuspended in TE (pH 8.0) and purified through a phenol-chloroform extraction and Agencourt bead purification. Pooled samples were run on a 2% agarose gel and fragments between 250-500 bp were selected and purified. Two ng of each pooled sample was amplified for 14-16 PCR cycles with custom indexed primers allowing us to pool ~180 samples for sequencing on one Illumina lane. Due to multiplexing with other libraries, samples were sequenced on a total of four Illumina HiSeq 2000 lanes with v3 chemistry. All raw data is available through the NCBI Sequence Read Archive (SRA Accession: SRX544941).

Raw reads were parsed by index and barcode; the number of reads per individual ranged from 0.4-2.8 million reads, with a median of 900,000 reads. After parsing, 101 bp reads were trimmed to remove low quality bases (Phred quality score<20) and reads with fewer than 30 bp of contiguous high quality bases. If an individual had more than 2 million reads, reads in excess of 2 million were excluded to improve the speed of the MSG analysis pipeline.

MSG analysis pipeline

The following parameters were specified in the MSG configuration file: recombination rate recRate=240, rfac=3, *X. birchmanni* error (deltapar1)=0.05, *X.*

503 *malinche* error (deltapar2)=0.04. See below for details on parameters and parameter
504 determination. All individuals were initially analyzed with naïve priors (probability of
505 ancestry for parent 1 = 0.33, parent 2 = 0.33, and heterozygous = 0.33) with the MSG
506 v0.2 pipeline (https://github.com/JaneliaSciComp/msg). Based on genome-wide ancestry
507 proportions given these priors, population-specific priors were calculated for Tlatemaco
508 (homozygous *X. malinche*=0.49, heterozygous=0.42, homozygous *X. birchmanni* =0.09)
509 and Calnali (homozygous *X. malinche*=0.04, heterozygous=0.32, homozygous *X.*
510 *birchmanni* =0.64). These estimates were used as new priors for a subsequent run of the
511 MSG pipeline. This resulted in genotype information at 1,179,187 ancestry informative
512 markers (~50% of the total number of ancestry informative markers). MSG ancestry
513 posterior probability files were thinned to exclude markers that were missing or
514 ambiguous in >15% of individuals leaving 469,400 markers (~820 markers/Mb).
515 Following this initial culling, markers were further thinned using a custom Python script
516 to exclude adjacent markers that did not differ in posterior probability values by +/- 0.1.
517 This resulted in 12,269 markers for linkage disequilibrium analysis; the median distance
518 between thinned markers was 2 kb (mean=48 kb). Individuals were considered hybrids if
519 at least 10% of their genome was contributed by each parent; using this definition, 100%
520 of individuals collected from Tlatemaco and 55% of individuals collected from Calnali
521 were hybrids. Three more individuals from Calnali were excluded because their hybrid
522 index was not within the 99% CI of the mean hybrid index. Including them resulted in a
523 significant deviation from the expected R^2 between unlinked sites of 1/2n (n – number of
524 sampled individuals).
525

Estimates of hybrid zone age

The expected number of generations since initial hybridization was estimated using the LD decay with distance method described in Hellenthal et al. (2014). First, we used genome sequences of 5 *Xiphophorus* outgroups (*X. maculatus*, *X. hellerii*, *X. clemenciae*, *X. variatus* and *X. nezahualcoyotl*) to identify autapomorphic loci in *X. malinche* and *X. birchmanni* respectively. We limit the analyses to only the autapomorphic loci for the minor parental species in each hybrid zone. We then fit an exponential curve D = a*exp(-T*x), where D is disequilibrium, a is a coefficient, T is time since hybridization in generations and x is the physical distance between markers in Morgans (scripts are available in the Dryad data repository under DOI doi:10.5061/dryad.q6qn0). Because we do not have access to a recombination map for our species, we assumed a uniform recombination rate of 1 cM/378 kb (Walter et al. 2004). This assumption can underestimate the time since hybridization in some cases (Sankararaman et al. 2012), but better estimates await more detailed genetic map information.

Quantifying LD and establishing significance

For each pairwise combination of markers (Figure 2 - figure supplement 2), we used a custom php script to calculate R, the correlation coefficient. R is typically defined as D, the disequilibrium coefficient, scaled by the square root of the product of the allele frequencies at the two loci (Hartl and Clark 1997). We use the methods outlined in Rogers & Huff (2009) for calculation of R using unphased data. We recorded whether R was positive or negative, corresponding to conspecific versus heterospecific association.

549 To assess significance of correlations, we used a Bayesian ordered logistic regression as
550 implemented in the R package bayespolr
551 (http://rss.acs.unt.edu/Rdoc/library/arm/html/bayespolr.html) to estimate Student's t; we
552 used this estimate to determine the two-sided p-value for the correlation. We only
553 considered pairs of markers belonging to different linkage groups; intrachromosomal
554 comparisons were excluded due to concerns about false positives caused by
555 recombination rate variation (scripts are available in the Dryad data repository under DOI
556 doi:10.5061/dryad.q6qn0).
557 To determine our expected false discovery rates (FDRs) associated with given p-
558 value significance thresholds, we used a simulation approach. For LD analysis, we
559 surveyed 12,269 markers (reduced from 1.2 million, see above), but many of these
560 markers are tightly clustered. We used the Matrix Spectral Decomposition method
561 described in Li & Ji (2005) as implemented in the program matSpDlite (Nyholt 2004), to
562 determine the effective number of markers. We used the correlation matrix for each pair-
563 wise marker from Tlatemaco for these calculations; calculations based on the correlation
564 matrix from Calnali resulted in a similar but slightly lower number of tests. We
565 determined based on this analysis that the effective number of markers is 1087. Based on
566 these results, we randomly selected 1087 markers from our dataset, randomly shuffled
567 genotypes within columns, calculated R^2 and p-values for the entire dataset, and
568 determined the expected number of false positives at different p-value thresholds. We
569 repeated this procedure 1,000 times. We compared the average number of false positives
570 to the total number of positives in the actual dataset at a number of p-value thresholds.
571 Based on this analysis, we determined that $p<0.013$ in both populations resulted in an

572 expected false discovery rate (FDR) of 0.05 for 1087 independent markers (excluding
573 within chromosome comparisons), while $p<0.007$ resulted in an expected FDR of 0.02.
574 Our analyses focused on the FDR=0.05 dataset, but we repeated these analyses with a
575 more restricted dataset (FDR=0.02, Tables 1 & 2). We also performed simulations to
576 investigate the potential effects of demographic processes on p-value distributions (see
577 below).
578

579 *Establishing the number of independent LD pairs*

580 In most cases, dozens to hundreds of contiguous markers showed the same
581 patterns of LD. In order to cluster these markers and delineate between independent and
582 non-independent LD blocks, we used an approach designed to conservatively estimate the
583 number of LD pairs. Within adjacent clusters on the same chromosome, we tested for
584 independence between clusters of sites by determining the p-value for R^2 between a focal
585 site and the last site of the previous LD cluster. If $p>0.013$ (our FDR=0.05 significance
586 threshold), we counted the focal site as the first site in a new cluster.
587

588 *Excluding mis-assemblies as causes of long range LD*

589 If regions of the *Xiphophorus* genome are misassembled, incorrect assignment of
590 contigs to different linkage groups could generate strong cross-chromosomal linkage
591 disequilibrium (see for e.g. Andolfatto et al. 2011). To evaluate this possibility, we
592 focused on markers at the edges of identified LD blocks and examine patterns of local
593 LD in these regions (Figure 4 – figure supplement 4). If markers had neighboring
594 markers within 300 kb (86%), we evaluated whether the marker had stronger LD with

neighboring markers than detected in any cross-chromosomal comparisons. Only 1.5% of markers in Calnali and 0.6% in Tlatemaco had stronger cross-chromosomal LD than local LD.

Analysis of potential hybrid incompatibilities

Selection against hybrid incompatibilities is expected to generate an excess of conspecific associations. To investigate whether regions in significant LD were more likely to have conspecific associations, we determined the direction of association between markers in each population. We compared the sign of R in pairs in significant LD (327 pairs at FDR=0.05) to the sign of R in 1000 datasets of the same size composed of randomly selected pairs from the genomic background ($p>0.013$ in each population).

Simulations of selection on hybrid incompatibilities

To investigate what levels of selection might be required to generate the patterns we observe, we use a model of selection on locus pairs following Karlin (1975) and a regression approach to approximate Bayesian inference using summary statistics as implemented in the program ABCreg (Beaumont et al. 2002; Jensen et al. 2008; Thornton 2009). We focus only on sites that have positive R (207 pairs) since these sites are expected to be enriched for hybrid incompatibilities.

Because selection on two-locus interactions results in changes in the frequency of particular genotypes, we used the frequency of double homozygous genotypes for the major parent (Tlatemaco: MM_MM, Calnali: BB_BB), frequency of homozygous-heterozygous genotypes for the minor parent (Tlatemaco: MB_BB and BB_MB, Calnali: MM_MB and MB_MM), and average final ancestry proportion as summary statistics.

Under the BDM incompatibility model, two distinct fitness matrices are possible (Figure 5 – figure 5 supplement 1). Because these models are equally likely (Coyne and Orr 2004) we used the random binomial function in R to assign the 207 conspecific-associated locus pairs to each fitness matrix for each simulation. To simplify our simulations, we assume that selection is equal on all genotype combinations that have not previously been exposed to selection in ancestral populations (see Figure 5 – figure 5 supplement 1). For each simulation, we drew from uniform prior distributions for 4 parameters. Limits on the prior distribution for admixture proportions for the two populations were determined as 0.5 to A, where A is the 95% CI of 1,000 bootstrap resamplings of population ancestry from the observed data. Each simulated replicate was generated as follows:

1) Draw values for *s* (0-0.1), initial admixture proportions (Tlatemaco 0.5-0.72, Calnali 0.18-0.5), number of generations of selection (Tlatemaco 40-70, Calnali 20-50), and hybrid population size (50-5,000).
2) Random assignment of 207 pairs to each of two possible BDM incompatibility fitness matrices
3) Calculate expected frequencies of each two-locus genotype using these priors and the methods described by Karlin (1975), introducing drift at each generation as sampling of 2N gametes.
4) After iterating through step 3 for t generations, we multinomially sampled expected frequencies from step 3 for n individuals. To account for variation in sample size, we simulated the actual distribution of sample sizes in the observed data.

5) Calculate the mean of each summary statistic

6) Repeat for 1,000,000 simulations

7) Run ABCreg with a tolerance of 0.5%

These simulations produced well-resolved estimates of the selection coefficient, *s,* and hybrid population size, N (Figure 5). We also repeated these simulations using a model of selection against all hybrid genotypes (Figure 5 – figure supplement 1C). These simulations also resulted in well-resolved posterior distributions of *s* and N and similar maximum a posteriori (MAP) estimates for both populations (Tlatemaco s=0.015, N=4360; Calnali s=0.043, N=270). This model may be more consistent with incompatibilities arising from co-evolving loci or selection against hybrid phenotypes. Scripts for this analysis are available in the Dryad data repository under DOI doi:10.5061/dryad.q6qn0.

To check the consistency of our simulations with the observed data we performed posterior predictive simulations by randomly drawing 100 values from the joint posterior (of N, s, generations of selection, and admixture proportions) with replacement for each population. For each draw we then simulated selection using these parameters, applying the same significance threshold as we applied to the real data, until 207 pairs had been generated. Departures from expectations under random mating were compared to departures in the real data (Figure 5, Figure 5 – figure supplement 3).

Genome divergence analyses

Regions involved in hybrid incompatibilities are predicted to be more divergent than other regions of the genome for a number of reasons (see main text). To evaluate

levels of divergence relative to the genomic background, we compared divergence (calculated as number of divergent sites/length of region) between *X. malinche* and *X. birchmanni* at 207 regions in significant conspecific LD (FDR=0.05) in both populations to 1000 datasets of the same size generated by randomly sampling regions throughout the genome that were not in significant LD ($p>0.013$ with all unlinked regions) using a custom perl script and the program fastahack (https://github.com/ekg/fastahack). For LD regions that included only 1 marker (n=60), we included the flanking region defined by the closest 5' and 3' marker. To analyze coding regions, we extracted exons from these regions and calculated dN, N, dS and S for each gene using codeml in PAML with the F3x4 codon model (Yang 1997, scripts are available in the Dryad data repository under DOI doi:10.5061/dryad.q6qn0). For a phylogenetically independent comparison, we repeated this analysis using pseudogenomes for two swordtail species for which we previously collected genome sequence data, *X. hellerii* and *X. clemenciae* (Schumer et al. 2013). Repeating all analyses for the full dataset (i.e. including pairs in heterospecific LD in one or both populations) did not substantially change the results (Supplementary File 1C).

Gene ontology analysis of genomic regions in conspecific LD

To determine whether there is significant enrichment of certain gene classes in our dataset, we annotated regions in significant LD. We only considered LD regions that contained 10 genes or fewer to limit our analysis to regions that are reasonably well-resolved. After excluding regions with no genes, this resulted in 242 regions for analysis containing 202 unique genes. We used the ensembl annotation of the *X. maculatus*

688 genome (http://www.ensembl.org/Xiphophorus_maculatus) to identify the HUGO
689 Genome Nomenclature Committee (HGCN) abbreviation for all the genes in each region.
690 Using the GOstats package in R, we built a custom *Xiphophorus* database using the
691 HGCN gene names listed in the genome and matching each of these to Gene Ontology
692 (GO) categories as specified in the annotated human genome database (in bioconductor
693 "org.Hs.eg.db"). This resulted in a total of 12,815 genes that could be matched to GO
694 categories. We then tested for functional enrichment in GO categories, using the GOstats
695 and GSEABase packages in R and a p-value threshold of <0.01.
696

697 *Modeling the effect of demographic processes on LD*

698 Demographic processes such as bottlenecks and continued migration can affect
699 LD measures and could potentially increase our false discovery rate. To explore how
700 demographic changes might influence LD p-value distributions, we performed coalescent
701 simulations using the MAP estimate of hybrid population size (co-estimated with other
702 parameters using ABC, see above).
703 We used Hudson's *ms* (Hudson 1990) to simulate an unlinked pair of regions in
704 two populations. We calculated time of admixture relative to the time of speciation using
705 the relationship $Tdiv_{4N}=(1/2)((D_{xy}/\theta) -1)$, where D_{xy} is the average number of
706 substitutions per site between species; we previously estimated ρ, the population
707 recombination rate, θ, the population mutation rate ($\rho = \theta = 0.0016$ per site), and N_e, the
708 effective population size (N_e =10,500), for *X. birchmanni* based on the whole genome
709 sequences (Schumer et al. 2013). Because parameter estimates were similar for *X.*
710 *malinche*, for simplicity, we used these estimates for both parental populations. For

711 population 1, we set the time of admixture ($t_{admix/4N}$) to 0.0014 generations, the proportion
712 *X. malinche* to 0.7 (derived from the average hybrid index in samples from Tlatemaco),
713 and the sample size to 170 diploids. For population 2, we set $t_{admix/4N}$ to 0.000875, the
714 proportion *X. malinche* to 0.2 (derived from the average hybrid index in samples from
715 Calnali), and the sample size to 143 diploids. We specified a bottleneck at $t_{admix/4N}$
716 reducing the population to 2% its original size in simulations of Calnali and 18% its
717 original size in simulations of Tlatemaco (based on results of ABC simulations, see
718 Results). In each simulated replicate, we selected one substitution from each unlinked
719 region and accepted a pair of sites if they were fixed for different states between parents
720 (evaluated by generating 30 chromosomes of each species per simulation). Simulations
721 were performed until 100,000 pairs had been simulated (scripts are available in the Dryad
722 data repository under DOI doi:10.5061/dryad.q6qn0); we used the rate of false positives
723 from these simulations to calculate the total expected number of false positives given our
724 effective number of tests. Based on these simulations, our expected FDR at $p<0.013$ is
725 ~10%, slightly larger than our expected FDR based on permutation of the data.
726 Because we do not have information about the migration history of these
727 populations, we use simulations of migration only to explore how continued migration or
728 multiple admixture events might affect our false discovery rate. We simulate
729 unidirectional migration of individuals from each parental population to the hybrid
730 population per generation ($4Nm=80$ for each parental population). Under this migration
731 scenario, our expected FDR at $p<0.013$ is 15%. In addition to scenarios of ongoing
732 migration we simulated migration bursts. For a short time interval that corresponds to ~1
733 generation starting ~10 generations ago ($t_{mig/4N}$=0.00025-0.000275) we set the migration

734 rate to 4Nm=4000, or 10% of the population made up of migrants. We simulated three
735 scenarios: 1) migration from the major parent (expected FDR 15%), 2) migration from
736 the minor parent (expected FDR 11%), and 3) migration from both parents (4Nm=2000
737 for each parental population, expected FDR 11%). None of these demographic scenarios
738 increased expected FDR above 15%.
739
740 *MSG parameter determination and power simulations*
741 Because MSG has not previously been used to analyze natural hybrids, we
742 evaluated performance at a range of parameters and performed power simulations. To
743 optimize the Hidden Markov Model (HMM) parameters of MSG for analysis of natural
744 hybrids, we used a combination of empirical data and simulations. We set the error rate
745 parameter (deltapar) for each parent based on the genome wide average rate of calls to
746 the incorrect parent in the 60 parental individuals of each species analyzed (deltapar=0.05
747 and 0.04 for *X. birchmanni* and *X. malinche*, respectively). The transition probability of
748 the HMM in MSG is determined by the mean genome-wide recombination rate
749 multiplied by a scalar (rfac). We set the recombination rate to 240 based on the estimate
750 of approximately 1 recombination event per chromosome per meiosis (24, 1 cM/378 kb,
751 (Walter et al. 2004), and an a prior expectation of at least 10 generations of hybridization.
752 We then increased the recombination factor step-wise to the maximum value that did not
753 induce false breakpoints in parental individuals; we determined that rfac could be set to 3
754 without leading to spurious ancestry calls for parental individuals.
755 At some point, ancestry blocks will be too small for the HMM to detect given our
756 density of ancestry informative markers. To determine the ancestry block size at which

757 sensitivity decreases, we used the pseudogenomes to generate 25 Mb chromosomes with
758 a homozygous block for the alternate parent randomly inserted (40, 60, 80, 100, 120 kb).
759 We simulated 100 replicates of each size class and generated 1 million reads at MseI sites
760 genome-wide. We then analyzed these simulated individuals using the MSG pipeline and
761 determined whether the homozygous segment was detected. Based on our simulations,
762 we determined that we had low power to detect ancestry blocks smaller than 80 kb
763 (probability of detection <=80%) and high power to detect blocks 120 kb or larger
764 (probability of detection >= 97%). To determine how much of the genome we are failing
765 to detect in small ancestry segments, we fit an exponential distribution to the observed
766 block sizes in the real data for each parent and for each population, generated samples
767 from an exponential distribution with the lambda of the observed distribution, and
768 determined what percent of bases pairs were found in ancestry blocks below our
769 detection threshold. Based on this analysis, we determined that in both hybrid zones, less
770 than 5% of base pairs in the genome are likely to fall into undetectable segments.
771 As an additional evaluation of MSG's effectiveness in genotyping hybrid
772 genomes with similar properties to ours, we simulated a 25 Mb admixed chromosome for
773 100 individuals, drawing ancestry size blocks from the block size distribution observed in
774 the real data. We then generated MSG data in silico for each simulated individual (1
775 million reads), and compared MSG ancestry calls to true ancestry. We found that on
776 average 91.4% of raw calls were made to the correct genotype; if ambiguous calls were
777 excluded (posterior probability <= 0.95, 7% of sites), MSG's accuracy increased to
778 >98%. The median size of regions for which incorrect calls were made was 29 kb, much
779 smaller than the median block size in the whole dataset.

Acknowledgments

We would like to thank Molly Przeworski, Graham Coop and Priya Moorjani for statistical advice, Bridgett vonHoldt, Stephen Wright and Ying Zhen for reading the manuscript, Dale Nyholt for providing the source code for matSpDlite, John Postlethwait for access to the *Xiphophorus maculatus* genome, and the federal government of Mexico for permission to collect fish.

Competing Interests

The authors declare no conflicts of interest.

801 **References**

802

803 Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, *et al*.
804 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping.
805 *Genome Res* **21**: 610-617. doi: 10.1101/gr.115402.110.
806 Barbash DA, Roote J, Ashburner M. 2000. The *Drosophila melanogaster* hybrid male
807 rescue gene causes inviability in male and female species hybrids. *Genetics* **154**:
808 1747-1771.
809 Barbash DA, Siino DF, Tarone AM, Roote J. 2003. A rapidly evolving MYB-related
810 protein causes species isolation in *Drosophila*. *Proc Natl Acad Sci USA* **100**:
811 5302-5307. doi: 10.1073/pnas.0836927100.
812 Bayes JJ, Malik HS. 2009. Altered Heterochromatin Binding by a Hybrid Sterility
813 Protein in Drosophila Sibling Species. *Science* **326**: 1538-1541. doi:
814 10.1126/science.1181756.
815 Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, Dangl JL, et al. 2007.
816 Autoimmune response as a mechanism for a Dobzhansky-Muller-type
817 incompatibility syndrome in plants. *PLoS Biol* **5**: 1962-1972. doi:
818 10.1371/journal.pbio.0050236
819 Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang X, Barbash DA. 2006. Two
820 Dobzhansky-Muller genes interact to cause hybrid lethality in *Drosophila*.
821 *Science* **314**: 1292-1295. doi: 10.1126/science.1133953.
822 Charlesworth B. 1998. Measures of divergence between populations and the effect of
823 forces that reduce variability. *Mol Biol Evol* **15**: 538-543.

824

825

826 Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF. 2013. Genetic
827 incompatibilities are widespread within species. *Nature* **504**: 135-137.
828 doi:10.1038/nature12678.
829 Coyne JA, Orr HA. 2004. *Speciation*. Sinaeur Associates, Sunderland, MA.
830 Cui R, Schumer M, Kruesi K, Walter R, Andolfatto P, Rosenthal G. 2013.
831 Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes.
832 *Evolution* **67**: 2166–2179. doi: 10.1111/evo.12099.
833 Culumber ZW, Fisher HS, Tobler M, Mateos M, Barber PH, Sorenson MD, et al. 2011.
834 Replicated hybrid zones of *Xiphophorus* swordtails along an elevational gradient.
835 *Mol Ecol* **20**: 342-356. doi: 10.1111/j.1365-294X.2010.04949.x.
836 Culumber ZW, Ochoa OM, Rosenthal GG. In press. Assortative mating and the
837 maintenance of population structure in a natural hybrid zone. *Amer Nat*.
838 Culumber ZW, Shepard DB, Coleman SW, Rosenthal GG, Tobler M. 2012.
839 Physiological adaptation along environmental gradients and replicated hybrid
840 zone structure in swordtails (Teleostei: *Xiphophorus*). *J Evol Biol* **25**: 1800-1814.
841 doi: 10.1111/j.1420-9101.2012.02562.x.
842 Cutter AD. 2012. The polymorphic prelude to Bateson-Dobzhansky-Muller
843 incompatibilities. *Trends Ecol Evol* **27**: 209-218. doi:
844 http://dx.doi.org/10.1016/j.tree.2011.11.004.

845 Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, et al. 2012. The
846 genomic landscape of species divergence in Ficedula flycatchers. *Nature* **491**:
847 756-760. doi:10.1038/nature11584.
848 Fang S, Yukilevich R, Chen Y, Turissini DA, Zeng K, Boussy IA, et al. 2012.
849 Incompatibility and Competitive Exclusion of Genomic Segments between
850 Sibling Drosophila Species. *PLoS Genet* **8**. doi: 10.1371/journal.pgen.1002795.
851 Fernandez AA, Bowser PR. 2010. Selection for a dominant oncogene and large male size
852 as a risk factor for melanoma in the *Xiphophorus* animal model. *Mol Ecol* **19**:
853 3114-3123. doi: 10.1111/j.1365-294X.2010.04738.x.
854 Fernandez AA, Morris MR. 2008. Mate choice for more melanin as a mechanism to
855 maintain a functional oncogene. *Proc Natl Acad Sci USA* **105**: 13503-13507. doi:
856 10.1073/pnas.0803851105.
857 Fisher HS, Mascuch SJ, Rosenthal GG. 2009. Multivariate male traits misalign with
858 multivariate female preferences in the swordtail fish, *Xiphophorus birchmanni*.
859 *Anim Behav* **78**: 265-269. doi: 10.1016/j.anbehav.2009.02.029.
860 Fisher HS, Wong BBM, Rosenthal GG. 2006. Alteration of the chemical environment
861 disrupts communication in a freshwater fish. *Proc R Soc London Ser B* **273**: 1187-
862 1193. doi: 10.1098/rspb.2005.3406.
863 Gardner K, Buerkle A, Whitton J, Rieseberg L. 2000. Inferring epistasis in wild
864 sunflower hybrid zones. In: *Epistasis and the Evolutionary Process* (ed. JB Wolf,
865 ED Brodie III, MJ Wade), pp. 264-279. Oxford University Press, New York.
866 Gavrilets S. 1997. Single locus clines. *Evolution* **51**: 979-983.

867 Gompert Z, Parchman TL, Buerkle CA. 2012. Genomics of isolation in hybrids. *Phil*
868 *Transs R Soc B* **367**: 439-450. doi: 10.1098/rstb.2011.0196.
869 Hartl DL, Clark AG. 2007. *Principles of population genetics*, Fourth edition. Sinauer
870 Associations, Sunderland.
871 Harushima Y, Nakagahra M, Yano M, Sasaki T, Kurata N. 2001. A genome-wide survey
872 of reproductive barriers in an intraspecific hybrid. *Genetics* **159**: 883-892.
873 Hastings A. 1981. Disequilibrium, selection, and recombination - limits in 2 - locus, 2 -
874 allele models. *Genetics* **98**: 659-668.
875 Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A
876 Genetic Atlas of Human Admixture History. *Science* **343**: 747-751. doi:
877 10.1126/science.1243518.
878 Hohenlohe PA, Bassham S, Currey M, Cresko WA. 2012. Extensive linkage
879 disequilibrium and parallel adaptive divergence across threespine stickleback
880 genomes. *Philos Trans R Soc London Ser B* **367**: 395-408.
881 doi:10.1098/rstb.2011.0245.
882 Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxford Surveys in*
883 *Evolutionary Biology* **7**: 1-44.
884 Jensen JD, Thornton KR, Andolfatto P. 2008. An Approximate Bayesian Estimator
885 Suggests Strong, Recurrent Selective Sweeps in *Drosophila*. *Plos Genetics* **4**:
886 e1000198. doi: 10.1371/journal.pgen.1000198.
887 Johnson JB. 2013. The architecture of phenotypes in a naturally hybridizing complex of
888 *Xiphophorus* fishes. Doctor of Philosophy in Biology, p. 103. Texas A&M
889 University, College Station, TX.

890 Kallman KD. 1984. A new look at sex determination in poeciliid fishes. In: *Evolutionary*
891 *genetics of fishes* (ed. Turner BJ), pp. 95-171. Plenum Press, New York.
892 Karlin S. 1975. General 2-locus selection models - Some objectives, results and
893 interpretations. *Theor Popul Biol* **7**: 364-398.
894 Lawniczak MKN, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, et al. 2010.
895 Widespread Divergence Between Incipient *Anopheles gambiae* Species Revealed
896 by Whole Genome Sequences. *Science* **330**: 512-514.
897 doi:10.1126/science.1195755.
898 Lee H-Y, Chou J-Y, Cheong L, Chang N-H, Yang S-Y, Leu J-Y. 2008. Incompatibility
899 of Nuclear and Mitochondrial Genomes Causes Hybrid Sterility between Two
900 Yeast Species. *Cell* **135**: 1065-1073. doi: 10.1016/j.cell.2008.10.047.
901 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
902 transform. *Bioinformatics* **25**:1754-1760. doi: 10.1093/bioinformatics/btp324.
903 Lunter G, Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast
904 mapping of Illumina sequence reads. *Genome Res* **21**:936-939. doi:
905 10.1101/gr.111120.110
906 Mallet J. 2005. Hybridization as an invasion of the genome. *Trends Ecol Evol* **20**: 229-
907 237. doi:10.1016/j.tree.2005.02.010.
908 Masly JP, Presgraves DC. 2007. High-resolution genome-wide dissection of the two rules
909 of speciation in *Drosophila. PLoS Biol* **5**: 1890-1898.
910 doi:10.1371/journal.pbio.0050243.

911 Matute DR, Butler IA, Turissini DA, Coyne JA. 2010. A Test of the Snowball Theory for
912 the Rate of Evolution of Hybrid Incompatibilities. *Science* **329**: 1518-1521. doi:
913 10.1126/science.1193440
914 Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL. 2010. Widespread
915 genomic divergence during sympatric speciation. *Proc Natl Acad Sci USA* **107**:
916 9724-9729. doi: 10.1073/pnas.1000939107.
917 Moyle LC, Graham EB. 2006. Genome-wide associations between hybrid sterility QTL
918 and marker transmission ratio distortion. *Mol Biol Evol* **23**: 973-980. doi:
919 10.1093/molbev/msj112.
920 Moyle LC, Nakazato T. 2010. Hybrid Incompatibility "Snowballs" Between *Solanum*
921 Species. *Science* **329**: 1521-1523. doi: 10.1126/science.1193063.
922 Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, et al.
923 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies
924 identified by large-scale targeted sequencing. *Philos Trans R Soc London Ser B*
925 **367**: 343-353. doi: 10.1126/science.1193063.
926 Nairn RS, Kazianis S, McEntire BB, DellaColetta L, Walter RB, Morizot DC. 1996. A
927 CDKN2-like polymorphism in *Xiphophorus* LG V is associated with UV-B-
928 induced melanoma formation in platyfish-swordtail hybrids. *Proc Natl Acad Sci*
929 *USA* **93**: 13042-13047.
930 Noor MAF, Bennett SM. 2009. Islands of speciation or mirages in the desert? Examining
931 the role of restricted recombination in maintaining species. *Heredity* **103**: 439-
932 444. doi: 10.1038/hdy.2009.151.

933 Nyholt DR. 2004. A simple correction for multiple testing for single-nucleotide
934 polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* **74**:
935 765-769. doi:10.1086/383251.
936 Orr HA. 1989. Genetics of sterility in hybrids between 2 subspecies of *Drosophila*.
937 *Evolution* **43**: 180-189.
938 Orr HA. 1995. The population genetics of speciation - the evolution of hybrid
939 incompatibilities. *Genetics* **139**: 1805-1813.
940 Orr HA, Coyne JA. 1989. The genetics of postzygotic isolation in the *Drosophila virilus*
941 group. *Genetics* **121**: 527-537.
942 Orr HA, Presgraves DC. 2000. Speciation by postzygotic isolation: forces, genes and
943 molecules. *Bioessays* **22**: 1085-1094. doi: 10.1002/1521-
944 1878(200012)22:12<1085::AID-BIES6>3.0.CO;2-G.
945 Orr HA, Turelli M. 2001. The evolution of postzygotic isolation: Accumulating
946 Dobzhansky-Muller incompatibilities. *Evolution* **55**: 1085-1094. doi:
947 10.1111/j.0014-3820.2001.tb00628.x.
948 Payseur BA. 2010. Using differential introgression in hybrid zones to identify genomic
949 regions involved in speciation. *Mol Ecol Resour* **10**: 806-820. doi:
950 10.1111/j.1755-0998.2010.02883.x.
951 Payseur BA, Hoekstra HE. 2005. Signatures of reproductive isolation in patterns of single
952 nucleotide diversity across inbred strains of mice. *Genetics* **171**: 1905-1916. doi:
953 10.1534/genetics.105.046193

954 Payseur BA, Krenz JG, Nachman MW. 2004. Differential patterns of introgression across
955 the X chromosome in a hybrid zone between two species of house mice.
956 *Evolution* **58**: 2064-2078. doi: 10.1554/03-738.
957 Presgraves DC. 2002. Patterns of postzygotic isolation in Lepidoptera. *Evolution* **56**:
958 1168-1183. doi: 10.1111/j.0014-3820.2002.tb01430.x.
959 Presgraves DC. 2003. A fine-scale genetic analysis of hybrid incompatibilities in
960 *Drosophila*. *Genetics* **163**: 955-972.
961 Presgraves DC. 2008. Sex chromosomes and speciation in *Drosophila*. *Trends Genet* **24**:
962 336-343. doi: 10.1016/j.tig.2008.04.007.
963 Presgraves DC, Balagopalan L, Abmayr SM, Orr HA. 2003. Adaptive evolution drives
964 divergence of a hybrid inviability gene between two species of *Drosophila*.
965 *Nature* **423**: 715-719. doi:10.1038/nature01679.
966 Pryke SR. 2010. Sex chromosome linkage of mate preference and color signal maintains
967 assortative mating between interbreeding finch morphs. *Evolution* **64**: 1301-1310.
968 doi: 10.1111/j.1558-5646.2009.00897.x.
969 Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, et al. 2013. Genomic
970 islands of divergence are not affected by geography of speciation in sunflowers.
971 *Nat Commun* **4**. doi:10.1038/ncomms2833.
972 Rogers AR, Huff C. 2009. Linkage Disequilibrium Between Loci With Unknown Phase.
973 *Genetics* **182**: 839-844. doi: 10.1534/genetics.108.093153.
974 Rosenthal GG, de la Rosa Reyna XF, Kazianis S, Stephens MJ, Morizot DC, Ryan MJ, et
975 al. 2003. Dissolution of sexual signal complexes in a hybrid zone between the
976 swordtails *Xiphophorus birchmanni* and *Xiphophorus malinche* (Poeciliidae).

977 *Copeia* **2003**: 299-307. doi: http://dx.doi.org/10.1643/0045-
978 8511(2003)003[0299:DOSSCI]2.0.CO;2.
979 Ross JA, Koboldt DC, Staisch JE, Chamberlin HM, Gupta BP, Miller RD, Baird SE,
980 Haag ES. 2011. *Caenorhabditis briggsae* Recombinant Inbred Line Genotypes
981 Reveal Inter-Strain Incompatibility and the Evolution of Recombination. *PLoS*
982 *Genet* **7**:e1002174. doi: 10.1371/journal.pgen.1002174.
983 Sankararaman S, Patterson N, Li H, Paeaebo S, Reich D. 2012. The Date of Interbreeding
984 between Neandertals and Modern Humans. *PLoS Genet* **8**:e1002947. doi:
985 10.1371/journal.pgen.1002947
986 Scascitelli M, Whitney KD, Randell RA, King M, Buerkle CA, Rieseberg LH. 2010.
987 Genome scan of hybridizing sunflowers from Texas (*Helianthus annuus* and *H.*
988 *debilis*) reveals asymmetric patterns of introgression and small islands of genomic
989 differentiation. *Mol Ecol* **19**: 521-541. doi: 10.1111/j.1365-294X.2009.04504.x.
990 Schartl M. 2004. Sex chromosome evolution in non-mammalian vertebrates. *Curr Opin*
991 *Genet Dev* **14**: 634-641. doi:10.1016/j.gde.2004.09.005.
992 Schartl M. 2008. Evolution of Xmrk: an oncogene, but also a speciation gene? *Bioessays*
993 **30**: 822-832. doi: 10.1002/bies.20807.
994 Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, et al. 2013. The genome
995 of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary
996 adaptation and several complex traits. *Nat Genet* **45**: 567-U150. doi:
997 10.1038/ng.2604.

998 Schumer M, Cui R, Boussau B, Walter R, Rosenthal G, Andolfatto P. 2013. An
999 evaluation of the hybrid speciation hypothesis for *Xiphophorus clemenciae* based
1000 on whole genome sequences. *Evolution* **67**: 1155-1168. doi: 10.1111/evo.12009.
1001 Schumer M, Cui R, Powell D, Dresner R, Rosenthal GG, Andolfatto P. Data from: High-
1002 resolution Mapping Reveals Hundreds of Genetic Incompatibilities in Hybridizing
1003 Fish Species. Dryad Digital Repository. doi:10.5061/dryad.q6qn0.
1004 Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel
1005 CL, Saetre G-P, Bank C, Braennstroem A et al. 2014. Genomics and the origin of
1006 species. *Nature Reviews Genetics* **15**: 176-192. doi:10.1038/nrg3644.
1007 Seidel HS, Rockman MV, Kruglyak L. 2008. Widespread genetic incompatibility in *C.*
1008 *elegans* maintained by balancing selection. *Science* **319**: 589-594. doi:
1009 10.1126/science.1151107.
1010 Smith LM, Bomblies K, Weigel D. 2011. Complex Evolutionary Events at a Tandem
1011 Cluster of *Arabidopsis thaliana* Genes Resulting in a Single-Locus Genetic
1012 Incompatibility. *PLoS Genet* **7**: 1-14. doi: 10.1371/journal.pgen.1002164.
1013 Sperling FAH. 1994. Sex-linked genes and species-differences in Lepidoptera. *Can*
1014 *Entomol* **126**: 807-818. doi:10.4039/Ent126807-3.
1015 Sweigart AL, Fishman L, Willis JH. 2006. A simple genetic incompatibility causes
1016 hybrid male sterility in mimulus. *Genetics* **172**: 2465-2479. doi:
1017 10.1534/genetics.105.053686.
1018 Tang S, Presgraves DC. 2009. Evolution of the *Drosophila* Nuclear Pore Complex
1019 Results in Multiple Hybrid Incompatibilities. *Science* **323**: 779-782. doi:
1020 10.1126/science.1169123.

1021 Thornton KR. 2009. Automating approximate Bayesian computation by local linear
1022 regression. *BMC Genet* **10**. doi:10.1186/1471-2156-10-35.
1023 Ting CT, Tsaur SC, Wu ML, Wu CI. 1998. A rapidly evolving homeobox at the site of a
1024 hybrid sterility gene. *Science* **282**: 1501-1504. doi:
1025 10.1126/science.282.5393.1501.
1026 Turelli M, Barton NH, Coyne JA. 2001. Theory and speciation. *Trends Ecol Evol* **16**:
1027 330-343. doi:10.1016/S0169-5347(01)02177-2.
1028 Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles*
1029 *gambiae*. *PLoS Biol* **3**: 1572-1578. doi: 10.1371/journal.pbio.0030285.
1030 Verzijden MN, Culumber ZW, Rosenthal GG. 2012. Opposite effects of learning cause
1031 asymmetric mate preferences in hybridizing species. *Behav Ecol* **23**: 1133-1139.
1032 doi: 10.1093/beheco/ars086.
1033 Verzijden MN, Rosenthal GG. 2011. Effects of sensory modality on learned mate
1034 preferences in female swordtails. *Anim Behav* **82**: 557–562.
1035 doi:10.1016/j.anbehav.2011.06.010.
1036 Vonholdt BM, Stahler DR, Bangs EE, Smith DW, Jimenez MD, Mack CM, et al. 2010. A
1037 novel assessment of population structure and gene flow in grey wolf populations
1038 of the Northern Rocky Mountains of the United States. *Mol Ecol* **19**: 4412-4427.
1039 doi: 10.1111/j.1365-294X.2010.04769.x.
1040 Walter RB, Rains JD, Russell JE, Guerra TM, Daniels C, Johnston DA, et al. 2004. A
1041 microsatellite genetic linkage map for *Xiphophorus*. *Genetics* **168**: 363-372. doi:
1042 10.1534/genetics.103.019349.

1043 Yang ZH. 1997. PAML: a program package for phylogenetic analysis by maximum
1044 likelihood. *Comput Appl Biosci* **13**: 555-556. doi: 10.1093/molbev/msm088.

1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065

Figure legends

Figure 1. Hybrids between *X. malinche* and *X. birchmanni*. (**A**) Parental (*X. malinche* top, *X. birchmanni* bottom) and (**B**) hybrid phenotypes with sample MSG genotype plots for linkage groups 1-3 (see figure supplement 1 for more examples) for each population shown in the right panel. Hybrid individuals from Tlatemaco (B top) have *malinche*-biased ancestry while hybrids from Calnali (B bottom) have *birchmanni*-biased ancestry.

Figure 1-figure supplement 1. MSG ancestry plots for parental and hybrid individuals. Representative parental individuals of *X. birchmanni* (**A**) and *X. malinche* (**B**) in linkage groups 1-3, shows that parental individuals are called as homozygous throughout the chromosome. Tick markers indicating calls to the other parent are the result of undetected polymorphism or error. More representatives from the parental panel are shown in (**C**). Hybrids from Tlatemaco (left) and Calnali (right) are shown in panel (**D**).

Figure 2. R^2 distribution and p-value distributions of the sites analyzed in this study. (**A**) Genome-wide distribution of randomly sampled R^2 values for markers on separate chromosomes (see figure supplement 1 for R^2 decay by distance; figure supplement 2 for a genome-wide plot). Blue indicates the distribution in Tlatemaco while yellow indicates the distribution in Calnali. Regions of overlapping density are indicated in green. The average genome-wide R^2 in Tlatemaco is 0.003 and in Calnali is 0.006. (**B**) qq-plots of –

1088 $\log_{10}$(p-value) for a randomly selected subset of unlinked sites analyzed in this study in
1089 each population; expected p-values are drawn from p-values of the permuted data.
1090
1091 **Figure 2-figure supplement 1. Decay in linkage disequilibrium.** Average decay of R^2
1092 over distance in Tlatemaco (black) and Calnali (red) in 100 kb windows (plot generated
1093 from 1000 randomly selected sites).
1094
1095 **Figure 2-figure supplement 2. Genome-wide linkage disequilibrium plot for**
1096 **Tlatemaco.** This plot demonstrates that there are few regions in the *X. birchmanni-*
1097 *malinche* genomes that are inverted relative to the *X. maculatus* genome. Red indicates
1098 regions of R^2 near 1, while green indicates low R^2. Regions outlined in dark blue appear
1099 to be inverted based on the analysis in both Tlatemaco and Calnali, regions highlighted in
1100 gray appear to be inverted only in the Tlatemaco analysis.
1101
1102 **Figure 3. Number of unlinked pairs in significant linkage disequilibrium and**
1103 **expected false discovery rates.** Plot showing number of pairs of sites in significant LD
1104 in both populations in the stringent and relaxed datasets (light blue). The expected
1105 number of false positives in each dataset is shown in dark blue, and was determined by
1106 simulation (see main text; figure supplement 1).
1107
1108 **Figure 3-figure supplement 1. False discovery rate (FDR) at different p-value**
1109 **thresholds.** (**A**) The number of pairs of loci in LD in both populations in black (y-axis
1110 left) versus the expected false discovery rate in red (y-axis right) at different p-value

thresholds. (**B**) Estimated number of true positives in the dataset at different p-value thresholds for all pairs (black), conspecific pairs (blue), and heterospecific pairs (red). Expected false discovery rate was determined by 1000 simulations randomly permuting markers from the real data in both populations.

Figure 4. Distribution of sites in significant linkage disequilibrium throughout the *Xiphophorus* genome. Schematic of regions in significant LD in both populations at FDR 5%. Regions in blue indicate regions that are positively associated in both populations (conspecific in association), regions in black indicate associations with different signs of R in the two populations, while regions in red indicate those that are negatively associated in both populations (heterospecific in association). Chromosome lengths and position of LD regions are relative to the length of the assembled sequence for that linkage group; most identified LD regions are <50 kb (figure supplement 1; figure supplement 2; figure supplement 3). Analysis of local LD excludes mis-assemblies as the cause of these patterns (figure supplement 4).

Figure 4-figure supplement 1. Log_{10} distribution of LD region length in base pairs. The dotted line indicates the median length of regions in cross-chromosomal LD (45 kb).

Figure 4-figure supplement 2. Plot of the number of recombination breakpoints detected along linkage group 2. Number of breakpoints in Tlatemaco are indicated in blue and Calnali in red. (**A**) Breakpoints counted in 1 Mb windows and (**B**) 100 kb

1133 windows. The high density of recombination events allows for the identification of
1134 narrow regions in linkage disequilibrium.
1135
1136 **Figure 4-figure supplement 3. Example of the use of data from two populations to**
1137 **narrow candidate regions in cross-chromosomal LD.** P-values for linkage
1138 disequilibrium between a marker on linkage group 2 and an interval on linkage group 16
1139 (blue: Calnali, black: Tlatemaco). Overlapping significant intervals from the two
1140 populations allows us to narrow candidate regions.
1141
1142 **Figure 4-figure supplement 4. Regions in cross-chromosomal LD are also in LD with**
1143 **their neighbors.** Decay in R^2 of markers at the edge of LD blocks in both populations
1144 (black lines) compared to 95% confidence intervals of 1000 markers randomly selected
1145 from the genomic background (blue) in Tlatemaco (**A**) and Calnali (**B**). Fewer than 5% of
1146 markers fall outside of the 95% confidence intervals in each 100 kb window in both
1147 populations. Average R^2 and 95% CI for regions in significant cross-chromosomal LD
1148 are shown in purple. LD blocks without neighboring markers within 300 kb of the focal
1149 marker are excluded from this figure.
1150
1151 **Figure 5. Loci in significant conspecific linkage disequilibrium show patterns**
1152 **consistent with selection against hybrid incompatibilities.** (**A**) Posterior distributions
1153 of the selection coefficient and hybrid population size from ABC simulations for
1154 Tlatemaco and (**B**) Calnali. The range of the x-axis indicates the range of the prior
1155 distribution, maximum a posteriori estimates (MAP) and 95% CI are indicated in the

inset. (**C**) Departures from expectations under random mating in the actual data (top- blue points indicate LD pairs, black points indicate random pairs from the genomic background) and samples generated by posterior predictive simulations (bottom, see Materials and Methods). The mean is indicated by a dark blue point; in the real data (top) smears denote the distribution of means for 1,000 simulations while in the simulated data (bottom) smears indicate results of each simulation. Genotypes with the same predicted deviations on average under the BDM model have been collapsed (figure supplement 1, but see figure supplement 3) and are abbreviated in the format locus1_locus2. These simulations show that the observed deviations are expected under the BDM model. The posterior distributions for *s* and hybrid population size are correlated at low population sizes (figure supplement 2). Deviations in Calnali also follow expectations under the BDM model (figure supplement 3).

Figure 5-figure supplement 1. Different fitness matrices associated with selection against hybrid incompatibilities. In a classic BDMI model (**A & B**), hybrid genotypes potentially under selection (indicated in red) are determined by the locus and order in which mutations occur. In a model of co-evolution between loci or extrinsic selection against hybrid phenotypes (**C**), more genotype combinations are potentially under selection. (**A**) Interaction between a mutation in locus 1 *malinche* and locus 2 *birchmanni* (two-lineage model) or first substitution occurring in locus 1 *birchmanni* or locus 2 *malinche* (one-lineage model). (**B**) Interaction between a mutation in locus 1 *birchmanni* and locus 2 *malinche* (two-lineage model) or first substitution occurring in locus 1

malinche or locus 2 *birchmanni* (one-lineage model). Format of genotypes is as follows: haplotype1_locus1-haplotype1_locus2/haplotype2_locus1-haplotype2_locus2.

Figure 5-figure supplement 2. Joint posterior distribution of hybrid population size and selection coefficient. Posterior distributions of hybrid population size and *s* indicate a relationship between these parameters in both populations (**A**-simulations of Tlatemaco, **B**- simulations of Calnali).

Figure 5-figure supplement 3. Deviations in genotype combinations compared to expected values under a two-locus selection model in both populations. Average deviation of genotype combinations (dark blue point) from expectations under random mating at conspecific LD pairs (top) compared to posterior predictive simulations (bottom, see Materials and Methods) in Tlatemaco (**A**) and Calnali (**B**). The light blue smears indicate the distribution of means for 1,000 bootstrap samples in the real data and result of individual simulations in the simulated data. Labels on the x-axis indicate the genotype in the format locus1_locus2.

Figure 6. Divergence of LD pairs compared to the genomic background in two species comparisons. (**A**) Regions identified in *X. birchmanni* and *X. malinche* and (**B**) orthologous regions in *X. hellerii* and *X. clemenciae*. The blue point shows the average divergence for genomic regions within significant LD pairs, and whiskers denote a 95% confidence interval estimated by resampling genomic regions with replacement. The

histogram shows the distribution of the average divergence for 1000 null datasets generated by resampling the genomic background without replacement.

Tables

Table 1. **Comparison of results for sites in significant LD at two different p-value thresholds.** P-values were determined resampling the genomic background, see main text for details.

Dataset	Number of pairs	Proportion of pairs with conspecific associations
Stringent (FDR<2%)	150	Tlatemaco: 72% ($p<0.001$) Calnali: 94% ($p<0.001$)
Relaxed (FDR 5%)	327	Tlatemaco: 67% ($p<0.001$) Calnali: 94% ($p<0.001$)

Table 2. Sites in significant LD are more divergent than the genomic background. Results shown here are limited to regions that had conspecific associations in both populations (stringent dataset: 200 regions, relaxed dataset: 414 regions). P-values were determined by resampling the genomic background.

Mutation type	Median divergence Genomic Background	Median divergence Stringent	Median divergence Relaxed
All sites	0.0040	0.0045 ($p<0.001$)	0.0044 ($p<0.001$)
Nonsynonymous	0.00040	0.00065 ($p=0.001$)	0.00040 ($p=0.6$)
Synonymous	0.0040	0.0048 ($p<0.001$)	0.0045 ($p=0.004$)

List of Supplementary Materials

Figure supplements:

Figure 1-figure supplement 1. MSG ancestry plots for parental and hybrid individuals.

Figure 2-figure supplement 1. Decay in linkage disequilibrium.

Figure 2-figure supplement 2. Genome-wide linkage disequilibrium plot for Tlatemaco.

Figure 3-figure supplement 1. False discovery rate (FDR) at different p-value thresholds.

Figure 4-figure supplement 1. Log_{10} distribution of LD region length in base pairs.

Figure 4-figure supplement 2. Plot of the number of recombination breakpoints detected along linkage group 2.

Figure 4-figure supplement 3. Example of the use of data from two populations to narrow candidate regions in long-range LD.

Figure 4-figure supplement 4. Regions in long-range LD are also in LD with their neighbors.

Figure 5-figure supplement 1. Different fitness matrices associated with selection against hybrid incompatibilities.

Figure 5-figure supplement 2. Joint posterior distribution of hybrid population size and selection coefficient.

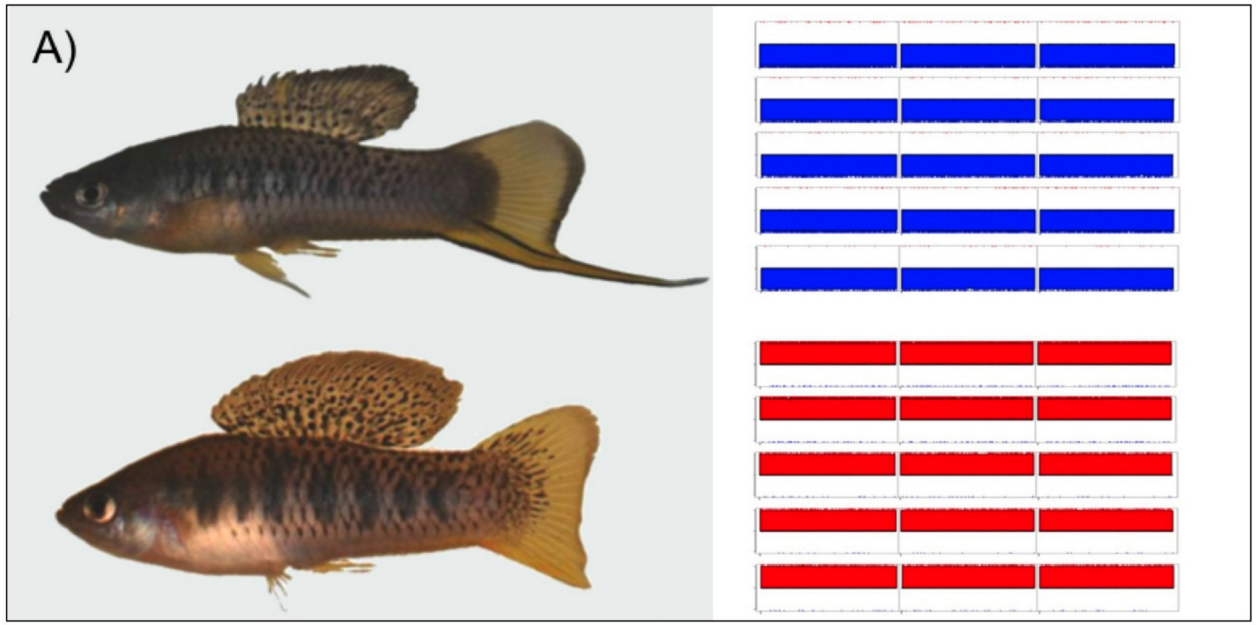
Figure 5-figure supplement 3. Deviations in genotype combinations compared to expected values under a two-locus selection model in both populations.

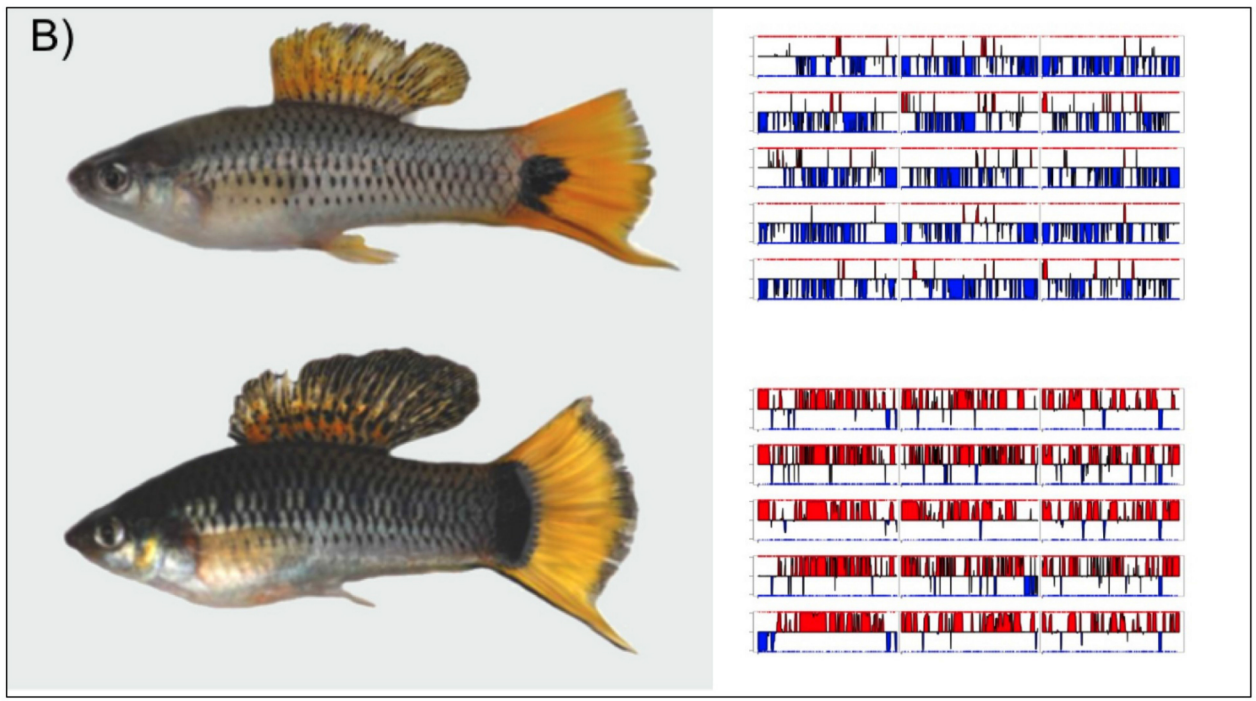
Supplementary Tables

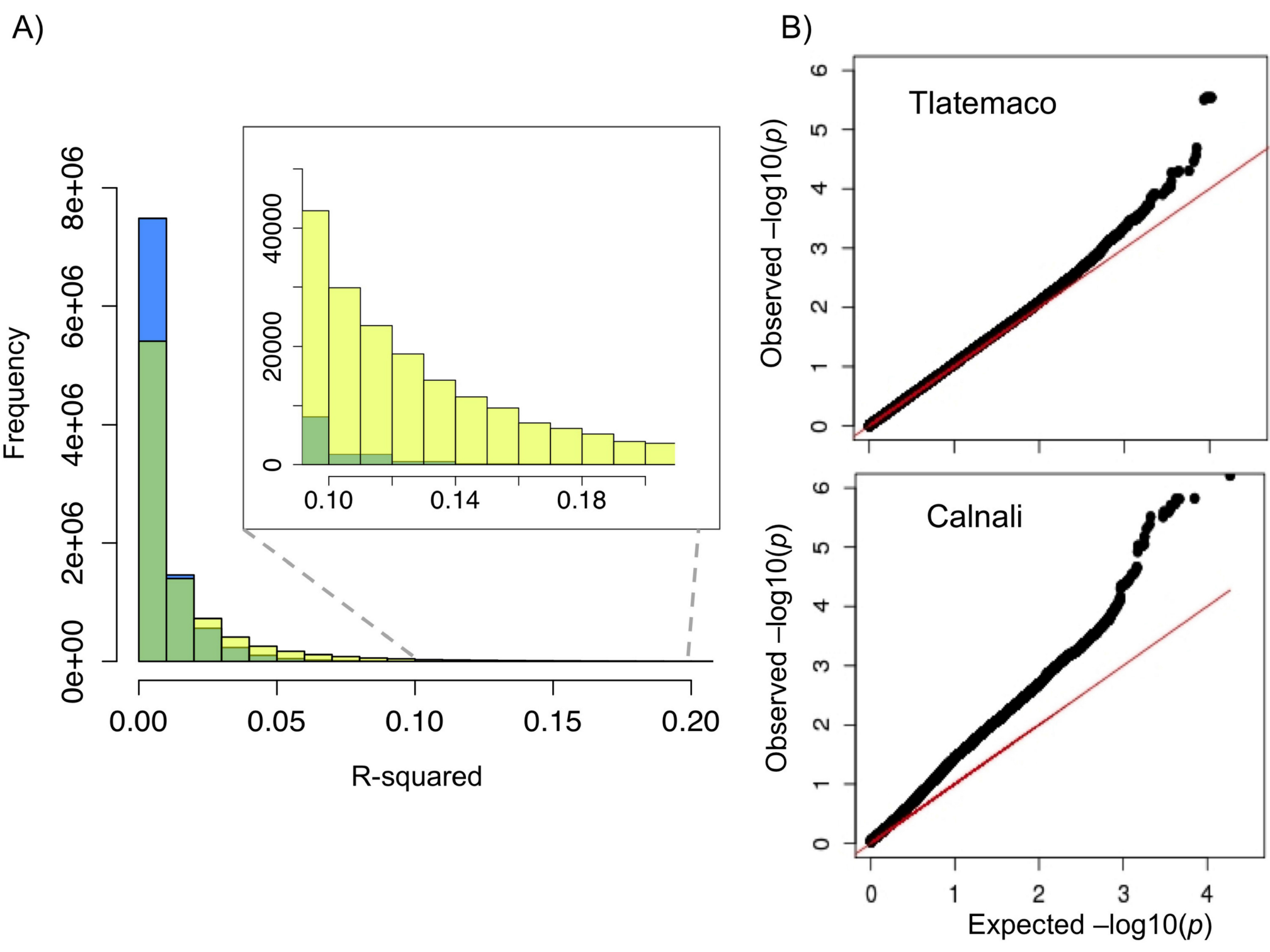
Supplementary File 1A: Pairs of regions in significant linkage disequilibrium (FDR 5%).

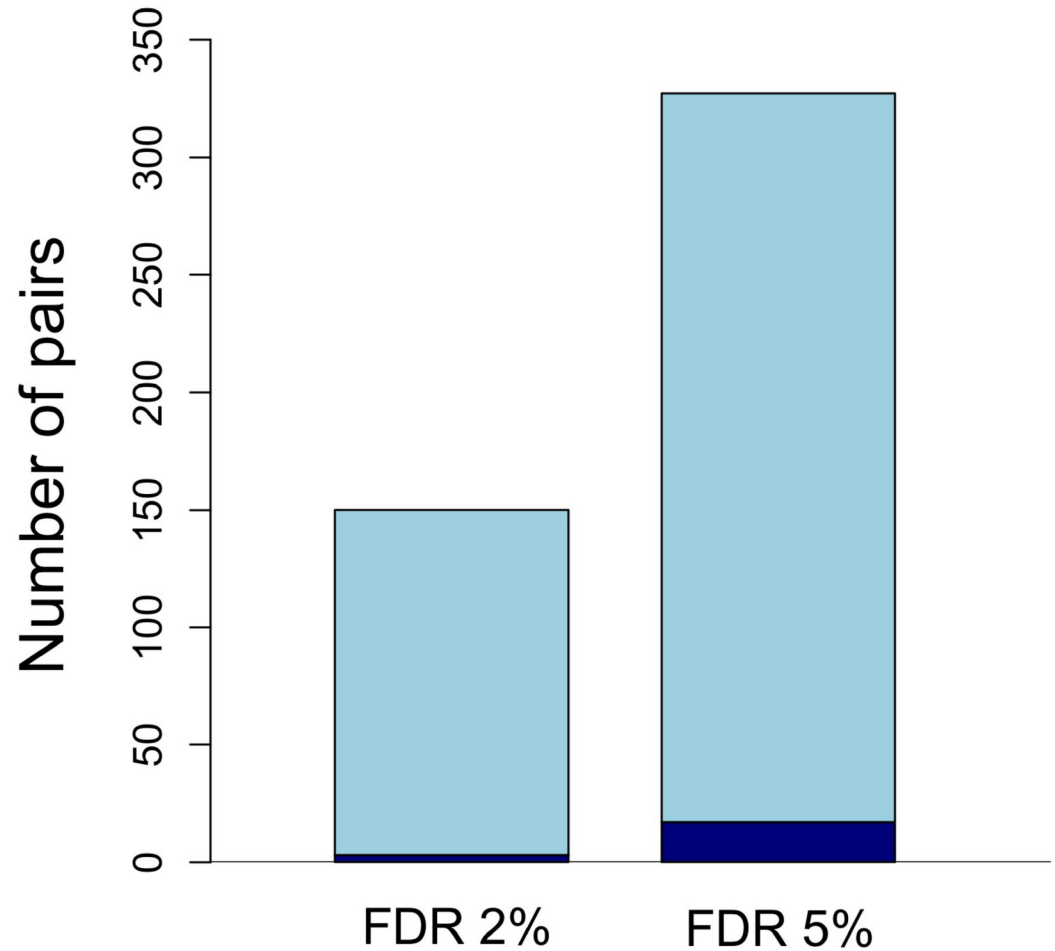
Supplementary File 1B: Pairs of LD regions (FDR 5%) that have single-gene resolution.

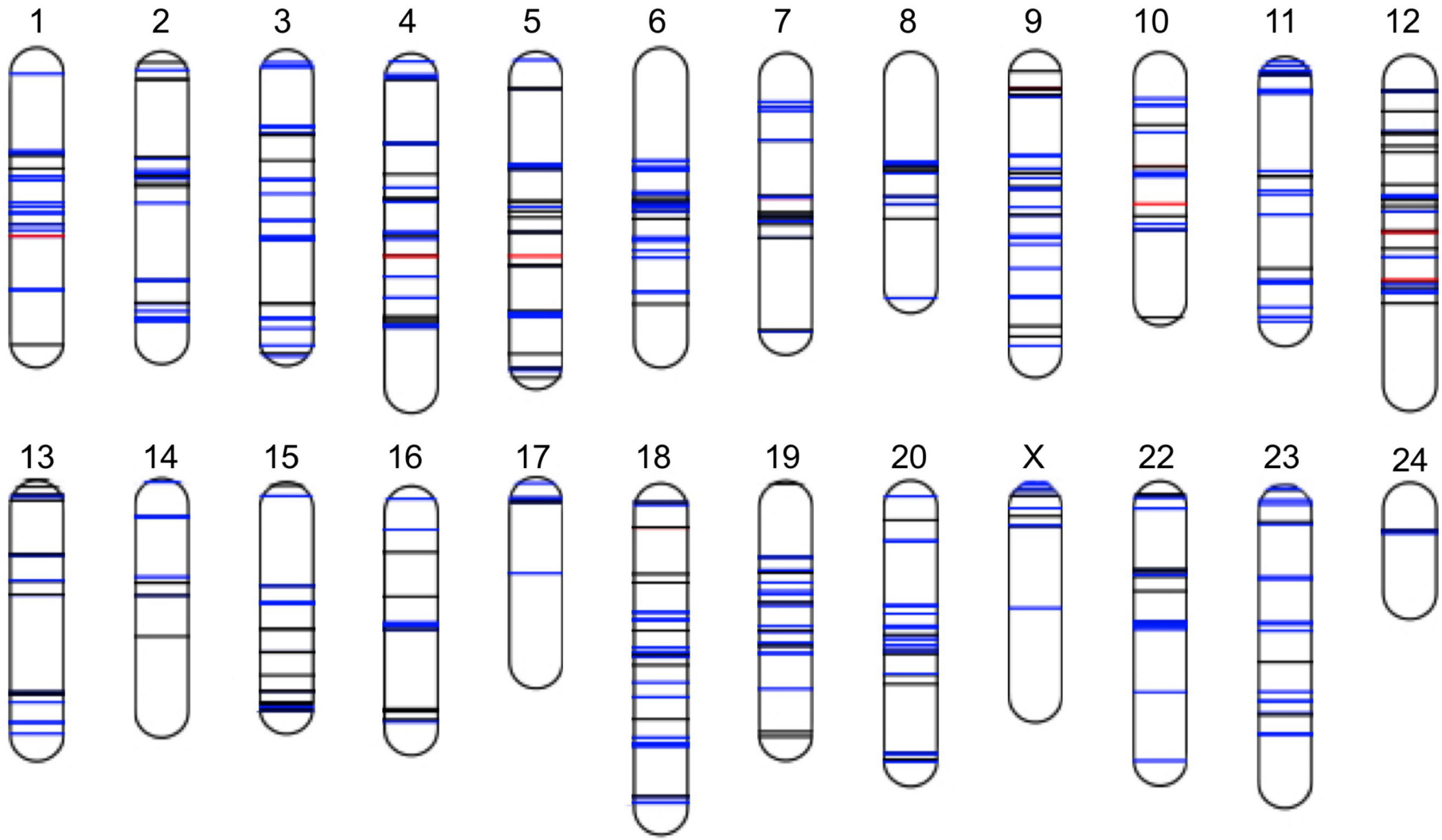
Supplementary File 1C: Divergence analysis for full dataset including pairs heterospecific in one or both populations.

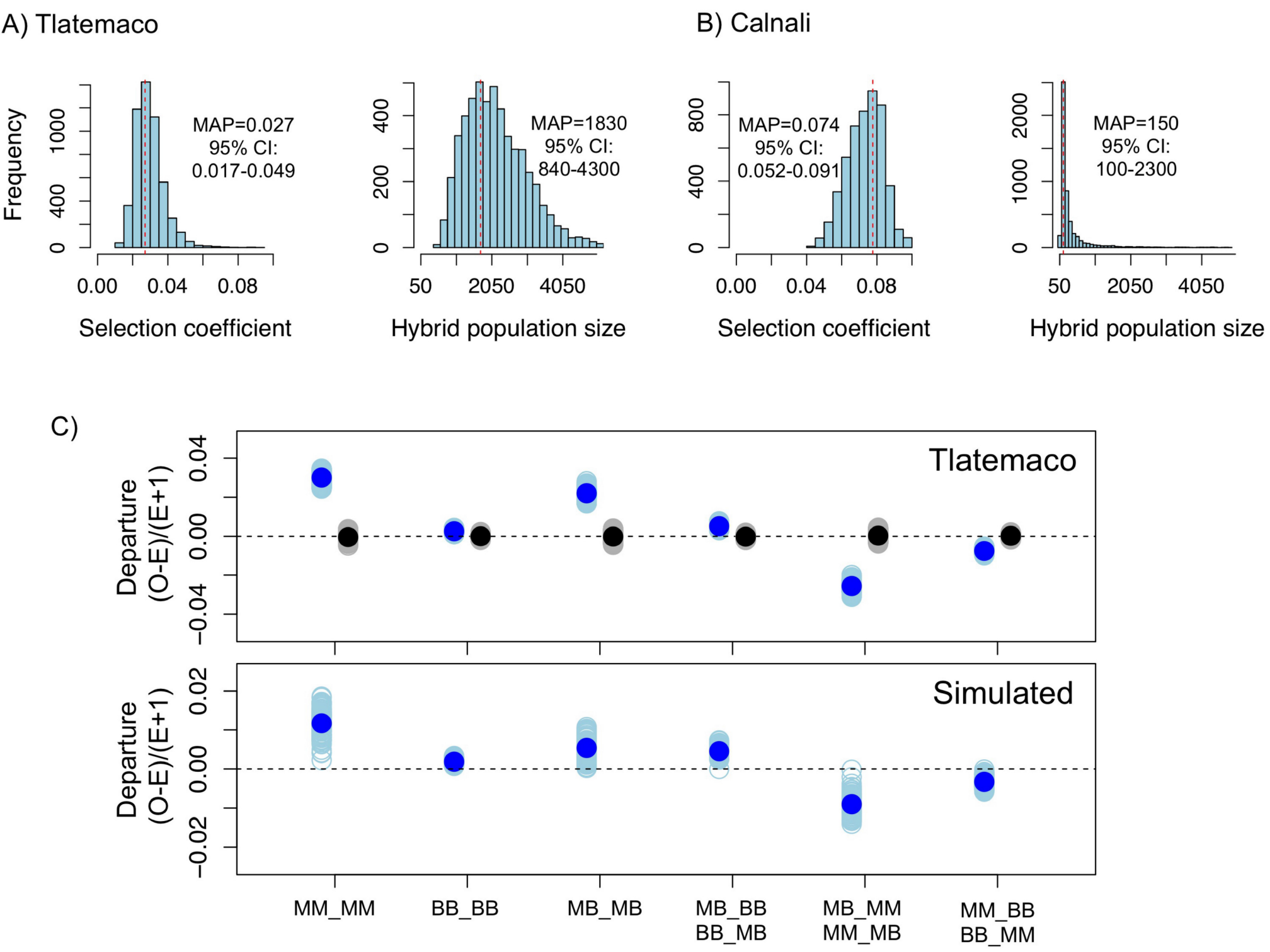

A)

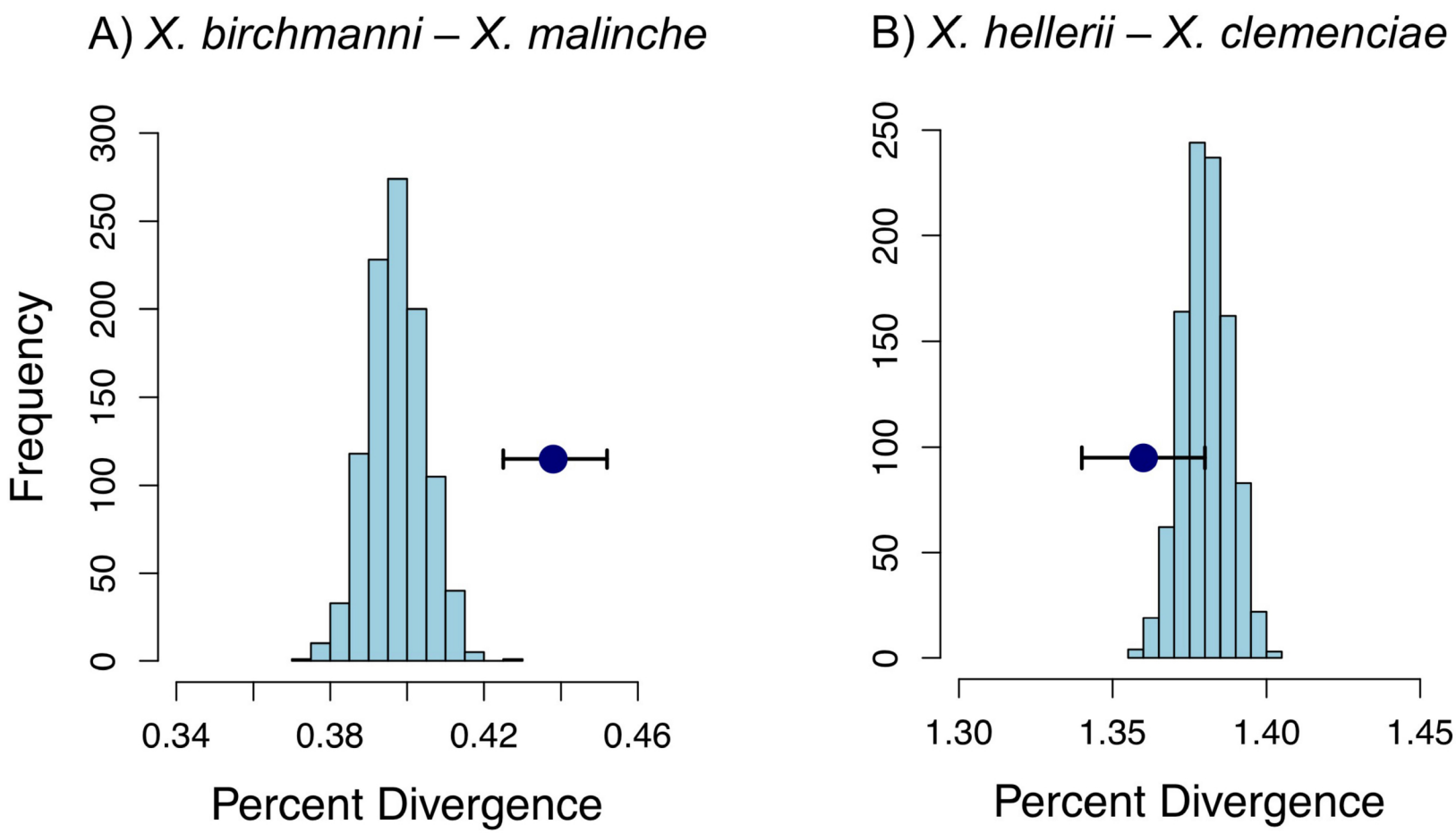

B)


A)
B)
Frequency
0e+00
2e+06
4e+06
6e+06
8e+06
0.00
0.05
0.10
0.15
0.20
R-squared
0
20000
40000
0.10
0.14
0.18
Tlatemaco
Calnali
Observed –log10(p)
Expected –log10(p)
0
1
2
3
4
5
6


Number of pairs
0
50
100
150
200
250
300
350
FDR 2%
FDR 5%


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
X
22
23
24


A) Tlatemaco
B) Calnali
Frequency
MAP=0.027
95% CI:
0.017-0.049
Selection coefficient
MAP=1830
95% CI:
840-4300
Hybrid population size
MAP=0.074
95% CI:
0.052-0.091
Selection coefficient
MAP=150
95% CI:
100-2300
Hybrid population size
C)
Tlatemaco
Simulated
Departure
(O-E)/(E+1)
MM_MM
BB_BB
MB_MB
MB_BB
BB_MB
MB_MM
MM_MB
MM_BB
BB_MM


A) X. birchmanni – X. malinche
Frequency
0
50
100
150
200
250
300
0.34
0.38
0.42
0.46
Percent Divergence
B) X. hellerii – X. clemenciae
0
50
100
150
200
250
1.30
1.35
1.40
1.45
Percent Divergence