Genome divergence and gene flow between *Drosophila simulans* and *D. mauritiana*

Sarah B. Kingan, Anthony J. Geneva, Jeffrey P. Vedanayagam, and Daniel Garrigan

Department of Biology, University of Rochester, Rochester, New York

Running title: Gene flow between allopatric *Drosophila*

Key words: *Drosophila*; genome; introgression, speciation

Corresponding author:

Daniel Garrigan

Department of Biology

University of Rochester

Rochester, New York 14627

Phone: +1-585-276-4816

Email: dgarriga@ur.rochester.edu

ABSTRACT

The fruit fly *Drosophila simulans* and its sister species *D. mauritiana* are a model system for studying the genetic basis of reproductive isolation, primarily because interspecific crosses produce sterile hybrid males and their phylogenetic proximity to *D. melanogaster*. We present an analysis of whole-genome patterns of polymorphism and divergence that shows, on average, the genomes of the two species differ at slightly more than 1% of nucleotide positions and an estimated 40% of autosomal and 60% of X-linked loci are reciprocally monophyletic. However, the analysis also identifies 21 major genomic regions, comprising ~1% of the genome, in which one species is segregating for haplotypes that are more similar to haplotypes from the other species than expected, given the levels of sequence divergence in that genomic region. This disjoint distribution of interspecific coalescence times is consistent with recent introgression between the cosmopolitan *D. simulans* and the island endemic *D. mauritiana*. We find that the putatively introgressed regions are more likely to have significantly higher rates of crossing-over and are enriched for genes with significantly slower rates of protein evolution. We also uncover instances in which genes experiencing lineage-specific positive selection closely interact with genes experiencing introgression. Finally, we find that a large introgressing region on the X chromosome has experienced a strong selective sweep in *D. mauritiana* and also has high levels of homozygosity in *D. simulans*. A detailed analysis reveals that the introgressing X chromosome haplotypes are closely associated with the presence of the *MDox* locus, which is the progenitor of the Winters *sex-ratio* meiotic drive genes. These results highlight how genetic systems that evolve rapidly in allopatry, including selfish meiotic drive elements, remain robust in natural hybrid genotypes and do not systematically promote reproductive isolation.

INTRODUCTION

In their original formulations of the biological species concept, Dobzhansky and Mayr envisioned the process of speciation as being complete only when gene flow ceases to occur between two diverging populations (DOBZHANSKY 1937; MAYR 1942). Yet both authors later revised their thinking to incorporate the possibility of limited gene flow between taxa that nonetheless exhibit partial reproductive isolation (DOBZHANSKY 1951; MAYR 1963). In these circumstances, many of the architects of the Modern Synthesis assumed that gene flow would necessarily be limited because natural selection was expected to remove most introgressing genomic material, so as not to "destroy the integrity of the two species" (WRIGHT 1978). The modern view on this problem emphasizes that species boundaries can be considered a continuum; one that is based on how individual genetic systems diverge and interact over time (HARRISON 1990; WU 2001).

Under a simple model, reproductive isolation evolves as an incidental byproduct of divergence in allopatry (COYNE AND ORR 2004). Upon secondary contact and hybridization, the magnitude of introgression may be variable and depend upon both the frequency of hybridization and the genomic distribution of substitutions that cause postzygotic isolation. There are only a handful of well-characterized examples of formerly allopatric populations, which have evolved substantial postzygotic isolation, coming into secondary contact and experiencing genomic introgression. The genomes of populations that form stable hybrid zones and presumably experience high levels of gene flow tend to show only localized regions of autosomal differentiation and high overall levels of differentiation on sex chromosomes (*e.g.*, TEETER *et al.* 2008). Among more diverged taxa, such as reproductively isolated types of the sea quirt *Ciona intestinalis*, there may be ample opportunity for hybridization, but the observed patterns of genomic introgression tend

to be restricted (ROUX *et al.* 2013). When introgression is limited by either high densities of incompatibilities or low frequency of hybridization events, there are examples of natural selection driving introgression events (BRAND *et al.* 2013). However, the generality of the role played by natural selection in driving introgression between populations at an advanced stage of speciation remains unclear.

The allopatric species pair of *Drosophila simulans* and *D. mauritiana* are a classic model system for studying the genetic basis of hybrid sterility (COYNE 1984). Together with *D. sechellia*, these two species are also the closest known relatives of the model organism *D. melanogaster*. The cosmopolitan human commensal *D. simulans* is estimated to have diverged from the island endemic *D. mauritiana* approximately 240 thousand years before the present (LACHAISE *et al.* 1986; KLIMAN *et al.* 2000; GARRIGAN *et al.* 2012). Crosses between the two species yield sterile F1 hybrid (*XY*) males and fertile F1 hybrid (*XX*) females and studies estimate that there may be at least 60 genes involved in hybrid male sterility between these species (TAO *et al.* 2003a; TAO AND HARTL 2003; TAO *et al.* 2003b). Likewise, a smaller number of factors have been recovered that cause inviability and female sterility (TRUE *et al.* 1996b). The two species also exhibit asymmetrical sexual isolation, based largely on female preference, in which female *D. mauritiana* discriminate against *D. simulans* males (MOEHRING *et al.* 2004). Yet despite substantial levels of both pre- and postzygotic reproductive isolation, both mitochondrial (BALLARD 2000) and nuclear (GARRIGAN *et al.* 2012) sequence data suggest introgression has occurred recently between these two species. These putative introgression events offer a valuable opportunity to study both the genomic context and potential forces driving gene flow between taxa at an advanced stage of reproductive isolation.

In this study, we use whole-genome resequence data from 10 *D. mauritiana* and 20 *D. simulans* inbred lines to identify and characterize the genomic regions that have experienced recent gene flow. Our approach to detecting introgression is purposefully conservative, as we seek to minimize type II error and focus primarily on recent introgression events that have occurred well after the evolution of reproductive isolation between these two species. In addition, we leverage the vast amount of functional annotation available for *D. melanogaster* to ask whether the putatively introgressing regions share any distinguishing functional characteristics. In total, we examine eleven different functional characteristics and find that the introgressed regions are located in highly recombining genomic intervals and are enriched for genes with a reduced rate of protein evolution compared to the genomic background. Finally, we uncover compelling evidence that a meiotic drive gene has recently introgressed between *D. simulans* and *D. mauritiana*. The introgressed haplotype containing the *MDox*/*Dox* meiotic drive element has swept to fixation in the *D. mauritiana* population and occurs at intermediate frequency in *D. simulans*. The high frequency of the haplotype in both species suggests that it has recently experienced drive and the pattern of introgression suggests that selfish genetic elements are not deterministic factors promoting postzygotic reproductive isolation.

MATERIALS AND METHODS

Samples and short read alignment: We sample 10 lines of *D. mauritiana*: nine wild isolates and the inbred laboratory line, *mau w* 12, (14021-0241.60), 20 lines of *D. simulans*: 10 wild isolates from Kenya (14021-0251.302-311), 9 wild isolates from Madagascar (14021-0251.293-301), and the genome reference strain, w^{501} (14021-0251.011), the reference strain of *D. sechellia*, (14021-0248.25), and the reference strain of *D. melanogaster* (*y;cn bw sp*). Data are reported previously (GARRIGAN *et al.* 2012; GARRIGAN *et al.* 2014; ROGERS *et al.* 2014). We

perform short read alignment against the *D. mauritiana* genome (version 2) using the "aln/sampe" functions of the BWA short read aligner and default settings (LI AND DURBIN 2009). Reads flanking indels are realigned using the SAMTOOLS software (LI *et al.* 2009). Individual BAM files are merged and sorted with SAMTOOLS.

Genome scans

Within and between population summary statistics: Both within- and between-population summary statistics are estimated in 10 kb windows using the software package POPBAM (GARRIGAN 2013). Within both the *D. mauritiana* and *D. simulans* populations, the estimated statistics include: unbiased nucleotide diversity π (NEI 1987), the summary of the folded site frequency spectrum Tajima's *D* (TAJIMA 1989), and the unweighted average pairwise value of the r^2 measure of linkage disequilibrium, Z_{nS}, excluding singletons (KELLY 1997). The between population statistics include: two measures of nucleotide divergence between populations, D_{XY} and net divergence, D_A (NEI 1987), the ratio of the minimum between-population nucleotide distance to the average, $G_{\min}$ (GENEVA *et al.* 2015), and the traditional fixation index, F_{ST} (WRIGHT 1951). From a total of 11083 scanned 10 kb windows, we only analyze windows with at least 50% of aligned sites passing the default quality filters in POPBAM, which results in a final alignment for 10443 scanned 10 kb windows. POPBAM output is formatted for use in the R statistical computing environment using the package, POPBAMTools (https://github.com/geneva/POPBAMTools). All statistics and data visualization of genome scans are also performed using the R software package (R CORE TEAM 2014).

Analysis of coding regions: We generate sequence alignments for the CDS of all known transcripts in *D. mauritiana*. Transcripts are annotated as described previously (GARRIGAN *et al.* 2014), except that reads are mapped to version 2 of the *D. mauritiana* reference genome

sequence. Variable sites are called using the POPBAM "snp" function with default settings. Sequence alignments comprising the 20 *D. simulans* lines, 10 *D. mauritiana* lines, and the *D. melanogaster* reference strain are constructed for all identified CDS (longest transcript per gene) using the Perl script, PBsnp2fa.pl (https://github.com/skingan/PBsnp2fa.pl).

The McDonald-Kreitman (MK) framework (McDonald and Kreitman 1991) is used to test a neutral model of protein evolution for 11530 CDS alignments. Unpolarized MK tests are performed using the Perl script CDS2SFS.pl (https://github.com/skingan/CDS2SFS), which generates the unfolded site frequency spectrum (included the fixed class) for synonymous and nonsynonymous sites for the 20 *D. simulans* lines using the *D. melanogaster* reference genome as the outgroup. Furthermore, for each gene alignment, we determine the significance of each MK test using a one-tailed Fisher's exact test, which explicitly tests whether there is an excess of nonsynonymous substitutions. To determine significance of the deviations from expectations, we correct for multiple tests, using the method described in Garrigan *et al*. (2012). We use $P < 0.001$ as the threshold for determination of significance, which corresponds to a proportion of tests that are truly null to be 0.984 and a false discovery rate of 18.3%. We do not perform MK tests that contrast divergence and polymorphism between *D. simulans* and *D. mauritiana*, because our primary motivation for the analysis is to determine whether putatively introgressed regions are more or less subject to natural selection over the long-term than a random sample from the genome. Introgression would have the net effect of reducing the number of substitutions between *D. simulans* and *D. mauritiana*, while inflating the count of polymorphic sites in the population that receives the introgressing sequences.

Analysis of introgression

Identification of introgressed regions: We use the G_{min} sequence measure to scan the genome for intervals that have recent common ancestry between *D. simulans* and *D. mauritiana* (GENEVA *et al.* 2015). G_{min} is defined as the ratio of the minimum number of nucleotide differences per aligned site between sequences from different populations to the average number of nucleotide differences per aligned site between populations. The G_{min} statistic is calculated in 10 kb intervals across each major chromosome arm, using the same filtering criteria that are used for all other summary statistics. From these values, we wish to calculate the probability of the observed G_{min} under a model of allopatric divergence, conditioned on the divergence time. For each 10 kb interval, the significance of the observed G_{min} value is tested via Monte Carlo coalescent simulations of two populations diverging in allopatry. All mutations in the simulations are assumed to be neutral. The population divergence time is set to $1.21 \times 2N_{sim}$ generations before the present, in which N_{sim} is current effective population size of *D. simulans* (GARRIGAN *et al.* 2012). In the simulations, the observed local value of D_{XY} is used to determine the neutral population mutation rate for that 10 kb interval. To account for uncertainty in local population recombination rate, for each simulated replicate, the rate is drawn from a normally distributed prior (truncated at zero) with the mean calculated from the empirical crossing-over frequencies estimated by (TRUE *et al.* 1996a). The empirical crossing-over estimates are converted from cM to ρ (the population crossing-over rate, $4N_{sim}c$, by assuming $N_{sim} \approx 10^6$). The effective population sizes of both species are assumed to be equal and constant. While this assumption and that of selective neutrality of mutations may not approximate reality very well, it does serve to make the test more conservative, because any factor that decreases the effective population size also increases the rate of coalescence within populations, which has the net effect

of increasing G_{min} and decreasing its variance, since fewer ancestral lineages will be available for inter-specific coalescence in the ancestral population (GENEVA *et al.* 2015). For each 10 kb interval, 10^5 simulated replicates are generated and the probability of the observed G_{min} value is estimated from the simulated cumulative density. The threshold for statistical significance is corrected for multiple tests using the same method as described for the Fisher's exact test *P*-values from the MK tests. We observe that the Monte Carlo-based test is much more liberal than the Fisher's exact test and the threshold for significance in this case is adjusted to $P < 0.0001$, at this threshold the proportion of null tests is 0.986 with a false discovery rate of 14.67%. Finally, maximum likelihood phylogenies are constructed for each of the putative introgression intervals using RAxML v. 8.1.1 (STAMATAKIS 2014).

Functional analysis of introgressed regions: Eleven different functional characteristics are measured for differences between the putatively introgressed genomic regions, or genes, compared to the background genomic distribution and are listed in **Table 1**. For the gene-based comparisons, gene ontology (GO) terms of the putatively introgressed regions are examined for both enrichment and underrepresentation of biological process and molecular function terms and the analysis uses the *D. melanogaster* annotation as a reference with correction for multiple tests (MI *et al.* 2013). The number of protein-protein interactions for the putatively introgressed coding sequences is obtained from the results of a previously published coaffinity purification coupled with mass spectrometry analysis of 4385 proteins in *D. melanogaster* (GURUHARSHA *et al.* 2011). Three different aspects of gene expression were examined: magnitude of expression, sex-bias, and tissue-specificity. For the first two aspects, we use sex-specific gene expression estimate from whole flies using previously published RNA-seq data (GARRIGAN *et al.* 2014). Short read data are mapped to version 2 of the *D. mauritiana* genome using the TopHat (version

2.0.13) of the Cufflinks software suite (KIM *et al.* 2013). Gene expression estimates in units of fragments per kilobase of transcript per million reads (FPKM) are obtained for single replicates of pooled mated and virgin adult males and females at the gene-level using Cuffdiff v.2.2.1 (TRAPNELL *et al.* 2013). We test for differences in the median sex-specific expression between genes in candidate introgression windows *versus* the whole genome. Significance of the test is assessed with 10^3 bootstrap resampling of *k* genes without replacement from the genome. Sex-biased genes are defined as having a four-fold or greater difference in expression between the sexes. A Fisher's exact test was used to compare observed *versus* expected counts of sex-biased genes in introgression windows where expectations are based on the proportion of sex-biased genes in the genome. Tissue-specific gene expression estimates for 11331 genes across 21 tissues for *D. melanogaster* are obtained from MEIKLEJOHN AND PRESGRAVES (2012), which is based on the FlyAtlas data (CHINTAPALLI *et al.* 2007). We estimate the tissue-specificity of each gene using the metric, τ, which ranges from 0 (broadly expressed) to 1 (tissue-specific) (YANAI *et al.* 2005) and designate tissue specific genes as having $\tau \geq 0.90$. We test for a difference in median tissue-specificity in introgression regions *versus* the whole genome using the same bootstrapping method as the sex-specific expression analysis. In addition, the relative age of each gene in the *Drosophila* phylogeny is contrasted between putatively introgressed and the background set of genes using data from ZHANG *et al.* (2010) and a Fisher's exact test based on observed and expected counts of genes in each age class.

We also ask whether the genes encoded in the putatively introgressed 10 kb intervals are enriched for genes experiencing either elevated or reduced rates of protein evolution. The rate of protein evolution is measured using the "Direction of Selection" (DoS) statistic, which quantifies the difference between the proportion of substitutions between *D. simulans* and *D. melanogaster*

that are nonsynonymous and the proportion of polymorphism in *D. simulans* that are nonsynonymous. The DoS measure is calculated as $[D_N/(D_N + D_S)] - [P_N/(P_N + P_S)]$, where D_N and D_S are the counts of nonsynonymous and synonymous substitutions, respectively, between *D. melanogaster* and *D. simulans* and P_N and P_S are the respective counts of nonsynonymous and synonymous polymorphic sites in *D. simulans*. The DoS statistic is more robust to sparse marginal counts than are other summaries of the traditional MK 2 × 2 contingency table (STOLETZKI AND EYRE-WALKER 2011) and its value is negative when there is an excess of nonsynonymous polymorphisms relative to nonsynonymous substitutions and positive when there is a relative excess of nonsynonymous substitutions. Our analysis asks whether the DoS in the pooled set of putatively introgressed genes is significantly greater or less than the DoS calculated from a pooled set of genes randomly sampled from the genome. This is achieved by summing D_N, D_S, P_N, and P_S for all genes in the pooled set and calculating DoS from the sums. The null hypothesis that there is no difference in the rate of protein evolution for the putatively introgressed genes is tested via a bootstrap analysis in which a set of *k* genes are sampled with replacement from the genome and the DoS is calculated for that resampled set. The two-tailed probability of the observed DoS is determined from the bootstrap cumulative distribution, based on 10^6 replicates. Lastly, a modified version of this test includes testing whether the introgressed genes are significantly different from the remainder of the genome in the proportion of significant unpolarized MK one-tailed Fisher's exact tests, as described in the preceding subsection.

For the region-based analyses, we consider whether putatively introgressed regions have different protein-coding potential than the genome at large. This hypothesis is tested by tallying the percent of each 10 kb interval that is included in a CDS annotation in the *D. mauritiana*

genome (GARRIGAN *et al.* 2014). We also ask whether introgressing regions have systematically different substitution rates than the rest of the genome. To test this hypothesis, we calculate net divergence (D_A) from *D. melanogaster* for 15394 short introns, a sequence class with low levels of functional constrains (PARSCH *et al.* 2010), 161 of which occur in introgression windows using POPBAM. We analyze introns shorter than 70 bp, excluding the first six and the last two bases, which contain splice signals and are thus constrained (HALLIGAN *et al.* 2004). Only short intron regions with at least 50% of sites passing the default POPBAM quality filters are included. We calculate D_A with *D. melanogaster* separately for *D. simulans* and *D. mauritiana* and analyze the X and autosomes both separately and combined. We compare the median D_A for introns within introgression windows to the genomic median and use 10^3 bootstrap replicates without replacement to determine significance. Finally, we ask whether the introgressed regions have significantly different rates of crossing-over, as measured by TRUE *et al.* (1996a). This test is conservative due to the low density of markers genotyped in that study. We did not contrast other proxies for recombination frequency (for example, Z_{nS}), because these within-population statistics may be biased by the presence of long, recently introgressed haplotypes.

Post hoc analysis of the *MDox*/*Dox* region

Genotyping the Winters *sex-ratio* genes: Genomic DNA from single male flies is extracted using the Qiagen Dneasy Blood and Tissue Kit. The meiotic drive genes of the Winters *sex-ratio* system (TAO *et al.* 2007a), *Dox* and *MDox*, are PCR amplified as previously described (KINGAN *et al.* 2010). The PCR product for *Dox* is digested with the restriction enzyme, StyI (NEB) to assay the presence or absence of the *Dox* gene insertion. Similarly, the *MDox* PCR products are digested with the StuI restriction enzyme (NEB). In both cases, only haplotypes containing the gene insertions have restriction sites as confirmed by samples with known genotypes. Finally,

the digests are run on a 1% agarose gel stained with EtBr and the band size was estimated using the GeneRuler 1 kb plus ladder (Thermo Scientific).

Quantitative PCR for *Dox*/*MDox* expression in fly testes: Expression of the *Dox* and *MDox* loci in fly testes is measured in both *D. simulans* strain MD63 and *D. mauritiana* strain *mau w* 12 using quantitative PCR. Approximately 15 testes are extracted in Ringer's solution from 5-10 day old flies. Total RNA is then extracted from the testes samples using Nucleospin RNA XS kit (Macherey-Nagel, Germany), and cDNA is synthesized using both poly dT oligos and random hexamers using Superscript III RT cDNA synthesis kit (Invitrogen, CA). Finally, qPCR is performed on a Bio-Rad Real-time PCR machine using the following cycling conditions: 95° C for 3 mins.; 40 cycles of 95° C for 10s, 58° C for 30s, and 72° C for 30s. The primer sequences used for qPCR are provided in **Table S1**.

RESULTS

Previous studies of the *D. simulans* clade characterized levels of both whole-genome divergence (GARRIGAN *et al.* 2012) and polymorphism individually within *D. mauritiana* (NOLTE *et al.* 2013; GARRIGAN *et al.* 2014) and *D. simulans* (BEGUN *et al.* 2007; ROGERS *et al.* 2014). Here, we present the first analysis of whole-genome variation both within and between *D. mauritiana* and *D. simulans*. The primary advantage of this approach is that it enables a fully integrated analysis of the effects of natural selection on genome polymorphism and divergence, and has the added benefit of increased power to detect low and intermediate frequency introgressed haplotypes.

Patterns of sequence polymorphism and divergence: We measure sequence polymorphism and divergence within and between *D. simulans* and *D. mauritiana* in non-overlapping 10 kb windows across each of the four major autosomal arms and the X chromosome (**Figure 1**). The

median number of pairwise sequence differences per site (D_{XY}) between the two species is 0.013 for the autosomes and 0.010 for the X chromosome (**Figure 2**). However, because the X chromosome has substantially lower levels of within-species polymorphism, the median net divergence (D_A) between the species is −0.0005 for the autosomes and 0.0007 for the X chromosome. A negative value of D_A on the autosomes indicates that, on average, levels of within-species polymorphism are higher than between-species divergence. These estimates correspond to approximately 41.8% of autosomal and 59.7% of X-linked windows being reciprocally monophyletic ($D_A > 0$). The cosmopolitan *D. simulans* has approximately 1.23 times greater polymorphism on the autosomes than does the island endemic *D. mauritiana* and 1.44 times greater polymorphism on the X chromosome. The median ratio of polymorphism on the X chromosome *versus* the autosomes is 0.656 for *D. mauritiana* and 0.778 for *D. simulans*. Additionally, both the autosomes and X chromosome of *D. simulans* have more low frequency polymorphisms than *D. mauritiana*, as evidenced by a more negative median value of Tajima's *D* statistic (**Figure 2**). In *D. simulans*, the median Tajima's *D* on the autosomes is −1.127 and −1.218 on the X chromosome, while in *D. mauritiana* the median Tajima's *D* on the autosomes is −0.359 and on the X chromosome is −0.536. In both cases, *D. simulans* has significantly more negative Tajima's *D* values than does *D. mauritiana* (Mann-Whitney *U* test; $P_{\mathrm{MWU}} < 0.001$).

Levels of linkage disequilibrium are substantially higher in *D. mauritiana*, across all compartments of the genome (**Figure 2**). For *D. simulans*, the median measure of linkage disequilibrium, calculated as the unweighted average r^2 (Z_{nS}), is 0.058 and 0.056 for the autosomes and X chromosome, respectively. In contrast, the median *D. mauritiana* Z_{nS} for the autosomes is 0.129 and 0.122 for the X chromosome, both of which are significantly higher than *D. simulans* ($P_{\mathrm{MWU}} < 0.001$). Finally, the index of differentiation, F_{ST}, is also higher for the X

chromosome (median F_{ST} = 0.378) than the autosomes (median F_{ST} = 0.279) and the ratio of X to autosomal F_{ST} is 1.355, close to the expected neutral value of 4/3 (if the effective population size of the X chromosome is 3/4 that of the autosomes). A genome-wide scan for F_{ST} is presented in **Figure S1**.

Analysis of introgression

Identifying introgressing regions of the genome: Introgressing regions of the genome are identified using the G_{min} measure of between-species sequence divergence. G_{min} is the ratio of the minimum pairwise sequence distance between species to the average pairwise distance between species. This measure is designed to be sensitive to coalescent genealogies with recent gene flow, especially when the introgressed sequence is low or intermediate frequency (GENEVA *et al.* 2015). The median and median absolute deviation of G_{min} estimates across 10 kb intervals for each chromosome arm varies from 0.761 ± 0.0537 for chromosome *3L* to 0.785 ± 0.0531 for chromosome *X* (**Figure 3**). As all loci in the genome approach reciprocal monophyly, G_{min} will approach one with zero variance, because when divergence is higher there will only be two ancestral lineages from each population available for inter-specific coalescence in the ancestral population and, by definition, the minimum distance (numerator) will be equal to the mean pairwise distance (denominator). Genomic intervals that are inconsistent with a purely allopatric divergence model are identified by a Monte Carlo simulation procedure that assumes that the true population divergence time is constant across all 10 kb intervals, separately for the X and the autosomes. Local variation in recombination rate is accounted for by assuming a prior distribution and integrating over all sampled parameter values.

In total, 138 of the 10443 10 kb genomic intervals (1.32%) have significantly more recent common ancestry between *D. simulans* and *D. mauritiana* than expected in a purely allopatric

divergence model, as indicated by significantly low values of G_{min}. These 138 significant intervals are arrayed into 21 distinct contiguous genomic regions (**Figure 1**). The length of the inferred introgression tracts varies from a single 10 kb interval to 230 kb in low-recombination pericentric regions (**Table 2**). However, the largest putatively introgressed region that occurs in the middle of a chromosomal arm spans 200 kb (2.944 cM) between coordinates 13870000-14070000 on chromosome *3R*. Intriguingly, this large region is immediately proximal to a region that has previously been shown to have experienced gene flow and simultaneous positive selection in both *D. simulans* and *D. sechellia* (BRAND *et al.* 2013). The putatively introgressed genomic intervals contain 143 genes with homologs in *D. melanogaster* (**Table S2**). Phylogenetic trees of all sequences in each of the 21 putatively introgressed windows reveals that, in the majority of cases, the low observed values of G_{min} are consistently the result of interspecific clustering of one to four *D. mauritiana* sequences within an otherwise monophyletic *D. simulans* clade (**Figure S2**).

Using RNA-seq data (GARRIGAN *et al.* 2014) and functional annotations from *D. melanogaster*, we perform several tests to determine whether there are any common characteristics to the 143 genes that introgress between *D. simulans* and *D. mauritiana*. We divide our analysis of function into two separate categories: analysis of the functional properties of the genomic regions themselves and functional analysis of the genes located within those regions.

Functional analysis of introgression regions: First, we consider whether the genomic region that are identified as introgressed bear any distinguishing features when compared to the background genomic distribution (**Table 1**). For this region-based comparison, we test whether the putatively introgressed genomic intervals differ with respect to three properties: 1) the amount of coding sequence in the interval, 2) the overall net substitution rate between

D. simulans and *D. melanogaster*, and 3) the recombination rate, as determined from a coarse-scale map of crossing-over events in a laboratory population. The assumption that inter-genic, non-coding sequence should have fewer deleterious fitness effects in hybrid individuals predicts that putatively introgressed regions should contain fewer genes. However, we find that 14.75% of base positions in introgressed regions encode CDS sequence, which does not significantly differ from the average of 14.73%, for both the X and the autosomes ($P_{MWU} > 0.05$). In addition, to test whether introgressed regions may have a lower mutation rate, we compared the median net divergence (D_A) between *D. simulans* and *D. melanogaster* for short introns, a relatively unconstrained sequence class (PARSCH *et al.* 2010). The net substitution rate per site in introgressed windows is 3.51%, which is not significantly different from the whole genome estimate of 3.38% (95% CI: 2.26-4.05%). We find no deviation in the introgressed regions when the autosomes and X are analyzed separately, nor when the analysis is performed for the *D. mauritiana-D. melanogaster* species pair (**Table S3).** Finally, we do find that the putatively introgressed genomic regions have significantly elevated rates of crossing-over than the overall genomic background rate (bootstrap $P < 0.05$ with 10^4 replicates; **Table 1**).

Functional analysis of introgressed genes: In addition to measuring the properties of genomic regions that are putatively introgressed, we also measure properties of the 143 genes located in the introgressed regions. We examine seven gene-based characteristics: 1) gene ontology, 2) the number of protein-protein interactions, 3) male and female gene expression levels 4) tissue-specificity of gene expression, 5) gene age, 6) rate of amino acid substitution, and 7) proportion of genes subject to positive natural selection.

Of the 143 genes located in putatively introgressed regions, 130 are associated with both gene ontology biological process and molecular function terms. A statistical analysis using the

PANTHER classification system finds a significant enrichment for the biological process term "cell development" (GO:0048468; $P = 2.91 \times 10^{-5}$; FDR $q = 0.026$). Likewise, there is a significant enrichment for the molecular function term "DNA binding" (GO:0003677; $P = 1.16 \times 10^{-4}$; FDR $q = 0.0344$). However, all of the genes responsible for both of the above GO enrichment results are located in the same genomic interval on chromosome *2L*, between coordinates 2790000-2860000 (**Figure S3**) and likely do not represented repeated, independent introgression events.

The Bateson-Dobzhansky-Muller (BDM) model of reproductive isolation assumes incompatibilities arise due to failed epistatic interactions between genes carrying lineage-specific substitutions (ORR 1997). One prediction of the BDM model is that introgressing genes are more likely to encode proteins with fewer number of interacting partners, so as to minimize the probability of reduced fitness due to failed epistatic interactions. We find that the median number of protein-protein interactions for 49 of the putatively introgressed genes is 9, compared to the median of 13 interactions for a background set of 4385 genes. On this basis, we conclude that the introgressed genes do not significantly differ in the number of protein-protein interactions from a large genomic sample ($P_{MWU} = 0.2029$). Interestingly, several of the putatively introgressed genes are members of large protein complexes and encode highly connected protein products, including the phosphoglycerate mutase enzyme *Pglym78* ($n_I = 46$), the proteasome subunit *Rpn2* ($n_I = 35$), and the cytoskeletal protein *Act78B* ($n_I = 27$).

The magnitude and pattern of gene expression is predictive of the rate of evolution of a gene, both in terms of sequence evolution and divergence in gene expression. For example, highly expressed genes tend to diverge slowly both in primary sequence and expression pattern (DRUMMOND *et al.* 2006; LIAO *et al.* 2006; ZHANG AND YANG 2015), whereas genes with high

tissue-specificity (*i.e.*, not broadly expressed) show high rates of protein evolution (LARRACUENTE *et al.* 2008), but low rates of expression evolution (LIAO *et al.* 2006). Finally, sex-biased genes tend to be rapidly evolving at the sequence level (ELLEGREN AND PARSCH 2007), with male-biased genes also showing rapid evolution of gene expression (MEIKLEJOHN *et al.* 2003). Characterizing how patterns of gene expression may differ in introgression windows is important because BDM incompatibilities may also be caused by changes to gene regulation (LANDRY *et al.* 2007), resulting in transgressive expression, which has been implicated in some cases of hybrid male sterility (GOMES AND CIVETTA 2015).

To determine whether putatively introgressed genes show any bias in the magnitude of gene expression, we analyze RNA-seq data from whole-body male and female samples of the genome reference strain of *D. mauritiana* (GARRIGAN *et al.* 2014). We find that genes in the introgression windows do have a higher median expression level when compared to the rest of the genome (**Figure S4**). This elevation is statistically significant for male-specific expression (12.91 FPKM for the introgression genes *versus* 8.42 FPKM for the genome; $P = 0.009$) and maximum sex-specific expression (20.46 FPKM for the introgression genes *versus* 14.68 FPKM for the genome average; $P = 0.043$) but not for female-specific expression (8.42 FPKM for the introgression genes *versus* 6.54 FPKM for the genome; $P = 0.240$).

Because sex-biased genes, and male-biased genes in particular, tend to evolve rapidly, we may expect that this class of sequence will be less likely to introgress between species with partial reproductive isolation. In the putative introgression windows, we observe fewer male-biased genes than expected: 27 compared to 40, although this difference is not significant ($P_{FET} = 0.074$; **Table S4**). In contrast, there is no deviation from expectations for female-biased genes (13

observed *versus* 13 expected). The reduction in male-biased genes in introgression windows is less extreme when higher fold changes are used to define sex-specific expression (**Table S4**).

To assess the tissue-specificity of introgressed genes, we analyze expression data from *D. melanogaster* for 21 different tissue types (CHINTAPALLI *et al.* 2007; MEIKLEJOHN AND PRESGRAVES 2012). Unexpectedly, genes in the introgression windows show slightly higher tissue-specificity than the genome as a whole (**Figure S5**), although this difference is not statistically significant. The median tissue specificity, τ, for introgressed genes is 0.656 compared to the genome median of 0.546 ($P = 0.130$, 10^3 bootstrap replicates). More than 47% of the tissue-specific genes are testes-specific, which tend to be rapidly evolving in *Drosophila* and other organisms (ELLEGREN AND PARSCH 2007). If the testes-specific genes are removed and the above analysis is repeated, we find a significant elevation of τ for introgressed genes ($\tau = 0.580$ for the introgressed genes and $\tau = 0.413$ for the genomic background; $P = 0.016$). These results are consistent with tissue-specific genes having lower rates of gene expression divergence and thus being less likely to cause regulatory incompatibilities.

To test whether introgressed regions are more or less likely to harbor newer (more recently evolved) genes, we utilize previously published data that places > 11000 genes along seven different branches of various ages in the *Drosophila* phylogeny (ZHANG *et al.* 2010). Despite the prediction that introgressed genes are more likely to be older and conserved (YUAN *et al.* 2012), we find no bias in the number of introgressed genes originating along each branch (**Table S5**).

Finally, we ask whether the putative introgressing coding sequences are more or less likely to have a history of positive natural selection. This general hypothesis is addressed using two different approaches. Both approaches are based on the framework of polarized MK tests using counts of nonsynonymous and synonymous polymorphisms from the *D. simulans* population

sample and counts of nonsynonymous and synonymous substitutions between *D. simulans* and *D. melanogaster* reference genome. First, we test whether the "Direction of Selection" (DoS) statistic among the pooled group of introgressed genes differs from random samples of genes from the genome. Among the 88 introgressed genes for which we could unambiguously create CDS alignments, the pooled DoS = 0.0129. A null genomic background distribution of pooled DoS for 88 randomly selected genes is created using 10^6 replicates. The null distribution has a mean DoS = 0.0839 (**Table 1**). Consequently, the pooled DoS for introgressed genes is significantly lower than a random sample of genes (P = 0.00002). A second complementary approach asks whether the introgressed regions are enriched or depauperate for genes that are individually deemed to have histories of recurrent positive natural selection. This analysis indicates that none of the 88 introgressed genes, for which a MK test result exists, deviate significantly from neutral expectations, following correction for multiple tests. From a total of 8415 one-tailed Fisher's exact tests, there are 141 tests with $P < 0.001$ (1.67%). Since no putatively introgressed genes result in a significant test, we are not able to determine whether the introgressed regions are depauperate in positively selected genes, however we can reject the alternative hypothesis that the introgression regions are enriched for positively selected genes over the genomic background (P_{FET} = 0.4067).

When we closely examine the sequences of the putatively introgressed genes, we can document instances in which a gene encodes amino acid differences that are fixed between the introgressed and the endogenous *D. mauritiana* haplotypes. We find that nine of the 143 introgressed genes have nonsynonymous substitutions between the introgressed and *D. mauritiana* sequences. These genes are *Papilin, Dhc98D, Syncrip, CG17271, TweedleB, TweedleD, TweedleP, lilliputian,* and *mangetout*. We perform a more detailed analysis on five of these coding sequences, excluding

the Tweedle genes (which are tandem arrayed in a single 10 kb interval on chromosome *3R* and are excluded to avoid any uncertainty in variant calls due to paralogous sequence alignment) and the *CG17271* gene, which lacks functional annotations. **Figure S6** shows maximum likelihood phylogenetic trees of the complete coding sequences from each of the five above introgressed genes. *Dhc98D* is a dynein heavy chain involved in microtubule movement for which there is one *D. mauritiana* line that carries the *D. simulans*-like haplotype (**Figure S6A**). The introgressed haplotype of *Dhc98D* differs by three amino acid substitutions in an ATPase and D4 functional domains. The *D. simulans* haplotype of the *lilliputian* developmental transcription factor is present in three *D. mauritiana* lines and differs by three amino acid substitutions from the endogenous *D. mauritiana* haplotype, one in the AT hook DNA-binding domain (**Figure S6B**). The *mangetout* gene is a G-protein coupled receptor that has been implicated in response to insecticide (MITRI *et al.* 2009). The shared *mangetout* haplotype is present in one *D. mauritiana* line and has a single amino acid substitution in a seven-transmembrane domain (**Figure S6C**). The *Papilin* gene is a component of pericellular matrices in *Drosophila* embryos; the *D. simulans* haplotype is present in a single *D. mauritiana* line and shows 29 amino acid substitutions from the endogenous haplotype (**Figure S6D**). Finally, the *Syncrip* gene is involved in RNA localization during development; the *D. simulans* haplotype is present in one *D. mauritiana* line and has eight amino acid differences from the endogenous haplotype, three of which occur in RNA recognition motifs (**Figure S6E**).

Introgression of the Winters *sex-ratio* meiotic drive genes: There are two adjacent putatively introgressed regions on the X chromosome occurs between 8560000-8580000 and 8630000-8690000. Both of these regions harbor the Winters *sex-ratio* meiotic drive genes: *Distorter on the X* (*Dox*) and its progenitor gene, *Mother of Dox* (*MDox*) (TAO *et al.* 2007b). In

D. simulans, the presence of *Dox* and *MDox* results in biased transmission of the X chromosome during meiosis, causing males to produce mostly daughters. Both *Dox* and *MDox* are located within arrays of 359 bp satellite repeats, which may have facilitated the duplication and insertion of *MDox* ~70 kb downstream to produce the *Dox* locus. The Winters drivers are suppressed by an autosomal gene, *Not much yin* (*Nmy*), through a post-transcriptional RNA-interference mechanism (TAO *et al.* 2007b). The *Nmy* gene originated as a retrotransposed copy of *Dox* and bears sequence homology to both driver genes. In *D. simulans*, all three genes are nearly fixed, although ancestral haplotypes at each locus, which lack the gene insertions, are present at low frequency (KINGAN *et al.* 2010). All three genes also show a history of selective sweeps due to the transmission advantage of *Dox*/*MDox* and subsequent selection at *Nmy* to restore equal sex ratios (KINGAN *et al*. 2010).

We assay the genotypes of our sequenced strains for the presence/absence of both *MDox* and *Dox* to determine if the introgressed haplotype is associated with the presence of these two genes. The presence of *MDox* is almost perfectly associated with the introgressed haplotype: across both *D. simulans* and *D. mauritiana*: all but one of the introgressed haplotypes in this region carry *MDox* (only *D. simulans* strain NS79 has the introgressed haplotype, but does not have *MDox*). All *D. mauritiana* strains have both the introgressed haplotype and *MDox*, but are polymorphic for the presence of *Dox* (only two *D. mauritiana* strains have *Dox*, while all 10 lines have the introgressed haplotype, **Figure 4**; **Table S6**). In contrast, all seven *D. simulans* strains with the introgressed haplotype carry the *Dox* insertion. The highest frequencies of the introgressed haplotypes occur at X:8560000-8575000 (near *MDox*), where all 10 *D. mauritiana* and 6-7 *D. simulans* lines have the shared haplotype, and position X:8635000 (near *Dox*), where all 10 *D. mauritiana* and 13 *D. simulans* lines have the introgressed haplotype. In both cases, the

peak in introgressed haplotype frequency overlaps a driver gene (**Figure 4**). Finally, we are able to amplify *MDox* cDNA from both *D. mauritiana* and *D. simulans*, indicating that the gene is expressed.

If *MDox* did, in fact, reach fixation in *D. mauritiana* because of meiotic drive, then it must be actively suppressed in *D. mauritiana*, since there is no known manifestation of the *sex-ratio* phenotype in this species. Furthermore, if the population is to survive, there must be a concomitant selective sweep at the suppressor locus. Because we are able to recover *MDox* cDNA, we infer that drive must be suppressed at the post-transcriptional level in *D. mauritiana*, as it is in *D. simulans*. Assuming that *Nmy* (or another locus with homology to *MDox*) acts as the suppressor via a similar RNA interference mechanism, we ask whether any such locus in the genome also shows a strong signal of a selective sweep. Although *Nmy* suppresses *MDox*/*Dox* drive in *D. simulans*, polymorphism and divergence in 5 kb increments of the 100 kb surrounding the *Nmy* gene shows little evidence for either a strong selective sweep or recent introgression (**Table S7**).

All three genes of the Winters *sex-ratio* system share sequence homology due to their origin through gene duplication and retrotransposition (TAO *et al.* 2007a). We attempt to locate any additional suppressor, driver, or modifier genes in the *D. mauritiana* genome by performing Blast searches of the predicted mRNA sequence of *D. mauritiana MDox* to both the *D. mauritiana* chromosome-level assembly and all scaffolds from the *de novo* genome assembly (GARRIGAN *et al.* 2014). There are three additional regions of the X chromosome with sequence homology at coordinates 9.24 Mb, 15.21 Mb and 15.38 Mb, the last of which was identified previously (TAO *et al.* 2007a). The 15.21 Mb region contains one expressed gene (TCONS_00014884) with no known ortholog in *D. melanogaster*. Interestingly, 621 bp

downstream from the region of homology with *MDox* is a Tudor domain that has homology with the gene *krimper*. The remaining X-linked regions have no expressed transcripts detectible in *D. mauritiana* whole-body mRNA samples. Within these three regions, estimates of $G_{\min}$ again show no evidence of introgression (**Table S7**), polymorphism is reduced in *D. simulans* for the 15.21 Mb region (average $\pi_{\text{sim}} = 0.0040$; $P < 0.05$), linkage disequilibrium is significantly elevated in all three regions in both species ($P < 0.05$), and Tajima's *D* is significantly reduced in *D. mauritiana* for the regions at 9.24 Mb and 15.38 Mb ($P < 0.05$). While these results are mixed, there is no definitive evidence for a selective sweep at either *Nmy* or these additional regions of homology at a scale seen at the *MDox*/*Dox* region. However it remains possible that other loci with homology to *MDox* are located in sweep regions, but they simply do not appear in our genome assembly.

DISCUSSION

The current geographical ranges of *Drosophila mauritiana* and *D. simulans* do not overlap and surveys show that *D. simulans* is absent on Mauritius, whereas *D. mauritiana* is one of the most abundant *Drosophila* species on the island (DAVID *et al.* 1989). The absence of an established *D. simulans* population on Mauritius suggests that it likely that opportunities for gene flow between *D. simulans* and *D. mauritiana* are limited and sporadic. Yet despite their largely allopatric history and the evolution of pre- and postzygotic barriers to hybridization, we find evidence for restricted levels of recent gene flow between these two taxa. This gene flow results in recombination between two genomes that differ at approximately 1% of nucleotide positions and are reciprocally monophyletic at more than 40% of autosomal and nearly 60% of X-linked loci. We use relatively conservative criteria to identify 21 distinct genomic regions as putatively

introgressed (totaling 1.38 Mb or 7.92 cM). This estimate corresponds to 1.3% of the total length of aligned sequence on the five major chromosome arms.

Hybridization between populations that have evolved partial postzygotic isolation creates an inherent tension in the genome (*sensu lato*, BARTON AND HEWITT 1985). The likelihood of genomic introgression is elevated by demographic factors, such as an increased frequency of hybrid mating, or by fitness-related factors such as heterosis. Alternatively, introgression is constrained by reduced hybrid fitness due to interactions between incompatible mutations. This tension is mediated by recombination and independent assortment progressively partitioning the genome into genotypes with levels of relative fitness that allow sustained rates of natural backcrossing. Our study documents which blocks of the genome are permitted by natural selection to introgress to intermediate frequency, despite a non-trivial density of loci causing both hybrid inviability and hybrid male sterility.

Because *D. mauritiana* and *D. simulans* are the closest known relatives of *D. melanogaster*, we are able to leverage the vast amount of functional genomic annotations from *D. melanogaster* to ask whether the introgressed regions bear some distinguishing characteristics. We find that genomic regions with higher rates of crossing-over are more likely to introgress between species. This is not necessarily unexpected because these regions are more frequently decoupled from incompatible interactions in the early generations of backcrossing. Two fundamental parameters of the BDM model are the number of substitutions and the probability that any two substitutions are incompatible (TURELLI AND ORR 2000). In support of the predictions of the BDM model, we find that proteins with reduced rates of amino acid substitution are over-represented in the introgressing genomic regions. Interestingly, we also find that genes in the introgression regions have higher expression levels, which is predicted based on the observation that gene expression

level inversely correlates with the rate of evolution in a variety of taxa (DRUMMOND *et al.* 2006; KOONIN AND WOLF 2006). In addition, we observe that genes in introgression regions have higher tissue-specificity, which is associated with lower rates of divergence in gene expression (LIAO *et al.* 2006) and thus, fewer regulatory incompatibilities.

However, we find other patterns that are not readily predicted by the BDM model. Under the BDM model, loss of hybrid fitness is caused by dysfunctional molecular interactions between substitutions that fixed independently and have never been tested together by natural selection. Thus, genes with larger numbers of interactions should have a higher probability of evolving an incompatibility than lowly interacting genes. However, we find there is no significant difference in the number of protein-protein interactions for genes in the putatively introgressed regions, compared to the genome average. One caveat to this is that larger genetic networks may also incorporate more redundancy and may be more robust to perturbation due to mutation (WAGNER AND WRIGHT 2007).

It is often thought that genes experience lineage-specific natural selection have an increased probability of causing postzygotic reproductive isolation (*e.g.*, PRESGRAVES 2010). Since we observe that the genes located in the putatively introgressed genomic regions are not disproportionately subject to positive natural selection and these genes also have a significantly slower rate of protein evolution, our results cannot fully address this hypothesis. However, our analyses does highlight a situation in which an introgressing gene is able to maintain a molecular interaction with a gene experiencing lineage-specific positive selection. We find that the *Chd1* gene has recently introgressed from *D. simulans* into *D. mauritiana* and its protein product has direct physical interactions with that of the *Hira* gene, which has experienced a recent selective sweep in *D. mauritiana* (GARRIGAN *et al.* 2014). A haplotype carrying the introgressed *Chd1*

gene occurs at intermediate frequency in *D. mauritiana* (**Figure S2**) and it differs by two fixed nonsynonymous substitutions from the endogenous *D. mauritiana* sequences. Generally, the *Chd1* gene is known to evolve rapidly in *Drosophila* (LEVINE AND BEGUN 2008) and its protein product has been shown to deposit maternally-derived histones onto the condensed male pronucleus prior to the first mitosis following fertilization (KONEV *et al.* 2007). To perform this function, the Chd1 protein interacts with the protein encoded by the *Hira* gene. Hira chaperones maternal H3.3 histones to Chd1 for deposition and there is a direct physical interaction between the two proteins. A previous analysis of this data set shows that the *Hira* gene has undergone a strong recent selective sweep in *D. mauritiana* (GARRIGAN *et al.* 2014), possibly to escape cytoplasmic incompatibility caused by *Wolbachia* (ZHENG *et al.* 2011). This example demonstrates that, despite rapid evolution in the components of a non-redundant interaction critical to reproduction, this system remains robust to perturbation due to interspecific divergence.

Unlike introgression events documented between *D. simulans* and *D. sechellia* (BRAND *et al.* 2013), we find no clear evidence for adaptive introgression between *D. simulans* and *D. mauritiana*. However, we do find that the two adjacent introgressing regions on the X chromosome both occur at high frequency in both *D. simulans* and *D. mauritiana*. These likely do not represent a case of a globally advantageous mutation, but instead the history of a recently active selfish genetic element. The introgressing regions on the X chromosome contains the two driver genes of the functionally-characterized Winters *sex-ratio* meiotic drive system and our analysis uncovers compelling circumstantial evidence that this drive system has recently been active in both *D. simulans* and *D. mauritiana* (**Figure 4**). The idea that meiotic drive lies at the heart of an evolutionary arms race— one that may ultimately lead to hybrid male sterility

(FRANK 1991; HURST AND POMIANKOWSKI 1991)— was initially viewed with skepticism (COYNE AND ORR 1993), but has more recently found qualified instances of empirical support (TAO *et al.* 2001; PHADNIS AND ORR 2009). In fact, the sweep at the *MDox*/*Dox* region in *D. mauritiana* has been cited as supporting the idea that intra-genomic conflict can contribute to speciation (NOLTE *et al.* 2013; SEEHAUSEN *et al.* 2014). Although it is theoretically difficult to imagine how a locus, that simultaneously causes both hybrid male sterility and sex chromosome segregation distortion, could readily invade and sweep through a divergent population, our data represent one observation that stands contrary to this idea of meiotic drive as a mechanism that facilitates reproductive isolation between allopatric populations. If it is generally true that, when given the opportunity, meiotic drive elements can readily traverse species boundaries and become fixed in the recipient population (MORITA *et al.* 1992), then this scenario suggests something quite the opposite— a questionable role for meiotic drive in directly promoting reproductive isolation.

The restricted pattern of genomic introgression between *D. simulans* and *D. mauritiana* suggests these two species' genomes have a high density of substitutions that cause postzygotic isolation or that there has been a limited opportunity for hybridization due to geographical considerations or substantial prezygotic isolation. While we are essentially documenting some fraction of the genome that does not contain hybrid incompatibilities, this data set has the potential to address the role of natural selection in driving the evolution of intrinsic postzygotic isolating factors. Ongoing work to map loci that cause hybrid male sterility when experimentally introgressed from *D. simulans* into a *D. mauritiana* background will allow a detailed analysis of the genomic consequences of the inherent tension between gene flow and reproductive isolation.

ACKNOWLEDGMENTS

We would like to thank H. Allen Orr and Daven Presgraves for insightful comments on previous drafts of the manuscript. We are grateful to Peter Andolfatto and Kevin Thornton for sharing their *D. simulans* data. This work was supported by National Institutes of Health (grant number R01 ODO1054801).

LITERATURE CITED

Ballard, J. W., 2000 When one is not enough: introgression of mitochondrial DNA in *Drosophila*. Mol Biol Evol 17**:** 1126-1130.

Barton, N. H., and G. M. Hewitt, 1985 Analysis of hybrid zones. Ann Rev Ecol Syst 16**:** 113-148.

Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y. P. Poh *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5**:** e310.

Brand, C. L., S. B. Kingan, L. Wu and D. Garrigan, 2013 A selective sweep across species boundaries in *Drosophila*. Mol Biol Evol 30**:** 2177-2186.

Chintapalli, V. R., J. Wang and J. A. Dow, 2007 Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet 39**:** 715-720.

Coyne, J. A., 1984 Genetic basis of male sterility in hybrids between two closely related species of *Drosophila*. Proc Natl Acad Sci USA 81**:** 4444-4447.

Coyne, J. A., and H. A. Orr, 1993 Further evidence against meiotic-drive models of hybrid sterility. Evolution 47**:** 685-687.

Coyne, J. A., and H. A. Orr, 2004 *Speciation*. Sinauer Associates, Sunderland, MA.

David, J. R., S. F. McEvey, M. Solignac and L. Tsacas, 1989 Drosophila communities on Mauritius and the ecological niche of *Drosophila mauritiana* (Diptera, Drosophilidae). Rev Zool Afr 103**:** 107-116.

Dobzhansky, T., 1937 *Genetics and the Origin of Species*. Columbia University Press, New York, NY.

Dobzhansky, T., 1951 *Genetics and the Origin of Species*. Columbia University Press, New York, NY.

Drummond, D. A., A. Raval and C. O. Wilke, 2006 A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23**:** 327-337.

Ellegren, H., and J. Parsch, 2007 The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet 8**:** 689-698.

Frank, S. A., 1991 Divergence of meiotic drive-suppression systems as an explanation for sex-biased hybrid sterility and inviability. Evolution 45**:** 262-267.

Garrigan, D., 2013 POPBAM: tools for evolutionary analysis of short read sequence alignments. Evol Bioinform 9**:** 343-353.

Garrigan, D., S. B. Kingan, A. J. Geneva, P. Andolfatto, A. G. Clark *et al.*, 2012 Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. Genome Res 22**:** 1499-1511.

Garrigan, D., S. B. Kingan, A. J. Geneva, J. P. Vedanayagam and D. C. Presgraves, 2014 Genome diversity and divergence in *Drosophila mauritiana*: multiple signatures of faster X evolution. Genome Biol Evol 6**:** 2444-2458.

Geneva, A. J., C. A. Muirhead, S. B. Kingan and D. Garrigan, 2015 A new method to scan genomes for introgression in a secondary contact model. PLoS One 10**:** e0118621.

Gomes, S., and A. Civetta, 2015 Hybrid male sterility and genome-wide misexpression of male reproductive proteases. Sci Rep 5.

Guruharsha, K. G., J.-F. Rual, B. Zhai, J. Mintseris, P. Vaidya *et al.*, 2011 A protein complex network of *Drosophila melanogaster*. Cell 147**:** 690-703.

Halligan, D. L., A. Eyre-Walker, P. Andolfatto and P. D. Keightley, 2004 Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. Genome Res 14**:** 273-279.

Harrison, R. G., 1990 Hybrid zones: windows on evolutionary process, pp. 69-128 in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyma and J. Antonovics. Oxford University Press, New York, NY.

Hurst, L. D., and A. Pomiankowski, 1991 Causes of sex ratio bias may account for unisexual sterility in hybrids: a new explanation of Haldane's rule and related phenomena. Genetics 128**:** 841-858.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics 146**:** 1197-1206.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley *et al.*, 2013 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14**:** R36.

Kingan, S. B., D. Garrigan and D. L. Hartl, 2010 Recurrent selection on the Winters *sex-ratio* genes in *Drosophila simulans*. Genetics 184**:** 253-265.

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman *et al.*, 2000 The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics 156**:** 1913-1931.

Konev, A. Y., M. Tribus, S. Y. Park, V. Podhraski, C. Y. Lim *et al.*, 2007 CHD1 motor protein is required for deposition of histone variant H3.3 into chromatin in vivo. Science 317**:** 1087-1090.

Koonin, E. V., and Y. I. Wolf, 2006 Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol 17**:** 481-487.

Lachaise, D., J. R. David, F. Lemeunier, L. Tsacas and M. Ashburner, 1986 The reproductive relationships of *Drosophila sechellia* with *Drosophila mauritiana, Drosophila simulans* and *Drosophila melanogaster* from the Afrotropical region. Evolution 40**:** 262-271.

Landry, C. R., D. L. Hartl and J. M. Ranz, 2007 Genome clashes in hybrids: insights from gene expression. Heredity 99**:** 483-493.

Larracuente, A. M., T. B. Sackton, A. J. Greenberg, A. Wong, N. D. Singh *et al.*, 2008 Evolution of protein-coding genes in *Drosophila*. Trends Genet 24**:** 114-123.

Levine, M. T., and D. J. Begun, 2008 Evidence of spatially varying selection acting on four chromatin-remodeling loci in *Drosophila melanogaster*. Genetics 179**:** 475-485.

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

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25**:** 2078-2079.

Liao, B.-Y., N. M. Scott and J. Zhang, 2006 Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol 23**:** 2072-2080.

Mayr, E., 1942 *Systematics and the Origin of Species*. Columbia University Press, New York.

Mayr, E., 1963 *Animal Species and Evolution*. Harvard University Press, Cambridge, MA.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351**:** 652-654.

Meiklejohn, C. D., J. Parsch, J. M. Ranz and D. L. Hartl, 2003 Rapid evolution of male-biased gene expression in Drosophila. Proc Natl Acad Sci USA 100**:** 9894-9899.

Meiklejohn, C. D., and D. C. Presgraves, 2012 Little evidence for demasculinization of the *Drosophila* X chromosome among genes expressed in the male germline. Genome Biol Evol 4**:** 1007-1016.

Mi, H., A. Muruganujan and P. D. Thomas, 2013 PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41**:** D377-D386.

Mitri, C., L. Soustelle, B. Framery, J. Bockaert, M. L. Parmentier *et al.*, 2009 Plant insecticide L-canavanine repels *Drosophila* via the insect orphan GPCR DmX. PLoS Biol 7**:** e1000147.

Moehring, A., J. Li, M. Schug, S. Smith, M. deAngelis *et al.*, 2004 Quantitative trait loci for sexual isolation between *Drosophila simulans* and *D. mauritiana*. Genetics 167**:** 1265-1274.

Morita, T., H. Kubota, K. Murata, M. Nozaki, C. Delarbre *et al.*, 1992 Evolution of the mouse *t* haplotype: recent and worldwide introgression to *Mus musculus*. Proc Natl Acad Sci USA 89**:** 6851-6855.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nolte, V., R. V. Pandey, R. Kofler and C. Schlotterer, 2013 Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. Genome Res 23**:** 99-110.

Orr, H. A., 1997 Haldane's rule. Ann Rev Ecol Syst 28**:** 195-218.

Parsch, J., S. Novozhilov, S. S. Saminadin-Peter, K. M. Wong and P. Andolfatto, 2010 On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. Mol Biol Evol 27**:** 1226-1234.

Phadnis, N., and H. A. Orr, 2009 A single gene causes both male sterility and segregation distortion in *Drosophila* hybrids. Science 323**:** 376-379.

Presgraves, D. C., 2010 The molecular evolutionary basis of species formation. Nat Rev Genet 11**:** 175-180.

R Core Team, 2014 R: A Language and Environment for Statistical Computing, pp. R Foundation for Statistical Computing, Vienna, Austria.

Rogers, R. L., J. M. Cridland, L. Shao, T. T. Hu, P. Andolfatto *et al.*, 2014 Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. Mol Biol Evol 31**:** 1750-1766.

Roux, C., G. Tsagkogeorga, N. Bierne and N. Galtier, 2013 Crossing the species barrier: genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species. Mol Biol Evol 30**:** 1574-1587.

Seehausen, O., R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman *et al.*, 2014 Genomics and the origin of species. Nat Rev Genet 15**:** 176-192.

Stamatakis, A., 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30**:** 1312-1313.

Stoletzki, N., and A. Eyre-Walker, 2011 Estimation of the neutrality index. Mol Biol Evol 28**:** 63-70.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123**:** 585-595.

Tao, Y., L. Araripe, S. B. Kingan, Y. Ke, H. Xiao *et al.*, 2007a A *sex-ratio* meiotic drive system in *Drosophila simulans*. II: An X-linked distorter. PLoS Biol 5**:** 2576-2588.

Tao, Y., S. Chen, D. L. Hartl and C. C. Laurie, 2003a Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana*. I. Differential accumulation of hybrid male sterility effects on the X and autosomes. Genetics 164**:** 1383-1397.

Tao, Y., and D. L. Hartl, 2003 Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana*. III. Heterogeneous accumulation of hybrid incompatibilities, degree of dominance, and implications for Haldane's rule. Evolution 57**:** 2580-2598.

Tao, Y., D. L. Hartl and C. C. Laurie, 2001 Sex-ratio segregation distortion associated with reproductive isolation in *Drosophila*. Proc Natl Acad Sci USA 98**:** 13183-13188.

Tao, Y., J. P. Masly, L. Araripe, Y. Ke and D. L. Hartl, 2007b A *sex-ratio* meiotic drive system in *Drosophila simulans*. I: an autosomal suppressor. PLoS Biol 5**:** 2560-2575.

Tao, Y., Z.-B. Zeng, J. Li, D. L. Hartl and C. C. Laurie, 2003b Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana*. II. Mapping hybrid male sterility loci on the third chromosome. Genetics 164**:** 1399-1418.

Teeter, K. C., B. A. Payseur, L. W. Harris, M. A. Bakewell, L. M. Thibodeau *et al.*, 2008 Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res 18**:** 67-76.

Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn *et al.*, 2013 Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31**:** 46-53.

True, J. R., J. M. Mercer and C. C. Laurie, 1996a Differences in crossover frequency and distribution among three sibling species of *Drosophila*. Genetics 142**:** 507-523.

True, J. R., B. S. Weir and C. C. Laurie, 1996b A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*. Genetics 142**:** 819-837.

Turelli, M., and H. A. Orr, 2000 Dominance, epistasis and the genetics of postzygotic isolation. Genetics 154**:** 1663-1679.

Wagner, A., and J. Wright, 2007 Alternative routes and mutational robustness in complex regulatory networks. Biosystems 88**:** 163-172.

Wright, S., 1951 The genetical structure of populations. Ann Eugen 15**:** 323-354.

Wright, S., 1978 *Evolution and the Genetics of Populations, Vol. 4: Variability Within and Among Natural Populations.* University of Chicago Press, Chicago, IL.

Wu, C.-I., 2001 The genic view of the process of speciation. J Evol Biol 14**:** 851-865.

Yanai, I., H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar *et al.*, 2005 Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21**:** 650-659.

Yuan, Y., Y. Xu, J. Xu, R. L. Ball and H. Liang, 2012 Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data. Bioinformatics 28**:** 1246-1252.

Zhang, J., and J.-R. Yang, 2015 Determinants of the rate of protein sequence evolution. Nat Rev Genet 16**:** 409-420.

Zhang, Y. E., M. D. Vibranovski, B. H. Krinsky and M. Long, 2010 Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. Genome Res 20**:** 1526-1533.

Zheng, Y., P.-P. Ren, J.-L. Wang and Y.-F. Wang, 2011 *Wolbachia*-induced cytoplasmic incompatibility is associated with decreased *Hira* expression in male *Drosophila*. PLoS One 6**:** e19512.

Table 1. Comparison of functional characteristics between regions of the genome experiencing recent introgression and the genomic background. Functional characteristics are categorized as "gene-based" meaning they are properties of the protein coding genes located in those regions, and "region-based", which means they are properties of the entire introgressing genomic region, irrespective of coding sequence properties. Statistically significant comparisons are shown in boldface type.

Type	Characteristic	Background	Introgression
Gene-based	**Gene ontology**	–	**Enriched for cell development process**
	Mean number of protein-protein interactions	19.95	17.61
	Median gene expression (FPKM)	**14.68**	**20.46**
	Proportion of male-biased genes	0.39	0.26
	Median tissue specificity (τ)	**0.413**	**0.580**
	Proportion of genes in oldest branch	0.925	0.942
	Direction of selection	**0.0839**	**0.0129**
	Proportion positively selected	0.0167	0
Region-based	Proportion coding sequence	0.1473	0.1475
	Median net divergence from *D. melanogaster*	0.0338	0.0351
	Median population crossing-over rate	**70.0223**	**107.301**

Table 2. Major genomic regions showing evidence of recent introgression.

chrom	begin	end	length (kb)	length (cM)	Percent CDS	genes
2L	1810000	1830000	20	0.0826	21.42	1
2L	2790000	2860000	70	0.2889	32.36	12
2L	9040000	9050000	10	0.0643	44.97	4
2L	15150000	15260000	110	0.0924	8.73	7
2R	3760000	3790000	30	0.1860	23.48	3
2R	13700000	13710000	10	0.0525	31.25	8
3L	20760000	20910000	150	0.2702	11.50	13
3L	21220000	21450000	230	1.7186	3.74	12
3R	3350000	3360000	10	0.0239	86.01	1
3R	4570000	4680000	110	0.1083	14.47	17
3R	5020000	5070000	50	0.0492	5.05	2
3R	13620000	13630000	10	0.1472	19.03	2
3R	13870000	14070000	200	2.9441	11.11	18
3R	21470000	21480000	10	0.0504	0	0
3R	21530000	21670000	140	0.7061	11.15	12
3R	23350000	23420000	70	0.3530	29.54	4
3R	23730000	23740000	10	0.0504	39.89	5
3R	23980000	24020000	40	0.2017	41.32	13
3R	24100000	24120000	20	0.1009	9.14	1
X	8560000	8580000	20	0.1082	49.89	4
X	8630000	8690000	60	0.3245	23.73	4

FIGURE LEGENDS

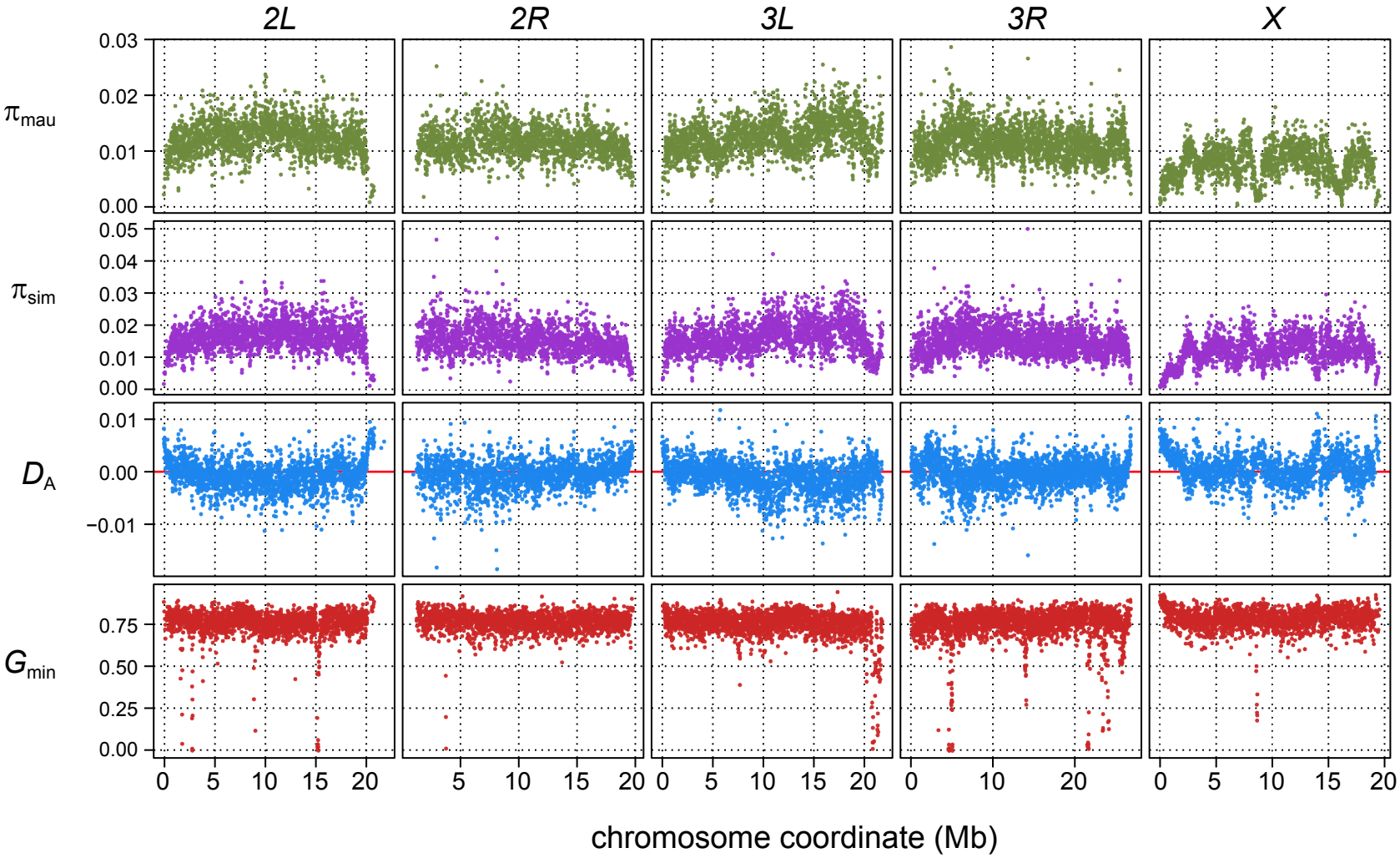
Figure 1. Genome scans for polymorphism and divergence in 10~kb windows. The top row of panels (green dots) shows nucleotide diversity in a sample of 10 inbred strains of *D. mauritiana* (π_{mau}) across each of the five major chromosome arms. The second row of panels (purple dots) shows nucleotide diversity in a sample of 20 inbred strains of *D. simulans* (π_{sim}). The third row (blue dots) shows net nucleotide divergence (D_A) between the *D. mauritiana* and *D. simulans* samples. The bottom row (red dots) plots G_{min}, which is the ratio of the minimum number of nucleotide differences between *D. mauritiana* and *D. simulans* to the average number of differences, a measure that is sensitive to introgression.

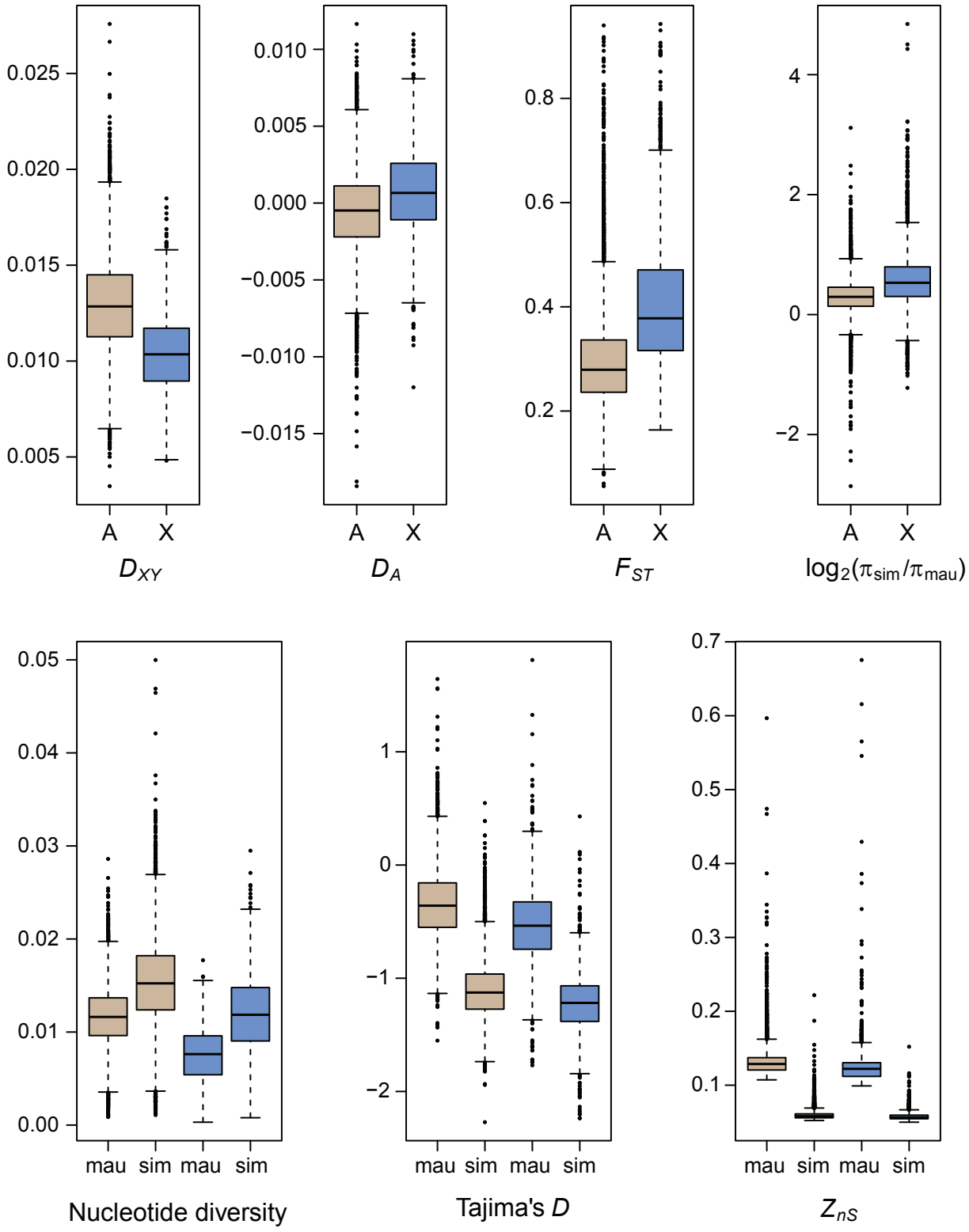
Figure 2. Box plots of polymorphism and divergence statistics. In all plots, the autosomal distributions are shown in brown and the X chromosome distributions are shown in blue. The top row of four plots shows the distributions of between-population measures: the average number of nucleotide differences per site between populations (D_{XY}), net divergence (D_A), the fixation index (F_{ST}), and the logarithm of the ratio of nucleotide diversity in *D. simulans* to that in *D. mauritiana*. The bottom row of plots contrast within-species measures of polymorphism (π), the allele frequency spectrum (Tajima's *D*), and linkage disequilibrium (Z_{nS}).

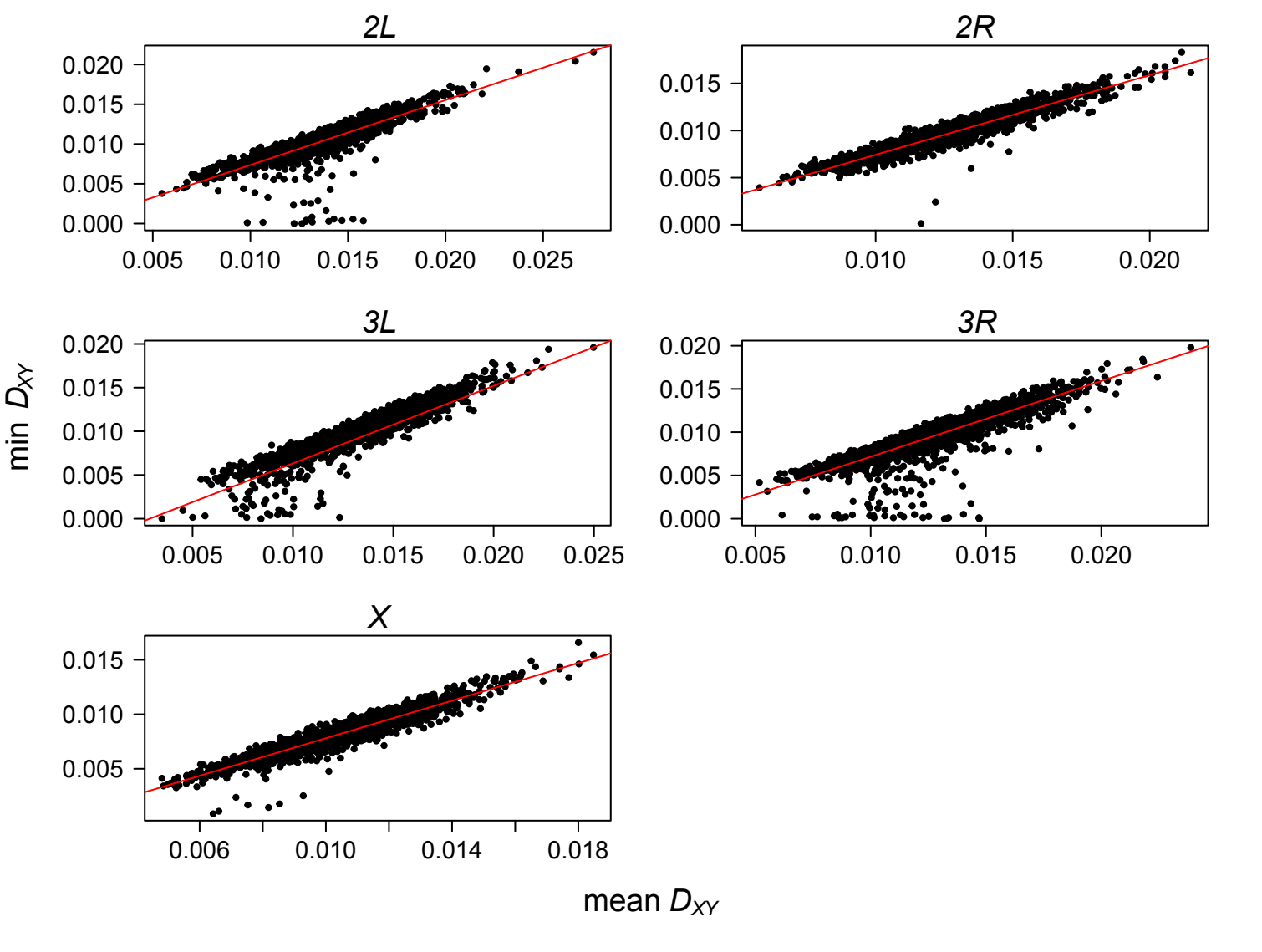
Figure 3. Plots of minimum D_{XY} *versus* mean D_{XY} for non-overlapping 10 kb windows on each chromosome arm, illustrating that windows with reduced G_{min} do not exclusively occur in regions of low divergence. Red lines plot the linear regression.

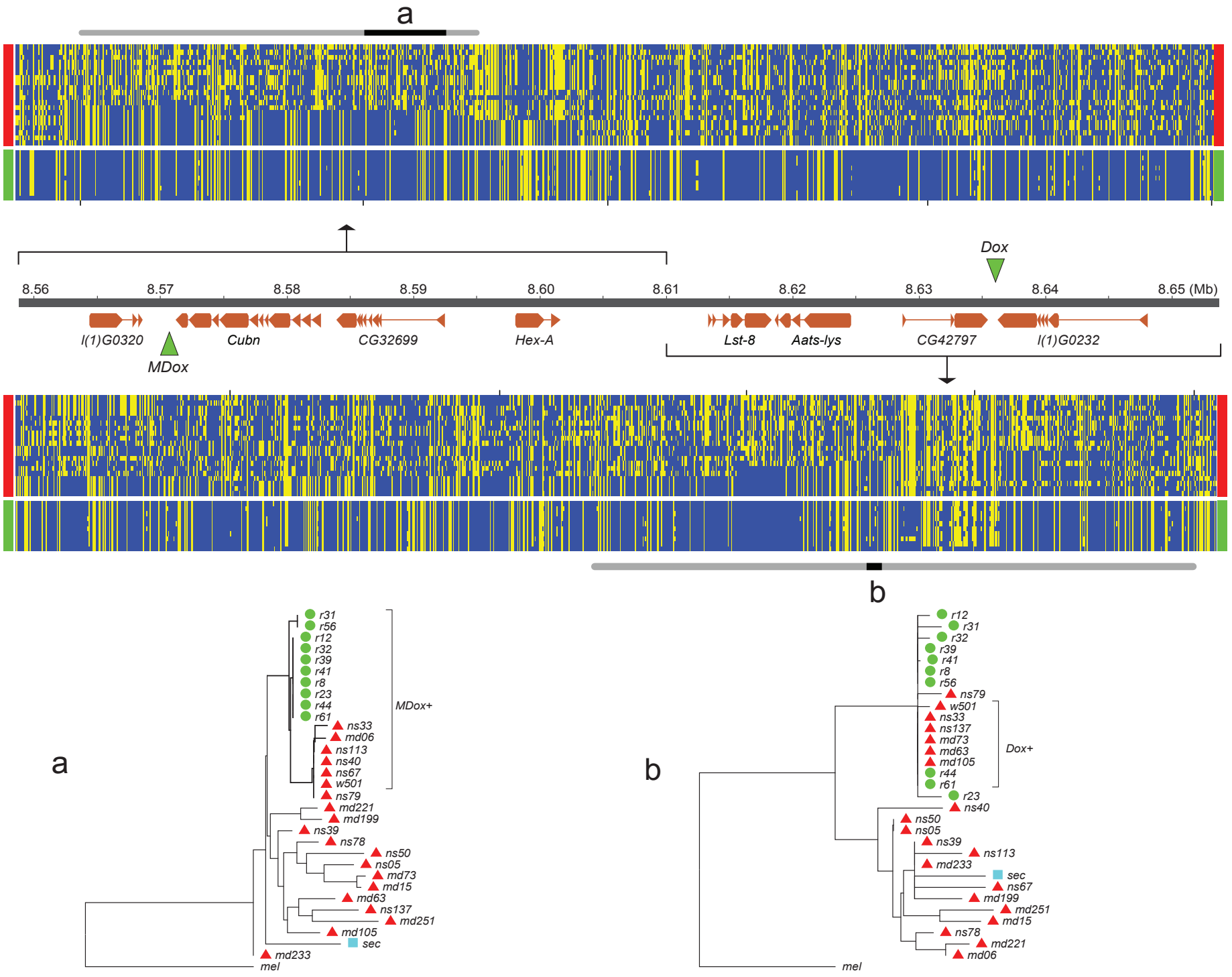
Figure 4. Polymorphism and divergence in the *MDox*/*Dox* region of chromosome *X*. The top half of the figure shows two polymorphism tables: the top table represents the *MDox* region and the bottom for the *Dox* region. Within the tables, yellow squares denote the derived allele and blue squares indicate the ancestral allele. The top 20 rows of each table (labeled with a red bar)

correspond to the *D. simulans* samples, while the bottom 10 rows (labeled with a green bar) correspond to the *D. mauritiana* samples. The map between the polymorphism tables shows the gene models for this region (orange boxes) and the locations of the *Dox* and *MDox* genes (green triangles). The grey bars on the top and bottom of the polymorphism tables mark the sites that occur within an introgressed region, with black sections labeled “a” and “b”, which mark the location of the *MDox* and *Dox* genes, respectively. Below the polymorphism tables are two maximum likelihood phylogenetic trees labeled “a” and “b”, again corresponding to the *MDox* and *Dox* regions.


2L
2R
3L
3R
X
π_{mau}
π_{sim}
D_A
G_{min}
0.03
0.02
0.01
0.00
0.05
0.04
0.03
0.02
0.01
0.00
0.01
0.00
−0.01
0.75
0.50
0.25
0.00
0 5 10 15 20
5 10 15 20
0 5 10 15 20
0 10 20
0 5 10 15 20
chromosome coordinate (Mb)




2L
2R
3L
3R
X
min D_{XY}
mean D_{XY}


a
8.56
8.57
8.58
8.59
8.60
8.61
8.62
8.63
8.64
8.65 (Mb)
Dox
l(1)G0320
MDox
Cubn
CG32699
Hex-A
Lst-8
Aats-lys
CG42797
l(1)G0232
b
a
r31
r56
r12
r32
r39
r41
r8
r23
r44
r61
MDox+
ns33
md06
ns113
ns40
ns67
w501
ns79
md221
md199
ns39
ns78
ns50
ns05
md73
md15
md63
ns137
md251
md105
sec
md233
mel
b
r12
r31
r32
r39
r41
r8
r56
ns79
w501
ns33
ns137
md73
md63
md105
r44
r61
Dox+
r23
ns40
ns50
ns05
ns39
ns113
md233
sec
ns67
md199
md251
md15
ns78
md221
md06
mel