

# Evaluation of the Genomic Extent of Effects of Fixed Inversion Differences on Intraspecific Variation and Interspecific Gene Flow in *Drosophila pseudoobscura* and *D. persimilis*

Carlos A. Machado,<sup>\*,1</sup> Tamara S. Haselkorn\* and Mohamed A. F. Noor<sup>†</sup>

<sup>\*</sup>Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721 and <sup>†</sup>DCMB Group/Biology Department, Duke University, Durham, North Carolina 27708

Manuscript received August 15, 2006  
Accepted for publication December 4, 2006

## ABSTRACT

There is increasing evidence that chromosomal inversions may facilitate the formation or persistence of new species by allowing genetic factors conferring species-specific adaptations or reproductive isolation to be inherited together and by reducing or eliminating introgression. However, the genomic domain of influence of the inverted regions on introgression has not been carefully studied. Here, we present a detailed study on the consequences that distance from inversion breakpoints has had on the inferred level of gene flow and divergence between *Drosophila pseudoobscura* and *D. persimilis*. We identified the locations of the inversion breakpoints distinguishing *D. pseudoobscura* and *D. persimilis* in chromosomes 2, XR, and XL. Population genetic data were collected at specific distances from the inversion breakpoints of the second chromosome and at two loci inside the XR and XL inverted regions. For loci outside the inverted regions, we found that distance from the nearest inversion breakpoint had a significant effect on several measures of divergence and gene flow between *D. pseudoobscura* and *D. persimilis*. The data fitted a logarithmic relationship, showing that the suppression of crossovers in inversion heterozygotes also extends to loci located outside the inversion but close to it (within 1–2 Mb). Further, we detected a significant reduction in nucleotide variation inside the inverted second chromosome region of *D. persimilis* and near one breakpoint, consistent with a scenario in which this inversion arose and was fixed in this species by natural selection.

CHROMOSOMAL rearrangements, such as inversions or translocations, may allow factors conferring adaptation or reproductive isolation to be genetically linked to each other even when they are not physically proximate along chromosome arms. Crossover products fail to be recovered from inversion heterozygotes, allowing these segments to be inherited as blocks. If inverted regions disproportionately bear alleles under divergent selection or conferring reproductive isolation, one can predict that many collinear regions of the genome will introgress between hybridizing species more readily than rearranged regions because of weak or no linkage to alleles conferring adaptation, mating discrimination, or hybrid dysfunction. Rearranged regions will more often be completely linked to such alleles, and introgression between species in such regions will be limited (ORTIZ-BARRIENTOS *et al.* 2002; BUTLIN 2005). The same effect should be observed in other regions of reduced recombination.

These predictions have been confirmed in numerous empirical studies. For example, rates of gene flow are higher between collinear than between rearranged chromosomes in sunflower hybrid zones (RIESEBERG *et al.* 1999), between races of *Rhagoletis* fruit flies (FEDER *et al.* 2003, 2005), and between two *Drosophila* species (*Drosophila pseudoobscura* and *D. persimilis*) (MACHADO *et al.* 2002; MACHADO and HEY 2003; HEY and NIELSEN 2004). Analogous results have also been observed in the low-recombination pericentromeric regions of *Anopheles* mosquitoes (STUMP *et al.* 2005a,b; TURNER *et al.* 2005) and house mice (PANITHANARAK *et al.* 2004).

Although evidence is mounting that chromosomal rearrangements in general, and inversions in particular, may facilitate the formation or persistence of new species, the extent of these effects inside and outside the inverted regions has not been studied in detail (but see NAVARRO *et al.* 1997). First, some loci outside of inversions may bear alternate alleles that are favored in the two divergent populations, and therefore gene flow will be reduced or eliminated at (and very near) those loci. These loci would introduce “noise” in the overall pattern in that positions outside inversions may occasionally bear striking differences between species. Second, the suppression of crossovers in inversion heterozygotes

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. EF041145–EF041478.

<sup>1</sup>Corresponding author: 1041 E. Lowell St., BSW 310, Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721. E-mail: cmachado@email.arizona.edu

extends some distance outside the inverted region [previously shown in our target species (ORTÍZ-BARRIENTOS *et al.* 2006)]. Third, some gene flow can occur between inverted regions through double crossovers or gene conversion, and these processes may be more frequent farther from the inversion breakpoints (NAVARRO *et al.* 1997; LAAYOUNI *et al.* 2003; SCHAEFFER and ANDERSON 2005). Finally, crossover rates in inversion heterozygotes are sometimes elevated in collinear parts of the genome, termed the “interchromosomal effect” (SCHULTZ and REDFIELD 1951).

Here we address these questions in the *D. pseudoobscura* species group, which perhaps constitutes the most extensively studied system with respect to the association of inversions with reproductive isolation and introgression. The two North American species *D. pseudoobscura* and *D. persimilis* differ by fixed chromosomal inversions in the second chromosome (Muller’s element E) and the left arm of the X chromosome (XL) (Muller’s element A). Further, an inversion in the right arm of the X chromosome (XR) (Muller’s element D) is fixed among *D. pseudoobscura* and non-sex-ratio (SR) XR *D. persimilis* strains. Finally, the third chromosome (Muller’s element C) is highly polymorphic for inversions in both species, and one abundant arrangement (“standard”) is shared. *D. pseudoobscura* is distributed across the western half of North America, with large amounts of gene flow among populations (JONES *et al.* 1981; SCHAEFFER and MILLER 1992b). *D. persimilis* occurs in sympatry with *D. pseudoobscura* in the western Pacific coast states. Divergence between these species is estimated to have begun 0.55 million years ago (WANG *et al.* 1997), and all hybrid males between these species are sterile. Strong species mate discrimination by females is observed in the laboratory (MERRELL 1954; NOOR 1996), and there is evidence for the evolution of reinforcement of isolating mechanisms (NOOR 1995).

Previous molecular evolutionary analyses of 14 loci revealed a striking pattern of extensive gene flow between the North American species *D. pseudoobscura* and *D. persimilis* in collinear regions of the genome but limited gene flow in regions within or near fixed inverted regions (MACHADO *et al.* 2002; MACHADO and HEY 2003; HEY and NIELSEN 2004). Concurrent genetic studies of sexual isolation, hybrid sterility, and hybrid inviability in these species found that these traits were primarily or exclusively mediated by genes within the inverted regions (NOOR *et al.* 2001a,b). Interestingly, when one compares the allopatric South American subspecies *D. pseudoobscura bogotana* to *D. persimilis*, both the molecular and the phenotypic signatures of interspecies gene flow are reduced or absent (MACHADO *et al.* 2002; MACHADO and HEY 2003; BROWN *et al.* 2004).

The genome sequence of *D. pseudoobscura* has been published recently (RICHARDS *et al.* 2005) and an assembly of the *D. persimilis* genome has also been posted online (GILBERT 2005). These resources allow us the

opportunity to (1) precisely identify the locations of the inversion breakpoints distinguishing *D. pseudoobscura* and *D. persimilis*, (2) obtain polymorphism and divergence data at specific, known, distances from the inversion breakpoints, and (3) compare levels of intraspecific variation and levels of gene flow between these species at varying distances from the inversion breakpoints, both inside and outside the inverted regions. These are the aims of our study. We focus on new and published population genetic data from 18 loci across the second chromosome because of its more complete assembly relative to the other chromosomes bearing fixed inversion differences between these two species. We also briefly examine a third chromosome locus and loci just inside two other inversions that distinguish these species (in XL and XR) for polymorphism and divergence comparisons.

## MATERIALS AND METHODS

**Drosophila stocks and DNA extractions:** We used inbred lines established for a previous study (MACHADO *et al.* 2002). Flies were collected from locations in the western United States (*D. pseudoobscura*, *D. persimilis*, and *D. miranda*) and in the vicinity of Bogotá, Colombia (*D. pseudoobscura bogotana*). The sequences from *D. miranda* were primarily used to root the variation found among the other species. The isofemale lines went through 12–17 generations of full-sib mating prior to DNA sequencing. Genomic DNA from these inbred lines was extracted from adult male flies using the single-fly squish protocol (GLOOR and ENGELS 1992). Sex-ratio *D. persimilis* sequences (obtained only for locus *X\_2102* on the XR chromosome arm) were derived from F<sub>1</sub> or F<sub>2</sub> male progeny of flies collected in Mather and Mount St. Helena, California, in 1997. These males were paired with females individually in food-containing vials and then frozen, and the offspring sex ratio was scored. Those males producing all-female offspring were kept and designated as SR. We also confirmed the SR assignment by the presence of one *Esterase-5* SNP previously shown to be associated with the SR arrangement in *D. persimilis* (BABCOCK and ANDERSON 1996). DNA was extracted from these males using the Puregene DNA Isolation kit (Gentra Systems, Research Triangle Park, NC).

**Identification of all fixed chromosomal inversion breakpoints:** The second chromosome inversion breakpoints distinguishing *D. pseudoobscura* and *D. persimilis* were previously localized by ORTÍZ-BARRIENTOS *et al.* (2006) to 200,000–400,000 base regions using the linkage maps of these species. We used a “leapfrogging” approach between the published *D. pseudoobscura* genome sequence (RICHARDS *et al.* 2005) and the *D. persimilis* genome sequence traces available in the NCBI trace archive to more precisely identify the breakpoints within these intervals, as well as to localize those of the XL and XR chromosomes. The *D. persimilis* genome sequence traces are archived as “matepairs,” which are equivalent to paired “forward” and “reverse” sequencing reactions off a plasmid insert of a known size (4000, 10,000, or 40,000 bp).

We took 500-base segments from *D. pseudoobscura* in regions known to have inversion breakpoints and BLASTed them to the *D. persimilis* trace archive (ALTSCHUL *et al.* 1990). From the best matches, we obtained the matepair sequences and the insert sizes. We then BLASTed the matepair sequences back to *D. pseudoobscura* and identified their locations. From this

**TABLE 1**  
**Approximate location of the sequenced loci in the assembled second chromosome of *D. pseudoobscura***  
**(genome release 2.0) (RICHARDS *et al.* 2005)**

Locus	Location (in Mb)	Location relative to inversion <sup>b</sup>	Gene ID no.	<i>D. melanogaster</i> homolog
2001 <sup>a</sup>	4.80	Telomere	—	—
2M05	5.75	Telomere	GA17984	CG4141, <i>Pi3K92E</i>
2M06	6.50	Telomere	—	—
2M07	7.24	Telomere	GA17887	CG4029, <i>jumu</i>
2002 <sup>a</sup>	8.06	Telomere	—	—
<i>BrkPt</i> <sup>c</sup>	9.42	Breakpoint	—	—
2M12	12.0	Inside	GA16110	CG31235
2M13	13.12	Inside	GA15148	CG1951
2M14	14.25	Inside	GA16040	CG31140
2M15	15.50	Inside	—	—
2M16	16.0	Inside	—	—
2M17	17.0	Inside	—	—
2M17.3	17.3	Centromere	GA15244	CG2104
2M18	18.50	Centromere	—	—
<i>bcd</i> <sup>d</sup>	19.73	Centromere	GA10255	CG1034, <i>bcd</i>
2M20	20.83	Centromere	GA19991	CG6966
<i>rh1</i> <sup>a</sup>	26.20	Centromere	GA18249	CG4550, <i>ninaE</i> <sup>d</sup>
2003 <sup>a</sup>	29.20	Centromere	—	—

<sup>a</sup> Data are from MACHADO *et al.* (2002).

<sup>b</sup> Markers outside inversion are located toward the telomeric or centromeric side of inversion.

<sup>c</sup> Distal breakpoint sequence of the fixed inversion separating *D. pseudoobscura* and *D. persimilis*.

<sup>d</sup> *ninaE* is a synonym of *rh1* in *D. melanogaster* (O'TOUSA *et al.* 1985).

information, we determined whether the distance between the matepairs in *D. pseudoobscura* was comparable to that in *D. persimilis*. This process was repeated throughout the putative breakpoint regions until we identified positions <40,000 bases apart in *D. persimilis* that were several megabases apart in *D. pseudoobscura*.

A misassembly in the *D. pseudoobscura* genome sequence could give a false signature of such an inversion breakpoint. However, leapfrogging through the telomeric 200-kb interval highlighted through linkage mapping provided a direct link to the previously highlighted centromeric 400-kb interval. A misassembly would have only a  $\leq 1\%$  probability of linking directly from one interval to the other (rather than a random location), so it is more parsimonious that these are the true inversion breakpoint regions distinguishing these species. The same approach was conducted on the XL and XR chromosome arms and, again, complemented by supporting recombinational linkage map data from within the two species (ORTÍZ-BARRIENTOS *et al.* 2006).

**Second chromosome loci sequenced:** The second chromosome has the most complete and accurate assembly of the *D. pseudoobscura* genome, with a length of  $\sim 30$  Mb. We extracted sequences 3 kb long every megabase that were interspersed with sequences from five loci previously studied (see below) (MACHADO *et al.* 2002) (Table 1). The sequences were aligned to the *D. pseudoobscura* genome using the BLAST tool from the DroSpeGe website (<http://insects.eugenes.org/species/>) and sequence regions that aligned only to the same unique genomic location (*i.e.*, without repetitive regions) were used for primer design. Primers were designed using Primer3 (ROZEN and SKALETSKY 2000) and BLASTed against the *D. pseudoobscura* genome to ensure that their sequences were unique. The primer sequences used to amplify and sequence each locus are available from the authors upon request. Markers were PCR amplified, cleaned with Millipore (Bedford, MA) filters, and sequenced in both directions in an ABI 3700

sequencer using standard protocols. Nucleotide sequences were deposited in GenBank with accession nos. EF041145–EF041478.

Table 1 lists the 18 sequenced regions, their approximate location in release 2.0 of the assembled second chromosome of *D. pseudoobscura*, and their relative locations with respect to the inversion. The sequences from five of those loci have been previously published (MACHADO *et al.* 2002): the coding regions *rh1* and *bcd* (312 and 397 codons, respectively) and the noncoding regions 2001, 2002, and 2003. Those loci were sequenced in 10–20 inbred lines of each of the three species, as well as in 1–4 lines of the outgroup *D. miranda* (MACHADO *et al.* 2002). We obtained new sequences from 12 additional loci spread across the second chromosome and from the distal breakpoint inversion region (2M05, 2M06, 2M07, 2M12, 2M13, 2M14, 2M15, 2M16, 2M17, 2M17.3, 2M18, 2M20, and *BrkPt*). Data for those loci were collected for up to 8 inbred lines of *D. pseudoobscura* and *D. persimilis*, for up to 6 inbred lines of *D. pseudoobscura bogotana*, and for 1 or 2 lines of *D. miranda*. Names of the newly sequenced loci refer to their approximate locations, in megabases, in the assembled second chromosome. Loci 2M06, 2M15, 2M16, 2M18, and *BrkPt* correspond to noncoding DNA on the basis of the current annotation of the *D. pseudoobscura* genome and comparison to the annotated *D. melanogaster* genome.

Locus 2M05 corresponds to the 3' end of GA17984, with four predicted exons (175 codons). We found an error in the predicted FlyBase annotation of this gene. One intron is not included in the original annotation. All our sequences had a stop codon in the overlooked intron. This stop codon was generated by a T present in every sequence from all four species (and also shared by the *D. persimilis* genome), which was a C in the *D. pseudoobscura* genome sequence. In addition, the alignment of the predicted protein to its homolog in *D. melanogaster* (CG4141) shows a large gap precisely at the location of the intron. Locus 2M07 corresponds to the middle

of GA17887, with one intron and one exon (154 codons). Locus *2M12* corresponds to the 5' end of GA16110 in the middle of a very long intron (sequence is entirely intronic). Locus *2M13* corresponds to GA15148, with four exons and three introns close to the 3' end of the gene (238 codons). Locus *2M14* corresponds to GA16040, with four exons and three introns in the middle of the gene (140 codons). Locus *2M17.3* corresponds to the 3' end of GA15244, with part of the last exon and the 3'-UTR (113 codons). Locus *2M20* corresponds to the 3' end of GA19991, with three exons and two introns (250 codons). Finally, we obtained sequences from the distal breakpoint region of the fixed inversion difference separating *D. pseudoobscura* and *D. persimilis* in the second chromosome (*BrkPt*).

**Other chromosome loci sequenced:** We also examined sequence variation at loci just inside two other inversions that distinguish *D. pseudoobscura* and *D. persimilis* (in XL and XR). Marker X\_2102 is the region just inside the XR inversion breakpoint, and X\_2105 is just inside the XL inversion breakpoint (both on the centromeric side in *D. pseudoobscura*). The X\_2102 data have the addition of the sex-ratio *D. persimilis* strains, which appear homosequential with *D. pseudoobscura* on the basis of polytene chromosome preparations. These two markers were selected for not being in known or predicted coding regions as well as for general proximity to, and position between, the inversion breakpoints. The names derive from the old names of the *D. pseudoobscura* contigs from which the original sequences were identified. Primers for PCR amplification were designed from the *D. pseudoobscura* genome sequence available at the Baylor College of Medicine Human Genome Sequencing Center's *Drosophila* Genome project (<http://hgsc.bcm.tmc.edu/projects/Drosophila/>). The primer sequences used to amplify and sequence each locus are available from the authors upon request.

Additionally, we sequenced two strains of *D. persimilis* for the *vg* locus (ID no. GA17716) on the third chromosome, which has been studied recently in *D. pseudoobscura* (SCHAEFFER and ANDERSON 2005). We limited our analyses to these two strains because they bear the standard (ST) chromosome arrangement, which is shared by the two species. We did not explore sequences in other arrangements because the very old and rich arrangement polymorphisms within both of these species would greatly complicate interpretation of any analyses. We use the published sequences for *D. pseudoobscura* ST strains and *D. miranda* in our analyses (GenBank accession nos. AF476881–AF476896 and AF476921). All but the 38 bp at the 3' end of the sequenced region were intronic. All products were amplified via PCR, purified via the Qiaquick Gel Extraction kit (QIAGEN, Valencia, CA), and sequenced in both directions in an ABI 3730xl DNA sequencer (Perkin-Elmer, Norwalk, CT) using standard protocols.

**Data analyses:** Sequences from each data set were edited and aligned with the program Sequencher (Genecodes, Ann Arbor, MI). Some alignments required manual improvement due to the presence of multiple indels. Basic polymorphism analyses were performed with the programs SITES (HEY and WAKELEY 1997) and DnaSP version 4.10 (ROZAS and ROZAS 1999). Indels were not included in the polymorphism analyses. Single-locus neutrality tests with (FU and LI 1993; FAY and WU 2000) or without (TAJIMA 1989) an outgroup were conducted in DnaSP, and significance was determined with 10,000 coalescent simulations without recombination. The average intra-locus linkage disequilibrium was estimated using  $Z_{ns}$  (KELLY 1997) and significance was determined with 10,000 coalescent simulations without recombination as implemented in DnaSP. Occurrence of gene conversion was explored using BETRAN *et al.*'s (1997) method as implemented in DnaSP 4.10. We determined whether the amounts of polymorphism and divergence

across loci in the second chromosome are correlated, as expected under neutrality, using the HKA test (HUDSON *et al.* 1987). The significance of the HKA test statistic was determined using a distribution generated from 10,000 coalescent simulations that were conducted under three assumptions: (a) constant population size of the two species, (b) the population size of the ancestor is the average of that of the two species, and (c) no recombination. Using the program HKA (Jody Hey, Rutgers University), HKA tests were applied to data from each of the three species using a single outgroup sequence from *D. miranda* or using all sequences from one pair of species. Locus *BrkPt*, located in the distal breakpoint region of chromosome 2, was analyzed separately and not included in the HKA analyses. Because *D. miranda* has the same chromosomal arrangement of *D. pseudoobscura* and *D. p. bogotana*, the divergence between *D. miranda* and *D. persimilis* at *BrkPt* is not presented because partially nonhomologous regions were sequenced (Table 2). It is thus not possible to compare polymorphism and divergence across loci using *BrkPt*. We also applied the McDonald–Kreitman test (McDONALD and KREITMAN 1991) to every coding region using DnaSP. This test examines whether the ratio of silent to replacement variation is the same for polymorphisms as it is for fixed differences between species. Under the assumption that these two kinds of variation are selectively neutral, the ratios are expected to be the same.

The polymorphism data were fitted to a model of speciation with no gene flow (WAKELEY and HEY 1997) using the WH program (WANG *et al.* 1997). The test statistic [Wang–Wakeley–Hey (WWH)] is the difference between the highest and lowest number of fixed differences across loci plus the difference between the highest and lowest number of shared polymorphisms (WANG *et al.* 1997). In addition, a  $\chi^2$ -statistic was computed by comparing the observed and expected values of shared polymorphisms, exclusive polymorphisms, and fixed differences as described by KLIMAN *et al.* (2000). To determine the level of significance, 10,000 coalescent simulations were conducted. The *P*-value for both the  $\chi^2$ - and the WWH-test statistics corresponds to the proportion of simulated values that are greater than or equal to those observed. The tests are one tailed because the focus is on detecting a departure from the model in the direction expected if historical gene flow had occurred. Tests were performed for different combinations of loci on the basis of their location with respect to the inversion.

Fixed differences between species and  $F_{ST}$ -based estimates of migration rate ( $Nm$ ) (HUDSON *et al.* 1992) were obtained using SITES (HEY and WAKELEY 1997). In the case of the comparison between *D. pseudoobscura* and *D. p. bogotana*, it is accepted that there is no current gene flow between them due to their fully allopatric distribution and that historically there is no evidence of gene flow during their divergence (MACHADO *et al.* 2002). Estimates of  $Nm$  for that subspecies pair are thus reflecting just levels of shared ancestral variation. We also applied FEDER *et al.*'s (2005) “relative node depth” (RND) test to the DNA sequence data for evidence of introgression (see also PATTERSON *et al.* 2006). This test was done for each locus separately by dividing the average pairwise sequence difference between *D. pseudoobscura* and *D. persimilis* sequences by the average pairwise sequence difference between these species and *D. miranda* (averaging that of *D. pseudoobscura* and *D. persimilis*). Assuming a molecular clock, RND presents a proportion of divergence that occurred after the split of these species relative to the total divergence from *D. miranda*. If there has been recent gene flow between *D. pseudoobscura* and *D. persimilis* at certain loci, RND will be reduced relative to that of other loci. Here, we specifically tested for a general pattern of lower RND with greater physical distance from the inversion breakpoints in collinear regions, suggesting greater introgression further from the inversion. Standard statistical analyses

**TABLE 2**  
**Polymorphism statistics of the second chromosome loci**

Locus	Species	<i>n</i> <sup>a</sup>	<i>L</i> <sup>b</sup>	<i>S</i> <sup>c</sup>	Syn <sup>d</sup>	Rep <sup>d</sup>	$\hat{\theta}$ <sup>e</sup>	$\pi$ <sup>f</sup>	<i>D</i> <sup>g</sup>	<i>Z</i> <sub>ns</sub> <sup>h</sup>	Diff. <sup>i</sup>
2M05	<i>pseudoobscura</i>	8	886.0	13	3	0	0.00566	0.00427	-1.2403	0.169	0.02765
	<i>bogotana</i>	6	886.0	4	1	0	0.00198	0.00150	-1.2950	1.0**	0.02897
	<i>persimilis</i>	8	884.3	13	3	0	0.00567	0.00404	-1.4567	0.303	0.01695
	<i>miranda</i>	2	886.0	2	0	0	0.00226	0.00226	—	—	—
2M06	<i>pseudoobscura</i>	8	840.4	32	—	—	0.01469	0.01232	-0.8543	0.177	0.01657
	<i>bogotana</i>	6	841.1	4	—	—	0.00208	0.00285	2.0058*	1.0**	0.01519
	<i>persimilis</i>	8	816.9	24	—	—	0.01133	0.00861	-1.2601	0.173	0.01328
	<i>miranda</i>	2	841.0	3	—	—	0.00357	0.00357	—	—	—
2M07	<i>pseudoobscura</i>	6	849.5	53	17	1	0.02733	0.02551	-0.4269	0.211	0.03461
	<i>bogotana</i>	6	835.4	16	6	1	0.00839	0.00734	-0.7713	0.412	0.03140
	<i>persimilis</i>	6	803.4	20	2	1	0.01090	0.01087	-0.0184	0.252	0.03419
	<i>miranda</i>	1	798	—	—	—	—	—	—	—	—
2M12	<i>pseudoobscura</i>	8	840.6	79	—	—	0.03625	0.03246	-0.5667	0.148	0.05332
	<i>bogotana</i>	6	830.7	31	—	—	0.01634	0.01966	1.2874	0.942**	0.05454
	<i>persimilis</i>	7	843.0	15	—	—	0.00726	0.00576	-1.1424	0.169	0.04470
	<i>miranda</i>	2	821.0	16	—	—	0.01949	0.01949	—	—	—
2M13	<i>pseudoobscura</i>	6	908.0	11	6	4	0.00531	0.00595	0.7312	0.358	0.01780
	<i>bogotana</i>	6	908.0	5	2	1	0.00241	0.00250	0.1965	0.370	0.01836
	<i>persimilis</i>	8	908.0	10	4	2	0.00425	0.00334	-1.0539	0.162	0.01666
	<i>miranda</i>	2	908.0	2	1	1	0.00220	0.00220	—	—	—
2M14	<i>pseudoobscura</i>	7	900.5	26	2	0	0.01178	0.01089	-0.4294	0.185	0.01593
	<i>bogotana</i>	6	900.8	5	0	0	0.00243	0.00318	1.7300	0.700	0.01407
	<i>persimilis</i>	8	891.0	6	0	0	0.00260	0.00200	-1.0643	0.128	0.01145
	<i>miranda</i>	2	905.0	1	0	0	0.00110	0.00110	—	—	—
2M15	<i>pseudoobscura</i>	8	883.0	10	—	—	0.00437	0.00392	-0.5038	0.219	0.01047
	<i>bogotana</i>	5	866.0	3	—	—	0.00166	0.00162	-0.1747	0.444	0.01156
	<i>persimilis</i>	8	878.3	3	—	—	0.00132	0.00085	-1.4475	0.020	0.01415
	<i>miranda</i>	2	878.0	1	—	—	0.00114	0.00114	—	—	—
2M16	<i>pseudoobscura</i>	6	876.9	21	—	—	0.01049	0.00988	-0.3610	0.277	0.03225
	<i>bogotana</i>	6	882.3	5	—	—	0.00248	0.00189	-1.3370	0.232	0.02951
	<i>persimilis</i>	6	861.9	16	—	—	0.00813	0.00665	-1.1244	0.243	0.03220
	<i>miranda</i>	1	865.0	—	—	—	—	—	—	—	—
2M17	<i>pseudoobscura</i>	8	813.0	37	—	—	0.01755	0.01375	-1.1564	0.144	0.03293
	<i>bogotana</i>	6	818.0	0	—	—	0	0	—	—	0.02550
	<i>persimilis</i>	7	1306.4	24	—	—	0.00750	0.00733	-0.1297	0.230	0.03718
	<i>miranda</i>	1	1294.0	—	—	—	—	—	—	—	—
2M17.3	<i>pseudoobscura</i>	6	814.1	29	0	6	0.01560	0.01458	-0.4156	0.301	0.03940
	<i>bogotana</i>	6	770.4	20	1	4	0.01137	0.01064	-0.3988	0.267	0.03806
	<i>persimilis</i>	7	808.7	22	0	2	0.01110	0.00930	-0.9142	0.239	0.03205
	<i>miranda</i>	2	849.0	9	0	4	0.01060	0.01060	—	—	—
2M18	<i>pseudoobscura</i>	8	800.5	35	—	—	0.01686	0.01245	-1.3952	0.155	0.03035
	<i>bogotana</i>	6	806.0	8	—	—	0.00435	0.00529	1.2839	1.0**	0.03322
	<i>persimilis</i>	8	802.5	22	—	—	0.01057	0.00783	-1.3559	0.208	0.03499
	<i>miranda</i>	1	809.0	—	—	—	—	—	—	—	—

(continued)

were conducted on JMP (SAS Institute, Cary, NC). Phylogenetic analyses were conducted using version 4.0b1 of PAUP\* (SWOFFORD 1998). Gene trees were reconstructed using the neighbor-joining (NJ) algorithm with Tamura–Nei distances (TAMURA and NEI 1993).

RESULTS

**Identification of all fixed chromosomal inversion breakpoints:** We used a leapfrogging approach to iden-

tify the breakpoints of the XL, XR, and second chromosome inversions distinguishing *D. pseudoobscura* and *D. persimilis* from the assembled genome sequence of *D. pseudoobscura* and the genome sequence traces of *D. persimilis*. We characterize these breakpoint regions elsewhere. Some slight imprecision in the estimation of the breakpoints results from repetitive sequences found in these regions.

The second chromosome inversion breakpoints are near positions 9,449,800 and 17,094,500 in release 2.01

**TABLE 2**  
(Continued)

Locus	Species	$n^a$	$L^b$	$S^c$	Syn <sup>d</sup>	Rep <sup>d</sup>	$\hat{\theta}^e$	$\pi^f$	$D^g$	$Z_{ns}^h$	Diff. <sup>i</sup>
2M20	<i>pseudoobscura</i>	8	872.0	21	15	2	0.00929	0.00782	-0.8234	0.217	0.01690
	<i>bogotana</i>	6	872.0	8	6	0	0.00402	0.00306	-1.4083	1.0**	0.01716
	<i>persimilis</i>	8	872.0	11	10	0	0.00487	0.00426	-0.6217	0.137	0.01659
	<i>miranda</i>	1	806.0	—	—	—	—	—	—	—	—
bcd <sup>j</sup>	<i>pseudoobscura</i>	20	1367.9	47	25	11	0.0097	0.0074	-0.9633	0.076	0.0192
	<i>bogotana</i>	14	1370.8	14	8	4	0.0032	0.0030	-0.2488	0.288	0.0191
	<i>persimilis</i>	14	1375.4	30	18	7	0.0068	0.0058	-0.6620	0.104	0.0176
	<i>miranda</i>	2	1369.0	0	0	0	0	0	—	—	—
rh1 <sup>j</sup>	<i>pseudoobscura</i>	17	1443.0	52	14	0	0.0106	0.0071	-1.3903	0.091	0.0151
	<i>bogotana</i>	11	1200.0	3	3	0	0.0008	0.0007	-0.7494	0.028	0.0148
	<i>persimilis</i>	14	1443.1	41	12	0	0.0089	0.0069	-1.0102	0.090	0.0155
	<i>miranda</i>	2	1443.0	1	1	0	0.0007	0.0007	—	—	—
2003 <sup>j</sup>	<i>pseudoobscura</i>	18	512.8	18	—	—	0.0102	0.0068	-1.2721	0.097	0.0195
	<i>bogotana</i>	14	528.3	7	—	—	0.0042	0.0027	-1.2767	0.202	0.0143
	<i>persimilis</i>	14	532.3	11	—	—	0.0065	0.0038	-1.6374	0.094	0.0176
	<i>miranda</i>	3	507.0	1	—	—	0.0013	0.0013	—	—	—
2002 <sup>j</sup>	<i>pseudoobscura</i>	19	924.8	66	—	—	0.0204	0.0152	-1.0396	0.076	0.0193
	<i>bogotana</i>	13	922.8	17	—	—	0.0059	0.0071	0.8372	0.434	0.0162
	<i>persimilis</i>	13	905.2	16	—	—	0.0057	0.0038	-1.3807	0.124	0.0171
	<i>miranda</i>	3	918.3	21	—	—	0.0152	0.0152	—	—	—
2001 <sup>j</sup>	<i>pseudoobscura</i>	17	678.1	38	—	—	0.0166	0.0108	-1.4482	0.069	0.0210
	<i>bogotana</i>	13	694.4	5	—	—	0.0023	0.0018	-0.8419	0.102	0.0222
	<i>persimilis</i>	14	677.2	20	—	—	0.0093	0.0073	-0.9079	0.104	0.0175
	<i>miranda</i>	1	690.0	—	—	—	—	—	—	—	—
Bkpt <sup>k</sup>	<i>pseudoobscura</i>	12	541.1	57	—	—	0.0349	0.0258	-1.2009	0.108	0.0599
	<i>bogotana</i>	5	571.0	6	—	—	0.0050	0.0056	0.7637	0.528	0.0623
	<i>persimilis</i>	6	569.6	4	—	—	0.0030	0.0037	1.1806	0.600	— <sup>k</sup>
	<i>miranda</i>	1	922.0	—	—	—	—	—	—	—	—

—, values could not be obtained for small samples or for groups of sequences with few informative sites. \*, significant at  $P < 0.05$ ; \*\*, significant at  $P < 0.01$ .

<sup>a</sup>Number of lines sequenced.

<sup>b</sup>Average length (base pairs) of the sequences from each species.

<sup>c</sup>Number of polymorphic sites.

<sup>d</sup>Number of synonymous (syn) and replacement (rep) polymorphisms in the coding regions.

<sup>e</sup>Estimate of  $4N\mu$ /bp using the number of polymorphic sites (WATTERSON 1975).

<sup>f</sup>Estimate of  $4N\mu$  using the average number of nucleotide differences per site (TAJIMA 1983).

<sup>g</sup>TAJIMA's (1989) statistic. Significance was determined using 10,000 coalescent simulations with no recombination.

<sup>h</sup>Average intralocus linkage disequilibrium (KELLY 1997). Significance was determined using 1000 coalescent simulations with no recombination.

<sup>i</sup>Average pairwise difference per base pair between alleles from each taxon and the alleles of *D. miranda*.

<sup>j</sup>Data from MACHADO *et al.* (2002).

<sup>k</sup>Distal breakpoint sequence of the fixed inversion separating *D. pseudoobscura* and *D. persimilis*. Divergence between *D. miranda* and *D. persimilis* is not presented because *D. miranda* has a similar arrangement to *D. pseudoobscura* and therefore partially non-homologous regions were compared.

of the *D. pseudoobscura* genome sequence, which is consistent with the coarse localizations we published earlier (ORTÍZ-BARRIENTOS *et al.* 2006). The sequenced region *Bkpt* was designed in the former of these two locations. We use the identified positions of these breakpoints and the *D. pseudoobscura* genome sequence assembly (RICHARDS *et al.* 2005) to determine the distance of our sequenced regions from the inversion breakpoints. These regions correspond to roughly positions 11,030,000 and 3,259,000 in scaffold 0 of the *D. persimilis* genome sequence assembly comparative analysis freeze 1 (GILBERT 2005). On the basis of these results, loci 2M12,

2M13, 2M14, 2M15, 2M16, and 2M17 are located inside the inverted region. Locus 2002 was considered by MACHADO *et al.* (2002) to be located inside the inverted region on the basis of *in situ* hybridizations. However, our mapping shows that this marker is ~1.4 Mb outside the inversion (ORTÍZ-BARRIENTOS *et al.* 2006).

The XL inversion breakpoints are located near scaffold XL\_group1e position 6,945,800 and scaffold XL\_group1a position 7,006,900. These locations correspond to *D. persimilis* scaffold 13 position 1,611,500 and scaffold 25 position 262,400. The XR inversion breakpoints are near *D. pseudoobscura* scaffold XR\_group6

approximate position 6,160,000 and scaffold XR\_group8 position 7,190,000 (observable in the *D. pseudoobscura* BLAST positions of *D. persimilis* genome sequence traces 814098501 and 817773456 relative to their matepairs). These locations correspond to *D. persimilis* scaffold 36, position 80,000, and the origin of scaffold 48. The XR breakpoints were less precisely localized ( $\pm 15,000$  bp) than the others because of thousands of bases of repetitive sequence in the breakpoint regions.

**DNA polymorphism in the second chromosome:** Polymorphism statistics for all the second chromosome loci are presented in Table 2. Locus *2M17* is the only one exhibiting complete lack of variation in one species, *D. p. bogotana*. The weighted average values of nucleotide diversity per base pair estimated using the number of segregating mutations,  $\hat{\theta}$  (WATTERSON 1975), or the average number of pairwise nucleotide differences,  $\pi$  (TAJIMA 1983), for the 18 loci from the second chromosome are, respectively, 0.0154 and 0.0120 for *D. pseudoobscura*, 0.0067 and 0.0054 for *D. persimilis*, and 0.0049 and 0.0041 for *D. p. bogotana*. These results agree with previous inferences of a larger historic effective population size for *D. pseudoobscura* than for its two close relatives (RILEY *et al.* 1992; SCHAEFFER and MILLER 1992a; WELLS 1996; HAMBLIN and AQUADRO 1999; MACHADO *et al.* 2002) and a smaller historical effective size for *D. p. bogotana* consistent with a founder effect (PRAKASH *et al.* 1969; MACHADO *et al.* 2002). A previous survey of 14 loci located in all chromosome arms of these three species (MACHADO *et al.* 2002) found the weighted average values of  $\hat{\theta}$  for autosomal loci of *D. pseudoobscura*, *D. persimilis*, and *D. p. bogotana* to be 0.0148, 0.0097, and 0.0059, respectively. There is thus a slight, albeit nonsignificant, reduction of nucleotide variation in *D. persimilis* in the second chromosome relative to the other autosomes (Mann–Whitney *U*-test:  $\hat{\theta}$ ,  $Z = 1.74$ ,  $P = 0.081$ ;  $\pi$ ,  $Z = 1.49$ ,  $P = 0.136$ ).

More importantly, in *D. persimilis* the average nucleotide variation inside the inverted region on the second chromosome ( $\hat{\theta} = 0.0048$ ) is significantly lower than the average variation outside the inversion ( $\hat{\theta} = 0.0083$ ) for loci in the same chromosome (Mann–Whitney *U*-test:  $\hat{\theta}$ ,  $Z = -2.08$ ,  $P = 0.037$ ;  $\pi$ ,  $Z = -1.99$ ,  $P = 0.046$ ). This is, however, not the case in *D. pseudoobscura* ( $\hat{\theta}$ ,  $Z = 0.09$ ,  $P = 0.927$ ;  $\pi$ ,  $Z = 0.27$ ,  $P = 0.786$ ) or in *D. p. bogotana* ( $\hat{\theta}$ ,  $Z = -0.54$ ,  $P = 0.587$ ;  $\pi$ ,  $Z = -0.36$ ,  $P = 0.717$ ). Further, variation inside the inversion is significantly different from that in the rest of the remaining loci in chromosome 2 plus loci on all the other autosomes for *D. persimilis* ( $\hat{\theta}$ ,  $Z = -2.32$ ,  $P = 0.020$ ;  $\pi$ ,  $Z = -2.11$ ,  $P = 0.034$ ) but not for *D. pseudoobscura* ( $\hat{\theta}$ ,  $Z = 0.07$ ,  $P = 0.943$ ;  $\pi$ ,  $Z = 0.211$ ,  $P = 0.832$ ) or *D. p. bogotana* ( $\hat{\theta}$ ,  $Z = -0.84$ ,  $P = 0.398$ ;  $\pi$ ,  $Z = -0.74$ ,  $P = 0.459$ ). Figure 1 shows the variation in nucleotide diversity across loci along the second chromosome and provides a graphic representation of the higher diversity observed in *D. pseudoobscura* and the reduction of variation

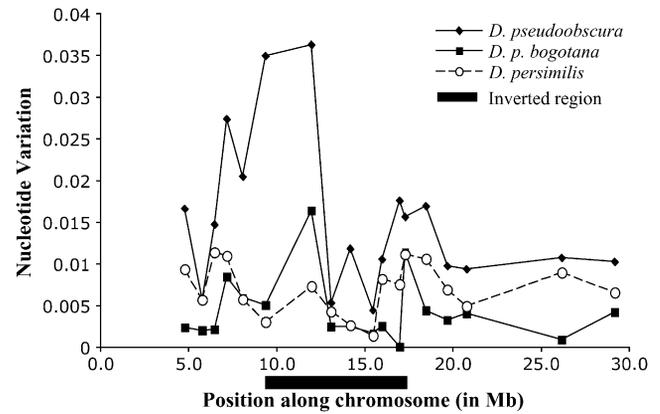


FIGURE 1.—Plot of nucleotide variation ( $\theta$ ) vs. chromosome location for the 18 second chromosome loci of the three species included in this study. The location of the inverted region is indicated.

in the inverted region of *D. persimilis*. The significant reduction in variation in the derived inverted arrangement of *D. persimilis* suggests a scenario in which this inverted arrangement arose and was rapidly fixed possibly through natural selection, leading to a reduction of variation at all linked loci inside the inversion. This scenario is consistent with predictions of a model presented by NAVARRO *et al.* (2000) on the establishment of inversion polymorphisms. A potential complication with respect to this conclusion is the observation of three shared polymorphisms in *2M12* (Table 3), a locus located inside the inversion. Under our proposed scenario, the shared polymorphisms are homoplasies that could have

TABLE 3

The number of shared polymorphisms and fixed differences between species in second chromosome loci

Locus	<i>pseudoobscura</i> / <i>bogotana</i>		<i>pseudoobscura</i> / <i>persimilis</i>		Location relative to inversion
	Shared	Fixed	Shared	Fixed	
<i>2001</i>	1	3	6	0	Outside
<i>2M05</i>	4	0	0	18	Outside
<i>2M06</i>	1	1	7	0	Outside
<i>2M07</i>	9	0	3	9	Outside
<i>2002</i>	8	0	0	6	Outside
<i>2M12</i>	21	1	3	6	Inside
<i>2M13</i>	2	1	0	7	Inside
<i>2M14</i>	3	0	0	6	Inside
<i>2M15</i>	0	0	0	6	Inside
<i>2M16</i>	1	1	0	5	Inside
<i>2M17</i>	0	3	0	21	Inside
<i>2M17.3</i>	12	1	1	20	Outside
<i>2M18</i>	3	1	1	5	Outside
<i>bcd</i>	4	0	5	0	Outside
<i>2M20</i>	1	2	1	1	Outside
<i>rh1</i>	1	2	8	0	Outside
<i>2003</i>	1	0	2	0	Outside

arisen only through recurrent mutation. The probability of observing a given number of shared polymorphisms by independent mutation at the same sites in two isolated species can be estimated using a hypergeometric distribution (CLARK 1997). On the basis of the levels of polymorphism in *2M12* in each species and the sequence length, there is a 15.97% chance that the two species would share three polymorphic sites by recurrent mutation, and thus our scenario is still consistent with the data at this locus.

**Testing the neutral model in the second chromosome loci:** We observed no evidence of rejection of the neutral model using the HKA test in any comparison (*D. pseudoobscura/D. miranda*,  $\chi^2 = 4.66$ ,  $P = 0.926$ ; *D. persimilis/D. miranda*,  $\chi^2 = 7.18$ ,  $P = 0.910$ ; *D. p. bogotana/D. miranda*,  $\chi^2 = 15.59$ ,  $P = 0.347$ ; *D. pseudoobscura/D. persimilis*,  $\chi^2 = 16.85$ ,  $P = 0.839$ ; *D. pseudoobscura/D. p. bogotana*,  $\chi^2 = 10.67$ ,  $P = 0.988$ ; *D. persimilis/D. p. bogotana*,  $\chi^2 = 34.00$ ,  $P = 0.152$ ). Thus, despite the significant reduction in nucleotide diversity in the inverted region of *D. persimilis*, there is no evidence of a recent selective sweep in this region using the HKA test. We also applied the McDonald–Kreitman test (MCDONALD and KREITMAN 1991) to every coding region. In the large majority of cases, the contingency tests could not be applied due to the presence of two or more 0 class cells in the data (not shown). In the cases in which Fisher's exact test could be applied, the test was not significant (*D. pseudoobscura/D. persimilis*, *2M05*,  $P = 0.839$ ; *2M07*,  $P = 1.0$ ; *2M13*,  $P = 0.55$ ; *2M17.3*,  $P = 0.47$ ; *2M20*,  $P = 1.0$ ; *D. pseudoobscura/D. p. bogotana*, *2M05*,  $P = 1.0$ ; *2M13*,  $P = 1.0$ ).

We also examined whether the pattern of variation at each locus within each species was consistent with the neutral model. Confirming the HKA results, the single-locus tests do not reveal evidence of rejection of the neutral model in loci located inside the inverted region in *D. persimilis*. The value of Tajima's *D*-statistic (TAJIMA 1989) was negative in almost all loci (Table 2), but its value was significantly different from zero only in the *2M06* data set from *D. p. bogotana* ( $D = 2.0058$ ,  $P < 0.05$ ). The significant positive value of Tajima's *D* coupled to the significant average intralocus linkage disequilibrium ( $Z_{NS} = 1.0$ ,  $P < 0.01$ ) suggests that this region is under balancing selection in *D. p. bogotana*. Single-locus tests of the neutral model using an outgroup (FU and LI 1993; FAY and WU 2000) were also significant in *D. p. bogotana* at *2M05* ( $H = -3.73$ ,  $P = 0.013$ ), *2M12* ( $D = 1.85$ ,  $P = 0.02$ ;  $F = 2.07$ ,  $P = 0.02$ ), and *2M18* ( $D = 1.73$ ,  $P = 0.03$ ). None of these tests were significant in *D. pseudoobscura*, and in only one locus (*2M05*) were they significant in *D. persimilis* ( $D = -2.46$ ,  $P = 0.003$ ;  $F = -2.66$ ,  $P = 0.002$ ). That locus has an unusually large number of fixed differences (18) between *D. pseudoobscura* and *D. persimilis* (Table 3) despite being well outside the inverted region. Interestingly, although the sequenced region encompasses four exons of a coding

gene (GA17984), of the 18 fixed differences 14 are non-coding, 3 are silent, and only one corresponds to a replacement substitution. Because of the evidence for selection in *D. persimilis* and *D. p. bogotana*, locus *2M05* was not included in some of the analyses described below.

To test whether the average value of Tajima's *D* across loci departs significantly from zero, we conducted a test using coalescent simulations. Ten thousand neutral genealogies were simulated for each locus conditioned on the number of observed polymorphisms and assuming constant population size and no recombination. For *D. pseudoobscura* and *D. persimilis*, the observed mean values of Tajima's *D* were less than all of the means found in 10,000 simulations ( $D = -0.797$ ,  $P < 0.007$ ;  $D = -1.011$ ,  $P < 0.0001$ , respectively), while the *D. p. bogotana* value was not significantly different from zero ( $D = -0.073$ ,  $P = 0.459$ ). For *D. pseudoobscura* and *D. persimilis* the mean values of Fu and Li's *D* (FU and LI 1993) were less than all of the means found in 10,000 simulations ( $D = -1.129$ ,  $P < 0.0001$ ;  $D = -1.445$ ,  $P < 0.0001$ , respectively), while the *D. p. bogotana* value was positive and not significantly different from zero ( $D = 0.324$ ,  $P = 0.948$ ). These results are consistent with previous observations of a skewed mutation frequency spectrum across multiple loci of *D. pseudoobscura* and *D. persimilis* (MACHADO *et al.* 2002), which suggests that these two species are undergoing an expansion in their population size.

**Nucleotide variation in the distal breakpoint region of chromosome 2:** Nucleotide variation at the *BrkPt* locus located in the distal breakpoint region of the second chromosome inversion is 11 times higher in *D. pseudoobscura* than in *D. persimilis* ( $\hat{\theta}$ , 0.0349 vs 0.0030) (Table 2). This locus has the third lowest level of variation in *D. persimilis*, after *2M15* ( $\hat{\theta} = 0.0013$ ) and *2M14* ( $\hat{\theta} = 0.0026$ ), which are located inside the inverted region, but, more importantly, it has the lowest relative level of variation of any loci compared to *D. pseudoobscura*. Interestingly, *BrkPt* is the only locus with a positive Tajima's *D* in *D. persimilis* (1.1806), although this value is not significant. The positive value of Tajima's *D* is possibly the result of the observed haplotype structure. This locus has the highest average intralocus linkage disequilibrium (KELLY 1997) in *D. persimilis* ( $Z_{NS} = 0.600$ ,  $P = 0.281$ ), with three haplotypes having polymorphic bases almost in complete linkage disequilibrium.

Only Fay and Wu's test statistic was significant for this locus ( $H = -1.33$ ,  $P = 0.04$ ). However, this test may not be appropriate in this situation (BAUDRY and DEPAULIS 2003) because of the large and artifactual distance between *D. persimilis* and the outgroup *D. miranda* given their unalignability resulting from the inversion (Table 2). Here, all polymorphisms in *D. persimilis* appear to be derived and at high frequency, leading to significant values of the *H*-statistic. Therefore, we do not consider

**TABLE 4**  
**Estimates of population migration rates and net divergence in the second chromosome loci**

Locus	Net divergence/bp <sup>a</sup>		Population migration rate ( <i>Nm</i> ) <sup>b</sup>	
	<i>pseudoobscura</i> / <i>bogotana</i>	<i>pseudoobscura</i> / <i>persimilis</i>	<i>pseudoobscura</i> / <i>bogotana</i>	<i>pseudoobscura</i> / <i>persimilis</i>
2001	0.00374	0.00282	0.417	0.799
2M05	0.00111	0.02179	0.652	0.048
2M06	0.00376	0.00138	0.504	1.921
2M07	0.01025	0.01393	0.404	0.339
2002	0.00286	0.00822	0.978	0.295
2M12	0.00568	0.01433	1.162	0.339
2M13	0.00355	0.01064	0.297	0.109
2M14	0.00108	0.01074	1.631	0.151
2M15	0.00224	0.00845	0.313	0.071
2M16	0.00285	0.00723	0.517	0.288
2M17	0.00526	0.02876	0.326	0.112
2M17.3	0.00276	0.03222	1.161	0.096
2M18	0.00458	0.00852	0.484	0.297
<i>bcd</i>	0.00310	0.00250	0.418	0.657
2M20	0.00445	0.00374	0.306	0.404
<i>rh1</i>	0.00148	0.00140	0.780	1.247
2003	0.00116	0.00158	1.037	0.842

<sup>a</sup> Calculated using Equation 10.21 of NEI (1987).

<sup>b</sup> Estimated using the method of HUDSON *et al.* (1992).

this result to provide evidence of recent hitchhiking at this locus in *D. persimilis*.

**Multilocus analyses of species divergence:** Tables 3 and 4 present basic statistics used to describe species divergence using multilocus data sets. The number of fixed differences and shared polymorphisms across loci are shown in Table 3. Under a null model of speciation, recently diverged species are expected to share variation that is lost over time through genetic drift. This process leads to an accumulation of fixed differences and a loss of shared polymorphisms as divergence time increases. When considering data from multiple loci, it is thus expected to see a negative correlation between fixed differences and shared polymorphisms across loci. This pattern is expected even if gene flow has occurred at some loci. The correlation between fixed and shared variation in our second chromosome data (Table 3) is negative and significant for the *D. pseudoobscura*/*D. persimilis* comparison (Spearman's  $\rho = -0.6501$ ,  $P = 0.0047$ ; Kendall's  $\tau = -0.5308$ ,  $P = 0.0073$ ) and negative but not significant for the *D. pseudoobscura*/*D. p. bogotana* comparison (Spearman's  $\rho = -0.3980$ ,  $P = 0.113$ ; Kendall's  $\tau = -0.3441$ ,  $P = 0.0944$ ). Three loci (2M05, 2M17, and 2M17.3), two of which have coding sequences (2M05 and 2M17.3), show a large number (18–21) of fixed differences between *D. pseudoobscura* and *D. persimilis* (Table 3). These differences are about twice as large as what had been reported earlier for this species pair at a different set of loci (MACHADO *et al.* 2002). As discussed above, 2M05 was the only locus in which the neutral model was rejected in *D. persimilis*, and the observation of increased numbers of fixed differences

may be a consequence of directional selection. In the case of 2M17.3, although 7 of the 20 fixed differences between *D. pseudoobscura* and *D. persimilis* are replacement differences, there is no evidence of natural selection driving the fixation of those differences on the basis of the McDonald–Kreitman test (see above). That locus seems to be under relaxed constraints for amino acid polymorphism and substitution.

Estimates of net divergence and population migration rates between the three taxa are shown in Table 4. Net divergence for each locus [average divergence from the outgroup minus average intraspecific polymorphism (NEI 1987)] is expected to be proportional to the time since species divergence. If introgression has occurred since the species started to diverge, we expect a large variance in levels of net divergence across loci. Further, when comparing net divergence values of *D. pseudoobscura*/*D. persimilis* with *D. pseudoobscura*/*D. p. bogotana*, inside and outside the inverted region, we expect all *D. pseudoobscura*/*D. persimilis* values to be larger inside the inversion but some values to be smaller outside the inversion due to gene flow at some loci. That is the pattern observed. Net divergence values are significantly larger in *D. pseudoobscura*/*D. persimilis* for all loci inside the inversion (Wilcoxon's signed rank test,  $Z = 10.5$ ,  $P = 0.031$ , d.f. = 5) but not for loci outside the inversion ( $Z = 14.0$ ,  $P = 0.24$ , d.f. = 10). However, when considering all second chromosome loci, net divergence values are significantly larger in the *D. pseudoobscura*/*D. persimilis* contrast ( $Z = 57.5$ ,  $P = 0.005$ , d.f. = 16). Similarly, population migration rates (*Nm*), estimated using average divergences, are significantly lower for the

TABLE 5

## Isolation model fitting for the second chromosome loci

Species 1 Species 2	Loci	$\chi^2$	$P\chi^2$	WWH	$P_{WWH}$
<i>pseudoobscura</i> <i>persimilis</i>	All	152.7	0.3562	29	0.3026
<i>pseudoobscura</i> <i>persimilis</i>	Inversion	46.0	0.3127	19	0.5114
<i>pseudoobscura</i> <i>persimilis</i>	Inversion plus close <sup>a</sup>	64.2	0.4223	19	0.6631
<i>pseudoobscura</i> <i>persimilis</i>	Outside inversion	82.6	0.3427	28	0.1205
<i>pseudoobscura</i> <i>persimilis</i>	Outside centromere	43.13	0.5328	27	0.0888
<i>pseudoobscura</i> <i>persimilis</i>	Outside telomere	27.1	0.1507	16	0.2225
<i>pseudoobscura</i> <i>bogotana</i>	All	90.4	0.9473	15	0.8808

Due to evidence of selection (see text) locus *2M05* was not included in the *D. pseudoobscura/D. persimilis* analyses, and loci *2M05*, *2M06*, *2M12*, and *2M18* were not included in the *D. pseudoobscura/D. p. bogotana* analyses.

<sup>a</sup>Close to inversion includes loci outside the inversion that are located within 1.5 Mb of the breakpoint. *P*-values are not corrected for multiple tests.

*D. pseudoobscura/D. persimilis* contrast inside the inversion ( $Z = -10.5$ ,  $P = 0.031$ , d.f. = 5) but not different outside the inverted region ( $Z = -2.0$ ,  $P = 0.89$ , d.f. = 11). When all loci in the second chromosome are considered, *Nm* values are not significantly larger in the *D. pseudoobscura/D. persimilis* contrast than in the *D. pseudoobscura/D. p. bogotana* comparison ( $Z = -29.5$ ,  $P = 0.174$ , d.f. = 16). These results for net divergence and *Nm* are consistent with a model of species divergence in which gene flow between *D. pseudoobscura* and *D. persimilis* has occurred at some loci outside the inverted region but has not occurred inside the inversion since species divergence (MACHADO *et al.* 2002).

We fitted the polymorphism data to a model of species divergence with no gene flow (isolation model) (WAKELEY and HEY 1997), using the WH program (WANG *et al.* 1997). We conducted a series of analyses of the isolation model including different loci combinations to contrast patterns in the inverted region with patterns outside the inverted region (Table 5). Loci showing evidence of selection were not used in the analyses. On the basis of previous results (WANG *et al.* 1997; MACHADO *et al.* 2002), we predict that in the *D. pseudoobscura/D. persimilis* comparison the isolation model may be rejected for all loci and for the suite of loci outside the inversion, but will not be rejected for loci inside the inversion. Further, we predict that the isolation model will not be rejected for the *D. pseudoobscura/D. p. bogotana* comparison. However, we could not reject the null isolation model in any of the comparisons (Table 5). This result is perhaps

unexpected for the *D. pseudoobscura/D. persimilis* comparisons given that we observe a large variance in shared variation and fixed differences across loci (Table 3). However, an important difference from two previous studies in this species pair (WANG *et al.* 1997; MACHADO *et al.* 2002) is that none of the loci presented here has a large number of shared polymorphisms. As the value of the WWH test statistic depends on the largest and smallest numbers of shared polymorphisms and fixed differences across loci (see MATERIALS AND METHODS) (WANG *et al.* 1997), this test tends to be overly conservative when the variance in shared variation is low across loci. Unfortunately, application of a more sensitive multilocus test of introgression (NIELSEN and WAKELEY 2001; HEY and NIELSEN 2004) is difficult due to high levels of recombination observed in our data.

**The effects of distance to inversion on interspecific gene flow and fixed genetic variation:** To assess the effect of distance from inversion breakpoint on gene flow and fixed genetic variation between *D. pseudoobscura* and *D. persimilis*, we plotted several variables against distance to nearest inversion breakpoint and conducted logarithmic regression analyses. We predict that this relationship should be logarithmic rather than linear because the inversion should affect recombination only for loci close to the breakpoint, while loci located far from it should introgress and recombine freely between species and thus show no notable effect regardless of their distance to the inversion breakpoint. While a standard linear regression would not capture that effect, a log-linear regression would. Although we test this prediction separately for loci located inside and outside the inversion, we expect a better fit for the loci outside the inversion. Any signature of gene conversion inside the inverted region may not exhibit a linear or log-linear form, but may instead show only a very steep effect right by the inversion breakpoint. Thus, for loci inside the inversion we use the log regression as an approximation but not as a strict perfect expectation.

The percentage of fixed differences per locus decreases dramatically with distance to inversion breakpoint (Figure 2). The logarithmic regression is significant for all loci outside the inversion ( $N = 10$ ,  $r^2 = 0.77$ ,  $P = 0.0008$ ). Interestingly, the regression is also significant for the loci inside the inversion ( $N = 6$ ,  $r^2 = 0.76$ ,  $P = 0.02$ ), although this pattern is partially due to the large fraction of fixed differences in locus *2M17*, which is located very close to the breakpoint. If this locus is removed the relationship is only marginally statistically significant ( $N = 5$ ,  $r^2 = 0.74$ ,  $P = 0.06$ ). The latter pattern is not caused by the occurrence of gene conversion events at the center of the inversion and no conversion at or very near the inversion breakpoint, because a standard test (BETTRAN *et al.* 1997) shows no evidence of gene conversion in any of the loci inside the inverted region (not shown). Among the loci outside the inversion, the relationship is significant for those loci

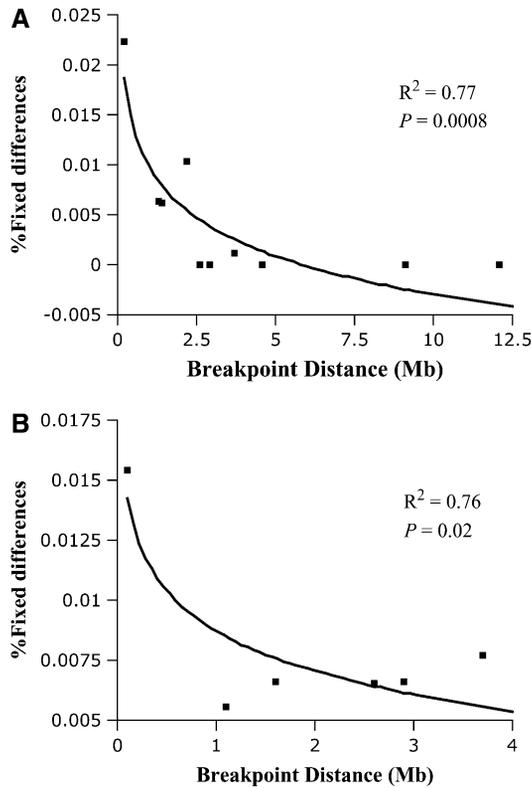


FIGURE 2.—Plot of percentage of fixed differences *vs.* distance to inversion for the group of second chromosome markers outside the inversion (A) or inside the inversion (B).

located on the centromeric side ( $N = 6$ ,  $r^2 = 0.84$ ,  $P = 0.009$ ) but not on the telomeric side ( $N = 4$ ,  $r^2 = 0.44$ ,  $P = 0.33$ ).

The regression on migration rate (estimated from  $F_{ST}$ ) on distance to breakpoint is not significant for loci outside ( $N = 10$ ,  $r^2 = 0.30$ ,  $P = 0.10$ ) or inside the inversion ( $N = 6$ ,  $r^2 = 0.04$ ,  $P = 0.70$ ) (Figure 3). The lack of significance for loci outside the inversion is due to locus *2M06*, located on the telomeric side, which has an unusually large  $Nm$  value (Table 4); without this locus the regression becomes highly significant ( $N = 9$ ,  $r^2 = 0.70$ ,  $P = 0.0048$ ). Further, the regression is significant for loci outside the inversion that are on the centromeric side ( $N = 6$ ,  $r^2 = 0.70$ ,  $P = 0.036$ ).

**RND comparisons:** We calculated FEDER *et al.*'s (2005) RND for all loci on the second chromosome studied. We specifically tested for decreasing RND with increasing physical distance from the inversion breakpoints. We predict that this relationship should be logarithmic rather than linear, as all loci readily recombining from the inverted region should introgress between the hybridizing species freely.

For the combined data set outside the inverted region of the second chromosome, there was a statistically significant log-linear relationship between RND and physical distance to the inversion breakpoint ( $N = 10$ ,  $r^2 = 0.65$ ,  $P = 0.005$ ) (Figure 4). This relationship was

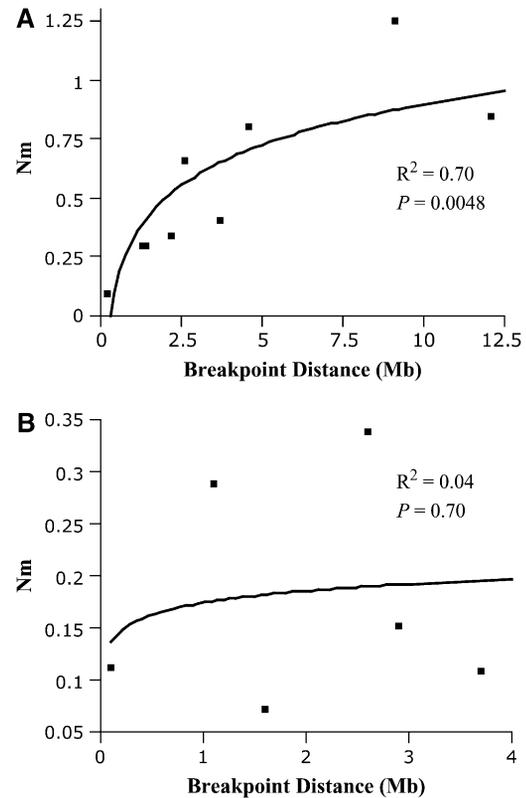


FIGURE 3.—Plot of migration rate ( $Nm$ ) *vs.* distance to inversion for the group of second chromosome markers outside the inversion (A) or inside the inversion (B). In this plot, marker *2M06* is not included (see text).

also statistically significant if only those loci on the centromeric side were considered ( $N = 6$ ,  $r^2 = 0.76$ ,  $P = 0.024$ ) and borderline significant for the loci on the telomeric side ( $N = 4$ ,  $r^2 = 0.88$ ,  $P = 0.06$ ). There was no obvious pattern for RND for the loci within the inverted region ( $N = 6$ ,  $r^2 = 0.086$ ,  $P = 0.57$ ), although the mean was nonsignificantly higher than that for the collinear regions (mean inverted = 0.90, mean collinear = 0.71, Mann-Whitney  $U$ -test,  $P = 0.22$ ).

It is noteworthy, however, that two of the three highest RND values identified were immediately flanking the inversion breakpoint (one inside, one outside, and both on the centromeric side in *D. pseudoobscura*). This is consistent with the results of several studies showing the absence of gene conversion and double crossover immediately adjacent to inversion breakpoints (reviewed in ANDOLFATTO *et al.* 2001).

**Comparison to inverted regions in other chromosomes:** We sought to briefly compare the results obtained for loci within the second chromosome to two loci just within the inverted regions distinguishing *D. pseudoobscura* and *D. persimilis* on the XL and XR chromosomes (*X\_2105* and *X\_2102*) and a locus on the third chromosome (*vg*). The *D. persimilis* XR bears both a unique arrangement (hereafter, "regular *D. persimilis*") as well as an arrangement similar to the type found in

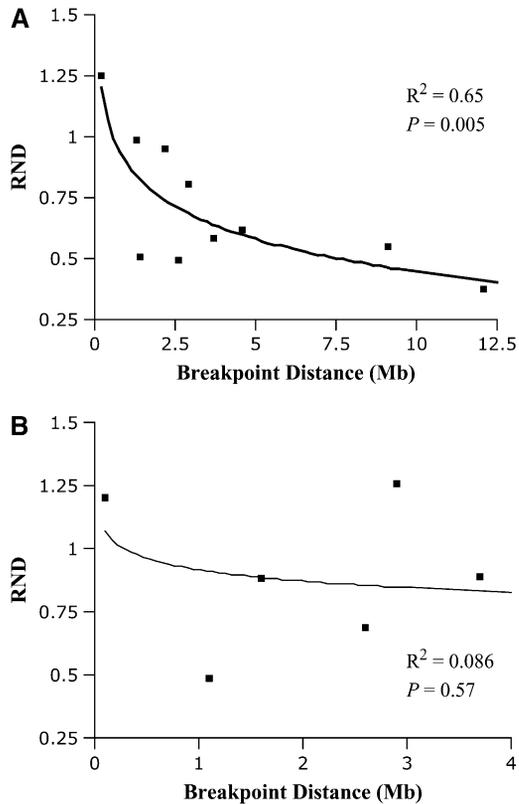


FIGURE 4.—Plot of relative node depth (RND) vs. distance to inversion for the group of second chromosome markers outside the inversion (A) or inside the inversion (B).

*D. pseudoobscura*, found in rare lines with sex-ratio distortion (hereafter, SRper). Similarly, the third chromosome bears one abundant shared arrangement (“ST”) and several arrangements unique to each species.

Table 6 shows the general polymorphism statistics for these loci. *X\_2105* and *X\_2102* show, respectively, 3 and 5 times more nucleotide variation in *D. pseudoobscura* than in regular *D. persimilis*, while the SRper arrangement in *D. persimilis* shows a 51-fold reduction in variation compared to *D. pseudoobscura* and a 10-fold reduction compared to regular *D. persimilis*. Nucleotide variation at locus *vg* in *D. p. bogotana* [Tree Line/Santa Cruz (TL/SC) arrangements] is higher than variation in the ST arrangement of *D. pseudoobscura*, agreeing with recent results showing that the TL arrangement had higher variability at this locus than the ST arrangement in *D. pseudoobscura* (SCHAEFFER and ANDERSON 2005). The neutral model could not be rejected for *vg* or *X\_2105* using different standard single-locus tests. *X\_2102* was not tested because of possible complications arising from meiotic drive affecting this chromosome arm.

We compared RND between the X chromosome loci and the loci within the second chromosome inverted region. High consistency among these RND values would suggest that the initial divergence between these

species was allopatric and that the inverted regions arose around or after the time of this initial divergence. If, on the other hand, initial divergence was sympatric then a high consistency among RND values would require the unlikely assumption that all three inversions arose at about the same time. Interestingly, we observed that RND between *D. pseudoobscura* and *D. persimilis* for the XL and XR loci (excluding SRper) was very similar to the mean observed for loci in the second chromosome inverted region [second chromosome inversion mean, 0.90; XL locus (*X\_2105*), 0.94; XR locus (*X\_2102*), 0.99].

On the XR chromosome arm, SRper and regular *D. persimilis* were very distinct from each other, with an average pairwise difference of 0.0784. However, SRper was also somewhat distinct from *D. pseudoobscura*, to which it is collinear, with an average pairwise difference of 0.0623. Phylogenetic analysis places at least one strain of *D. pseudoobscura* in a weakly supported cluster with SRper (Figure 5), suggesting an old introgression event or possibly segregation of shared ancestral polymorphism. Similarly, BETRAN *et al.*'s (1997) method for detecting gene conversion identified two putative exchanges between SRper and strains of *D. pseudoobscura*, but none between regular *D. persimilis* and *D. pseudoobscura*.

If *D. pseudoobscura* and *D. persimilis* have experienced recent gene exchange, we also predicted that the third chromosome ST arrangement of these species should have an RND estimate similar to that observed in the collinear regions of the second chromosome far (>1.5 Mb) from the inversion breakpoints. This predicted pattern was observed (RND at *vg* = 0.668, mean RND in collinear region >1.5 Mb from second chromosome breakpoint = 0.624). However, there were 12 purportedly fixed differences between the two species at *vg*, in contrast to none or very few in the second chromosome collinear regions far from the inversion breakpoint. These apparently fixed differences very likely reflect our poor sampling of *D. persimilis* (two strains), since we were limited to our only two cultured strains bearing the ST gene arrangement. Greater sampling may demonstrate that some or all of the apparently *D. pseudoobscura*-specific bases are present in *D. persimilis*.

## DISCUSSION

We identified the approximate location of the breakpoints of the fixed chromosome inverted regions that differentiate *D. pseudoobscura* and *D. persimilis*. On the basis of the assembled sequence of the second chromosome of *D. pseudoobscura*, the size of the inverted region is 7.6 Mb, thus encompassing ~25% of the length of this chromosome. The length of the XL and XR inverted regions cannot be estimated with the same precision because the genome assembly of those chromosome arms is not as complete as that of the second chromosome. Information on the precise location of the second

TABLE 6

Polymorphism statistics of loci from the XR and XL breakpoints and the ST arrangement (third chromosome)

Locus	Species	$n^a$	$L^b$	$S^c$	Syn <sup>d</sup>	Rep <sup>d</sup>	$\hat{\theta}^e$	$\pi^f$	$D^g$	$Z_{nS}^h$	Diff <sup>i</sup>
Vg (3)	ST <i>pseudoobscura</i>	15	512.5	28	1	0	0.01680	0.01752	0.1800	0.165	0.05039
	SC/TL <i>bogotana</i> <sup>j</sup>	4	431.0	16	0	0	0.02025	0.01856	-0.8490	1.0	—
	ST <i>persimilis</i>	2	518.0	1	0	0	0.00193	0.00193	—	—	0.06687
	<i>miranda</i>	1	564.0	—	—	—	—	—	—	—	—
X_2105 (XL)	<i>pseudoobscura</i>	6	600.8	28	—	—	0.02041	0.01809	-0.7198	0.209	0.03235
	<i>bogotana</i>	4	604.0	2	—	—	0.00181	0.00221	1.8931	1.0	0.03483
	<i>persimilis</i>	6	605.0	11	—	—	0.00796	0.00639	-1.1937	0.283	0.02810
	<i>miranda</i>	1	608.0	—	—	—	—	—	—	—	—
X_2102 (XR)	<i>pseudoobscura</i>	10	541.4	113	—	—	0.07378	0.05189	— <sup>k</sup>	0.120	0.07407
	<i>bogotana</i>	5	664.0	2	—	—	0.00145	0.00181	— <sup>k</sup>	0.444	0.06248
	<i>persimilis</i>	5	619.6	18	—	—	0.01394	0.01243	— <sup>k</sup>	0.215	0.08092
	SR <i>persimilis</i>	6	608.3	2	—	—	0.00144	0.00142	— <sup>k</sup>	0.400	0.07955
	<i>miranda</i>	1	900.0	—	—	—	—	—	—	—	—

—, values could not be obtained for small samples or for groups of sequences with few informative sites.

<sup>a</sup> Number of lines sequenced.

<sup>b</sup> Average length (base pairs) of the sequences from each species.

<sup>c</sup> Number of polymorphic sites.

<sup>d</sup> Number of synonymous (syn) and replacement (rep) polymorphisms in the coding regions.

<sup>e</sup> Estimate of  $4N\mu$ /bp using the number of polymorphic sites (WATTERSON 1975).

<sup>f</sup> Estimate of  $4N\mu$  using the average number of nucleotide differences per site (TAJIMA 1983).

<sup>g</sup> TAJIMA's (1989) statistic. Significance was determined using 10,000 coalescent simulations with no recombination.

<sup>h</sup> Average intralocus linkage disequilibrium (KELLY 1997). Significance was determined using 1000 coalescent simulations with no recombination.

<sup>i</sup> Average pairwise difference per base pair between alleles from each taxon and the alleles of *D. miranda*.

<sup>j</sup> The *D. p. bogotana* data presented here should not be used for comparison, as they do not bear the ST arrangement.

<sup>k</sup> Tajima's *D* not included because of possible complications arising from meiotic drive affecting this chromosome arm in *D. persimilis*.

chromosome inversion was used to guide collection of new population genetic data from 13 loci across this chromosome that were combined with 5 additional loci from a previous study (MACHADO *et al.* 2002) to explore the effect that distance to the nearest inversion breakpoint has had on the amount of inferred interspecific gene flow and divergence between these two species. Here, we conducted an empirical test of theoretical predictions of a model originally proposed by NAVARRO *et al.* (1997) on the effect of inversions on recombination and gene flux in inversion heterokaryotypes.

**The effect of distance to inversion breakpoint on divergence and introgression:** Distance to the nearest inversion breakpoint shows a significant effect on several measures of divergence and shared polymorphism between *D. pseudoobscura* and *D. persimilis* (Figures 2A and 4A). Percentage of fixed differences per locus, migration rate, and RND change significantly with distance for loci located outside the inverted region, corroborating the predictions of NAVARRO *et al.* (1997). The data fit well a logarithmic relationship, showing that the suppression of crossovers in inversion heterozygotes

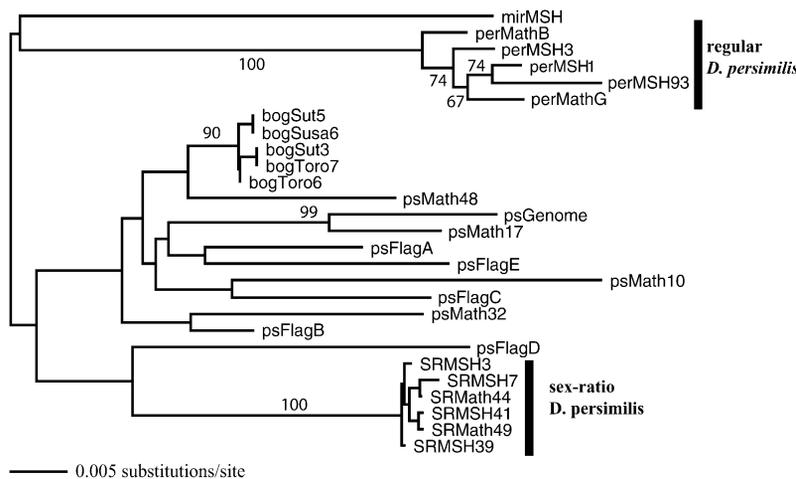


FIGURE 5.—NJ tree of *X\_2102* (chromosome XR) showing the genealogy of the regular and sex-ratio *D. persimilis* sequences. Individuals with the prefix “ps” are *D. pseudoobscura*, and those with prefix “bog” are *D. p. bogotana*. Numbers associated with branches represent bootstrap values >65% based on 500 pseudoreplications. Mid-point rooting was used to place the outgroup.

also extends to loci located outside the inversion but close to it (within 1–2 Mb). Although gene flow between *D. pseudoobscura* and *D. persimilis* may have stopped some time ago at nuclear loci (MACHADO *et al.* 2002), there is still a strong signal of significant increases in several measures of shared variation and gene flow that indicate that loci located far from the inverted region are not affected by the suppression of recombination generated in interspecies inversion heterozygotes. We find that the effects are significant for loci outside the inversion that are located in the centromeric side but not in the telomeric side of the chromosome. There are at least two explanations for this contrast. First, we sampled more loci on the centromeric side, which is also bigger than the telomeric side (13 Mb *vs.* 9 Mb). Although we sampled six loci in the centromeric side and five loci in the telomeric side, we analyzed only four loci in the latter due to evidence of selection at locus *2M05* in *D. persimilis*. Second, two of the loci on the telomeric side had unexpected patterns of variation. Locus *2M06* has an unusually large *N<sub>m</sub>* value on the basis of very low *F<sub>ST</sub>*, and locus *2M05* has a large number of fixed differences (18) and shows evidence of reduction in variation due to positive selection in *D. persimilis*. That pattern suggests linkage of *2M05* to a locus or suite of loci with alternative alleles favored in each species that would therefore reduce or eliminate interspecific gene flow at that chromosomal location. Determining whether we have in fact found a “speciation island” (TURNER *et al.* 2005) surrounding locus *2M05*, which is thus not located inside one of the fixed inversions, will require additional sampling of loci close to *2M05*.

The NAVARRO *et al.* (1997) model also predicts that gene flow can occur between inverted regions through double crossovers or gene conversion. Gene flow is expected to be more frequent further from the inversion breakpoints and toward the center of the inversion especially in large inversions. This effect has been detected in previous studies in *Drosophila* species that have inversion polymorphisms (LAAYOUNI *et al.* 2003; SCHAEFFER and ANDERSON 2005). LAAYOUNI and HASSON (2003) found evidence of gene flow in the middle of the second chromosome inversion of *D. buzzatti* among different inverted arrangements. SCHAEFFER and ANDERSON (2005) also found evidence of intraspecific gene flow in the middle of the polymorphic third chromosome inversion of *D. pseudoobscura* among five different inverted arrangements. Here, we observed some patterns of variation suggesting evidence of interspecific gene flow inside the second chromosome inversion (large fraction of fixed differences and high RND very close to the breakpoint), although the tested effects of distance to inversion breakpoint for loci inside the inversion were mostly nonsignificant (Figures 2B and 4B) and we did not see any evidence of gene conversion. This difference in the strength of the effects observed in intraspecific *vs.* interspecific comparisons

likely results from the very low frequency of interspecific inversion heterozygotes (hybrids) relative to intraspecific inversion heterozygotes, hence greatly reducing the opportunity for gene flux between inversions. Alternatively, it could result from hybrid incompatibility genetic factors being evenly spread across the second chromosome inversion, which effectively decreases the possibility of gene flow between *D. pseudoobscura* and *D. persimilis* in the middle of the second chromosome inverted region.

**The origin of the fixed inverted regions:** There is a statistically significant, twofold reduction in nucleotide variation inside the inverted region ( $\hat{\theta} = 0.0048$ ) of *D. persimilis* compared to the rest of the autosomal loci sampled ( $\hat{\theta} = 0.0096$ ). This reduction in variation is not observed in *D. pseudoobscura* and is thus not due to mutation constraints in the sampled loci. This significant reduction in polymorphism is consistent with a scenario in which the *D. persimilis* arrangement was quickly fixed in this species, possibly by natural selection (KIRKPATRICK and BARTON 2006). DOBZHANSKY and TAN (1936) observed that *D. persimilis* (*D. pseudoobscura* race B, in their article) has a derived arrangement in the second chromosome because the arrangement of the outgroup *D. miranda* is closer to that of *D. pseudoobscura*. That observation suggests a scenario in which the second chromosome inversion arose and was fixed in *D. persimilis*, thus being consistent with our inference of rapid fixation and consequent loss of variation in the inverted region. Despite the inferred rapid fixation of the inverted region of *D. persimilis*, there is no evidence of a recent selective sweep in that region using any standard test of neutrality either for individual loci or for the group of all loci inside the inverted region. Further, average linkage disequilibrium inside the inversion, measured using the  $Z_{nS}$ -statistic (KELLY 1997), is not significantly different than linkage outside the inversion ( $Z = 0.05$ ,  $P = 0.96$ ), indicating that recombination has been able to shuffle variation inside the inverted region in *D. persimilis*. Therefore, if selection was involved in the fixation of this inversion in *D. persimilis*, the selective sweep was not recent (at least not within the last  $0.2N_c$  generations).

To further explore the hypothesis of a selection-driven fixation of the inversion in *D. persimilis*, we polarized the fixed differences in all loci inside the inversion using *D. miranda* as the outgroup. This choice of outgroup is problematic for some genes inside the inversion (*2M14*, *2M17*) because of the allele-sorting problem presented in Figure 6, where *D. miranda* alleles appear closer to *D. persimilis* alleles in some genes. In those cases it will be incorrect to polarize the changes using *D. miranda*, and we decided not to do this for loci *2M14* and *2M17*. For all the remaining loci we found more fixations in the lineage leading to *D. persimilis*: *2M12* (*pseudoobscura*, 0 fixations/*persimilis*, 5 fixations), *2M13* (3/4), *2M15* (1/5), and *2M16* (2/3). The results

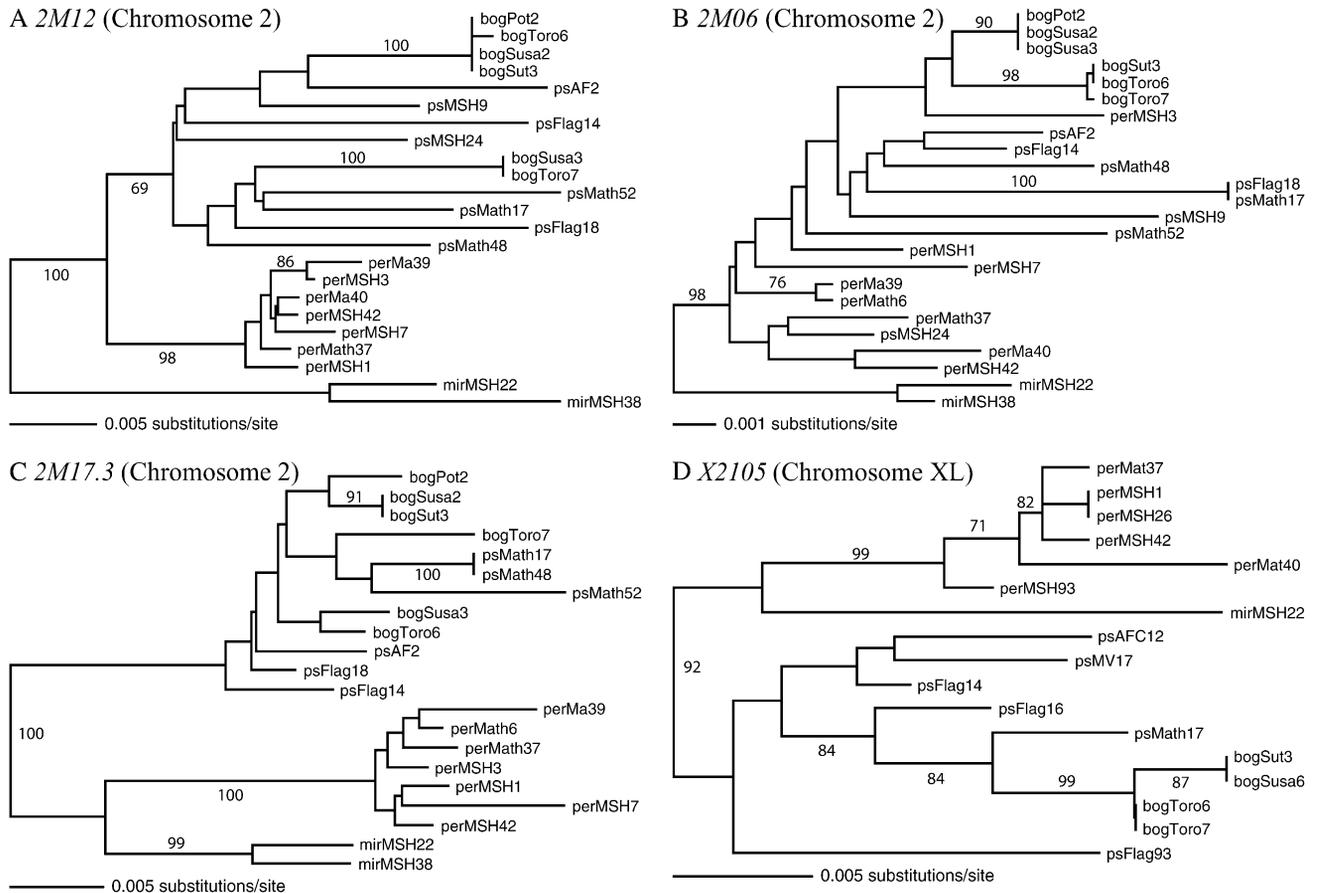


FIGURE 6.—NJ trees of 3 second chromosome loci (A–C) and one XL locus (D) showing three different patterns of phylogenetic association. Individuals with the prefix “ps” are *D. pseudoobscura*, those with prefix “per” are *D. persimilis*, those with prefix “bog” are *D. p. bogotana*, and those with prefix “mir” are *D. miranda*. Numbers associated with branches represent bootstrap values >65% based on 500 pseudoreplications. (A) *2M12*; (B) *2M06*; (C) *2M17.3*; (D) *X\_2105*. Midpoint rooting was used to place the outgroup in all trees.

are striking in *2M12* and *2M15*, with an overwhelming excess of *D. persimilis* fixations. Although the observed excess of *D. persimilis* fixations is not significant ( $Z = -5$ ;  $P = 0.062$ ), the results are consistent with our proposed scenario of a rapid fixation of the inverted arrangement in *D. persimilis*. Some of the fixations are possibly the result of captured rare mutations that were segregating in the ancestral *D. persimilis*, while others could have just accumulated faster in the *D. persimilis* lineage due to the smaller effective size of the inverted region soon after its fixation.

We also observed an 11-fold reduction of variation in the distal breakpoint region of the second chromosome in *D. persimilis* compared to *D. pseudoobscura*. This reduction is expected under a scenario in which the inverted region was fixed through a selective sweep. NAVARRO *et al.* (1997, 2000) showed that regions close to the breakpoints of recently arisen inversions will be the most affected by the selective sweep as the inversion increases in frequency or becomes fixed. Therefore, nucleotide variation in the breakpoint region is expected to be very low, compared to variation in the original ar-

range, and linkage disequilibrium is expected to be high and be maintained long after the derived inversion reaches its equilibrium frequency (NAVARRO *et al.* 2000). These effects are the consequence of a high reduction in gene flow at the breakpoints due to stronger recombination suppression there than in regions toward the center of the inversion. For instance, DEPAULIS *et al.* (1999, 2000) have shown that hitchhiking effects are enhanced by proximity to inversions in *D. melanogaster*. Our results are consistent with the predictions of NAVARRO *et al.* (1997, 2000) and are similar, although less striking, to those made by HASSON and EANES (1996), who documented a 20-fold reduction in the level of nucleotide variation in the derived breakpoint with respect to the standard breakpoint region in *In(3L)Payne* of *D. melanogaster*.

All of the evidence presented here suggests a scenario in which the inverted second chromosome arrangement arose and was quickly fixed in *D. persimilis*, possibly through natural selection. The fixation event likely occurred sufficiently long ago, thus leaving few detectable effects on current patterns of polymorphism besides the

significant reduction in variation at the sampled loci. Besides the twofold significant reduction in variation inside the inversion, none of the observed patterns of variation are inconsistent with neutrality. The fixation of the inverted arrangement could have occurred at about the time of species divergence, and, since then, variation has been recovered inside the inverted region of *D. persimilis*, although variation in the inversion is still well below levels typical of the rest of the autosomal regions in this species. The lower variability inside the inversion could result from little time ( $<2N_e$  generations) having passed since the fixation event for this region to reach mutation–drift equilibrium (NEI and LI 1976; PERLITZ and STEPHAN 1997). Alternatively, the low variability could also be caused by the lack of introgression and low influx of shared variation in the inverted region from the more polymorphic *D. pseudoobscura* into *D. persimilis*. In other words, variation inside this inversion may reflect the “true” level of variability in *D. persimilis*, while that in collinear regions of the genome may be inflated by exchange with *D. pseudoobscura*. On the basis of these results, we predict that average polymorphism levels inside the XL and XR inversions will also be significantly lower than those in the collinear regions of the genome, assuming that those inversions swept to fixation at about the same time as the second chromosome inversion. Lower variation in the regions close to the XL and XR breakpoints also suggests that those *D. persimilis* arrangements may have been fixed by selection, although this conclusion is based only on single loci and needs to be substantiated with a larger sample.

#### **The divergence of *D. pseudoobscura* and *D. persimilis*:**

The high (near unity) RND estimates both for the three inverted regions and for collinear regions very near the inversion breakpoint suggest that the divergence of *D. pseudoobscura* and *D. persimilis* began around the same time as the split of these species from *D. miranda*. The high level of divergence between SRper and regular *D. persimilis* suggests that the derived SRper arrangement (shared with *D. pseudoobscura*) arose perhaps slightly prior to, but close in time to, the split of *D. pseudoobscura* and *D. persimilis*. However, we emphasize that there are large variances in these estimates, and only single loci were sampled within the XL and XR inverted regions, so these conclusions are necessarily tentative. There is, however, additional evidence based on phylogenetic reconstructions that also suggests such a scenario for the divergence of this group.

Figure 6 shows phylogenies of four markers that represent the three typical types of topologies observed for the data (see also MACHADO and HEY 2003). Figure 6A (2M12) represents the typical topology of regions close to or inside inverted regions, in which *D. persimilis* sequences are well differentiated from *D. pseudoobscura*. Figure 6B (2M06) represents the typical topology observed in collinear regions of the genome, in which sequences from *D. persimilis* and *D. pseudoobscura* strains

are not reciprocally monophyletic and well mixed, suggesting the occurrence of introgression during their divergence. More interestingly, although most genes show a clear differentiation of *D. miranda* sequences as a clear outgroup, there are several loci in which *D. persimilis* sequences show high affinity and bootstrap support with *D. miranda* sequences (2M17.3, X\_2105: Figure 6, C and D). We had not observed this phylogenetic pattern in our previous study (MACHADO and HEY 2003), possibly because no sequences surveyed were within the inverted region, but here we have observed it in markers located close to or inside the second chromosome inverted region (2M13, 2M14, 2M17, 2M17.3) and the XL inverted region (X\_2105), as well as in 2M05, the locus in the telomeric collinear region of the second chromosome that has a large number of fixed differences between the two species and that shows evidence of selection in *D. persimilis*. This is an odd pattern given that, in the most accepted scenario for the divergence of this species group, the ancestor of *D. persimilis* and *D. pseudoobscura* split from the ancestor of *D. miranda* ~2 MYA (WANG *et al.* 1997; MACHADO and HEY 2003).

Our observations, and especially the fact that this phylogenetic pattern is associated with markers in and adjacent to the inverted region, suggest an alternative scenario for the divergence of this species group. In this scenario, the ancestral species separated into three populations (A, B, and C) fairly close in time. Subsequently, an inversion (or a group of inversions) arose in B, and, sometime later, populations A and B started to exchange some genes in the collinear regions not linked to the inversion(s). Under this scenario, RND (or any phylogenetic clustering) would put collinear regions between A and B tightly with each other and C as the outgroup. In contrast, phylogenetic reconstructions would not easily resolve the phylogenetic locations of the three taxa A, B, and C in inverted regions or in regions tightly linked to the inversion(s) or would make inconsistent pairs on the basis of lineage sorting prior to the separation event. If A is *D. pseudoobscura*, B is *D. persimilis*, and C is *D. miranda*, this hypothetical scenario explains the bulk of our observations.

We thank Richard Kliman for his insightful comments and two anonymous reviewers for their suggestions. Sarah Cavanaugh helped with the data collection. T.S.H. was supported by a University of Arizona—National Science Foundation (NSF) Integrative Graduate Education Research Training grant Genomics Initiative (DGE-0114420). Research was supported by grants from the NSF to C.A.M. (DEB-0520535) and to M.A.F.N. (DEB-0509780 and DEB-0549893).

#### LITERATURE CITED

- ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.
- ANDOLFATTO, P., F. DEPAULIS and A. NAVARRO, 2001 Inversion polymorphisms and nucleotide variability in *Drosophila*. *Genet. Res.* **77**: 1–8.
- BABCOCK, C. S., and W. W. ANDERSON, 1996 Molecular evolution of the sex-ratio inversion complex in *Drosophila pseudoobscura*: analysis of the *Esterase-5* gene region. *Mol. Biol. Evol.* **13**: 297–308.

- BAUDRY, E., and F. DEPAULIS, 2003 Effect of misoriented sites on neutrality tests with outgroup. *Genetics* **165**: 1619–1622.
- BETRAN, E., J. ROZAS, A. NAVARRO and A. BARBADILLA, 1997 The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. *Genetics* **146**: 89–99.
- BROWN, K. M., L. M. BURK, L. M. HENAGAN and M. A. F. NOOR, 2004 A test of the chromosomal rearrangement model of speciation in *Drosophila pseudoobscura*. *Evolution* **58**: 1856–1860.
- BUTLIN, R. K., 2005 Recombination and speciation. *Mol. Ecol.* **14**: 2621–2635.
- CLARK, A. G., 1997 Neutral behavior of shared polymorphism. *Proc. Natl. Acad. Sci. USA* **94**: 7730–7734.
- DEPAULIS, F., L. BRAZIER and M. VEUILLE, 1999 Selective sweep at the *Drosophila melanogaster* *Suppressor of Hairless* locus and its association with the *In(2L)t* inversion polymorphism. *Genetics* **152**: 1017–1024.
- DEPAULIS, F., L. BRAZIER, S. MOUSSET, A. TURBE and M. VEUILLE, 2000 Selective sweep near the *In(2L)t* inversion breakpoint in an African population of *Drosophila melanogaster*. *Genet. Res.* **76**: 149–158.
- DOBZHANSKY, T., and C. C. TAN, 1936 Studies on hybrid sterility III. A comparison of the gene arrangement in two species, *Drosophila pseudoobscura* and *Drosophila miranda*. *Z. Indukt. Abstamm. Vererbungslehre* **72**: 88–114.
- FAY, J. C., and C.-I. WU, 2000 Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.
- FEDER, J. L., J. B. ROETHELE, K. FILCHAK, J. NIEBALSKE and J. ROMERO-SEVERSON, 2003 Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics* **163**: 939–953.
- FEDER, J. L., X. XIE, J. RÜLL, S. VELEZ, A. FORBES *et al.*, 2005 Mayt, Dobzhansky, and Bush and the complexities of sympatric speciation in *Rhagoletis*. *Proc. Natl. Acad. Sci. USA* **102**: 6573–6580.
- FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- GILBERT, D. G., 2005 DroSpeGe, a public database of *Drosophila* species genomes (<http://insects.eugenes.org/DroSpeGe/>).
- GLOOR, G. B., and W. R. ENGELS, 1992 Single-fly DNA preps for PCR. *Dros. Inf. Serv.* **71**: 148–149.
- HAMBLIN, M. T., and C. F. AQUADRO, 1999 DNA sequence variation and the recombinational landscape in *Drosophila pseudoobscura*: a study of the second chromosome. *Genetics* **153**: 859–869.
- HASSON, E., and W. F. EANES, 1996 Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. *Genetics* **144**: 1565–1575.
- HEY, J., and R. NIELSEN, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.
- HEY, J., and J. WAKELEY, 1997 A coalescent estimator of the population recombination rate. *Genetics* **145**: 833–846.
- HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–159.
- HUDSON, R. R., M. SLATKIN and W. P. MADDISON, 1992 Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.
- JONES, J. S., S. H. BRYANT, R. C. LEWONTIN, J. A. MOORE and T. PROUT, 1981 Gene flow and the geographical distribution of a molecular polymorphism in *Drosophila pseudoobscura*. *Genetics* **98**: 157–178.
- KELLY, J. K., 1997 A test of neutrality based on interlocus associations. *Genetics* **146**: 1197–1206.
- KIRKPATRICK, M., and N. BARTON, 2006 Chromosome inversions, local adaptation and speciation. *Genetics* **173**: 419–434.
- 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.
- LAAYOUNI, H., E. HASSON, M. SANTOS and A. FONTDEVILA, 2003 The evolutionary history of *Drosophila buzzatii*. XXXV. Inversion polymorphism and nucleotide variability in different regions of the second chromosome. *Mol. Biol. Evol.* **20**: 931–944.
- MACHADO, C. A., and J. HEY, 2003 The causes of phylogenetic conflict in a classic *Drosophila* species group. *Proc. R. Soc. Lond. Ser. B* **270**: 1193–1202.
- MACHADO, C. A., R. M. KLIMAN, J. A. MARKERT and J. HEY, 2002 Inferring the history of speciation using multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Mol. Biol. Evol.* **19**: 472–488.
- MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–654.
- MERRELL, D. J., 1954 Sexual isolation between *Drosophila persimilis* and *Drosophila pseudoobscura*. *Am. Nat.* **88**: 93–99.
- NAVARRO, A., E. BETRAN, A. BARBADILLA and A. RUIZ, 1997 Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics* **146**: 695–709.
- NAVARRO, A., A. BARBADILLA and A. RUIZ, 2000 Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in *Drosophila*. *Genetics* **155**: 685–698.
- NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- NEI, M., and W. H. LI, 1976 The transient distribution of allele frequencies under mutation pressure. *Genet. Res.* **28**: 205–214.
- NIELSEN, R., and J. WAKELEY, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**: 885–896.
- NOOR, M. A., 1995 Speciation driven by natural selection in *Drosophila*. *Nature* **375**: 674–675.
- NOOR, M. A., 1996 Absence of species discrimination in *Drosophila pseudoobscura* and *D. persimilis* males. *Anim. Behav.* **52**: 1205–1210.
- NOOR, M. A., K. L. GRAMS, L. A. BERTUCCI and J. REILAND, 2001a Chromosomal inversions and the reproductive isolation of species. *Proc. Natl. Acad. Sci. USA* **98**: 12084–12088.
- NOOR, M. A. F., K. L. GRAMS, L. A. BERTUCCI, Y. ALMENDAREZ, J. REILAND *et al.*, 2001b The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. *Evolution* **55**: 512–521.
- ORTIZ-BARRIENTOS, D., J. REILAND, J. HEY and M. A. NOOR, 2002 Recombination and the divergence of hybridizing species. *Genetica* **116**: 167–178.
- ORTÍZ-BARRIENTOS, D., A. S. CHANG and M. A. F. NOOR, 2006 A recombinational portrait of the *Drosophila pseudoobscura* genome. *Genet. Res.* **87**: 23–31.
- O'TOUSA, J. E., W. BAEHR, R. L. MARTIN, J. GIRSH, W. L. PAK *et al.*, 1985 The *Drosophila ninaE* gene encodes an opsin. *Cell* **40**: 839–850.
- PANITHANARAK, T., H. C. HAUFFE, J. F. DALLAS, A. GLOVER, R. G. WARD *et al.*, 2004 Linkage-dependent gene flow in a house mouse chromosomal hybrid zone. *Evolution* **58**: 184–192.
- PATTERSON, N., D. J. RICHTER, S. GNERRE, E. S. LANDER and D. REICH, 2006 Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**: 1103–1108.
- PERLITZ, M., and W. STEPHAN, 1997 The mean and variance of the number of segregating sites since the last hitchhiking event. *J. Math. Biol.* **36**: 1–23.
- PRAKASH, S., R. C. LEWONTIN and J. L. HUBBY, 1969 A molecular approach to the study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal and isolated populations of *Drosophila pseudoobscura*. *Genetics* **61**: 841–858.
- RICHARDS, S., Y. LIU, B. R. BETTENCOURT, P. HRADECKY, S. LETOVSKY *et al.*, 2005 Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene and cis-element evolution. *Genome Res.* **15**: 1–18.
- RIESEBERG, L. H., M. A. ARCHER and R. K. WAYNE, 1999 Transgressive segregation, adaptation and speciation. *Heredity* **83**: 363–372.
- RILEY, M. A., S. R. KAPLAN and M. VEUILLE, 1992 Nucleotide polymorphism at the xanthine dehydrogenase locus in *Drosophila pseudoobscura*. *Mol. Biol. Evol.* **9**: 56–69.
- ROZAS, J., and R. ROZAS, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analyses. *Bioinformatics* **15**: 174–175.
- ROZEN, D. E., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers, pp. 365–386 in

- Bioinformatics Methods and Protocols: Methods in Molecular Biology*, edited by S. KRAWETZ and S. MISENER. Humana Press, Totowa, NJ.
- SCHAEFFER, S. W., and W. W. ANDERSON, 2005 Mechanisms of genetic exchange within the chromosomal inversions of *Drosophila pseudoobscura*. *Genetics* **171**: 1729–1739.
- SCHAEFFER, S. W., and E. L. MILLER, 1992a Molecular population genetics of an electrophoretically monomorphic protein in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* **132**: 163–178.
- SCHAEFFER, S. W., and E. L. MILLER, 1992b Estimates of gene flow in *Drosophila pseudoobscura* determined from nucleotide sequence analysis of the alcohol dehydrogenase region. *Genetics* **132**: 471–480.
- SCHULTZ, J., and H. REDFIELD, 1951 Interchromosomal effects on crossing over in *Drosophila*. Cold Spring Harbor Symp. Quant. Biol. **16**: 175–197.
- STUMP, A. D., M. C. FITZPATRICK, N. F. LOBO, S. TRAORE, N. SAGNON *et al.*, 2005a Centromere-proximal differentiation and speciation in *Anopheles gambiae*. *Proc. Natl. Acad. Sci. USA* **102**: 15930–15935.
- STUMP, A. D., J. A. SHOENER, C. COSTANTINI, N. SAGNON and N. J. BESANSKY, 2005b Sex-linked differentiation between incipient species of *Anopheles gambiae*. *Genetics* **169**: 1509–1519.
- SWOFFORD, D. L., 1998 *PAUP\**. *Phylogenetic Analysis Using Parsimony (\*and Other Methods)*. Sinauer Associates, Sunderland, MA.
- TAJIMA, F., 1983 Evolutionary relationships of DNA sequences in finite populations. *Genetics* **105**: 437–460.
- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**: 512–526.
- TURNER, T. L., M. W. HAHN and S. V. NUZHIDIN, 2005 Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* **3**: e285.
- WAKELEY, J., and J. HEY, 1997 Estimating ancestral population parameters. *Genetics* **145**: 847–855.
- WANG, R. L., J. WAKELEY and J. HEY, 1997 Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* **147**: 1091–1106.
- WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**: 256–276.
- WELLS, R. S., 1996 Nucleotide variation at the Gpdh locus in the genus *Drosophila*. *Genetics* **143**: 375–384.

Communicating editor: S. W. SCHAEFFER