Gene Flow and Natural Selection in the Origin of *Drosophila pseudoobscura* and Close Relatives

Rong Lin Wang,[1] John Wakeley and Jody Hey

Biological Sciences, Rutgers University, Nelson Labs, Piscataway, New Jersey 08855-1059

Manuscript received April 2, 1997
Accepted for publication July 2, 1997

ABSTRACT

The divergence of *Drosophila pseudoobscura* and close relatives *D. persimilis* and *D. pseudoobscura bogotana* has been studied using comparative DNA sequence data from multiple nuclear loci. New data from the *Hsp82* and *Adh* regions, in conjunction with existing data from *Adh* and the *Period* locus, are examined in the light of various models of speciation. The principal finding is that the three loci present very different histories, with *Adh* indicating large amounts of recent gene flow among the taxa, while little or no gene flow is apparent in the data from the other loci. The data were compared with predictions from several isolation models of divergence. These models include no gene flow, and they were found to be incompatible with the data. Instead the DNA data, taken together with other evidence, seem consistent with divergence models in which natural selection acts against gene flow at some loci more than at others. This family of models includes some sympatric and parapatric speciation models, as well as models of secondary contact and subsequent reinforcement of sexual isolation.

DROSOPHILA pseudoobscura and close relatives *D. persimilis, D. miranda*, and subspecies *D. pseudoobscura bogotana* may provide an opportunity to study species divergence in the presence of gene flow between species. With the exception of *D. p. bogotana*, which is restricted to regions near Bogota, Colombia (DOBZHANSKY *et al.* 1963), these species occupy large and partially sympatric ranges in western North America. In the laboratory, reproductive isolation between *D. miranda* and its sibling species is complete (DOBZHANSKY and EPLING 1944), but fertile hybrids are formed in crosses between *D. pseudoobscura* and *D. p. bogotana* (PRAKASH 1972; ORR 1989a), between *D. pseudoobscura* and *D. persimilis* (DOBZHANSKY and EPLING 1944), as well as between *D. persimilis* and *D. p. bogotana* (H. A. ORR, personal communication).

Whether gene flow occurs among these taxa was a question of long standing interest to DOBZHANSKY (DOBZHANSKY and EPLING 1944; DOBZHANSKY 1973). DOBZHANSKY and colleagues did, in fact, find direct evidence of gene flow: a total of three backcross hybrids collected from nature, although this took many years and over 30,000 chromosomal preparations (DOBZHANSKY 1973; POWELL 1983). Other attempts to address questions of gene flow have relied on patterns of shared genetic variation. An apparent absence of divergence for mitochondrial DNA between *D. pseudoobscura* and *D. persimilis* that were collected from regions of sympatry was regarded as evidence of gene flow (POWELL 1983). However, a similar study concluded that the species do not share variation and that there was no evidence of mitochondrial gene flow (HALE and BECKENBACH 1985). The wealth of allozyme data on these species is also difficult to interpret in terms of gene flow (PRAKASH 1972; AYALA and DOBZHANSKY 1974; SINGH 1983), since the presence of shared alleles can be due to gene flow or to the persistence of alleles since the time of common ancestry. A recent study of DNA sequence variation at the *X*-linked *Period* locus found evidence for very limited gene flow between *D. pseudoobscura* and *D. persimilis* (WANG and HEY 1996).

In this study we extend the nuclear gene comparative DNA approach to include two more loci. We report new results for a heatshock locus *Hsp82* and from the Alcohol dehydrogenase (*Adh*) region that has already been studied extensively within *D. pseudoobscura* and *D. p. bogotana*.

Hsp82 encodes a heatshock protein that is highly conserved among Drosophila species at the amino acid level (BLACKMAN and MESELSON 1986). It is located within a puff of chromosome region 23, on the right arm of the *X* chromosome, of *D. pseudoobscura* (BLACKMAN and MESELSON 1986; SEGARRA *et al.* 1996). We sequenced a region of ~2000 base pairs (bp), much of it from the large intron.

The *Adh* region lies on chromosome *4*, an autosome (SCHAEFFER and AQUADRO 1987), and includes both *Adh* and *Adh-Dup*, a fairly old and divergent duplication of *Adh* (SCHAEFFER and AQUADRO 1987). In a series of papers, SCHAEFFER and MILLER (1991, 1992a,b, 1993) have described the pattern of variation within *D. pseudoobscura* for a span of >3500 bp. They have also studied the divergence between *D. pseudoobscura* and *D. p. bogo-*

Corresponding author: Jody Hey, Biological Sciences, Rutgers University, Nelson Hall, P.O. Box 1059, Piscataway, NJ 08855-1059.
E-mail: hey@mbcl.rutgers.edu www: http://heylab.rutgers.edu/

[1] *Present address:* Department of Plant Biology, University of Minnesota, 1445 Gortner Ave., St. Paul, MN 55108-1020.

tana and sequenced one copy from each of *D. persimilis* and *D. miranda* in this same region (SCHAEFFER and MILLER 1991). We have sequenced five additional lines of *D. persimilis* for this same region.

MATERIALS AND METHODS

***Hsp82* sequencing:** The fly samples are identical to those used for the *Period* locus study (see Table 1 of WANG and HEY 1996). DNA from individual male flies was extracted according to protocol 48 of ASHBURNER (1989). From each sample of genomic DNA, a section of the *Hsp82* locus (between positions −11 and 2279 of BLACKMAN and MESELSON 1986) was PCR amplified. Additional DNA preparation and sequencing followed the protocol used by KLIMAN and HEY (1993). Both strands were sequenced for each strain. A total of 10 20-bp-long sequencing primers, spaced ~200 bp apart, were used on each strand. The final length of the sequenced portion was ~2 kilobases (kb), covering exon I (not translated), the only intron, and exon II and spanning positions 873–2872 of BLACKMAN and MESELSON (1986) inclusive. The sequences have been submitted to GenBank (accession numbers AF006529–AF006563).

***Adh* sequencing:** *D. persimilis* lines 40, 42, 44, 49, and 50 were used (see Table 1 of WANG and HEY 1996). All of these lines were originally from the National Drosophila Species Resource Center (NDSRC, Bowling Green, OH), and they represent a geographically diverse sample. To avoid sequencing heterozygous DNA samples, the lines were first inbred via full sib-mating for 10 or 11 generations. Genomic DNA was prepared from individual flies using protocol 48 of ASHBURNER (1989). STEVE SCHAEFFER kindly provided the PCR and DNA sequencing primers that he designed and used for the generation of the large *D. pseudoobscura* and *D. p. bogotana* data sets (SCHAEFFER and MILLER 1991, 1992a). With these primers, the five *D. persimilis* lines were sequenced for the same 3.5 kb as had previously been done in *D. pseudoobscura* and *D. p. bogotana*. Sequencing was done in both directions, and no evidence of heterozygosity was observed within samples. The sequences have been submitted to GenBank (accession numbers AF006564–AF006568).

Data analysis: The large majority of the DNA sequences were assembled and aligned visually. For two difficult portions of the *Adh* region, and in order to align the *D. persimilis Adh* sequences with those from *D. pseudoobscura* and *D. p. bogotana*, the multiple sequence alignment program PILEUP of the Genetics Computer Group Sequence Analysis Software Package was also used. Most polymorphism and recombination analyses were carried out using the SITES computer program (HEY and WAKELEY 1997). Gene tree estimates were carried out with the PHYLIP computer program package (FELSENSTEIN 1993).

Isolation model fitting: WAKELEY and HEY (1997) developed a method for fitting a general model of speciation via isolation to polymorphism data that come from two closely related populations or species. This model assumes that two descendant populations formed from an ancestral population at a single time point and that there was no gene flow between the populations beyond that time. Each of the three populations have constant sizes, though they may be different from one another. The input data are the counts of four types of polymorphic base positions: polymorphisms that are exclusive to species 1, the same for species 2, polymorphisms that are shared by the two species, and polymorphisms that appear as fixed differences between the two species. The method yields estimates of the population mutation parameter θ, which is equal to $4Nu$, where N is the effective population size and u is the neutral mutation rate. Since there are three species (species 1, species 2, and the ancestral species) each of which may have a unique effective population size, there are three population mutation parameters, θ_1, θ_2, and θ_A. The method also yields an estimate of the time since isolation T, in units of $2N_1$ generations (note that WAKELEY and HEY primarily used a slightly different measure of time, τ, which is easily converted to T by the relation $T = \tau/\theta_1$). In the original report, a method was not provided for the case when data come from multiple loci with varying sample sizes. Here we describe a modified method that addresses three aspects of multilocus data sets: (1) samples from different loci may be of different sizes; (2) different loci may have inherently different effective population sizes if, for example, some are autosomal and others are *X*-linked; and (3) different loci may have different neutral mutation rates or different lengths.

Assume that l loci have been sampled and that the sample sizes for locus i in the two populations are $n_1^{(i)}$ and $n_2^{(i)}$. Point (1) above is that there may be l different $n_1^{(i)}$ and $n_2^{(i)}$. Next a scaling factor must be included to account for different models of inheritance among loci; this is point (2) above. Let $g^{(i)}$ be the ratio of the effective copy number of locus i to that of an autosomal locus. Thus $g^{(i)} = 1$ for autosomal loci, $g^{(i)} = {}^3/_4$ for *X*-linked loci, and $g^{(i)} = {}^1/_4$ for uniparentally inherited loci (*e.g.*, organellar or *Y*-linked genes). Both the $n^{(i)}$ and the $g^{(i)}$ are known at the outset of the analysis and are not to be estimated. After adjusting by $g^{(i)}$, the model parameters for locus i, $\theta_1^{(i)}$, $\theta_2^{(i)}$, $\theta_A^{(i)}$, and $T^{(i)}$, may vary among loci depending on the neutral mutation rate at each locus. This is point (3) above and is addressed by introducing a new parameter, f, which does need to be estimated from the data. Thus $f^{(i)}$ is defined as the fraction of the total neutral mutation rate that is attributable to locus i. Since

$$\sum_{i=1}^{l} f^{(i)} = 1,$$

there are just $l - 1$ independent $f^{(i)}$ to be estimated.

If θ_1, θ_2, θ_A, and T are the total parameters for all l loci combined, the single locus values are $\theta_j^{(i)} = g^{(i)} f^{(i)} \theta_j$, where j is either 1, 2, or A, and $T^{(i)} = f^{(i)} T$. The expectations of the numbers of exclusive, shared, and fixed polymorphic sites at each locus ($S_{X1}^{(i)}$, $S_{X2}^{(i)}$, $S_S^{(i)}$, and $S_F^{(i)}$, respectively) depend on these parameters and are given in WAKELEY and HEY (1997). With multilocus data, estimates are obtained for the the four total parameters plus $l - 1$ values of $f^{(i)}$. These are obtained by numerical solution of the following system of $4 + l - 1$ equations.

$$\sum_{i=1}^{l} S_{X1}^{(i)} = \sum_{i=1}^{l} E(S_{X1}^{(i)} \mid n_1^{(i)}, n_2^{(i)}, g^{(i)}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_A, \hat{T}^{(i)}, \hat{f}^{(i)}),$$

$$\sum_{i=1}^{l} S_{X2}^{(i)} = \sum_{i=1}^{l} E(S_{X2}^{(i)} \mid n_1^{(i)}, n_2^{(i)}, g^{(i)}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_A, \hat{T}^{(i)}, \hat{f}^{(i)}),$$

$$\sum_{i=1}^{l} S_S^{(i)} = \sum_{i=1}^{l} E(S_S^{(i)} \mid n_1^{(i)}, n_2^{(i)}, g^{(i)}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_A, \hat{T}^{(i)}, \hat{f}^{(i)}),$$

$$\sum_{i=1}^{l} S_F^{(i)} = \sum_{i=1}^{l} E(S_F^{(i)} \mid n_1^{(i)}, n_2^{(i)}, g^{(i)}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_A, \hat{T}^{(i)}, \hat{f}^{(i)}), \text{ and}$$

$$\begin{aligned} S_{X1}^{(i)} + S_{X2}^{(i)} + S_S^{(i)} + S_F^{(i)} &= E(S_{X1}^{(i)} \mid n_1^{(i)}, n_2^{(i)}, g^{(i)}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_A, \hat{T}^{(i)}, \hat{f}^{(i)}) \\ &\quad + E(S_{X2}^{(i)} \mid n_1^{(i)}, n_2^{(i)}, g^{(i)}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_A, \hat{T}^{(i)}, \hat{f}^{(i)}) \\ &\quad + E(S_S^{(i)} \mid n_1^{(i)}, n_2^{(i)}, g^{(i)}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_A, \hat{T}^{(i)}, \hat{f}^{(i)}) \\ &\quad + E(S_F^{(i)} \mid n_1^{(i)}, n_2^{(i)}, g^{(i)}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_A, \hat{T}^{(i)}, \hat{f}^{(i)}), \end{aligned}$$

for $1 \le i \le l - 1$.

RESULTS

Polymorphism summary: The polymorphisms for *Hsp82* are shown in Figure 1, while those for *D. persimilis Adh* are shown in Figure 2. A summary of the numbers of polymorphisms and of estimators of the neutral mutation parameter θ (equal to $4Nu$ for autosomal genes and $3Nu$ for *X*-linked genes) is given on a per base pair basis in Table 1. The overall pattern appears to be one in which *Adh* is the most polymorphic locus, followed by *Period*, and then *Hsp82*. Among taxa, *D. pseudoobscura* is the most variable, followed by *D. persimilis*, and then *D. p. bogotana*. The one exception to these patterns is *Period* in *D. p. bogotana*, which revealed very little variation. This pattern was statistically significant in HKA tests, suggesting that natural selection had removed variation from *D. p. bogotana* at *Period* (WANG and HEY 1996).

As in the case of *Period* (WANG and HEY 1996), both *Hsp82* and *Adh* showed the greatest divergence in comparisons between *D. miranda* and the other species. At *Hsp82*, net divergence per base pair between *D. miranda* and the other taxa was ~0.023 in each of the contrasts. At *Adh*, the net divergence values involving *D. miranda* ranged from 0.025 to 0.028 changes per base pair (SCHAEFFER and MILLER 1991). These values can be compared with those from other species contrasts in Table 2. The finding that *D. miranda* is the most distant member of this group is consistent with the original reports of morphological and chromosomal differences between *D. miranda* and *D. pseudoobscura* (DOBZHANSKY and EPLING 1944) as well as numerous reports of genetic differences among species.

Analyses of the differences among *D. pseudoobscura*, *D. p. bogotana*, and *D. persimilis* are shown in Tables 2 and 3. Interestingly, the pattern of divergence is not the same for all loci. Table 2 shows the levels of net divergence, which is the average pairwise divergence between species, minus the average of the within-species average pairwise variation (NEI 1987). Under a simplistic speciation model with no gene flow, and in which the ancestral species has a population size that is the average of that of its descendants, net divergence is expected to be proportional to the time since speciation (HUDSON *et al.* 1987). Variation among loci for net divergence should mirror variation among loci for polymorphism levels within species (Table 1). Furthermore, the ranking of net divergence levels among species pairs should be the same for all loci. However, neither of these expectations are borne out by the data. *Adh* reveals per base pair values of net divergence that are on par with or less than that for *Hsp82* (Table 2), even though the *Adh* locus shows considerably more variation per base pair within species than the other loci (Table 1). Also, based on net divergence at *Adh*, *D. pseudoobscura* and *D. persimilis* are the most closely related species pair. This pattern conflicts with the other loci and with mitochondrial, protein electrophoretic, and chromosomal inversion data (DOBZHANSKY *et al.* 1963; PRAKASH 1969, 1972; SINGH 1983; ORR 1989b; BARRIO *et al.* 1992). The pattern of wide variation among loci for measures of divergence is different from observations made in a similar study on the *D. melanogaster* species complex, where different loci showed similar patterns of divergence (HEY and KLIMAN 1993).

Similar patterns can be seen in estimates of the population migration parameter *Nm* (Table 2). An *Fst*-based estimate of *Nm* can be generated from the observed pairwise differences within and between populations or taxa (HUDSON *et al.* 1992), assuming an equilibrium model of constant population size and constant rates of gene flow. For each species pair, the estimate of *Nm* is roughly an order of magnitude higher for *Adh* than for *Period* and *Hsp82*. A small part of this difference is expected because of the autosome *vs.* *X* chromosome difference. However adjusting the *Period* and *Hsp82* values upwards by $^4/_3$ does not appreciably change the pattern. Even more striking is that the estimated migration between *D. pseudoobscura* and *D. persimilis* at *Adh* is two to three times the corresponding value for *Adh* in the other species contrasts.

That the different loci have different histories is also apparent from the numbers of shared polymorphisms and fixed differences (Table 3). In general, populations that have just recently diverged from a common ancestor or are sharing genes via migration are expected to share polymorphic sites. In contrast, populations that have not shared ancestry recently and are not engaged in gene flow will have gene trees that coalesce more recently than the time of species divergence and will have fixed differences between species (HEY 1991; WAKELEY and HEY 1997). *Adh* reveals only a single fixed difference between *D. persimilis* and *D. p. bogotana* and none in the other species contrasts. *Adh* does reveal a very large number of shared polymorphisms. In fact, 32 polymorphisms are found in all three taxa. In contrast, *Hsp82* and *Period* primarily reveal fixed differences and relatively few shared polymorphisms. One exception is at *Period* between *D. pseudoobscura* and *D. persimilis*, where six shared polymorphisms and two fixed differences were found. Most of these shared polymorphisms were due to a *D. persimilis* sequence that closely resembled *D. pseudoobscura* sequences over a portion of its length. This sequence probably represents an instance of gene flow, sometime in the distant past (although more recent than the speciation event between these taxa) (WANG and HEY 1996).

Table 4 shows estimates of the population recombination rate and the number of recombination events per mutation event. For *Hsp82* there was evidence of recombination only in *D. pseudoobscura* and that was apparently at a lower rate than found for the other loci. *Adh* and *Period* reveal evidence of high levels of recombination. These high levels of recombination are especially note-

```
Base
position                  1111122 2222222223 3333333334 4444444444 5555555555 5555555555 5555555555 5556666666
          123334455 7991144800 3568889990 1111118990 2222234577 1345556666 6778888888 8889999999 9990000033
         8060271524 4452718016 3770491233 0125678134 3456794856 4577890123 4890123456 7890123456 7890567824
I/R/S    IIIIIIIIII I  IIIIIII IIIIIIIIII IIIIIIIIII      IIIII III      I   I   I            I          I
Indel               DD                                DDDDD        DDDDDDD DDDDDDDDDD DDDDDDDDDD DDDDDDDDD D
Consensus TTGGTGACGG G--TTAATGT AGATTGTCTC TTGACCATCA -----AGTCG GCC----GTC TTTTGCCACT C---------- ---A----TC
PSEUDO1   .......... .......... .......... .......A.. .......... .......--- -......... .ATGCATACA CTC.......
PSEUDO2   ...A..T... .TA......A .....A.... .......A.. .....T.... .T........ .......... .ATGCATAAA CTC.......
PSEUDO3   .......... .......... .......... .......A. GGTTCT.... .T.GTAT... .......... .ATGCATACA CTC.......
PSEUDO4   .......... .......... .......... .......... .......... .......--- -......... .ATGCATACA CTC.......
PSEUDO5   .......... .......... .......... .......... .......... .........A .......... .ATGCATACA CTC.......
PSEUDO6   .......... .......... .......... .......... .......... .......--- -......... .ATGCATACA CTC.......
PSEUDO7   .......... .......... ..G.A..... ..C....A.. .....T.... .......--- -......... .ATGCATACA CTC.......
PSEUDO8   .......... .......... .......... .......A.. .......... .......--- -......... .ATGCATACA CTC.......
PSEUDO9   .......... .......... .......A.. .......A. .....T.... .T........ .......... .ATGCATACA CTC.......
PSEUDO10  ..T....... .......... .......... .......A. .....T.... .T........ .......... .ATGCATACA CTC.......
PSEUDO11  .......... .......... .......... .......A. .....TC... .T........ .......... .ATGCATACA CTC.......
BOGOTA60  .........A .TA.A..... .......A.. .......A.. .......... .......--- -......... .......... ..........
BOGOTA61  .......... .TA....... G......A.. .......A.. .......... .......--- -......... .......... ..........
BOGOTA62  .......... .TA....... G......A.. .......A.. .......... .......--- -......... .......... ..........
BOGOTA63  .........A .TA.A..... .......A.. .......A.. .......... .......--- -......... .......... ..........
BOGOTA67  ........NA .TA.A..... .......A.. .......A.. .......... .......--- -......... .......... ..........
BOGOTA70  ........TA .TA.A..... .......A.. .......A.. .......... .......--- -......... .......... ..........
BOGOTA73  .........A .TA.A..... .......A.. .......A.. .......... .......--- -......... .......... ..........
BOGOTA74  .........A .TA.A..... .......A.. .......A.. .......... .......--- -......... .......... ..........
BOGOTA77  ........TA .TA.A..... .......A.. .......A.. .......... .......--- -......... .......... ..........
PERSIMI40 .......... ......G... .......... .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI41 .......... ......G... .......... .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI42 .......... ......G... .......... .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI43 .......... ......G... .......... .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI44 .......... ......G... ......AT.. .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI45 .......... ......G... .......... .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI46 .......... ......G... ......AT.. .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI47 .......... ......G... .......... .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI48 .......... ......G.A. .......... .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI49 .......... ......G... .......... .......... .......... .TA....... .--------- -......... ...-....A-
PERSIMI50 .......... ......G... .......... .......... .......... .TA....... ..G.------ -......... ........A-
MIRANDA22 AC..AT.A.. A..G.T.C.. .A.AGAG.CA GGTCAAT..T .......CTT A......... ......A... .ATGCATACA CTC.CATAA.
MIRANDA23 AC..AT.A.. A..G.T.C.. .A.AGAG.CA GGTCAAT..T .......CTT A......... ......A... .ATGCATACA CTC.CATAA.
MIRANDA24 NC..AT.A.. A..G.T.C.. ...AGAG.CA GGTCAAT..T ........TT A......... ......A... .ATGCATACA CTC.CATAA.
MIRANDA25 AC..AT.A.. A..G.T.C.. ...AGAG.CA GGTCAAT..T ........TT A......... ......A... .ATGCATACA CTC.CATAA.
```

FIGURE 1.—Polymorphic sites of 35 samples from *D. pseudoobscura* (PSEUDO), *D. p. bogotana* (BOGOTA), *D. persimilis* (PERSIMI), and *D. miranda*. Base position refers to the base number in the aligned set of sequences of variable positions. Base 1 corresponds to base 873 of BLACKMAN and MESELSON (1986). I/R/S, intron base substitution, coding region amino acid replacement substitution, and coding region synonymous substitution, respectively. For the line labeled Indel, a D refers to a base position where a sequence had a gap relative to the consensus sequence. Gaps are indicated with a hyphen.

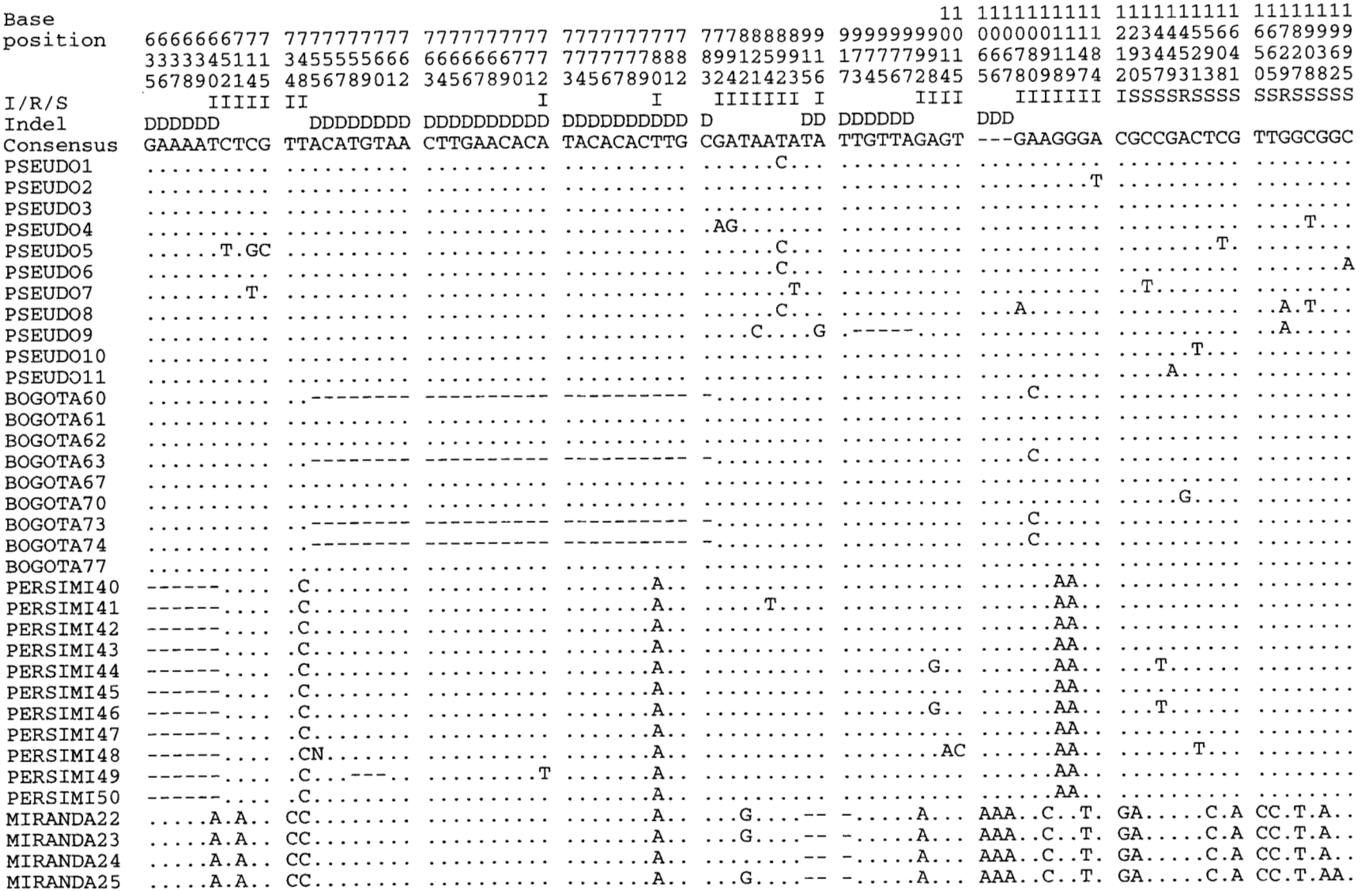

FIGURE 1.—*Continued*

```
Base                                                                     111111111
position       111111111 1111111222 2222223333 3333444444 4455577789 9000000000
               334555566 6667889124 5577880127 8899005778 9912911738 8444444444
              3478478901 2347127862 2845566901 0258379671 0283645833 4012345678
I/R/S         IIIII         IIIIIII IIIIIIIIII IIIIII III IIII  IIIR R I  I
Indel              DDDDD DDD                         D        DDD     DDDDDDDDD
Consensus     TTAGCACTCG GTCAACGATG GAGGGCACTA TCCGTT-TTT GACGGCGACC GACGAA----
DPERS         .......... .....T.... T.TC..TTC. .G......G. ....-....G C-----GAAA
40            .C..T----- ---CG.AG.. .........T ...AC..G.A T......... ..........
42            ..G..----- ---......T .G...T.... ...AC.A... .G....A... ..........
44            C......... .....TAG.. .......... ..T...A... .G...--G.. ..A.......
49            C.GA...... .......... ....TT.... ..T....... ..TC....T. ..A.......
50            ..G..----- ---C....C. T.TC..TTC. C.T..G...A .....--... .....G....

              1111111111 1111111111 1111111111 1111111111 1111111111 1111111111
              0000000000 0000000011 1122555555 5555555555 5555555555 5555555666
              4555555555 5678889900 4556122233 3333333344 4444444455 7777788444
              9012345678 9025694702 4697958901 2345678901 2345678901 2345901345
I/R/S                      I II III SSRSII                       I          S R
Indel         DDDDDDDDDD DD D  D          DDDD DDDDDDDDDD DDDDDDDDDD DDDDDD  D
Consensus     ---------- --CAATATTA TAGATA---- ---------G TTTTTCCTAC TTTG---C-A
DPERS         GAGCAATACC GT.-TA-..T ....ATAACT ATCCCAAAA- ---------- ----TAT.AG
40            .......... ..A.....AT .......... .......... .......... ..........
42            .......... ........AT .G........ .......... .......... ..........
44            .......... .......C.. .......... .......... .......C.. ........TA.
49            .......... .......C.. .......... .......... .......... ..........
50            .......... .......A.. A.CCA.AACT A........- ----...... ..........

              1111111111 1111111111 1122222222 2222222222 2222222222 2222222222
              7789999999 9999999999 9900000112 2233333333 3333333333 3333333333
              0640666666 6677777777 7799999003 9900000000 0011111222 2223333666
              0644234567 8901234567 8912389079 8901234567 8901234456 7890123678
I/R/S         SSS                           IS
Indel            DDDDDDD DDDDDDDDDD DDDDDDDD   DDDDDDDDDD DDDDDDDDDD DDDDDDDDDD
Consensus     ACCT------ ---------- --------TC GTAGAG---T GGTGTAGAGT GTTCCGATCT
DPERS         G...ATTCTC TTTTATGGAA GG.......T .-----...- -------... ..........
40            .T........ .......... .....TGGG. .......... .......--- -------...
42            .T........ .......... ..TCG..... ------...- ------.... .......---
44            ..T....... .......... ..TCG....T .......... .......... .......---
49            G..-ATTCTC TTTTATGGAA GG........ .......... .......... ..........
50            G......... .......... .......... ..-...AGG. .......--- -------...

              2222222222 2222222222 2222222222 2222222222 2233333333 3333333333
              3333333333 3333333333 3444444455 5566668888 9900000000 0111112222
              6777777778 8888888889 9000000888 9901360079 2603345566 7377890344
              9012467890 1234567897 8013456089 0106031767 4357930338 6589908134
I/R/S             I                 IIII    I    IISSSSSS SSSIIIIIII ISRRSRRSSR
Indel         DDDD DDDDD DDDDDDDDDD     DDD DD DD
Consensus     AGTCCAGTCT CT-------T GTGT---TCA TAACCTCGTT TAGACAAGGT CTGCAACGCC
DPERS         ....T..... .......... .......G.. ...T...... ...G..C... T.A..T....
40            .......... ..AGTCTCT. .......... .......... ..C.GG.... ......G..A
42            ----...... .......... ........-- --....T.GA ..C.GG.... ...AGTG..A
44            ----...... .......... .......... ..CT...T.. .G.....C.G TC....G...
49            .....----- --.......- ACACAGC... ....TC...A C........C ........T.
50            .....----- --........ .......... .......... .G......AC ...AGT.T..

              33333
              34445
              10880
              68252
I/R/S         RSIII
Indel
Consensus     TCGCA
DPERS         GG.TT
40            ..TT.
42            ...T.
44            .G..T
49            .G...
50            .....
```

FIGURE 2.—Polymorphic sites of six samples of *D. persimilis Adh* region sequences. Line numbers refer to *D. persimilis* strains listed in Table 1 of (WANG and HEY 1996). The DPERS line was done previously by SCHAEFFER and MILLER (1991). The complete alignment of *D. persimilis* sequences with the larger data set of *D. pseudoobscura* and *D. p. bogotana* (SCHAEFFER and MILLER 1991, 1992a,b) is available upon request to J.H. See Figure 1 legend for additional explanation.

TABLE 1

Polymorphism summaries

Species	*Period*				*Hsp82*				*Adh*[a]			
	n	S	$\hat{\theta}$	π	n	S	$\hat{\theta}$	π	n	S	$\hat{\theta}$	π
pseudoobscura	11	48	0.0112	0.0084	11	34	0.0059	0.0042	99	400	0.0225	0.0105
p. bogotana	9	3	0.0008	0.0009	9	6	0.0016	0.0012	8	61	0.0068	0.0066
persimilis	11	36	0.0083	0.0070	11	10	0.0018	0.0012	6	94	0.0119	0.0118
miranda	4	9	0.0033	0.0032	4	4	0.0011	0.0012	1	—	—	—

n, number of DNA sequences in the sample; S, number of polymorphic sites; $\hat{\theta}$, WATTERSON's estimate of θ (WATTERSON 1975; TAJIMA 1993); π, average number of pairwise differences, also an estimate of θ (TAJIMA 1993); for both $\hat{\theta}$ and π, the value for each complete locus has been divided by the number of base pairs for that locus.

[a] The *D. pseudoobscura* sequences for *Adh* were reported in a series of papers by SCHAEFFER and MILLER (1991, 1992a,b). SCHAEFFER and MILLER (1991 also reported the sequences for *D. p. bogotana, D. miranda,* and one strain of *D. persimilis.*

worthy for their effect on the variance of other estimates. Recombination within a locus reduces the stochastic variance of the genealogical history of a locus (HUDSON 1983), so that the pattern of variation is expected to be closer to the average of that for all loci. Thus estimators of θ and Nm are expected to be more accurate, on average, when the recombination rate is high.

Testing speciation models: The patterns of variation within and among loci suggest that *D. pseudoobscura* and close relatives may have been sharing genes subsequent to speciation and that the rate of gene flow may vary among different parts of the genome. If true, then the data suggest a speciation model that is quite interesting and more complicated than one in which gene flow has been absent for all loci for the same length of time. In general, the simplest model of speciation is an isolation model in which two populations become completely separated at a single point in time, with no gene exchange thereafter. This model corresponds roughly to allopatric models of speciation, and it is one for which coalescent models of divergence are tractable (TAKAHATA and NEI 1985; HUDSON *et al.* 1987; HEY 1991, 1994; WAKELEY and HEY 1997). To test the fit between the data for the three loci and this kind of isolation model (with no gene flow) we carried out the following procedure: (1) a test statistic, a measure of variation in fixed and shared differences, was calculated from the data; (2) population size parameters and speciation times were estimated from the data sets assuming a simple isolation model; (3) population recombination rates were estimated from the data using the method of HEY and WAKELEY (1997) (see Table 4); (4) simulated values of the test statistic were found by carrying out coalescent simulations using the estimated parameters, including the estimated recombination rates; and (5) the observed test value was compared to the distribution of values generated by the simulations.

We considered two primary criteria in selecting a test statistic: sensitivity to variation among loci for gene flow, and simplicity. For loci that are not engaged in gene flow, a basic finding is that the expected number of fixed differences will increase for greater divergence times (HEY 1991) and that the expected numbers of shared polymorphisms will decrease for greater divergence times (WAKELEY and HEY 1997). Thus these two polymorphism measures are expected to negatively covary; indeed, in the absence of recombination, a locus can only reveal either fixed differences or shared polymorphisms (or neither, of course). If we now consider a locus with a relatively large divergence time *and* some gene flow, then the gene flow may introduce shared polymorphisms that would not otherwise be expected. This line of reasoning suggests a test statistic that would have a high value when there is lots of variation among loci for fixed differences *and* when there is lots of variation among loci for shared polymorphisms. The test statistic we used was the difference between the highest and lowest values of fixed differences among the three loci *plus* the difference between the highest and lowest values of shared polymorphisms. This quantity is eaasily calculated, and it is expected to be sensitive to variation in both fixed and shared differences.

The first speciation model tested was that used by HUDSON *et al.* (1987), in which the ancestral species has a population size that is the average of the two descendant species. The tests of this Hudson, Kreitman, and Aguadé (HKA) isolation model are shown in Table 5. In all species pairs, the observed values of the test statistic are shown to be very unlikely, and the speciation model does not fit the data.

The HKA isolation model imposes an assumption that the ancestral population size was the average of the two descendants. A rejection of this model (Table 5) may just represent a failure of this restrictive assumption. We also tested a more general isolation model in which the ancestral population size does not depend on that of the descendant species (WAKELEY and HEY 1997). This model is similar to the HKA model, but it includes an additional parameter, $\theta_A = 4N_A u$, which is the population mutation parameter for the ancestral species prior to the time of speciation. WAKELEY and HEY (1997) describe a procedure for estimating model

TABLE 2

Divergence and migration

Locus	Net divergence per base pair			Population migration rate estimate (Nm)		
	pseudoobscura/ p. bogotana	*pseudoobscura/ persimilis*	*p. bogotana/ persimilis*	*pseudoobscura/ p. bogotana*	*pseudoobscura/ persimilis*	*p. bogotana/ persimilis*
Adh	0.00200	0.00122	0.00329	1.075	2.293	0.703
Period	0.00879	0.00967	0.01537	0.131	0.198	0.064
Hsp82	0.00176	0.00413	0.00571	0.386	0.165	0.054

Net divergence is calculated using expression 10.21 of NEI (1987). For migration rate estimation, N is the effective population size and m is the fraction of individuals that are migrants each generation. Nm was estimated using expression 4 of HUDSON *et al.* (1992), with the exception that a factor of $^1/_4$ replaces a factor of $^1/_2$ so that the estimate applies to the case of diploidy. For the *X*-linked loci, *Period* and *Hsp82*, the estimates in the table can be multiplied by $^4/_3$ for comparison with the diploid *Adh*.

parameters using data on exclusive, shared and fixed polymorphisms (see also MATERIALS AND METHODS). The same basic test procedure that was used for the HKA isolation model as shown in Table 5 was done for this more general isolation model. The results are shown in Table 6. One of the effects of having some loci with large numbers of fixed differences and others with large numbers of shared differences is to generate a very large value for the estimated population size for the ancestral species. This effect is especially extreme for the *D. p. bogotana/ persimilis* contrast (Table 6). Statistical tests using these estimated parameter values also indicate that the isolation model is not consistent with the data, though the data fit better than under the HKA model assumptions. Simulations could not be conducted for the *D. p. bogotana/ persimilis* contrast because of difficulties in implementing recombination under extreme population sizes in the common ancestor. However the extreme parameter estimates by themselves suggest that the isolation model is not appropriate for this species pair.

The conclusion from these tests is that neither isolation model is consistent with the data. The deviation is in the direction of increased variation among loci, which is consistent with a history that includes gene flow (WAKELEY 1996). It is possible that another kind of history without gene flow, perhaps with some pattern of changes in population size, could explain this large degree of variation. However, the relative generality of the isolation model (WAKELEY and HEY 1997), in that it allows for some changes in population size, should decrease the chance of a spurious result. Especially when considered together with other evidence (see DISCUSSION), these tests indicate a history of gene flow among these species.

TABLE 3

Numbers of shared polymorphisms and fixed differences

Locus	*pseudoobscura/ p. bogotana*		*pseudoobscura/ persimilis*		*p. bogotana/ persimilis*	
	Shared	Fixed	Shared	Fixed	Shared	Fixed
Adh	52	0	67	0	33	1
Period	1	6	6	2	0	16
Hsp82	0	0	1	8	0	11

Gene tree estimation: Figure 3 shows an estimated gene tree for *Hsp82*. With the exception of the tree spanning the *D. pseudoobscura* samples, which contains a subtree for the *D. p. bogotana* samples, each of the species samples form monophyletic groups. This tree is probably a good estimate of the true *Hsp82* genealogy, except within *D. pseudoobscura* where there has been some recombination (Table 4): boot strap values for deep branches among taxa are >80% (Figure 4); and a maximum parsimony analysis on a reduced data set (with just one sequence representing *D. pseudoobscura*) returned a single most parsimonious tree with consistency index 1.0 (results not shown).

Hsp82 lies in chromosome section 23 of *X*R, the right arm of the *X* chromosome (BLACKMAN and MESELSON 1986; SEGARRA *et al.* 1996). This chromosome section also contains the *Esterase-5* gene cluster (BABCOCK and ANDERSON 1996), a region that was also the subject of a recent comparative DNA sequence study in this species group. BABCOCK and ANDERSON (1996) examined a 500-bp intergenic region in *D. pseudoobscura*, *D. persimilis*, and *D. miranda* (though not *D. p. bogotana*). Among the non-Sex-Ratio chromosomes in that study, the gene tree relationships among the three taxa are similar to those in Figure 3. One difference is that, at *Esterase-5*, the sample of *D. persimilis* sequences revealed no variation and formed a cluster that fell within a larger tree of *D. pseudoobscura* sequences. Like *Hsp82*, the *Esterase-5* data showed no evidence of gene flow between *D. pseudoobscura* and *D. persimilis*.

Gene trees can be a useful tool for studying migration or the admixture of sequences among populations (SLATKIN and MADDISON 1989). However, the *Adh* region has experienced high levels of recombination (SCHAEFFER and MILLER 1993), so that the true genealogy is a complex network and not a bifurcating tree. It is possible to estimate trees for short portions of the sequence that do not appear to have experienced much recombination. Figure 5A shows a neighbor-joining tree

TABLE 4

Recombination estimates

Species	*Period*		*Hsp82*		*Adh*	
	γ	$\gamma/\hat{\theta}$	γ	$\gamma/\hat{\theta}$	γ	$\gamma/\hat{\theta}$
D. pseudoobscura	0.0271	2.411	0.0026	0.436	0.0605	2.694
D. p. bogotana	0.0	0.0	0.0	0.0	0.0149	2.182
D. persimilis	0.0226	2.728	0.0	0.0	0.0798	6.681

γ is an estimate of the population recombination rate $4Nc$, where c is the recombination rate per generation per base pair (HEY and WAKELEY 1997). For the *X*-linked loci *Period* and *Hsp82*, γ is an estimate of $3Nc$. The ratio of recombination rate per base pair to neutral mutation rate per base pair is estimated by dividing γ by $\hat{\theta}$. γ could not be determined for *D. miranda* for *Period* and *Hsp82* because of low levels of variation and for *Adh* because only a single line was sequenced.

(SAITOU and NEI 1987) for a region that showed very little evidence of recombination by the criteria of HUDSON and KAPLAN (1985). Figure 5B shows a maximum parsimony tree for a shorter region that showed no evidence of recombination. Although both trees reveal a tendency for sequences to cluster by the taxon designations, both trees also reveal multiple instances where sequences do not cluster by taxon. Note also the nearly complete lack of concordance between the two trees. Migration rate estimates can be generated using either the migration counting method of SLATKIN and MADDISON (1989) or the *Fst*-based method of HUDSON *et al.* (1992). Counts of the minimum numbers of migration events required in each of the trees of Figure 5 are given in the legend of that figure. From comparison with Table 1 of SLATKIN and MADDISON (1989) these counts correspond roughly to the following values of *Nm*: *pseudo./p. bogotana*, $0.5 < Nm < 1.5$; *pseudo./persimilis*: $2 < Nm < 4$; and *persimilis/p. bogotana*, $Nm < 0.5$. The *Fst*-based assessments for the two regions in Figure 5 are 0.648 for *pseudo./p. bogotana*, 3.87 for *pseudo./persimilis*, and 0.747 for *persimilis/p. bogotana*.

The trees in Figure 5 are intended as examples of the kinds of gene trees that exist for short intervals. However, because they are based on short sequences and because these regions were selected for their low homoplasy, it is difficult to assess the confidence of the these estimates. It is important to note that the *Fst*-based estimates for the regions in Figure 5 are similar to those in Table 4 for the entire *Adh* region, so these two short regions are not atypical of the *Adh* region with respect to apparent gene flow.

Speciation times: Of the three loci studied, *Hsp82* shows the least evidence of gene flow. The numbers of shared polymorphisms and the *Nm* estimates are low (Tables 2 and 3) and the gene trees show no evidence of gene flow (Figures 3 and 4). If we assume that divergence at *Hsp82* is typical of loci that did not experience gene flow since the time of speciation, then we may use the data from this locus to estimate speciation times.

SHARP and LI (1989) estimated the synonymous substitution rate for *Hsp82* and other Drosophila genes with high codon bias to be 8×10^{-9} per year [the estimated rate was double this for low-bias genes (SHARP and LI 1989)]. Then, following the method used by KLIMAN and HEY (1993), the net divergence between *D. pseudoobscura* and *D. miranda* at *Hsp82* per silent site is 0.042. If the data of the Sophophoran radiation is 40 mya (THROCKMORTON 1975), then these values correspond to an estimated speciation time of 2.63 mya (the estimate is 1.97 mya if the Sophophoran radiation was 30 mya). There is too little synonymous site divergence among *D. pseudoobscura*, *D. persimilis*, and *D. p. bogotana* to estimate speciation dates in the same way. However,

TABLE 5

HKA isolation model tests

Species pair	$\hat{\theta}_1$	$\hat{\theta}_2$	T	Test value	P
pseudoobscura/p. bogotana	35.8	17.2	0.376	43	0.008
pseudoobscura/persimilis	34.6	33.6	0.427	53	0.016
p. bogotana/persimilis	17.6	35.6	1.528	48	0.002

$\hat{\theta}_1$ is the estimate of the population mutation parameter for the first species listed in the species pair in column 1, estimated for the *Adh* locus. $\hat{\theta}_2$ is the same quantity estimated for the second species. For the other loci, the ratio of N_1 and N_2 is the same as for *Adh*, though the estimate of the relative neutral mutation rate is different (HUDSON *et al.* 1987). T is the estimated speciation time in units of $2N_1$ generations. The observed test value was calculated from the observations in Table 3. It is the difference between the highest and lowest values of fixed differences among the three loci plus the difference between the highest and lowest values of shared polymorphisms (see text). P is the probability of observing a more extreme simulated test value than observed, based on 1000 coalescent simulations.

TABLE 6

Wakeley and Hey isolation model tests

Species pair	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_A$	T	Test value	P
pseudoobscura/p. bogotana	46.0	7.7	88.1	0.16	43	0.055
pseudoobscura/persimilis	28.7	24.9	102.9	0.482	53	0.023
p. bogotana/persimilis	0.004	0.009	130.7	1.1	48	—[a]

$\hat{\theta}_A$ is the estimate of the population mutation parameter estimate for the ancestral population (WAKELEY and HEY 1997). Other parameters are as in Table 5.

[a] Simulations could not be conducted for *p. bogotana/persimilis,* because of difficulties in implementing recombination under extreme population sizes in the common ancestor.

there is divergence within the large *Hsp82* intron and this can be used in conjunction with the estimated time for the *D. pseudoobscura/D. miranda* divergence. Net divergence values per base pair for the intron between *D. pseudoobscura* and the other species are 0.0330, 0.0069, and 0.0029 (for *D. miranda, D. persimilis*, and *D. p. bogotana*, respectively). By scaling to the estimated divergence time between *D. pseudoobscura* and *D. miranda* of 2.63 mya, the estimated time for the split between *D. pseudoobscura* and *D. persimilis* is 0.55 mya and the estimated time for the origin of *D. p. bogotana* is 0.23 mya. These estimates are rough, but the values for *D. miranda* and *D. persimilis* are very similar to those based on other loci (AQUADRO *et al.* 1991; BABCOCK and ANDERSON 1996). The estimate for the divergence between *D. pseudoobscura* and *D. p. bogotana* is greater than the estimate of 0.155 mya based on a different method applied to *Adh* (SCHAEFFER and MILLER 1991).

DISCUSSION

At the core of several species concepts are the ideas that the organisms within a species share in some set of defining properties and that these qualities cannot be easily disturbed by gene exchange with organisms from other species. Indeed, under the biological species concept these ideas are joined: a species is defined by interbreeding *and* isolating mechanisms that prevent gene flow with other species (MAYR 1942; DOBZHANSKY 1951). Similarly, under the recognition species concept, the organisms of a species share in common fertilization systems and thus tend not to hybridize with organisms of other species (PATERSON 1993). TEMPLETON (1989, 1994) builds on these concepts, arguing that species are entities with phenotypic and genetic cohesion, and that cohesion can arise from a variety of demographic and population genetic causes. A common thread of these and other species concepts is that species are not easily undone by gene flow, even though some gene flow may occur.

The apparent conflict between the ideas of phenotypically homogeneous species and of gene flow between species is resolved by invoking natural selection. Depending on the number of genes and linkage relationships among genes that are divergent between species because of natural selection (perhaps due to adaptation to local circumstances or to evolution to limit gene flow), gene flow may be absent for some regions of the genome and present for others. A pattern that includes divergence *and* gene flow can be most easily envisioned in a model of sympatric speciation. In general, if speciation occurs and some hybrids are formed and reproduce and if the same goes for subsequent generations of backcross progeny, then some portions of the genome will cross the species boundary. A famous finding of population genetics theory is that very little gene flow between populations is required to maintain genetic equanimity (WRIGHT 1931). Thus a simple prediction of sympatric (and parapatric speciation models) in which phenotypic cohesion and mate recognition are due to a small subset of loci, is that sister taxa may share much of their genetic variation.

In general, speciation models based on a small number of loci, and that include the presence of hybridization, need not preclude gene flow between species at those loci that are *not* associated with species specific adaptations or assortative mating. One of the most interesting, and least explored, manifestations of oligo-locus speciation models is that species can become divergent over just a subset of the genome and may continue to share variation at other parts of the genome.

The data presented here, including new data from *Hsp82* and *Adh,* in conjunction with *Period* locus data (WANG and HEY 1996) and a larger *Adh* data set (SCHAEFFER and MILLER 1991, 1992a,b), are consistent with a speciation model in which species continue to exchange genes at some loci and not at others. The three genes present conflicting portraits of divergence: *Adh* reveals evidence of relatively large amounts of gene flow involving all three taxa; the *Period* data suggest limited, perhaps relatively ancient, gene exchange between *D. pseudoobscura* and *D. persimilis* (WANG and HEY 1996); while only *Hsp82* reveals a pattern consistent with a simple divergence model of speciation, in which gene exchange ceases at the time of species formation.

The contrasts among loci, and the apparently high level of migration at *Adh*, are especially striking given the high levels of recombination that have occurred in the histories of the two genes in the *Adh* region (Table

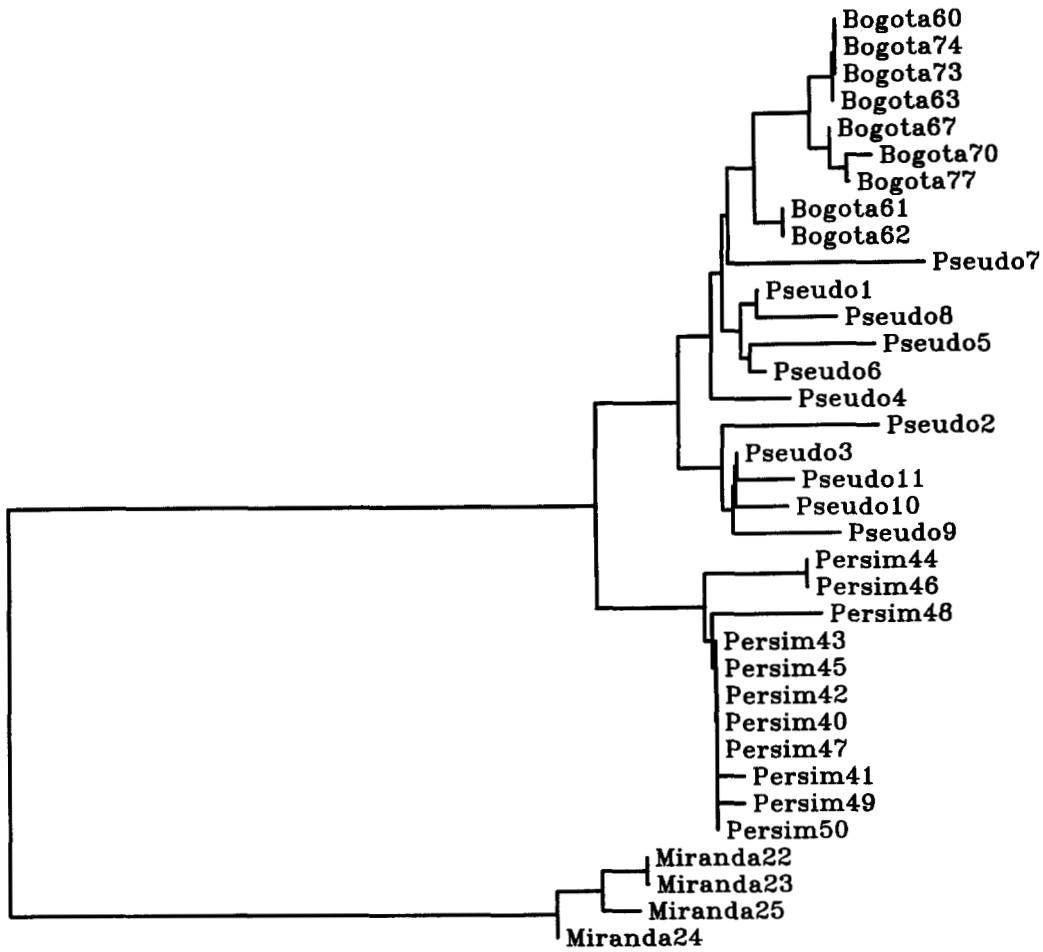

FIGURE 3.—A neighbor joining tree (SAITOU and NEI 1987) of the *Hsp82* sequences constructed using the PHYLIP computer programs DNADIST and NEIGHBOR (FELSENSTEIN 1993). Sequence names are given in Table 1 of WANG and HEY (1996). For reference, the lower deepest branch between the base of the tree and *MIRANDA24* has a length of 0.01 changes per base pair.

4). High recombination causes the estimates of variation and migration for one locus to be closer to the average of that for all loci. Put another way, the probability that one locus appears to have a different history from other loci, whether due to natural selection or by chance, is much reduced if that locus has had considerable recombination. The high level of historical recombination in the *Adh* samples also bears on the high migration rates and the kinds of forces that could contribute to migration. If the apparently high migration rate were due to an unusual pattern of natural selection on the *Adh* region, then only a relatively small portion of the sequence would be affected, because of the high recombination rate (HUDSON and KAPLAN 1988). For example, if some kind of selection created the pattern in Figure 5A, then a different force (*e.g.*, selection on a different base position) would have to be invoked for the pattern in Figure 5B (which has an almost completely different topology from Figure 5A) because of recombination between the two regions represented by these figures.

The *Adh* data are consistent with gene flow among all three taxa, *D. pseudoobscura*, *D. persimilis*, and *D. p. bogotana*. However present day gene flow between *D. persimilis* and *D. p. bogotana* is probably not possible because of their disjoint geographic distributions. It is possible that the large amount of shared polymorphism at *Adh* between these taxa is due to past gene flow, if geographic distributions have changed considerably and recently. Perhaps more likely is that the gene flow between these two has occurred through *D. pseudoobscura*. Certainly, if *D. pseudoobscura* is exchanging *Adh* sequences with both taxa, then *D. pseudoobscura* could be a conduit for variation. In the remainder of the DISCUSSION we consider just two divergences or speciation events: between *D. pseudoobscura* and *D. persimilis* and between *D. pseudoobscura* and *D. p. bogotana*.

The divergence of *D. pseudoobscura* and *D. persimilis*: Differences between these two taxa have been documented for a variety of different traits: they exhibit nonidentical geographic ranges, chromosome inversion differences (DOBZHANSKY and EPLING 1944), and subtle morphological differences (RIZKI 1951). There is also clear and strong evidence for reproductive isolation and thus that natural selection is acting to keep these taxa separate from each other. The two species exhibit considerable postmating reproductive isolation (ORR 1987, 1989b), and there exists geographic variation in *D. pseudoobscura* for the degree of premating isolation (NOOR 1995b). Somehow our model of speciation must reconcile the species differences and the reproductive isolation with the conclusion that gene flow has

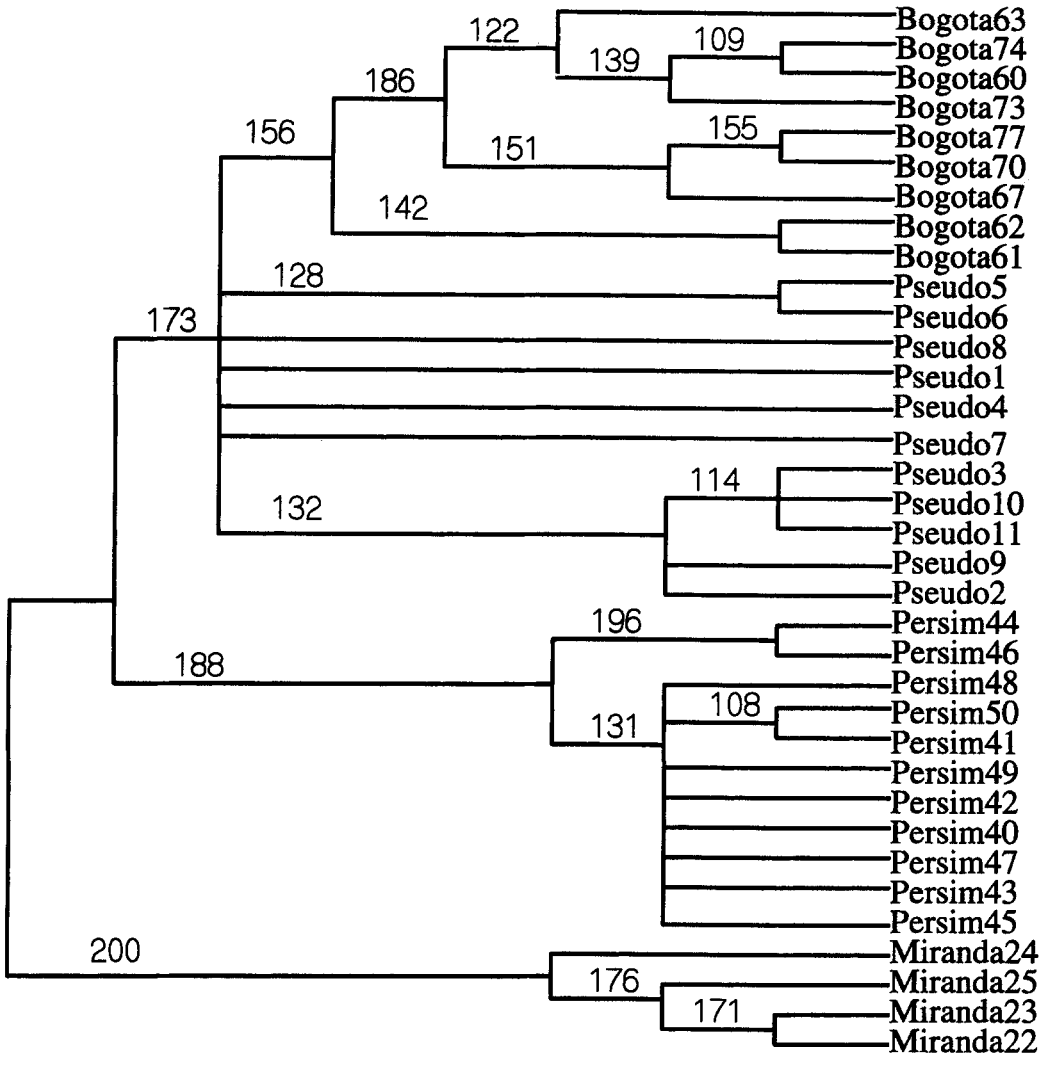

FIGURE 4.—A majority rule consensus tree for *Hsp82* generated with 200 bootstrap replications (FELSENSTEIN 1985). The tree was constructed using the PHYLIP computer programs SEQBOOT, DNADIST, NEIGHBOR, and CONSENSE (FELSENSTEIN 1993). Only those branches that appeared in >50% of the trees are shown. The numbers of trees supporting a branch are shown above the branch.

been occurring (see RESULTS: *Testing speciation models*). One possible explanation is that gene flow ceased not very long ago and that the reproductive isolation and those traits that distinguish the species have arisen very recently. However two kinds of evidence suggest that gene flow is either ongoing or has continued until recently: the occurrence of backcross hybrids in nature (DOBZHANSKY 1973; POWELL 1983) and the spacing of nodes that indicate migration in gene trees from the *Adh* region (Figure 5). Some of the most recent nodes in these trees indicate migration events because they represent ancestors of sequences collected from multiple species (SLATKIN and MADDISON 1989).

If all the evidence is considered together, including evidence of genetic differentiation and reproductive isolation between these species and the evidence of gene flow and the rejection of isolation models of speciation, there is strong reason to conclude that gene flow is occurring at some loci *and* that natural selection is preventing gene flow for other loci.

However, a finding of natural selection does not necessarily mean that those loci that showed less gene flow (*e.g.*, *Hsp82*) are closely linked to sites where natural selection is preventing gene flow. Among loci that experience limited gene flow, there is expected to be a wide variance in the depths of gene trees and the apparent level of divergence between species (WAKELEY 1996). In general, a model of divergence via isolation will generate less variance among loci for gene tree depths than will a model of divergence via limited gene flow (WAKELEY 1996). Thus, while the data presented here cannot be reconciled with an isolation model, it may be difficult to reject a model in which the different loci in the study are subject to similar (and low) levels of gene flow. In short, the conclusion of gene flow is based on the data from *Adh*, *Hsp82*, and *Period*, but the conclusion of natural selection maintaining the distinctness of the species is based on the list of other notable interspecies differences that would not be expected if there was gene flow but no natural selection.

Some circumstances do suggest that natural selection is acting to limit gene flow near the *Hsp82* and *Period* loci. Both loci show low levels of estimated gene flow at *Period* and *Hsp82* between all species pairs (and *Adh*

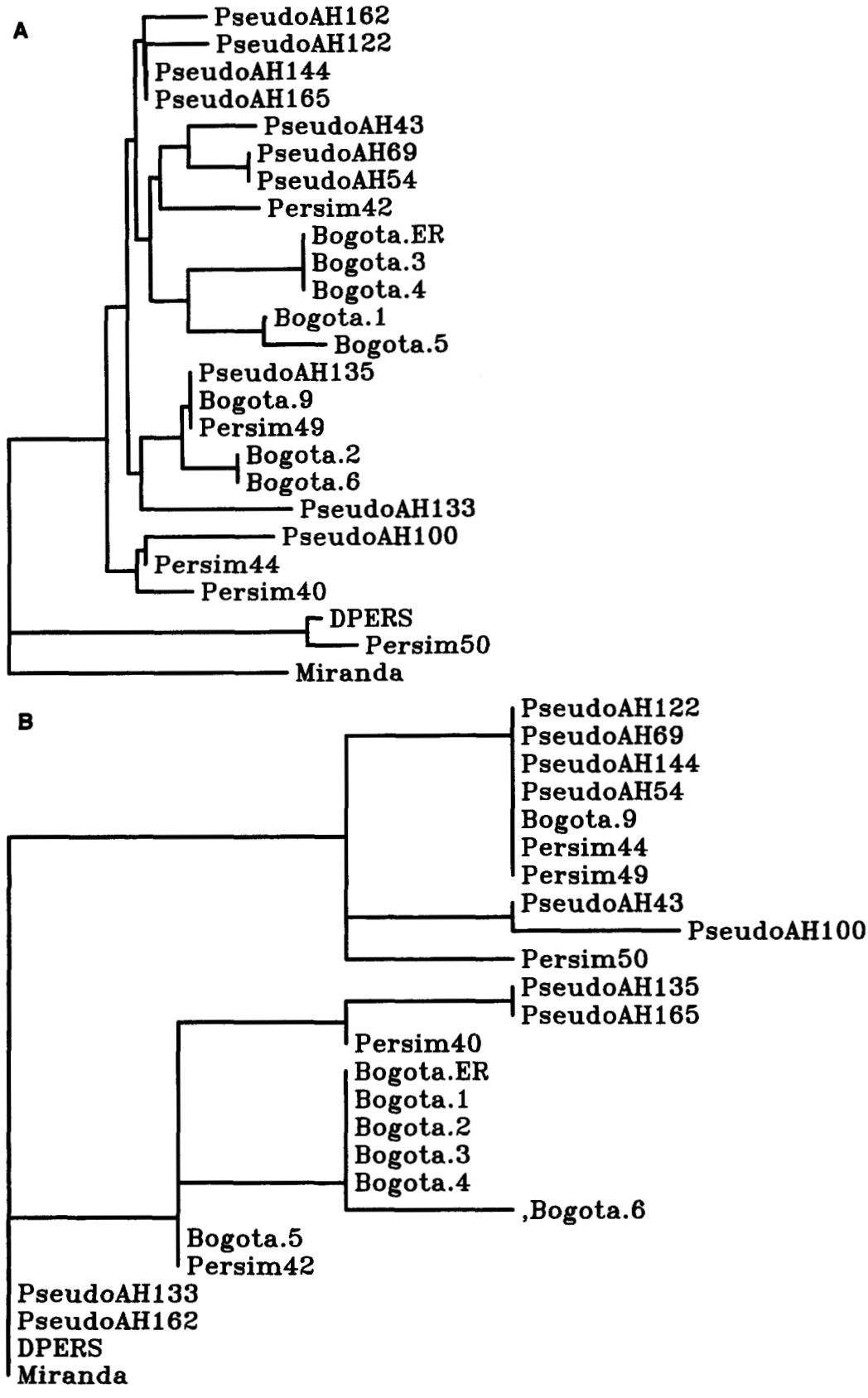

FIGURE 5.—Tree estimates for portions of the *Adh* region. Sequence labels beginning with "Persim" are those reported in this article, and the strains are identified in Table 1 of WANG and HEY (1996). All other sequences are from SCHAEFFER and MILLER (1991). (A) A neighbor joining tree for the *Adh* region between positions 244 and 435 of the aligned sequences. Following the method of SLATKIN and MADDISON (1989), the minimum numbers of migration events suggested by this tree are as follows: three for *pseudo./p. bogotana*, four for *pseudo./persimilis*, and one for *persimilis/p. bogotana*. For reference, the length of the bottom branch that connects *Miranda* is 0.023 changes per base pair. (B) The maximum parsimony tree (length = 11, consistency index = 1.0) for the *Adh* region between positions 1158 and 1268 of the aligned sequences. Except for one deep branch of length 2 near the top of the figure, all branches shown have length 1 corresponding to one change at one polymorphic site. The minimum numbers of migration events suggested by this tree are as follows: two for *pseudo./p. bogotana*, three for *pseudo./persimilis*, and two for *persimilis/p. bogotana*.

shows high levels between all species pairs). This correspondence across different speciation events is not necessarily expected if the high variance among loci is simply due to similar but limited gene flow at all loci. On the other hand, *ad hoc* selection models may invoke similar selection at or near the same loci in separate

cases of speciation. Another reason to think that selection has limited gene flow at the *X*-linked genes *Hsp82* and *Period* is the very high level of recombination apparent at *Adh*. This high recombination means that the estimates of gene flow (as well as other parameters) in this region have relatively low variance. Thus it is possible that the estimates of *Nm* based on *Adh* (Table 2) may accurately reflect the amount of gene flow that would be observed at *Period* and *Hsp82* were there no selection occurring near these genes. If so, this level of gene flow is fairly high and we would not expect to see such low estimates of *Nm* and so many fixed differences at *Period* and *Hsp82*.

One possible factor that could reduce gene flow for *Period* or *Hsp82* between *D. pseudoobscura* and *D. persimilis* is if they are linked to chromosome inversions. Both the *X*L and *X*R elements of the *X* chromosome have been reported to be sites of paracentric inversions that distinguish the species; while no species differences have been reported for chromosome *4* (the site of *Adh*) (Dobzhansky and Epling 1944; Anderson *et al.* 1977; Moore and Taylor 1986; Segarra and Aguadé 1992; Segarra *et al.* 1996). In the case of *Hsp82*, tight linkage to an inversion can be ruled out. This gene has been localized to chromosome section 23 of *X*R (Blackman and Meselson 1986; Segarra *et al.* 1996), which is not near a species-specific inversion. However, this location is near a breakpoint for a segregating Sex-Ratio (SR) inversion in *D. pseudoobscura*, and it is possible that this reduces the effective population size for this locus and others near it (Babcock and Anderson 1996). The physical location of the *Period* locus is not yet known, though based on the strong conservation of chromosome homologies among Drosophila species, it is almost certainly on one of the arms of the *X* chromosome (Muller 1940; Steinemann *et al.* 1984; Segarra and Aguadé 1992; Segarra *et al.* 1995, 1996). It is possible that it is linked to one of the inversions that distinguish the species and that selection on an inversion has limited gene flow for *Period*.

Another consideration regarding the *X*-linked genes is the observation of a large *X*-chromosome effect on sterility in Drosophila species hybrids (Coyne and Orr 1989). For *pseudoobscura/persimilis* hybrids (Orr 1987), as well as for many Drosophila species pairs, a large portion of the postzygotic barrier to mating maps to the *X* chromosome (Coyne and Orr 1989).

The findings of gene flow, variable selection against gene flow, and the findings that natural selection may be acting to reinforce mate choice in regions of sympatry between *D. pseudoobscura* and *D. persimilis* (Noor 1995b) are consistent with a sympatric speciation model. Perhaps the current sympatry persists since the onset of divergence, and the current degree of isolation is just a stage of a speciation process that originated as functional and behavioral differences due to a small number of loci. Other models with initial but limited divergence under allopatry and subsequent sympatry are also consistent with the observations.

The divergence of *D. pseudoobscura* and *D. p. bogotana*: In contrast to the case of *D. pseudoobscura* and *D. persimilis*, conclusions regarding natural selection and gene flow between *D. pseudoobscura* and *D. p. bogotana* must be fairly tenuous. These two taxa exhibit no fixed chromosomal inversion differences (Dobzhansky *et al.* 1963), and the only hybrids that exhibit fertility loss are males with *D. p. bogotana* mothers (Orr 1989a). Also, premating barriers to mating are absent (Prakash 1972) or very slight (Noor 1995a). Suppose that *D. pseudoobscura* and *D. p. bogotana* exchange genes regularly at a low rate and that natural selection against gene flow is not occurring. Then it is expected that there will be some divergence and few fixed differences, as is seen in the three loci studied here, as well as in allozyme data (Singh 1983) and chromosomal inversion data (Dobzhansky and Epling 1944). The fixed differences that are observed are mostly limited to the *Period* locus and may have been caused by a recent selective sweep near this gene in *D. p. bogotana* (Wang and Hey 1996). In general, an observation of divergence between populations, or candidate taxa, can be explained with an isolation model or with a model in which gene flow has been present at low levels indefinitely into the past. Also the high variance that we observed among loci is consistent with a model of long-term limited gene flow, with no set time for the onset of divergence (Wakeley 1996). In short, it seems possible that *D. pseudoobscura* and *D. p. bogotana* are not separate species but rather are linked by low levels of gene flow. At present the best evidence against this are the observations of relatively weak pre- and postmating barriers (Orr 1989a; Noor 1995a).

The results of this multilocus study on *D. pseudoobscura* and close relatives differ considerably from those on the *D. melanogaster* species complex. In a five locus study of variation within and between the four taxa of the *D. melanogaster* complex, one major finding was that different loci showed consistent levels of polymorphism and divergence among taxa (Hey and Kliman 1993; Kliman and Hey 1993; Hilton *et al.* 1994). An exception to this was that two loci in regions of low recombination exhibited less divergence than expected, possibly due to limited gene flow (Hilton *et al.* 1994). The most closely related species of the *D. melanogaster* complex are *D. simulans*, *D. mauritiana*, and *D. sechellia*, which probably diverged from one another ~0.75 mya. *D. simulans* (like *D. melanogaster*) is a cosmopolitan species that lived historically in continental Africa. *D. mauritiana* and *D. sechellia* are both island endemic species. Thus the basic finding of little or no gene flow and the divergence portraits that are similar across loci are consistent with the current geographical distribution and a simple allopatric speciation model. In contrast, the ranges of *D. pseudoobscura*, *D. persimilis*, and *D. p.*

bogotana are not nearly so disjunct or isolated. With a geography that is more permissive of gene flow, it is perhaps not surprising to find evidence of gene flow and to find that speciation has probably involved an interaction between natural selection and gene flow.

Speciation, gene flow, and geography: Over the range of *D. persimilis, D. pseudoobscura* and *D. persimilis* are sympatric. Also, according to original reports on range limits, neither species co-occurs with *D. p. bogotana*, though *D. pseudoobscura* has been collected as far south as Guatemala (DOBZHANSKY and EPLING 1944). This geographic pattern fits well with the observations in this paper and some recent reports on premating isolation among the species (NOOR 1995a,b). Estimated migration rates for *Period* and *Adh* are higher between the sympatric species *D. pseudoobscura* and *D. persimilis* than for the other species pairs, despite the wealth of evidence that *D. pseudoobscura* and *D. p. bogotana* are the most recently diverged taxa.

NOOR (1995b) found that in mate-choice experiments, female *D. pseudoobscura* from regions of sympatry with *D. persimilis* were more discriminatory against *D. persimilis* males than were female *D. pseudoobscura* from regions of allopatry. This is exactly the pattern expected if natural selection, in the form of partial postmating reproductive failure, is acting as a selective force for the evolution of mate discrimination. This reinforcement could only occur in regions where the two species are sympatric. The finding that *D. pseudoobscura* and *D. persimilis* have experienced considerable gene flow at the *Adh* region (estimated *Nm* levels for *Adh* are higher than between *D. pseudoobscura* and *D. p. bogotana*, Table 2) is consistent with this reinforcement scenario. If *D. pseudoobscura* and *D. persimilis* did not exchange genes in nature, then selection for stronger mate discrimination in regions of sympatry for mate choice could not be said to contribute to the speciation process (simply because speciation is complete if the taxa are not exchanging genes). However, if the taxa are engaged in moderate levels of gene flow, the species are not entirely reproductively isolated and speciation, in the sense of the biological species concept (MAYR 1942; DOBZHANSKY 1951), is not complete. Thus it seems quite plausible that natural selection for mate choice in regions of sympatry is contributing to the evolution of isolation of these taxa.

We thank MOHAMED NOOR and STEVE SCHAEFFER for helpful comments. This study was funded in part by a grant from the National Science Foundation to J.H. (DEB-9306625) and a postdoctoral fellowship to R.L.W. by the Bureau of Biological Research, Rutgers University. J.W. was funded by a National Institutes of Health National Research Service Award (GM-17745). STEVE SCHAEFFER was supported by the National Institutes of Health (GM-42472).

LITERATURE CITED

ANDERSON, W. W., F. J. AYALA and R. E. MICHOD, 1977 Chromosomal and allozymic diagnosis of three species of *Drosophila*. J. Hered. **68:** 71–74.

AQUADRO, C. F., A. L. WEAVER, S. W. SCHAEFFER and W. W. ANDERSON, 1991 Molecular evolution of inversions in *Drosophila pseudoobscura*: the amylase gene region. Proc. Natl. Acad. Sci. USA **99:** 305–309.

ASHBURNER, M., 1989 *Drosophila, a Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

AYALA, F. J., and T. DOBZHANSKY, 1974 A new subspecies of *Drosophila pseudoobscura* (Diptera:Drosophilidae). Pan-pac. Entomol. **50:** 211–219.

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.

BARRIO, E., A. LATORRE, A. MOYA and F. J. AYALA, 1992 Phylogenetic reconstruction of the *Drosophila obscura* group, on the basis of mitochondrial DNA. Mol. Biol. Evol. **9:** 621–635.

BLACKMAN, R. K., and M. MESELSON, 1986 Interspecific nucleotide sequence comparisons used to identify regulatory and structural features of the Drosophila *hsp82* gene. J. Mol. Biol. **188:** 499–515.

COYNE, J. A., and H. A. ORR, 1989 Two rules of speciation, pp. 180–207 in *Speciation and Its Consequences*, edited by D. OTTE and J. A. ENDLER. Sinauer, Sunderland, MA.

DOBZHANSKY, T., 1951 *Genetics and the Origin of Species*, 3rd ed. Columbia University Press, New York.

DOBZHANSKY, T., 1973 Is there gene exchange between *Drosophila pseudoobscura* and *Drosophila persimilis* in their natural habitats? Am. Nat. **107:** 312–314.

DOBZHANSKY, T., and T. EPLING, 1944 Taxonomy, geographic distribution and ecology of *Drosophila pseudoobscura* and its relatives, pp. 1–46 in *Contributions to the Genetics, Taxonomy, and Ecology of Drosophila Pseudoobscura and its Relatives*, vol. 554, edited by T. DOBZHANSKY and T. EPLING. Carnegie Institute of Washington, Washington, DC.

DOBZHANSKY, T., A. S. HUNTER, O. PAVLOVSKY, B. SPASSKY and B. WALLACE, 1963 Genetics of an isolated marginal population of *Drosophila pseudoobscura*. Genetics **48:** 91–103.

FELSENSTEIN, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 783–791.

FELSENSTEIN, J. 1993 PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

HALE, L. R., and A. T. BECKENBACH, 1985 Mitochondrial DNA variation in *Drosophila pseudoobscura* and related species in Pacific northwest populations. Can. J. Genet. Cytol. **27:** 357–364.

HEY, J., 1991 The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. Genetics **128:** 831–840.

HEY, J., 1994 Bridging phylogenetics and population genetics with gene tree models, pp. 435–449 in *Molecular Approaches to Ecology and Evolution*, edited by B. SCHIERWATER, B. STREIT, G. WAGNER and R. DESALLE. Birkhäuser-Verlag, Basel, Switzerland.

HEY, J., and R. M. KLIMAN, 1993 Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol. Biol. Evol. **10:** 804–822.

HEY, J., and J. WAKELEY, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

HILTON, H., R. M. KLIMAN and J. HEY, 1994 Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* complex. Evolution **48:** 1900–1913.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 819–829.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 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.

KLIMAN, R. M., and J. HEY, 1993 DNA sequence variation at the

period locus within and among species of the *Drosophila melanogaster* complex. Genetics **133:** 375–387.

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

MOORE, B. C., and C. E. TAYLOR, 1986 Drosophila of southern California III. Gene arrangements of *Drosophila persimilis*. J. Hered. **77:** 313–323.

MULLER, H. J., 1940 Bearings of the Drosophila work on systematics, pp. 185–268 in *The New Systematics*, edited by J. HUXLEY. Clarendon Press, Oxford.

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

NOOR, M. A., 1995a Incipient sexual isolation in *Drosophila pseudoobscura bogotana* Ayala & Dobzhansky (Diptera:Drosophilidae). Pan-pac. Entomol. **71:** 125–129.

NOOR, M. A., 1995b Speciation driven by natural selection in Drosophila. Nature **375:** 674–675.

ORR, H. A., 1987 Genetics of male and female sterility in hybrids of *Drosophila pseudoobscura* and *D. persimilis*. Genetics **116:** 555–563.

ORR, H. A., 1989a Genetics of sterility in hybrids between two subspecies of *Drosophila*. Evolution **43:** 180–189.

ORR, H. A., 1989b Localization of genes causing postzygotic isolation in two hybridizations involving *Drosophila pseudoobscura*. Heredity **63:** 231–237.

PATERSON, H. E. H., 1993 *Evolution and the Recognition Concept of Species: Collected Writings*. Johns Hopkins University Press, Baltimore.

POWELL, J. R., 1983 Interspecific cytoplasmic gene flow in the absence of nuclear gene flow: evidence from *Drosophila*. Proc. Natl. Acad. Sci. USA **80:** 492–495.

PRAKASH, S., 1969 Genic variation in a natural population of *Drosophila persimilis*. Genetics **62:** 784–788.

PRAKASH, S., 1972 Origin of reproductive isolation in the absence of apparent genetic differentiation in a geographic isolate of *Drosophila pseudoobscura*. Genetics **72:** 143–155.

RIZKI, M. T. M., 1951 Morphological differences between two sibling species, *Drosophila pseudoobscura* and *Drosophila persimilis*. Proc. Natl. Acad. Sci. USA **37:** 156–159.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SCHAEFFER, S. W., and C. F. AQUADRO, 1987 Nucleotide sequence of the *Adh* region of *Drosophila pseudoobscura*: evolutionary change and evidence for an ancient gene duplication. Genetics **117:** 61–73.

SCHAEFFER, S. W., and E. L. MILLER, 1991 Nucleotide sequence analysis of Adh genes estimates the time of geographic isolation of the Bogota population of *Drosophila pseudoobscura*. Proc. Natl. Acad. Sci. USA **88:** 6097–6101.

SCHAEFFER, S. W., and E. L. MILLER, 1992a Estimates of gene flow in *Drosophila pseudoobscura* determined from nucleotide sequence analysis of the alcohol dehydrogenase region. Genetics **132:** 471–480.

SCHAEFFER, S. W., and E. L. MILLER, 1992b 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, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135:** 541–552.

SEGARRA, C., and M. AGUADÉ, 1992 Molecular organization of the *X* chromosome in different species of the obscura group of Drosophila. Genetics **130:** 513–521.

SEGARRA, C., E. R. LOZOVSKAYA, G. RIBÓ, M. AGUADÉ and D. L. HARTL, 1995 P1 clones from *Drosophila melanogaster* as markers to study the chromosomal evolution of Muller's A element in two species of the *obscura* group of *Drosophila*. Chromosoma **104:** 129–136.

SEGARRA, C., G. RIBÓ and AGUADÉ, 1996 Differentiation of Muller's chromosomal elements D and E in the Obscura group of Drosophila. Genetics **144:** 139–146.

SHARP, P. M., and W. H. LI, 1989 On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28:** 398–402.

SINGH, R. S., 1983 Genetic differentiation for allozyme and fitness characters between mainland and Bogota populations of *Drosophila pseudoobscura*. Can. J. Genet. Cytol. **25:** 590–604.

SLATKIN, M., and W. P. MADDISON, 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:** 603–613.

STEINEMANN, M., W. PINSKER and D. SPERLICH, 1984 Chromosome homologies within the Drosophila obscura group probed by *in situ* hybridization. Chromosoma **91:** 46–53.

TAJIMA, F., 1993 Measurement of DNA polymorphism, pp. 37–59 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARKE. Sinauer Associates, Sunderland, MA.

TAKAHATA, N., and M. NEI, 1985 Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110:** 325–344.

TEMPLETON, A. R., 1989 The meaning of species and speciation: a genetic perspective, pp. 3–27 in *Speciation and Its Consequences*, edited by D. OTTE and J. A. ENDLER. Sinauer Associates, Sunderland, MA.

TEMPLETON, A. R., 1994 The role of molecular genetics in speciation studies, in *Molecular Approaches to Ecology and Evolution*, edited by B. SCHIERWATER, B. STREIT, G. WAGNER and R. DESALLE. Birkhäuser-Verlag, Basel, Switzerland.

THROCKMORTON, L. H., 1975 The phylogeny, ecology and geography of Drosophila, pp. 421–470 in *Handbook of Genetics, Volume 3, Invertebrates of Genetic Interest*, edited by R. C. KING. Plenum Publishing, New York.

WAKELEY, J., 1996 The variance of pairwise nucleotide differences in two populations with migration. Theor. Popul. Biol. **49:** 39–57.

WAKELEY, J., and J. HEY, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

WANG, R. L., and J. HEY, 1996 The speciation history of *Drosophila pseudoobscura* and close relatives: inferences from DNA sequence variation at the period locus. Genetics **144:** 1113–1126.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–275.

WRIGHT, S., 1931 Evolution in mendelian populations. Genetics **16:** 97–159.

Communicating editor: A. G. CLARK