Selection on *X*-linked Genes During Speciation in the *Drosophila athabasca* Complex

Michael J. Ford and Charles F. Aquadro

Section of Genetics and Development, Cornell University, Ithaca, New York 14853-2703

Manuscript received February 24, 1996
Accepted for publication June 26, 1996

ABSTRACT

We present the results of a restriction site survey of variation at five loci in *Drosophila athabasca*, complimenting a previous study of the *period* locus. There is considerably greater differentiation between the three semispecies of *D. athabasca* at the *period* locus and two other *X*-linked genes (*no-on-transient-A* and *E74A*) than at three autosomal genes (*Xdh*, *Adh* and *RC98*). Using a modification of the HKA test, which uses fixed differences between the semispecies and a test based on differences in *Fst* among loci, we show that the greater differentiation of the *X*-linked loci compared with the autosomal loci is inconsistent with a neutral model of molecular evolution. We explore several evolutionary scenarios by computer simulation, including differential migration of *X* and autosomal genes, very low levels of migration among the semispecies, selective sweeps, and background selection, and conclude that *X*-linked selective sweeps in at least two of the semispecies are the best explanation for the data. This evidence that natural selection acted on the *X*-chromosome suggests that another *X*-linked trait, mating song differences among the semispecies, may have been the target of selection.

THE role played by natural selection during speciation is a persistent and unanswered question in evolutionary biology (MAYNARD SMITH 1966; ENDLER 1977; LANDE and KIRKPATRICK 1988; BUTLIN 1989; HOWARD 1993; NOOR 1995). One tool that has recently been brought to bear on questions of selection and speciation is surveys of genetic variation at the DNA level between closely related species and populations (*e.g.*, HEY and KLIMAN 1993; KLIMAN and HEY 1993; FORD *et al.* 1994; HILTON *et al.* 1994). For instance, some forms of selection leave "footprints" in patterns of genetic variation that are clearly recognizable. One such form of selection is a "selective sweep," which occurs when a strongly advantageous allele at low frequency, such as a new mutation, rises rapidly in frequency due to its selective advantage and quickly replaces all other alleles at that locus in the population. Any neutral variants that are linked to the selected site will also be fixed, resulting in a reduction of variation in the region surrounding the selected site due to "hitchhiking" (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989). The size of the region with reduced variation will depend on both the strength of selection and the rate of recombination (KAPLAN *et al.* 1989; WIEHE and STEPHAN 1993). In this paper, we make use of this expected "footprint" of a selective sweep to address several questions of selection during speciation in *Drosophila athabasca*.

Corresponding author: Michael J. Ford, National Marine Fisheries Service, Northwest Fisheries Science Center, Coastal Zone and Estuarine Studies Division, 2725 Montlake Boulevard East, Seattle, WA 98112. E-mail: mford@nwfsc.noaa.gov

D. athabasca consists of three semispecies, called "western-northern," "eastern-A," and "eastern-B," which are distributed across most of northern North America (Figure 1). The three semispecies are reproductively isolated (premating) from each other to different degrees, with eastern-A being almost completely reproductively isolated from the other two as assessed by no-choice crosses in the laboratory (MILLER 1958; MILLER and WESTPHAL 1967; YOON 1991). Western-northern and eastern-B cross relatively easily in the laboratory under no choice conditions. One of the only phenotypic differences among the semispecies, and a likely candidate for the mechanism of their reproductive isolation, is a difference in male courtship song (MILLER *et al.* 1975; YOON 1991). A study of F_1, F_2 and backcross hybrids among the semispecies showed that two characters of male mating song that differ among the semispecies, interpulse interval and the relative abundance of low repetition rate and high repetition rate song, show patterns of segregation consistent with a strong effect of the *X* chromosome (YOON 1991). A large *X* chromosome effect has also been observed for similar song differences between *D. pseudoobscura* and *D. persimilis* (EWING 1969). The semispecies share allozyme polymorphism at most loci surveyed (JOHNSON 1978, 1985), and crosses among the semispecies produce viable, fertile progeny in the laboratory (MILLER 1958; YOON 1991; M. J. FORD and C. F. AQUADRO, personal observation). This, and the extremely low divergence seen at mtDNA, suggests that the behavioral differences that define the semispecies must have arisen very recently probably by the action of natural selection (YOON and AQUADRO 1994).

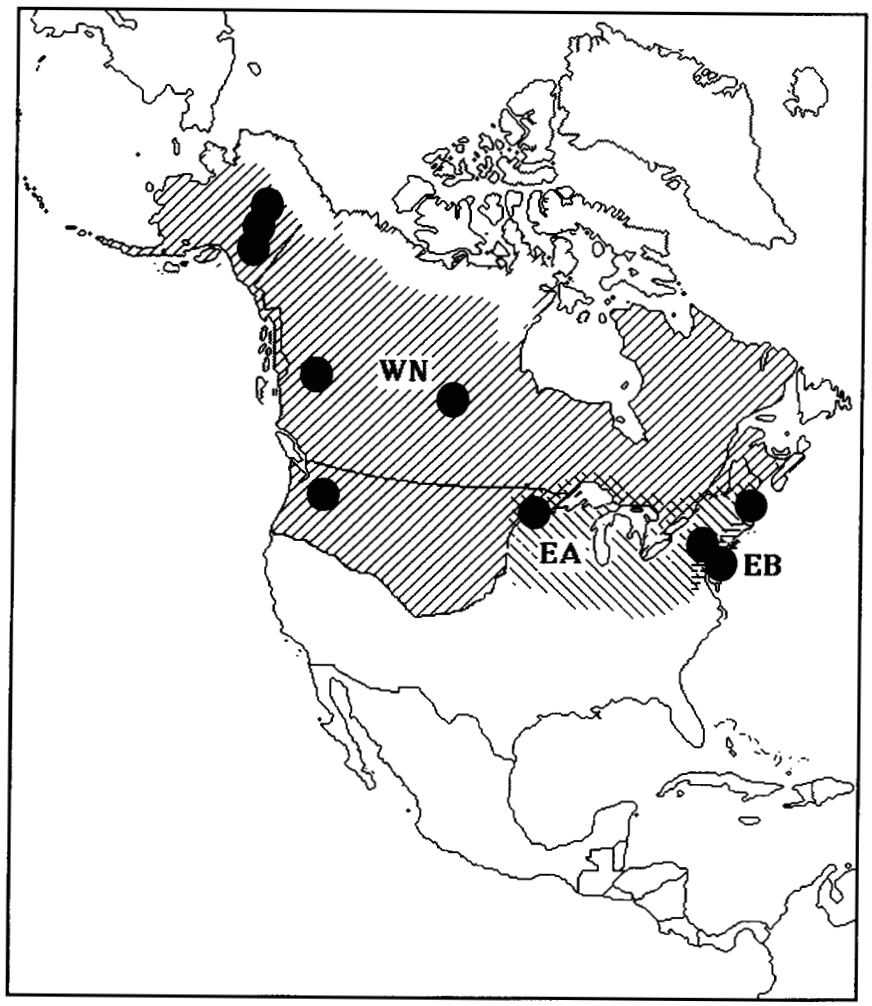

FIGURE 1.—Range map of *D. athabasca*. Sampling localities are represented by black dots.

In this report, we are interested in addressing the following questions: (1) In our study of the *period* locus (FORD *et al.* 1994), we found greater differentiation among the semispecies than had been previously reported for most allozyme loci (JOHNSON 1978, 1985). Does this reflect greater differentiation at DNA markers in general (*e.g.*, KARL and AVISE 1992; POGSON *et al.* 1994) or is this limited to *period* or just the *X* chromosome? (2) We also found evidence of a selective sweep at *period* in the western-northern semispecies. Is this limited to *period*, or is this effect seen at other *X*-linked loci or other candidate song genes? (3) Is there evidence of selection in the two eastern semispecies or just in western-northern?

To address these questions, we have surveyed restriction site variation at an additional five loci: *no-on-transient-A* (*nona*), *E74A*, *Xdh*, *Adh*, and *RC98*. *Nona*, like *period*, is a gene that affects male mating song when mutated (KULKARNI *et al.* 1988; JONES and RUBIN 1990; RENDAHL *et al.* 1992). *E74A* is a gene involved in ecdysone response (JONES *et al.* 1991). *Xdh* (*rosy* in *D. melanogaster*) is the gene encoding the enzyme xanthine dehydrogenase (COTE *et al.* 1986). *Adh* encodes the enzyme alcohol dehydrogenase (CHIA *et al.* 1985), and *RC98* is a single copy random clone. Like *period*, *nona* is on the *X* chromosome in *D. melanogaster* and, based on MULLER's elements (cited in ASHBURNER 1989b, p. 23), is expected to be on the short arm of the *X* in *D. athabasca*. *E74A* is on 3L in *D. melanogaster*, which places it tentatively on the long arm of the *X* in *D. athabasca* (ASHBURNER 1989b, p. 23). *Xdh* and *Adh* are both autosomal in *D. melanogaster* and are expected to be autosomal in *D. athabasca* (the *E* and *B* chromosomes, respectively). *RC98* is autosomal in *D. athabasca* (see RESULTS).

MATERIALS AND METHODS

Sampling of individuals: A total of 41 western-northern, 19 eastern-A, and 18 eastern-B individuals were used in this study, although not every individual was scored for variation at every locus. All individuals were male and were either sampled directly from the wild or taken from different isofemale lines (Figure 2). Seven of the western-northern individuals were used in a previous study of the *period* locus (FORD *et al.* 1994): ac-1, ac-5, cc-4, cc-6, s-16, s-4, and s-12. Six eastern-A individuals (sf-44, gsc-5, gsc-30, gsc-33, gsc-2 and gsc-23) and 10 eastern-B individuals (sf-49, sf-39, sf-12, sf-56, sf-35, sf-9, sf-69, gsc-20, gsc-46, and sf-62) came from the same isofemale lines as those individuals with the same names surveyed for the *period*

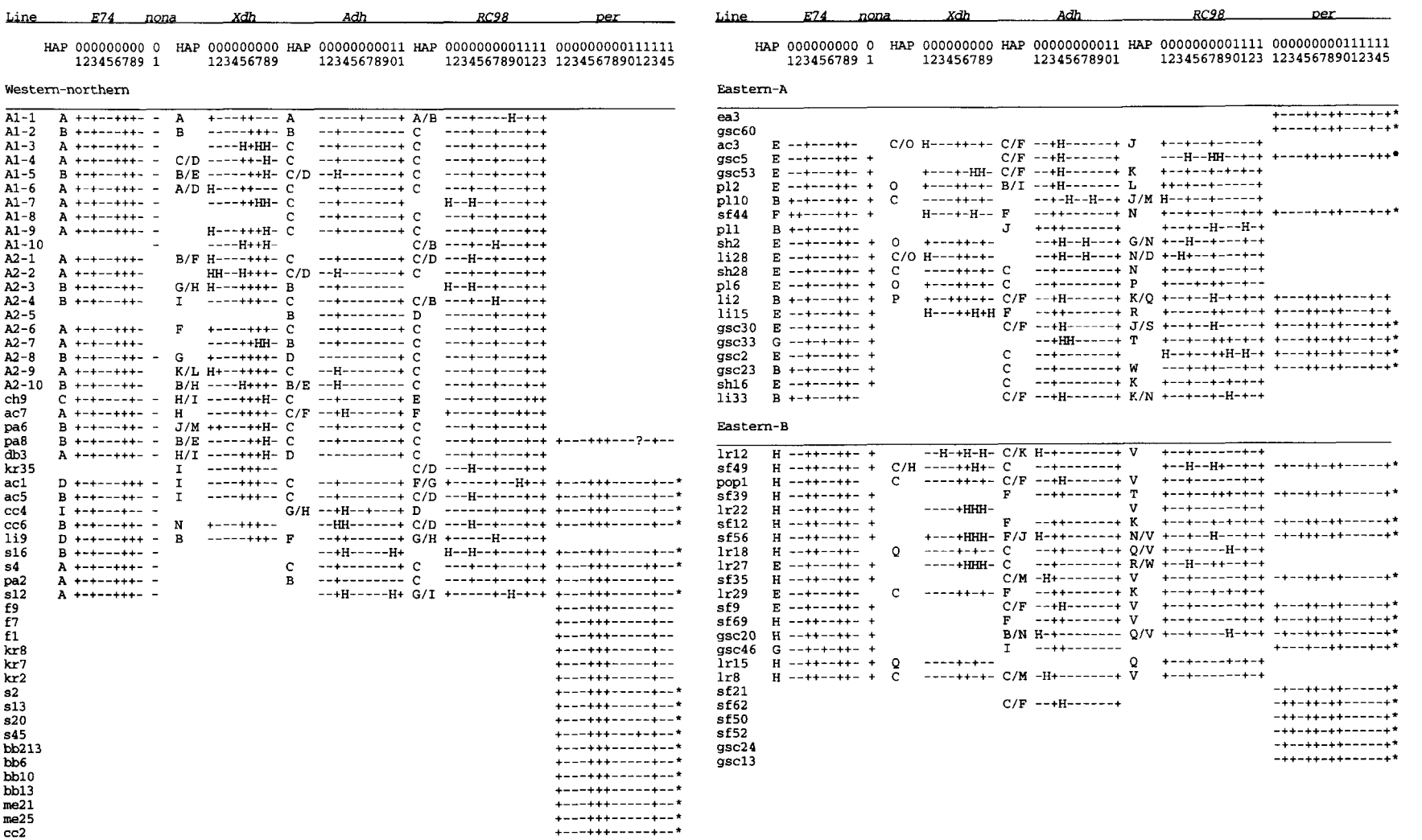

FIGURE 2.—Restriction site genotypes. A + indicates the presence of a site, a − an absence, an H indicates that the individual is heterozygous at the site. HAP refers to the multisite haplotype for each individual, excluding individuals heterozygous at more than one site. Blank areas indicate that the individual was not scored for that locus. The restriction enzyme for each site is as follows: *E74A:* 1–2) *Alu*I, 3) *Msp*I, 4–5) *Taq*I, 6) *Rsa*I, 7–9) *Dde*I. *nona:* 1) *Scrf*I. *Xdh:* 5′ fragment: 1) *Sau*3AI. 3′ fragment: 2–4) *Rsa*I, 5–6) *Scr*FI, 7) *Dde*I, 8) *Hae*III, 9) *Alu*I. *Adh:* 1–2) *Scr*FI, 3–5) *Rsa*I, 6–10) *Dde*I, 11) *Alu*I. *RC98:* 1–3) *Hin*f1, 2) *Scr*FI, 3) *Msp*I, 4) *Dde*I, 5–6) *Sau*3AI, 7–8) *Hae*III, 9) *Alu*I, 10–11) *Rsa*I. The locations and collection dates of the individuals are as follows: A1 and f, Fairbanks, AK, 1993; A2- and kr, Kayakuk River, AK, 1993; ch9, Cassiar Highway, B.C., 1993; ac, Acadia, ME, 1992; pa, Prince Albert Provincial Park, Saskatchewan, 1993; db3, Devil's Backbone State Park, IA, 1993; cc, Cobscook Bay State Park, ME, 1993; li, Lake Itaska State Park, MN, 1993; s, Snoqualmie Pass, WA, 1989; bb, Burnaby, B.C., 1989; me, Blue Hill, ME, 1989; gsc, Poughkeepsie, NY, 1989; ea, Brooktondale, NY, 1989; pl and sh, Ithaca, NY, 1992; sf, Princeton, NJ, 1989; lr and pop, Arlington, VA, 1993. *Data from FORD *et al.* (1994).

locus. All other individuals were collected subsequently for this study. Flies were collected over a large part of the species's range (Figure 1). All individuals except those mentioned above were collected during the summer of 1993, except for "pl" and "sh" individuals, which were collected during the summer of 1992. Lines were assigned to semispecies on the basis of male mating song and/or capture location.

Method of surveying variation: The basic strategy for surveying genetic variation at the five loci of this study is the same as for our survey of variation at *period* (FORD *et al.* 1994). In short, we used the polymerase chain reaction (PCR) to amplify a fragment or fragments of a locus from genomic DNA obtained from single male flies (protocol 48 in ASHBURNER 1989a). At some of the loci (see below), we performed two amplifications; a primary amplification from genomic DNA and then a secondary amplification from a dilution of the primary amplification using primers internal to the original primers. The resulting amplified fragments were then aliquoted into either eight or nine aliquots, and each aliquot was cut with a different 4-bp recognition restriction enzyme. The resulting fragments were run out on either 2% agarose or 8% polyacrylamide gels. The bands were visualized by staining with ethidium bromide and then photographed on an ultraviolet light box. The restriction enzymes were the same as were used for the study of *period*: *Hin*fI, *Rsa*I, *Taq*I, *Msp*I, *Sau*3A1, *Scr*FI, *Hae*III, *Dde*I, and *Alu*I.

Lacking complete *D. athabasca* sequences for the loci surveyed in this study, we could not readily infer the locations of the restriction sites relative to each other. This does not affect the analysis or interpretation of our results, since we are only interested in comparing levels and patterns of variation among different loci not among regions within each locus. The size of a particular band on a gel can differ among individuals either because of a gain or loss of a restriction site or because of an insertion or deletion (indel). The two can usually be distinguished by comparing the sizes and number of bands among individuals. In some cases, it was difficult to tell from an individual enzyme whether a particular band shift was due to an indel or a change in a restriction site. These cases could usually be resolved by examining the fragment patterns produced by the other enzymes, since an indel is usually detectable in more than one enzyme, whereas a change in a particular restriction site is limited to a single enzyme. Very small indels (<5 bp) are not distinguishable from base substitutions with this method.

***Xdh*:** A plasmid clone encompassing all of the putative *D. athabasca Xdh* structural gene was kindly provided by C. K. YOON (YOON 1991). We confirmed that this clone did indeed contain DNA homologous to *Xdh* by sequencing ~300 bp (data not shown) using primers kindly provided by M. RILEY designed from the *D. pseudoobscura Xdh* sequence (RILEY 1989). Based on the *D. athabasca* sequences, we designed PCR primers to amplify an ~2.5-kb fragment from near the 5′ end of the first exon to approximately the middle of the second exon (nucleotides 920–2880 in RILEY 1989) and an ~1.6-kb fragment from near the 5′ end of the third exon to beyond the end of the 3′ end of the fifth (and last) exon (starts at nucleotide 5214 of RILEY 1989).

***Nona*:** We screened at low stringency ~52,000 plaques of a *D. athabasca* (western-northern) EMBL4 phage library (kindly provided by C. K. YOON) with a fragment of *D. melanogaster nona* DNA cut from a plasmid kindly provided by J. C. HALL (clone pHΔB235R11 from JONES and RUBIN 1990). We isolated several putatively positive plaques, and screened secondary and tertiary libraries created from these initial positives with the same probe. From the tertiary library, we picked several positive single plaques for further analysis. From these, we extracted DNA, which we then cut with *Eco*RI, *Sal*I, *Pst*I and *Hin*dIII. The resulting fragments were electrophoresed through 1.8% agarose gels in TBE buffer and then probed at high stringency with a 3-kb *Pst*I fragment containing the *D. melanogaster nona* coding sequence isolated from the same clone used to screen the library. This probe hybridized to 1.6- and 6.0-kb *Sal*I fragments, both of which we sub-cloned into a plasmid vector (Bluescript II KS+). By sequencing the ends of these two clones, we determined that the clones contained sequences homologous to *nona*.

Based on the alignment of the sequences with *D. melanogaster nona* (JONES and RUBIN 1990), the two fragments are contiguous and share the *Sal*I site at nucleotide 4436–4440 present in the *D. melanogaster nona* sequence (JONES and RUBIN 1990). We subcloned additional fragments, and between sequencing the ends of these fragments and generating sequence from primers designed from the newly generated *D. athabasca* sequence, we determined the sequence of most of the second, third and fourth exons, and most of the first, second, third and fourth introns. Most of the exon sequence of the *D. athabasca nona* gene is ~80% similar to the *D. melanogaster* homologue and was easily aligned by eye. None of the intron sequences, however, were alignable. The *D. athabasca nona* sequences are available from GenBank (accession numbers U37550-U37556).

Based on the *D. athabasca* sequence, we designed PCR primers that amplified an ~1.4-kb fragment running from the middle of the second intron to the middle of the fourth intron. It is this fragment that we used for the restriction site survey in this report. We also examined variation in an ~1-kb fragment immediately 5′ of the surveyed fragment. In many individuals, however, PCR reactions amplifying this 5′ fragment produced two bands of slightly differing size. When these reaction products were partially sequenced, two distinct sequence ladders were found, differing at ~5% of the bases sequenced (data not shown). We interpret this as evidence for a duplication of *nona* in *D. athabasca*. We determined that the variation seen in the fragment studied for this report, however, is allelic and *X*-linked (see below).

***Adh*:** We designed PCR primers to amplify an ~1.4-kb fragment encompassing most of the *Adh* and *Adh-related* structural genes (bases 1047–2557 in MARFANY and GONZALES-DUARTE 1992) from conserved regions in an alignment of six Drosophilid *Adh* and *Adh-related* genes (*D. Madeirensis, D. Teissieri, D. Guanche, D. Immigrans, D. Subobscura,* and *D. Melanogaster,* Genbank accession numbers obtained from RUSSO *et al.* 1995). We sequenced ~200 bp of the resulting fragment and determined that this product was indeed homologous to *Adh* (data not shown). From these sequences, we designed internal primers that were used for a secondary amplification and subsequent restriction enzyme digestions.

***E74A*:** Primers directly on either side of the seventh intron were designed from an alignment of *D. melanogaster, D. pseudoobscura,* and *D. virilis* cDNA sequences (JONES *et al.* 1991). These were used in a PCR reaction to amplify an ~1.3-kb fragment, presumably homologous to the *E74A* seventh intron. The ends of this fragment were sequenced, and new PCR primers were designed from the sequences. In the subsequent population survey, the original primers were used in a primary PCR reaction from genomic DNA, followed by a secondary PCR reaction with the internal primers.

***RC98*:** We extracted genomic DNA from ~30 *D. athabasca* (eastern-A) adults. Approximately 200 ng of this DNA was cut with *Eco*RI and ligated into the *Eco*RI site of Bluescript II KS+ (Stratagene), whose ends had been dephosphorylated to prevent self religation (SAMBROOK *et al.* 1989). A small amount of the ligation mixture was used to transform competent *Escherichia coli* (DH5-α, Gibco-BRL), which were then plated on LB plates containing ampicillin, X-gal and IPTG

(SAMBROOK *et al.* 1989). DNA was prepared from 100 randomly picked white colonies and digested with *Eco*RI to determine the size(s) of the insert(s). Twenty clones contained apparently single inserts of the desired size (between 1 and 3 kb). We sequenced the ends of most of these, and designed PCR primers for five of them. Of these, only one, *RC98,* was not obviously present in multiple copies in the *D. athabasca* genome. For this clone, we designed additional internal PCR primers for use in a secondary reaction (see above). Based on the restriction patterns present in natural populations and from a specific cross (see below), this sequence appears to be single copy.

Statistical analysis and computer simulations: The individuals used in this study fall into two categories: those that were sampled directly from the wild and those that were sampled from different isofemale lines that had been maintained in the laboratory for ≤2 yr. For the *X*-linked genes, this distinction is not relevant, since each male carries only a single *X* chromosome. For the autosomal loci, however, one must make a distinction between wild-caught flies and flies sampled from isofemale lines. This is because when an individual sampled from an isofemale line is homozygous at a locus, it is not possible to know if the two alleles at the locus are recently identical by descent or really represent two alleles sampled from nature. This is not a problem for heterozygous individuals, since barring a mutation while in culture, if the two alleles are different, they must represent two distinct alleles sampled from nature. Although there are several possible ways around this problem, we chose to simply analyze the data twice, once assuming that every homozygous individual at a particular locus counted as two alleles and once assuming that such individuals counted as a single allele. The difference in measures of variation such as π (NEI 1987) and θ (WATTERSON 1975) among these two estimates were very small and for subsequent analysis we simply used the average of the two estimates.

Statistical tests of neutrality: We use two new statistical tests to test for departures from a strict neutral model. The first is a modification of HUDSON *et al.*'s test (1987, the HKA test), which we refer to as the fixed sites (FS) test. In this test, we use the number of fixed differences (HEY 1991) as a measure of divergence, as opposed to picking a random allele from each population (HUDSON *et al.* 1987). Unlike the latter measure of divergence, the number of fixed differences is affected by selection at linked sites (HEY 1991), and therefore the FS test can be more powerful than the HKA test when used on closely related populations that are not expected to have any fixed differences under neutrality (M. J. FORD and C. F. AQUADRO, unpublished data). Like the HKA test, the FS test works by estimating the neutral parameter θ for each locus, the divergence time, T, measured in units of $2N$ generations, between the two populations involved in the test, and a parameter f, which is the factor by which the size of the first population needs to be multiplied to get the size of the second population. These parameters were estimated from the number of segregating sites at each locus in each population and the average number of differences between each pair of populations by the method of HUDSON *et al.* (1987), modified to take into account the difference in population size between *X*-linked and autosomal loci (BEGUN and AQUADRO 1991). These estimates are then used to generate the expected values and variances of the number of segregating sites at each locus within each population and the number of fixed differences at each locus among populations. The latter were calculated according the formulas derived by HEY (1991), modified to take into account differences in population size between populations and loci. In all cases, we assume that *X*-linked loci have an effective population size three-quarters that of autosomal loci (see BEGUN and AQUADRO 1991). The FS test statistic is constructed from the observed and expected values of segregating sites and fixed differences in the same way as the HKA test statistic (HUDSON *et al.* 1987). If the number of segregating sites and the number of fixed differences are distributed approximately normally and independently, the test statistic will be approximately chi-square distributed with $2L$-1 degrees of freedom, where L is the number of loci used in the test (HUDSON *et al.* 1987). The actual distribution of the FS test statistic for the case of $L = 2$ is fully explored in another paper (M. J. FORD and C. F. AQUADRO, unpublished data). Here, we note that although the FS statistic is not chi-square distributed when T is small, the critical values derived from a chi-square distribution are nonetheless approximately correct.

We call the second new test of neutrality the DFst test, and this test is based on the difference in *Fst* between two loci or the average difference between two groups of loci. We estimate *Fst* as D_{net}/D_{ab}, where D_{ab} is average pairwise divergence between two populations and D_{net} is D_{ab} minus k, the average number of pairwise differences within each population (NEI 1987). This is equivalent to the estimate of *Fst* used by HUDSON *et al.* (1992). To determine the distribution of the DFst statistic under several different evolutionary scenarios, we used a computer to simulate the coalescent process in a subdivided population with and without migration and with and without selection. This was done in a way similar to that described in HUDSON (1990), whereby the simulation moves backward in time with coalescent and migration events distributed exponentially. Only alleles in the same population can coalesce, and when the simulation reaches the predetermined time, T, of population splitting, the coalescent process continues in the single ancestral population. Assumptions about population size are identical to those described above for the FS test. The parameters (in addition to the sample sizes) needed to fully simulate the neutral coalescent process, θ for each locus, T, and f, were estimated in exactly the same way as for the FS and HKA tests (HUDSON *et al.* 1987). Differences in effective population size between loci and populations were taken into account by modifying the exponential coalescence and migration parameters by the appropriate factors, again assuming that the effective population size of an *X*-linked locus is three-quarters of the effective population size of an autosomal locus. The P value for the test is calculated as the proportion of the simulated test statistics greater than or equal to the observed test statistic. Since we had an *a priori* expectation that the genes on the *X* chromosome might have a higher *Fst* (FORD *et al.* 1994), we performed one tailed tests.

RESULTS

***Nona* and *E74A* are *X*-linked:** Chromosomal banding homologies (ASHBURNER 1989b, p. 23) predict that *E74A* is on the long arm of the *X*-chromosome and *period* and *nona* are on the short arm. A western-northern female and eastern-B male differing at a *Scr*FI site in *nona* and a *Taq*I site in *E74A* were crossed. All five male F_1 progeny scored had the female parent's genotype at these two sites, and all five female F_1 progeny scored were heterozygous for both parental genotypes at both sites, confirming that these two loci are *X*-linked in *D. athabasca*. F_2 progeny from this cross were scored for the two sites above, plus site *Rsa*I 668 in the 5′ *period* fragment (FORD *et al.* 1994). Out of 28 F_2 progeny scored, we found a single recombinant, between *E74A* and (*period, nona*). Western-northern

and eastern-B are fixed for a different series of overlapping paracentric inversions on both arms of the *X*-chromosome, but there are no pericentric inversions among the semispecies (MILLER and VOELKER 1969a,b). Although the number of meioses scored was small, the fact that we did observe a recombination event between *E74A* and (*period, nona*) but none between *period* and *nona* is consistent with these patterns of chromosomal inversion variation among the semispecies. It is important to note that the level of recombination we have measured is among semispecies and does not necessarily reflect the level of recombination within any of the semispecies.

Patterns of restriction site variation: All five of the loci surveyed were variable in our sample, but the five loci varied considerably in both the overall level of variation and in how the variation was partitioned within and among the semispecies. In particular, although there is no tendency for the *X*-linked loci to be either more or less variable than the autosomal loci, there is a clear trend for the *X*-linked loci to have much of their variation partitioned among the semispecies, whereas the autosomal loci tend to have most of their variation partitioned within the semispecies.

Figure 2 summarizes the polymorphic sites for each locus, as well as, for completeness, the *period* locus (FORD *et al.* 1994), including 12 individuals scored for *period* subsequently to our original report. We calculated the effective number of sites surveyed by the method of HUDSON (1982). The estimates of the number of nucleotides surveyed ranged from ~150 (*Adh*) to >550 (*Xdh*, Table 1). By normalizing measures of variation by the effective number of sites surveyed, levels of variation can be directly compared across loci.

Levels of variation per nucleotide site vary considerably across loci (Table 1, Figure 3). Considering the sample as a whole, estimates of π (the average number of pairwise differences per site, NEI 1987) vary from a high of nearly 0.01 at *E74A* and *RC98*, to a low of just over 0.002 at *nona* (Table 1, Figure 3). Again considering the sample as a whole, there is no tendency for the *X*-linked loci to be either more or less variable than autosomal loci. There are, however, considerable differences among loci in the absolute and relative levels of variability when considering each semispecies separately (Figure 3). In this case, there is a tendency for the three *X*-linked loci to be less variable than the three autosomal loci. There are several potential causes for this, which we discuss below.

Like estimates of polymorphism, estimates of divergence among semispecies vary considerably across loci. The average number of pairwise differences per site among semispecies (D_{ab}) is greatest at *E74A*, and least at *Xdh*, and there is no tendency for either the autosomal or the *X*-linked genes to be more divergent (Table 2, Figure 4, dark lines). There is, however, a clear trend for the three *X*-linked loci to have a greater level of net divergence (Table 2, Figure 4, wide gray lines). This is reflected in the estimates of *Fst:* the average *Fst* of the three *X*-linked loci is 0.765 compared with 0.196 for the three autosomal loci (Table 2). This can also be seen when considering the number of fixed differences among semispecies at each locus. *Nona* has only a single segregating site in the form of a fixed difference between western-northern and the two eastern semispecies. *Period* has four fixed differences between western-northern and the eastern semispecies, and *E74A* has a single fixed difference between western-northern and eastern-B. In contrast, although they are on average just as variable as the *X*-linked loci, the three autosomal loci do not have any fixed differences among the semispecies.

We can use the FS and DFst tests to test the hypothesis that the differences in how variation is partioned within and among the semispecies can be explained under neutrality, taking into account the expected difference in population size between *X*-linked and autosomal loci. Table 3 contains the estimates of θ for each locus as well as T and f for each possible pairing of the *D. athabasca* semispecies. Of the three pairwise six-locus FS tests (Table 4), the test considering western-northern and eastern-A is significant assuming a chi-square distribution ($\chi^2 = 50.89$, d.f. = 11, $P < 0.001$). By far, the largest contribution to the test statistic comes from the presence of fixed differences at *period* and *nona*. The fixed difference at *nona* alone accounts for 56% of the value of the test statistic. Given the estimate of θ for *nona* and divergence time between western-northern and eastern-A, there is a very low probability of observing a fixed difference. The four fixed differences at *period* are also very improbable, accounting for 36% of the test statistic. Normal HKA tests using the number of differences between two randomly chosen alleles from each population as a measure of divergence failed to reject neutrality in any of the six-locus comparisons, although several two-locus comparisons were significant (data not shown). Although there are just as many fixed differences at *nona* and *period* between western-northern and eastern-B as between western-northern and eastern-A, the test between western-northern and eastern-B is not significant. This appears to be because of the larger value of D_{ab} between these two semispecies (Table 2, Figure 4), which results in a larger estimate of T for which fixed differences are more likely than in the case of western-northern and eastern-A. All of the two-locus FS tests among these semispecies involving *nona* and any of the autosomal loci did, however, reject neutrality (data not shown). Eastern-A and eastern-B have no fixed differences between them, so it is not surprising that the FS test failed to reject neutrality in this comparison.

We also conducted DFst tests for each comparison among the semispecies. The average values of DFst between the *X*-linked and the autosomal loci are reported in Table 5. Using computer simulations (see MATERIALS

TABLE 1

Polymorphism for all loci

Locus	*n*	*S*	*k*	Sites	π	θ
E74						
Total	68	9	1.428	148	0.00965	0.01270
WN	32	3	0.760	172	0.00442	0.00433
EB	17	2	0.500	176	0.00284	0.00336
EA	19	3	0.771	172	0.00449	0.00499
nona[a]						
Total	45	1	0.454	188	0.00241	0.00121
WN	15	0	0	192	0	0
EB	14	0	0	192	0	0
EA	16	0	0	192	0	0
period[b]						
Total	51	15	3.600	740	0.004865	0.004505
WN	26	2	0.153	792	0.000195	0.000662
EB	15	6	1.714	788	0.002175	0.002733
EA	10	3	0.755	776	0.000973	0.001171
period, nona, E74						
Total	26	22	5.587	1076	0.005193	0.005305
WN	9	4	1.333	1156	0.001153	0.001223
EB	9	7	2.555	1156	0.002211	0.002140
EA	8	7	2.035	1140	0.001786	0.002259
RC98						
Total	97 (140)[c]	13	1.991	204	0.009762	0.01212
WN	48 (70)	6	1.093	232	0.004712	0.00559
EB	21 (32)	4	1.389	240	0.005791	0.00437
EA	27 (38)	11	2.432	212	0.011475	0.01287
Xdh						
Total	79 (98)	9	2.128	508	0.004191	0.003503
WN	47 (56)	5	1.811	524	0.003458	0.002109
EB	15 (20)	5	1.864	524	0.003558	0.002764
EA	16 (22)	6	1.677	520	0.003225	0.003270
Adh						
Total	96 (136)	11	1.105	140	0.007898	0.014949
WN	42 (66)	6	0.982	160	0.006143	0.008261
EB	24 (32)	5	1.187	164	0.007240	0.007793
EA	30 (38)	6	0.995	160	0.006221	0.009129

Measures of polymorphism for the five genes in this study as well the *period* gene. *n* is the number of alleles sampled, counting homozygous individuals as one allele and heterozygous individuals as two alleles; *S* is the number of polymorphic restriction sites; *k* is the average number of restriction site differences between alleles (NEI 1987); sites is an estimate of the number of nucleotides surveyed (HUDSON 1982); π is *k* divided by sites; and θ is WATTERSON's (1975) measure of diversity per site.

[a] It is not clear whether the single polymorphic site at *nona* is a nucleotide substitution or a length variant. In either case, the infinite sites model of molecular evolution is not likely to be violated.

[b] The data from the *period* gene includes those individuals surveyed by FORD *et al.* (1994) as well as an additional nine western-northern individuals collected from north of Fairbanks, AK (see METHODS).

[c] Sample size counting both heterozygotes and homozygotes as two individuals. The values of *k*, π and θ are based on the average of those obtained counting homozygotes as one allele and those obtained counting homozygotes as two alleles. The differences between these two estimates were in all cases small.

AND METHODS), we determined the probability of observing DFst values this large or larger under several different evolutionary scenarios. The first hypothesis we tested is a model of strict neutrality, with no migration between the semispecies and all population parameters (θs, *T*, *f*) estimated from all the loci considered together (Table 3). This hypothesis takes into account the differences in population size between an *X*-linked and an autosomal locus (see MATERIALS AND METHODS) and was rejected in all three comparisons (H_0 in Table 5), indicating the differences in *Fst* between the two groups of loci are too great to be consistent with a strict neutral model with no migration. The fact that all three pairwise DFst tests reject the null hypothesis implies that a nonneutral process is occurring in at least two of the semispecies.

One way to generate different amounts of differentiation at different sets of loci is different levels of gene flow at the different loci. This could occur if hybrids among the semispecies are unfit and if the genes that

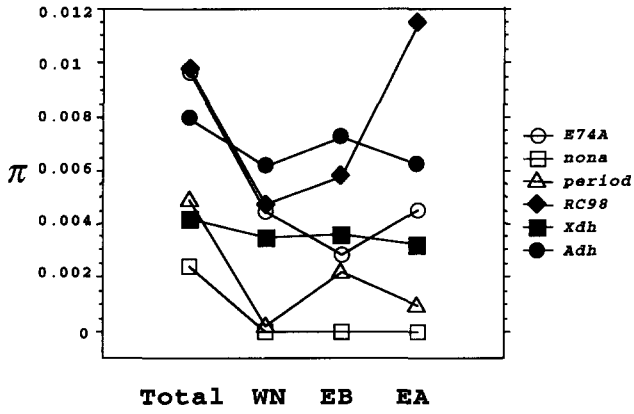

FIGURE 3.—Estimates of nucleotide diversity (π) for each locus in each semispecies and in the entire sample.

contribute to hybrid maladaptation are located on the *X* chromosome (see *e.g.*, HILTON *et al.* 1994). Under this scenario, the level of divergence of the *X*-linked loci better reflects the time of divergence between the semispecies and the relative lack of differentiation on the autosomal loci better reflects the effects of gene flow at those loci. This hypothesis was tested by examining the distribution of the DFst statistic under the assumption of no *X*-linked gene flow and a level of gene flow on the autosomes estimated from the average autosomal *Fst* values. The gene flow parameter is $M = 4Nm$, where m is the proportion of the population that is migrants. Under the island model of migration (WRIGHT 1978), an estimate of M is $1/Fst - 1$ (HUDSON *et al.* 1992). Using this estimate of M as the migration parameter for the autosomal loci (estimated separately for each of the three tests) and assuming a divergence time estimated from the *X*-linked loci, we simulated the DFst statistic for the six loci as before. Unlike the strictly neutral, no migration model, this hypothesis is not rejected for the comparisons between western-northern and either of the two eastern semispecies (H_1, Table 5). The hypothesis is rejected, however, for the comparison between the two eastern semispecies. This appears to be because the estimated divergence time between these two groups, even when estimated from the *X*-linked genes alone, is not large enough to generate the value of *Fst* seen at the *X*-linked loci.

A third scenario is that one or the other of the populations in each comparison has recently experienced a selective sweep on the *X* chromosome that has affected patterns of variation at each of the three *X*-linked loci through hitchhiking, increasing net divergence at the expense of polymorphism. We tested this hypothesis by

TABLE 2

Total and net divergence and *Fst*

	D_{ab}	D_{net}	D_{fix}	*Fst*
E74A				
WN-EA	1.498355 (0.008711)	0.732350 (0.004258)	0 (0)	0.489
WN-EB	2.470779 (0.014200)	1.849739 (0.010631)	1 (0.005747)	0.749
EA-EB	1.291022 (0.007420)	0.655057 (0.003765)	0 (0)	0.507
nona				
WN-EA	1.0 (0.00500)	1.0 (0.00500)	1 (0.00500)	1.0
WN-EB	1.0 (0.00500)	1.0 (0.00500)	1 (0.00500)	1.0
EA-EB	0.0 (0.00)	0.0 (0.00)	0 (0.00)	—
period				
WN-EA	4.476923 (0.005710)	4.022222 (0.005130)	4 (0.001276)	0.899
WN-EB	6.876923 (0.008705)	5.942857 (0.007522)	4 (0.008861)	0.864
EA-EB	3.200000 (0.004092)	1.965079 (0.002513)	0 (0)	0.614
period, nona, E74				
WN-EA	7.152778 (0.006231)	5.468254 (0.004763)	5 (0.004355)	0.764
WN-EB	9.888889 (0.008641)	7.944444 (0.006872)	6 (0.005190)	0.803
EA-EB	4.208333 (0.003666)	1.912698 (0.001666)	0 (0)	0.454
RC98				
WN-EA	2.347771 (0.01057)	0.584761 (0.002587)	0 (0)	0.249
WN-EB	2.339285 (0.009912)	1.147709 (0.004863)	0 (0)	0.490
EA-EB	2.217305 (0.009811)	0.306010 (0.001378)	0 (0)	0.138
Xdh				
WN-EA	2.514852 (0.004799)	0.770378 (0.001476)	0 (0)	0.306
WN-EB	2.257257 (0.004308)	0.418843 (0.000799)	0 (0)	0.186
EA-EB	2.097348 (0.004003)	0.326552 (0.000626)	0 (0)	0.156
Adh				
WN-EA	1.109706 (0.006936)	0.120598 (0.000754)	0 (0)	0.109
WN-EB	1.251555 (0.007726)	0.166436 (0.001027)	0 (0)	0.133
EA-EB	1.086805 (0.006709)	−0.004508 (−0.000028)	0 (0)	0

Values in parentheses are n per site.

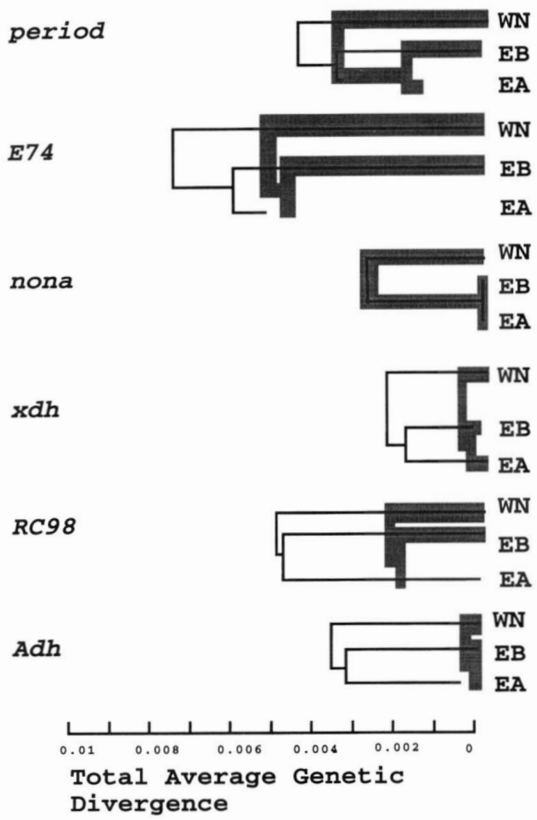

FIGURE 4.—Average and net divergence. Dark black lines are D_{ab}; light gray lines are D_{net}.

conducting simulations with the divergence time between the semispecies estimated from the autosomal loci alone and simulating a strong selective sweep at each *X*-linked locus at time $T_s = 0.04N$ generations in the past. Under this hypothesis, none of the three comparisons is significant, although all are close to being significant. The time of the sweep was chosen because it is slightly less than the estimated divergence time between the two eastern semispecies. We caution, however, that this choice is truly arbitrary, and this test should be viewed more as an exploration of a scenario than a true test of a hypothesis. Other selective scenarios with more recent sweep times or a sweep in each population are likely to fit the data better. Also note that although we simulated a sweep at each *X*-linked locus separately, this is the same as assuming a sweep at any one locus with no recombination among loci.

WAKELEY (1996) has recently shown that at migration/drift equilibrium low levels of migration can potentially greatly inflate the variance of a number of pairwise measures of polymorphism and divergence (k, D_{ab} and D_{net}) compared with that expected under complete isolation. This can lead to the rejection of the neutral model in several tests of neutrality even though no selection is present (M. J. FORD and C. F. AQUADRO, unpublished data). To test for this effect on the DFst statistic, we conducted simulations based on the parameters estimated for all the loci, but assuming migration/drift equilibrium instead of complete isolation. As WAKELEY (1996) has pointed out, this level of migration is equal to $1/T$ assuming complete isolation. To conduct the simulations, we set the divergence time among the populations equal to $200N$ generations (large enough to ensure that the assumption of migration/drift equilibrium is met), and M equal to the reciprocal of the estimates of T in Table 3 ($M = 3.33$, 0.55 and 0.88 for the EA-EB, WN-EB and WN-EA tests, respectively). We found that this level of migration cannot explain the large values DFst for any of the three comparisons (H_3, Table 5). We also conducted simulations with a very low migration rate ($M = 0.01$) but with divergence times estimated from the data. This very low level of migration also cannot explain the high DFst values (data not shown).

LEWONTIN and KRAKAUER (1973) also proposed a test

TABLE 3

Estimates of parameters for the HKA FS, and DFst tests

	$\hat{\theta}$ *X*-linked loci			$\hat{\theta}$ Autosomal loci					
	E74	*nona*	*period*	*RC98*	*Xdh*	*Adh*	$\hat{T}$	$\hat{f}$	Estimates
EA-EB	0.68	0.00	1.99				0.85	1.22	*X*-linked only
	0.89	0.00	1.93	2.19	1.86	1.99	0.30	0.77	All loci
				2.45	2.09	1.68	0.05	0.64	Autosomal only
WN-EB	0.46	0.07	1.21				4.93	1.88	*X*-linked only
	0.70	0.10	1.43	1.05	1.09	1.78	1.81	1.22	All loci
				1.28	1.32	1.27	0.50	1.03	Autosomal only
WN-EA	0.56	0.08	1.08				3.11	1.51	*X*-linked only
	0.64	0.09	0.92	1.50	1.12	1.30	1.13	1.57	All loci
				1.59	1.20	1.08	0.23	1.60	Autosomal only

The parameters θ (for each locus), T, and f are estimated from S and D_{ab} by solving the series of simultaneous equations described in HUDSON *et al.* (1987). In the row "*X*-linked only," the parameters are estimated from *E74A, nona,* and *period* only. In the row "all loci," the parameters are estimated from all six loci, and in the row "Autosomal only" the parameters are estimated from *RC98, Xdh,* and *Adh* only.

TABLE 4

Fixed sites tests

Comparison	X-linked loci			Autosomal loci		
	E74A	*nona*	*period*	*RC98*	*Xdh*	*Adh*
EA-EB						
Obs S EA	3	0	3	11	6	6
Exp S EA	2.33	0	4.09	8.52	6.30	7.95
Var S EA	3.04	0	7.32	16.24	11.78	14.33
Obs S EB	2	0	6	4	5	5
Exp S EB	2.26	0	4.70	7.98	6.17	7.51
Var S EB	2.96	0	8.01	15.65	11.64	13.87
Obs Fixed	0	0	0	0	0	0
Exp Fixed	0.002902	0	0.011267	0.000594	0.001155	0.000425
Var Fixed	0.004990	0	0.030106	0.001492	0.002752	0.001002
WN-EA						
Obs S WN	3	0	2	6	5	6
Exp S WN	1.93	0.22	2.63	6.66	4.95	5.59
Var S WN	2.30	0.23	3.40	10.31	6.98	8.33
Obs S EA	3	0	3	11	6	6
Exp S EA	1.68	0.22	1.95	5.78	3.72	5.15
Var S EA	2.04	0.23	2.69	9.40	5.70	7.87
Obs Fixed	0	1	4	0	0	0
Exp Fixed	0.210822	0.032423	0.342710	0.233205	0.194903	0.201198
Var Fixed	0.377973	0.035988	0.730767	0.838721	0.577180	0.652571
WN-EB						
Obs S WN	3	0	2	6	5	6
Exp S WN	2.11	0.24	4.09	4.66	4.81	7.66
Var S WN	2.56	0.25	5.94	6.45	6.74	12.79
Obs S EB	2	0	6	4	5	5
Exp S EB	1.77	0.24	3.49	3.78	3.54	6.65
Var S EB	2.21	0.25	5.30	5.53	5.42	11.72
Obs Fixed	1	1	4	0	0	0
Exp Fixed	0.737976	0.098183	1.525571	0.834483	0.891628	1.406667
Var Fixed	1.127514	0.104733	3.160698	1.938804	2.106639	4.566734

EA-EB: $\chi^2 = 2.79$, d.f. = 11, $P < 0.995$; WN-EA: $\chi^2 = 50.89$, d.f. = 11, $P < 0.001$; WN-EB: $\chi^2 = 14.78$, d.f. = 11, $P < 0.25$.

of neutrality based on differences in *Fst* among loci. This test was criticized for failing to take into account potentially large variances in *Fst* associated with nonisland models of migration (NEI and MARUYAMA 1975; ROBERTSON 1975). In examining the distribution of the DFst test statistic, we also have only explored a single

TABLE 5

DFst tests

Comparison	DFst	*P* value				
		H_0	H_1	H_2	H_3	H_4
EA-EB	0.463	0.01	0.04	0.18	0.004	0.05
WN-EB	0.601	0.006	0.72	0.07	0.001	0.007
WN-EA	0.575	0.008	0.60	0.07	0.004	0.03

H_0: strict neutrality, no migration, parameters estimated from all loci considered together. H_1: no *X*-linked gene flow; gene flow present on autosomes. Divergence time estimated from *X*-linked loci alone, migration parameter M estimated from autosomal loci alone. H_2: no migration, selective sweep of *X*-linked genes at time 0.04 in first population in comparison. H_3: strict neutrality, migration/drift equilibrium with $M = 1/T$ in Table 3. H_4: background selection with f_0 on *X*-chromosomes = 0.32.

TABLE 6

Nonindependence of sites

Locus	Four gametic types	V	C
E74A	0/6	0.037601	437.4
period	0/24	0.475769	3.8
period+ E74A+ nona (WN, EA, and EB)	0/78	0.313208	11.5
period+ E74A+ nona (EA and EB only)	0/10	0.291528	14.7
Xdh	8/15	0.022698	4750.1
Adh	4/15	−0.046917	∞
RC98	10/21	0.009626	>10000

"Four gametic types" is the proportion of the total number of pairwise comparisons between sites at each locus where all four gametic types were seen. The total number of comparisons was calculated after removing all doubly or triply heterozygous individuals, and all variants present in only a single individual. $V = (S^2 - \Sigma h_j + \Sigma h_j^2)/\theta^2$, where $S^2 = \Sigma\Sigma(k_{ij} - K)^2/n^2$, k_{ij} is the number of restriction site differences between the ith and jth alleles, $K = \Sigma\Sigma k_{ij}/n^2$, Σh_j is the sum of the individual site heterozygosities, Σh_j^2 is the sum of each site heterozygosity squared, and θ is estimated by $\Sigma h_j[n/(n-1)]$ (Equation 4 in HUDSON 1987). For the *X*-linked loci, C estimates $3Nc$, where c is the recombination rate between sites. For the autosomal loci, C estimates $4Nc$. C is estimated by solving the equation $V = g(C,n)$ as described in HUDSON (1987). Unless noted in the table, these statistics were calculated from the total sample.

model of migration (Table 5), and it is possible that more complicated models, involving for instance more than two populations, might lead to greater variances and hence spurious rejection. We feel that this is unlikely, however, for the same reason that low levels of migration in the two population model do not lead to rejection even though at equilibrium low migration leads to greatly increased variances in k, D_{ab} and D_{net} (WAKELEY 1996, see above). This is because so long as divergence times are small, as they are in the case of *D. athabasca,* shared history will dominate patterns of neutral variation among populations, making the model of migration or the migration rate irrelevant.

Another difference between the *X*-linked and autosomal variation is the degree to which the polymorphic sites are segregating independently. Within all three autosomal loci, all four gametic types are present at a proportion of all possible pairwise comparisons among sites (Table 6). At *E74A,* like *period* (FORD *et al.* 1994), in no comparison were all four gametic types seen. In fact, if we examine the 26 individuals (or in some cases lines, see Figure 2) for which we have scored all three *X*-linked loci, we still do not see all four gametic types in any of the 78 possible comparisons (Table 6).

We also estimate the parameter $C = 4Nc$ ($3Nc$ for *X*-linked genes), where c is the recombination rate among sites, for each locus using the method of HUDSON (1987). This method is based on the expected relationship between pairwise disequilibria among sites and a statistic related to the variance of the number of pairwise differences among alleles (see Table 6 footnote). For small θ, this estimate is very inaccurate, however, there is clearly a greater variance in the number of pairwise differences, and hence a lower estimated recombination parameter, at *period* and at *period, nona* and *E74A* combined than at the autosomal loci or *E74A* alone. It is important to note that we are examining linkage disequilibrium in the total sample, which includes individuals from at least three distinct populations. As such, the estimates of C are not true estimates of $4Nc$ (or $3Nc$), as this assumes a single randomly mating population. In fact, high levels of linkage disequilibrium are expected when individuals are sampled from two or more genetically distinct populations, and the difference in the degree of linkage disequilibrium between the *X*-linked loci and the autosomal loci is presumably a reflection of the greater degree of differentiation at the *X*-linked genes.

The difference in levels of linkage disequilibrium between the *X*-linked loci and the autosomal loci is also reflected in the cladograms depicting relationships among haplotypes. Like *period,* the restriction site haplotypes at *E74A* can be joined together in a single most parsimonious network with no homoplasy (data not shown). In fact, the haplotypes defined by the variants at all three *X*-linked genes together can be joined in a single most parsimonious network, again with no homoplasy (Figure 5). The variation at the autosomal loci (each considered separately), in contrast, produces hundreds of equally parsimonious trees for each locus (data not shown).

DISCUSSION

Comparison of differentiation at DNA and allozymes: The *D. athabasca* semispecies have been surveyed for ~13 polymorphic allozyme loci (JOHNSON 1978, 1985). Although the populations sampled are not the same as those we sampled for our surveys, it is nonetheless instructive to compare the patterns of variation seen at these two different types of molecular variants. This is especially true in light of recent reports of discrepancies between levels of differentiation seen at allozyme *vs.* DNA markers (KARL and AVISE 1992; POGSON *et al.* 1995). The average estimate of *Fst* across the surveyed allozyme loci and all three semispecies is 0.255 (from JOHNSON 1978), which is similar to the average estimate of *Fst* from the three autosomal loci in our survey (0.196, Table 2) and much lower than the average estimates of *Fst* from the *X*-linked loci (0.765).

The estimate of *Fst* for *Xdh* alone is greater for the restriction sites (0.22) than for the allozyme (0.087, from JOHNSON 1978). The opposite is true for *Adh,* however. At the allozyme level, ADH is one of only two loci to be fixed for different variants between western-

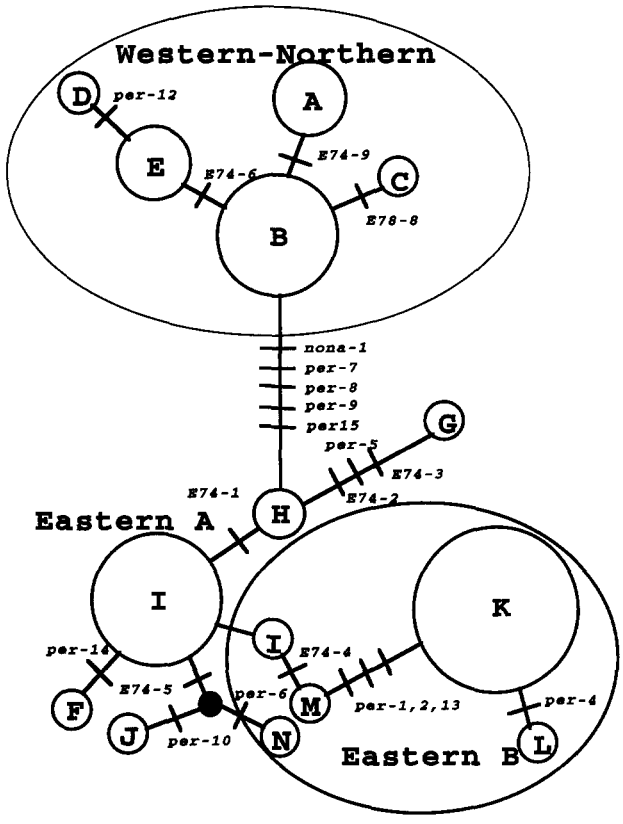

FIGURE 5.—Cladistic network of X-linked haplotypes. At 22 steps, this is the single most parsimonious network linking the 14 haplotypes. Each circle represents a haplotype and the area of the circle is proportional to the frequency of the haplotype in the sample.

northern and the eastern semispecies (JOHNSON 1978). In our survey of restriction site variation, however, *Adh* is the least differentiated among the semispecies of all of the six loci surveyed ($Fst = 0.08$), with no fixed differences. There are several possible explanations for this. One is that *Adh* is duplicated in *D. athabasca*, and we have not surveyed the locus that encodes the allozyme surveyed by JOHNSON (1978). A second possibility is that there really is a fixed amino acid difference between western-northern and the eastern semispecies at the *Adh* locus, but this difference does not result in a restriction site change at any of the enzymes we used in our survey. The fact that we observed evidence for recombination within the *Adh* locus in the history of our sample makes this scenario plausible, since there could be a fixed difference at one part of the locus and shared polymorphisms at other parts of the locus.

KARL and AVISE (1992) and POGSON *et al.* (1995) have interpreted greater differentiation at DNA markers than allozyme markers as evidence for some form of balancing selection maintaining relatively constant allele frequencies at allozyme loci. In the case of the *D. athabasca* semispecies, however, we think it is more likely that the greater differentiation of the X-linked DNA markers is due to selection, whereas the autosomal DNA variants we have scored, and probably most of the allozyme markers, are evolving more in accord with neutrality.

Is the pattern of variation observed at period unique to that locus, or part of a larger X-chromosome phenomenon? Although we observed only a single segregating site at *nona*, it is in the form of a fixed difference between western-northern and the eastern semispecies. This is consistent with the pattern observed at *period*, where much of the variation in the total sample was partitioned among the semispecies and in contrast to the patterns of variation seen at the autosomal loci of this study. *Nona*, like *period*, however, is both X-linked and a candidate song gene, and the greater differentiation at *nona* compared with the autosomal loci could potentially be due to either of these factors. *E74A* is X-linked, but not known to be involved in male mating song. The patterns of variation at *E74A* are in several ways intermediate among those seen at the other X-linked loci and the three autosomal loci. The pairwise estimates of *Fst* at *E74A* (Table 2) are higher than those for the autosomal loci, but lower than for *period* and *nona*. There is a fixed difference at *E74A* between western-northern and eastern-B, but it contributes almost nothing to the FS test statistic among these semispecies (Table 4). The variation at *E74A* is in complete linkage disequilibrium with *period* and *nona*, and there are no cases where all four gametic types are seen within *E74A*, but the variance in the distribution of pairwise differences is more similar to the autosomal loci than it is to *period* (Table 6). A possible explanation for this intermediate pattern of variation is that the loci that were the targets of a selective sweep are located on the short arm of the *X* chromosome where *period* and *nona* putatively reside. Since *E74A* is probably on the long arm of the *X* chromosome, the intermediate patterns of variation at this locus may be due to looser linkage to the selective sites(s). The fact that we observed an easily measurable level of recombination between *E74A* and (*period, nona*) is consistent with this idea. The issue could be resolved by surveying variation at additional X-linked loci, on both the short and long arms of the chromosome.

Evidence for selection at the X-linked loci: The results of the FS and DFst tests indicate that the contrasting patterns of variation at the X-linked *vs.* autosomal loci are inconsistent with a strict neutral model with or without migration (Tables 4 and 5). Several selective models could conceivably explain the data, and we examine each in turn. The first is the possibility of differential gene flow on the *X vs.* autosomes (H_1 in Table 5). This model is consistent with the patterns of variation between western-northern and both eastern semispecies but not between the two eastern semispecies. For western-northern and eastern-B, this scenario is biologically unrealistic, since these two semispecies are allopatric and any gene flow among them would have to be via eastern-A, which seems unlikely. A second

scenario, and the one we favor, is that the greater differentiation at the *X*-linked genes is the result of a selective sweep or sweeps on the *X* chromosome in at least two of the semispecies (H_2 in Table 5). This scenario also explains the greater number of fixed inversion differences on the *X*-chromosome and is consistent with the finding that several of the song characters that differ among the semispecies are *X*-linked (YOON 1991).

Selective sweeps, however, are not the only process that can potentially reduce variation and increase divergence among closely related species. CHARLESWORTH *et al.* (B. CHARLESWORTH *et al.* 1993; D. CHARLESWORTH *et al.* 1995) and HUDSON and KAPLAN (1995) have shown that in regions of low recombination, deleterious mutations can also reduce polymorphism below neutral expectations. This process can also lead to greater differentiation among closely related populations or species and can lead to a rejection of neutrality with HKA, FS and DFst tests (M. J. FORD and C. F. AQUADRO, unpublished data). There are three reasons why we think the process of background selection cannot explain the greater differentiation on the *X*-chromosome in *D. athabasca,* however. The first is that the process of background selection requires that recombination rates be low over a large portion of the chromosome. Although there are many polymorphic chromosomal inversions in *D. athabasca* that could potentially suppress recombination, there is no evidence that within the semispecies this suppression is expected to be any greater on the *X*-chromosome than on the autosomes. In western-northern, for instance, MILLER and VOELKER (1969a,b) found only three small inversions in western-northern sampled from throughout its range. Based on the frequencies of the inversions found in a single population in Minnesota (MILLER and VOELKER 1969a,b) and assuming random mating and that suppression of recombination in inversion heterozygotes is complete but limited to the area of the inversion, recombination on the long arm and short arm of the *X*-chromosome is expected to be reduced in western-northern by only 0.2 and 10%, respectively, compared with the level expected without inversion polymorphism. Comparable data on the actual frequencies of the autosomal inversions is not published, but in western-northern there are at least as many inversions on each autosome as on the *X*-chromosome (MILLER and VOELKER 1968, 1972), so it is likely that inversions contribute to a reduction of autosomal recombination at least as much as they do to the *X*-chromosome.

A second reason that background selection is an unlikely explanation for our data is that, given equal rates of recombination, the model of CHARLESWORTH *et al.* (1993) predicts that background selection will have less of an effect on the *X*-chromosome than on the autosomes, assuming that deleterious mutations are on average partially recessive. Our finding of greater differentiation on *X*-chromosome than the autosomes is inconsistent with this prediction. The third reason background selection seems unlikely is that it is an equilibrium process. This means that the process was probably occurring in the ancestral population of the semispecies as well as in the current populations. If this is the case, then the ancestral population should also have had reduced polymorphism at the *X*-linked genes compared with the autosomal genes. Figure 3 indicates that this was not the case: the levels of variation at the *X*-linked genes in the total sample are on average just as variable as the autosomal genes. This test between background selection and hitchhiking only works if most of the divergence among the semispecies reflects the effects of differential drift (or hitchhiking) and not the accumulation of new mutations along independent lineages. This appears to be the case in the *D. athabasca* semispecies, as the levels of variation for the putatively neutral autosomal loci are on average nearly as great within each semispecies as within the total sample (Figure 3).

In addition to the arguments against background selection above, we can also test the hypothesis more rigorously via computer simulation. To do this, we have estimated f_o, which is the estimated population size of genes on the *X* chromosome compared with genes on the autosomes, assuming that background selection is reducing variation to a greater extent on the *X* chromosome than the autosomes. We estimated f_o as the average within semispecies value of π of the *X*-linked loci divided by the average within semispecies estimate of π of the autosomal loci, which results in an estimate of $f_o = 0.32$. We then performed computer simulations as described for the DFst tests (Table 5), with the assumption that the genes on the *X*-chromosome have a population size 0.32 times that of the autosomal genes. The hypothesis of background selection with this value of f_o was rejected in all three comparisons among the semispecies (H_4 in Table 5). Taken with the arguments above, this indicates that positive selection on *X*-linked genes is a better explanation than background selection for the greater degree of *X*-linked than autosomal differentiation in *D. athabasca.*

What does the evidence for selection on the *X* chromosome imply about speciation in *D. athabasca*? CHARLESWORTH *et al.* (1987) have suggested that if beneficial mutations are often at least partially recessive, adaptations among recently diverged populations will often involve genes on the *X* chromosome. The patterns of variation seen at the *X*-linked loci of *D. athabasca* support this hypothesis. Given our interpretation of recent selection, any phenotypic difference among the semispecies that maps to the *X* chromosome is a candidate for the target of selection. There are only two known phenotypic differences among the semispecies: duration of copulation and male courtship song. By far, the best characterized of these is male courtship song. Males of each semispecies sing a distinct song, and at least some of the charac-

ters that distinguish the songs are *X* linked (MILLER *et al.* 1975; YOON 1991). Based on recency of divergence and the fact that the semispecies that is doubly sympatric (eastern-A) has the most derived song and is the most reproductively isolated, YOON (1991) and YOON and AQUADRO (1994) suggested that there has been selection for increased reproductive isolation among the semispecies. Our finding that the chromosome to which the song differences map has recently experienced selection in at least two of the semispecies strengthens this hypothesis. Based on the results of this study and the fact that western-northern has the least diverged song and eastern-A the most (YOON 1991), we propose the following scenario: the first divergence event was between the common ancestor of eastern-A and eastern-B and the ancestor of western-northern. Based on the estimated divergence time between western-northern and eastern-A from the autosomal loci (Table 3) and assuming a population size of $\sim 10^6$ and ~ 10 generations per year, this would have happened ~23,000 yr ago. This roughly corresponds to the last glacial retreat of ice across much of northern North America (PIELOU 1991) and is about the earliest time western-northern could have expanded into much of its current range. This range expansion may have resulted in new selection pressures (*e.g.*, for a modified circadian clock), and at some point during this time, the ancestor of western-northern experienced a selective sweep at an *X*-linked locus on a rare inversion type. Hybrids between the newly adapted western-northern populations and the more ancestral eastern populations may have been maladapted to both environments. This created a selection pressure for both types to mate assortatively, and the eastern type experienced a selective sweep at a song gene or genes on the *X* chromosome in at least part of its range, resulting in increased reproductive isolation among the groups due to an increase in mating song interpulse interval. This new song type may have been similar to that currently characteristic of eastern-B. Finally, ~5000 yr ago (Table 3), an isolated southern population experienced a selective sweep on the *X* chromosome, and hybrids between the newly adapted southern population (eastern-B) and the other eastern populations were now maladapted. The central eastern populations responded to this selection and diverged even further in several mating song characters, becoming what we now call eastern-A. This scenario is of course speculative, but is consistent with our molecular data and the patterns of song variation and mating behavior. Further support for this scenario could come from a strengthening of our understanding of the relationship between reproductive isolation and courtship song, and a better understanding of the genetics of the reproductive isolation and adaptation among the semispecies.

We thank CASEY BERGMAN for collecting much of the data on *Adh*, KATY SIMONSEN and RICHARD HUDSON for help with computer simulations and calculating the parameter $4Nc$, JOHN WAKELEY for help with fixed differences, and MICHAEL NACHMAN, MARTHA HAMBLIN, RICHARD HARRISON, RON HOY, ANDREW CLARK and STEPHEN SCHAEFFER for useful comments on an earlier version of this manuscript. This research was supported in part by Olin and Howard Hughes Medical Institute Predoctoral fellowships to M.J.F. and National Institutes of Health grant GM-36431 to C.F.A.

LITERATURE CITED

ASHBURNER, M., 1989a *Drosophila: A Laboratory Manual.* Cold Spring Harbor Press, Cold Spring Harbor, New York.

ASHBURNER, M., 1989b *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Press, Cold Spring Harbor, New York.

BEGUN, D., and C. F. AQUADRO, 1991 Molecular population genetics of the distal portion of the *X* chromosome in *Drosophila:* evidence for genetic hitchhiking of the *yellow-achaete* region. Genetics **129:** 1147–1158.

BUTLIN, R., 1989 Reinforcement of premating isolation, In *Speciation and its Consequences,* edited by D. OTTE and J. ENDLER. Sinauer, Sunderland, MA.

CHARLESWORTH, B., J. COYNE and N. BARTON, 1987 The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. **130:** 113–146.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

CHARLESWORTH, D., B. CHARLESWORTH and M.T. MORGAN, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619–1632.

CHIA, W., R. KARP, S. MCGILL and M. ASHBURNER, 1985 Molecular analysis of the ADH region of the genome of *Drosophila melanogaster.* J. Mol. Biol. **186:** 689–706.

COTE, B., W. BENDER, D. CURTIS and A. CHOVNICK, 1986 Molecular mapping of the *rosy* locus in *Drosophila melanogaster.* Genetics **112:** 769–784.

ENDLER, J., 1977 *Geographic Variation, Speciation, and Clines.* Princeton University Press, Princeton, NJ.

EWING, A. W., 1969 The genetic basis of sound production in *Drosophila pseudoobscura* and *D. persimilis.* Anim. Behav. **17:** 555–560.

FORD, M. F., C. K. YOON and C. F. AQUADRO, 1994 Molecular evolution of the *period* gene in *Drosophila athabasca.* Mol. Biol. Evol. **11:** 169–182.

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., 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.

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

HOWARD D. J., 1993 Reinforcement: origin, dynamics, and fate of an evolutionary hypothesis, in *Hybrid Zones and the Evolutionary Process,* edited by R. G. HARRISON. Oxford University Press, New York.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

HUDSON, R. R., 1982 Estimating genetic variability with restriction endonucleases. Genetics **100:** 711–719.

HUDSON, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

HUDSON, R. R., and N. L. KAPLAN, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

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.

JOHNSON, D., 1978 Genetic differentiation in two members of the *Drosophila athabasca* complex. Evolution **32:** 798–811.

JOHNSON, D., 1985 Genetic differentiation in the *Drosophila athabasca* complex. Evolution **39:** 467–472.

JONES C. W., M. D. DALTON and L. H. TOWNLEY, 1991 Interspecific comparisons of the structure and regulation of the *Drosophila* ecdysone-inducible gene E74. Genetics **127:** 535–543.

JONES, K. R., and G. M. RUBIN, 1990 Molecular analysis of *no-on-transient A*, a gene required for normal vision in Drosophila. Neuron **4:** 711–723.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

KARL, S. A., and J. C. AVISE, 1992 Balancing selection at allozyme loci in oysters: implications from nuclear RFLPs. Science **256:** 100–102.

KLIMAN, R. M., and J. HEY, 1993 Sequence variation at the *period* locus within and among species of the *Drosophila melanogaster* complex. Genetics **133:** 375–387.

KULKARNI, S., A. STEINLAUF and J. C. HALL, 1988 The *dissonance* mutant of courtship song in *Drosophila melanogaster:* isolation, behavior and cytogenetics. Genetics **118:** 267–285.

LANDE, R., and M. KIRKPATRICK, 1988 Ecological speciation by sexual selection. J. Theor. Biol. **133:** 85–98.

LEWONTIN, R. C., and J. KRAKAUER, 1973 Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. Genetics **74:** 175–195.

MARFANY, G., and R. GONZALEZ-DUARTE, 1992 The *Drosophila subobscura Adh* genomic region contains valuable evolutionary markers. Mol. Biol. Evol. **9:** 261–277.

MAYNARD SMITH, J., 1966 Sympatric speciation. Am. Nat. **100:** 637–650.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

MILLER, D., 1958 Sexual isolation and variation in mating behavior within *Drosophila athabasca.* Evolution **12:** 72–81.

MILLER, D., and R. VOELKER, 1968 Salivary gland chromosome variation in the *Drosophila affinis* subgroup I. The C chromosome of "western" and "eastern" *Drosophila athabasca.* J. Hered. **59:** 86–98.

MILLER, D., and R. VOELKER, 1969a Salivary gland chromosome variation in the *Drosophila affinis* subgroup III. The long arm of the *X* chromosome in "western" and "eastern" *Drosophila athabasca.* J. Hered. **63:** 230–238.

MILLER, D., and R. VOELKER, 1969b Salivary gland chromosome variation in the *Drosophila affinis* subgroup IV. The short arm of the *X* chromosome in "western" and "eastern" *Drosophila athabasca.* J. Hered. **60:** 306–311.

MILLER, D., and R. VOELKER, 1972 Salivary gland chromosome variation in the *Drosophila affinis* subgroup V. The B and E chromosomes of "western" and "eastern" *Drosophila athabasca.* J. Hered. **63:** 2–10.

MILLER, D., and WESTPHAL, N., 1967 Further evidence on sexual isolation within *Drosophila athabasca.* Evolution **21:** 479–492.

MILLER, D., R. GOLDSTEIN and P. R. ANDERSON, 1975 Semispecies of *Drosophila athabasca* distinguishable by male courtship sounds. Evolution **29:** 531–544.

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

NEI, M., and T. MARUYAMA, 1975 Lewontin-Krakauer test for neutral genes. Genetics **80:** 395.

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

PIELOU, E. C., 1991 *After the Ice Age: The Return of Life to Glaciated North America.* University of Chicago Press, Chicago.

POGSON, G. H., K. A. MESA and R. G. BOUTILIER, 1995 Genetic population structure and gene flow in the Atlantic cod *Gadus morhua:* a comparison of allozyme and nuclear RFLP loci. Genetics **139:** 375–385.

RENDAHL, K. G., K. R. JONES, S. J. KULKARNI, S. H. BAGULLY and J. H. HALL, 1992 The *dissonance* mutation at the *no-on-transient-A* locus of *D. melanogaster:* genetic control of courtship song and visual behaviors by a protein with putative RNA-binding motifs. J. Neurosci. **12:** 390–407.

RILEY, M. A., 1989 Nucleotide sequence of the XDH region in *Drosophila pseudoobscura* and an analysis of the evolution of synonymous codons. Mol. Biol. Evol. **6:** 22–52.

ROBERTSON, A., 1975 Remarks on the Lewontin-Krakauer test. Genetics **80:** 396.

RUSSO, C. A. M., N. TAKEZAKI and M. NEI, 1995 Molecular phylogeny and divergence times of Drosophila species. Mol. Biol. Evol. **12:** 391–404.

SAMBROOK, J., E. F. FRITSCH and T. MANIATIS, 1989 *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Press, NY.

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

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

WIEHE, T. H. E., and W. STEPHAN, 1993 Analysis of a genetic hitchhiking model and its application to DNA polymorphism data from *Drosophila melanogaster.* Mol. Biol. Evol. **10:** 842–854.

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

YOON, C. K., 1991 Molecular and behavioral evolution in the semispecies of *Drosophila athabasca.* Ph.D. dissertation, Cornell University, Ithaca, NY.

YOON, C. K., and C. F. AQUADRO, 1994 Mitochondrial DNA variation within and among the *Drosophila athabasca* semi-species and *D. affinis.* J. Hered. **85:** 421–426.

Communicating editor: A. G. CLARK