Evolution, 48(6), 1994, pp. 1900–1913

USING HITCHHIKING GENES TO STUDY ADAPTATION AND DIVERGENCE DURING SPECIATION WITHIN THE *DROSOPHILA MELANOGASTER* SPECIES COMPLEX

HOLLY HILTON, RICHARD M. KLIMAN[1], AND JODY HEY
Rutgers University, Nelson Labs, P.O. Box 1059, Piscataway, New Jersey 08855-1059
E-mail: HEY@MBCL.rutgers.edu

Abstract.—Several studies of intraspecific and interspecific DNA sequence variation from *Drosophila* loci have revealed a pattern of low intraspecific variation from genomic regions of low recombination. The mechanisms consistently invoked to explain these patterns are the selective sweep of advantageous mutations together with genetic hitchhiking of linked loci. To examine the effect of selective sweeps on genetic divergence during speciation, we studied two loci in different genomic regions thought to be subject to selective sweeps. We obtained DNA sequences from 1.1-kb pair portions of the fourth chromosome locus *cubitus interruptus Dominant* (ci^D) and from the *asense* locus near the telomere of the X chromosome. At ci^D, we found very low variation among multiple lines of *Drosophila mauritiana* and *D. sechellia*. This finding is consistent with an earlier report of very low variation in *D. melanogaster* and *D. simulans* at ci^D and supports the conclusion of selective sweeps and genetic hitchhiking on the nonrecombining fourth chromosome. The pattern of variation found at *asense* suggests that a selective sweep has occurred recently at the tip of the X chromosome in *D. simulans*, but not in *D. melanogaster* or *D. mauritiana*. The data from ci^D and *asense* are compared with data from three X chromosome loci (*period, zeste,* and *yolk protein 2*) that experience normal levels of recombination. By examining estimated genealogies and the rates at which different classes of mutations have accumulated, we conclude that selective sweeps are common occurrences on the fourth chromosome but less common near the tip of the X chromosome. An interesting pattern of low variation at ci^D among *D. simulans, D. mauritiana,* and *D. sechellia* suggests that a selective sweep may have occurred among these forms even after divergence into separate species had begun.

Key words.—*asense, cubitus interruptus Dominant,* phylogeny, speciation.

Received September 7, 1993. Accepted March 10, 1994.

Perhaps the simplest type of adaptation experienced by a population is when an advantageous mutation occurs and, by virtue of higher reproductive success conferred upon carriers, increases in frequency and replaces all other copies of the gene. Adaptations of this type are expected to leave a genetic footprint in the population, a region of zero or reduced variation among DNA sequences at and near the site of the original mutation. The formation of this footprint has been called genetic hitchhiking (Maynard Smith and Haigh 1974; Kaplan et al. 1989), and the size of the footprint depends on how much recombination within heterozygotes occurred near the site during the fixation process. Recent estimates of DNA sequence variation from genomic regions of low or no recombination in *Drosophila* have supported predictions from models of genetic hitchhiking. These studies have found very little DNA sequence variation within species, despite finding levels of interspecific variation similar to those reported for genomic regions with regular recombination (Begun and Aquadro 1991; Berry et al. 1991; Martin-Campos et al. 1992; Stephan and Mitchell 1992; Langley et al. 1993). Additional support for the genetic hitchhiking model comes from a compilation of data from 20 gene regions from across the *D. melanogaster* genome, which revealed a positive correlation between nucleotide diversity and local rates of recombination (Begun and Aquadro 1992).

This paper describes research on the role of this type of adaptation (i.e., characterized by selective sweeps and genetic hitchhiking) in the divergence of recently formed species. Our approach has been to measure DNA sequence variation at genomic hitchhiking regions within and between recently diverged species and to compare these patterns to those previously reported for other genomic regions. We report on two loci, from regions of low recombination and thought to be subject to genetic hitchhiking, in the four species of the *D. melanogaster* species complex. The ci^D locus is on the nonrecombining fourth chromosome (Hochman 1976), and a 1.1-kb pair region had previously been sequenced by Berry

[1] Present address: Department of Biology, Radford University, Radford, Virginia 24124.

et al. (1991) in ten lines of *D. melanogaster* and nine lines of *D. simulans*. They found the typical hitchhiking pattern of little polymorphism within species with normal amounts of divergence between species. We sequenced a 1.1-kb region of ci^D from six individuals of *D. mauritiana* and four individuals of *D. sechellia*; both species are island endemics and are very closely related to *D. simulans* (Lachaise et al. 1988). We also sequenced a 1.1-kb stretch of the *asense* locus in six lines in each of *D. melanogaster, D. simulans* and *D. sechellia* and in five lines of *D. mauritiana*. The *asense* gene is part of the *yellow-achaete-scute* complex (Gonzalez, et al. 1989), located on the tip of the X chromosome and, therefore, also subject to reduced recombination (Ashburner 1989, p. 453). There have been several reports of reduced variation within the *yellow-achaete-scute* complex in *D. melanogaster* and *D. simulans* (Aguade et al. 1989; Beech and Leigh Brown 1989; Begun and Aquadro 1991; Martin-Campos et al. 1992).

The ci^D and *asense* loci were also chosen because they come from genomic regions that experience different levels of recombination. The fourth chromosome, the site of ci^D, has not been found to undergo meiotic recombination under normal conditions (see Hochman 1976), so that a selective sweep at any locus would be expected to effect the entire chromosome. From estimates of DNA content in polytene bands (Sorsa 1988, chap. 16, table 1), ci^D is completely linked to about 1.4% of the genome. In contrast, *asense* is found near one end of a large chromosome, and in a region where recombination is measurable. The *yellow-achaete-scute* complex has at least seven genes and is approximately 90-kb long. Recombination in the region is thought to occur at a roughly 17-fold to 20-fold reduction relative to other regions of the genome (Dubinin et al. 1937; Begun and Aquadro 1991; Aguade et al. 1989). In contrast to ci^D, the chance that *asense* undergoes hitchhiking depends on the proximity of the selected locus *and* the strength of selection. More strongly selected mutations will go to fixation more quickly, on average, and thus will cause hitchhiking in a larger region (Maynard Smith and Haigh 1974).

The ci^D and *asense* data sets are comparable to recently reported studies of loci from regularly recombining regions of the X chromosome. Kliman and Hey (1993a) studied variation in a 1.9-kb region of the *period* (*per*) locus from six individuals of each of the four species; and Hey and Kliman (1993) reported data from approximately 1-kb regions of the *zeste* and *yolk protein 2* (*yp2*) loci, again with six sequences from each species. None of these loci showed evidence of either recent selective sweeps or recent balancing selection. Together, the data from *per, zeste,* and *yp2* support a historical model in which *D. simulans* has a large population size and has changed relatively slowly since the origin of the island endemics. The separations of *D. sechellia* and *D. mauritiana* from ancestral *D. simulans* appear to have occurred recently and at similar times. By extrapolating from estimated absolute rates of silent substitution, the time since formation of the island endemics was estimated at between 0.58 and 0.86 Mya, with *D. sechellia* having diverged from ancestral *D. simulans* about 0.1 My before *D. mauritiana* (Hey and Kliman 1993). The time since the split from *D. melanogaster* was estimated at 2.5–3.4 Mya.

Most hitchhiking studies have compared *D. melanogaster* and *D. simulans* (Begun and Aquadro 1991; Berry et al. 1991; Martin-Campos et al. 1992; Langley et al. 1993). By extending the research to include *D. mauritiana* and *D. sechellia* and comparing our results to data from three nonhitchhiking loci (*per, zeste,* and *yp2*), we can explore speciation among very recently diverged species while examining the effect of adaptive selective sweeps on the process.

Materials and Methods

Sources of Flies

All strains were also used for *per* (Kliman and Hey 1993a), *zeste,* and *yp2* (Hey and Kliman 1993). The *asense* sequences were generated using the identical DNA preparations (and thus the same X chromosomes) used in the *per* study (Kliman and Hey 1993a), with the exception of one line of *Drosophila mauritiana* (MA-1). The ci^D sequences were generated from new DNA preparations of a subset of the isofemale lines used in the other studies. The six *D. mauritiana* lines are the same as before, and the four lines of *D. sechellia* were SE-C1, SE-C2, SE-P1 and SE-P3 (Kliman and Hey 1993a).

DNA Preparation

DNA preparations were made from single male flies (protocol 48 in Ashburner 1989). At ci^D, a 1.1-kb region was PCR-amplified using the same 20-mer oligonucleotide primers as Berry et al. (1991), starting at positions 1897 ("+" primer 5′ base) and 3003 ("−" primer 5′ base) in the

TABLE 1. Variable sites for ci^D among *Drosophila simulans, Drosophila mauritiana* and *Drosophila sechellia.* This table corresponds closely to table 1 of Berry et al. (1991). With the exception of sites 11, 2154, 2338, and 2765, all of the substitutions listed in table 1 of Berry et al. (1991) show *D. mauritiana* and *D. sechellia* to have the same base as reported for *D. simulans.* With just two exceptions, position 2279 in *D. mauritiana* and position 2338 in *D. simulans* (Berry et al. 1991), there was no intraspecific variation.

		Species/DNA sequence				Species/amino acid sequence			
Region	Position	*mel*	*sim*	*mau*	*sec*	*mel*	*sim*	*mau*	*sec*
Introl 2	11	C	T	C	C				
Exon	2154	G	A	G	G	Pro	Pro	Pro	Pro
	2279	T	T	T/A	T	Leu	Leu	Leu/His	Leu
	2338	T	T/G	T	T	Leu	Leu/Val	Leu	Leu
	2572	G	G	G	C	Ser	Ser	Ser	Ser
	2608	A	A	C	A	Arg	Arg	Arg	Arg
	2765	T	C	T	T	Leu	Pro	Leu	Leu
	2775	G	G	C	G	Thr	Thr	Thr	Thr
	2822	T	T	A	C	Phe	Phe	Tyr	Ser

original sequence (Orenic et al. 1990). At *asense,* a 1.1-kb region was PCR-amplified using 20-mer oligonucleotide primers corresponding to bases 2082–2101 ("+" primer) and 3353–3372 ("−"primer) of the published sequence of Gonzalez et al. (1989). PCR and DNA sequencing methods were identical to those of Kliman and Hey (1993a).

Simulations

For HKA tests, the distribution of X^2, the test statistic, was generated by a multispecies coalescent simulation. The parameters for the simulation are the HKA parameter estimates generated by applying the HKA test to the actual data. The simulation protocol is very similar to that for a conventional coalescent simulation (Hudson 1983, 1990): (1) the simulation proceeds backward in time within each of the two species until the time of speciation is reached; and then (2) the remaining lines from each of the two species are coalesced as if from a single species. For each of 20,000 rounds of simulation, the HKA test is applied to the simulated data, and a distribution of values for X^2 is generated. The actual value of X^2 is then placed within this distribution to assess the significance level of the observation.

For K tests (see RESULTS, *Selective Sweeps during Speciation,* for descriptions of K_1 and K_2), a modified HKA test was developed for the case of three species (details available on request). The parameter estimates, including time estimates for two cases of speciation, were input to three-species coalescent simulations. These estimates were conducted in the same way as those described above for the conventional HKA test, with the addition of extending the simulations through two speciation events. For each of 5000 rounds of simulation, K_1 and K_2 were calculated for each locus and an X^2 statistic was calculated for all of the loci in the test. The actual value of X^2, calculated from the K_1 and K_2 estimates from the real data, was then compared to the distribution of X^2 values generated by the simulations.

TABLE 2. The average number of pairwise differences per base pair within species (based on all base-pair differences). The *per* data are from Kliman and Hey (1993a). The *zeste* and *yp2* data are from Hey and Kliman (1993). The ci^D data from *Drosophila melanogaster* and *D. simulans* is from Berry et al. (1991).

	Species			
Locus	*melano-gaster*	*simulans*	*mauritiana*	*sechellia*
zeste	0.0025	0.0078	0.0045	0.0000
yp2	0.0052	0.0011	0.0012	0.0003
per	0.0062	0.0115	0.0118	0.0009
asense	0.0021	0.0000	0.0023	0.0000
ci^D	0.0000	0.0002	0.0003	0.0000

RESULTS

DNA Sequence Variation Summary

cubitus interruptus.—The *D. mauritiana* and *D. sechellia* sequences cover the same 1075-bp region of ci^D sequenced by Berry et al. (1991). Table 1 summarizes all differences found within the *D. simulan–D. mauritiana–D. sechellia* triad (hereafter referred to as the *simulans* complex). All four *D. sechellia* sequences were identical. The six *D. mauritiana* sequences revealed one polymorphism, a single base change that distinguished one of the six sequences.

TABLE 3. Variation between species. Net divergence is equal to the average pairwise divergence between species less the mean of the intraspecific values given for the two species in table 2 (Nei and Tajima 1981). Fixed differences are the number of base-pair positions at which all of the sequences from species 1 are different from all of the sequences of species 2 (see table 1).

Species 1–species 2	*zeste*	*yp2*	*per*	*asense*	*ci*^{D}
Net divergence per base pair					
simulans–mauritiana	0.0033	0.0029	0.0069	0.0013	0.0056
simulans–sechellia	0.0084	0.0033	0.0114	0.0019	0.0047
mauritiana–sechellia	0.0109	0.0066	0.0167	0.0032	0.0037
melanogaster–simulans	0.033	0.023	0.026	0.024	0.050
melanogaster–mauritiana	0.034	0.025	0.031	0.024	0.050
melanogaster–sechellia	0.037	0.025	0.035	0.023	0.049
Fixed differences					
simulans–mauritiana	1	2	3	1	6
simulans–sechellia	4	4	18	2	5
mauritiana–sechellia	10	6	21	3	4
melanogaster–simulans	29	23	37	26	54
melanogaster–mauritiana	32	25	44	27	54
melanogaster–sechellia	36	27	60	28	53

Intraspecific variation is summarized in table 2, and interspecific divergence is summarized in table 3. By comparing the hitchhiking loci with *per, zeste,* and *yp2,* we can look for patterns expected of recent selective sweeps. Table 2 shows that *D. mauritiana ci*^{D} sequences exhibit the same pattern of very reduced levels of variation, relative to other loci, as were seen in *D. melanogaster* and *D. simulans* at *ci*^{D} (Berry et al. 1991). *Drosophila mauritiana* is endemic to the single island of Mauritius, and its level of variation at *per, zeste,* and *yp2* suggests an effective population size (N_e) even larger than that of *D. melanogaster,* though less than *D. simulans* (Hey and Kliman 1993). We also found no intraspecific variation in *D. sechellia*; however, variation is low, in general, in this species (Cariou et al. 1990; Hey and Kliman 1993; Kliman and Hey 1993a).

asense.—Figure 1 summarizes all nucleotide differences found in *asense.* There was no polymorphism among the six *D. simulans* sequences, nor among the six *D. sechellia* sequences. For *D. melanogaster,* the four North American lines (ME-NJ1, ME-NJ2, ME-LI1, and ME-LI2) were identical. The African lines, ME-K1 and ME-K2, differed at multiple sites from the North American lines (two and five sites, respectively) and from each other (five sites). Among the five *D. mauritiana* sequences, there were four distinct sequences (MA-2 and MA-3 were identical) and five polymorphic sites. One line, MA-6, contained a 9-base deletion, within a region of repeats, that was otherwise unique to the *D. sechellia* sequences. This may represent an old-length polymorphism that is fixed as a deletion in *D. sechellia,* fixed as present in *D. simulans,* and still segregating in *D. mauritiana* (fig. 1).

At *asense,* both *D. mauritiana* and *D. melanogaster* have levels of variation within or near the range observed at the other X-linked loci. In *D. simulans,* however, we found no intraspecific variation, which suggests that there may have been a recent selective sweep at this locus. We can rule out the possibility that the six *asense* sequences of *D. simulans* are similar to each other by accidental sampling of closely related lines, because we observed substantial polymorphism at other X-linked loci using the same DNA preparations (table 2). In the case of *D. sechellia,* the lack of variation, although consistent with genetic hitchhiking, is most simply explained by a very small N_e (Hey and Kliman 1993).

Genealogical Inference

Maximum parsimony analysis was carried out using PAUP (Swofford 1985) with the branch-and-bound option (Hendy and Penny 1982). The gene trees built from the *ci*^{D} and *asense* data can be compared to those from *per* (Kliman and Hey 1993a), *yp2* and *zeste* (Hey and Kliman 1993) and examined within the phylogenetic and speciation context developed with those loci. Together the three nonhitchhiking genes support a history in which first *D. sechellia,* and then *D. mauritiana,* emerged from a very large population of *D. simulans.*

cubitus interruptus.—For *ci*^{D}, exactly three most parsimonious gene trees were found, all having

```
Base                    1111111112  2   2   23333444456
position  4   59   9    0000000000  0   2   31255048961
          7   23   7    0123456780  9   1   80225051620
comment   r   sr   r    ddddddddddr  r   r   sssssssssr
ME-NJ1    T(S)TG(R)C(H)CCAGCAGAAA(S)C(H)A(S)GGCCAAATCTT(D)
ME-NJ2    -(-)--(-)-(-)----------(-)-(H)-(-)----------(-)
ME-K1     -(-)--(-)-(-)----------(-)-(H)-(-)----------(-)
ME-K2     -(-)--(-)-(-)----------(-)-(H)-(-)----------(-)
ME-LI1    -(-)--(-)-(-)----------(-)-(H)-(-)----------(-)
ME-LI2    -(-)--(-)-(-)----------(-)-(H)-(-)----------(-)
SI-CA1    C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCT-G(E)
SI-CA2    C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCT-G(E)
SI-K1     C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCT-G(E)
SI-K2     C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCT-G(E)
SI-LI1    C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCT-G(E)
SI-LI2    C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCT-G(E)
SE-C1     C(P)CA(Q)G(Q)*********T(C)A(S)G(G)A-TTGG-CT-G(E)
SE-C1     C(P)CA(Q)G(Q)*********T(C)A(S)G(G)A-TTGG-CT-G(E)
SE-P1     C(P)CA(Q)G(Q)*********T(C)A(S)G(G)A-TTGG-CT-G(E)
SE-P2     C(P)CA(Q)G(Q)*********T(C)A(S)G(G)A-TTGG-CT-G(E)
SE-P3     C(P)CA(Q)G(Q)*********T(C)A(S)G(G)A-TTGG-CT-G(E)
SE-P4     C(P)CA(Q)G(Q)*********T(C)A(S)G(G)A-TTGG-CT-G(E)
MA-2      C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCGCG(E)
MA-3      C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCGCG(E)
MA-4      C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCGCG(E)
MA-5      C(P)CA(Q)G(Q)---------T(C)A(S)G(G)A-TTGGGCG-G(E)
MA-6      C(P)CA(Q)G(Q)*********T(C)A(S)G(G)AATTGGGCG-G(E)

Base                                                1   1
position  6   66   677   77   89   9   99   99      0   0
          2   25   925   88   50   1   25   67      1   1
          0   81   712   48   35   0   87   17      1   8
comment   r   sr   ssr   sr   sr   r   sr   sr      r   s
ME-NJ1    A(T)AT(L)AAA(T)AC(L)TA(T)G(Q)GT(V)TA(S)A(Q)T
ME-NJ2    -(-)--(L)---(-)--(-)--(-)-(-)--(-)--(-)-(-)-
ME-K1     -(-)G-(-)--G(A)-A(I)C-(-)-(-)-C(A)--(-)-(-)-
ME-K2     -(-)G-(-)---(-)--(-)--(-)-(-)--(-)--(-)-(-)C
ME-LI1    -(-)--(-)---(-)--(-)--(-)-(-)--(-)--(-)-(-)-
ME-LI2    -(-)--(-)---(-)--(-)--(-)-(-)--(-)--(-)-(-)-
SI-CA1    T(S)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SI-CA2    T(S)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SI-K1     T(S)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SI-K2     T(S)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SI-LI1    T(S)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SI-LI2    T(S)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SE-C1     C(P)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SE-C1     C(P)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SE-P1     C(P)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SE-P2     C(P)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SE-P3     C(P)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
SE-P4     C(P)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
MA-2      T(S)GC(P)G-G(A)G-(-)-G(A)-(-)AC(A)CG(G)C(P)-
MA-3      T(S)GC(P)G-G(A)G-(-)-G(A)-(-)AC(A)CG(G)C(P)-
MA-4      T(S)GC(P)G-G(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
MA-5      T(S)GC(P)GTG(A)G-(-)--(-)-(-)AC(A)CG(G)C(P)-
MA-6      T(S)GC(P)G-G(A)G-(-)--(-)T(H)AC(A)CG(G)C(P)-
```

FIG. 1. Variable sites at *asense*. The first rows indicate the base position of variable sites within the sequenced region. The first and last bases sequenced correspond to positions 2149 and 3216, respectively, of Gonzalez et al. (1989). In the comment row, s, synonymous substitution; r, amino acid replacement substitution; and d, deletion. The ME-NJ1 sequence is used as the reference. Nucleotides identical to the reference in the remaining 22 lines are indicated by a dash. At amino acid replacement sites, the nucleotide is followed in parentheses by the one letter code for the resulting amino acid (S, Ser; P, Pro; R, arg; Q, Gln; H, his; C, Cys; G, Gly; D, Asp; E =glu; T, thr; L, Leu; I, ile; A, Ala). Length variation is indicated by an asterisk (*) in sequences shortened relative to others.

a length of 61 steps (including both informative and noninformative sites) and a consistency index of 1.0. The ambiguity results from one of the informative sites having three character states (position 2822 in table 1). The trees differed in the position of the *D. sechellia* sequence: as a sister group to *D. mauritiana*; as a sister group to *D. simulans*; or as a sister group to a *D. mauritiana–D. simulans* pair. Regardless of which of the three trees is correct, the data suggest a virtual trichotomy, and are portrayed as such in a strict consensus tree in figure 2.

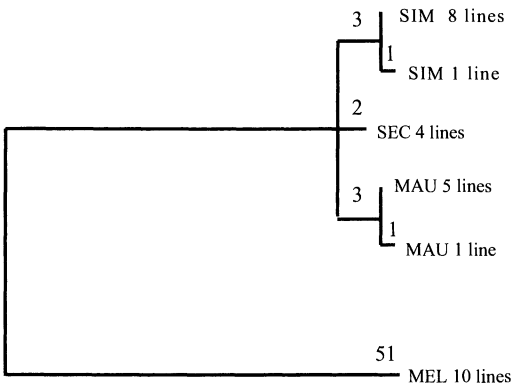

FIG. 2. Unrooted consensus of three most parsimonious trees for *ci*^{D}. Numbers adjacent to each branch are the number of inferred mutational steps.

Divergence among the species of the *simulans* complex appears different for *ci*^{D} than for the other loci. For the X-linked loci, including *asense,* in both net divergence and fixed differences, the divergence between *D. mauritiana* and *D. sechellia* is approximated by the sum of the divergence values between each of these species and *D. simulans.* It is as if *D. simulans* has evolved very slowly since the formation of the island endemics (table 3). In contrast, at *ci*^{D} there are very similar numbers of pairwise differences found between the species of the *simulans* complex; and this is consistent with the trichotomy genealogy for *ci*^{D} in the *simulans* complex.

asense.—With the *asense* data, three most parsimonious gene trees were found, each having lengths of 39 steps (from both informative and noninformative sites) and consistency index values of 0.97 (from 30 informative sites). Three trees were found because there are no informative sites to distinguish among three of the *D. mauritiana* haplotypes. One obvious difference between the *asense* tree and those for the other X-linked loci is the collapse of the variation at *D. simulans* into a single lineage, as expected if there has been a recent hitchhiking event.

The strict consensus tree for *asense* (fig. 3) is consistent with the ordering of speciation events inferred from the *per, zeste,* and *yp2* trees, with *D. sechellia* arising first, followed by a split between *D. simulans* and *D. mauritiana* (Hey and Kliman 1993). The only polymorphic site that cannot be placed on the tree in figure 3 as a single mutation (thus reducing the consistency index to 0.97) is the 9-bp deletion. This deletion appears relatively ancient, much like the *per* changes that

are shared between *D. mauritiana* and *D. simulans*. The deletion is segregating in *D. mauritiana*, fixed in *D. sechellia*, and absent from *D. simulans*.

Tests of Natural Selection

The essential criteria for an inference of genetic hitchhiking is that levels of variation within species appear reduced and inconsistent with levels of variation found between species. A statistical framework is available in the form of the HKA test (Hudson et al. 1987), which employs a neutral model connecting intraspecific and interspecific variation under a common set of parameters. For a pair of species, and an arbitrary number of loci, the method provides estimates of Θ ($4N_e\mu$; μ is the neutral mutation rate; $3N_e\mu$ is more accurate for sex-linked loci) for each species and locus, as well as an estimate of the time since the two species diverged from their common ancestor.

Table 4 lists the results of several tests on various subsets of the five-locus, four-species data set. A large number of tests are possible, but we have focused on those contrasting *asense* and *ci*^{D} with the other loci, and on the species pairs *D. melanogaster–D. simulans* and *D. simulans–D. mauritiana*. The contrast between *D. melanogaster* and *D. simulans* was chosen because it is a frequently made comparison (e.g., Begun and Aquadro 1991; Berry et al. 1991; Langley et al. 1993). For a contrast within the *simulans* complex, we focused on *D. simulans* and *D. mauritiana*. Although this comparison is not independent of that between *D. melanogaster* and *D. simulans*, we chose not focus on *D. mauritiana* and *D. sechellia* because of the lack of variation at all loci in the latter.

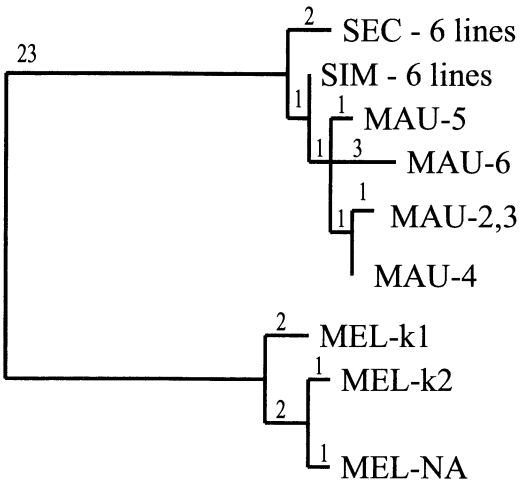

FIG. 3. Unrooted consensus of the three most parsimonious tree for *asense*. Numbers adjacent to each branch are the number inferred mutational steps. ME-NA refers to the identical sequences found from ME-NJ1, ME-NJ2, ME-LI1 and ME-LI2

For each contrast in table 4, the test statistic, X^2, has been compared with both a χ^2 distribution and with a distribution generated from simulations using the parameter estimates. In the original application of the HKA test (Hudson et al. 1987), the distribution of the test statistic, X^2, was found by simulation to be in excellent fit with the χ^2 distribution. To see if this held true for our data, we also carried out simulations. These simulations were carried out in a coalescent fashion (Hudson 1990) in strict accord with

TABLE 4. HKA tests (Hudson et al. 1987). $\hat{\Theta}$ is an estimate of $3N_eu$ for species 1 for sex-linked loci and $4N_eu$ for autosomal loci (i.e., *ci*^{D}). Population size corrections for including X-linked loci with autosomal loci followed Begun and Aquadro (1991). $\hat{T}$ is an estimate of the time since the common ancestor of the species in units of 3/2 N_e generations, where N_e is the effective population size for species 1. $\hat{f}$ is an estimate of the scaler by which estimates of $3N_eu$ for species 1 are multiplied to get those for species 2. X^2 is the goodness-of-fit statistic. P_{χ^2} is the probability of observing a X^2 value greater than or equal to the actual value, assuming a χ^2 distribution. P_{sim} is the probability according to a distribution generated via simulation (20,000 replicates; see Materials and Methods).

Species 1–species 2	$\hat{\Theta}$ *per*	*zeste*	*yp2*	*ci*^{D}	*asense*	$\hat{T}$	$\hat{f}$	X^2	P_{χ^2}	P_{sim}
melanogaster–simulans	12.03	5.09	3.47	—	—	5.07	1.60	4.40	0.355	0.275
melanogaster–simulans	9.36	3.95	2.70	3.69	—	8.48	1.63	27.0	0.00014**	0.00015**
melanogaster–simulans	12.01	5.08	3.47	—	2.66	5.60	1.42	9.69	0.138	0.057
simulans–mauritiana	24.33	6.48	2.03	—	—	0.57	0.81	0.91	0.924	0.782
simulans–mauritiana	23.45	6.25	1.97	1.36	—	0.76	0.82	7.64	0.265	0.080
simulans–mauritiana	23.42	6.24	1.97	—	1.24	0.60	0.88	4.41	0.622	0.290

** $P < 0.01$.

TABLE 5. Relative rates of change at multiple loci in the *Drosophila melanogaster* complex. Changes were determined for a data subset with one sequence randomly drawn from each species (see text). Silent sites were calculated by considering, for each base position of a *D. simulans* sequence, the fraction of possible base changes (⅓, ⅔, or ⅓) that would not affect the amino-acid sequence. These values were then summed across all exon base positions, and the total was rounded to the nearest integer. Intron, synonymous, and replacement changes are simply the total number of variable sites of those types observed across the four sequences (see Results, *Evolutionary Constraint*). *I* is the number of intron changes divided by the total intron length. *S* is the number of synonymous changes divided by the number of silent sites. *R* is the number of replacement changes divided by the number of replacement sites (i.e., exon length minus silent sites). CAI is the codon adaptation index (Sharp and Li 1987) calculated with the codon usage table for "high bias" genes identified in table 2 of (Shields et al. 1988). For those codons in table 2 of Shields et al. (1988) with zero counts, we followed the suggestion of Bulmer (1988) and used a relative usage level of 0.01. A single *D. simulans* sequence was used for the calculation of CAI, although the value changes very little if the sequences from the other species are used. Intron G-C content was calculated from a *D. melanogaster* sequence.

Locus	Intron length	Exon length	Silent sites	Intron changes	Synonymous changes	Replacement changes	*I*	*S*	*R*	CAI	Intron G-C content (%)
zeste	182	805	167	20	22	2	0.110	0.132	0.003	0.467	37.7
yp2	63	1051	234	10	19	8	0.159	0.081	0.010	0.697	37.5
per	192	1679	386	28	77	4	0.145	0.200	0.003	0.490	51.0
asense	0	1067	208	—	15	14	—	0.072	0.016	0.226	—
ci^{D}	117	958	208	10	26	22	0.085	0.125	0.029	0.160	15.5

the assumptions of the HKA test (Hudson et al. 1987; see MATERIALS AND METHODS, *Simulations*). In contrast to the original application of the HKA test, there is often a considerable discrepancy between the two significance levels for the tests in table 4, especially for the *D. simulans–D. mauritiana* tests. These discrepancies are probably caused by the numerous cases of low values for intraspecific and interspecific variation. The use of the χ^2 distribution entails an assumption, that measures of variation are approximately normally distributed, which holds much better for high values.

When just *per, zeste,* and *yp2* are considered, the data are consistent with the neutral model (table 4) (Hey and Kliman 1993). Berry et al. (1991) had shown that *ci*^{D} from *D. melanogaster* and *D. simulans,* in HKA tests involving genomic regions at and near the *Adh* locus, does not fit the neutral model. From table 4, this also holds when *ci*^{D} is paired with *per, zeste,* and *yp2.* When *D. simulans* and *D. mauritiana* are considered with *per, zeste, yp2,* and *ci*^{D}, the neutral model is not rejected. The reason for this is that although there is low variation within the species of the *simulans* complex at *ci*^{D}, there is also very low between species variation (table 3, fig. 2).

For *asense,* the HKA tests generally show the neutral model cannot be rejected; however, P_{sim} is very close (0.057, table 4) in the *D. melanogaster–D. simulans* case. If we examine the actual departures from expectations in this test, we find that out of a total X^2 value of 9.69, the two greatest contributions came from *asense* variation within *D. simulans* (2.52; observation less than expected) and *asense* variation between *D. simulans* and *D. melanogaster* (2.66; observation greater than expected). Thus, the directions of the greatest departure from the neutral model fit a pattern of recent hitchhiking in *D. simulans.*

Evolutionary Constraint

To compare rates of evolutionary change among loci, a single sequence was randomly picked from each species and the total number of changes among the four sequences were tabulated for each locus (table 5). The single-sequence approach was taken to help ensure that the length of the gene tree, in terms of time (not mutations), is the same for all loci. Specifically, we avoid much of the difficulty presented by loci thought to have had recent selective sweeps and which will have shortened gene trees within species as a result of the recent common ancestry caused by the selective sweep. However, using a single sequence does not completely remove the issue of variation among loci in the time depth of gene trees. Gene trees from loci with recurrent selective sweeps are also expected to have slightly shorter time depths between species (see *Selective Sweeps during Speciation*).

The different loci vary considerably for levels of amino acid replacement changes (table 5, fig. 4), with *asense* and *ci*^{D} exhibiting higher levels than the other loci. This variation is significantly greater than expected by chance (G = 32.51, 4

df, $P \leq 1.5 \cdot 10^{-6}$). In individual pairwise comparisons with the other loci, *ci*^{D} has significantly more replacement changes than all except *asense* (*per*: $G = 25.10$, $P \leq 6.7 \cdot 10^{-5}$; *zeste*: $G = 16.2$, $P \leq 1.1 \cdot 10^{-5}$; *yp2*: $G = 7.9$, $P \leq 0.0049$; *asense*: $G = 2.9$, $P \leq 0.084$; 1 df in all cases). In individual pairwise comparisons, *asense* is significantly different from *zeste* and *per* for replacement changes (*per*: $G = 10.6$, $P \leq 0.001$; *zeste*: $G = 6.85$, $P \leq 0.009$; *yp2*: $G = 1.35$, $P \leq 0.245$; 1 df in all cases).

The pattern of elevated levels of amino acid replacements in *ci*^{D} and *asense* is predicted by population genetic models, if many of those replacements are slightly deleterious. In general, natural selection is expected to be less effective in removing deleterious variation from regions of low recombination (Muller 1964; Hill and Robertson 1966; Felsenstein 1974; Li 1987; Birky and Walsh 1988; B. Charlesworth et al. 1993; D. Charlesworth et al. 1993). This effect should be especially marked for a region undergoing regular selective sweeps at which selection against slightly deleterious mutations is dominated by the fixation process of linked favorable mutations. The increased number of replacement changes in *ci*^{D} probably reflects an increase in the fixation rate of slightly deleterious mutations that purifying selection was unable to detect. The intermediate position of *asense* between *ci*^{D} and the other loci is consistent with this model if *asense* experiences fewer selective sweeps than does *ci*^{D}.

A second measure of evolutionary constraint is codon bias, the unequal usage of synonymous codons. Codon bias may be positively associated with gene-expression levels in *Drosophila* (Shields et al. 1988) and presumably reflects the efficiency at which a gene is able to be translated (Ikemura 1985). As the selective advantage of each optimal codon is very small, codon bias should also be reduced in areas of low recombination, where the effectiveness of natural selection is expected to be reduced. This prediction was confirmed in a study of 385 *D. melanogaster* loci (Kliman and Hey 1993b), and the pattern also fits the loci considered in this report. Codon bias in *ci*^{D} and *asense* is lower than in *per, zeste,* and *yp2* (table 5). These loci also have codon bias scores typical for genes of their genomic region. From Kliman and Hey (1993b), the mean of all genes was 0.430; eight loci near the tip of the X chromosome (including *asense*) had an average codon bias score of 0.30; and three fourth-chromosome genes (including *ci*^{D}) had an average score of 0.18.

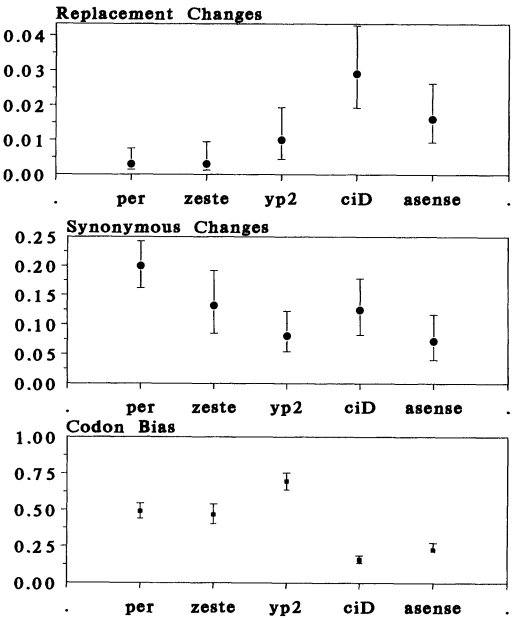

FIG. 4. Synonymous changes, Amino acid replacement changes and codon bias. Top and middle: the levels of synonymous changes and amino acid replacement changes given in table 5 together with 95% confidence limits. Bottom: values of codon bias (CAI) from table 5 with 95% confidence limits. The confidence limits were determined from the 2.5% and 97.5% positions within a ranking of 1000 CAI values created by simulation. For each gene the simulations were carried out by constructing genes of the same length as the actual sequence. For each simulated gene, both amino acids and codons were sampled, with replacement, from the distributions observed in the actual sequenced region.

The high numbers of replacement changes and the low codon bias strongly suggest that populations experience reduced levels of N_e for genomic regions near *ci*^{D}. The somewhat higher numbers of replacement changes and low codon bias in *asense* also suggests reduced N_e, though the effect appears to be less than on the fourth chromosome. Natural selection, while inhibited, may be more effective at the tip of the X chromosome, because there is some recombination and, hence, shorter regions of tight linkage.

If codon bias variation among loci is due to variation in the effectiveness of selection in screening suboptimal codons, then it is expected that *asense* and *ci*^{D} should also have higher levels of synonymous changes. From table 5 and figure 4, it is clear that they do not. Among the five loci, the variation in the level of synonymous variants is significant ($G = 20.9$, 3 df, $P \leq 0.0003$) and is largely due to the high level at *per* (without

per, $G = 4.9$, 3 df, $P \leq 0.29$). A possible explanation for low synonymous variation at ci^D is that the locus is very A+T rich. Kliman and Hey (1993b) surveyed G+C content in 142 *D. melanogaster* intron sequences from Genbank and found an average of 36.8% with a range of 22.6%–55.2%. The *per*, *zeste*, and *yp2* loci fall within this range (table 5), whereas ci^D, has a level of 15.5%. Thus, it may be that ci^D has a reduced rate of synonymous variation because of underlying mutational constraints. However, this explanation cannot be used for *asense* because this gene has no introns (Gonzalez et al. 1989) and because the G+C content of the coding region (50%) is higher than that of ci^D (45.8%), and in the range of *per* (54.8%), *zeste* (53.6%), and *yp2* (49.7%).

Selective Sweeps during Speciation

To explore the role of adaptations characterized by selective sweeps and genetic hitchhiking in the process of speciation, we will consider the different types of gene trees that are expected under several different models. To begin, consider a locus that has not been subject to selective sweeps and consider a single DNA sequence taken from each of two recently diverged species. The time since the most recent ancestral sequence is a function of the time since the species formed and ceased gene exchange, as well as the amount of divergence between the two ancestral sequences at the time of speciation. The latter component is a function of the effective population size of the ancestral species prior to speciation, in exactly the same way that the divergence between two randomly selected sequences from an extant species is expected to reflect the effective population size of that species (see, e.g., Gillespie and Langley 1979).

The component of interspecific variation that is due to ancestral polymorphism may be missing if the locus is repeatedly subject to selective sweeps and genetic hitchhiking. If a selective sweep occurred within the ancestral species prior to speciation, then the hitchhiking pattern of low intraspecific variation will have been present in the ancestral species at the time of speciation. Both *asense* and ci^D were selected because their location in regions of low recombination may correspond to a history of repeated selective sweeps and genetic hitchhiking. Therefore, we expect shorter gene trees between species, in terms of time (not mutations; see Results, *Evolutionary Constraint*, for a discussion of why hitchhiking loci may have less constraint and a higher rate of accumulation of mutations), at these loci than at other loci from regions of high recombination. This reduction in time depth will seem slight if the time since speciation is large relative to the expected depth of intraspecific genealogies, as may well be the case when *D. simulans* and *D. melanogaster* are compared. Indeed, this effect has not been noted in studies reporting evidence of hitchhiking in comparisons between these species (Begun and Aquadro 1991; Berry et al. 1991; Martin-Campos et al. 1992; Langley et al. 1993). However, for the recently formed species of the *simulans* complex, reduced interspecific divergence because of a lack of ancestral polymorphism at hitchhiking loci may be apparent in comparisons with other loci.

Figure 5a depicts, with widely spaced parallel lines, the splitting of a single species into two. The time at which the two species are formed and gene exchange ceases is depicted with a horizontal dashed line. We assume that gene flow ceased at the same time for all loci under study. This model serves as a null model in which natural selection associated with the speciation event has not affected gene flow of the loci under investigation (alternative models are described below). Within the wide lines of the "species tree" in figure 5a are two genealogies, each representing the history of a sample of two gene sequences. The dotted line represents a locus that undergoes repeated hitchhiking and that had a short genealogy within the ancestral species prior to speciation. The solid line represents a locus that has not undergone repeated hitchhiking and that had a relatively deep genealogy within the ancestral species.

We can now consider two different situations in which the gene tree of the hitchhiking loci may be different from that depicted in figure 5a. (1) Consider a model of speciation in which gene exchange does not cease for all loci at the same time, but rather that reproductive isolation develops because of natural selection against gene flow. Furthermore, assume that this selection results from the presence of different locally adaptive alleles in the different populations. Under this model, gene flow ceases first for those loci that undergo adaptations serving only one of the two incipient species. In this situation, selective sweeps may play a causative role in the formation of the species if they are caused by the local fixation of adaptive alleles. The gene trees of selected and linked loci will extend deeper into the

ancestral species than those of other loci not linked to this type of selective sweep. This model closely resembles the situation apparent at hitchhiking loci in populations of *D. ananassae* (Stephan and Mitchell 1992). Figure 5b depicts a deeper gene tree for a hitchhiking gene than for another locus that experienced gene flow more recently. (2) If speciation occurs in such a way that gene flow remains possible but is generally prevented by natural selection, then some selective sweeps may act against the further divergence of the species. In this view, there may be some mutations that are favorable within both species and, if reproductive isolation is not complete, a selective sweep may proceed through both species. The genealogy for a locus that recently underwent such a "transspecies" sweep might be much shorter than for other loci (fig. 5c).

If the null model (fig. 5a) is correct, then, during the formation of the *simulans* complex species, gene flow ceased at the same time for *ci*^{D}, *asense, per, zeste,* and *yp2*. To compare the divergence levels among the loci and assess our null model, we require a measure of interspecific divergence that does not include the variation caused by ancestral polymorphism. We focus on net interspecific divergence (Nei and Tajima 1981, table 2), which equals the average pairwise divergence between species less the average of the two species' intraspecific variation. If the effective size of the ancestral population prior to speciation was equal to the average effective population size of the descendant species, then net divergence equals twice the average number of mutations that have occurred since the speciation event (e.g., see Hudson et al. 1987). This same population size assumption that we now add to the null model is also used for the HKA test (Hudson et al. 1987).

We now describe the use of net divergence in a statistical test of whether the pattern of variation among a set of loci from different species is consistent with a model in which the timing of the cessation of gene flow during species formation is the same for all of the loci. For each locus, we determined the quantity K_1, the net divergence between *D. mauritiana–D. sechellia* divided by two, which is an estimate of the number of mutations that have accumulated on a lineage within a species since the formation of the species. These values cannot be compared among loci, however, unless we control for sequence length and mutation rate. To do so, we determined the quantity K_2, an estimate of the

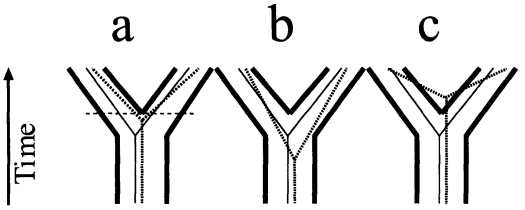

FIG. 5. Graphical representations of species trees with hitchhiking and nonhitchhiking genealogies. See text for detailed explanation. For each model, the dotted line represents an expected genealogy for two DNA sequences of a locus that is repeatedly subject to selective sweeps and genetic hitchhiking, the solid line represents an expected genealogy of a nonhitchhiking locus.

number of mutations that separate *D. melanogaster* from the species ancestral to *D. mauritiana* and *D. sechellia* at the time of their most recent common ancestor. K_2 was calculated by averaging the net divergence between *D. melanogaster* and each of the island endemics and subtracting K_1 from this. The statistical test then follows a simple contingency table design (i.e., 2 × the number of loci). The values for K_1 and K_2 are shown in table 6. The most apparent outlier is *ci*^{D}, which has a low value of K_1 relative to K_2. This apparent discrepancy is in the direction expected of the model of "transspecies" sweeps depicted in figure 5c. The hypothesis of equanimity among all five loci for the relative levels of K_1 and K_2 is rejected ($G = 13.31$, 4 df, $P \leq 0.01$). This finding holds up if we just compare *ci*^{D} to *per, zeste,* and *yp2* ($G = 11.80$, 3 df, $P \leq 0.0081$). The pattern for *asense* is in the same direction as *ci*^{D}, although the test of *asense* with the other X-linked loci is not significant ($G = 5.423$, 3 df, $P \leq 0.1433$). For completeness we carried out the test on just *per, zeste,* and *yp2,* and it is not significant ($G = 2.39$, 2 df, $P < 0.303$).

The use of the G statistic in the contingency-table test assumes that all of the mutations that contribute to K_1 and K_2 accrue independently of each other (Sokal and Rohlf 1981, p. 696). However, this is not strictly the case within each class, and there is expected to be a significant stochastic variance among loci for net divergence (Tajima 1983). To avoid this in the statistical test, multispecies coalescent simulations were carried out (see Materials and Methods, *Simulations*) and an X^2 statistic was compared with a distribution of 5000 simulated values. The conclusions remain, though the probability associated with the

TABLE 6. Distances from the *mauritiana–sechellia* divergence. K_1 is an estimate of the distance from the *mauritiana–sechellia* ancestor to present day *Drosophila mauritiana* and *D. sechellia*. K_2 is an estimate of the distance from the *mauritiana-sechellia* ancestor to present-day *D. melanogaster*.

	Locus					
Divergence	*zeste*	*per*	*yp2*	*adh*	*asense*	*ci*^{D}
K_1	5.4	15.7	3.7	8.0	1.7	2.0
K_2	29.2	46.3	24.4	35.3	23.7	51.5

G tests are higher than for the simulated values: for all five loci, $P < 0.0248$; for *per, zeste, yp2,* and *ci*^{D}, $P < 0.0288$; and for *per, zeste, yp2,* and *asense,* $P < 0.204$.

Table 6 also shows values for *alcohol dehydrogenase* (*Adh,* located on chromosome 2) calculated from the data of Kreitman (1983), Bodmer and Ashburner (1984), and Cohn and Moore (1988). The relative level of K_1 to K_2 for *Adh* clearly resembles the patterns for *per, zeste,* and *yp2* more so than for the hitchhiking loci. Thus, the significant findings involving *ci*^{D} do not appear to be an artifact of differences between the X chromosome and autosomes.

DISCUSSION

Genetic hitchhiking is frequently invoked to explain the reduced levels of intraspecific variation found in genomic regions of low recombination (Aguade et al. 1989; Stephan 1989; Stephan and Langley 1989; Begun and Aquadro 1991, 1992; Berry et al. 1991; Martin-Campos et al. 1992; Langley et al. 1993). B. Charlesworth et al. (1993) have recently proposed an alternative explanation based on the expectation that, for regions of the genome under tight linkage, the proportion of haplotypes that are linked to deleterious mutations will be larger than that for high recombination regions. They suggest that, in large regions of tight linkage, N_e would be reduced by that proportion of haplotypes carrying deleterious mutations. The "background selection" model is discussed specifically within the context of the observation of low variation in *D. melanogaster.* The authors conclude that, although deleterious mutations may make a significant contribution to the reduced levels of variation, the model cannot account for the complete lack of variation found in some portions of the genome and in some populations.

The background selection model also does not fit the pattern of variation seen at *asense,* in which we observed reduced variation in *D. simulans* but not in the closely related *D. mauritiana.* Under the B. Charlesworth et al. (1993) model, the number of haplotypes in the population not carrying deleterious mutations is the zero class sampled from a Poisson distribution. Since the sample size will typically be of or near the order of the actual population size, the variance of the size of this class will be negligible. In other words, the Charlesworth et al. model predicts that effective population sizes be reduced by an essentially constant fraction, across species and populations. If both *D. simulans* and *D. mauritiana* have the same loci and similar levels of recombination on the tip of the X chromosome, then they should all experience the background selection effect to a similar degree.

cubitus interruptus

The original report of Berry et al. (1991) revealed a marked reduction of DNA sequence variation in *D. melanogaster* and *D. simulans.* We have found the same pattern in *D. mauritiana,* where base-pair heterozygosity is 0.00028 (because of a single polymorphism), whereas the average (weighted by gene length) for *zeste, per,* and *yp2* is 0.00898. We also found no variation in *D. sechellia,* but we have little power to determine whether *D. sechellia* has undergone a selective sweep, as the species has low intraspecific variation at other loci (Cariou et al. 1990; Hey and Kliman 1993; Kliman and Hey 1993a). Thus, in all three species examined (discounting *D. sechellia*), *ci*^{D} appears to have undergone recent hitchhiking. Further evidence that *ci*^{D} is frequently linked to selective sweeps comes in the form of very low codon bias and relatively high levels of amino acid replacement changes. This finding is consistent with the population-genetic prediction that regular selective sweeps of this region would not allow purifying selection to purge slightly deleterious mutations (i.e., suboptimal silent and replacement changes).

asense

Our data for *D. melanogaster* and *D. simulans* are consistent with several recent studies of variation within the *yellow-achaete-scute* region (Aguade et al. 1989; Eanes et al. 1989; Begun and Aquadro 1991, 1993; Martin-Campos et al. 1992). Two recent RFLP studies found that variation in this region is very reduced in *D. simulans*

and somewhat reduced in North American and European populations of *D. melanogaster* (Begun and Aquadro 1991; Martin-Campos et al. 1992). Consistent with these reports are our finding of zero polymorphisms in *D. simulans* and the conclusion of a probable recent hitchhiking event in that species and our finding of no variation among North American sequences of *D. melanogaster*. Our finding of five polymorphic sites between two *D. melanogaster* sequences from Kenya is also consistent with previous reports. Other studies with African populations of *D. melanogaster* on genes in the *yellow-achaete-scute* region have reported higher levels of variation compared to findings from European and North American populations (Eanes et al. 1989; Begun and Aquadro 1993).

The major contrast with ci^D is that just one species (*D. simulans*) shows evidence for recent hitchhiking near *asense,* whereas the other species do not. When compared with the gene trees from *per, zeste,* and *yp2,* the *asense* gene tree (fig. 3) seems congruent with the patterns of speciation suggested by those loci. The branching pattern of the *asense* gene tree is similar to those of *per, zeste,* and *yp2,* except for the collapse of *D. simulans* to a single lineage. *Drosophila mauritiana* is also segregating a potentially old length polymorphism which is fixed in *D. sechellia* yet absent from the *D. simulans* sample. This is consistent with the suggestion of Hey and Kliman (1993) that the formation of *D. mauritiana* involved a large population size, and that the species is still segregating ancient polymorphisms.

We suggest that the *asense* region of the genome is subject to selective sweeps, but not at such a high rate as suggested for the fourth chromosome by ci^D. First, the levels of within-species variation at *asense* for *D. melanogaster* and *D. mauritiana* were within or near the range observed for other loci, whereas in *D. simulans* the within-species variation was reduced to zero. Second, the effective amount of evolutionary constraint on *asense,* based on levels of codon bias and of amino acid replacement changes, appears to lie between that of ci^D and the more freely recombining X-linked loci. We suggest, therefore, that natural selection is more effective at the tip of the X chromosome than on the fourth chromosome because there is some recombination in the former and, hence, less threat of fixation of suboptimal haplotypes via genetic hitchhiking associated with selective sweeps of highly advantageous linked alleles.

Selective Sweeps and Species Formation

A major motivation of this research was the complex evolutionary question: to what extent does the process of adaptation by natural selection contribute to or detract from the accumulation of reproductive barriers during species formation? In taking a genealogical perspective, we are restricted to those types of adaptations characterized by rapid selective fixation of mutations. Furthermore, we are restricted to those portions of the genome where recombination is low and where selective sweeps are expected to determine the genealogy of relatively large linked portions of the genome.

One possible outcome (fig. 5b; see Results, *Selective Sweeps and Species Formation*) does not emerge from our data. If selective sweeps near *asense* or ci^D had contributed to the differential adaptation of interbreeding populations that went on to become species, then divergence between species at these genes should be greater than for other nonhitchhiking loci. However, among the *simulans* complex species, *asense,* and ci^D have less divergence than *zeste, per,* and *yp2* (table 3). Stephan and Mitchell (1992) describe a pattern of variation at two hitchhiking loci, *v* and *fw,* in two Asian populations of *D. ananassae* that is in some ways consistent with this model. Both *v* and *fw* appear to have recently undergone different selective sweeps, and the two populations share no polymorphism. In contrast, the two populations share several RFLPs at the *Om(1D)* (Stephan 1989) and *f* (Stephan and Langley 1989) loci.

We do have limited evidence for a model (fig. 5c) in which a fourth-chromosome selective sweep was shared by the incipient species of the *simulans* complex, at a time when divergence at other loci had already begun to accumulate. The ci^D gene tree (fig. 2) has a trichotomy for the sequences from these three species, and the distance from this node to the present is short relative to the distance to present day *D. melanogaster.* The same comparison for *zeste, per,* and *yp2* suggests a more ancient split for the *simulans* complex species. It appears as if a selective sweep of the fourth chromosome occurred in the *simulans* complex relatively late in the speciation process that led to present day *D. simulans, D. sechellia* and *D. mauritiana.* The ci^D data also does not show the pattern of divergence among the *simulans* complex species exhibited by the X-linked loci. In that pattern, the distance be-

tween *D. mauritiana* and *D. sechellia* was close to the sum of the distances between each of these species and *D. simulans*. At ci^D, the sequences of all three species appear to diverge from the same point in time and at similar rates, though very few mutations have accumulated.

The conclusion of a fourth-chromosome "transspecies" sweep is tentative. Because of the relatively slow accumulation of mutations in *D. simulans* since the formation of the *simulans* complex, the *K*-ratio test was done using *D. mauritiana* and *D. sechellia* (table 6). The test is not significant if either of the island endemic species is paired with *D. simulans*. The *K*-ratio test also makes the assumption that mutations that contribute to K_2 accumulate at the same rate as those that contribute to K_1. We cannot rule out the possibility that the high value for K_2 at ci^D is the result of an accelerated mutation rate along that part of the tree.

If this interpretation of a "transspecies" sweep is true, it follows that fourth chromosome selective sweeps had a homogenizing effect and acted to reduce the divergence among the emerging species. It would also suggest that natural selection favoring reproductive isolation was not so strong that the selective differential associated with a fourth chromosome selective sweep was sufficient to overcome barriers to gene flow.

Acknowledgments

This research was supported by National Science Foundation grant BSR8918164 to J.H.

Literature Cited

Aguade, M., N. Miyashita, and C. H. Langley. 1989. Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics 122:607–615.

Ashburner, M. 1989. *Drosophila:* a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Beech, R. N., and A. J. Leigh Brown. 1989. Insertion-deletion variation at the *yellow-achaete-scute* region in two natural populations of *Drosophila melanogaster*. Genetical Research 53:7–15

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.

———. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356:519–520.

———. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365:548–550.

Berry, A. J., J. W. Ajioka, and M. Kreitman. 1991. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics 129: 1111–1117.

Birky, C. W., and J. B. Walsh. 1988. Effects of linkage on rates of molecular evolution. Proceedings of the National Academy of Sciences, USA 85:6414–6418.

Bodmer, M., and M. Ashburner. 1984. Conservation and change in the DNA sequences coding for *alcohol dehydrogenase* in sibling species of *Drosophila*. Nature 309:425–430.

Bulmer, M. 1988. Are codon usage patterns in unicellular organisms determined by selection-mutation balance? Journal of Evolutionary Biology 1:15–26.

Cariou, M.-L., M. Solignac, M. Monnerot, and J. R. David. 1990. Low allozyme and mtDNA variability in the island endemic species *Drosophila sechellia* (*D. melanogaster* complex). Experientia 46: 101–104.

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

Charlesworth, D., M. T. Morgan, and B. Charlesworth. 1993. Mutational accumulation in finite outbreeding populations. Genetical Research 61:39–56.

Cohn, V. H., and G. P. Moore. 1988. Organization and evolution of the *alcohol dehydrogenase* gene in *Drosophila*. Molecular Biology and Evolution 5:154–166.

Dubinin, N. P., N. N. Sokolov, and G. G. Tiniakov. 1937. Crossing over between the genes *yellow, achaete,* and *scute*. Drosophila Information Service 8:76.

Eanes, W. F., J. Labate, and J. W. Ajioka. 1989. Restriction map variation associated with the *yellow-achaete-scute* region in five populations of *Drosophila melanogaster*. Molecular Biology and Evolution 6:492–502.

Felsenstein, J. 1974. The evolutionary advantage of recombination. Genetics 78:737–756.

Gillespie, J. H., and C. H. Langley. 1979. Are evolutionary rates really variable. Journal of Molecular Evolution 13:27–34.

Gonzalez, F. S., Romani, P. Cubas, J. Modolell, and S. Campuzano. 1989. Molecular analysis of the asense gene, a member of the *achaete-scute* complex of *Drosophila melanogaster,* and its novel role in optic lobe development. EMBO 8:3553–3562.

Hendy, M. D., and D. Penny. 1982. Branch and bound algorithms to determine minimal evolutionary trees. Mathematical Bioscience 59:277–290.

Hey, J., and R. M. Kliman. 1993. Population genetics and phylogenetics of the DNA sequence variation at multiple loci within the *Drosophila melanogaster* complex. Molecular Biology and Evolution 10:804–822.

Hill, W. G., and A. Robertson. 1966. The effect of linkage on limits to artificial selection. Genetical Research 8:269–294.

Hochman, B. 1976. The fourth chromosome of *Drosophila melanogaster*. Pp. 903–928 *in* A. Ashburner and E. Novitski, eds. The genetics and biology of *Drosophila,* vol. 1B. Academic Press, New York.

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology 23:183–201.

———. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* P. H. Harvey and L. Partridge, eds. Oxford Surveys in Evolutionary Biology, Vol. 7. Oxford University Press, New York.

Hudson, R. R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Molecular Biology and Evolution 5:568–583.

Kaplan, N., R. R. Hudson, and C. H. Langley. 1989. The "hitchhiking effect" revisited. Genetics 123: 887–899.

Kliman, R. M., and J. Hey. 1993a. DNA sequence variation at the *period* locus within and among species of the *Drosophila melanogaster* complex. Genetics 133:375–387.

———. 1993b. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Molecular Biology and Evolution 10:1239–1258.

Kreitman, M. 1983. Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. Nature 304:412–417.

———. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphic divergence. Genetics 127:565–582.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas, and M. Ashburner. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. Evolutionary Biology 22:159–225.

Langley, C., J. M. MacDonald, N. Miyashita, N. and M. Aguade. 1993. Lack of correlation between interspecific divergence and intraspecific polymorphism at the *suppressor of forked* region in *Drosophila melanogaster* and *Drosophila simulans*. Proceedings of the National Academy of Sciences, USA 90:1800–1803

Li, W.-H. 1987. Models of nearly neutral mutations with particular implication for nonrandom usage of synonymous codons. Journal of Molecular Evolution 24:337–345.

Martin-Campos, J. M., J. M. Comeron, N. Miyashita, and M. Aguade. 1992. Intraspecific and Interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. Genetics 130:805–816.

Maynard Smith, J., and J. Haigh. 1974. The hitchhiking effect of a favorable gene. Genetical Research 23:23–35.

Muller, H. J. 1964. The relation of recombination to mutational advance. Mutation Research 1:2–9.

Nei, M., and F. Tajima. 1981. DNA polymorphism detectable by restriction endonucleases. Genetics 97:145–163.

Orenic, T. V., D. C. Slusarski, K. L. Kroll, and R. A. Holmgren. 1990. Cloning and characterization of the segment polarity gene *cubitus interruptus Dominant* of *Drosophila*. Genes and Development 4:1053–1067.

Sharp, P. M., and W.-H. Li. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15:1281–1295.

Shields, D., P. M. Sharp, D. G. Higgins, and F. Wright. 1988. Silent sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Molecular Biology and Evolution 5:704–716.

Sokal, R. R. and J. F. Rohlf. 1981. Biometry. W. H. Freeman, San Francisco.

Stephan, W. 1989. Molecular genetic variation in centromeric region of the X chromosome in three *Drosophila ananassae* populations. II. The *Om(1D)* Locus. Molecular Biology and Evolution. 6:624–635.

Stephan, W., and C. H. Langley. 1989. Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between *vermillion* and *forked* loci. Genetics 121:89–99.

Stephan, W., and S. J. Mitchell. 1992. Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. Genetics 132:1039–1045.

Sorsa, V. 1988. Chromosome maps of *Drosophila*, Vol. II. CRC Press, Boca Raton, Fla.

Swofford, D. L. 1985. *PAUP version 2. 4.* Illinois Natural History Survey, Champaign.

Tajima, F. 1983. Evolutionary relationships of DNA sequences in finite populations. Genetics 105:437–460.

Corresponding Editor: W. Stephan