The Population Genetics of the Origin and Divergence of the *Drosophila simulans* Complex Species

Richard M. Kliman,*[,1] Peter Andolfatto,[†,2] Jerry A. Coyne,[‡,3] Frantz Depaulis,[§,2] Martin Kreitman,[‡,3] Andrew J. Berry,[‡,4] James McCarter,[‡,5] John Wakeley[,4] and Jody Hey***

**Department of Genetics, Rutgers University, Piscataway, New Jersey 08854-8082, [†]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, [‡]Department of Ecology, Evolution and Behavior, Princeton University, Princeton, New Jersey 08544, [§]Laboratoire d'Ecologie, Université Paris 6, CNRS-UMR 7625 Case 237, 75252 Paris Cedex 05, France and **Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138*

Manuscript received April 14, 2000
Accepted for publication September 11, 2000

ABSTRACT

The origins and divergence of *Drosophila simulans* and close relatives *D. mauritiana* and *D. sechellia* were examined using the patterns of DNA sequence variation found within and between species at 14 different genes. *D. sechellia* consistently revealed low levels of polymorphism, and genes from *D. sechellia* have accumulated mutations at a rate that is ~50% higher than the same genes from *D. simulans*. At synonymous sites, *D. sechellia* has experienced a significant excess of unpreferred codon substitutions. Together these observations suggest that *D. sechellia* has had a reduced effective population size for some time, and that it is accumulating slightly deleterious mutations as a result. *D. simulans* and *D. mauritiana* are both highly polymorphic and the two species share many polymorphisms, probably since the time of common ancestry. A simple isolation speciation model, with zero gene flow following incipient species separation, was fitted to both the *simulans/mauritiana* divergence and the *simulans/sechellia* divergence. In both cases the model fit the data quite well, and the analyses revealed little evidence of gene flow between the species. The exception is one gene copy at one locus in *D. sechellia*, which closely resembled other *D. simulans* sequences. The overall picture is of two allopatric speciation events that occurred quite near one another in time.

SEVERAL hundred thousand years ago one species of Drosophila gave rise to three that today we call *Drosophila simulans*, *D. mauritiana*, and *D. sechellia*. Today the three species are morphologically distinct (primarily on the basis of male genitalia), partially intersterile (male hybrids are sterile, female hybrids fertile), and largely allopatric (*D. simulans* is a nearly cosmopolitan human commensal, while the other two are island endemics). The combination of clear phenotypic distinction, partial infertility, and recent coancestry (not to mention their evolutionary proximity to *D. melanogaster*) has made this little species complex our most thoroughly studied speciation model system (Coyne and Kreitman 1986; Coyne 1992; Wu and Palopoli 1994; Coyne and Charlesworth 1997).

Corresponding author: Jody Hey, Department of Genetics, Rutgers University, 604 Allison Rd., Piscataway, NJ 08854-8082.
E-mail: jhey@mbcl.rutgers.edu

[1]*Present address:* Department of Biological Sciences, Kean University, 1000 Morris Ave., Union, NJ 07083.

[2]*Present address:* ICAPB, University of Edinburgh, King's Bldg., W. Mains Rd., Edinburgh EH9 3JT, United Kingdom.

[3]*Present address:* Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637.

[4]*Present address:* Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138.

[5]*Present address:* Genome Sequencing Ctr., Box 8501, Washington University School of Medicine, 4444 Forest Park Pkwy., St. Louis, MO 63108.

Historically there have been two main approaches to the genetic study of species divergence. The classical approach is to genetically map traits that are thought to be important in speciation. Such traits tend to fall into one of three categories.

1. The most straightforward are those for which the species exhibit characteristic differences and that probably represent species-specific adaptations. Major lifestyle or life history adaptations can, in principle, play a large direct role in speciation, particularly if those changes arise first as polymorphisms within the ancestral species (Bush 1969; Rice and Hostert 1993). For example, the preferred host of *D. sechellia*, *Morinda citrifolia*, is toxic to the other species of the *D. melanogaster* complex, and the genes that confer resistance can be mapped in the species hybrids (Jones 1998).
2. A second class of "speciation" traits are those that are features of mating pairs of organisms. In recent years, a host of interesting Drosophila mating phenotypes have come under focus, including species-specific mate detection pheromones (Coyne *et al.* 1994; Coyne and Charlesworth 1997), sperm competition (Snook *et al.* 1994; Price *et al.* 1999), and female mediation of sperm competition (Price 1997).
3. The third class of speciation traits are those that appear almost exclusively in species hybrids. Exam-

ples of these are hybrid inviability and hybrid sterility. In general such traits can be genetically studied only in species where postzygotic isolation is incomplete, which is necessary for the production of F_2s or backcrosses.

A different genetic approach to understanding speciation is to study the history of species divergence as it is revealed in the polymorphism pattern at randomly selected genes. In recent years, comparative DNA sequence data (especially mitochondrial) have been frequently used to address basic questions about the relatedness of close sister taxa and populations (Avise 1989). Conceptually, the idea is a direct extension of basic population genetic questions (*i.e.*, questions about population subdivision, gene flow, and natural selection) to the species level. However, the use of DNA sequence data also permits the use of genealogical coalescent models, which incorporate classical population genetic parameters (*e.g.*, effective population size and migration rate) within a gene tree framework (Hudson 1990), as well as the entire suite of tools used by molecular phylogeneticists. These methods become even more informative when data come from multiple loci and thus can be used to distinguish forces that act on all genes from those, like natural selection, that affect individual loci (Hudson *et al.* 1987; Hey 1994). For the sake of a useful label, we refer to this general approach in the remainder of the article as divergence population genetics (DPG).

The two approaches (the mapping of speciation phenotypes and DPG) have historically been directed at very different questions. The gene mapping studies address the genetic architecture of phenotypes that may have been important in speciation; however, these maps bear no direct connection to the demographic factors that have caused species, and they may not have a direct connection to the selective factors that have actually caused species. All of the speciation phenotypes listed above can arise during or following speciation that is primarily caused by selection on other phenotypes. Thus, for example, hybrid sterility and inviability may arise as epistatic by-products of independent adaptations in the separate incipient species (Dobzhansky 1936; Muller 1940). In contrast to the gene map approach, divergence studies can focus directly on evolutionary forces, particularly those demographic factors that affect all of the genes in the genome.

The two approaches can greatly complement one another, such as when interpretations of the evolution of phenotypic traits are laid upon an understanding of phylogenetic history. For example, recent attempts to demonstrate sympatric speciation and assess its frequency have strongly relied upon accurate branching phylogenies (Schliewen *et al.* 1994; Barraclough and Vogler 2000; Coyne and Price 2000). However, phylogenetic history, particularly for recent speciation events, may not be well represented by simple branching trees. Most real species probably emerge gradually whereas a branching model entails the assumption of instantaneous splitting. This can be misleading particularly if speciation has been recent or there have been multiple speciation events that overlap in their time course. In such cases we must consider "phylogeny" more broadly as concerning the genesis of phyla, however complex or slow that process might have been. If speciation events have been recent, and if they have been complex and not instantaneous, it may be possible to reveal the complexities using a population genetics approach.

This report brings together the efforts of several investigators interested in the speciation events that have led to our current *simulans* complex species. To date, DPG studies on the *simulans* complex have been done for five different nuclear loci (Hey and Kliman 1993; Kliman and Hey 1993a; Hilton *et al.* 1994). Here we report on patterns of DNA sequence divergence at an additional nine loci. Together these data permit a broad, genome-wide assessment of speciation.

MATERIALS AND METHODS

The data for five loci, *per*, *yp2*, *z*, *ase*, and *ci*, have previously been described (Hey and Kliman 1993; Kliman and Hey 1993a; Hilton *et al.* 1994). DNA sequences were collected from multiple lines of each species of the *simulans* complex for each of nine additional loci. For some of these genes, new data were aligned with existing, previously reported data for some species. For all loci, at least one sequence from *D. melanogaster* was also available.

***Zw*, *Adh*, and *est-6*:** DNA sequences had previously been reported for *D. melanogaster* and *D. simulans* for these genes (Kreitman 1983; Cooke and Oakeshott 1989; McDonald and Kreitman 1991; Eanes *et al.* 1993, 1996; Karotam *et al.* 1993). DNA was extracted from single individuals of *D. mauritiana* and *D. sechellia* drawn from isofemale lines that had been in the laboratory for >200 generations. PCR on these genomic DNAs was done to generate a 1.3-kb region of *Zw*, a 0.87-kb region of *Adh*, and a 1.6-kb region of *est-6*. The sequenced region of *Zw* corresponds to sites 148–1460 in a previously reported *D. simulans* sequence (Eanes *et al.* 1993). The sequenced region of *Adh* corresponds to sites 2195–2900 in a previously reported *D. simulans* sequence (Cohn *et al.* 1984). The sequenced region of *est-6* corresponds to sites 157–1679 of a previously reported *D. melanogaster* sequence (Cooke and Oakeshott 1989). Primary PCR products were purified from an agarose gel slice using the QIAquick gel extraction kit (QIAGEN, Valencia, CA) and were then used as the template for subsequent amplification of shorter regions for DNA sequencing. These reactions employed one primer carrying an M13 forward tail, while the other primer carried an M13 reverse tail. Subsequent sequencing reactions were done in both directions simultaneously using fluorescently labeled M13 primers on a Li-Cor (Lincoln, NE) automated DNA sequencer. For *D. mauritiana*, six sequences were used. Four of these (lines 105, 197, 152, and 207) were the same as those used previously for *per*, *yp2*, *z*, *ase*, and *ci* (designated in those articles as MA-3, MA-4, MA-5, and MA-6, respectively). For *D. sechellia*, two sequences were obtained, one from SE-C1 (also called strain 24) and one from SE-P1 (also called strain ss77; Kliman and Hey 1993a).

***janus*:** Two loci, *janus-A* and *janus-B*, are overlapping and related by an ancient duplication (Yanicostas *et al.* 1989). Three lines of *D. mauritiana* and *D. simulans* were sequenced, as well as one line each from *D. melanogaster* and *D. sechellia*. To isolate just one allele from each isofemale line, single males were crossed with a virgin female from a balanced lethal strain of *D. melanogaster* [Df(3R)X3F/TM3Sb + P[ry + .RP49].84F] deficient for a region including the *janus* loci. Following DNA extraction from single F_1 individuals, PCR was conducted using the primers from positions 639–658 of GenBank record DMSRYG1 and positions 1736–1717 of GenBank record DROJAN. DNA sequencing was conducted with internal primers spaced approximately every 250 bp, using an Amersham (Piscataway, NJ) sequenase T7 kit and corresponding protocol. The final sequence spanned bases 429–1491 of Yanicostas *et al.* (1989), including most of the *janus-A* locus and part of the *janus-B* locus. Throughout the analyses, these regions were treated as a single locus. *D. simulans* strains were provided by C. Montchamp Moreau; *D. mauritiana* strains and the *D. sechellia* strain were provided by F. Lemeunier.

***hb*, *mt:ND5*, *Sxl*, and *w*:** Portions of each of these loci were sequenced from single flies drawn from inbred lines of each species that were collected or obtained from others. For *D. mauritiana*, 1 line was obtained from H. Robertson and 9 from O. Kitagawa. For *D. sechellia*, 8 isofemale lines were collected from the Seychelles in 1985. These 8 lines were sequenced for each of these genes. In addition, 6 lines of *D. sechellia* were collected from the Seychelles in 1989. These lines were sequenced only for the *Sxl* locus. For *D. simulans*, 3 lines (from France, Tunisia, and Kenya) were obtained from the Drosophila species stock center, and 13 lines were collected from diverse locations, including Florida City, FL; Beltsville, MD; Murakata City, Japan; Palmer Island, Australia; Ottawa, Canada; Cairns, Australia; Capetown, South Africa; Brazzaville, Congo; Morven, GA; and Praslin, Seychelles.

DNA sequencing was done using templates generated via PCR on genomic DNA. PCR was done using a kinased primer and was followed by treatment with λ exonuclease to degrade one strand (Higuchi and Ochman 1989). The DNAs were sequenced with the dideoxy method with [^{35}S]dATP label (Sanger *et al.* 1977).

For *hb* (hunchback), the sequenced region corresponds to intronic sequence from positions 7769–8052 of *D. melanogaster* GenBank record U17742. For *mt:ND5* (mitochondrial NADH-ubiquinone oxidoreductase chain 5), the sequenced region corresponds to positions 7256–7472 of *D. melanogaster* GenBank record U37541. For *Sxl* (Sex Lethal), the sequenced region corresponds to intronic sequence from positions 241722–241977 of *D. melanogaster* GenBank record AE003439. For *w* (white), the sequenced region corresponds to intronic sequence from positions 12260–12478 of *D. melanogaster* GenBank record X02974.

***In(2L)t*:** *In(2L)t* refers to the *D. simulans*/*D. mauritiana*/*D. sechellia* homologue of the proximal breakpoint site of the *In(2L)t* inversion that segregates in natural populations of *D. melanogaster* (Andolfatto *et al.* 1999). *D. simulans* isofemale lines were collected from Arena Farms, Maryland. The lines for *D. sechellia* include the original "Robertson" isofemale line collected and described by Tsacas and Baechli (1981) and provided by Hugh Robertson and two lines collected from Cousin Island, Seychelles, in 1985. The *D. mauritiana* lines were provided by Chung-I Wu. To obtain alleles from *D. simulans* (Arena Farms, Maryland) and *D. mauritiana* populations, multiple males from each isofemale line were crossed to virgin female *D. melanogaster In(2L)t* homozygotes. The resulting hybrid progeny (all female) were heterozygous for *In(2L)t*. This allowed the recovery of individual (one per isofemale line) *D. simulans* and *D. mauritiana* alleles by PCR with a standard arrangement-specific primer pair (Andolfatto *et al.* 1999). For *D. sechellia*, genomic DNA was prepared, one individual per isofemale line. Due to the unexpectedly high degree of similarity between one *D. simulans* (ar07) and one *D. sechellia* allele (from the Robertson line), male genitalia were checked for both lines and both alleles were resampled and sequenced. Polyethylene glycol-precipitated PCR products were directly sequenced on both strands using a Rhodamine Terminator cycle sequencing kit (Applied Biosystems, Foster City, CA) and run on an ABI377XL automated sequencer.

GenBank accession numbers: Accession numbers are as follows: for *Zw*, *Adh*, and *est-6*, AF284474–AF284497; *janus*, AF284453–AF284459; *hb*, AF295808–AF295835; *mt:ND5*, AF295836–AF295861; *Sxl* and *w*, AF295862–AF295921; *In(2L)t*, AF294398–AF294409 and AF217926–AF21791.

RESULTS

Polymorphism summaries: Sample sizes and basic statistics of the loci studied are listed in Table 1. DNA sequence variation is summarized in Table 2. A simple weighted average of nucleotide diversity per base pair shows *D. simulans* to be the most variable, followed by *D. mauritiana* and *D. sechellia*. For the autosomal loci the weighted average values of $\hat{\theta}$/bp were 0.015, 0.011, and 0.003 for these three species, respectively. The corresponding X chromosome values were slightly less than one-half of the autosomal values, at 0.007, 0.005, and 0.001. The rankings are unchanged from those originally reported for fewer loci (Hey and Kliman 1993). The simplest interpretation of these patterns is that the historical effective population sizes have been largest in *D. simulans* and smallest in *D. sechellia*. Both *D. simulans* and *D. mauritiana* have higher levels of polymorphism than reported for *D. melanogaster* (Hey and Kliman 1993; Moriyama and Powell 1996).

Tests of selective neutrality: To focus on demographic factors associated with the divergence of species, we first addressed whether the data show evidence that natural selection has shaped levels of variation. Table 3 shows the results of contingency table tests in which variable sites are classified both with respect to whether they are polymorphisms within species or fixed differences between species and whether they occurred at synonymous or replacement sites within the protein-coding regions. Under a model in which all mutations are either deleterious or neutral, the expected ratio of synonymous to amino acid replacement variation should be the same for polymorphisms and for fixed differences between species. Three loci (*est6*, *janus*, and *Zw*) revealed a poor fit to the neutral model (Table 3), and in each case the direction is the same as had previously been reported for *Adh* (McDonald and Kreitman 1991) and *Zw* (Eanes *et al.* 1993) in contrasts involving *D. simulans* and *D. melanogaster*. If we assume that the pattern of synonymous site variation is close to that expected for neutral mutations, then the direction of departure for these tests is one in which the number of fixed replacement differences between species is higher

TABLE 1

Gene and sample summary statistics

Locus	n_{sim}[a]	n_{mau}	n_{sec}	Map[b]	E_{nc}[c]	Rec[d]	Length[e]	Coding	Noncoding
Adh	6	6	2	2 35B2	31.2	0.0020	698	567	131
ase	6	5	6	X 1B4	55.7	0.0000	1067	1067	0
ci	9	6	4	4 101F1	51.8	0	1075	957	118
est6	4	6	2	3 69A1	54.6	0.0027	1520	1470	50
hb	10	9	8	3 85A6	45.2	0.0006	252	0	252
In(2L)t	11	9	3	2 34A8	NA	0.0026	680	0	680
janus	3	3	1	3 99D4	54.6	0.0037	1072	558	514
mt:ND5	9	8	8	mt	25.2	0	277	277	0
per	6	6	6	X 3B2	41.2	0.0029	1870	1677	193
Sxl	13	9	14	X 6F5	51.5	0.0038	236	0	236
w	9	7	6	X 3C2	47.7	0.0034	194	0	194
yp2	6	6	6	X 9A2	33.1	0.0029	1100	1044	56
z	6	6	6	X 3A3	45.9	0.0027	980	799	181
Zw	12	6	2	X 18E	34.5	0.0023	1298	1164	134

[a] The sample sizes for each locus and species.

[b] The genomic location of the gene in *D. melanogaster*. The physical maps of these species are all identical to each other and nearly identical to *D. melanogaster* with the exception of one inverted segment on the third chromosome (sections 84F–93F; Lemeunier and Ashburner 1976). The chromosome (X, 2, 3, or 4) is shown followed by the polytene chromosome band number (Bridges 1938) as listed in Flybase (http://flybase.bio.indiana.edu:82/). The *mt:ND5* gene is located in the mitochondrial (mt) genome.

[c] The effective number of codons (Wright 1990). Values near the upper limit of 61 have a low codon bias, whereas values near the lower limit of 20 have a high codon bias. Values were taken from Kliman and Hey (1993b) or calculated from the complete coding portion of the gene as reported in the GenBank record. NA, not applicable; *In(2L)t* was positionally cloned and has no associated protein coding gene.

[d] The estimated recombination rate per generation as reported for *D. melanogaster* in the Appendix of Kliman and Hey (1993b). Units are centimorgans per kilobase pair.

[e] The average number of aligned base positions from all pairwise comparisons. Values are shown for total length, as well as for amino acid coding regions, and noncoding regions.

than expected. This pattern would result if directional natural selection has caused some amino acid mutations to become fixed within species.

Similar in principle to the McDonald-Kreitman tests in Table 3, the Hudson-Kreitman-Aguadé (HKA) test examines whether the relative levels of observed polymorphism and divergence are consistent across multiple loci. Figure 1 shows the results of HKA tests (Hudson *et al.* 1987). Rather than rely on the assumption that the test statistic follows a χ^2 distribution (Hudson 1987), the overall test statistic was compared with a distribution generated from 10,000 coalescent simulations. Figure 1 shows, for each locus, whether or not the observed values of polymorphism and divergence are higher or lower than expected, and it shows the contribution from each data point to the overall test statistic. In each case the overall test statistic indicates a rejection of the neutral model: *D. simulans*, $\chi^2 = 25.31$, $P = 0.0010$; *D. mauritiana*, $\chi^2 = 18.61$, $P = 0.0390$; and *D. sechellia*, $\chi^2 = 52.84$, $P = 0.0010$. For *D. simulans*, Figure 1 shows how *ase* and *ci* make large contributions to the test statistic, as expected from previous reports (Berry *et al.* 1991; Hilton *et al.* 1994). These two genes are also the only ones in the study from low recombination portions of the genome (as identified in *D. melanogaster*; see Table 1) and thus have probably been subject to collateral selective effects via linkage—genetic hitchhiking (Maynard Smith and Haigh 1974) or background selection (Charlesworth *et al.* 1993). The effect of this indirect selection, whether via beneficial or deleterious mutations, is to reduce polymorphism levels in regions of low recombination while leaving divergence levels typical of those seen among loci (Begun and Aquadro 1992). In the case of *D. mauritiana*, the *ci* and *In(2L)t* loci contributed a large amount to the test statistic. *D. sechellia* presents an interesting situation, for it carries very low polymorphism levels at nearly all loci, almost certainly due to a small effective population size (Cariou *et al.* 1990; Hey and Kliman 1993). Again, both *ase* and *ci* have low polymorphism, but, in this species, neither locus appears unusual, as all loci but one have low polymorphism levels. The exception is *In(2L)t*, which revealed 23 polymorphisms in *D. sechellia*. Upon inspection of the three sequences, two were revealed to be very similar to each other (four differences), while the third closely resembled sequences from *D. simulans*. It is this *simulans*-like sequence that contributes most of the polymorphisms to the *D. sechellia* sample for *In(2L)t*. There are two possible explanations for the observation: limited gene flow in the wild and recent admixture in the laboratory. Gene flow seems reasonable in that *D. sechellia* and *D. simulans* are partially interfertile and

TABLE 2

Polymorphism statistics

Locus	Species	S^a	$\hat{\theta}^b$	$\hat{\theta}$/bp	γ^c	D^d	Div.^{e}
Adh	*sim*	13	5.69	0.0082	—	−0.82	0.0281
	mau	5	2.19	0.0031	—	−1.34	0.0326
	sec	0	0.0	0.0	—	—	0.0279
ase	*sim*	0	0.0	0.0	—	—	0.0252
	mau	5	2.40	0.0023	—	0.00	0.0267
	sec	0	0.0	0.0	—	—	0.0244
ci	*sim*	1	0.37	0.0003	—	−1.09	0.0504
	mau	1	0.44	0.0004	—	−0.93	0.0504
	sec	0	0.0	0.0	—	—	0.0494
est6	*sim*	69	37.64	0.0247	149.0	0.19	0.0507
	mau	36	15.77	0.0104	24.5	−0.50	0.0503
	sec	0	0.0	0.0	—	—	0.0507
hb	*sim*	11	3.89	0.0155	—	−1.89*	0.0324
	mau	6	2.21	0.0084	—	−1.12	0.0353
	sec	0	0.0	0.0	—	—	0.0328
In(2L)t	*sim*	30	10.24	0.015	4.3	0.52	0.0382
	mau	44	16.19	0.0243	41.3	−0.34	0.0407
	sec	23	15.33	0.0239	—	—	0.0415
janus	*sim*	33	22.00	0.0215	23.9[f]	—	0.0517
	mau	29	19.33	0.0190	49.2[f]	—	0.0494
	sec	—	—	—	—	—	0.0586
mt:ND5	*sim*	1	0.37	0.0013	—	−1.09	0.0393
	mau	0	0.0	0.0	—	—	0.0505
	sec	1	0.39	0.0014	—	−1.05	0.0510
per	*sim*	54	23.65	0.0127	44.3	−0.59	0.0349
	mau	48	21.02	0.0113	84.6	0.34	0.0402
	sec	4	1.75	0.0009	—	−0.06	0.0391
Sxl	*sim*	8	2.58	0.0101	42.7	0.33	0.0362
	mau	4	1.47	0.0057	—	−1.15	0.0353
	sec	1	0.31	0.0015	—	0.25	0.0572
w	*sim*	19	7.00	0.0361	12.2	−0.15	0.0704
	mau	3	1.22	0.0066	—	−1.89*	0.0626
	sec	2	0.88	0.0042	—	−1.13	0.0760
yp2	*sim*	3	1.31	0.0012	—	−0.45	0.0258
	mau	4	1.75	0.0016	—	−1.30	0.0285
	sec	1	0.44	0.0004	—	−0.93	0.0280
z	*sim*	18	7.88	0.0080	6.0	−0.07	0.0386
	mau	9	3.94	0.0040	—	0.03	0.0373
	sec	0	0.00	0.0	—	—	0.0387
Zw	*sim*	10	3.31	0.0026	9.3	1.33	0.0360
	mau	10	4.38	0.0033	—	−1.16	0.0364
	sec	0	0.0	0.0	—	—	0.0411

For each gene, the top row shows values for *D. simulans* (*sim*), the second row shows values for *D. mauritiana* (*mau*), and the third row shows values for *D. sechellia* (*sec*). (—)Estimates cannot be obtained for small samples or for samples with few informative polymorphic sites.

[a] The number of polymorphic sites.

[b] An estimate of the population mutation rate $2Gu$, where G is the effective number of gene copies and u is the mutation rate for the region (Watterson 1975). Also shown is the estimate of $2Gu$/bp.

[c] An estimate of the population recombination rate, $2Gc$, where c is the rate of crossing over per generation for the region (Hey and Wakeley 1997).

[d] Tajima's statistic of equality between different estimates of $2Gu$ (Tajima 1989b). Negative values indicate an excess of low frequency polymorphisms. * indicates the value exceeds the expected 95% confidence interval.

[e] Average divergence, per base pair, between the sample sequences and a sequence from *D. melanogaster*.

[f] One outgroup sequence from *D. melanogaster* was included with the small *janus* samples to estimate γ.

that both have been collected on the large island of Mahé (Cariou *et al.* 1990; R'Kha *et al.* 1991). However, no other loci show a pattern suggestive of recent gene flow. The second explanation, recent mixing in the laboratory, also does not fit the observed pattern at *In(2L)t* in a simple way, as the *D. sechellia* line from which

TABLE 3

Contrasting levels of synonymous and replacement variation

Locus	Syn poly	Rep poly	Syn fixed	Rep fixed	G^a	*P*
Adh	9	0	3	1	1.576	0.2093
ase	3	2	2	1	0.029	0.8628
ci	0	2	4	2	2.716	0.0994
est–6	68	18	16	12	4.792	0.0286*
(*sim vs. mau*)	68	18	1	3	4.237	0.0396*
(*sim vs. sec*)	53	11	12	9	5.136	0.0234*
(*mau vs. sec*)	24	9	14	11	1.705	0.1916
janus	14	1	4	4	5.040	0.0248*
(*sim vs. mau*)	14	1	0	0	NA	NA
(*sim vs. sec*)	9	1	4	4	3.306	0.0690
(*mau vs. sec*)	9	1	4	4	3.306	0.0690
mt:ND5	2	0	13	1	0.117	0.7323
per	71	7	16	2	0.068	0.7943
yp2	4	2	2	2	0.237	0.6264
Zw	14	0	8	4	6.240	0.0125*
(*sim vs. mau*)	14	0	2	3	7.953	0.0048**
(*sim vs. sec*)	7	0	8	4	3.738	0.0532
(*mau vs. sec*)	7	0	6	1	0.974	0.3237

McDonald-Kreitman tests of amino acid replacement and synonymous polymorphisms, within and between species (McDonald and Kreitman 1991): Syn poly, synonymous polymorphisms; Rep poly, replacement polymorphisms; Syn fixed, synonymous fixed differences; Rep fixed, replacement fixed differences. Tests include all three species. A site is counted as polymorphic if it is variable in any of the species. For those tests that were statistically significant (* $P < 0.05$; ** $P < 0.01$), tests on individual species pairs are also shown (*sim*, *D. simulans*; *mau*, *D. mauritiana*; *sec*, *D. sechellia*).

[a] Williams' correction is applied to *G*-tests (Sokal and Rohlf 1981).

this sequence arose has normal viability and normal male genitalia for this species. Thus, neither explanation can be directly supported nor ruled out. Also, despite the evidence from the McDonald-Kreitman tests (Table 3) of excess replacement differences between species at *est6*, *janus*, and *Zw*, we do not find evidence that these loci have overall levels of polymorphism and divergence that are inconsistent with the neutral model (Figure 1).

The HKA test was repeated with the exclusion of just those loci that showed the strongest departures from expectations. As expected, the value of the overall test statistics dropped markedly, though that for *D. sechellia* was still significant (*D. simulans*, $\chi^2 = 10.77$, $P = 0.1308$; *D. mauritiana*, $\chi^2 = 13.92$, $P = 0.1520$; *D. sechellia*, $\chi^2 = 21.90$, $P = 0.0459$*). In the case of *D. sechellia* the still significant departure is primarily due to two loci (*mt:ND5* and *w*) that revealed two polymorphisms where none were expected.

We also considered Tajima's *D* statistic (Table 2) of the difference between different estimators of the population mutation rate, $\theta = 2Gu$, where *G* is the effective number of gene copies and *u* is the mutation rate (Tajima 1989b). For a diploid species of effective population size *N*, $G = 2N$ for an autosomal locus; $G = 3N/2$ for an X-linked locus; and $G = N/2$ for a sex-limited, effectively haploid locus found on the mitochondria or the Y chromosome. Under a neutral model of constant population size, the expected value of *D* is very near zero (Tajima 1989b). A negative *D* results when more than the expected number of polymorphic sites have low frequencies in the sample, a pattern that can be caused either by recent selection that has removed variation or by a recent population size expansion. For *D. simulans*, the values of *D* vary considerably and one (*hb*) is significantly less than zero. However, for *D. mauritiana*, nine values were negative while only two were positive (two could not be calculated and one was equal to zero), and again one of the values was significantly different from zero (*w*). To check whether such an overall negative pattern of *D* values is very unlikely by chance, the average observed value of *D* was calculated (weighted by locus length) and compared to the distribution of the same quantity generated by computer simulation. The simulations were the same ones used for the HKA tests and included 10,000 independent standard coalescent simulations using estimates of divergence time and θ for each locus that were generated from the observed polymorphism and divergence levels. For each simulation, we noted whether the absolute value of the observation was greater than the absolute value of the simulated value (two-tailed test). For *D. mauritiana*, the weighted average of *D* was −0.677 and only 2% ($P = 0.020$) of the simulations generated a more extreme value. For *D. simulans* and *D. sechellia*, the same analysis revealed a weighted value of *D* that fell near the middle of the

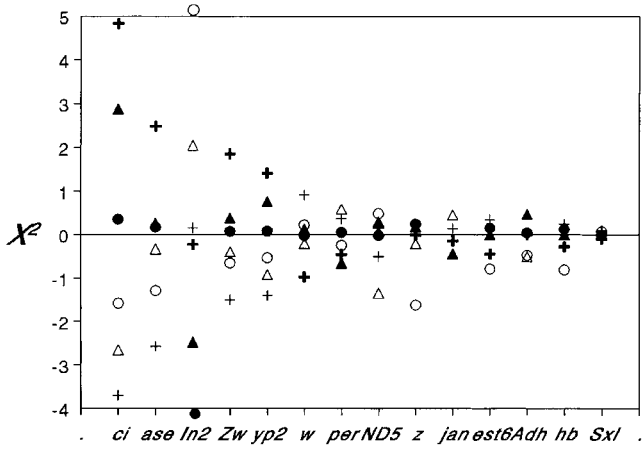

FIGURE 1.—Three multilocus HKA tests were done (HUDSON *et al.* 1987), one per species. (+) *SimP*, (+) *SimD*, (△) *MauP*, (▲) *MauD*, (○) *SecP*, (●) *SecD*. In each case, polymorphism within species (as listed in Table 2) and divergence from a single *D. melanogaster* sequence were used for the test. Shown are the contributions to the overall χ^2 test statistic by the polymorphism and divergence observations for each locus. Thus, for example, *SimP* refers to the standardized departure from expectations for polymorphism within *D. simulans* and *SimD* refers to the same quantity for divergence from *D. melanogaster*. If the observed value was greater than the expected, then the point is placed above the line; otherwise it is placed below the line. In the case of *SecP* for *In(2L)t* the value was 38.4, and in the case of *SecD* for *In(2L)t* the value was 7.39. These extreme values are represented by points outside the graph. Loci are ordered from left to right in rough accord with their degree of departure from expectations.

simulated distribution (results available upon request). The overall negative pattern of *D* values from *D. mauritiana* suggests that recent population demographics have shaped the polymorphism pattern, with the simplest explanation being recent population size expansion (TAJIMA 1989a).

Divergence of genes: The three species of the *simulans* complex are closely related to one another, and much of the history of gene samples drawn from the *simulans* complex predates the origins of the three species (HEY and KLIMAN 1993). This recent complicated history precludes any simple analysis in which gene divergence is equated with species divergence (see below). However, we can ask some simple questions about how individual gene copies have diverged. In particular, we can use sequences from *D. melanogaster* to root the differences between pairs of sequences drawn from the *simulans* complex and ask whether genes drawn from different species have accumulated mutations at the same rate.

Relative rate tests were conducted for pairs of *simulans* complex sequences, rooted by a *D. melanogaster* sequence, using the method of WU and LI (1985). Because each test involved just a pair of sequences, and there are many such pairs, we have done two types of summaries. In the first place, we examined all of the data for the 14 genes by doing all possible pairwise comparisons of sequences drawn from different species and summarized the results for each species pair and each gene in Table 4. Half of the genes (*hb, In(2L)t, janus, per, Sxl, w,* and *Zw*) had some pairs of sequences in which the substitution rate difference seemed excessive under the null model of no rate variation. These significant comparisons tended to show up in all three pairwise species contrasts. Another pattern is that for all these genes where some tests were significant, there was an average substitution rate excess for *D. sechellia* relative to *D. simulans* and relative to *D. mauritiana* (Δ columns in Table 4). In the comparisons between *D. simulans* and *D. mauritiana* the direction of departure varied evenly among genes.

The second method of summarizing was applied to the data, prior to the relative rate tests, so as to have results that are not complicated by so many multiple comparisons. This analysis employed just one single constructed sequence from each species. The following genes have been sequenced, at least in part, from at least one individual from each of the three *simulans* complex species, as well as *D. melanogaster*: the proximal *Amylase* gene (*Amy-P*; SHIBATA and YAMAZAKI 1995); *Amyrel* (DA LAGE *et al.* 1998); the *Cecropin* gene cluster (RAMOS-ONSINS and AGUADÉ 1998); *dynein* (*Dhc-Yh3*; ZUROVCOVA and EANES 1999); *glutathione-S-transferase D1* (*GstD1*; M. T. HARGIS and J. B. COCHRANE, unpublished sequences in GenBank); *myosin alkali light chain* (*Mlc1*; LEICHT *et al.* 1995); male accessory gland peptide genes *Mst26Aa* and *Mst26Ab* (AGUADÉ *et al.* 1992); *nullo* (CACCONE *et al.* 1996); *Cu-Zn superoxide dismutase* (*Sod*; K. ARXONTAKI, P. KASTANIS, S. TSAKAS, M. LOUKAS and E. ELIOPOULOS, unpublished sequences in GenBank); and *serendipity* (*sry*-α; CACCONE *et al.* 1996). The sequence from each species, for each of these genes, was aligned by eye and concatenated. To these sequences were added one randomly drawn sequence from each species from the 14 genes listed in Table 1. The final data set included concatenated data from 23 gene regions, with a total length for each sequence of 28,692 bases. For each of the three possible comparisons, the *D. melanogaster* sequence was used to root the divergence between the sequences from each of the other two species and to obtain estimates of the substitution rate per base pair since that root point. For the comparisons between *D. simulans* and *D. sechellia* the relative rate test yielded values of 0.0092 and 0.0137 changes per site, respectively, which are highly significantly different ($P < 0.0001$). For the *simulans/mauritiana* comparison, the values are 0.0102 and 0.0122, respectively ($P < 0.05$); and for the *mauritiana/sechellia* contrast the values are 0.0121 and 0.0145, respectively ($P < 0.05$). On balance it appears that genes from *D. sechellia* have been evolving ~50% more quickly than have genes in *D. simulans* and that genes in *D. mauritiana* have an average rate of mutation accumulation that is in between that of the other species. Put another way, if we consider

TABLE 4

Relative rate test results

	sim and *sec*			*sim* and *mau*			*sec* and *mau*		
Loci	No. pairs[a]	No. sig[b]	Δ[c]	No. pairs	No. sig	Δ	No. pairs	No. sig	Δ
Adh	12	0	0.0002	36	0	−0.0047	12	0	−0.0049
ase	36	0	0.0008	30	0	−0.0016	30	0	−0.0024
ci	36	0	0.0011	54	0	0.0000	24	0	−0.0011
est6	8	0	0.0010	22	0	0.0007	12	0	−0.0003
hb	80	6	−0.0022	90	9	0.0004	70	2	0.0026
In(2L)t	32	1	−0.0045	96	3	−0.0030	26	1	0.0014
janus	3	1	−0.0075	9	1	0.0027	3	1	0.0102
mt:ND5	72	0	−0.0126	72	0	−0.0120	64	0	0.0006
per	36	14	−0.0055	36	17	−0.0063	36	0	−0.0007
Sxl	143	77	−0.0201	117	3	0.0011	99	53	0.0212
w	54	6	0.0180	62	7	0.0098	42	8	0.0050
yp2	36	0	−0.0020	36	0	−0.0028	36	0	−0.0008
z	36	0	−0.0001	36	0	0.0013	36	0	0.0014
Zw	24	12	−0.0057	72	1	−0.0006	12	8	0.0051

Each sequence from each locus was compared against all of the other sequences from each of the other species in a relative rate test (Wu and Li 1985).

[a] The number of pairwise comparisons between the two species being compared.

[b] The number of comparisons that were statistically significant at $P \leq 0.05$.

[c] The difference between the estimated substitution rate, per base pair, for the first species listed at the head of the column and the same quantity for the second species listed at the head of the column. If Δ is positive then a sequence from the first species has had a higher substitution rate than the sequence from the second species.

just the 112 Mb of DNA sequence recently reported for the *D. melanogaster* genome project, then a random copy of the *D. sechellia* genome had >500,000 more mutations accumulate than a comparable copy of the *D. simulans* genome since the various times at which the different genes had common ancestors.

The ranking of mutation accumulation rates inversely mirrors the ranking of estimated effective population sizes—the larger the effective population size, the lower the rate of mutation accumulation. This pattern is consistent with the slightly deleterious model of mutation accumulation, in which more mutations are effectively neutral when population sizes are smaller (Ohta 1972, 1973). If the slightly deleterious mutation model does explain the differing rates of mutation accumulation, then we would also expect this to be reflected in the ways that synonymous mutations have accumulated in the different species. Synonymous codon usage in Drosophila does appear to have been shaped, in part, by natural selection (Kliman and Hey 1993b; Akashi 1994, 1995; Duret and Mouchiroud 1999), and the degree to which prefered codons (*vs.* unprefered codons) accumulate can be taken as a measure of the efficacy of natural selection on codon usage. Thus, for example, Akashi (1995) found that fixations of unpreferred codons were significantly more numerous than fixations of preferred codons in *D. melanogaster*, indicating that selection on codon usage may have become ineffective in this species subsequent to its split from *D. simulans*. Using *D. melanogaster* as an outgroup, we identified by parsimony the ancestral and derived states for fixed synonymous substitutions unique to each of the three *simulans* complex species (see Table 5). *D. simulans* and *D. mauritiana* had too few fixed synonymous substitutions with which to conduct a test, but there are 29 such fixations in *D. sechellia*. Of these, 17 substitute an unpreferred codon for an ancestral preferred codon,

TABLE 5

Fixed synonymous mutations

Ancestral codon	Derived codon	*sim*	*mau*	*sec*
Preferred	Unpreferred	0	2	17
Unpreferred	Unpreferred	1	2	5
Preferred	Preferred	1	0	1
Unpreferred	Preferred	0	3	6

The ancestral state at each site was inferred by parsimony, using *D. melanogaster* sequence(s) as an outgroup. Fixed differences are those base positions where a derived base is unique to just one of the three *D. simulans* complex species. In other words, the inferred derived state is fixed in only one species, while only the ancestral state is found in the others (including *D. melanogaster*). The ancestral and derived states were classified as preferred or unpreferred on the basis of synonymous codon preferences established by Akashi (1995) for *D. melanogaster* and its closest relatives.

TABLE 6

Shared polymorphisms and fixed differences

	sim and *sec*		*sim* and *mau*		*sec* and *mau*	
Locus	Shared	Fixed	Shared	Fixed	Shared	Fixed
Adh	0 (0.00)	7	1 (0.09)	4	0 (0.00)	14
ase	0 (0)	2	0 (0)	1	0 (0)	3
ci	0 (0)	5	0 (0.00)	6	0 (0)	4
est6	0 (0)	23	11 (1.63)	4	0 (0)	27
hb	0 (0)	0	1 (0.26)	0	0 (0)	2
In(2L)t	8 (1.01)	0	9 (1.94)	0	5 (1.59)	0
In(2L)t[a]	0 (0.18)	11	9 (1.94)	0	0 (0.26)	10
janus	NA	NA	11 (0.89)	0	NA	NA
mt:ND5	0 (0.00)	8	0 (0)	8	0 (0)	12
per	1 (0.12)	18	11 (1.39)	3	0 (0.01)	21
Sxl	0 (0.07)	1	1 (0.14)	0	0 (0.03)	5
w	0 (0.19)	1	0 (0.29)	0	0 (0.03)	4
yp2	0 (0.00)	4	0 (0.01)	2	0 (0.00)	6
z	0 (0)	4	0 (0.17)	1	0 (0)	10
Zw	0 (0)	15	0 (0.08)	6	0 (0)	9

Shared polymorphisms are those DNA base positions in which two species share two or more segregating bases. Fixed differences are those positions where all the sequences from one species are different from all those of the second species. Observed values are shown for each of the three two-species contrasts. Values in parentheses are the expected number of shared polymorphisms calculated using expression (2) in the text. An expected value of (0) arises if one or both species have no polymorphic sites. NA, not applicable.

[a] Two sets of values are shown for *In(2L)t*. The upper values are based on inclusion of all three *D. sechellia* sequences, while the lower value is based on just those two that do not resemble *D. simulans* (see text).

while only 6 show the opposite pattern. These values differ significantly from equality ($G = 5.48$, $P = 0.0019$), consistent with the hypothesis that selection on silent sites has been ineffective in *D. sechellia* in the time since coancestry with the other species.

The evidence of reduced effective population size and reduced efficiency of natural selection, in *D. sechellia* relative to the other species, is also consistent with the finding that *D. sechellia* bears many fewer genes that contribute to hybrid sterility in crosses with *D. simulans* than does *D. mauritiana*. Though this pattern was once interpreted as evidence that *D. simulans* and *D. sechellia* are the most closely related species pair (Palopoli *et al.* 1996), it is also consistent with a greater rate of adaptation in *D. mauritiana*, as might occur with a larger effective population size.

Divergence of species: As incipient species begin to diverge from one another they can be expected to share genetic variation that was common to their ancestral species. If neither incipient species experiences a strong population bottleneck, then these shared polymorphisms may persist for a long period of time, particularly at those genes that are not associated with adaptive divergence (and are not linked to such genes). Table 6 shows the numbers of shared polymorphisms and fixed differences found between each species pair. Both *D. simulans* and *D. mauritiana* are highly polymorphic, and even though the number of sequences sampled is small, we find that the two species share polymorphisms at a majority of the loci. In contrast, species comparisons that involve *D. sechellia* generally revealed no shared polymorphisms, as expected given the low level of polymorphism found within this species. The exceptions involving *D. sechellia* are a single shared polymorphism between *D. simulans* and *D. sechellia* at *per* and the abundance of shared polymorphisms at *In(2L)t* due to a single *D. sechellia* sequence (see above).

To assess how many of the shared polymorphisms could be expected to arise just by recurrent mutation, we conducted a simple calculation under the assumption that mutations occur randomly and independently with equal probability at all sites. If $s1$ and $s2$ polymorphic sites were observed in each of two historically independent species over a common region of length L, then the probability that exactly ss of those polymorphisms fall on the same base positions in the two samples is given by the hypergeometric probability

$$P(ss|L, s1, s2) = \frac{\binom{L - s1}{s2 - ss}\binom{s1}{ss}}{\binom{L}{s2}} \quad (1)$$

(Clark 1997). The expected number of shared polymorphisms $E(ss)$ is equal to

$$\sum_{j=0}^{\mathrm{Min}(s1,s2)} j \times P(j|L, s1, s2) = s1 \times s2/L. \quad (2)$$

In the case of the *simulans* complex data, the expected

values of shared polymorphisms are quite low, generally near zero, and even when the values are >1, they are still a small fraction of the observed number (Table 6). The reasons for this are simply that mutations are rare and that there are a very large number of available sites. This method assumes that all sites are equally likely to mutate, and so it is likely to underestimate shared polymorphisms that arise via multiple mutations. However, even if the analysis is repeated by first breaking down the observed values of *L*, *s*1, and *s*2 into components due to replacement, synonymous, and intron sites, the overall expected values do not approach the observations for those cases when shared polymorphisms were observed (results available upon request).

Another way to check whether mutations are occurring randomly and fairly uniformly across sites is to compare observations with a Poisson distribution. An approximate check can be made by asking whether the number of sites that support a 2-, 3-, or 4-base polymorphism is consistent with a Poisson distribution, given the number of sites that revealed no polymorphic sites. Fitting a Poisson distribution to the *D. simulans* data set returned expected values of 12,271, 275, 3, and 0 positions with 1, 2, 3, and 4 segregating bases, respectively (sites with 1 segregating base are invariant). The observed values were 12,271, 271, 7, and 0. The good fit of the Poisson distribution suggests that overall the data set has just a small number of sites where recurrent mutations have occurred.

Also revealed in the comparison between *D. simulans* and *D. mauritiana* is the negative correlation, across loci, that is expected between fixed differences and shared polymorphisms. In the absence of recombination and recurrent mutation, a gene tree for one locus can support either fixed differences or shared polymorphisms, but not both (neither may occur as well), as a simple byproduct of the possible gene tree topologies (Wakeley and Hey 1997). However, if recombination has been occurring, then different portions of a locus have different gene trees and it is possible for both shared polymorphisms and fixed differences to occur.

Figure 2 shows the results of cluster analyses for most of the genes (similar diagrams for the remaining genes were reported previously). These diagrams should not be equated with gene tree estimates, as most loci showed evidence of recombination and thus do not have a bifurcating gene tree history. However, these diagrams do serve to show the variable patterns of similarity that are found among genes and how those patterns are not consistent with simple phylogenetic relationships among species. As in the case of the original studies on *ase*, *ci*, *per*, *yp2*, and *z* (Hey and Kliman 1993; Kliman and Hey 1993a; Hilton *et al.* 1994), sequences from *D. simulans* show only a limited tendency to cluster by their taxonomic designation. The same kind of dispersed pattern is seen for *D. mauritiana* sequences at *hb*, *In(2L)t*, and *janus*. The *D. sechellia* samples do consistently cluster with one another (with the exception of one sequence at *In(2L)t*), but depending on the gene the *D. sechellia* cluster may fall almost anywhere within the diagram.

The frequent tendency for genes from *D. simulans* and *D. mauritiana* to cluster with those from the other species is entirely consistent with the presence of a large number of shared polymorphisms between these species (Table 6). These patterns are expected if multiple gene lineages persist in both species since the time of speciation (Clark 1997; Wakeley and Hey 1997).

Fitting a speciation model: We compared the data to what would be expected under a simple speciation model, called an "isolation model," in which an ancestral constant size population splits over a very short period of time into two populations, each of constant size. There are four primary parameters to the model, including three θ's, or population mutation rates (one for the ancestral population and one for each descendant), and a time since the splitting event. The model fitting requires the counts of shared polymorphisms and fixed differences, as well as counts of the numbers of unique polymorphisms. The method is outlined in Wakeley and Hey (1997) and Wang *et al.* (1997).

Table 7 shows the results of fitting the isolation model to four different data sets. The first case includes *D. simulans* and *D. mauritiana* and, as in the original application of the method for this species pair, the ancestral species appears to have had a size intermediate between the descendants and to have occurred not very long ago (Wakeley and Hey 1997). The second case is the same as the first except that the numbers of shared polymorphisms were reduced by the number expected by chance and independent mutation as shown in Table 6. The isolation model parameter estimates are very similar to the first case. The third and fourth applications are to the case of *D. simulans* and *D. sechellia* (with and without the shared sequence of *In(2L)t*, respectively). It is interesting that the removal of that sequence does not have a large effect on the parameter estimates. In both cases *D. sechellia* has a low estimated value for θ, while the ancestral species estimate is considerably larger than that for either descendant species. The reason for the similarities, with and without the *In(2L)t* sequence, is that this sequence is not the only locus where a shared polymorphism was found (one also occurred at *per*; Table 6). Thus, in both applications, the model must still reconcile the presence of divergence between the taxa, low polymorphism within *D. sechellia*, and the presence of shared polymorphism. The combined effect of all three is to drive up the estimate of the size of the ancestral species (Wang *et al.* 1997).

We also performed statistical tests of the quality of fit between the expected levels of polymorphism under the isolation model and the observed values (Wang *et al.* 1997). The test proceeds by conducting coalescent simulations based on the parameter estimates and then

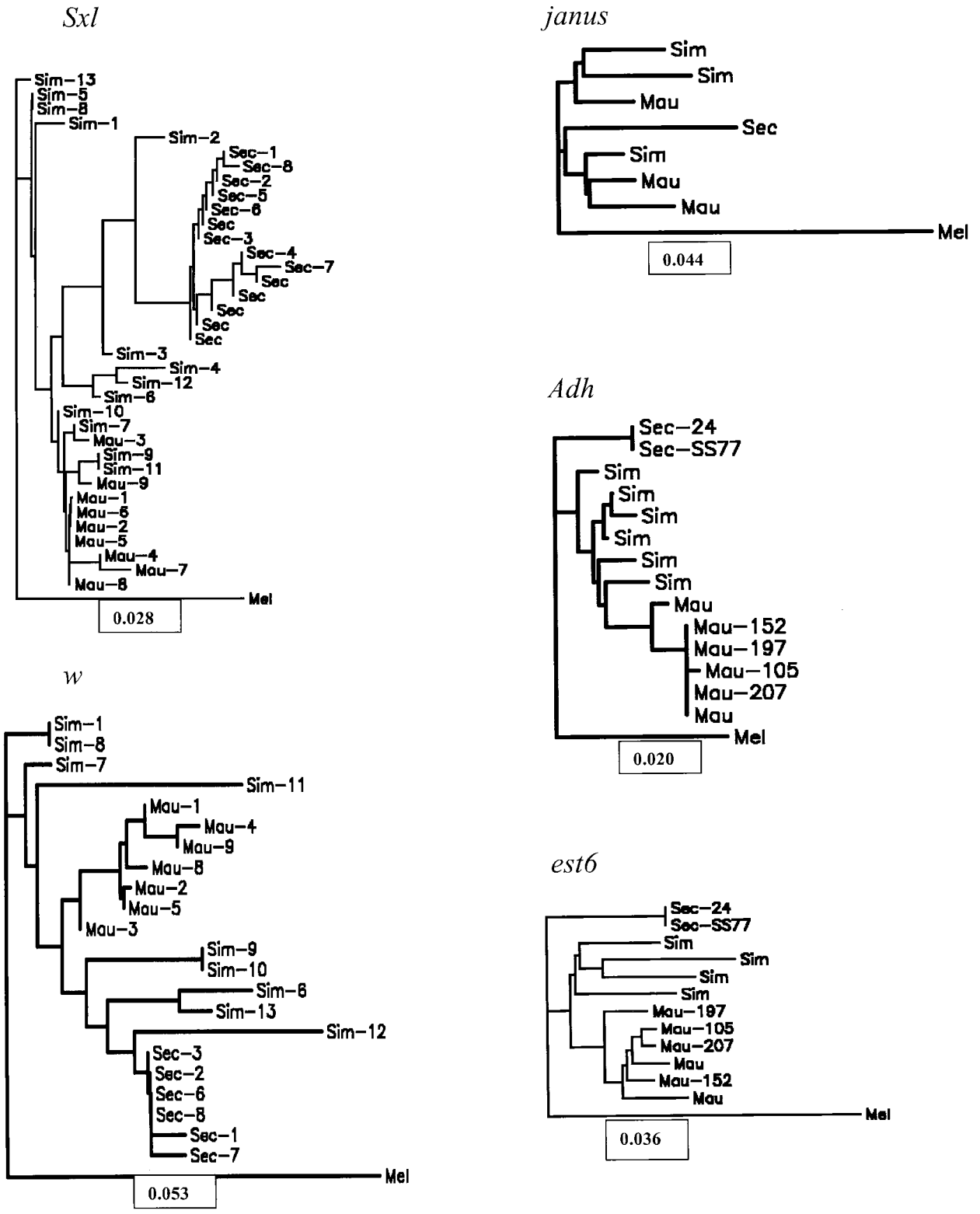

FIGURE 2.—Distance trees for nine loci. The length of the branch to the outgroup sequence of *D. melanogaster* is shown in units of estimated changes per base pair. Comparable trees for the remaining loci (*ase*, *ci*, *per*, *yp2*, and *z*) were reported previously (HEY and KLIMAN 1993; KLIMAN and HEY 1993a; HILTON *et al.* 1994). For each locus, DNA sequences were aligned by eye and clustering was done using the neighbor-joining algorithm (SAITOU and NEI 1987). Where lines were not common to multiple loci, lines are labeled only to species. For *Adh*, *est6*, and *Zw*, the lines with specific line numbers were the same as some of those used in earlier reports on *ase*, *ci*, *per*, *yp2*, and *z* (see MATERIALS AND METHODS). For *hb*, *mt:ND5*, *Sxl*, and *w*, most lines came from a common set (as described in MATERIALS AND METHODS) and these lines are numbered within each species.

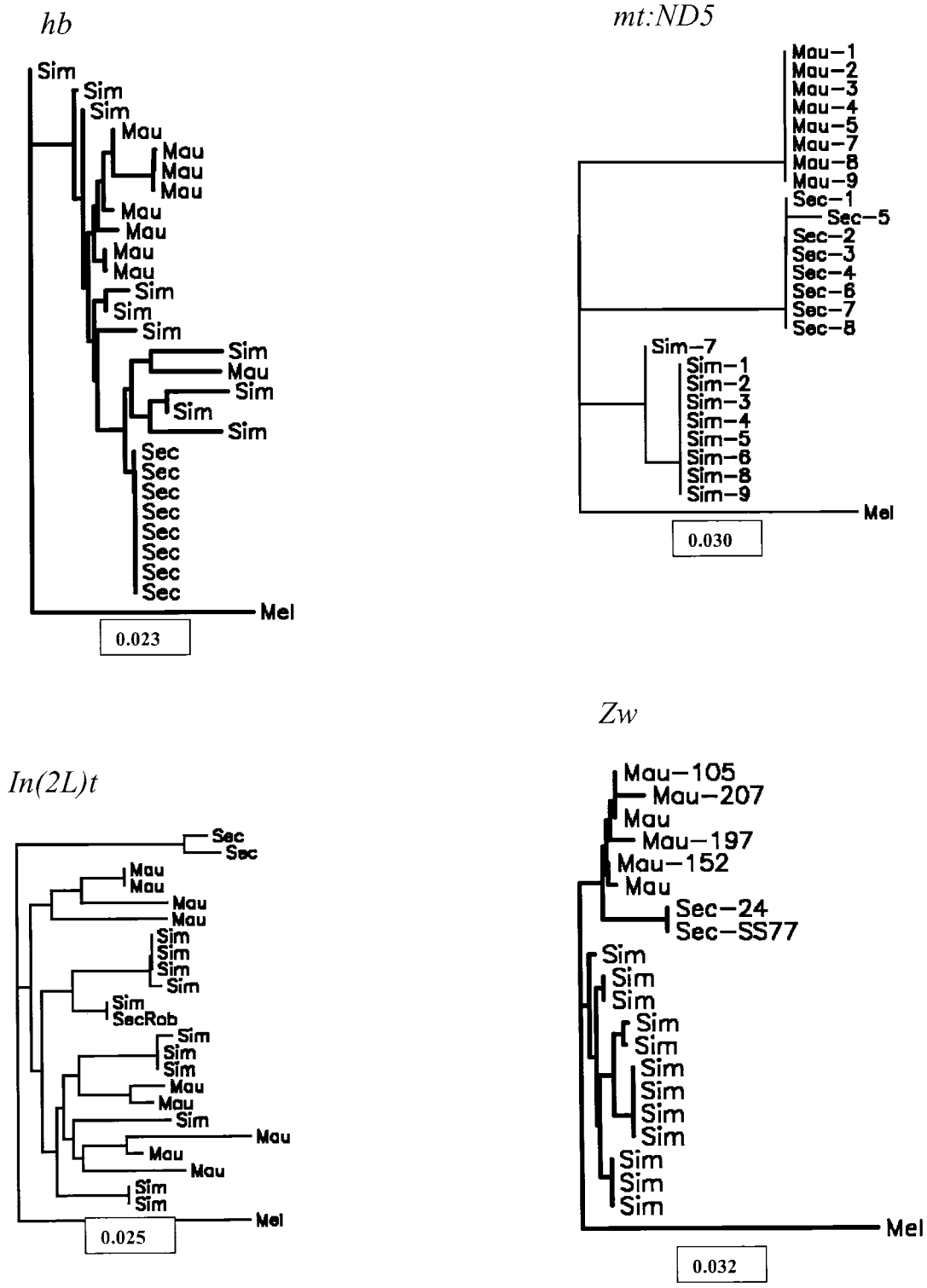

FIGURE 2.—*Continued.*

by comparing the distribution of results from 10,000 such simulations with the actual data. These simulations incorporated recombination, at rates based on the estimated amount of recombination that occurs within each gene in each species, as this strongly affects the degree to which shared polymorphisms and fixed differences covary. We used the γ estimate of the population recombination rate (HEY and WAKELEY 1997) as determined for each species and locus. Just as with the actual data, each simulated data set is partitioned into the four categories of polymorphic sites (polymorphisms exclusive to species 1, those exclusive to species 2, shared polymorphisms, and fixed differences) for each locus, and these quantities are used to generate isolation model parame-

TABLE 7

Isolation model fitting

Species 1	Species 2	Test	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_A$	T	P_{χ^2}	P_{WWH}
sim	*mau*	1	143.7	79.5	139.9	0.42	0.138	0.827
			75.7–280.5	46.0–129.7	85.5–203.0	0.22–0.64		
sim	*mau*	2	152.3	83.3	127.1	0.44	0.118	0.781
			87.2–293.3	50.7–134.0	75.7–186.1	0.23–0.67		
sim	*sec*	3	32.3	7.75	221.1	0.51	0.216	0.465
			3.76–187.1	0.95–25.8	0.00–283.4	0.30–0.97		
sim	*sec*	4	55.7	7.98	201.7	0.56	0.434	0.810
			17.1–303.7	2.33–17.0	0.0–259.3	0.35–1.52		

For each contrast, the data were fit to the isolation model as described (Wang *et al.* 1997). The estimated values for the primary parameters are shown, along with the 95% confidence intervals determined by simulation (see text). Four contrasts are shown: (1) 12 loci (all excluding *ase* and *ci*); (2) same as (1) but observed shared polymorphisms were decreased by that amount expected by chance (see Table 6); (3) 11 loci (all excluding *ase*, *ci*, and *janus*); and (4) same as (3) except that one *D. sechellia* sequence for *In(2L)t* data was excluded (see text). The *P* values, for both the χ^2 and the Wang, Wakely, and Hey (WWH) test statistics, are the proportion of simulated values greater than or equal to the observed. The test is one-tailed because the focus is on detecting a departure from the model in the direction expected if historical gene flow had occurred.

ter estimates and expected values for each of the quantities. The simulations were also used to generate 95% confidence intervals for the parameter estimates.

We considered two test statistics. One was a simple χ^2 statistic that summed the discrepancies between observations and expectations for each locus and each polymorphism type. If we denote the counts of the four types of polymorphisms for locus i as $S_{i,j}$, with $j = 1 \ldots 4$, and if there are L loci, then

$$\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{4}\frac{(S_{i,j} - E(S_{i,j}))^2}{E(S_{i,j})}. \quad (3)$$

The second test statistic was that used by Wang *et al.* (1997), which is equal to the difference between the highest and lowest counts of shared polymorphisms observed across loci, plus the difference between the highest and lowest counts of fixed differences, observed across loci.

From comparison of the first two rows of Table 7, it is clear that adjusting the observed numbers of shared polymorphism by the number expected by chance has little effect on the parameter estimates or the quality of the fit of the isolation model. Similarly, from rows three and four, we see that the effect of including the unusual sequence of *In(2L)t* within the *D. sechellia* sample has little effect on the parameter estimates. There is an effect on the fit between the model and the data (the model fits better when the sequence is excluded), but in neither case is the model rejected.

DISCUSSION

Our basic approach has been to extend DNA sequence-based population genetics to questions associated with relatively ancient speciation events. The divergence of the *simulans* complex species probably began hundreds of thousands of years ago (Hey and Kliman 1993), yet in the patterns of mutation accumulation and in the patterns of shared and fixed differences, we can still assess the effects of population sizes and assess the historical presence of gene flow between species. For two quite different reasons, the DPG approach to the study of recent speciation events becomes considerably more informative the more that comparative DNA sequence data are available from multiple independently segregating genes. First, multiple loci permit the assessment of different evolutionary forces. The historical portraits that are developed for each locus can be compared to see whether different loci are consistent with a common historical model. Thus multiple loci can be used to distinguish those demographic forces that have acted on many genes (*e.g.*, population splitting, population size changes, and migration) from those that have acted just on smaller parts of the genome (*e.g.*, natural selection). The second benefit of multiple loci concerns sampling effort. For populations or species that have been diverging for some time, the gene trees within species may not extend back to the time of the common ancestor, and even if they do, only a small minority of lineages are expected to be of that age. Thus, only a small portion of the true genealogical history for a species, at a locus, may extend from the time period under investigation. When this is the case, repetitive sampling within species tends to include that history even in a small sample. Put another way, the older nodes of a species' true genealogy, for a locus, tend to be revealed in a small sample, whereas more recent portions are, on average, only revealed as the sample size per locus grows large (Kliman and Hey 1993a). This basic feature of genealogical sampling necessarily dictates an optimal strategy that is shifted away from multiple sequences per locus and toward multiple loci—each with few sequences. In the extreme, it is even possible to study the sizes of ancestral species by using just one sequence

from each species, so long as many loci are studied (Takahata 1986).

Departures from the neutral model: The major assumptions of the basic null model that is used as a heuristic guide for many analyses, and as an explicit baseline in the statistical tests, are that mutations are neutral and that population sizes are constant (the McDonald-Kreitman test does not rely upon the latter). We observed four distinct kinds of departures from null expectations: an overall negative value of Tajima's *D* for *D. mauritiana* (Table 2), suggestive of a recently expanding population size in this species (Tajima 1989a); significant McDonald-Kreitman tests at *est6*, *janus*, and *Zw* (Table 3), suggestive of an accumulation of excess amino acid replacement differences between species at these loci (Eanes *et al.* 1993); an excess accumulation of mutations in *D. sechellia*, many of which are probably slightly deleterious (Tables 5 and 6); and significant HKA tests, primarily due to low variation within species at genes in low recombination portions of the genome (Figure 1). On balance, our null model does not fare very well, though we do learn a great deal from each of these exceptions. These findings necessarily lessen the applicability of the isolation model fitting, which strongly relies upon the neutral model. In recent years, particularly with growing data on polymorphism and divergence, there have come many reports on exceptions to the neutral model, particularly for the well-studied *D. melanogaster* (Kreitman 1996; Moriyama and Powell 1996; Ohta 1996; Hey 1999).

Speciation: Throughout this report, the three *simulans* complex species are considered to be biological entities within which evolutionary forces of natural selection and genetic drift play out amid a recombining gene pool, and between which there is a near absence of gene exchange (Dobzhansky 1937). If we wish, this starting point can be taken as an assumption under test. Thus, for example, consider that the three *simulans* complex taxa have been represented by DNAs prepared from organisms that were taxonomically identified on morphological grounds. Then our evolutionary investigation amounts to a testing of the hypothesis that the taxonomic samples have indeed come from biological species. In particular, we can ask two questions: (1) whether the patterns of DNA sequence variation for any one taxon are consistent with that taxon being a single biological species and (2) whether the patterns of DNA sequence variation between taxa are consistent with genetic isolation. The first question can be partly assessed by asking whether single taxa show evidence of multiple separate gene pools. An example of this would be if multiple genes show evidence of relatively ancient population subdivision among the samples from a single taxon. For example, a pattern like this was found for multilocus samples of *D. novamexicana* (Hilton and Hey 1996, 1997). The second question can be partly assessed by asking whether multiple taxa have experienced substantial gene flow at multiple loci. Hilton and Hey (1996, 1997) also found a case wherein two cytospecies, *D. americana americana* and *D. a. texana*, had experienced considerable gene exchange at multiple loci and thus should not be considered as species (see also McAllister and Charlesworth 1999).

When *simulans* taxa are considered from the standpoint of just one single locus, then we often do find that taxa are poorly reflected by the patterns of similarity among individual gene copies. Pairs of gene copies drawn from *D. simulans* and *D. mauritiana* vary widely in the degree to which they differ, and gene tree estimates for individual genes show that these taxa are highly paraphyletic when represented by multiple gene copies for a single locus (Satta and Takahata 1990; Hey and Kliman 1993; Kliman and Hey 1993a; Hilton *et al.* 1994; Figure 2). However, when the same set of multiple strains from each species was studied at multiple loci, there was no tendency for ancient population subdivision within species and no strong evidence of recent gene flow at multiple loci (Hey and Kliman 1993; Hilton *et al.* 1994). In the branching cluster diagrams for *z*, *yp2*, and *per*, the *D. sechellia* sequences clustered, as did those for *D. mauritiana*, while those for *D. simulans* tended to be spread out across the diagram and to come together mostly at the deepest parts of the diagram (Hey and Kliman 1993). However, despite these common patterns among taxa across loci, the clustering patterns varied widely within taxa across loci—even though the same set of inbred lines had been studied for each locus. For some of the newly studied loci (*hb*, *w*, *mt:ND5*, and *Sxl*), a common set of inbred lines was used, and again we see that there are common patterns among taxa (very similar to what was found with *z*, *yp2*, and *per*), and again we see that within taxa the relationships among particular lines vary widely across loci (Figure 2). These results are just what we expect for taxa that represent real, recently diverged, biological species, within which there is recombination. The overall picture is one in which all three species diverged at about the same time, in which both *D. mauritiana* and *D. simulans* have had large effective population sizes and still carry shared polymorphisms since divergence, and in which *D. sechellia* has a small effective population size.

At the crux of many speciation discussions is the question of whether or not natural selection plays a direct creative role in forming species. In the simplest models of allopatric speciation it does not, and speciation is a byproduct of the evolution that proceeds in physically separated populations. Thus, for example, in the classic speciation model of Dobzhansky and Muller, each of two separate populations accumulates adaptations one by one. However, it turns out that when given the chance to hybridize, the mutations fixed in one species are incompatible with the novel genome of the other species. In other words, they are epistatic and deleterious

when expressed in a genetic background other than the one in which they arose (Dobzhansky 1936; Muller 1940). However, speciation may also arise in a more dynamic context where natural selection promoting regional specialization is in a tug-of-war against recombination and gene flow that break down associations both among genes and between genes and geography. Under these circumstances, where diverging populations are sympatric or parapatric, natural selection is acting directly to shape the species barrier.

Whether or not natural selection promotes species formation directly or indirectly depends on whether or not gene exchange was occurring among incipient species. Thus, research on the historical demographic processes associated with species divergence may reveal evidence of ancient gene flow and, therefore, illuminate the kinds of natural selection and the kinds of phenotypes that might have existed during the beginning stages of species formation. Of course, if gene flow and natural selection were important factors for just a short period of time at the beginning of speciation, then patterns of variation may not be indistinguishable from those expected under the isolation model, particularly if those events were long ago. However, population genetic methods can sometimes reveal recent or ongoing gene flow between species that are otherwise long diverged. With such findings our understanding of species as evolutionary entities undergoes a significant adjustment; for it is then that natural selection can be seen as having maintained the phenotype, by which we recognize the species, in the face of that gene flow.

Assessing gene flow: Variation can be shared between species either by gene flow or by dual persistence since the time of population splitting. These historical alternatives can be difficult to distinguish, relying primarily on two kinds of observations. First, if most gene sequences suggest moderate or high divergence, but a minority are identical for two species, then the simplest explanation may be population splitting long ago and limited recent gene flow. This kind of observation is essentially one of an appearance of a sequence that is "atypical" for its taxon. An example of this pattern was found at the *per* locus in *D. pseudoobscura* and *D. persimilis* (Wang and Hey 1996). Among the loci studied here, only *In(2L)t* showed an example of this. The second kind of observation that can be suggestive of gene flow is if loci vary widely in the degree to which they share polymorphisms. A model of limited gene flow is expected to give wide variation among loci in apparent divergence (Wakeley 1996; Wang *et al.* 1997). However, such variation among genes must be quite high, and assessment of it is strongly dependent upon recombination rates. Thus we could not reject the model of no gene flow, in the testing of the isolation model, between *D. simulans* and *D. sechellia*, despite the appearance of haplotype sharing at *In(2L)t*. A third explanation of shared variation is not genealogical, but mutational—recurrent mutations can also cause shared variation. Given the overall recent low levels of DNA sequence divergence, this has probably not been a large factor in this study. However, recurrent mutation could well have been the cause of the single polymorphism that is shared by *D. sechellia* and *D. simulans* at the *per* locus.

Limited evidence of gene flow among these species also comes from a study of *ase* and *ci*. In this case the observation did not involve shared variation (polymorphism is nearly absent in these genes) but rather that divergence between the *simulans* complex species was less than expected given what had been found at other loci (Hilton *et al.* 1994). On balance, the data are largely consistent with an absence of gene flow. Only *In(2L)t* revealed the kind of pattern expected of a very recent gene flow event, and in this case laboratory admixture cannot be ruled out. Overall the levels of shared polymorphisms and fixed differences are consistent with an isolation model. However, given the difficulty of distinguishing gene flow from shared ancestral variation, we cannot rule out a speciation model that included a period of gene flow following population splitting.

Phylogeny: As an important model system for the study of speciation, the *D. simulans* complex has been the subject of many efforts to infer phylogeny. Indeed, all three possible pairs of taxa have been proposed as the most closely related species pair, including *simulans/sechellia* (Cariou 1987; Palopoli *et al.* 1996) and *simulans/mauritiana* (Lachaise *et al.* 1986; Joly 1987; Hey and Kliman 1993; Coyne and Charlesworth 1997; Harr *et al.* 1998; Ting *et al.* 2000), and also the *sechellia/mauritiana* pairing (Caccone *et al.* 1988, 1996), which seems unlikely on the basis that it would require a colonization from one remote island to another, whereas the other models simply require two colonizations from the mainland.

The difficulty of the phylogeny problem can be seen both from the standpoint of the data and from the standpoint of theory. Regarding data, a simple appraisal of the cluster diagrams for the 14 genes shows how difficult it could be to try to discern an overall species branching history. Thus consider from the standpoint of *D. sechellia* sequences, which always cluster together [excepting *In(2L)t*], and ask whether the next most similar sequence is from *D. simulans*, or *D. mauritiana*, or whether it is a node that joins sequences from both of these species. A plurality of genes pair the *D. sechellia* cluster with a mix of *simulans* and *mauritiana* gene copies, including *ase* and *ci* (Hilton *et al.* 1994), as well as *est6*, *hb*, *In(2L)t*, *janus*, and *yp2* (Hey and Kliman 1993). Five genes reveal a *simulans* gene copy, or a cluster of *simulans* copies, as the next most similar to the *D. sechellia* cluster, including *Adh*, *per* (Kliman and Hey 1993a), *Sxl*, *w*, and *z* (Hey and Kliman 1993). Just 2 genes, *mt:ND5* and *Zw*, have a *sechellia/mauritiana* pairing. On balance, there is a suggestion that the origin of what we call *D. sechellia* arose prior to the splitting that gave

rise to our other species. This was the conclusion based on just 3 genes (Hey and Kliman 1993), and now with 14 genes we see that a plurality of the cluster diagrams favor this explanation. This conclusion is also supported by a recent study of the Odysseus (*OdsH*) locus that contributes to hybrid sterility between *D. mauritiana* and *D. simulans*. Multiple sequences from each of the three species revealed a striking pattern of very low polymorphism within species and multiple fixed differences between species (Ting *et al.* 2000). When considered in light of the relative paucity of fixed differences found at other genes, this pattern is strongly suggestive of multiple recurrent selective sweeps at this locus. The *OdsH* coding region sequences of the three taxa appear quite separate and distinct on the estimated gene tree, with those from *D. simulans* and *D. mauritiana* more closely related to each other than either are to those of *D. sechellia* (Ting *et al.* 2000).

It is noteworthy that what appears to be the most unlikely pairing for the most recent speciation event, on the basis of these cluster analyses and on biogeographic grounds (*D. sechellia* and *D. mauritiana*), was the favored topology in a study that brought together multiple comparative DNA sequence data sets (Caccone *et al.* 1996). Caccone *et al.* included data sets for which there were only single copies from each taxon, as well as those available data sets with multiple sequences from each. When multiple sequences were available the data were collapsed within taxa, so as to represent each taxon by just a single sequence, with polymorphisms represented using the IUPAC ambiguity codes (A. Caccone, personal communication). Different genes supported different topologies, but when all the data were combined into one large data set (*i.e.*, one long sequence for each species), the result was strong support for the *sechellia/mauritiana* pairing. This result was not sensitive to inclusion of the ambiguous (*i.e.*, polymorphic) positions. For three reasons, we do not further explore why Caccone *et al.*'s method of data combining would yield a network that is at odds with the data from most of the 14 genes studied here. First, it is difficult to assess whether the collapsing and combining of data from many genes, with widely varying histories, might lead to a misinterpretation of closely spaced speciation events. Second, the data presented here cannot rule out any particular bifurcating topology—though the *sechellia/mauritiana* pairing seems unlikely. Third, we have tried to avoid imposing a traditional phylogenetic model on our analyses. Such models necessarily employ assumptions of instantaneous splitting among distinct homogeneous entities. In the diversification of the *simulans* complex, we have the opportunity to understand phylogeny in a broader sense.

It is worth noting that the difficulty of inferring a branching species history is probably not a simple byproduct of too little data. The 14-locus data set comprises very nearly 220,000 bp of DNA sequence, not including the *D. melanogaster* outgroup sequences, and there are a total of 554 polymorphic sites, including 320 so-called "phylogenetically informative" polymorphisms (*i.e.*, the rarer base occurs more than once). Also, as these are very closely related DNA sequences, only a small fraction of these polymorphisms are expected to have occurred at the same site (see results). One might suppose that a data set with just three taxa and hundreds of informative sites (with little recurrent mutation) would permit a straightforward, traditional, phylogenetic resolution, but clearly it does not.

If we consider "phylogeny" as pertaining to the genesis of phyla then we have good reasons for eschewing most analyses that impose a simple bifurcating model on the history of these species. All three species are similarly related to one another, and the data suggest that all three have been evolving as separate entities for about the same amount of time. It also appears that divergence has been accruing in a manner consistent with allopatric speciation. If that is correct then we must also consider the likelihood that there was an extended period of time when multiple separate, but nonreproductively isolated, populations existed. The isolation model used here for some analyses assumes an instantaneous population splitting event, but even if that is accurate, neither that model nor any of our data help us to think about the origins of reproductive isolation. Given the recency of these speciation events, their evident proximity in time to one another, and the biological necessity that such events encompassed some time, there seems a large chance that we could misunderstand history if we were to take "speciation event" too literally as denoting an instance in time. For example, under allopatry and the Dobzhansky/Muller model (Dobzhansky 1936; Muller 1940), it would have taken some time for independent adaptive mutations to arise and sweep to fixation in the separate populations.

There are also a number of ways that the demographic circumstances associated with the origin of these taxa could positively mislead any attempt to impose a bifurcating model. For example, if the ancestral species consisted of multiple populations with limited gene exchange, with differentiation and local adaptation then the divergence of multiple species out of this ancestral species could be expected to reflect this structure. Indeed, there is evidence that *D. simulans* once had more population structure than we find at present (Hamblin and Veuille 1999). It is entirely possible that conclusions from a majority of gene trees, or a combined data set, might mistakenly reflect this population structure and fail to reflect the actual sequence of speciation events.

A synthesis: If we draw from the current biogeography and patterns of DNA sequence similarities, then it appears as if there were two island colonization events by flies that came from a large continental population. Given the large variation in DNA sequence similarities,

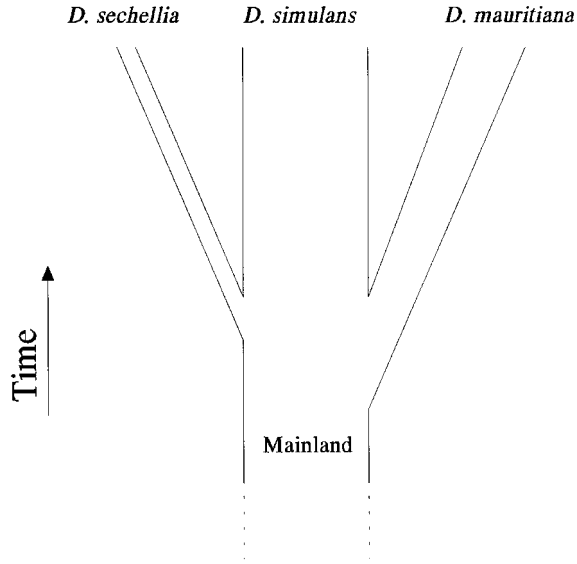

FIGURE 3.—Diagram of one mainland population of constant size giving rise to two island populations of different sizes.

particularly in the way that sequences from different species cluster, it seems nearly certain that a large amount of the variation that presently occurs among species includes samples of the variation that was present in ancestral species. If the two colonization events happened nearly at the same time, then different genes are expected to suggest different orders and topologies for these population splitting events.

Consider a model in which a large continental species gives rise to two smaller isolated populations on offshore islands, and that after formation these island populations are constant in size and exchange no genes with the mainland population (Figure 3). Then the expected amount of divergence between a gene copy from an island endemic and the mainland species can be expressed as a function of the time since splitting, the mutation rate since splitting, and the amount of variation within the mainland ancestral species. Let d_{im} be the average number of base pair differences between the island species (i) and the mainland species (m); let t be the time of island population formation; and let u_i and u_m be the respective mutation rates per year experienced by each. Then

$$d_{im} = tu_i + tu_m + \Pi_m, \quad (4)$$

and

$$t = (d_{im} - \Pi_m)/(u_i + u_m), \quad (5)$$

where Π_m is the average number of differences between two sequences in the mainland population at $t = 0$. Note that under the assumption of constant population size Π_m can be estimated by taking the average number of differences between two sequences from our mainland species, *D. simulans*. This quantity, summed across the 14 loci, is 123.7. The reason for including separate mutation rates after splitting is that we have clear evidence from the relative rate tests that *D. mauritiana* and *D. sechellia* have been evolving faster than *D. simulans* since the time of common ancestry. From the overall difference in branch lengths, we can estimate for *D. sechellia* that $u_i = u_m \cdot 1.49$ (*i.e.*, 0.0137/0.0092; see RESULTS, *Divergence of genes*). Similarly we can estimate for *D. mauritiana* that $u_i = u_m \cdot 1.20$ (*i.e.*, 0.0122/0.0102; see RESULTS, *Divergence of genes*). The estimate of d_{im} is simply the average number of pairwise differences between sequences from the island and mainland species, summed across the 14 loci; for *D. sechellia* it is 205.4 and for *D. mauritiana* it is 169.7. Substituting these quantities into expression (5) we find for the divergence between *D. simulans* and *D. sechellia* that $t = 32.8/u_m$. Similarly for the divergence between *D. simulans* and *D. mauritiana* we find $t = 20.9/u_m$.

The absolute time can be roughly assessed by assuming that u_m applies to the divergence between *D. melanogaster* and *D. simulans*. The average of the pairwise differences between these species, summed across these 14 loci, is 476.62. If we assume that the separation of these gene copies was ~3 million years ago (HEY and KLIMAN 1993), then $u_m = 476.62/(2 \cdot 3 \cdot 10^6) = 7.94 \cdot 10^{-5}$. Applying this rate we obtain an estimate of t, for *D. sechellia*, of 413,000 years and of t, for *D. mauritiana*, of 263,000 years. These dates scale linearly with any estimate of u_m, and it should be noted that the 3 million year date is very rough, as it relies upon a few amber fossils of early Drosophilids of somewhat uncertain age (THROCKMORTON 1975) and an assumption of a molecular clock (HEY and KLIMAN 1993; KLIMAN and HEY 1993a).

We thank Constantin Yanicostas for assistance with *janus*. R.M.K. and J.H. were supported by National Institutes of Health (NIH) grant R01GM58060. R.M.K. also received support from the Jeffress Memorial Trust. F.D. was supported by "Groupe de Recherche sur les Genomés" grant GREG92-392 to Michel Veuille. J.C. was supported by NIH GM 58260. J.W. was supported by National Science Foundation grant DEB-9815367.

LITERATURE CITED

AGUADÉ, M., N. MIYASHITA and C. H. LANGLEY, 1992 Polymorphism and divergence in the Mst26A male accessory gland gene region in Drosophila. Genetics **132:** 755–770.

AKASHI, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136:** 927–935.

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

ANDOLFATTO, P., J. D. WALL and M. KREITMAN, 1999 Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of *Drosophila melanogaster*. Genetics **153:** 1297–1311.

AVISE, J. C., 1989 Gene trees and organismal histories: a phylogenetic approach to population biology. Evolution **43:** 1192–1208.

BARRACLOUGH, T. G., and A. P. VOGLER, 2000 Detecting the geographical pattern of speciation from species-level phylogenies. Am. Nat. **155:** 419–434.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519–520.

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.

Bridges, C. B., 1938 A revised map of the salivary gland X-chromosome of *Drosophila melanogaster*. J. Hered. **29:** 11–13.

Bush, G. L., 1969 Sympatric host race formation and speciation in frugivorous flies of the genus *Rhagoletis* (Diptera, Tephritidae). Evolution **23:** 237–251.

Caccone, A., G. D. Amato and J. R. Powell, 1988 Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup. Genetics **118:** 671–683.

Caccone, A., E. N. Moriyama, J. M. Gleason, L. Nigro and J. R. Powell, 1996 A molecular phylogeny for the *Drosophila melanogaster* subgroup and the problem of polymorphism data. Mol. Biol. Evol. **13:** 1224–1232.

Cariou, M. L., 1987 Biochemical phylogeny of the eight species in the *Drosophila melanogaster* subgroup, including *D. sechellia* and *D. orena*. Genet. Res. **50:** 181–185.

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

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

Clark, A. G., 1997 Neutral behavior of shared polymorphisms. Proc. Natl. Acad. Sci. USA **94:** 7730–7734.

Cohn, V. H., M. A. Thompson and G. P. Moore, 1984 Nucleotide sequence comparison of the Adh gene in three Drosophilids. J. Mol. Evol. **20:** 31–37.

Cooke, P. H., and J. G. Oakeshott, 1989 Amino acid polymorphisms for esterase 6 in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **86:** 1426.

Coyne, J. A., 1992 Genetics of sexual isolation in females of the *Drosophila simulans* species complex. Genet. Res. **60:** 25–31.

Coyne, J. A., and B. Charlesworth, 1997 Genetics of a pheromonal difference affecting sexual isolation between *Drosophila mauritiana* and *D. sechellia*. Genetics **145:** 1015–1030.

Coyne, J. A., and M. Kreitman, 1986 Evolutionary genetics of two sibling species, *Drosophila simulans* and *D. sechellia*. Evolution **40:** 673–691.

Coyne, J. A., and T. D. Price, 2000 Little evidence for sympatric speciation in birds. Evolution (in press).

Coyne, J. A., A. P. Crittenden and K. Mah, 1994 Genetics of a pheromonal difference contributing to reproductive isolation in Drosophila. Science **265:** 1461–1464.

Da Lage, J. L., E. Renard, F. Chartois, F. Lemeunier and M. L. Cariou, 1998 Amyrel, a paralogous gene of the amylase gene family in *Drosophila melanogaster* and the Sophophora subgenus. Proc. Natl. Acad. Sci. USA **95:** 6848–6853.

Dobzhansky, T., 1936 Studies of hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics **21:** 113–135.

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

Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. USA **96:** 4482–4487.

Eanes, W. F., M. Kirchner and J. Yoon, 1993 Evidence for adaptive evolution of the G6pd gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. Proc. Natl. Acad. Sci. USA **90:** 7475–7479.

Eanes, W. F., M. Kirchner, J. Yoon, C. H. Biermann, I. N. Wang *et al.*, 1996 Historical selection, amino acid polymorphism and lineage-specific divergence at the *G6pd* locus in *Drosophila melanogaster* and *D. simulans*. Genetics **144:** 1027–1041.

Hamblin, M. T., and M. Veuille, 1999 Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. Genetics **153:** 305–317.

Harr, B., S. Weiss, J. R. David, G. Brem and C. Schlotterer, 1998 A microsatellite-based multilocus phylogeny of the *Drosophila melanogaster* species complex. Curr. Biol. **8:** 1183–1186.

Hey, J., 1994 Bridging phylogenetics and population genetics with gene tree models, pp. 435–449 in *Molecular Approaches to Ecology and Evolution*, edited by B. Schierwater, B. Streit, G. Wagner and R. DeSalle. Birkhäuser-Verlag, Basel.

Hey, J., 1999 The neutralist, the fly and the selectionist. Trends Ecol. Evol. **14:** 35–38.

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

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

Higuchi, R. G., and H. Ochman, 1989 Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucleic Acids Res. **17:** 5865.

Hilton, H., and J. Hey, 1996 DNA sequence variation at the period locus reveals the history of species and speciation events in the *Drosophila virilis* group. Genetics **144:** 1015–1025.

Hilton, H., and J. Hey, 1997 A multilocus view of speciation in the *Drosophila virilis* group reveals complex histories and taxonomic conflicts. Genet. Res. **70:** 185–194.

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

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

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by P. H. Harvey and L. Partridge. Oxford University Press, New York.

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

Joly, D., 1987 Between species divergence of cyst length distributions in the *Drosophila melanogaster* species complex. Jpn. J. Genet. **62:** 257–263.

Jones, C. D., 1998 The genetic basis of *Drosophila sechellia*'s resistance to a host plant toxin. Genetics **149:** 1899–1908.

Karotam, J., A. C. Delves and J. G. Oakeshott, 1993 Conservation and change in structural and 5′ flanking sequences of esterase 6 in sibling Drosophila species. Genetica **88:** 11–28.

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.

Kliman, R. M., and J. Hey, 1993b Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239–1258.

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

Kreitman, M., 1996 The neutral theory is dead. Long live the neutral theory. Bioessays **18:** 678–683.

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

Leicht, B. G., S. V. Muse, M. Hanczyc and A. G. Clark, 1995 Constraints on intron evolution in the gene encoding the myosin alkali light chain in Drosophila. Genetics **139:** 299–308.

Lemeunier, F., and M. Ashburner, 1976 Relationships within the melanogaster species subgroup of the genus Drosophila II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc. R. Soc. Lond. Ser. B Biol. Sci. **193:** 275–294.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

McAllister, B. F., and B. Charlesworth, 1999 Reduced sequence variability on the Neo-Y chromosome of *Drosophila americana americana*. Genetics **153:** 221–233.

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

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Muller, H. J., 1940 Bearings of the Drosophila work on systematics,

pp. 185–268 in *The New Systematics*, edited by J. Huxley. Clarendon Press, Oxford, UK.

Ohta, T., 1972 Evolutionary rate of cistrons and DNA divergence. J. Mol. Evol. **1:** 150–157.

Ohta, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature **246:** 96–98.

Ohta, T., 1996 The current significance and standing of neutral and neutral theories. Bioessays **18:** 673–677.

Palopoli, M. F., A. W. Davis and C. I. Wu, 1996 Discord between the phylogenies inferred from molecular *vs.* functional data: uneven rates of functional evolution or low levels of gene flow? Genetics **144:** 1321–1328.

Price, C. S. C., 1997 Conspecific sperm precedence in Drosophila. Nature **388:** 663–666.

Price, C. S., K. A. Dyer and J. A. Coyne, 1999 Sperm competition between Drosophila males involves both displacement and incapacitation. Nature **400:** 449–452.

Ramos-Onsins, S., and M. Aguadé, 1998 Molecular evolution of the Cecropin multigene family in Drosophila. Functional genes *vs.* pseudogenes. Genetics **150:** 157–171.

Rice, W. R., and E. F. Hostert, 1993 Laboratory experiments on speciation: what have we learned in 40 years. Evolution **47:** 1637–1653.

R'Kha, S., P. Capy and J. R. David, 1991 Host-plant specialization in the *Drosophila melanogaster* species complex: a physiological, behavioral, and genetical analysis. Proc. Natl. Acad. Sci. USA **88:** 1835–1839.

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

Sanger, F., S. Nicklen and A. R. Coulson, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

Satta, Y., and N. Takahata, 1990 Evolution of Drosophila mitochondrial DNA and the history of the *melanogaster* subgroup. Proc. Natl. Acad. Sci. USA **87:** 9558–9562.

Schliewen, U. K., D. Tautz and S. Paabo, 1994 Sympatric speciation suggested by monophyly of crater lake cichlids. Nature **368:** 629–632.

Shibata, H., and T. Yamazaki, 1995 Molecular evolution of the duplicated Amy locus in the *Drosophila melanogaster* species subgroup: concerted evolution only in the coding region and an excess of nonsynonymous substitutions in speciation. Genetics **141:** 223–236.

Snook, R. R., T. A. Markow and T. L. Karr, 1994 Functional nonequivalence of sperm in *Drosophila pseudoobscura*. Proc. Natl. Acad. Sci. USA **91:** 11222–11226.

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

Tajima, F., 1989a The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

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

Takahata, N., 1986 An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet. Res. **48:** 187–190.

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

Ting, C. T., S. C. Tsaur and C. I. Wu, 2000 The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *odysseus*. Proc. Natl. Acad. Sci. USA **97:** 5313–5316.

Tsacas, L., and G. Baechli, 1981 *Drosophila sechellia*, n.sp., huitieme espece du sous-goupe *melanogaster* des Iles Sechelles (Diptera, Drosophilidae). Rev. Fr. Entomol. **3:** 146–150.

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

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

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

Wang, R. L., J. Wakeley and J. Hey, 1997 Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. Genetics **147:** 1091–1106.

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

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

Wu, C. I., and W. H. Li, 1985 Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82:** 1741–1745.

Wu, C. I., and M. F. Palopoli, 1994 Genetics of postmating reproductive isolation in animals. Annu. Rev. Genet. **28:** 283–308.

Yanicostas, C., A. Vincent and J. A. Lepesant, 1989 Transcriptional and posttranscriptional regulation contributes to the sex-regulated expression of two sequence-related genes at the *janus* locus of *Drosophila melanogaster*. Mol. Cell. Biol. **9:** 2526–2535.

Zurovcova, M., and W. F. Eanes, 1999 Lack of nucleotide polymorphism in the Y-linked sperm flagellar dynein gene *Dhc-Yh3* of *Drosophila melanogaster* and *D. simulans*. Genetics **153:** 1709–1715.

Communicating editor: W. F. Eanes