DNA Variation at the *rp49* Gene Region of *Drosophila simulans*: Evolutionary Inferences From an Unusual Haplotype Structure

Julio Rozas, Myriam Gullaud,[1] Gaëlle Blandin[2] and Montserrat Aguadé

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, 08071 Barcelona, Spain

Manuscript received January 13, 2001
Accepted for publication April 11, 2001

ABSTRACT

An ~1.3-kb region including the *rp49* gene plus its 5′ and 3′ flanking regions was sequenced in 24 lines of *Drosophila simulans* (10 from Spain and 14 from Mozambique). Fifty-four nucleotide and 8 length polymorphisms were detected. All nucleotide polymorphisms were silent: 52 in noncoding regions and 2 at synonymous sites in the coding region. Estimated silent nucleotide diversity was similar in both populations ($\pi = 0.016$, for the total sample). Nucleotide variation revealed an unusual haplotype structure showing a subset of 11 sequences with a single polymorphism. This haplotype was present at intermediate frequencies in both the European and the African samples. The presence of such a major haplotype in a highly recombining region is incompatible with the neutral equilibrium model. This haplotype structure in both a derived and a putatively ancestral population can be most parsimoniously explained by positive selection. As the rate of recombination in the *rp49* region is high, the target of selection should be close to or within the region studied.

DROSOPHILA simulans, like *D. melanogaster*, is a cosmopolitan human commensal that originated in tropical Africa ~2.5 mya (Lachaise *et al.* 1988; Powell 1997). Populations of both species from that area could be considered ancestral populations and, therefore, neutral variation in those populations would be expected to be at mutation-drift equilibrium. On the other hand, populations from other areas would be derived populations and their variation might or might not be at equilibrium. In fact, there would have been ample room for adaptive evolution, *i.e.*, for the action of natural selection, during the dispersal of these species from tropical to temperate regions. Additionally, derived populations of these species might still reflect the possible founder events associated with the out-of-Africa expansion of these species.

Both demographic events and natural selection acting in a particular genomic region can have similar effects on the pattern of nucleotide variation in that region. However, population history has a genome-wide effect and should affect, therefore, all regions of the genome. In contrast, both directional and balancing selection are locus specific and affect neutral variation only at loci tightly linked to the locus under selection. In this sense, the level of both within-population variation and between-population differentiation is generally lower for morphological characters, allozymes, and mitochondrial DNA variation in *D. simulans* than in *D. melanogaster* (as reviewed in Singh and Long 1992); this might reflect a more recent expansion of the distribution area of *D. simulans* (but see Ballard *et al.* 1996 for mtDNA).

Initial surveys of nucleotide sequence variation in nuclear genes of *D. simulans* generally analyzed few sequences sampled from different populations (see Begun and Whitley 2000 for references). An unusual haplotype structure was detected in some of these surveys (Begun and Aquadro 1995; Eanes *et al.* 1996; Hasson *et al.* 1998; Labate *et al.* 1999), suggesting the presence of at least two old lineages in this species.

Hamblin and Veuille (1999) have studied variation in two *X*-linked gene regions (*vermilion* and *G6pd*) in several African and non-African populations of *D. simulans.* A strong haplotype substructure was detected both in West African and in all non-African populations studied. The haplotype structure observed in the *vermilion* region departed from predictions of the neutral theory for stationary populations in the non-African, but not in Central African populations. This departure from neutrality was considered evidence of a bottleneck in their rather recent foundation (Hamblin and Veuille 1999). The authors also suggested that the non-equilibrium haplotype distributions might be compatible with ancient population subdivision and recent admixture in populations of *D. simulans* ancestral to the European and American populations.

The paucity in the number of haplotypes and/or in haplotype diversity detected in three out of the four loci analyzed in non-African populations by Hamblin

Corresponding author: Julio Rozas, Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08071 Barcelona, Spain. E-mail: julio@bio.ub.es

[1] *Present address:* Laboratoire de Génétique Moléculaire de la Différenciation, Institut Jacques Monod, 75251 Paris Cedex 05, France.

[2] *Present address:* Unité de Génétique Moléculaire des Levures, Institut Pasteur, 75724 Paris Cedex 15, France.

and Veuille (1999) was considered to support a genome-wide phenomenon and, therefore, the population admixture hypothesis. In *vermilion*, however, the region surveyed was not randomly chosen, but its choice was based on the previous knowledge that it presented two divergent haplotypes (Begun and Aquadro 1995). Indeed, in the North American sample sequenced by Begun and Aquadro (1995), the haplotype test gave different results when applied to the complete region than when applied to the particular region chosen by Hamblin and Veuille (1999). To draw any general conclusion, it is important, therefore, to survey nucleotide variation in African and non-African samples of *D. simulans* in a larger number of randomly chosen regions.

We have analyzed variation in an ~1.3-kb region encompassing the *rp49* gene (named *RpL32* in the FlyBase *Drosophila* database; http://flybase.bio.indiana.edu) in a European and a Southeast African population of *D. simulans.* This gene is located in band 99D of *D. simulans* and encodes ribosomal protein 49 (ribosomal protein L32 in FlyBase). Similarly to the *vermilion* region surveyed by Hamblin and Veuille (1999), recombination in this autosomal region is expected to be high (Kliman and Hey 1993a; True *et al.* 1996). Extensive surveys of variation in the homologous region of *D. subobscura* in relation to chromosomal polymorphism (Rozas and Aguadé 1993, 1994; Rozas *et al.* 1999) indicate that, at least in this species, the *rp49* region is a neutrally evolving region with normal levels of nucleotide polymorphism. Surprisingly, our initial survey of nucleotide variation in a European population of *D. simulans* revealed an unusual haplotype structure showing one rather common haplotype with zero variation even though the complete sample had a normal level of variation. This motivated extending the survey to a putatively ancestral population of that species in an effort to discern between historical and selective explanations.

MATERIALS AND METHODS

Fly samples: Twenty-four lines randomly sampled from two natural populations of *D. simulans* were studied: 10 lines from Montblanc, Tarragona, Spain (SimS lines) and 14 from Maputo, Mozambique (SimMz lines). The European and African samples were collected in September 1993 and in August 1997, respectively. We obtained highly inbred lines after 10 generations of sibmating. We also used 1 line of *D. melanogaster* (line M66), which was collected in Montemayor, Córdoba, Spain in March 1990 and was subsequently made isochromosomal for the third chromosome by the standard series of crosses with the *TM6/MKRS* balancer stock.

DNA extraction, PCR amplification, and DNA sequencing: Genomic DNA was extracted using a modification of protocol 48 from Ashburner (1989). An ~1.4-kb fragment, which included the *rp49* gene (402 bp of coding region and a small intron of 59 bp) and its 5′ and 3′ flanking regions, was amplified by PCR (Saiki *et al.* 1988) using oligonucleotides designed on the published sequence of *D. melanogaster* (O'Connell and Rosbash 1984; GenBank accession no. X00848). Several oligonucleotides, designed at intervals of ~300 nucleotides, were used as primers for sequencing. The amplified fragments were cyclesequenced and separated on a Perkin-Elmer (Norwalk, CT) ABI PRISM 377 automated DNA sequencer following the manufacturer's instructions. For each line, the DNA was sequenced on both strands. The nucleotide sequences are available from the EMBL nucleotide sequence database under accession nos. Y13939 (*D. melanogaster*) and AJ309023–AJ309046 (*D. simulans*).

Data analysis: Nucleotide sequences were assembled using the SeqEd version 1.0.3 program (Applied Biosystems, Inc., Foster City, CA), multiply aligned using the Clustal W program (Thompson *et al.* 1994), and edited with the MacClade version 3.06 program (Maddison and Maddison 1992). Phylogenetic analysis was performed with genetic distances corrected according to the Jukes and Cantor model (Jukes and Cantor 1969) using the neighbor-joining algorithm (Saitou and Nei 1987) implemented in the MEGA version 2 (Kumar *et al.* 2000) program. The analysis was conducted using the *rp49* nucleotide sequence of *D. melanogaster* (line M66) as the outgroup; the bootstrap values were based on 1000 replicates.

The DnaSP version 3.50 software (Rozas and Rozas 1999) was used to estimate population genetic parameters and genetic distances and also to perform different neutrality tests. The confidence intervals (and the *P* values) of several test statistics were obtained by Monte Carlo simulations based on the coalescent process for a neutral infinite-sites model and assuming a large and constant population size (Kingman 1982a,b; Hudson 1983, 1990). The simulations were carried out either by assuming a value of θ ($\theta = 4Nu$, where N is the effective population size and u is the per gene mutation rate; Watterson 1975) or by fixing the number of segregating sites. As in both cases the simulations yielded similar results, we present only results based on the coalescent conditional on the number of segregating sites. The simulations were performed assuming either no intragenic recombination or intermediate levels of recombination (Hudson 1983, 1990). Each computer simulation was based on 10,000 (for no recombination) or 1000 (for intermediate levels of recombination) independent replicates. The empirical distribution of the corresponding statistic was thus generated and used to determine the confidence intervals.

The recombination parameter C (in Drosophila $C = 2Nc$, where N is the effective population size and c is the recombination rate per generation between the most distant sites) was estimated using the methods of Hudson and Kaplan (1985) and of Hudson (1987). The first method is based on R_M or the minimum number of recombination events in the sample; estimates of R_M were used to estimate C by coalescent simulations. The Hudson (1987) method is based on the variance of the number of differences between pairs of sequences; in that case, the estimate of C can be obtained numerically. We also estimated the minimum value of C compatible with the observed value of R_M (C_L); thus, C_L is an underestimate of the true C value. The C_L value was estimated as the lowest value of C for which the right tail (5%) of the R_M distribution contains values equal to or higher than the observed value of R_M.

An estimate of C based on the estimates of c (True *et al.* 1996) in the *rp49* region, or C_M, was also obtained (C_M = 49.2). This estimate was obtained following Andolfatto and Przeworski (2000) and considering that (1) the *rp49* and the *Tpi* regions (located in cytological bands 99D and 99E, respectively) have the same recombination rates (*i.e.*, $c = 0.92 \times 10^{-8}$), (2) in *D. simulans* $N = 2 \times 10^6$, and (3) the average length of the *rp49* region is 1337 bp.

The overall genetic association between polymorphic sites was measured by the Z_{nS} statistic (Kelly 1997), which is the average of r^2 (Hill and Robertson 1968) over all pairwise comparisons,

$$Z_{nS} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} r_{i,j}^2,$$

where S is the number of polymorphic sites and $r_{i,j}$ is the r estimator (Hill and Robertson 1968) between sites i and j. The confidence intervals of Z_{nS} were determined by computer simulations using the coalescent algorithm.

The effect of intragenic recombination on nucleotide variation was studied by analyzing the levels of linkage disequilibrium between polymorphic sites in relation to the physical distance. A new test statistic, *ZZ*, that is defined as

$$ZZ = Z_A - Z_{nS} \tag{1}$$

was developed, where

$$Z_A = \frac{1}{(S-1)} \sum_{i=1}^{S-1} r_{i,i+1}^2, \tag{2}$$

and Z_{nS} is the Kelly (1997) statistic. The Z_A statistic is the average of r^2 (Hill and Robertson 1968), but only between adjacent polymorphic sites. Because linkage disequilibrium decays with physical distance due to intragenic recombination, the *ZZ* statistic is expected to have larger positive values with increasing recombination, and eventually it could be used to estimate the recombination parameter *C*. Although in regions with high levels of recombination R_M estimates might be inflated by parallel mutation, *ZZ* values would probably not be affected. Confidence intervals of the *ZZ* statistic were determined by coalescent simulations. An algorithm for computing the *ZZ* statistic from DNA sequence data and for estimating its confidence intervals will be implemented in the next release of the DnaSP software (Rozas and Rozas 1999).

RESULTS

DNA sequence variation: The *rp49* gene plus its 5′ and 3′ flanking regions were sequenced in 24 lines of *D. simulans* (10 from Europe and 14 from Africa) and in one line of *D. melanogaster*. In *D. simulans*, a total of 54 polymorphic nucleotide sites (corresponding to 56 mutations) were identified over the 1292 bp examined (excluding all sites with alignment gaps). Polymorphisms at sites 626 and 713 (exon 2) were synonymous, whereas the rest were in noncoding regions (Figure 1). Eight insertion/deletion polymorphisms (ranging from 1 to 29 bp in length) were also detected in noncoding regions. Estimates of nucleotide variation are shown in Table 1.

Ten of the 24 *rp49* sequences surveyed were identical (for both nucleotide and insertion/deletion changes); there was an additional sequence (line SimMz7) that differed from this common haplotype by a single nucleotide substitution (Figure 1). These 11 sequences were designated as L#a. Most other sequences were designated as L#b. Two sequences (SimS13 and SimMz39) probably originated by recombination between the two divergent major haplotypes (L#a and L#b). Lines with the L#a haplotype were found both in Montblanc (six lines) and in Maputo (five lines); the frequency of this haplotype did not differ significantly between both populations (Fisher's exact test, $P = 0.41$).

Estimates of nucleotide divergence between populations ($d_{xy} = 0.0118$, and $d_a = 0.000$) and F_{ST} values ($F_{ST} = 0.020$) were consistent with weak population subdivision. The methods of Roff and Bentzen (1989) and of Hudson *et al.* (1992) were used for detecting genetic differentiation between populations (considering either all sequences or only L#b sequences). None of the tests performed detected any significant differentiation between populations (results not shown). This lack of genetic differentiation between populations is reflected in the neighbor-joining tree, where European and African sequences are interspersed (Figure 2).

Intragenic recombination and linkage disequilibrium: We tested the effect of intragenic recombination on nucleotide sequence variation (Table 2). The estimated *ZZ* values were significantly positive, suggesting that in this region intragenic recombination has played an important role in shuffling nucleotide variation among DNA sequences. Estimates of the recombination parameter *C* obtained by the methods of Hudson and Kaplan (1985) and of Hudson (1987) are shown in Table 2. The discrepancy between both estimates could be due to the particular structure of genetic variation found at the *rp49* gene region (see below). This structure would also cause the C_L values to be conservative; indeed, larger C_L values were obtained when a single L#a sequence was considered in the computer simulations (16.9 and 2.7 for Montblanc and Maputo, respectively).

The significance of the pairwise associations between polymorphic sites, or linkage disequilibrium, was established by the chi-square test. In the total sample, 332 out of 1326 pairwise comparisons showed a significant association; 86 of these comparisons remained significant after applying the Bonferroni procedure. No significant overall association between polymorphic sites was detected by using the Z_{nS} statistic (Kelly 1997) without recombination or introducing conservative recombination estimates ($C = C_L$; results not shown).

Neutrality tests: We tested whether the observed pattern of nucleotide variation is compatible with that expected under neutrality. We applied several tests that compare different estimates of θ either using only intraspecific data (Tajima 1989) or using intraspecific data and sequence information of another species (the outgroup) to determine the polarity of mutations (Fu and Li 1993; Fay and Wu 2000). All these tests failed to reject the neutral equilibrium model (Table 1). The HKA test (Hudson *et al.* 1987) was conducted to assess whether levels of polymorphism and divergence were correlated. We compared polymorphism (in *D. simulans*) and divergence (between *D. simulans* and *D. melanogaster*) in the *rp49* region (present results) and in the *vermilion* region in samples from North Carolina and the Congo (Begun and Aquadro 1995). None of the

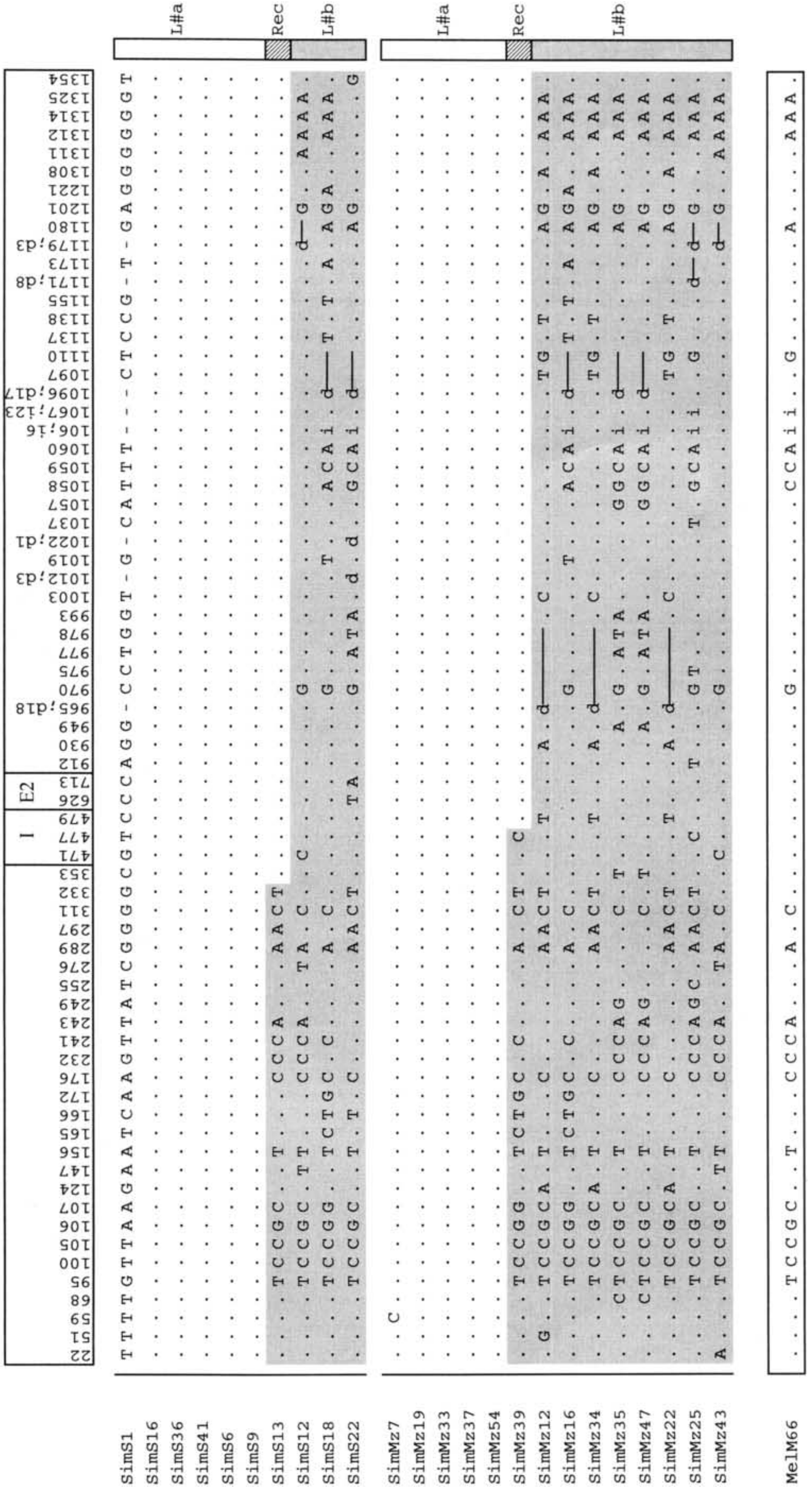

FIGURE 1.—Nucleotide polymorphism at the *rp49* gene region in *D. simulans*. Nucleotide numbering is according to the *rp49* sequence of the M66 line of *D. melanogaster*. Nucleotides identical to the first sequence are indicated by a dot. For length polymorphisms, the nucleotide position refers to the first site affected, and a dash indicates absence of the corresponding length variant. d, deletion; i, insertion; d#, deletion of #bp; i#, insertion of #bp. Polymorphic sites 471–479 correspond to the intron of the *rp49* gene (I), and polymorphic sites 626 and 713 to exon 2 (E2) of the *rp49* gene. SimS, *D. simulans* from Montblanc; SimMz, *D. simulans* from Maputo; MelM66, *D. melanogaster* line M66. A continuous line indicates a deletion affecting more than one polymorphic site. Shaded blocks indicate L#b sequence information. Rec, recombinant sequence. The last row gives the nucleotide information of the outgroup species *D. melanogaster* (putative ancestral state) for all polymorphic sites in *D. simulans*.

TABLE 1

Nucleotide polymorphism and divergence in the *rp49* region

	Montblanc	Maputo	Total
Sample size	10	14	24
S (η)	40 (42)	51 (53)	54 (56)
No. of sites	1318	1296	1292
No. of silent sites	1012.8	990.8	986.8
π	0.0103	0.0134	0.0120
π_S	0.0134	0.0175	0.0157
K^a	0.0553	0.0547	0.0543
K silent^b	0.0728	0.0724	0.0720
Tajima's D	−0.408 NS	0.184 NS	0.132 NS
Fu and Li's D	−0.690 NS	0.202 NS	0.842 NS
Fu and Li's F	−0.745 NS	0.241 NS	0.726 NS
Fay and Wu's H	−7.555 NS	5.275 NS	1.652 NS

S, number of segregating sites; η, total number of mutations; π, nucleotide diversity; π_s, silent nucleotide diversity; both noncoding positions and synonymous sites in the coding region were used to estimate π_s (Nei and Gojobori 1986; Nei 1987); NS, not significant.

[a] Average nucleotide divergence between *D. simulans* and *D. melanogaster* corrected by the Jukes and Cantor (1969) method.

[b] Average silent nucleotide divergence between *D. simulans* and *D. melanogaster* corrected by the Jukes and Cantor (1969) method.

HKA tests showed a significant deviation from neutral predictions (results not shown).

We tested by coalescent simulations whether the large number of identical sequences found in the sample was compatible with the neutral equilibrium model (see Hudson *et al.* 1994). We also investigated whether the presence of such a major haplotype was compatible with the equilibrium neutral model by analyzing the distribution of the number of haplotypes and of haplotype diversity (Ewens 1972; Strobeck 1987; Depaulis and Veuille 1998). According to the neutral model, both the number of haplotypes (and the number of identical sequences) and the haplotype diversity are a function of the sample size and of θ. The analyses were performed under conservative assumptions (under no recombination and using the conservative C_L estimate of the recombination parameter) and also under a more realistic assumption (using the C_M estimate that is based on the comparison of the physical and genetic maps). The analyses showed a significant (or nearly significant) excess of identical sequences and a significant (or nearly significant) reduction in the number of haplotypes and in the haplotype diversity values (Table 3). Values of Fu's F_S statistic (Fu 1997), which is related to Strobeck's S statistic, pointed in the same direction (results not shown).

DISCUSSION

In *D. simulans*, as in *D. melanogaster*, the *rp49* gene is located at band 99D where recombination is rather high (Kliman and Hey 1993a; True *et al.* 1996; Andolfatto

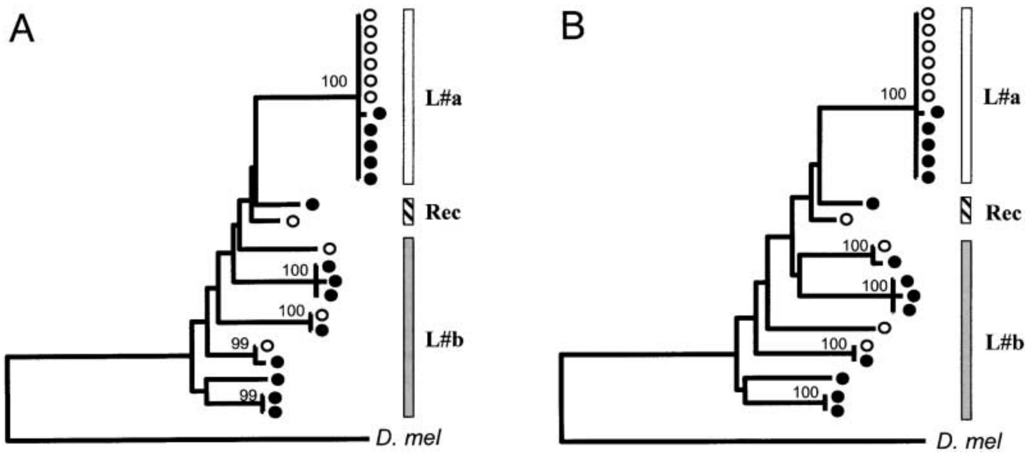

Figure 2.—Neighbor-joining tree of the *rp49* gene region sequences of *D. simulans.* Bootstrap values >90% are shown on the tree. *D. mel*, *D. melanogaster* (M66 line); Rec, recombinant sequence. The open and solid circles indicate lines from Montblanc and Maputo, respectively. (A) Tree built considering only nucleotide substitutions information. (B) Tree built using information on both nucleotide substitutions and length variants. Each indel was treated as a single mutational event.

TABLE 2

Estimates of the recombination parameter

	Montblanc	Maputo	Total
Sample size	10	14	24
R_M	4	3	5
C^a	31.8	10.7	18.4
C^b	2.4	27.4	11.8
$C_L{}^c$	7.3	2.4	5.5
ZZ (*P* value[d])	0.183 (0.000)	0.105 (0.005)	0.116 (0.002)
95% C.I. of *ZZ*	(−0.086, 0.091)	(−0.074, 0.078)	(−0.070, 0.076)

C.I., confidence interval.

[a] Estimate of the recombination parameter *C* from the minimum number of recombination events (Hudson and Kaplan 1985).

[b] Estimate of the *C* parameter from the variance of the number of differences between pairs of sequences (Hudson 1987).

[c] Minimum value of *C* (see text).

[d] Probability of obtaining values of the *ZZ* statistic equal or greater than the observed value; the probabilities were obtained from coalescent simulations with no recombination.

and Przeworski 2000; this article). The relatively high R_M values and the significant *ZZ* estimates obtained from our data would support the conclusion that recombination in this region is not reduced. Furthermore, *D. simulans* is monomorphic at the chromosomal level and thus, in this species, no local decrease of recombination is expected as a consequence of chromosomal polymorphism.

TABLE 3

Haplotype distribution tests

	Montblanc	Maputo	Total
Sample size	10	14	24
No. of identical lines, *l*	6	4	10
P value[a]; $C = 0$	0.005	0.205	0.014
P value[a]; $C = C_L$	0.001	0.162	0.004
P value[a]; $C = C_M$	0.000	0.023	0.000
No. of haplotypes, *k*	5	9	12
Strobeck's *S*[b]; $C = 0$	0.046	0.252	0.149
P value[c]; $C = C_L$	0.009	0.209	0.072
P value[c]; $C = C_M$	0.000	0.011	0.000
Haplotype diversity, *h*	0.667	0.912	0.826
P value[d]; $C = 0$	0.004	0.117	0.011
P value[d]; $C = C_L$	0.002	0.083	0.002
P value[d]; $C = C_M$	0.000	0.004	0.000

C, recombination parameter; C_L, lower bound estimate of *C*; C_M, estimate of *C* based on the physical and genetic maps comparison. The critical values for these tests were obtained by computer simulations based on the coalescent process, except for Strobeck's *S* statistic for which it was obtained analytically (Ewens 1972; Strobeck 1987).

[a] Probability of obtaining a number of identical lines equal to or greater than *l* (the observed value).

[b] Strobeck's *S* statistic represents the probability of observing a number of haplotypes equal to or lower than *k*.

[c] Probability of obtaining a number of haplotypes equal to or lower than *k* (the observed value).

[d] Probability of obtaining values of haplotype diversity equal to or lower than *h* (the observed value).

The estimated silent nucleotide variation in the *rp49* region ($\theta = 0.016$) was lower than estimates for other regions that were also located on the 3R chromosomal arm (the average θ for 19 genes was 0.035; Begun and Whitley 2000). However, most silent variation at the *rp49* region corresponds to variation at noncoding sites, while estimates in Begun and Whitley (2000) are based on synonymous variation. Our data conform, therefore, to the general observation of lower variation in noncoding flanking regions than at synonymous sites of coding regions (Moriyama and Powell 1996).

Haplotype substructure and demographic factors: Nucleotide variation at the *rp49* region in the two populations of *D. simulans* stands out because it is highly structured. Both the European and African samples present the same haplotype at intermediate frequency. They also share other minor haplotypes (Figures 1 and 2) and, in fact, no significant genetic differentiation was detected between populations.

In the subsample of lines that constitute the major haplotype in the *rp49* region (11 lines designated as L#a), there was a single polymorphism and its rarest variant was present in only one line. Forty-five polymorphisms segregated, however, in L#b lines ($n = 11$). There were nine fixed differences between L#a and L#b lines (Figure 1). The presence of such a major and divergent haplotype (L#a lines) at the *rp49* region is incompatible with the neutral equilibrium model, even in the absence of recombination.

Variation at the *vermilion* and *rp49* regions departed from neutral expectations in a similar way. There are, however, important differences between both sets of results. First, in the *vermilion* region, only non-African populations showed a significant reduction in haplotype number and/or haplotype diversity. Second, in that region there was no major haplotype common to all populations surveyed. Both observations in the *vermilion* region are compatible with an important founder effect in the origin of these derived populations. The haplotype structure detected in other regions surveyed in samples from non-African populations might also be the result of founder events (Hamblin and Veuille 1999; Labate *et al.* 1999; Andolfatto and Kreitman 2000; Duvernell and Eanes 2000). It has been also argued that the haplotype structure detected in most surveys of non-African populations could be the result of population subdivision and recent admixture.

The presence in the *rp49* region of the same major haplotype in an ancient and in a recently established population cannot be easily explained by founder events. It could be argued, however, that African populations of *D. simulans* were genetically differentiated (Hamblin and Veuille 1999) and that the Maputo population was a recently established population. Even in that case, it would be rather unlikely that the same haplotype (haplotype L#a) was present at relatively high frequency in both populations (Montblanc and Maputo). Variation at the *Acp26Aa* region in lines from the same collections does not show such a haplotype structure (M. Aguadé, unpublished results) suggesting that the observed haplotype structure in the *rp49* region is not the result of a genome-wide phenomenon. On the other hand, for the *rp49* region only one subset of lines is depleted of variation. If the pattern observed were due to population subdivision and recent admixture, one of the subpopulations should have been nearly monomorphic for this region. This is rather unlikely since the level of silent nucleotide diversity in regions with normal rates of recombination can be quite important, even in species with small effective population sizes such as *D. mauritiana* and *D. madeirensis* (*e.g.*, Hey and Kliman 1993; Kliman and Hey 1993b; Khadem *et al.* 2001; see also Gillespie 1999, 2000).

Haplotype structure and selective causes: In Drosophila, only a few surveys of DNA sequence variation in regions with normal (or high) levels of recombination have revealed a high proportion of sequences with zero (or nearly zero) variation in fragments longer than 1.3 kb. The first such pattern was detected in the *Sod* region of *D. melanogaster* (Hudson *et al.* 1994, 1997), and it was considered to reflect the hitchhiking effect of an advantageous mutation that was increasing in frequency. However, unlike in this study, only North American and European populations had been surveyed, which would always leave room for historical explanations. Similarly, a survey of DNA variation at the *runt* region of *D. simulans* in lines from different North American populations revealed that most lines (six out of eight) were identical or nearly identical (Labate *et al.* 1999).

Selection on favorable mutations can remove nucleotide variation at linked sites, causing a selective sweep or hitchhiking effect (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Barton 1998; Fay and Wu 2000; Gillespie 2000; Kim and Stephan 2000). Positive selection has thus been proposed to explain the pattern of nucleotide variation found at the *Sod* region in two populations of *D. melanogaster* (Hudson *et al.* 1994, 1997). The *rp49* data of *D. simulans* exhibits a pattern of nucleotide variation similar to that found for the *Sod* locus, *i.e.*, two sets of highly diverged sequences in a region with normal levels of recombination. Moreover, we have found the same haplotype structure not only in Europe, but also in a population from the putative ancestral distribution area. It is thus unlikely that the unexpected pattern of variation found at the *rp49* region was due to some founder event associated with the colonization of Europe by *D. simulans*. Therefore, positive selection would most parsimoniously explain the pattern of variation observed at this region. Although we did not detect the excess of high frequency-derived variants expected immediately after a selective sweep (Fay and Wu 2000; Kim and Stephan 2000), eight of the nine mutations fixed in L#a sequences (relative to L#b sequences) were derived (see Figure 1), which would support the selective hypothesis. There are, however, several selective scenarios that could explain the presence of a major haplotype with low variation: (1) the selected haplotype could be in its transient phase either to fixation or to an equilibrium frequency, (2) the pattern could reflect a very recently established balanced polymorphism, or (3) the advantageous mutation could have attained fixation, but the *rp49* region could be relatively far away from the advantageous mutation.

Because in *D. simulans* the *rp49* gene is located in a genomic region with high recombination, the fragment affected by the proposed selective sweep should be short (Kaplan *et al.* 1989; Stephan *et al.* 1992). In *D. melanogaster*, the genomic region encompassing the *rp49* gene presents a high density of coding regions: 12 coding regions have been identified in a 20-kb fragment spanning the *rp49* gene (FlyBase database). The conserved synteny between *D. melanogaster* and *D. simulans* allows prediction of a similar density in the latter species and, thus, any of these coding regions (or some regulatory regions) could have been the target of selection.

Time of hitchhiking: The time back to the hitchhiking event can be inferred from the amount of nucleotide variation present in the hitchhiked haplotype. For this inference, we need to know (1) the neutral mutation rate for the *rp49* region and (2) the expected topology of the gene genealogy. Assuming that silent substitutions (both at noncoding and synonymous sites) are neutral, the neutral mutation rate for the *rp49* region can be estimated from the estimated silent nucleotide divergence between *D. simulans* and *D. melanogaster* (K silent = 0.072; Table 1). Assuming that the split of the *D. melanogaster* and *D. simulans* lineages occurred 2.5

mya (Lachaise *et al.* 1988; Powell 1997), the silent mutation rate would be 1.4 × 10^{-8} per nucleotide and per year and 1.4 × 10^{-5} on a per sequence basis (as the *rp49* region includes 987 silent sites; Table 1). Several authors have shown (*e.g.*, Slatkin and Hudson 1991) that after a selective sweep the gene genealogy is star-like, *i.e.*, a genealogy compressed at the internal nodes. If mutations are Poisson distributed, the expected number of mutations on the genealogy is $\mu E(T)$, where T is the total length of the branches in the genealogy (in years), and μ is the mutation rate per sequence per year (Hudson 1990, Equation 1). In the *rp49* region, only one mutation (a singleton variant) was detected (site 59 in SimMz7 line) among all L#a sequences ($n = 11$). Assuming a star genealogy for our sample, T would be $11{*}t$ (where t is the time back to the hitchhiking event) and, consequently, t would be ∼6500 years. Thus, the proposed selective sweep would have occurred very recently. The lack of length variation in L#a sequences (Figures 1 and 2) would also support the conclusion that the hitchhiking event was rather recent.

Although hitchhiking would most consistently explain the pattern of variation observed in the *rp49* region, we have not definitively ruled out historical explanations. Indeed, the detection of some haplotype structure in other surveyed regions pointed to historical explanations. Only a multilocus approach using large population samples might allow discarding the admixture hypothesis. Also, analysis of variation across contiguous regions of the genome might be used to detect the differential signature of natural selection (Nurminsky *et al.* 2001). If the haplotype structure detected at the *rp49* region were due to hitchhiking, it would decay and eventually disappear at some distance of this region. Analysis of variation in genomic regions located at increasing distances from the *rp49* gene would, thus, allow contrasting of the selective hypothesis and, if confirmed, it would also allow delimiting the target of selection.

We thank Gema Blasco and David Salguero for technical support and Serveis Científico-Tècnics, Universitat de Barcelona, for automated sequencing facilities. We are especially grateful to A. Barbal and C. Arribas for collecting flies in Maputo, and S. O. Kolokotronis for his collaboration in the project. We also thank C. Segarra for valuable comments on the manuscript. G.B. and M.G. were supported by the Erasmus program between Université Denis Diderot, Paris, France, and Universitat de Barcelona, Spain. This work was supported by grants PB97-0918 from Comisión Interdepartamental de Ciencia y Tecnología, Spain and 1999SGR-25 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Catalonia, Spain, to M.A.

LITERATURE CITED

Andolfatto, P., and M. Kreitman, 2000 Molecular variation at the *In(2L)t* proximal breakpoint site in natural populations of *Drosophila melanogaster* and *D. simulans.* Genetics **154:** 1681–1691.

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257–268.

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Ballard, J. W. O., J. Hatzidakis, T. L. Karr and M. Kreitman, 1996 Reduced variation in *Drosophila simulans* mitochondrial DNA. Genetics **144:** 1519–1528.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. **72:** 123–133.

Begun, D. J., and C. F. Aquadro, 1995 Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans.* Genetics **140:** 1019–1032.

Begun, D. J., and P. Whitley, 2000 Reduced *X*-linked nucleotide polymorphism in *Drosophila simulans.* Proc. Natl. Acad. Sci. USA **97:** 5960–5965.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite site model. Mol. Biol. Evol. **15:** 1788–1790.

Duvernell, D. D., and W. F. Eanes, 2000 Contrasting molecular population genetics of four hexokinases in *Drosophila melanogaster, D. simulans* and *D. yakuba.* Genetics **156:** 1191–1201.

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.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Fu, Y.-X., 1997 Statistical tests of neutrality against population growth, hitchhiking and background selection. Genetics **147:** 915–925.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gillespie, J. H., 1999 The role of population size in molecular evolution. Theor. Popul. Biol. **55:** 145–156.

Gillespie, J. H., 2000 Genetic drift in an infinite population: the pseudohitchhicking model. Genetics **155:** 909–919.

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.

Hasson, E., I. N. Wang, L. W. Zeng, M. Kreitman and W. F. Eanes, 1998 Nucleotide variation in the triosephosphate isomerase (*Tpi*) locus of *Drosophila melanogaster* and *Drosophila simulans.* Mol. Biol. Evol. **15:** 756–769.

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.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

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

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

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

Hudson, R. R., D. D. Boos and N. L. Kaplan, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the Superoxide Dismutase (*Sod*) region of *Drosophila melanogaster.* Genetics **136:** 1329–1340.

Hudson, R. R., A. G. Sáez and F. J. Ayala, 1997 DNA variation at the *Sod* locus of *Drosophila melanogaster*: an unfolding story of natural selection. Proc. Natl. Acad. Sci. USA **94:** 7725–7729.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–120 in *Mammalian Protein Metabolism*, edited by H. W. Munro. Academic Press, New York.

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

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

Khadem, M., J. Rozas, C. Segarra and M. Aguadé, 2001 DNA variation at the *rp49* gene region in *Drosophila madeirensis* and *D. subobscura* from Madeira: inferences about the origin of an insular endemic species. J. Evol. Biol. (in press).

Kim, Y., and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155:** 1415–1427.

Kingman, J. F. C., 1982a The coalescent. Stochastic Processes and Their Applications **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

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

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

Kumar, S., K. Tamura, I. Jakobsen and M. Nei, 2000 *MEGA, Molecular Evolutionary Genetics Analysis*, version 2.0.

Labate, J. A., C. H. Biermann and W. F. Eanes, 1999 Nucleotide variation at the *runt* locus in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **16:** 724–731.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159–255.

Maddison, W. P., and D. R. Maddison, 1992 *MacClade: Analysis of Phylogeny and Character Evolution*. Version 3.0. Sinauer, Sunderland, MA.

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

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

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

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

Nurminsky, D., D. De Aguiar, C. D. Bustamante and D. L. Hartl, 2001 Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. Science **291:** 128–130.

O'Connell, P., and R. Rosbash, 1984 Sequence, structure and codon preference of the *Drosophila* ribosomal protein 49 gene. Nucleic Acids Res. **12:** 5495–5513.

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology. The Drosophila Model*. Oxford University Press, New York.

Roff, D. A., and P. Bentzen, 1989 The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol. Biol. Evol. **6:** 539–545.

Rozas, J., and M. Aguadé, 1993 Transfer of genetic information in the *rp49* region of *Drosophila subobscura* between different chromosomal gene arrangements. Proc. Natl. Acad. Sci. USA **90:** 8083–8087.

Rozas, J., and M. Aguadé, 1994 Gene conversion is involved in the transfer of genetic information between naturally occurring inversions of *Drosophila*. Proc. Natl. Acad. Sci. USA **91:** 11517–11521.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Rozas, J., C. Segarra, G. Ribó and M. Aguadé, 1999 Molecular population genetics of the *rp49* gene region in different chromosomal inversions of *Drosophila subobscura*. Genetics **151:** 189–202.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi *et al.*, 1988 Primer-directed enzymatic amplification of DNA with a thermostable polymerase. Science **239:** 487–491.

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

Singh, R. S., and A. D. Long, 1992 Geographic variation in *Drosophila*: from molecules to morphology and back. Trends Ecol. Evol. **7:** 340–345.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

Stephan, W., T. H. E. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.

Strobeck, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117:** 149–153.

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

True, J. R., J. M. Mercer and C. C. Laurie, 1996 Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics **142:** 507–523.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

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

Communicating editor: W. Stephan