Nucleotide Variation and Recombination Along the Fourth Chromosome in *Drosophila simulans*

Wen Wang,*,†,[1] Kevin Thornton,‡,[1] J. J. Emerson* and Manyuan Long*,‡,[2]

**Department of Ecology and Evolution, ‡The Committee on Genetics, University of Chicago, Chicago, Illinois 60637 and †CAS-Max Planck Junior Scientist Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, People's Republic of China*

Manuscript received December 10, 2003
Accepted for publication December 23, 2003

ABSTRACT

The fourth chromosome of *Drosophila melanogaster* and its sister species are believed to be nonrecombining and have been a model system for testing predictions of the effects of selection on linked, neutral variation. We recently examined nucleotide variation along the chromosome of *D. melanogaster* and revealed that a low average level of recombination could be associated with considerably high levels of nucleotide variation. In this report, we further investigate the variation along the fourth chromosome of *D. simulans*. We sequenced 12 gene regions evenly distributed along the fourth chromosome for a worldwide collection of 11 isofemale lines and 5 gene regions in a local population of 10 isofemale lines from South America. In contrast to predictions for regions of very low recombination, these data reveal that the variation levels in many gene regions, including an intron region of the *ci* gene, vary considerably along the fourth chromosome. Nucleotide diversity ranged from 0.0010 to 0.0074 in 9 gene regions interspersed with several regions of greatly reduced variation. Tests of recombination indicate that the recombination level is not as low as previously thought, likely an order of magnitude higher than that in *D. melanogaster*. Finally, estimates of the recombination parameters are shown to support a crossover-plus-conversion model.

INVESTIGATION of evolutionary forces at the sequence level often relies on the understanding of the relationship between recombination and variation. Of particular interest is the effect of natural selection on levels of neutral variation at linked sites. Both background selection (Charlesworth *et al.* 1993) and recurrent selective sweeps (Berry *et al.* 1991) reduce levels of neutral variation at linked sites. A significant positive correlation between the rate of recombination and nucleotide variation has been observed in Drosophila (Aguade *et al.* 1989a,b, 1994; Stephan and Langley 1989; Berry *et al.* 1991; Begun and Aquadro 1992; Aquadro *et al.* 1994), Mus (Nachman 1997), humans (Nachman 2001), and Aegilops (Dvorak *et al.* 1998). These empirical studies, bolstered by the theoretical interpretations of hitchhiking effects of selection, lead us to expect extremely low levels of variation in genomic regions of low recombination.

Two evolutionary genetic models have been proposed to account for the correlation. The classic model, selective sweep via the hitchhiking effect (Maynard-Smith and Haigh 1974; Kaplan *et al.* 1989; Hudson 1990), posits that recent directional selection purges variation on selected and linked loci, allowing current alleles little time to accumulate mutations. The low level of polymorphism, then, is a result of the young age of the alleles that originated after the selected allele dominated the population. An alternative model, background selection, was subsequently proposed to account for the same phenomenon (Charlesworth *et al.* 1993, 1995, 1997; Hudson and Kaplan 1994; Charlesworth 1996; Nordborg *et al.* 1996). This model posits that the negative selection against deleterious mutation will decrease the size of effective population by a fraction, depending on both the recombination rate and the mutation rate to deleterious alleles. The number of mutation-free alleles will significantly decrease in the regions of low recombination that contain high numbers of deleterious mutations. Consequently, the neutral variation, as a function of population size N (*e.g.*, nucleotide diversity $\pi = 4N\mu$), will be reduced. In these genetic models the effect of selection is so strong that in regions of extremely low recombination variation is severely reduced. Even though the fourth chromosome was previously thought to be virtually free of recombination, either selective sweep or background selection can reduce variation.

Genomic regions of low recombination have become a hunting ground for detection of selection, even though it is unclear that the selection is positive or negative. One such region is the small fourth chromosome of Drosophila. Although the chromosome contains ~5 Mbp, only one-quarter of it in the right arm of the chromosome is euchromatic, encoding only ~80 genes (Adams *et al.* 2000; Locke *et al.* 2000; Sun *et al.* 2000).

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Department of Ecology and Evolution, University of Chicago, 1101 East 57th St., Chicago, IL 60637.
E-mail: mlong@midway.uchicago.edu

Genetic analyses of the fourth chromosome in *Drosophila melanogaster* conducted by Bridges (1935) and Patterson and Muller (1930; for later reviews, see Hochman 1976; Ashburner 1989) failed to observe any crossovers in thousands of normal flies, except under special laboratory conditions like heat shock or artificially introduced interchromosomal effects. Here, the chromosome was concluded to be essentially nonrecombining; a conclusion with nontrivial consequences for subsequent evolutionary genetic analysis. The assumption of no recombination implies that the whole chromosome would evolve as a single unit. Qualitatively, both selective sweeps and background selection can explain low levels of variation.

The first direct observation of nucleotide variation on the fourth chromosome was an analysis of polymorphism at the *cubitus interruptus* (*ci*) gene in *D. melanogaster* and *simulans* (Berry *et al.* 1991). In these 1-kb exon sequences, no polymorphism was observed in 10 alleles from *D. melanogaster* and only one polymorphism was observed in 9 alleles from *D. simulans*. When the analysis of *ci* was extended to the other two sibling species, *D. sechellia* and *D. mauritiana*, similarly low levels of variation ($\theta = 0$–0.0003) were observed (Hilton *et al.* 1994). These observations, albeit on a single gene, suggest that the entire chromosome lacks variation and that different chromosome regions should not exhibit great differences in the level of variation. Especially in light of early conclusions of Bridges and Muller (Bridges 1935) that the chromosome is nonrecombining, this implicated selective sweeps as an important factor in reducing variation on the fourth chromosome. The conclusion that the fourth chromosomes in all three sibling species of *D. melanogaster* are nonrecombining was not tested by genetic experiments and inferred only on the basis of the early experiments on the *D. melanogaster* fourth chromosome. Since then, the Drosophila fourth chromosome has become a standard example of how directional selection could purge the variation in a whole chromosome. Meanwhile, theoretical models of background selection were also able to account for the deficit of variation in the chromosome (Charlesworth *et al.* 1993, 1995, 1997; Hudson and Kaplan 1994; Charlesworth 1996). Carr *et al.* (2001) reported several retrotransposon insertions with high frequencies in natural populations of *D. melanogaster*. Jensen *et al.* (2002) reported a low level of polymorphism in the *ankyrin* (*ank*) gene in *D. melanogaster* and *D. simulans* with frequency spectra that are unlikely to be explained by selective sweep. Furthermore, Jensen *et al.* (2002) and Sheldahl *et al.* (2003) found evidence for recombination in a few loci of the fourth chromosome.

The conclusion that the fourth chromosome is essentially nonrecombining was questioned in general when we analyzed the polymorphism data of *Sphinx* (*Spx*). A young gene that recently originated in the single lineage of *D. melanogaster*, *Spx* and its flanking region (Wang *et al.* 2002a) revealed a considerably high level of variation with a peculiar haplotype structure. In a chromosome-wide survey of polymorphisms in the *D. melanogaster* worldwide collections of both isofemale lines and local populations (Wang *et al.* 2002b), we found that the chromosome is variable in several regions that possess long chromosome domains (20–30% of the euchromatic arm) with highly variable dimorphic haplotype structures. This suggests that the chromosome does not evolve as a single unit. We also identified possible recombination events, using the four-gamete method (Hudson and Kaplan 1985), although the levels of recombination appeared two orders of magnitude lower than that in the other three chromosomes. Evolutionary analyses suggest that some form of positive selection, possibly the balancing selection, might govern the dimorphism domain. In this report, we extend the analysis to the fourth chromosome of *D. simulans* and document several findings that substantially differ from what we have previously known about the chromosome in either this species or *D. melanogaster*.

MATERIALS AND METHODS

Drosophila strains: Eleven *D. simulans* isofemale lines collected from different geographic areas were investigated for sequence variation (Figure 1) and have been kept in a lab for >15 years. DNA extracted from a single male of each stock served as the PCR template for amplification of all 12 regions surveyed. Ten isofemale lines of a local population from Quito, Ecuador (collected by W. Ballard in 2000) were also used in this study.

Gene regions and sequencing: We chose 12 gene regions to survey for nucleotide variation of the worldwide samples (Figure 1), which are evenly distributed along the chromosome. We also sequenced five gene regions for variation in the Ecuador population (Figure 2). These gene regions were PCR amplified using a high-fidelity DNA polymerase (Roche Molecular Biochemicals, Indianapolis) with the extension temperature at 68°, as recommended by the manufacturer (the primers are listed in Table 1). The PCR products were cleaned with the PCR product purification kit (QIAGEN, Valencia, CA) and sequencing was accomplished on an Applied Biosystems (Foster City, CA) 377 sequencer. Alignments were manually inspected and all polymorphisms were resequenced and carefully verified to avoid artifacts. All primers and fly strains are available upon request. No heterozygous nucleotide sites were found in worldwide-collected lines. In a few lines of the local population of Quito, some heterozygous sites were found, and one of two nucleotides in such sites was randomly chosen, although used in the analysis of variation only. Recombination and linkage disequilibrium were analyzed in the worldwide-collected lines, but not in the local population because of its heterozygous sites.

Gene order in the region between CG11153 and *Pho* genes: Locke *et al.* (2000) reported gene order in the region that differs from Adams *et al.* (2000). We performed fluorescence *in situ* hybridization (FISH) experiments to confirm their results in *D. melanogaster* and *D. simulans*. Our previous FISH experiments on *D. melanogaster* (Wang *et al.* 2002b) indicated that the gene order, between *Pho* and CG11153 genes, was erroneously inverted in the results of the Drosophila genome project (Adams *et al.* 2000), supporting a physical map of

TABLE 1

Sequenced gene regions and primers

Primer names	Primer sequences (5′ → 3′)	Sequenced gene regions
ciF	TGGCTCCATTTTCAGTTC	Intron 4 (partial) and 5, exon 5 and 6 (partial) of *ci* gene
ciR	ACACAAATGCACGGTAAGA	
ciF2	TAGCTAATTCGATCGGCGATT	Intron 1 of *ci* gene
Ciintron1R1	TAGTCGAGTTTCCCGACTA	
CrkF	GATTATGTACTTTGTGTAAGGT	Intron 2 of *Crk* gene
CrkR	GCTAACTTTTGTATCTTCTCT	
RadF2	GAGCCGTTGAATATCTCATT	Exon 2 (partial), 3, and 4 (partial); intron 2 and 3 of *Rad23* gene
RadR	ATTCGCGGTCAATCGGTTGAT	
bipintron1F	CATAGACATATGGAACACGGT	Intron 1 of *bip2* gene
bipintron1R	TCGGCTCTATCCGATTTGCTA	
CG1511F	ACATCCCGTTCATAGAGGTC	Partial exon 3 and 4, and intron 3 of CG1511 gene (*Ephrin*)
CG1511R	CTGGGATACGGATAAACAATG	
eyF	ACATTAGCGCTGGCACTGGAC	Partial exon 7 and 8, intron 7 of *eyeless* gene
eyR	TGGCGGGCATCAAAATACTCA	
CG11153F	GTTCCTTCCGGCCTTCCTACAGC	Partial exon 1 and intron of CG11153 gene
CG1153R	TTGCCCGTGATCCTCCAGAACAT	
uncF2	ATGCGGATATTGTGAAACTT	Exon 15 (partial), 16, 17, and 18 (partial); intron 15, 16 and 17 of *unc-13* gene
uncR2	GTGCGTAACGTAAACTCTGAA	
spxF1	AGATTGCCGCTGCTGGTA	Correspondent to the upstream region of the *sphinx* gene in *D. melanogaster* (Wang *et al.* 2002a,b)
spxR2	TCGAGATCGAAACATTTTCAG	
RfaF	TTGCCGTGCCGTTCCATACTG	Partial exon 4 of *RfaBp* gene
RfaR	CAACGTTACTCGGCCCAAGACATT	
phoF	AGGCAACCACTACGACTTACAC	Partial exon 1 and 5, exon 4, intron 3 and 4 of gene *pho*
phoR	TTCCACATTCTGCACAAACAT	

Locke *et al.* (2000). Recently, Locke's group reported that the whole right arm of the fourth chromosome of *D. simulans* is inverted compared to that of *D. melanogaster* (Podemski *et al.* 2001). To determine gene order and relative distance between genes, we did further FISH on the polytene chromosomes of *D. melanogaster* and *D. simulans*, using probes of *pho*, CG11153, and *spx* genes. PCR amplified DNA fragments of these genes, which were labeled with digoxigenin (DIG) or biotin (purchased from Roche Molecular Biochemicals). FISH was conducted according to the method of Wang *et al.* (2000).

Evolutionary genetic analysis: The major goal of the analysis is to test various statistical expectations on the basis of two conventional beliefs: (1) The *D. simulans* fourth chromosome lacks variation and (2) the *D. simulans* fourth chromosome, similar to the chromosome of *D. melanogaster*, is nonrecombining (Berry *et al.* 1991; True *et al.* 1996). We therefore conducted the following analyses:

1. *Estimation of summary statistics*: We first generated polymorphism data, including polymorphism created by both nucleotide substitutions and insertion-deletion (indel) changes. We then calculated the following statistics from the sequences and polymorphism data: Numbers of synonymous sites (silent sites), numbers of nonsynonymous sites (replacement sites), nucleotide diversity π (Nei 1987) in total sites and synonymous sites, Watterson's estimator θ (Watterson 1975) in total sites and synonymous sites, number of substitutions per synonymous site (K_S), and number of substitutions per nonsynonymous site (K_A). These statistics were computed by using DnaSP (Rozas and Rozas 1999).
2. *Statistical comparison between observed and expected values of variation*: Under the assumption of neutrality, a probability distribution of polymorphism for an expected variation parameter $\theta = 4N\mu$ was calculated following the recursion equation

$$P_n(s) = \sum_{i=0}^{s} P_{n-1}(s - i)\,Q_n(i)$$

(Hudson 1990), where N and μ are effective population size and mutation rate per nucleotide site, respectively, and

$$Q_n(i) = \left(\frac{l\theta}{l\theta + n - 1}\right)^i \frac{n - 1}{l\theta + n - 1},$$

in which l is length of DNA sequence and n is number of alleles. Two values of θ were used in the calculation: $\theta = 0.0002$ for the *ci* gene and $\theta = 0.0005$ for the *ank* gene (Jensen *et al.* 2002).
3. *Test of homogeneity across gene regions*: We tested a null hypothesis that the variation in different regions is similar, on the basis of the evolutionary hypothesis that the fourth chromosome evolved as a single unit. This was conducted by calculating a goodness-of-fit statistic in which the observed and the expected number of segregating sites at the ith region (S_i; Kreitman and Hudson 1991),

$$X^2 = \sum_{i=1}^{k} \frac{(\mathrm{Obs}(S_i) - \mathrm{Exp}(S_i))^2}{\mathrm{Var}(S_i)},$$

which follows the χ^2 distribution with $k - 1$ d.f. (k = number of gene regions). This statistic was shown to be insensitive to assumptions of recombination [see p. 572 of Kreitman and Hudson (1991)]. In this equation, the expectation $\mathrm{Exp}(S_i)$ and variance $\mathrm{Var}(S_i)$ of the segregating sites in the ith gene regions are

$$\mathrm{Exp}(S_i) = l_i\theta\sum_{j=1}^{n-1}\frac{1}{j}$$

and

$$\mathrm{Var}(S_i) = l_i\theta\sum_{j=1}^{n-1}\frac{1}{j} + (l_i\theta)^2\sum_{j=1}^{n-1}\frac{1}{j^2},$$

where θ was computed as nucleotide diversity using the total number of polymorphic sites of all k gene regions, l_i as the length of the ith gene region, and n as the sample size for polymorphism survey.

4. *Tests based on the site frequency spectrum*: By taking advantage of a large number of polymorphic sites [56 single-nucleotide polymorphisms (SNPs) out of 82 polymorphic sites including indels] pooled from all gene regions, we used Tajima's D (Tajima 1989) to test the prediction that selective sweeps on the fourth chromosome reduce variation such that the polymorphism spectrum is skewed toward rare variants. Braverman *et al.* (1995) and Simonsen *et al.* (1995) showed in simulation studies that such deviation from the spectra predicted from neutrality can distinguish either directional selection or demographic effects from background selection, although both types of forces can reduce the variation in regions of low recombination.

Recombination estimation: Given that our sequence data are from noncontinuous gene regions, we used two methods to detect recombination.

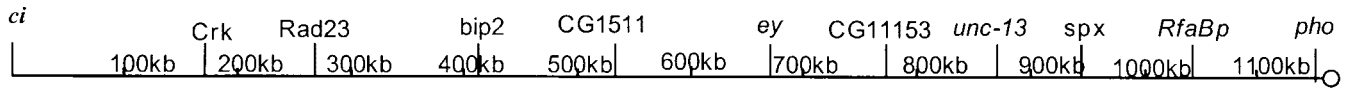
1. First, we used the four-gamete method (Hudson and Kaplan 1985) to compute the minimum number of possible recombination events, R_m. Because R_m is an increasing function of a sample's size and is well correlated with the change of $\rho = 4Nr$ (r is the rate of recombination caused by crossing over), as shown in the simulation (Hudson and Kaplan 1985), we computed the R_m density as the R_m value per kilobase of sequences per chromosome and compared with the R_m density in different regions. Because the value of R_m is dependent on the sample size, we made the comparison with other samples that have similar sample sizes with comparable levels of variation. For instance, Kreitman (1983) sampled 11 *Adh* alleles; the *D. melanogaster* fourth chromosome variation data are from 10 chromosomes. Both data sets have similar sample sizes compared to that of the *D. simulans* we used in this study (11). We show that the levels of nucleotide variation in the *D. simulans* fourth chromosome are not as low as previously thought; hence the comparisons among the different regions and samples under study provide insight.
2. In addition, we investigated estimates of the population recombination parameter ρ and the rate of gene conversion f ($f = g/r$, where g is the probability of a conversion at a nucleotide site per generation), using the composite-likelihood method of Hudson (2001; see also Frisse *et al.* 2001). The individual *D. simulans* fourth chromosome loci that we sequenced from the worldwide sample were assembled into a single polymorphism table for the entire chromosome. The positions of the segregating sites were estimated from the *D. melanogaster* fourth chromosome sequence (Adams *et al.* 2000), assuming that the distances between loci are the same in the two species. We note that the inversion on this chromosome between *D. melanogaster* and *D. simulans* involves the entire euchromatic region so that the relative gene order within the chromosome is conserved (Podemski *et al.* 2001).

The three parameters of interest are: ρ, the product of four times the effective population size times the per-generation recombination rate per site; f, the rate of gene conversion relative to crossing over; and t, the mean tract length of a conversion event. The model of gene conversion under which parameters are estimated is that of Wiuf and Hein (2000), which is a simple model of unbiased gene conversion with symmetric heteroduplexes and a geometric distribution of tract lengths. We obtained composite-likelihood estimates of ρ both without gene conversion and with gene conversion, over a range of tract lengths. In all cases, maximum composite-likelihood estimates (MCLE) were obtained by evaluating the composite likelihood of the data over a grid of parameter values, using software provided by Richard Hudson (http://home.uchicago.edu/rhudson1). We first estimated ρ assuming no gene conversion. Then, to estimate ρ and gene conversion parameters, we evaluated the composite-likelihood statistic on a grid of 500 values of ρ ranging from 0.0001 to 0.0011848, which is the estimate of ρ obtained assuming no conversion. The grid of f was 250 values in the range 0.001 to 500. For ρ and f, the distances between points on the grid were geometric rather than uniform to ensure coverage of low values of the two parameters. Eleven values for t were used (50 bp and 100–1000 bp in increments of 100 bp).

RESULTS

Levels of variation: Figure 1 lists nucleotide polymorphism in all 12 gene regions on the fourth chromosome in 11 worldwide lines, revealing unexpected levels of variation. The average levels of variation over all 12 gene regions are $\pi = 0.0024$ and $\theta = 0.0023$. These levels are significantly higher than the reported variation in *ci* and *ank* ($P < 0.0001$), although they still appear to be lower than the average variation in other chromosome regions (Begun and Whitley 2000; Andolfatto 2001). A more interesting result is that this average is accompanied with a great heterogeneity in the variation levels among 12 gene regions. Several regions do not show reduced variation, contrary to previous surveys in *ci* and *ank* in *D. simulans*. The nucleotide diversity at silent sites (introns and synonymous sites) in 9 gene regions ranges from 0.0019 to 0.0074 and only four genes, *Pho*, *RfaBP*, *unc-13*, and the exon region of *ci*, have a low level of variation ($\pi = 0$–0.00086 at silent sites). Table 2 lists the probabilities under a null hypothesis that the variation levels for all these regions are from the same distributions of polymorphism for $\theta = 0.0002$ (the *ci* gene; Berry *et al.* 1991) and for $\theta = 0.0005$ (*ank*; Jensen *et al.* 2002). These tests showed that the levels of nucleotide polymorphism for 9 of the 11 gene regions are significantly higher than those of the exon regions of *ci*, and for 8 of the 11 genes they are significantly higher than those of *ank*. It should be noted that the variation would be even higher than the expectation if the indel polymorphisms in the four gene regions are considered [the average indel polymorphism in 7 gene regions (0.74×10^{-3}) is threefold higher than the indel polymorphism (0.22×10^{-3}) in *ank* (Jensen *et al.* 2002)]. The distribution of the variation overlaps with the range of variation in other chromosomes that do not include genes in regions of low crossing over per physical length (Begun and Whitley 2000).

a



b

Gene	ci-exon	ci-intron	Crk	Rad23	bip2	CG1862	ey	CG11153	unc-13	spx (CG11091)	RfaBp	pho
Sequenced length (bp)	835	1085	1066	784	1063	967	1109	966	931	1172	1110	1027
		1111111111					1					
		1112 4666 78888 899990000000001	1334458	1337	1135 667888 9	23467	333570	223	8	2245 699	666	9
		3 60158 8779 80012 823450111123580	7 3890632	2463	41626 346788 6	81312	039184	1169	6	28866 437	003	3
		9 65805 6782 77871 274382457806082	8 3886516	8344	63472 162268 9	05390	533115	4856	2	14537 513	493	8
Cook island	0	d1GGAGd2GAGd3AGCGd4TTAAATACCTTTATC	i5A-TAAGA	GACT	TCC-i6CAAAAi7T	GCAAA	TAGCCA	CT-T	G	TAACi8ATC	CCG	A
Impala,Kenya	0	. ..... A... ....i4G.XXXXXXXXXXXXX	d5-A-.TTC	....	..... .G.... A	.G.CC	..C..G	....	.	AGGA. .A.	TGA	G
Lima,Peru	0	. ...	d5-A-GT..	TCT.	..ATd6AG.... T	.....	...TT.	...A	.	A..A. ...	...	.
Morro Bay,California	0	. ...	d5-A-.TT.	....	....d6AGTTGd7T	.....	.....	....	A	A..Ad8...	..A	.
New Guinea	0	. ...	d5-A.GT..	....	..A.. .G.... T	..TCA	.....	G...	.	....	..A	A
Sechelle island	0	. ...	d5-A-GT..	TCT.	..ATd6AG.... T	.....	...TT.	...A	.	A..A. ...	...	.
Tamazunchale,Mexico	0	. ...	. .....	....	AAA.. .GTTGd7T	.....	...TT.	....	A	A..Ad8...	..A	.
Trinidad	0	. ...	. .....	....	..ATd6AG.... T	.....	AT....	....	.	AGG.. .AT	..A	.
W66,Reunion island	0	. ATG.. A... .T... .....TG....A	d5-A-.TT.	...-	..... .GTTGd7T	.G.CC	NA	.AAA	.	A..A. G..	.GA	.
W74,Reunion island	0	. ...Ai2A... G.TC. GG.GCAGTTGGCTGC	d5-A-.TT.	...-	..... .GTTGd7T	.G.CC	.....	.AAA	.	A..A. G..	.GA	.
W75,Reunion island	0	i1ATG.. ATAi3..... ..GG.AGTTGGGTGC	d5-A-.TT.	....	....d6AGTTGd7T	-.TCC	.....	...A	.	A..A. ...	..A	.
4 gametes sites												
Outgroup state			d5AXXXXXA	GACX	TCA-X XXXXXX X	TCTCA	TXGCCG	CT-A	G	AGTAX AGT	CCA	A

c

Silent sites	254.1	1085	1053	600.3	1054	740.9	805.7	451.9	380	1172	240.83	309.0
replacement sites	579.6	-	-	182.6	-	225.0	302.3	513.0	549.5	-	865.8	717.5
Single nucleotide polymorphism:												
π (total sites, x1000	0	7.41	3.25	1.67	4.73	2.07	1.56	1.43	0.35	2.21	0.95	0.19
π (silent sites, x1000)	0	7.41	3.25	2.18	4.73	2.01	1.90	1.85	0.86	2.21	0	0.65
θ (total sites, x1000)	0	9.24	2.59	1.74	3.89	1.77	1.91	1.41	0.37	2.33	0.92	0.34
θ (silent sites, x1000)	0	9.24	2.59	2.27	3.89	1.84	2.19	2.27	0.90	2.33	0	1.14
ndel polymorphism:												
π (x1000)	0	0.84	1.70	0.42	1.44	0.19	0	0.34	0	0.28	0	0
θ (x1000)	0	1.57	1.28	0.43	0.96	0.35	0	0.35	0	0.29	0	0
Ks (x100)	11.08	9.24	14.0	12.85	12.28	17.00	13.49	11.83	8.76	17.8	16.13	15.79
Ka (x100)	1.57	-	-	3.94	-	3.17	1.67	3.29	0	-	2.58	4.67
Tajima's D	-	-1.0	1.04	-0.15	0.94	0.66	-0.76	0.04	-0.10	-0.23	0.10	-1.11

FIGURE 1.—Polymorphisms in the surveyed genes on the *Drosophila simulans* fourth chromosome. (a) The sequenced gene regions and their approximate positions on the *D. simulans* fourth chromosome, with the *ci* gene adjacent to the telomere and the *pho* gene close to the centromere. All the sequences are orientated from the 5′ to 3′ direction. The circle indicates the centromere. (b) Polymorphisms in 12 gene regions. Positions start from the first nucleotide in the forward PCR primers of each gene region (Table 1). The sequenced length is given beneath the name of each gene region. i1, CGAA; i2, TA; i3, GTTAAGTTAA; i4, 49 bp; i5, TAAACTGTAAAATT; d5, ATATAGT—; i6, GAATGCT; i7, GCAG; i8, TAA. –, one nucleotide deletion; X, the absence of nucleotide caused by a long deletion; 0, in the *ci*-exon region, no polymorphic sites found in its sequence of 835 nucleotides. The four-gamete sites (R_m) are labeled by adjacent double arrows. And the outgroup states are derived by comparison with the *D. melanogaster* sequences (ADAMS *et al.* 2000). (c) Summary statistics of polymorphisms. Total length sequenced is 12,091 bp, with Tajima's $D = -0.19169$. For all Tajima's *D*'s, $P > 0.10$. –, absence of information. The values of divergence between *D. melanogaster* and *D. simulans* in the *ci* intron (9.24%) and *bip2* intron (12.28%) are calculated in the alignable regions of these two intron sequences.

TABLE 2

Comparison between observed and expected polymorphisms of the *D. simulans* fourth chromosome

Regions	Length (bp)	S_{obs}	$P(S \geq S_{obs} \mid \theta = 0.0002)$	$P(S \geq S_{obs} \mid \theta = 0.0005)$
Ci intron	1085	32	2.0×10^{-16}	1.9×10^{-13}
Crk	1066	8	7.1×10^{-3}	0.0947
Rad23	784	4	2.6×10^{-3}	0.0433
Bip2	1063	12	2.7×10^{-7}	2.5×10^{-4}
CG1511	967	5	9.5×10^{-4}	0.0283
Ey	1109	6	3.2×10^{-4}	0.0171
CG11153	966	4	5.2×10^{-3}	0.0745
Unc-13	931	1	0.4100	0.7066
CG11091	1172	8	8.4×10^{-5}	0.0083
RfaBp	1110	3	0.0364	0.2311
Pho	1027	1	0.4356	0.7382

The number of alleles = 11; S_{obs} is the number of segregating sites including indel sites; CG11091 is homologous to the flanking region of *sphinx* in *D. melanogaster*. The genomic average θ is 0.019 for noncoding regions and 0.008 for coding regions (Powell 1997). Calculation of the probabilities in the three statistical tests conservatively excluded indel sites.

Variation in diversity levels: While a large proportion of gene regions are more variable than previously expected, several regions show low levels of variation. An exon region of the *ci* gene was previously found to have little variation: Of nine alleles for a 958-bp region, there was only one polymorphic site, a singleton (Berry *et al.* 1991). We confirmed this result in a worldwide sample, finding no polymorphism in the region. The other two regions, *unc-13* and *pho*, also show very low levels of variation. The substitutions at synonymous sites between *D. simulans* and *D. melanogaster* in the three genes are 0.11 (*ci* exon), 0.09 (*unc-13*), and 0.16 (*pho*). These are in the normal range of synonymous divergence between the two species (Powell 1997, Figure 10-5), revealing no significant reduction in mutation rates in these regions. Furthermore, tests of homogeneity ($\chi^2 = 39.58$, $P = 4.2 \times 10^{-5}$) show that the nucleotide variation levels in different regions are not homogenous. These analyses indicate that the chromosome does not evolve as a single unit; rather, different regions have different histories, as we found in the *D. melanogaster* fourth chromosome. These conclusions derived from the worldwide sample are consistent with the five genes we analyzed from the Ecuadorian population (Figure 2). Once again, the local population revealed variable and relatively high levels of variation and a high proportion of heterozygosity at a number of sites, which is a result of a higher level of variation.

Unexpectedly high variation in the *ci* intron: In contrast to the low levels of variation in the exon regions of this gene, as detected in previous work (Berry *et al.* 1991; Hilton *et al.* 1994), we found a high level of polymorphisms in the intron region of this gene. There are 32 segregating sites with the nucleotide diversity of $\theta = 0.0087$ (4 indels included). This level of variation is not significantly different from the average level of the genome ($\theta = 0.019$, $P = 0.1231$), in stark contrast to previous observations of exons in the same gene. The intron region is located toward centromere, whereas the exon with the low-variation region is toward the telomere. The sharp boundary between the exon and the intron reveals distinct histories of evolution in the two regions of the same gene.

Recombination and linkage equilibrium: The four-gamete analysis (Hudson and Kaplan 1985) reveals 10 exchanges in the history of the fourth chromosome samples here, as shown by the line-arrow signs in Figure 1b. Between-region recombination events often involve long tracts of haplotype structure. For example, the *crk-bip2* contains some haplotypes consisting of eight polymorphisms inside a 250-kb region. The hypothesis of parallel mutations may not explain such a highly organized polymorphism structure, suggesting recombination between two considerably diverged ancestral alleles. The hypothesis of gene conversions may have greater difficulty explaining this long tract of haplotype structure than the crossover hypothesis, because, in general, conversion tracts have been shown to be ~350 nucleotides in length (Hilliker *et al.* 1994; Langley *et al.* 2000).

Intralocus recombination events reflect rates of recombination in a given stretch of DNA better than recombination events detected between genes where only one recombination event is detectable although many additional recombination events are possible. Thus, we incorporated only six within-gene recombination events to calculate the R_m density 6/12.1 kb/11 chromosomes = 0.0451 R_m/kb/chromosome. This is 27.0% of the estimate of the moderately recombining *Adh* gene of *D. melanogaster* (0.1673 R_m/kb/chromosome; Kreitman 1983), which is 9 times the estimate of the *D. melanogaster* fourth chromosome (0.005 R_m/kb/chromosome; Wang

Loci	CG1862	ey	CG11093	plex A	pho
Length(bp)	987	1090	787	1097	1065
Positions		1			N
	1	1 7 8 0	7	6	O
	7	6 1 7 0 0	5 8	2	N
	2	1 6 3 1 9	3 2 2	1	E
EC01	G	T C A T C	C C A	G	0
EC02	.	. T T A .	G . .	A	0
EC03	.	. . . . .	. . .	.	0
EC04	.	. . . . .	. . .	.	0
EC05	A	C	. . T	A	0
EC06	.	C	. T .	A	0
EC07	.	. . . . .	. . .	A	0
EC08	A	C . T A .	. T .	A	0
EC09	A	C . . A .	. T .	A	0
EC10	A	C . T . A	. . T	A	0
θ(total sites, x 1000)	0.4	1.6	1.3	0.3	0.0
π(total sites, x 1000)	0.5	1.7	1.3	0.4	0.0

FIGURE 2.—Nucleotide variation in the fourth chromosome in the Ecuador population of *D. simulans*.

TABLE 3

Estimates of recombination parameters

T	ρ (= $4Nr$)	f	ln L
No conversion	0.011848	NA	−5466.821254
50	0.000100	265.656745	−5438.753087
100	0.000100	148.784990	−5437.376467
200	0.000100	92.591742	−5436.992478
300	0.000100	74.993312	−5437.373304
400	0.000100	64.026571	−5437.928844
500	0.000100	54.663565	−5438.502330
600	0.000100	49.195241	−5439.064229
700	0.000100	46.669771	−5439.631125
800	0.000100	44.273947	−5440.193060
900	0.000100	42.001114	−5440.728962
1000	0.000100	148.784990	−5437.376467

NA, not applicable.

et al. 2002b). This result suggests a recombination rate much higher than previously assumed on the basis of the genetic analysis of the *D. melanogaster* fourth chromosome. We estimated $4Nr = 0.01185$ per base for *D. simulans*, assuming no gene conversion, using the method of HUDSON (2001). We also estimated $4Nr = 0.00016$ per base for the *D. melanogaster* fourth chromosome, using the data of WANG *et al.* (2002b). It should be noted that this latter data set contains 50% of each of the two haplotypes in a dimorphic region, slightly different from the worldwide sample from which we previously estimated frequencies of 40 and 60%. However, the following analysis shows a revealing comparison. Thus, it appears that the recombination rate of *D. simulans* is 74 times higher, assuming that effective population sizes of the two species are similar. Furthermore, the estimated level of recombination in the *D. simulans* fourth chromosome means that a ratio of $4Nr/4N\mu = r/\mu = \sim 7$, higher than that of most previously observed *D. simulans* genes (ANDOLFATTO and PRZEWORSKI 2000). Given that the recombination rate in chromosomes X, 2, and 3 of *D. simulans* is only slightly (30%) higher than that in *D. melanogaster* (TRUE *et al* 1996), more than one order of difference in magnitude between the fourth chromosomes of *D. simulans* and *D. melanogaster* is unusual.

We attempted to estimate the population recombination and gene conversion rates from the data, using a composite-likelihood procedure (HUDSON 2001). For a model with no gene conversion, we estimate ρ to be 0.011848 (ln $L = -5466.821254$). With gene conversion, we always estimated ρ to be the minimum value used on the grid ($\rho = 0.0001$). Because this result suggests we did not fully explore the likelihood surface, we also evaluated the likelihood of the data for $\rho < 0.0001$ and $\rho = 0$. For values of $\rho < 0.0001$, the likelihood of the data decreases rapidly and the data are not compatible with a no-recombination model (data not shown). The effect of varying t is shown in Table 3, which suggests that the data are compatible with a range of ρ, conversion rates, and mean tract lengths. This is not surprising because ρ and f are confounded in the estimation procedure; with a fixed mean tract length, for a given set of values for ρ and f, another pair of values (with a higher ρ and lower f, or vice versa) can be found for which the probability of recombination (*i.e.*, crossover or conversion) between two adjacent sites is identical (see Equation 1 of FRISSE *et al.* 2001). However, Table 3 suggests that models with nonzero f and nonzero t are slightly more likely than models with no conversion. Given the expectation from laboratory experiments that the rate of crossing over is very low on the *D. melanogaster* fourth chromosome, it may not be unlikely that some recombination events are nonreciprocal exchanges (gene conversions). In general, our estimates of ρ are nearly an order of magnitude less than those obtained by WALL *et al.* (2002) from loci on the third chromosome of *D. simulans*.

We also estimated linkage disequilibrium using the correlation coefficient, r^2 (HARTL and CLARK 1997). Figure 3 reported the distribution of r^2 in the *D. simulans* fourth chromosomes. While most r^2 values are not high, a proportion of r^2 are significant at the levels of $P = 0.01$ and 0.05, suggesting some degrees of linkage disequilibrium among distant sites. Although the comparison among sites is not independent, some suggestive observations can be made. It is interesting to see a decay of r^2 for the polymorphism separated <1 kb apart (see Figure 3, inset). These results suggest a possibility of gene conversion in short distance.

Evolutionary forces: Our results show unexpected levels of DNA sequence polymorphism in genes on the fourth chromosome of *D. simulans* that were not predicted from previous reports or simple theoretical pre-

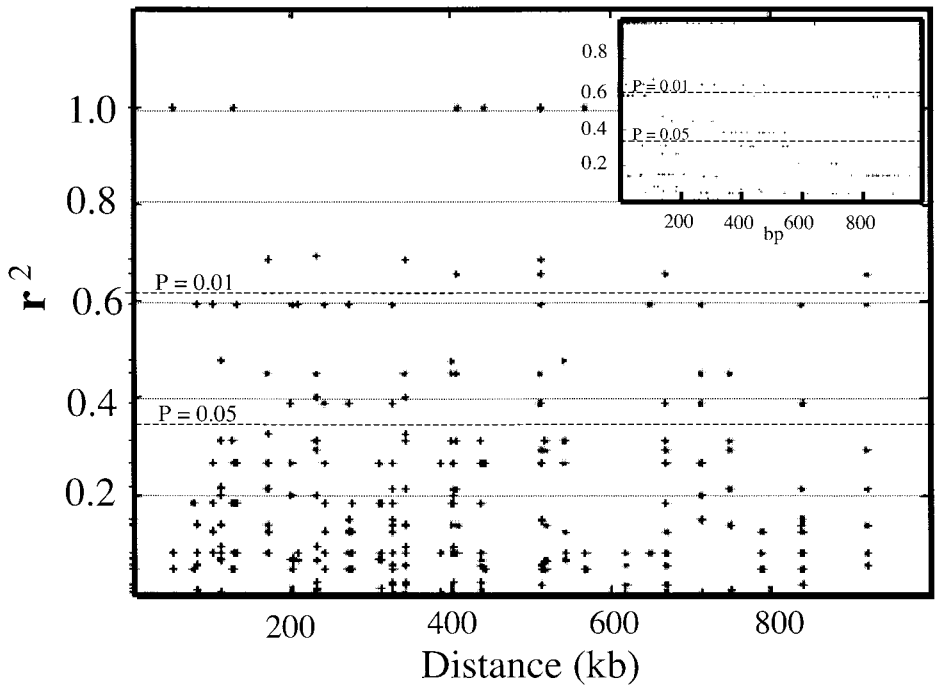


FIGURE 3.—Distribution of linkage disequilibrium over the nucleotide distances of gene pairs on the *D. simulans* fourth chromosome. The disequilibrium is measured as correlation coefficients (r^2). The inset is the distribution of r^2 for short distances. The dashed lines indicate the r^2 values for $P = 0.01$ and $P = 0.05$, which were determined by a chi-square distribution $\chi^2 = nr^2$, with d.f. = 1, where n is the number of alleles (HARTL and CLARK 1997).

dictions. Are there any detectable forces of evolution driving the evolution of the chromosome? We considered three possible forces hypothesized in previous studies: variable neutral mutation rates, recent selective sweep, and background selection.

Divergence between *D. melanogaster* and *D. simulans* at silent sites fell well within the range of normal values for these two species (POWELL 1997), suggesting that the neutral mutation rates of the fourth chromosome loci do not differ substantially from the genome average. We calculated Tajima's *D* for every gene region and pooled data over all gene regions (including the *ci*-intron region with the *ci*-exon region excluded; Figure 1). In 11 gene regions that contain polymorphisms, Tajima's *D* values range from −1.0 to 1.04, none of them significant at the 5% level. Tajima's *D* based on pooled data is −0.1917 with a probability of 0.10 under the assumption of neutrality. Thus, the hypothesis of a recent selective sweep is not supported in general, although this test does not exclude the possibility that the selective sweep could happen in some local regions, especially considering the limited detecting power with the sample size used in this study. On the other hand, we found that the variation in most gene regions is not as low as previously thought, but is an order of magnitude lower than the variation in the chromosome regions of a normal level of recombination (ANDOLFATTO 2001). This appears to be consistent with a lower than normal level of recombination in this chromosome. These two results are not unexpected from what background selection predicts: Regions of low recombination are expected to contain large regions of reduced variation.

Gene orders: Both ADAMS *et al.* (2000) and LOCKE *et al.* (2000) published the gene orders of the fourth chromosome in *D. melanogaster*, but with different gene orders in the region from CG11153 to *Pho*. We did FISH experiments to ascertain the order of the region, using several genes as probes (Figure 4), which shows that the gene order of LOCKE *et al.* (2000) is correct. We also mapped these probes on the strain used to sequence the *D. melanogaster* genome, which did not support the gene order of ADAMS *et al.* (2000) in the region. Recently, the same group investigated the gene orders on the fourth chromosomes in eight species of the *D. melanogaster* subgroup and concluded that all chromosomes are homosequential, *i.e.*, in intact gene orders (PODEMSKI *et al.* 2001). Interestingly, it was found that the entire right arm of the fourth chromosome in *D. simulans, D. mauritiana*, and *D. sechellia* was inverted, relative to the *D. melanogaster* fourth chromosome. Thus, the *ci* gene is adjacent to the centromere in *D. melanogaster*, but the *ci* gene in *D. simulans* is near the telomere. However, the patterns of polymorphisms in different gene regions of *D. simulans* and *D. melanogaster* (WANG *et al.* 2002b) seem independent of the inverted gene orders, suggesting that the fourth chromosomes in the two species evolved independently under no effect of gene orders. Finally, the physical length of the *D. simulans* fourth chromosome has not been measured. But the morphology and size of the chromosome seems similar to that of *D. melanogaster*; thus we used the gene orders and the distances between the genes in *D. melanogaster* as an approximate description of the *D. simulans* chromosome. The small actual differences in the total chromosomal length between the species should not change the estimate of $4Nr$ drastically.

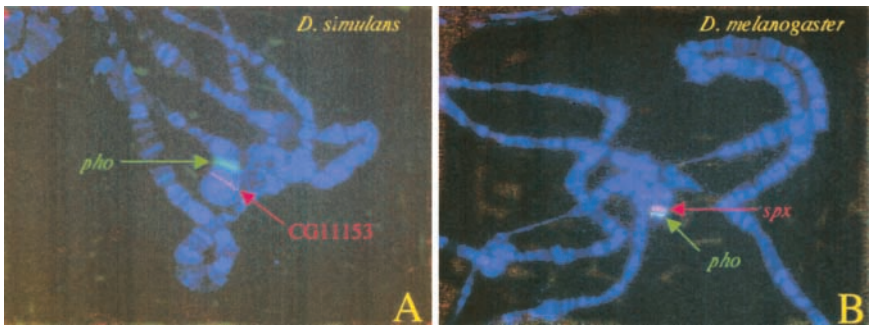


FIGURE 4.—FISH results showing the gene order between CG11153 and *Pho*. The arm of *D. simulans* chromosome 4 is inverted, compared to that of *D. melanogaster* (PODEMSKI *et al.* 2001). *Pho* (green) is close to the chromocenter in *D. simulans* (A) but adjacent to the telomere in *D. melonogaster* (B). Note that the positions of CG11153 and *Spx* are consistent with the depiction in Figure 1 drawn on the basis of the *D. melanogaster* map (ADAMS *et al.* 2000) with the correction of scaffold orientation between *Pho* and CG11153.

DISCUSSION

Variability in the *D. simulans* fourth chromosome: It is clear that the *D. simulans* fourth chromosome is not devoid of nucleotide variation, unlike the *ci* locus (BERRY *et al.* 1991) in which only one polymorphic site was detected from nine lines and *ank* (JENSEN *et al.* 2002) in which a 20-fold reduction in variation was observed. Instead, a high proportion of gene regions on the *D. simulans* fourth chromosome contain considerably high levels of variation, significantly higher than those typical of low-recombination regions. One region, the intron in the *ci* gene, even demonstrates levels of polymorphism close to the average variation of the *D. simulans* genome. However, a few regions of the fourth chromosome do indeed show low levels of variation, such as the exon region of *ci*, *pho*, *RfzBp*, and *unc-13*. Recently, JENSEN *et al.* (2002) reported that *ank*, a gene region between *ci* and *crk* (Figure 1), also demonstrates a low level of polymorphism, with a silent nucleotide diversity around 0.0005 in both *D. simulans* and *D. melanogaster*.

Recombination: The heterogeneity in levels of variation among different regions was shown to be highly significant. These data and analysis do not support the prevailing prediction that the entire fourth chromosome evolved as a single unit. Instead, they indicate that different gene regions have different evolutionary trajectories. To be consistent with this observation, recombination must be invoked.

The analysis of R_m revealed that the chromosome is recombining and the rate of recombination is not so low as previously assumed (BERRY *et al.* 1991; HILTON *et al.* 1994). Further estimation of recombination parameters using the likelihood method indicates that the level of recombination in the *D. simulans* fourth chromosome is more than one order of magnitude higher than that in *D. melanogaster*. We note that an equal haplotype sample in *D. melanogaster* (WANG *et al.* 2002b) might slightly bias the recombination ρ. However, the rate of recombination detected in the *D. simulans* fourth chromosome is too high to be explained by the level of recombination predicted from comparisons to *D. melanogaster* (TRUE *et al.* 1996), which revealed ~30% higher rates of recombination in the *D. simulans* genome, excluding chromosome 4. Thus, the previous assumption that the fourth chromosomes in the two species are similar in the level of recombination is incorrect. Finally, the level of recombination in the *D. simulans* fourth chromosome is also consistent with the observed levels of variation in the fourth chromosomal genes.

The mechanisms of meiotic recombination generally include gene conversion and crossing over. BERRY and BARBADILLA (1999) revealed the important role of gene conversion in the variation in Drosophila natural populations. LANGLEY *et al.* (2000) further inferred that there was high level of gene conversion in the *su(s)* and *su(w(a))* regions of low crossing over in the *D. melanogaster* X chromosome. Our analysis (Table 3) supports crossover-plus-conversion models. We can further ask whether or not the gene conversion alone can account for the detected recombination in the chromosome. Indeed, a model invoking only gene conversion at first seems plausible, because historically BRIDGES (1935) and others (*e.g.*, PATTERSON and MULLER 1930; HOCHMAN 1976) already concluded that there was no crossing over in *D. melanogaster*. Genetic analyses, as reviewed by HAWLEY *et al.* (1993), generally indicate no crossing over in normal laboratory conditions.

Nevertheless, crossing over in natural populations can be hypothesized to explain a part of the detected recombination. Several observations suggest the possibility of reciprocal exchanges. One difference between crossing over and conversion is that the former can create large regions or even chromosome-wide recombination while the gene conversion usually involves a short conversion tract. The linkage relationship between the sites outside conversion tracts does not change. A nonrecombining chromosome would show chromosome-wide linkage disequilibrium. In our linkage disequilibrium analysis across the chromosome, we observed some linkage disequilibrium among distant sites and a decay of the disequilibrium among the sites of short distance (Figure 3), suggesting a role of gene conversion. However, Figure 1 reveals that some sites displaying all four gametes showed long recombination tracts. The lengths of these tracts, >100 kb, are difficult to explain as the tracts of gene conversions that are more often several hundred nucleotide long (HILLIKER *et al.* 1994; LANGLEY *et al.* 2000) and may be a signature of crossing over. It has

been shown that some conditions, *e.g.*, heat shock and interchromosomal effects, could lead to crossing over in *D. melanogaster* (Grell 1971; Hochman 1976; Sandler and Szauter 1978). It is conceivable that some of the similar conditions may exist in nature, which would trigger crossing over in natural populations of the two Drosophila species.

Evolutionary forces: In above analyses of evolution of the fourth chromosome, the hypothesis of the neutrality that assumes variable mutation rates was excluded. The synonymous substitution rates of most gene regions (Figure 1) are from 0.0111 to 0.0178 in a normal level of between-species divergence (Powell 1997). This small range of variation in K_S is not consistent with large heterogeneity of the nucleotide diversity (π values for synonymous sites in the 12 gene regions are from 0.0000 to 0.0074, Figure 1). For example, the divergence levels in silent sites between *D. melanogaster and D. simulans* are 0.1110 in the *ci*-exon region and 0.0924 in an alignable region of *ci* intron of 277 nucleotides. However, *ci* exon has 0 polymorphisms in both species [in Berry *et al.* (1991), only 1 polymorphism was observed in nine alleles] whereas the alignable *ci*-intron region contains 16 single-nucleotide polymorphisms in *D. simulans*. A Hudson-Kreitman-Aguade test (Hudson *et al.* 1987) rejects the neutral hypothesis that polymorphism and divergence are coupled for the two regions of *ci* ($\chi^2 = 5.85$, $P = 0.0150$), suggesting that the level of polymorphism in the *ci*-intron region is not due to an elevated mutation rate. A recent selective sweep over the whole chromosome is neither supported by the high level of variation observed in many gene regions nor supported by results of Tajima's *D* tests for both individual genes and the pooled chromosomal polymorphism, which are not significant. On the other hand, the possibility of background selection cannot be ruled out, as both recombination and variation are lower than the levels of the chromosome regions of normal recombination.

In the *ci* gene, we found that a high level of variation in introns is in sharp contrast to the low level of nucleotide polymorphism in exons. Despite the small sample size, it can be seen that the polymorphism in the intron region seems to reflect geographic origins of the isofemale lines; *e.g.*, the three Reunion lines appear to have peculiar variations although their exons are devoid of variation. This observation questions the previous conclusion that a recent selective sweep has occurred in the *ci* gene (Berry *et al.* 1991), because such a scenario would predict a low level of variation in the intron too. It can be speculated that such a sweep may not be complete, *e.g.*, beyond the Reunion population, because the introns of non-Reunion populations contain little polymorphism. However, this partial selective sweep hypothesis would predict polymorphism in exon regions of the Reunion *ci*, which is not supported by the data. The *ci* gene has a low K_A/K_S ratio (0.14, a value lower than that of almost all other gene regions on the chromosome; Figure 1) and a K_A (0.0157) lower than the average K_A, 0.0185, for ~193 genes between the two species from Betancourt and Presgraves (2002; all *Acp* genes excluded) of the *ci* exon region; it may be more likely that the exon region of *ci* has been under purifying selection, consistent with the results of Jensen *et al.* (2002) that there is no strong evidence for sweeps in either *D. melanogaster* or *D. simulans*.

In this work, we revealed considerably high levels of the nucleotide variation associated with most of the genes we surveyed, suggesting that most parts of the chromosome are variable. The several low-variation regions that are scattered within the chromosome delineate the different chromosomal regions that have been governed by different evolutionary processes. Interpretation of such a level of variation with different evolutionary histories has to invoke recombination, probably crossing over. Furthermore, a level of recombination close to the *Adh* region of *D. melanogaster* was detected. The difference between *D. melanogaster* and *D. simulans* in the level of recombination is so large that the previous assumption that the chromosomes in the two Drosophila sibling species experience similar levels of recombination has to be reconsidered. Thus, our population genetic analysis demonstrates a level of recombination on the *D. simulans* fourth chromosome previously deemed unlikely in genetic analyses, allowing the persistence of a high level of natural variation.

We thank M.-L. Wu and W. Ballard for the *D. simulans* strains; K. Burtis for providing the strain used in the *D. melanogaster* genome sequencing project; J. Locke for valuable discussion of gene orders in *Drosophila*; R. R. Hudson, M. Kreitman, C.-I Wu, S. Hawley, and two anonymous reviewers for stimulating discussion and suggestions; C. H. Langley and M. Kreitman for critically reading the manuscript; and D. J. Begun for sending the published polymorphism data on *D. simulans*. We are especially indebted to R. R. Hudson for his help in computational analysis of recombination. We thank Jeff Wall for providing recombination rate estimates for *D. simulans*. This work was supported by grants from the National Science Foundation (MCB9977990 and CAREER award), National Institutes of Health (R01GM065429-01A1), and David and Lucile Packard Foundation (Packard Fellowship in Science and Engineering) to M.L.

LITERATURE CITED

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185–2195.

Aguade, M., N. Miyashita and C. H. Langley, 1989a Restriction-map variation at the zeste-tko region in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 123–130.

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

Aguade, M., W. Meyers, A. D. Long and C. H. Langley, 1994 Single-strand conformation polymorphism analysis coupled with stratified DNA sequencing reveals reduced sequence variation in the su(s) and su(wa) regions of the *Drosophila melanogaster* X chromosome. Proc. Natl. Acad. Sci. USA **91:** 4658–4662.

Andolfatto, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila* melanogaster and *Drosophila simulans*. Mol. Biol. Evol. **18:** 279–290.

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

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination, and DNA polymorphism in *Drosophila*, pp. 46–55 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

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

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.

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

Berry, A. J., and A. Barbadilla, 1999 Gene conversion is a major determinant of genetic diversity at the DNA level, pp. 102–123 in *Evolutionary Genetics: From Molecules to Morphology,* edited by R. Singh and C. Krimbas. Cambridge University Press, Cambridge UK.

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.

Betancourt, A. J., and D. C. Presgraves, 2002 Linkage limits the power of natural selection in *Drosophila*. Proc. Natl. Acad. Sci. USA **99:** 13616–13620.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

Bridges, C. B., 1935 The mutants and linkage data of chromosome four of *Drosophila melanogaster*. J. Biol. (Moscow) **4:** 401–420.

Carr, M., J. R. Soloway, T. E. Robinson and J. F. Brookfield, 2001 An investigation of the cause of low variability on the fourth chromosome of *Drosophila*. Mol. Biol. Evol. **18:** 2260–2269.

Charlesworth, B., 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet Res. **68:** 131–149.

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

Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res. **70:** 155–174.

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

Dvorak, J., M. C. Luo and Z. L. Yang, 1998 Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing aegilops species. Genetics **148:** 423–434.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

Grell, R. F., 1971 Heat induced exchange in the fourth chromosome of diploid females of *Drosophila melanogaster*. Genetics **69:** 523–527.

Hartl, D. L., and A. G. Clark, 1997 *Principles of Population Genetics*. Sinauer, Sunderland, MA.

Hawley, R. S., K. S. McKim and T. Arbel, 1993 Meiotic segregation in *Drosophila* melanogaster females: molecules, mechanisms, and myths. Annu. Rev. Genet. **27:** 281–317.

Hilliker, A. J., G. Harauz, A. G. Reaume, M. Gray, S. H. Clark *et al.*, 1994 Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. Genetics **137:** 1019–1026.

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

Hochman, B., 1976 The fourth chromosome of *Drosophila melanogaster*, pp. 903–928 in *The Genetics and Biology of Drosophila*, Vol. 1b, edited by M. Ashburner and E. Novitski. Academic Press, New York.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–42.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

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

Hudson, R. R., and N. L. Kaplan, 1994 Gene trees with background selection, pp. 140–153 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

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

Jensen, M. A., B. Charlesworth and M. Kreitman, 2002 Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. Genetics **160:** 493–507.

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

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

Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci *in Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen and J. M. Braverman, 2000 Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the *Drosophila melanogaster* X chromosome. Genetics **156:** 1837–1852.

Locke, J., L. Podemski, N. Aippersbach, H. Kemp and R. Hodgetts, 2000 A physical map of the polytenized region (101EF–102F) of chromosome 4 in *Drosophila melanogaster*. Genetics **155:** 1175–1183.

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

Nachman, M. W., 1997 Patterns of DNA variability at X-linked loci in *Mus domesticus*. Genetics **147:** 1303–1316.

Nachman, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17:** 481–485.

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

Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 The effect of recombination on background selection. Genet. Res. **67:** 159–174.

Patterson, J. T., and H. J. Muller, 1930 Are "progressive" mutations produced by X-rays? Genetics **15:** 495–578.

Podemski, L., C. Ferrer and J. Locke, 2001 Whole arm inversions of chromosome 4 in *Drosophila* species. Chromosoma **110:** 305–312.

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

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

Sandler, I., and P. Szauter, 1978 The effect of recombination-defective meiotic mutants on the fourth-chromosome crossing over in *Drosophila melanogaster*. Genetics **90:** 699–712.

Sheldahl, L. A., D. M. Weinreich and D. M. Rand, 2003 Recombination, dominance and selection on amino acid polymorphism in the Drosophila genome. Contrasting patterns on the X and fourth chromosomes. Genetics **165:** 1195–1208.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

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

Sun, F. L., M. H. Cuaycong, C. A. Craig, L. L. Wallrath, J. Locke *et al.*, 2000 The fourth chromosome of *Drosophila melanogaster*: interspersed euchromatic and heterochromatic domains. Proc. Natl. Acad. Sci. USA **97:** 5340–5345.

Tajima, F., 1989 Statistical method for testing 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.

Wall, J. D., P. Andolfatto and M. Przeworski, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203–216.

Wang, W., J. Zhang, C. Alvarez, A. Llopart and M. Long, 2000 The origin of the *Jingwei* gene and the complex modular structure of its parental gene, yellow emperor, in *Drosophila melanogaster*. Mol. Biol. Evol. **17:** 1294–1301.

Wang, W., F. Brunet, E. Nevo and M. Long, 2002a Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **99:** 4448–4453.

Wang, W., K. Thornton, A. Berry and M. Long, 2002b Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science **295:** 134–137.

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

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

Communicating editor: L. Harshman