Population, Evolutionary and Genomic Consequences of Interference Selection

Josep M. Comeron[*,†,1] **and Martin Kreitman**[*]

[*]*Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637 and* [†]*Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242*

Manuscript received August 15, 2001
Accepted for publication February 14, 2002

ABSTRACT

Weakly selected mutations are most likely to be physically clustered across genomes and, when sufficiently linked, they alter each others' fixation probability, a process we call interference selection (IS). Here we study population genetics and evolutionary consequences of IS on the selected mutations themselves and on adjacent selectively neutral variation. We show that IS reduces levels of polymorphism and increases low-frequency variants and linkage disequilibrium, in both selected and adjacent neutral mutations. IS can account for several well-documented patterns of variation and composition in genomic regions with low rates of crossing over in Drosophila. IS cannot be described simply as a reduction in the efficacy of selection and effective population size in standard models of selection and drift. Rather, IS can be better understood with models that incorporate a constant "traffic" of competing alleles. Our simulations also allow us to make genome-wide predictions that are specific to IS. We show that IS will be more severe at sites in the center of a region containing weakly selected mutations than at sites located close to the edge of the region. *Drosophila melanogaster* genomic data strongly support this prediction, with genes without introns showing significantly reduced codon bias in the center of coding regions. As expected, if introns relieve IS, genes with centrally located introns do not show reduced codon bias in the center of the coding region. We also show that reasonably small differences in the length of intermediate "neutral" sequences embedded in a region under selection increase the effectiveness of selection on the adjacent selected sequences. Hence, the presence and length of sequences such as introns or intergenic regions can be a trait subject to selection in recombining genomes. In support of this prediction, intron presence is positively correlated with a gene's codon bias in *D. melanogaster*. Finally, the study of temporal dynamics of IS after a change of recombination rate shows that nonequilibrium codon usage may be the norm rather than the exception.

THE general concept of effective population size (N_e), due to Wright (1931, 1938), is usually associated with species-specific factors such as inbreeding, unequal numbers of the two sexes, variance in mating success, temporal variation in population size, population subdivision, and pervasive selection. According to theory, these factors are expected to have a genome-wide influence on the standing crop of both weakly selected and selectively neutral mutations because evolutionary parameters governing these mutations, N_es (α) and $N_e\mu$ (β), respectively, include effective population size (see Table 1). However, polymorphism levels are not constant across the genome but rather are correlated with recombination rates, suggesting that additional factors may be influencing N_e (Aguadé *et al.* 1989; Stephan and Langley 1989, 1998; Begun and Aquadro 1992; Martín-Campos *et al.* 1992; Langley *et al.* 1993; Aguadé and Langley 1994; Aquadro *et al.* 1994; Stephan 1994; Nachman 1997; Dvorák *et al.* 1998; Kraft *et al.* 1998; Nachman *et al.* 1998; Przeworski *et al.* 2000). This has led to theoretical investigations of the effects of linkage and selection on neutral variation levels.

Further investigation on whether N_e can be viewed as varying across a genome takes advantage of its consequences on the effectiveness of weak selection. Theory predicts that the evolutionary dynamics of mutations whose selective effects are on the order of the reciprocal of population size (*i.e.*, $\alpha \approx 0.25$–2.5) are expected to be very sensitive to small shifts in N_e (Ohta and Kimura 1971; Ohta 1972, 1995; Li 1987). Two lines of evidence are congruent with N_e changing across genomes in Drosophila (Hilton *et al.* 1994). The first comes from studies on the biased usage of synonymous codons (codon bias), a paradigmatic example of weak selection for translational efficiency in many organisms (Grantham *et al.* 1981; Ikemura 1981; Bennetzen and Hall 1982; Grosjean and Fiers 1982; Kurland 1987; Sharp and Li 1987, 1989; Shields *et al.* 1988; Bulmer 1991; Moriyama and Hartl 1993; Akashi 1994, 1995, 1996; Eyre-Walker 1996; Comeron and Kreitman 1998; Comeron *et al.* 1999; McVean and Vieira 2001). Codon bias increases with recombination rates in *Drosophila melanogaster*, indicating an increase in the effectiveness of selection (and hence larger N_e) in regions of high

[1] *Corresponding author:* Department of Biological Sciences, University of Iowa, 433 Biology Bldg., Iowa City, IA 52242.
E-mail: josep-comeron@uiowa.edu

TABLE 1

Variables and parameters

N_e	Effective population size
s	Selection coefficient
μ	Mutation rate per site per generation
c	Recombination rate between adjacent sites per generation
α	$N_e s$
β	$N_e \mu$
ρ	$N_e c$

α_N, β_N, and ρ_N are the parameters corresponding to α, β, and ρ, respectively, when they refer to values used in computer simulations.

recombination (Kliman and Hey 1993; Comeron *et al.* 1999), an effect that is also observed using the complete *D. melanogaster* genome (J. M. Comeron, unpublished data). Second, in interspecific comparisons in the *D. melanogaster* species complex, the rate of replacement substitutions (K_a) is significantly smaller in regions of high recombination than in regions of low recombination (Hilton *et al.* 1994; Comeron and Kreitman 2000), suggesting stronger purifying selection against deleterious amino acid changes in genes located in regions of high recombination.

Several models of strong selection ($\alpha \gg 1$) have been proposed to explain why N_e is reduced in regions of low recombination: (i) the hitchhiking (HH) and pseudo-HH (pHH) models, which invoke frequent positive Darwinian selection (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Gillespie 2000); (ii) the background selection (BGS) model, which assumes a high and constant rate of deleterious mutations (Charlesworth *et al.* 1993; Charlesworth 1994, 1996; Hudson and Kaplan 1995); (iii) the joint effects of positively and negatively selected mutations (Kim and Stephan 2000); and (iv) models based on fluctuating selection (Gillespie 1997). The fixation of a favorable allele (*i.e.*, selective sweep or draft under the HH and pHH models, respectively) alters the probability of fixation of linked selected mutations present in the population, decreasing the probability of fixation of other advantageous mutations and increasing the chance of fixation of deleterious mutations. A similar outcome will be produced by the steady elimination of strongly deleterious alleles and the indirect uniform reduction of N_e under BGS. In all of these models, mutations under strong directional selection make linked mutations behave as if they are evolving under a smaller population size (Robertson 1961). The indirect effects of selected mutations on linked variation will be greatest in the absence of recombination and will decrease with increasing recombination.

However, strong selection is not a requirement for this effect. Hill and Robertson (1966), using simulations of a two-locus model, showed that the probability of fixation of a favorable mutation at one locus becomes reduced due to the *segregation* of alleles at a second selected locus. This so-called Hill-Robertson effect (Felsenstein 1974; Lewontin 1974) is caused by the presence of selected mutations in the genetic background and the consequent increment of stochastic sampling effects (*i.e.*, variance of offspring number and genetic drift) in finite populations (see also Robertson 1961; Birky and Walsh 1988). An increase in genetic drift under a Hill-Robertson scenario has the consequence of reducing the intensity of selection and it is regarded as equivalent to a reduction of N_e (Robertson 1961; Hill and Robertson 1966; Felsenstein 1974; Kliman and Hey 1993). These effects of interference apply mostly to mutations under moderate/weak selection because these mutations have much longer sojourn times than strongly selected mutations and this enhances the opportunity for weakly selected mutations to influence one another's fate.

Weakly selected mutations, taken individually, are not expected to have a measurable effect on population parameters or on the tree topology of linked neutral mutations (Golding 1997; Neuhauser and Krone 1997; Przeworski *et al.* 1999). Similarly, interference between pairs of weakly selected mutations should also be negligible because this interaction is expected to be a second-order magnitude effect. We are, however, interested in investigating interference due to segregating mutations under weak selection, which we call Interference Selection (IS), because for many organisms weakly selected mutations are both abundant and physically clustered within genomes, and hence the magnitude of IS interactions becomes measurably large. First, a large fraction of the mutations segregating in populations of many species may be weakly selected. Synonymous mutations, for example, are an abundant source of weakly selected mutations in Drosophila and other species. Weakly selected mutations may also be common in many species among amino acid replacement changes (Ohta 1995; Zeng *et al.* 1998) and in regulatory regions (Ludwig *et al.* 1998). Second, weakly selected mutations are likely to be physically clustered in the coding regions of genes and in regulatory regions, generating genomic regions with a high density of segregating weakly selected mutations. Reduced physical distance and hence recombination between these mutations creates the opportunity for multiple interference interactions. Nevertheless, it is not obvious whether IS can be rationalized simply in terms of N_e because the magnitude of its effect will depend jointly on the mutation rate to selected alleles, the intensity of selection, and the recombination rate between mutations under selection.

Li (1987), using simulations, showed that weakly selected mutations under complete linkage can interfere with one another to cause shifts in the frequency of favorable mutations at equilibrium. Later, Comeron *et al.* (1999) allowed the numbers of selected sites, selection coefficients, and recombination rates to vary across ranges of values thought to be biologically realistic for species such as Drosophila. This multilocus study showed that the level of codon bias (*i.e.*, used as a proxy for the effectiveness of selection and N_e) decreases when either the recombination rate decreases or the number of weakly selected sites increases for different intensities of weak selection. McVean and Charlesworth (2000) further investigated both codon bias and level of polymorphism in selected sites under IS, confirming that these traits are affected by IS not only for absolute linkage but for a range of recombination rates observed in outcrossing species. These studies proposed the Hill-Robertson effect [the small-scale Hill-Robertson (Comeron *et al.* 1999) or weak-selection Hill-Robertson (McVean and Charlesworth 2000) effect] and a reduction in N_e to explain the results. These two studies, however, analyzed only the consequences of IS on the sites under selection, where the outcome is the composite effect of both selection on the mutations themselves and IS (which at the same time alters the effectiveness of selection).

Tachida (2000) studied evolutionary effects that selection at many sites has on interlaced neutral sites with total linkage. Here we extend this approach with the analysis of many population and evolutionary parameters under IS both on selected mutations and on adjacent neutral mutations, using a broad range of numbers of selected sites, recombination rates, and weak selection coefficients. The effects of IS on adjacent neutral variability are key to gaining insight into the mechanism and evolutionary effects of IS and to appraising the importance of IS as a viable evolutionary force.

We also investigate two other consequences of IS under low rates of recombination. First, because genes (exons and regulatory regions) are embedded in a matrix of generally less severely constrained DNA, IS may occur at sites with well-defined boundaries along the DNA. Therefore, we study the expected consequences of IS along such intervals under selection. We hypothesize that the effects of IS should not be uniform across a gene, and we use simulations to generate predictions that can be tested with Drosophila genomic data. Second, neutral sequences located between groups or clusters of selected sites under weak/moderate selection (*e.g.*, introns within coding regions of genes) can be viewed as modifiers of recombination and hence can alter the effectiveness of selection (Comeron and Kreitman 2000; Comeron 2001). We investigate whether this indirect selection on the presence and length of intermediate "neutral" sequences is plausible. If true, this makes indels in introns and other noncoding regions potential targets for selection.

MATERIALS AND METHODS

Forward computer simulations: A Wright-Fisher model was simulated with N diploid individuals ($2N$ chromosomes) as previously described (Comeron *et al.* 1999; see also Li 1987) and with the following modifications. Every chromosome is composed of two adjacent stretches of sites, each of length L, with one stretch evolving neutrally (neutral sequence) and one evolving under weak selection (selected sequence). In each generation the total number of mutations and recombination events are drawn from a Poisson distribution with mean $4L\beta_N$ and $3L\rho_N$, respectively (see Table 1 for definitions). The number of recombination events per meiosis is not restricted (assuming no chiasma interference) to avoid underestimating the effect of recombination in long sequences when ρ_N is high. The mutation process allows only two allelic states at a site, and reversible mutation is permitted, mimicking the mutation process between preferred (p) and unpreferred (u) codons. Mutation rates from p to u and vice versa are w and v, respectively, where $w + v = \mu$ and γ ($\gamma = v/\mu$) is the mutational bias. Unless otherwise indicated, we applied a mutation rate of $\beta_N = 0.01$, with $\gamma = 0.45$. The previous assumption of only two allelic states (Li 1987; Comeron *et al.* 1999; McVean and Charlesworth 2000; Tachida 2000), one favorable and the other deleterious, is a reasonable approximation for weakly selected mutations at synonymous sites in organisms like Drosophila where G/C-ending codons are the preferred state (Shields *et al.* 1988; Akashi 1994, 1995). Based on preliminary studies (appendix), our simulated populations were composed of 1000 chromosomes; a sample size of 12 sequences was used to estimate population genetic parameters.

The fitness of each individual is based only on the selected sequence. The selection differential between p (preferred allele) and u (unpreferred allele) in the selected sequence is $+s$; fitness is multiplicative over sites and mutations are semidominant in their effect on fitness. Each new generation is obtained by first choosing N individuals ($2N$ chromosomes) with probability proportional to their relative fitness. The next generation is constituted by randomly pairing the $2N$ chromosomes (each composed by the selected and adjacent neutral sequence), which are possibly mutated and/or recombined to form N new diploid individuals.

As indicated in results, the time to reach base composition equilibrium under a mutation-selection-drift (MSD) balance and IS (*e.g.*, codon usage) when $\beta_N = 0.01$ may take on the order of ≈100–250 N generations. Accordingly, our study of IS at equilibrium begins after a minimum of 250 N generations to assure base composition equilibrium. Each independent population realization was analyzed every N generations for a minimum of 1000 N generations. Population parameters were estimated in 20 independent samples. All estimates of population and evolutionary parameters (heterozygosity, nucleotide diversity, frequency skew, fixation rates) as well as the frequency of preferred codons (P) were obtained by studying the same number of sites, 250, regardless of the total number of sites in the sequence when $L \geq 250$. These studied sites were homogeneously distributed across the sequence, unless explicitly indicated, to assure an average estimate of the parameters across the region. In every simulation both the selected and neutral sequences were analyzed. The ranges of parameter values we investigated for recombination, selection, and length were $0 \leq \rho_N \leq 0.4$, $0.25 \leq \alpha_N \leq 2.5$, and $125 \leq L \leq 2500$. The ranges of recombination rates under study are representative of most eukaryotes, including *D. melanogaster*. Assuming $N_e \approx 1 \times 10^6$ for *D. melanogaster* (Andolfatto and Przeworski 2000) and using laboratory-based rates of crossing over, the complete *D. melanogaster* genome would satisfy $\rho_N < 0.05$, $\rho_N < 0.1$ after taking into account gene conversion (Hilliker and Chovnick 1981; Andolfatto and Nordborg 1998; Andolfatto and Przeworski 2000). *D. melanogaster* genomic regions with $\rho_N \leq 0.004$ may contain ~15% of genes using rates of crossing over; these genomic regions are defined by the cytological bands 1A–2C/20C–20F (X chromosome), 21A/38B–40F/41A–44B/60D–60F (chromosome 2), 61A–61B/76B–80F/81A–84E (chromosome 3), and the complete fourth chromosome. When the contribution of gene conversion to the total recombination is taken into account, $\rho_N \leq 0.004$ may apply to >10% of *D. melanogaster* genes.

To evaluate the relative change in the effectiveness of selection on the selected sequences caused by varying the parameters (changing recombination rates and L and presence and length of intermediate regions) we compared the estimated value of the parameter α on the basis of the observed P. Following directly from the probability of fixation of p and u, and P at equilibrium under the infinitely many sites model and free recombination (Wright 1931; Crow and Kimura 1970; Ewens 1979), we have for diploid organisms

$$\alpha = \left(\frac{1}{4}\right)\mathrm{Ln}\left[\frac{P(1-\gamma)}{(1-P)\gamma}\right]$$

(see Li 1987; Bulmer 1991), when $\beta \ll 1$. We call this the single-site model based on MSD (SS-MSD).

Linkage disequilibrium (LD) was estimated as the average over all pairwise comparisons of polymorphic sites by using D' (LD-D'; Lewontin 1964); coupling gametes were composed of the most frequent alleles (Langley *et al.* 1974). We used LD-D' because this measure is the least affected by the frequency of the mutations compared to measures such as D (Lewontin and Kojima 1960) or Zns (Hill and Robertson 1968; Kelly 1997) in conditions of reduced or no recombination. However, LD-D' is not entirely independent of the frequency of the mutations (Lewontin 1988), and so we also studied the extent of LD within predefined frequency classes. We avoided using sequence-based estimates of recombination such as Hudson's C (C_{hud}; Hudson 1987) because its mean is influenced not only by the frequency of mutations but also by the number of segregating sites when this number is not large, a number that is changed by IS. To make comparable estimates of LD between sequences of different length, all estimates of LD were obtained using the polymorphisms detected within 250 contiguous sites; otherwise, the average distance between polymorphisms increases with L, biasing the comparisons of LD.

Analyses of the *D. melanogaster* genome: We studied the complete *D. melanogaster* genome (Adams *et al.* 2000; Version 2 (October 2000), http://www.ebi.ac.uk/genomes). The analyses were performed using all 9172 genes with empirical data supporting both their presence and specified exon/intron structure (*i.e.*, with complete mRNA information and no evidence of multiply spliced forms). The study of intergenic sequences was carried out using only those intergenic sequences (6271) that are flanked by two genes with empirical data supporting both their presence and specified gene structure.

Heterogeneous codon bias across exons: Two groups of genes were investigated. The first group (659 genes) was composed of all genes with a single long exon (>1000 bp or >333 amino acids). The second group (187 genes) included all genes with long coding regions (>333 amino acids) interrupted by introns and satisfying two criteria: (i) all (one or more) introns should be centrally located, dividing the coding region into two comparable regions (*i.e.*, introns located between 30 and 70% of the relative total length of the coding region), and (ii) at least one intron should be >100 bp. The synonymous codon usage bias was measured using the frequency of GC-ending codons (GC3), the frequency of GC-ending codons in four-fold degenerate amino acids (GC4), and the frequency of preferred codons in *D. melanogaster* (Akashi 1995). As indicated, GC3 is a good measure of codon bias in *D. melanogaster* and at the same time is equivalent to the *in silico* study of IS with two alleles.

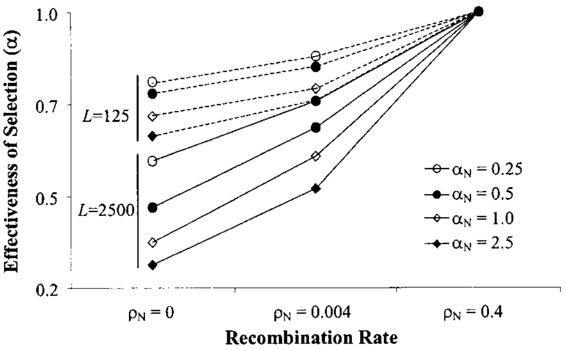

Figure 1.—Relative effectiveness of selection (α) per site based on the frequency of preferred sites (P) for different recombination rates (ρ_N) and scaled selection coefficients (α_N). Results are shown for two extreme cases of number of sites under selection (L), $L = 125$ and $L = 2500$.

Codon bias and the proportion of selected sites in a gene: The analysis was carried out using all 7499 complete genes (out of 9172) with introns. As a proxy for the relative number (or *density*) of selected sites in a gene, we used the proportion of the length of the coding region (PLCR) in a gene, measured as the ratio between the length of the coding region and the length of the coding region plus the total length of the introns.

The recombination rate for each gene in the *D. melanogaster* genome was estimated as previously described (see Comeron *et al.* 1999 for details) on the basis of the cytological position associated to each genomic scaffold sequence. All statistical analyses were carried out using Statistica for Windows 5.1 (1997).

RESULTS

Effects of IS on population and evolutionary parameters at selected and adjacent neutral sequences

Effectiveness of selection: We investigated the effectiveness of selection on weakly selected mutations by analyzing the proportion of preferred mutations at equilibrium (P; Li 1987; Comeron *et al.* 1999; McVean and Charlesworth 2000). In contrast to neutral polymorphism, for which measures such as heterozygosity (θ) are nearly linearly related to N_e, the relationship between P and the selection parameter α ($N_e s$) is strongly nonlinear. For instance, a 5% increase in P represents a 50, 21, and 27% increase in α when the original P is 0.5, 0.6, and 0.9. Therefore, although our simulations measure shifts in P with changes of parameters affecting IS, the magnitude of these shifts is better reflected by the change in the parameter α needed to account for the results under a no-interference model (SS-MSD). As Figure 1 shows, the effectiveness of selection, as measured by α, decreases as the recombination rate decreases, and this effect increases with L (see also Comeron *et al.* 1999 and McVean and Charlesworth 2000 for the effect of IS on measures of codon bias and P). In addition, the relative reduction in the effectiveness of selection due to IS increases with the strength of selection acting on individual mutations ($\alpha_N < 2.5$).

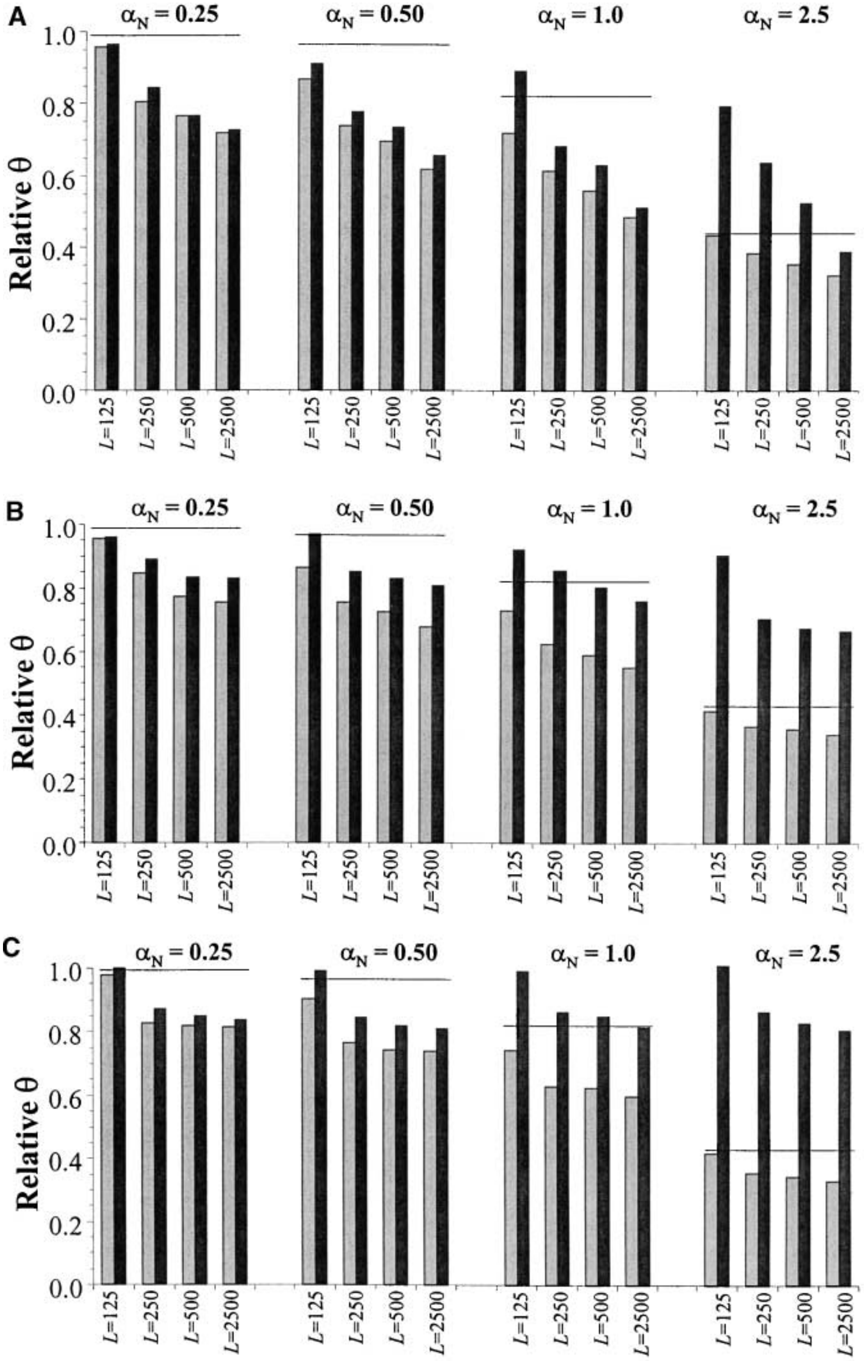

Figure 2.—Relative level of polymorphism (heterozygosity; θ) in selected (θ_s) and adjacent neutral sequences (θ_n). Results are shown for different recombination rates (ρ_N), scaled selection coefficients per site (α_N), and number of selected sites (L). Horizontal lines indicate the expected values under a single-site model and mutation-selection-drift balance (SS-MSD). Sample size (n) = 12. (▨) Selected (θ_s). (■) Neutral (θ_n). (A) $\rho_N = 0$, (B) $\rho_N = 0.004$, (C) $\rho_N = 0.4$.

Polymorphism levels: We studied the effect of IS on polymorphism levels, as measured by heterozygosity, in selected (θ_s) and neutral (θ_n) sequences. Under single-site models of weak selection (SS-MSD), the expectations are clear. A general reduction of the intensity of selection (α) predicts a relative increase of θ_s, making θ_s closer to θ_n. On the other hand, a reduction in N_e will cause a direct reduction in θ_n. The expected net consequence of reducing N_e, hence α, for mutations under SS-MSD is a reduction of θ_s because the reduction of θ_n is always greater than the expected increase of selected polymorphism due to a reduced selection, although this decrease of θ_s is not expected to be proportional to the reduction of N_e. For strong selection, a moderate reduction of N_e would not alter θ_s.

Our results (Figure 2) show that θ_s is below the levels expected on the basis of the imposed strength of selection acting on these mutations under SS-MSD. For all combinations of selection intensity (α_N) and recombination rates (ρ_N), θ_s decreases as the number of sites under selection (L) increases (see also McVean and Charlesworth 2000). The relative impact that increasing L

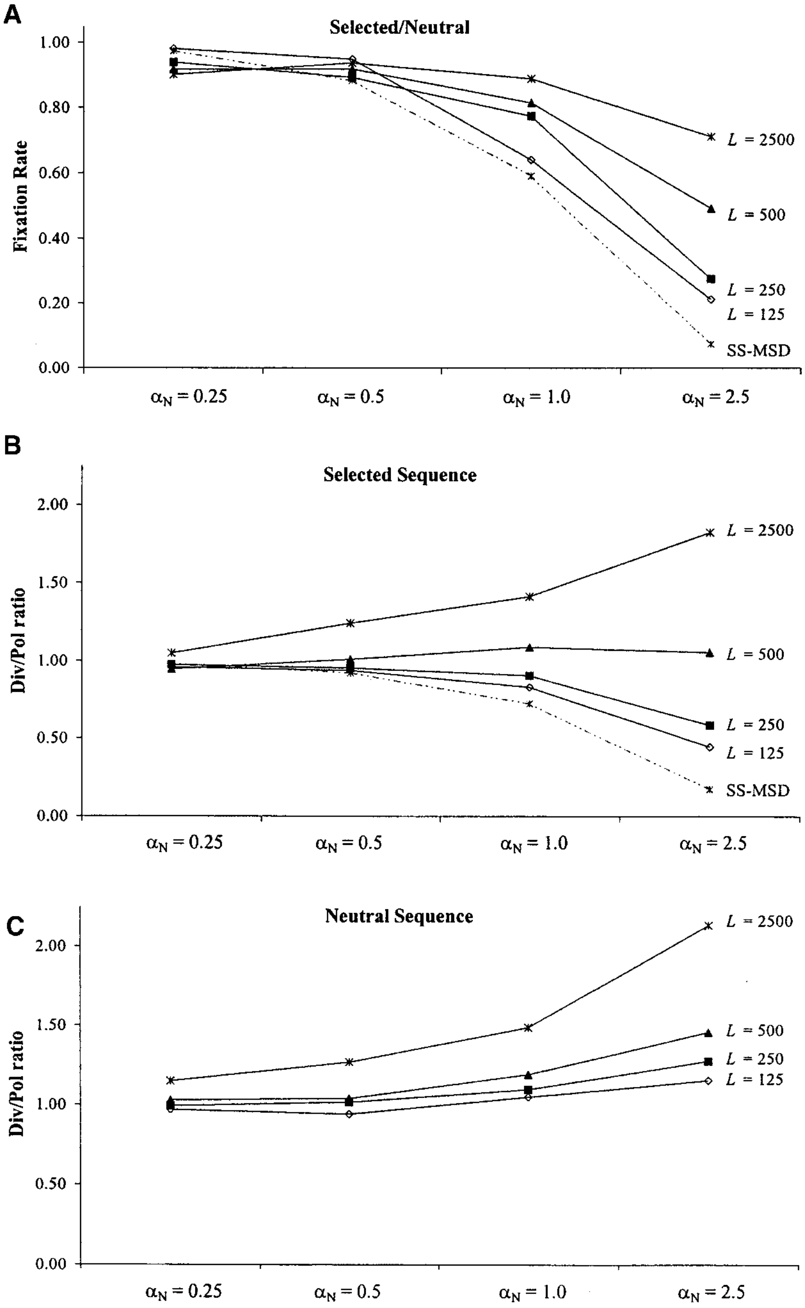

Figure 3.—Divergence (fixation rate) and divergence:polymorphism (Div/Pol) ratio under IS in selected and adjacent neutral sequences for $\rho_N = 0$. Results are shown for different scaled selection coefficients per site (α_N) and different number of selected sites (L). For all cases, values are relative to neutral expectations. (A) Fixation rate of selected sequences relative to adjacent neutral sequences. Dashed line indicates the expected fixation rate under a single-site model and mutation-selection-drift balance (SS-MSD). (B) Div/Pol ratio of selected sequences. Dashed line indicates the expected Div/Pol ratio under SS-MSD. (C) Div/Pol ratio in neutral sequences adjacent to sequences with L selected sites. $n = 12$.

has in reducing heterozygosity is similar for different selection intensities when $\rho_N = 0$ (*e.g.*, θ_s for $L = 2500$ is 70–75% of that for $L = 125$, both for $\alpha_N = 0.25$ and $\alpha_N = 2.5$), and this impact decreases, but is still noticeable, for very weak selection and high recombination (*e.g.*, $\alpha_N = 0.25$ and $\rho_N = 0.4$). For $\alpha_N = 0.25$, we also studied whether an even higher recombination rate ($\rho_N = 1.0$) would completely eliminate IS. The results show that $\rho_N = 1.0$ does eliminate most IS on θ_s when $L \leq 125$ compared to SS-MSD expectations, but IS is still detectable for larger L.

Heterozygosity is also reduced in the adjacent neutral sequences as a result of linkage to sites under weak selection. The most extreme reductions in neutral variation are observed for $\rho_N = 0$ (Figure 2A), where increasing either L or α_N in the selected sequence substantially reduces θ_n. The impact that the L has on θ_n increases with the intensity of selection. For example, when $L = 2500$, θ_n is ~75 and ~50% of that observed when $L = 125$ for $\alpha_N = 0.25$ and $\alpha_N = 2.5$, respectively. When $\rho_N = 0$ and L is large, there is a tendency to observe similar levels of polymorphism in selected (θ_s) and adjacent neutral mutations (θ_n), which are most evident for $\alpha_N = 0.25$ (when $L \leq 2500$), implying that linked selectively neutral sites and sites under weak selection may not be distinguishable by this criterion. Increasing L reduces θ_n even when the recombination rate is high (Figure 2, B and C). This observed reduction in θ_n is similar for

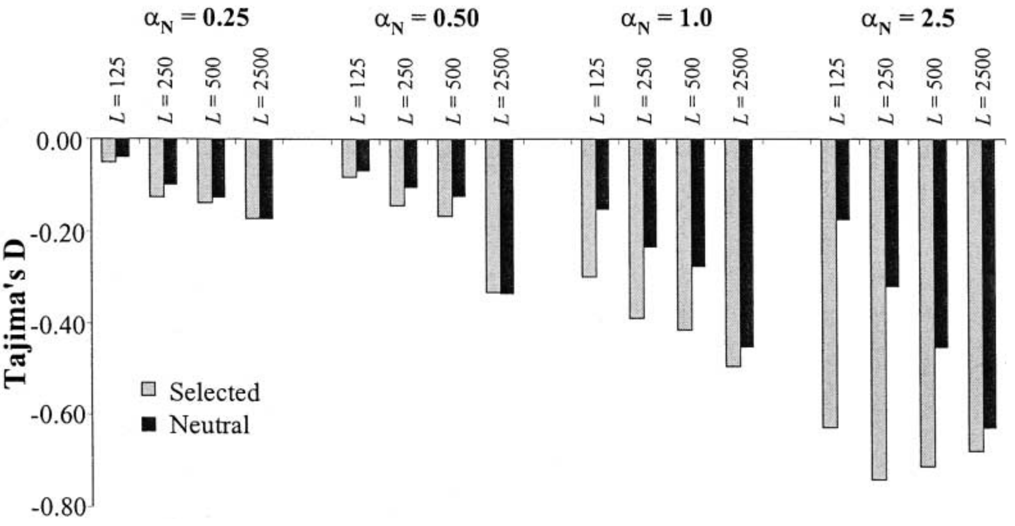

Figure 4.—Tajima's D statistic under IS in selected and adjacent neutral sequences for $\rho_N = 0$. Results are shown for different scaled selection coefficients per site (α_N) and different numbers of selected sites (L). $n = 12$.

different selection intensities when recombination is highest ($\rho_N = 0.4$), likely reflecting a recombination rate threshold for IS.

Divergence and divergence to polymorphism ratio: Under SS-MSD equilibrium the rate of fixation or divergence of selected mutations rapidly decreases with increasing selection intensity α. We focused on the rate of divergence when $\rho_N = 0$ to illustrate the effect of IS on this evolutionary parameter (Figure 3A). As expected, selection acting at linked sites does not influence divergence for neutral mutations. The fixation rate of mutations under selection increases with L for any given α_N, indicative of a reduction in the effectiveness of selection due to IS.

The SS-MSD model also predicts that the divergence:polymorphism ratio (Div/Pol) for weakly selected mutations decreases with increasing α because weak selection has stronger effects in reducing the rate of fixation than the level of polymorphism. Figure 3B shows the Div/Pol ratio for the region containing mutations under selection again for $\rho_N = 0$. For the case of $L = 125$, Div/Pol decreases with selection but to a lesser degree than that expected for a SS-MSD case. For the intermediate case of $L = 500$, Div/Pol is only barely affected by selection. More exceptional is the situation in which the number of sites under selection is moderate to large (*e.g.*, $L = 2500$): In these cases Div/Pol increases not only relative to single-site expectations but also relative to neutral expectations. This trend results from two opposing effects of IS on selected mutations, increasing divergence and reducing polymorphism. Neutral sites (Figure 3C) show a consistent increase in the Div/Pol ratio with increasing IS, caused by the effect that IS has in reducing levels of linked neutral polymorphism.

Mutation frequency spectrum: The SS-MSD models predict that, as α increases, weakly selected mutations will become less abundant and allele frequencies at polymorphic sites will decrease compared to neutral expectations. Figure 4 plots Tajima's D statistic, a measure of the skew of allele frequency compared to neutral frequency spectrum, for selected and neutral sequences under complete linkage. In the selected sequence, a more negative Tajima's D is observed with increasing selection intensity, as expected (see Tachida 2000). As L increases, and hence IS, a further excess of rare mutations is seen compared to single-site expectations for any given α_N. This trend does not hold, however, when α_N and L are large, (*i.e.*, $\alpha_N = 2.5$, $L > 500$), where Tajima's D becomes unaffected or even less negative. This is, however, not surprising because Tajima's D statistic is not entirely independent of the number of segregating sites: It tends toward zero as the number of segregating sites in a sample becomes small for any given (nonneutral) frequency of variants. Therefore, IS increases the relative frequency of rare variants (hence it induces a negative Tajima's D) but IS also decreases the number of segregating sites, thus biasing Tajima's D estimates closer to zero when IS and its reduction of the number of segregating sites are severe.

The frequency spectrum of neutral mutations departs from the neutral equilibrium expectation, showing an excess of low frequency alleles when IS occurs in the adjacent selected sequence. Tajima's D becomes more negative as the number of selected sites or α_N on these sites increases. When IS is strongest, the skew toward low frequencies becomes similar for both selected and neutral mutations, a trend we have also encountered for heterozygosity.

IS also influences the allele frequency of variants in the selected sequences when recombination is highest ($\rho_N = 0.4$) while the frequency spectrum of neutral variants in adjacent sequences remains mostly unaffected. The skew toward low frequency variants in the selected sequences increases with L, although to a lesser degree compared to $\rho_N \leq 0.004$, and this effect intensifies with increasing α_N.

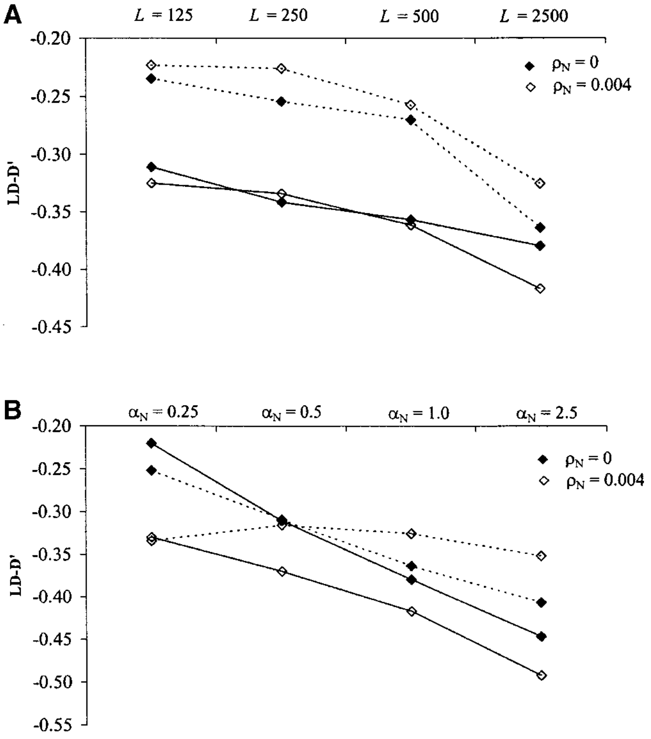

Figure 5.—Linkage disequilibrium measured by D' (LD-D') in selected (solid lines) and adjacent neutral sequences (dashed lines) for $\rho_N = 0$ and $\rho_N = 0.004$. (A) LD-D' for different numbers of sites (L) in the selected sequence and $\alpha_N = 1$. (B) LD-D' for different scaled selection coefficients per site (α_N) for $L = 2500$.

IS and linkage disequilibrium: Hill and Robertson (1966) showed that as recombination decreases (hence IS increases) there is an increment in repulsion associations between mutations under selection. McVean and Charlesworth (2000) further showed that this negative LD decreases with the recombinational distance between selected sites. Here, we investigated how the intensity of selection and the number of sites under selection—two contributors to IS—each influence the extent of LD, as measured by D' (LD-D'; Lewontin 1964). Again, we studied both selected and neutral mutations to better infer the mechanism influencing LD. Figure 5A shows that negative LD increases with L, for both intermediate- and no-recombination cases. The increment in LD is detected not only in the selected sequence but also in the adjacent neutral sequence. When L (and therefore IS) is small, LD-D' is clearly different for the selected and neutral sequences. But for larger values of L the magnitude of LD becomes more similar for the two classes of mutations. Increasing the intensity of selection α_N increases LD (with repulsion associations) in the selected sequences to a greater extent than in the adjacent neutral sequences (see Figure 5B for $L = 2500$). When recombination is very high ($\rho_N = 0.4$), LD-D' also varies in the selected sequences for different α_N, but not in the adjacent neutral sequences.

However, the conditions that increase negative LD are the very same conditions for which there is an excess of low-frequency variants (with more negative Tajima's D estimates). Because most measures of LD correlate, to some degree, with the frequency of the mutations under study (Lewontin 1988; see materials and methods), the observation of greater negative LD-D' in the same situations in which Tajima's D is also more negative might well be two faces of the same phenomenon: IS skews the frequency of mutations in the population. To assess whether the increment in D' with IS is only the result of an increasing skew in the mutational frequency spectrum, we studied the average LD-D' for different average Tajima's D classes (*i.e.*, as a proxy of different frequency classes); our interest centered on the results within each frequency class. Investigation of adjacent neutral sequences allowed us to eliminate the direct effects that selection has on the frequency spectrum of the selected mutations. Results, shown in Figure 6, show that LD-D' becomes increasingly negative for each frequency class as selection intensity α_N (and IS) increases. Therefore, the observed increase in LD with IS, as measured by D', is not only the consequence of a shift in the frequency spectrum.

Temporal dynamics after a change of recombinational environment: Genome-wide and/or gene-specific changes in recombination rates may be common in many evolutionary systems, and so it is important to study the time needed to reach new equilibria. For instance, in Drosophila there is extensive gene order shuffling within chromosomal arms between species (Lemeunier and Ashburner 1976; Pinsker and Sperlich 1984; Papaceit and Prevosti 1989; Whiting *et al.* 1989; Segarra and Aguadé 1992; Kress 1993; Lozovskaya *et al.* 1993; Gallego *et al.* 1999), which most likely causes frequent changes in recombination rates (Charlesworth 1994).

We studied the number of generations needed to reach equilibrium after a change in the recombination rate under IS conditions. For simplicity we assumed a population at equilibrium under the initial conditions and the instantaneous fixation of a randomly chosen allele in a new recombinational environment. Such a situation might apply to a gene located near the breakpoint of a chromosomal rearrangement that was quickly driven to fixation (possibly by natural selection). In accord with expectations, most population and evolutionary parameters reach their new equilibria much sooner than codon usage (Figure 7, A and B). Under neutrality, few N_e generations ($\approx 4N_e$ for diploid individuals) after a hitchhiking event are sufficient to achieve near-equilibrium levels of polymorphism (Perlitz and Stephan 1997). Under IS the time required for weakly selected mutations to reach values close to those at equilibrium for parameters such as θ_s or Tajima's D increases

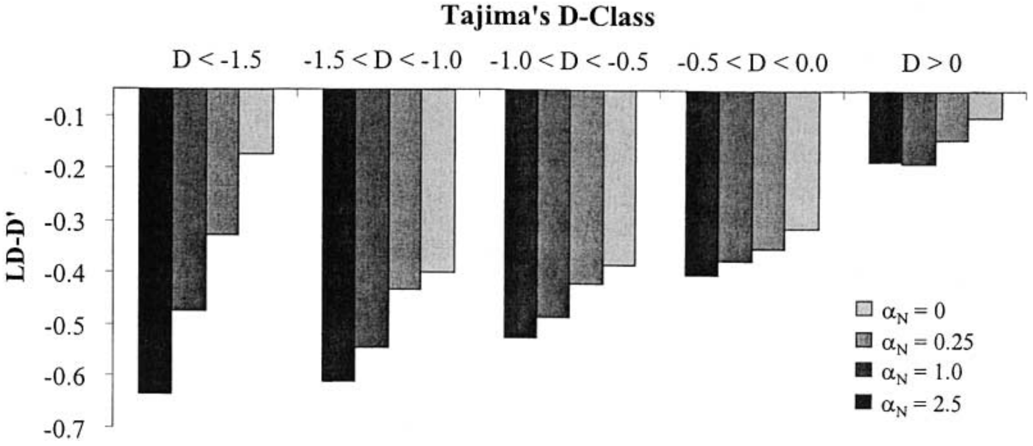

FIGURE 6.—Linkage disequilibrium measured by D' (LD-D') in neutral sequences adjacent to selected sequences in different Tajima's D classes for different scaled selection coefficients per site (α_N). $L = 2500$ and $\rho_N = 0$.

with α_N, requiring $\approx$20–40N generations when $\alpha_N = 2.5$ while this time is close to 4N generations when $\alpha_N \leq 0.5$.

Multilocus parameters, such as those that estimate codon usage, take a very large number of generations to reach a new equilibrium following perturbation. Indeed, codon usage requires $\approx$100–250N or $\sim$1–2.5/μ generations to reach the new equilibrium. This required time is not strongly dependent on the number of sites under selection, but it is dependent, as expected, on μ (data not shown). Two other features of codon bias evolution following a change of recombination environment were also observed. First, the change in codon bias is faster when the number of preferred mutations is increasing (*i.e.*, changing from low to high recombination) than when the number is decreasing. Second, the weaker the selection the longer the period required to reach the new base composition equilibrium in either direction. These two features can be easily explained by three factors: (i) The average time to fixation of weakly advantageous mutations is shorter than that expected for weakly deleterious mutations, (ii) the speed to fixation of preferred mutations increases with α, and (iii) mutational pressure toward unpreferred mutations is higher when the ancestral sequence has higher P.

We also observed a tendency for population and evolutionary parameters to "overshoot" their equilibrium values when a sequence changes from no recombination to a high recombination rate (stronger for $\alpha_N = 2.5$ than for $\alpha_N = 0.5$), but this effect is not detectable in the opposite direction. This overshoot creates a transient situation more nearly resembling a neutral equilibrium. Under the conditions we investigated, population parameters are closest to the neutral expectations $\approx$4–5N generations after the fixation of a single sequence in an environment with high recombination. For instance, in the case of $\alpha_N = 2.5$, θ_s changes through time from zero (right after the fixation event) to the new IS equilibrium under high recombination (after $\approx$ 40N generations) with an intermediate state 50% higher than that finally observed at equilibrium.

IS and its effect across regions under uniform selection: Consider an interval of a recombining genome in which segregating sites under selection result in IS, and further assume this interval is embedded in a region containing few additional mutations under selection. Since the magnitude of the IS effect acting at a particular site in this interval will be governed by the interactions between the segregating sites located on both sides of that site, it is reasonable to expect that IS will be stronger at sites embedded in the middle of a region under selection than at sites located close to an edge of this region. This situation may apply to many protein-coding regions in eukaryotic genomes, but it would be pertinent to any group of physically clustered sites under weak selection surrounded by largely unconstrained sites. Here, we investigate the magnitude of the "center" *vs.* "edge" effect for plausible rates of recombination and selection. The issue under scrutiny is whether IS differs measurably between the center and edge of a region under selection when both mutation rates and selection coefficients are uniformly distributed across the region.

We studied population and evolutionary parameters across sequences with 2500 sites under uniform selection, recombination, and mutation, with emphasis on a lateral and the central region of 250 sites each. Two indicators of IS are depicted in Figure 8 for the central and lateral regions: the proportion of preferred codons (P) and the divergence to polymorphism (Div/Pol) ratio. As expected, no effect is seen for the no-recombination case ($\rho_N = 0$). For intermediate recombination rates, IS differs between regions, with the central region showing stronger IS (central regions have lower P and higher Div/Pol ratio than lateral regions). This heterogeneous distribution of IS across regions (the center effect) decreases when recombination is very high, but it can still be seen for high recombination ($\rho_N = 0.4$) when selection intensity is weak ($\alpha_N = 0.5$).

The effect of neutral sequences between regions under selection: We studied whether or not small changes in the overall recombination rate *between* two regions

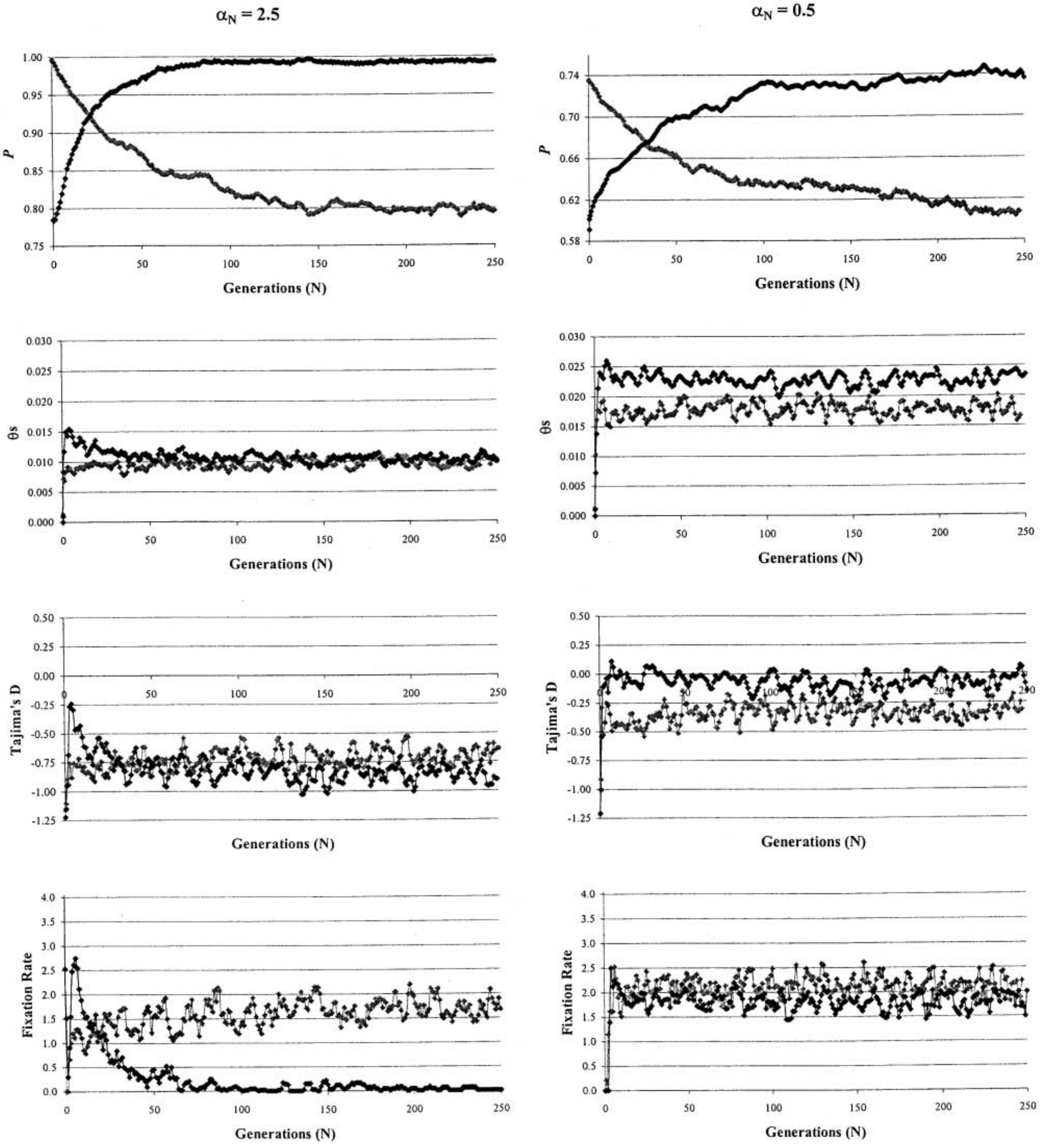

Figure 7.—Temporal dynamics in selected sequences across 250 *N* generations after a change of recombination environment (see text for details). Black lines indicate the change from total linkage ($\rho_N = 0$) to high recombination ($\rho_N = 0.4$) and gray lines from high recombination to total linkage. The frequency of preferred codons (*P*), polymorphism levels (θ_s, $n = 12$), Tajima's *D* ($n = 12$), and fixation rates are shown for $L = 2500$. Each value represents the average of 10 independent simulations.

under selection, caused only by a change in the physical distance between them, have a detectable effect on the overall IS. Here, the simulation procedure allowed us to generate a variable number of neutral sites (*i.e.*, middle or "spacer" sequence) between two sequences under selection. The two regions under selection were identical, with equal numbers of selected sites, selection coefficients per site, and mutation and recombination rates per site. Mutation and recombination rates per site in the spacer sequence are the same as in the flanking selected sequences, but the mutations were selectively neutral. Thus, with respect to IS the presence and length of the intermediate neutral region alters only the number of recombination events between the two selected

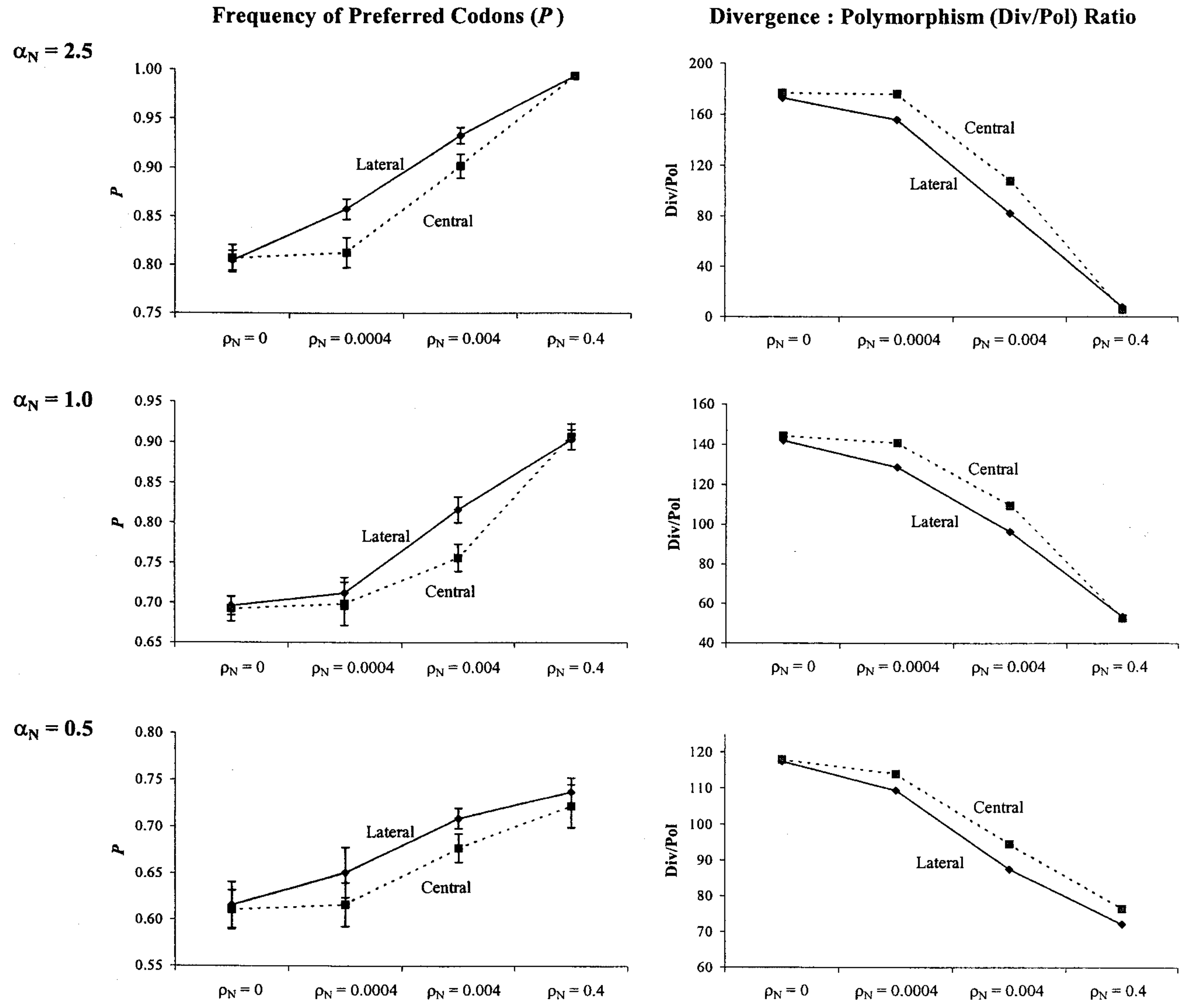

Figure 8.—Comparison between central (dashed line) and lateral (continuous line) regions of sequences under selection for different recombination rates (ρ_N) and scaled selection coefficients per site (α_N). Two indicators of IS are depicted: frequency of preferred codons (P) and divergence:polymorphism ratio (Div/Pol). The complete sequence under selection has 2500 sites and the lateral and central regions have 250 sites each. Selection coefficients and recombination and mutation rates are uniformly distributed across the entire sequence.

regions; it does not change directly any parameter on the flanking selected sequences.

We studied intermediate rates of recombination ($\rho_N = 0.004$ and $\rho_N = 0.0004$) for the case of a neutral sequence located between two sequences each of 500 selected sites. Figure 9A depicts the relative change of the effectiveness of selection (*i.e.*, α based on P) caused by the presence and length of the spacer sequence. The results show that the length of an intermediate neutral region has a detectable effect. In all cases, longer spacers lead to an increase of the effectiveness of selection (a reduction in IS) on the adjacent selected mutations. Serving as illustration, for the case of $\alpha_N = 1$ and $\rho_N = 0.004$ the presence of a 1000-bp-long region in the middle of the selected sequence is equivalent to a relative increase of 7.4% in the overall fitness associated with the selected sequences (*i.e.*, a gain of 2.3% preferred codons). A substantial fraction of the potential increment in fitness in regions of moderate to high recombination is achieved with short/intermediate sequences (<1000 bp), while for regions of more severely restricted recombination longer sequences are required to produce an equivalent increment in fitness. The maximum relative gain in fitness is higher for $\alpha_N = 2.5$ than for $\alpha_N = 1$ for the two rates of recombination investigated, as expected (see Figure 1).

Empirical tests of IS based on *D. melanogaster*'s genome

Distribution of codon bias within genes: As indicated in our simulations, IS is expected to be stronger in the center of regions under IS than in the margins of these regions. This leads to the first test prediction: *Codon usage bias, a measure of the effectiveness of selection, will be*

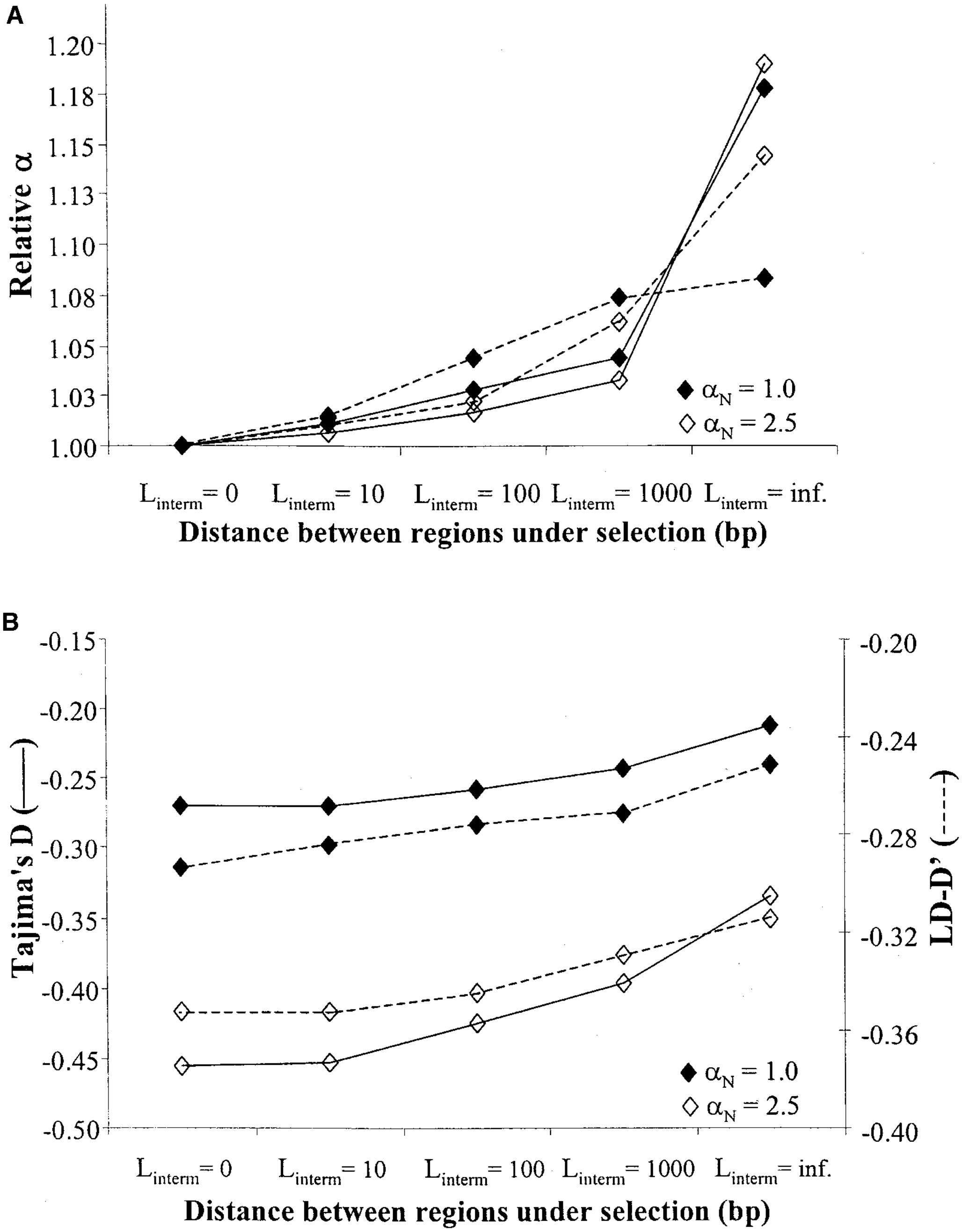

Figure 9.—Consequence on IS of the presence and length of an intermediate neutral sequence. Simulations are based on two sequences under selection of 500 sites each separated by an intermediate neutral sequence of L_{interm} sites (see text for details). Results are shown for $0 \leq L_{\mathrm{interm}} \leq 1000$ sites and for $L_{\mathrm{interm}} = \infty$ (inf.). Open and solid diamonds depict $\alpha_N = 2.5$ and $\alpha_N = 1.0$, respectively. (A) Selected region: relative effectiveness of selection (α) per site based on the frequency of preferred sites (*P*). Solid and dashed lines are $\rho_N = 0.0004$ and $\rho_N = 0.004$, respectively. (B) Neutral region: Tajima's *D* statistic (solid lines) and linkage disequilibrium *D′* (LD-*D′*; dashed lines) for $\rho_N = 0.0004$. $n = 12$.

lower in the middle of coding regions of genes than in the amino- or carboxy-terminal regions. Comparing codon bias levels within genes eliminates expression level and gene length as factors that can alter codon usage (Moriyama and Powell 1998; Comeron *et al.* 1999; Duret and Mouchiroud 1999). It also controls for any possible heterogeneity in mutational tendencies, either across the genome or associated with transcription rates. We have not attempted to control for heterogeneity in amino acid constraints within genes, but we expect this to obscure rather than to enhance the predicted IS effect (see also below).

We restricted our attention to the set of genes in the *D. melanogaster* genome composed of single long exons (>333 amino acids; see materials and methods), a total of 659 genes. The frequency of GC at the third position of codons (GC3) was used as a measure of codon usage bias (Shields *et al.* 1988; Akashi 1994, 1995; see materials and methods). For each gene, we measured GC3 in three sections of 100 codons, the first and last 100 codons (proximal and distal regions, respectively) and 100 codons in the middle of the gene (central region). As shown in Figure 10A, average GC3 frequencies differ significantly across the three regions (Friedman ANOVA test, $\chi^2 = 28.7$, $P < 1 \times 10^{-6}$), with a lower GC3 in the central region. A similar result is obtained when the average GC3 of the two lateral sections is compared to the central section ($\chi^2 = 26.0$, *P*

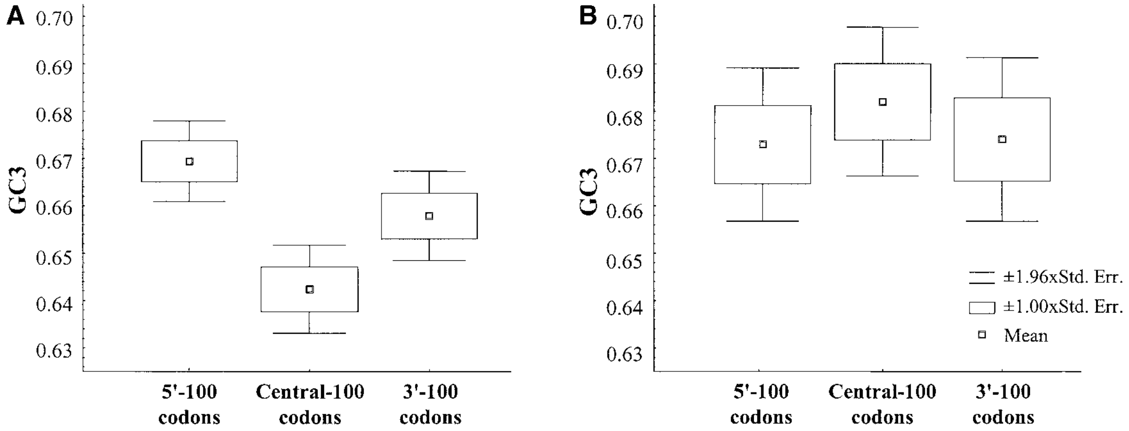

FIGURE 10.—GC content at the third position of codons (GC3), as a measure of codon bias, across the same gene in *D. melanogaster.* (A) Central and lateral (5′ and 3′) regions of 659 genes constituted by a single long exon (>333 amino acids). (B) Central and lateral (5′ and 3′) regions of 187 genes (>333 amino acids) containing only centrally located introns. Central and lateral regions are 100 codons long. The central region of genes without introns shows a significantly reduced GC3 and other measures of codon bias, compared to the lateral regions of the same genes (see text for details).

$< 1 \times 10^{-6}$) or when each lateral region is compared separately to the central section ($\chi^2 = 23.0$, $P < 1 \times 10^{-6}$ and $\chi^2 = 10.5$, $P = 0.001$, respectively). On average, the lower GC3 content in the central region of coding regions is equivalent to a reduction in α of ~10% on synonymous mutations compared to lateral regions.

IS simulations also show that the intensity of IS increases with the length of the gene region (and hence number of sites) subject to weak selection. This leads to the second test prediction: *The relative reduction in codon bias in the center of a gene will be positively correlated with the length of the coding region.* Consistent with this prediction, we find a highly significant positive correlation between the length of the coding region and the difference of GC3 between lateral and central regions in the same set of 659 genes analyzed above (Spearman's correlation $R = 0.207$, $P < 1 \times 10^{-6}$). Quantitatively similar results are obtained with the frequency of preferred codons and with GC content at fourfold degenerate sites; data not shown.

According to our simulations, the presence of neutrally evolving sequences placed in the center of a region subject to weak selection can relieve the IS effect. This leads to the third test prediction: *Centrally located introns will ameliorate the effect of IS in the central region of genes that contain them.* To test this prediction, we compared codon bias in these same 659 genes, which lack introns, with comparable genes with introns located in the center of the coding regions (see MATERIALS AND METHODS). Figure 10B shows the results for the 187 genes obtained from the genome database satisfying these criteria. For these genes, there is no apparent reduction of GC3 in the middle of the coding regions. Accordingly, we do not detect significant heterogeneity of GC3 between the three regions ($\chi^2 = 2.94$, $P > 0.20$) or a difference between the GC3 content of central and the two lateral regions ($P > 0.15$). In addition, the central sections of coding regions of genes with central intron(s) have a significantly higher GC3 content than the equivalent central sections in genes without introns (Mann-Whitney *U*-test, $P = 2 \times 10^{-6}$) whereas both lateral sections show similar frequencies ($P = 0.61$ and $P = 0.09$). Therefore, the lower GC3 frequency in the middle of the coding region in genes without introns cannot be the result of general relaxed selection on codon bias in the central part of coding regions.

Proportion of selected sites in a gene and codon bias: According to our simulations, the presence of neutral sequences embedded in a region under selection causes an increase in the effectiveness of selection on adjacent selected sequences, the length of such neutral sequences being positively correlated with the increment of the effectiveness of selection. We investigated, therefore, a fourth test prediction: *Codon bias will be positively correlated with measures of a gene's intron length and number.* As a first approximation, we studied the relationship between measures of codon bias (*e.g.*, GC3) and total intron length in the set of all genes with confirmed intron/exon structure. There is a weak positive relationship between GC3 and total intron length, both using all introns ($R = 0.040$, $P = 0.0007$) and after eliminating the small fraction of introns with detectable remnants of TE elements ($R = 0.041$, $P = 0.0005$).

We also studied the relationship between codon bias and measures of the proportion of sites under selection in a gene. As a simple measure of the *density* of sites under selection in a gene, we used the PLCR in a gene when embedded introns are included (see MATERIALS AND METHODS). The prediction under IS is again explicit: *Codon bias (as measured by GC3) will decrease as PLCR increases.* The analysis of all 7499 genes with introns reveals a significantly negative relationship between GC3 and PLCR ($R = -0.136$, $P < 1 \times 10^{-6}$); equivalent results are obtained using other measures of codon bias.

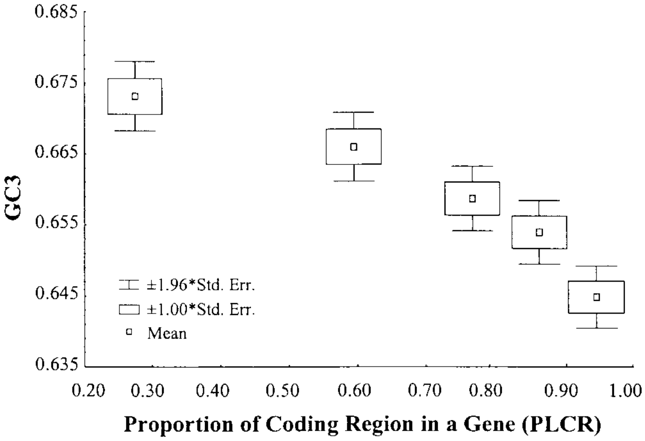

Figure 11.—Relationship between the proportion of the length of the coding region (PLCR) in a gene and GC3 in *D. melanogaster* (see materials and methods). The five PLCR classes divide the data set of 7499 genes into subsets of equivalent size ($n \approx 1500$ genes). The analysis of all genes with introns in *D. melanogaster* shows a significantly negative relationship between PLCR and GC3 (and other measures of codon bias; $R = -0.136$, $P < 1 \times 10^{-6}$; see text for details).

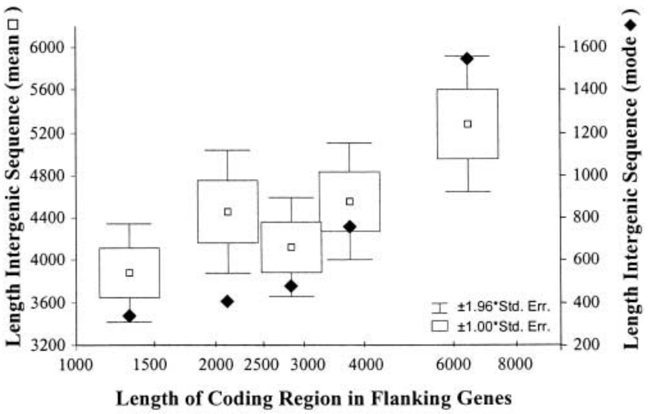

Figure 12.—Relationship between the total length of two adjacent coding regions and the size of the intergenic sequence between these two genes in *D. melanogaster*. The analysis of 6271 intergenic sequences flanked by well-defined genes (see text) shows a positive relationship between the length of these intergenic sequences and the length of flanking coding sequences ($R = 0.097$, $P < 1 \times 10^{-6}$).

Figure 11, a display of GC3 when genes are grouped with respect to PLCR into five sets of equal sample size, shows that the effect may be stronger when PLCR is medium/high.

Gene length may have a confounding effect on the relationship between GC3 and PLCR because the length of coding region is negatively related to codon bias (Moriyama and Powell 1998; Comeron *et al.* 1999; Duret and Mouchiroud 1999). Gene length is, for obvious reasons, positively correlated with PLCR ($R = 0.11$, $P < 1 \times 10^{-6}$). Multiple linear regression analysis corroborates that GC3 is negatively associated with PLCR (partial $r = -0.094$, $P < 1 \times 10^{-6}$) after controlling for the length of the coding region (multiple $r = 0.19$, $P < 1 \times 10^{-6}$). Thus, there is a significant negative relationship between GC3 and PLCR not attributable to gene length.

Gene length and intron presence: IS predictions of the favorable consequences of intermediate or spacer sequences forecast that *intron length will increase with the length of the coding region*. The average length of introns increases with the total length of the coding region ($R = 0.219$, $P < 1 \times 10^{-6}$). This relationship is not attributable to differences in either recombination rates or gene expression levels and remains significant ($P < 1 \times 10^{-6}$) after controlling for these variables. A greater number of introns are also observed in long genes ($R = 0.53$, $P < 1 \times 10^{-6}$) although this observation could be connected to causes other than IS.

Intergenic distance and gene length: In addition to having longer (and a greater number of) introns in relation to the length of a coding region, is there also evidence for *greater intergenic distance as a function of length of coding regions of adjacent genes*? This result is expected under a scenario where longer intergenic regions are favored when, otherwise, IS between adjacent genes would be enhanced, *i.e.*, when the lengths of the neighboring coding regions increase. To address this seventh test prediction, we investigated the length of intergenic regions separating well-defined genes (see materials and methods). The results, displayed in Figure 12, reveal a positive relationship between the length of the 6271 intergenic sequences investigated and the length of the flanking coding regions ($R = 0.097$, $P < 1 \times 10^{-6}$); a positive relationship is also observed in regions of high recombination ($>3 \times 10^{-8}$/bp/generation; $R = 0.104$, $P < 1 \times 10^{-6}$, $n = 2367$). Alternative explanations to this observation based on functional considerations might also be proposed, such as genes with longer coding regions, if they are functionally more complex, might require tighter gene regulation (and hence longer noncoding regions). But we are unaware of any explicit empirical support for this class of alternative explanations. If our interpretation of this correlation is correct, it would suggest that IS between adjacent genes might not be negligible in most of the range of recombination rates in Drosophila. This relationship is not an indirect consequence of the effect that recombination rates might have on both parameters: The length of the intergenic regions decreases with increasing recombination rates ($R = -0.034$, $P = 0.008$) but no relationship is detected between the length of coding regions and recombination ($P > 0.40$).

DISCUSSION

The Hill-Robertson effect, broadly defined, considers the reduction in the efficacy of selection as an indirect consequence of selection at a linked locus. This effect is generally interpreted as being equivalent to a reduction in N_e (Robertson 1961; Felsenstein 1974; Kliman and Hey 1993). Previous simulation studies on the evolutionary dynamics of tightly linked weakly selected mutations conform to this interpretation; *i.e.*, the relative strength of selection acting on a weakly selected mutation decreases from its parametric value as the number of additional linked mutations under selection increases (Comeron *et al.* 1999; McVean and Charlesworth 2000). But, as we show here, other consequences of this form of selection are not explicable by a reduction in N_e, and therefore they do not conform to general notions of the Hill-Robertson effect. Hence we favor the use of the term "interference selection" to distinguish it from other types of selection (see below). We also avoided using the term "hitchhiking" to describe IS because it is a term generally associated with strong positive selection.

Previous investigation of IS under plausible conditions of recombination, selection, and mutation allowed us to integrate two empirical observations about codon bias in Drosophila genes not easily explained by single-site models of selection (Comeron *et al.* 1999): (i) a positive relationship between codon bias and the level of nucleotide diversity in *D. melanogaster* (Moriyama and Powell 1996) and (ii) a negative relationship between gene length and both K_s and codon bias (Comeron and Aguadé 1996; Powell and Moriyama 1997; Moriyama and Powell 1998; Comeron *et al.* 1999; Duret and Mouchiroud 1999). The present study delves deeper into consequences of IS on the efficacy of selection, and it also considers its effects on linked neutral variability. We also use IS theory as a guide to discover new regularities in the genomic architecture of Drosophila. These genomic features can be specifically understood as consequences of IS.

IS and its evolutionary consequences: Polymorphism levels in the selected sequence (θ_s) are, in general terms, decreased by either increasing the number of selected sites or reducing recombination. When selection increases, θ_s becomes less affected by changes in recombination rates while linked neutral polymorphism (θ_n) and other parameters such as codon bias or rates of fixation vary substantially. This more modest response of θ_s to an increment of IS when selection increases is not surprising because the expected net reduction of θ_s due to smaller N_e under SS-MSD also decreases when selection increases (see results).

IS reduces polymorphism levels in adjacent neutral sequences and the effect increases with increasing any parameter that contributes to IS. The effect that IS has on θ_n is maximum for total linkage but it is also measurable when recombination occurs. Therefore, IS may be a contributing factor in the reduction of neutral polymorphism levels in regions of low recombination observed in a variety of organisms. But it is unlikely that IS alone can cause extreme reduction in levels of polymorphism in regions of low recombination. When IS is greatest (*i.e.*, complete linkage and large L), θ_s and θ_n may become similarly reduced, making the distinction between selected and linked neutral sequences uncertain.

Consistent with previous studies and with the idea that IS reduces N_e and the efficacy of selection, we find that polymorphism levels (both θ_s and θ_n) decrease and the fixation rates of weakly selected mutations increase with IS. But IS also increases the skew in the frequency spectrum of mutations under weak selection (*i.e.*, increasing the proportion of low-frequency variants). Furthermore, the study of linked neutral sequences shows that IS also creates a skew in the frequency spectrum of neutral mutations away from the equilibrium neutral distribution. These results reveal complexities in the evolutionary dynamics of IS that cannot be rationalized as being equivalent to a reduction in N_e in standard formulations of weak selection at equilibrium.

IS predicts a skew of allele frequencies away from the neutral equilibrium and to lower, nonpolarized, frequencies for both selected and linked neutral mutations. In common with the HH and pHH models, this skew will be most discernible in genomic regions with reduced recombination, causing a positive correlation between Tajima's *D* and rates of recombination (Braverman *et al.* 1995; Gillespie 2000). IS may, therefore, be a contributing factor to the observed excess of rare variants in regions of low recombination in *D. melanogaster* (Langley *et al.* 2000; Andolfatto and Przeworski 2001).

Hill and Robertson (1966) indicated that the smaller the recombination rate between selected sites, the greater the negative linkage disequilibrium (LD observed in repulsion associations; see McVean and Charlesworth 2000). Here, we studied the extent of this LD in selected and adjacent neutral sequences. In the selected sequences, LD becomes more negative as IS increases (increasing the number of weakly selected sites or strength of selection or reducing recombination). More informative are the results showing that neutral mutations adjacent to the mutations under selection also show increments in negative LD with the same causes of IS (most noticeable with increasing the number of adjacent sites under selection). In fact, the magnitudes of LD in selected and adjacent neutral sequences become nearly the same when IS is maximum. Moreover, the study of LD within frequency classes has allowed us to remove most of the confounding effect of allele frequency on measures of LD. These results are indicative of the allele perturbation or "traffic" phenomenon caused by IS. The magnitude of LD caused by IS is, however, expected to decrease faster than directly predicted by an increase in recombination rates alone because IS itself also decreases with recombination.

Andolfatto and Przeworski (2000) reported a genome-wide departure from the standard neutral model in *D. melanogaster* and *D. simulans*, with greater intralocus LD than expected, as measured by C_{hud} (Hudson 1987). This observation could not be explained by either the recombination or mutation rates found across the species' genomes or by current theories of selection and linkage based on strong selection (BGS and HH models; see Andolfatto and Przeworski 2000 for details). Models of selection with strongly favorable alleles increasing only to intermediate frequency (Hudson *et al.* 1994), a multilocus model with epistatic selection involving secondary structures of pre-mRNA (Kirby and Stephan 1995), or balancing selection at closely linked loci (Slatkin 2000) may all generate regional increases in LD as well as a high variance of the number of pairwise differences, on which C_{hud} is based, causing low estimates of C_{hud}. These scenarios, however, are not likely to be common as indicated by the fact that high/intermediate frequency variants (*i.e.*, positive Tajima's *D*) would be most conspicuous in regions of low recombination, which is contrary to the observations in *D. melanogaster* (see above). The simulation results of the IS model, revealing an allele perturbation or traffic scenario, together with the fact that consequences of IS are detected across *D. melanogaster*'s genome, point out that IS might also contribute to a general high variance in pairwise differences in this species.

In general, the application of the SS-MSD model to estimate the compound parameter α ($N_{\mathrm{e}}s$) when IS is present will lead to an underestimate of α for a given level of polymorphism or rate of divergence but it will lead to an overestimate of α on the basis of the frequency spectrum of mutations. Consequences of IS are also expected to be highly heterogeneous among genes and across genomic regions, on the basis of differences in recombination rates, gene densities, gene sizes, and gene structures. Hence, IS would cause large variances in many population and evolutionary parameters. This heterogeneity across genomes may be useful for differentiating IS from demographic causes, for which more homogeneously distributed effects are expected.

Variability in the intensity of IS has population genetics and evolutionary consequences that can easily be misinterpreted as indicating differences in selective regimes among genes. IS causes an increase in the rate of divergence of mutations under MSD, where the stronger the selection ($0.25 \leq \alpha_N \leq 2.5$), the more conspicuous the effect of IS on the rate of divergence. As a result, differences in rates of substitution, both K_{s} and K_{a} (and $K_{\mathrm{a}}/K_{\mathrm{s}}$ ratio), between genes can potentially be explained by variable IS with constant selection. Equivalently, variable IS will alter Div/Pol ratios without requiring differences in selection coefficients.

Temporal dynamics after a change of recombinational environment: Gene rearrangements within chromosomal arms are a recurrent characteristic of Drosophila micro- and macroevolution. Because recombination is not homogeneously distributed along Drosophila chromosomes (Lindsley and Sandler 1977; Ashburner 1989; Aquadro *et al.* 1994; True *et al.* 1996; Hamblin and Aquadro 1999; Takano-Shimizu 2001), a change in the chromosomal position is expected to change the recombination environment (Charlesworth 1994). Moreover, in Drosophila there is evidence of severe differences in the recombination rates of homologous chromosomal regions even between closely related species [*e.g.*, within the *D. melanogaster* group (True *et al.* 1996; Takano-Shimizu 1999, 2001)].

Our simulations indicate that the period of nonequilibrium base composition (codon usage) after a drastic change in recombination environment may be 100–$250N_{\mathrm{e}}$ generations (*i.e.*, equivalent to >10–25 mya in most Drosophila species) when $\beta_N = 0.01$. This suggests that nonequilibrium codon usage caused by frequent change in recombination rates may be the norm rather than the exception, at least in Drosophila evolution. Furthermore, the period required to reach the new (multisite) base composition equilibrium is expected to be longer for genes undergoing a reduction in codon bias than for those in which it is increasing. The nonequilibrium scenario is apparent in analyses of fixed synonymous mutations along *D. melanogaster/D. simulans* lineages, with codon bias declining both in *D. melanogaster* and, to a lesser degree, in *D. simulans* (Akashi 1996; Begun 2001). Also, genes located in genomic regions where crossing over is now severely attenuated in *D. melanogaster* show a reduction in codon bias not as extreme as expected on the basis of the highly reduced levels of neutral polymorphisms and estimated N_{e} (Comeron *et al.* 1999). Altogether, these observations are in agreement with a report of a recent reduction in crossover frequency in the *D. melanogaster/D. simulans* lineages, after the evolutionary split from *D. yakuba*. (Takano-Shimizu 1999, 2001), with at least a fraction of genes located in regions of very low recombination in *D. melanogaster* not yet at codon usage equilibrium (Comeron *et al.* 1999).

Population and evolutionary parameters, such as polymorphism levels and their frequencies or rates of evolution, reach estimates representing the new equilibrium faster than those of codon usage (Akashi 1996; Akashi and Schaeffer 1997; Kliman 1999). However, these population parameters might tend to overshoot the new equilibrium values when a substantial reduction in IS occurs after changing from low to high recombination. In this intermediate phase after a sharp increment in recombination, overall polymorphism levels and frequencies of mutations may be closer to neutral expectations than either are to precedent and future equilibrium levels, suggestive of a reduction of the effectiveness of selection, although the opposite might be the case.

The results also have a practical implication when estimates of the compound parameter α are estimated from divergence data. Estimates of α using divergence data of weakly selected mutations will tend to indicate their lower boundaries, mostly suggesting a shrinking N_e in nearly all lineages because periods with reduced recombination will contribute most of the substitutions. This effect will be in addition to the previously described general underestimation of α based on rates of fixation because of IS. The observed strong and relatively fast effects that changing recombination rates have on the rates of fixation of weakly selected mutations are congruent with the suggestion (Charlesworth 1994; Comeron *et al.* 1999) that changes in recombination environment may account for the high variance of substitutions rates of these mutations in Drosophila (Zeng *et al.* 1998).

Heterogeneous effect of IS across selected regions: The magnitude of IS is expected to be heterogeneously distributed across regions under uniform selection intensity and mutation rate as a consequence of the different number of weakly selected sites surrounding each site (*i.e.*, *density* of selected sites within a genetic distance). This prediction was confirmed using simulations that show that consequences of IS are stronger in the central regions of sequences under selection compared to lateral regions, and the effect, we believe, is empirically detectable in Drosophila genes (see below).

Enhanced IS in central regions of sequences under selection has the consequence that many population and evolutionary parameters will be also heterogeneously distributed across sequences. These spatial differences might be incorrectly interpreted as indicating variable selective regimes although constancy (spatial and temporal) in selection coefficients might be the case. Two examples can be briefly given. First, heterogeneously distributed IS across sequences will generate Div/Pol values higher in the central regions of sequences under constant selection, suggestive of past action of positive selection or relaxed constraints in these central regions. Second, stronger IS will also tend to generate higher substitution rates and vary K_a/K_s ratios in central regions of genes compared to the edges of genes and this can easily be misinterpreted as variable selective constraints across the gene.

U-shaped distribution of codon bias across long exons in *D. melanogaster*: Long undisrupted coding sequences in *D. melanogaster* have central sections with significantly reduced codon bias compared to lateral sections of the same coding region (Figure 9). The study and comparison of central and lateral sections of the same gene allow us to exclude many factors implicated as drivers of differential codon bias, including gene expression, gene length, genomic recombinational environment, and possible mutational differences associated with recombination or transcription rates. The comparison of codon bias in central regions in genes without introns and in genes with introns centrally located further allows us to rule out general relaxed constraints on codon bias in the central parts of proteins.

This observation fits with the outcome of the simulations that produce a U-shaped distribution of the effectiveness of selection and codon bias across sequences under selection, caused by stronger IS in central regions (Figure 4), and it is not predicted by other models of codon bias. For instance, models of selection on codon bias due to translational accuracy (Bulmer 1991; Akashi 1994; Eyre-Walker 1996) predict, if anything, an increase in codon bias with amino acid position and thus a J-shaped codon bias distribution. Conflicting selection pressures close to the start of the genes also generate a J-shaped distribution (Eyre-Walker and Bulmer 1993; Kliman and Eyre-Walker 1998). Therefore, we suggest that IS plays a significant role in shaping the effectiveness of selection on codon bias in Drosophila, not only among genes but also along coding sequences. Moreover, in accord with the IS hypothesis, the difference between central and lateral regions of the same coding region increases with the length of the coding region ($P < 1 \times 10^{-6}$). This difference is also significant in genes located in regions with the highest recombination in *D. melanogaster* (*e.g.*, $>3 \times 10^{-8}$/bp/generation; Friedman ANOVA, $\chi^2 = 7.28$, $P = 0.007$; $n = 231$). This last observation supports the proposal that IS might be detectable in genes across the entire range of recombination in *D. melanogaster* (Comeron *et al.* 1999), with simulation results revealing consequences of IS on codon bias even when the recombination rate is higher than that expected in this species (Comeron *et al.* 1999; McVean and Charlesworth 2000).

Our analysis of genes without introns confirms a prior report indicating a higher GC3 in 5′ sections of coding regions of *D. melanogaster* genes than 3′ sections (Kliman and Eyre-Walker 1998). This trend may be explained by variation in mutational tendencies or by biased mismatch repair after gene conversion events (Kliman and Eyre-Walker 1998). This last possibility could be associated with a relationship between transcription mechanism and initiation of meiotic recombination (Nicolas 1998). This relationship, if present in *D. melanogaster*, forecasts—under the IS model—a declining effective rate of recombination along genes (hence of effectiveness of selection and codon bias) when the distribution of meiotic gene conversion tract length (Hilliker *et al.* 1994) is taken into account.

Neutral regions as modifiers of recombination between selected regions: Granted that the evolutionary consequence of neutral intermediate sequences as modifiers of recombination is very small for plausible lengths (*i.e.*, <1000 bp), simulations show that under realistic rates of selection, recombination, and mutation, the presence of neutral intermediate (or spacer) sequences may have a measurable effect on the overall magnitude of IS in adjacent sequences under selection. Even very small increments in the number of recombination events between two regions under selection, obtained by increasing the physical distance between the two selected regions, can reduce IS when the recombination rate (per physical unit) is moderate/low. This reduction in IS instigates an increase of the effectiveness of selection together with the decline of all properties associated with allele perturbation or traffic. Therefore, in regions of reduced recombination, reasonably long intermediate sequences may be favored as a counterbalance to the reduced effectiveness of selection caused by tight linkage. Congruent with the simulation results, in *D. melanogaster* the presence of introns is associated with an increase in the effectiveness of selection. This result is observed using either the absolute length of introns or a measure of the relative length of introns, taking into account the length of the coding region. The difference in codon bias in central regions of coding regions between genes with introns centrally located and genes without introns also supports this interpretation.

Thus, genomic data in *D. melanogaster* support the hypothesis that the presence and length of "junk" DNA between clusters of selected sites may itself be a selective trait (Comeron and Kreitman 2000; Comeron 2001). These results are also congruent with the positive relationship between average intron length and the length of the coding region reported here and the observed negative relationship between intron length and recombination across *D. melanogaster* and human genomes that could not be explained by the presence of transposable elements or by mutational differences (Carvalho and Clark 1999; Comeron and Kreitman 2000). The selective advantage of neutral sequences embedded in sequences under selection to reduce IS may be one of the forces that oppose the apparent mutational deletion bias present in Drosophila (Petrov *et al.* 1996; Petrov and Hartl 1998; Comeron and Kreitman 2000).

Whether IS is restricted mainly to intervals containing single genes (both coding and regulatory regions) or, conversely, whether neighboring genes have detectable effects on each other will depend on the effective recombination rate between genes (involving physical distance and recombination rate per site) and gene lengths. Because the distance between genes in most eukaryotic species is usually several kilobases, IS between most adjacent genes will likely be negligible except for species/genomic regions with very low recombination rates. However, the results showing a positive relationship between intergenic distance and the length of the flanking coding regions suggest that, in *D. melanogaster*, IS may be influencing the size of intergenic sequences in some instances. Under this perspective, the study of IS and of the evolution of recombination and their effects on the effectiveness of selection should also incorporate the enormous plasticity that genomes have, involving gene structure, intron size, and gene density. Our studies on IS suggest that the apparent lack of biological function associated with many intronic or intergenic sequences might not always imply that they are devoid of evolutionary function.

Conclusions: Mutations of weak selective effect, when they are sufficient in number and are tightly linked, reduce the overall efficacy of selection (*e.g.*, cause an increase in the fixation rate of selected mutations) and, under the model we investigated, the consequences include a reduction of polymorphism. We found that the reduction of polymorphism extends not only to the sites under selection but to linked neutral mutations as well. These effects are similar to those seen in single-site models of selection, such as MSD, when N_e is reduced. Other consequences of IS, however, cannot be easily related to simpler models of selection, and perhaps these represent the unique signatures of IS. They include the increase in occurrence of rare alleles and negative linkage disequilibrium in both selected and linked neutral mutations.

The discovery of a genome-wide relationship between noncoding polymorphism levels and the recombination rate in Drosophila stimulated the investigation of models involving common forms of natural selection and the influence these forms of selection have on linked neutral variability. As with definitely deleterious mutation and BGS, weakly selected mutation and IS are also likely to be omnipresent throughout a species' genome. Both, therefore, have the potential for explaining variation in polymorphism levels associated with recombination rates. Likely, IS is distinguishable from models involving strong selection (BGS and HH/pHH models) in that consequences of clusters of weakly selected sites may have a much finer and patchier distribution across genomes than those caused by definitively selected mutations, and linked neutral variability will be reduced under IS only in small regions surrounding these clusters. Complete genome polymorphism studies of Drosophila, similar to those contemplated in humans, may allow us to distinguish IS effects from those caused by strongly selected mutations.

Perhaps the most exciting findings presented here are the empirical tests of IS. The seven predictions we generated for testing IS are, we believe, highly specific to IS and therefore are strong tests of this theory. Indeed the regularities we discovered in Drosophila genome architecture were investigated *because* we had a theory that led us to their predicted existence. Having made these discoveries now, we hope these genome-wide patterns stimulate the search for alternative explanations. *A priori* one might have thought that the attempt to find empirical support for a theory about weakly interacting mutations might be confined to the population genetic realm. Here we show that genome features may also be highly relevant to our population genetic theory. What this means is that genome features that have previously been attributed to ancient events and accidents of history may actually be features retained and sculpted by a common form of selection, underscoring the hidden treasures present in genome data. This trend—using genome-wide data to investigate population genetic models of selection—will only accelerate as additional species are sequenced.

We thank R. Hudson and A. Llopart for helpful discussions and suggestions. We also thank the anonymous reviewers and N. Takahata for useful comments. This research was supported by a National Institutes of Health grant GM-39355 to M.K.

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.

Aguadé, M., and C. H. Langley, 1994 Polymorphism and divergence in regions of low recombination in Drosophila, pp. 67–76 in *Non-Neutral Evolution*: *Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

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

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

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

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster.* Genetics **144:** 1297–1307.

Akashi, H., and S. W. Schaeffer, 1997 Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics **146:** 295–307.

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics **148:** 1397–1399.

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

Andolfatto, P., and M. Przeworski, 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster.* Genetics **158:** 657–665.

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination and DNA polymorphism in Drosophila, pp. 46–56 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., 2001 The frequency distribution of nucleotide variation in *Drosophila simulans.* Mol. Biol. Evol. **18:** 1343–1352.

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.

Bennetzen, J. L., and B. D. Hall, 1982 Codon selection in yeast. J. Biol. Chem. **257:** 3026–3031.

Birky, C. W., and J. B. Walsh, 1988 Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. USA **85:** 6414–6418.

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.

Bulmer, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

Carvalho, A. B., and A. G. Clark, 1999 Intron size and natural selection. Nature **401:** 344.

Charlesworth, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63:** 213–227.

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.

Comeron, J. M., 2001 What controls the length of noncoding DNA? Curr. Opin. Genet. Dev. **11:** 652–659.

Comeron, J. M., and M. Aguadé, 1996 Synonymous substitutions in the *Xdh* gene of Drosophila: heterogeneous distribution along the coding region. Genetics **144:** 1053–1062.

Comeron, J. M., and M. Kreitman, 1998 The correlation between synonymous and nonsynonymous substitutions in Drosophila: mutation, selection, or relaxed constraints? Genetics **150:** 767–775.

Comeron, J. M., and M. Kreitman, 2000 The correlation between intron length and recombination in Drosophila: dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175–1190.

Comeron, J. M., M. Kreitman and M. Aguadé, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory.* Harper & Row, New York.

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

Dvorák, 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.

Ewens, W. J., 1979 *Mathematical Population Genetics.* Springer-Verlag, Berlin.

Eyre-Walker, A., 1996 Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol. Biol. Evol. **13:** 864–872.

Eyre-Walker, A., and M. Bulmer, 1993 Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. **21:** 4599–4603.

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

Gallego, P., E. Juan and M. Papaceit, 1999 Chromosomal homologies between *Drosophila melanogaster* and *D. funebris* determined by in-situ hybridization. Chromosome Res. **7:** 331–339.

Gillespie, J. H., 1997 Junk ain't what junk does: neutral alleles in a selected context. Gene **205:** 291–299.

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

Golding, G. B., 1997 The effect of purifying selection on genealogies, pp. 271–285 in *Progress in Population Genetics and Human Evolution*, edited by P. Donnelly and S. Tavare. Springer-Verlag, New York.

Grantham, R., C. Gautier, M. Gouy, M. Jacobzone and R. Mercier, 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9:** R43–74.

Grosjean, H., and W. Fiers, 1982 Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18:** 199–209.

Hamblin, M. T., and C. F. Aquadro, 1999 DNA sequence variation and the recombinational landscape in *Drosophila pseudoobscura*: a study of the second chromosome. Genetics **153:** 859–869.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on the limits to artificial selection. Genet. Res. **8:** 269–294.

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

Hilliker, A. J., and A. Chovnick, 1981 Further observations on intragenic recombination in *Drosophila melanogaster.* Genet. Res. **38:** 281–296.

Hilliker, A. J., G. Harauz, A. G. Reaume, M. Gray, S. H. Clark and A. Chovnick, 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 adaptation and divergence during speciation within the *Drosophila melanogaster* species complex. Evolution **48:** 1900–1913.

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

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

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.

Ikemura, T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151:** 389–409.

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.

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

Kirby, D. A., and W. Stephan, 1995 Haplotype test reveals departure from neutrality in a segment of the white gene of *Drosophila melanogaster*. Genetics **141:** 1483–1490.

Kliman, R. M., 1999 Recent selection on synonymous codon usage in Drosophila. J. Mol. Evol. **49:** 343–351.

Kliman, R. M., and A. Eyre-Walker, 1998 Patterns of base composition within the genes of Drosophila melanogaster. J. Mol. Evol. **46:** 534–541.

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

Kraft, T., T. Sall, I. Magnusson-Rading, N. O. Nilsson and C. Hallden, 1998 Positive correlation between recombination rates and levels of genetic variation in natural populations of sea beet (*Beta vulgaris* subsp. *maritima*). Genetics **150:** 1239–1244.

Kress, H., 1993 The salivary gland chromosomes of *Drosophila virilis*: a cytological map, pattern of transcription and aspects of chromosome evolution. Chromosoma **102:** 734–742.

Kurland, C. G., 1987 Strategies for efficiency and accuracy in gene expression. 1. The major codon preference: a growth optimization strategy. Trends Biochem. Sci. **12:** 126–128.

Langley, C. H., Y. N. Tobari and K. I. Kojima, 1974 Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics **78:** 921–936.

Langley, C. H., J. Macdonald, N. Miyashita and M. Aguadé, 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism at the *suppressor of forked* region in *Drosophila melanogaster* and *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **90:** 1800–1803.

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.

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

Lewontin, R. C., 1964 Interaction of selection + linkage. I. General considerations—heterotic models. Genetics **49:** 49–67.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics **120:** 849–852.

Lewontin, R. C., and K. Kojima, 1960 The evolutionary dynamics of complex polymorphisms. Evolution **14:** 458–472.

Li, W.-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. **24:** 337–345.

Lindsley, D. L., and L. Sandler, 1977 The genetic analysis of meiosis in female *Drosophila melanogaster*. Philos. Trans. R. Soc. Lond. B Biol. Sci. **277:** 295–312.

Lozovskaya, E. R., D. A. Petrov and D. L. Hartl, 1993 A combined molecular and cytogenetic approach to genome evolution in Drosophila using large-fragment DNA cloning. Chromosoma **102:** 253–266.

Ludwig, M., N. Patel and M. Kreitman, 1998 Functional conservation of *even-skipped* stripe 2 enhancer in Drosophila. Development **125:** 949–958.

Martín-Campos, J. M., J. M. Comerón, N. Miyashita and M. Aguadé, 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. Genetics **130:** 805–816.

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

McVean, G. A., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929–944.

McVean, G. A., and J. Vieira, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157:** 245–257.

Moriyama, E. N., and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes in Drosophila. Genetics **134:** 847–858.

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

Moriyama, E. N., and J. R. Powell, 1998 Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. **26:** 3188–3193.

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

Nachman, M. W., V. L. Bauer, S. L. Crowell and C. F. Aquadro, 1998 DNA variability and recombination rates at X-linked loci in humans. Genetics **150:** 1133–1141.

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics **145:** 519–534.

Nicolas, A., 1998 Relationship between transcription and initiation of meiotic recombination: toward chromatin accessibility. Proc. Natl. Acad. Sci. USA **95:** 87–89.

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

Ohta, T., 1995 Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. **40:** 56–63.

Ohta, T., and M. Kimura, 1971 On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. **1:** 18–25.

Papaceit, M., and A. Prevosti, 1989 Differences in chromosome A arrangement between Drosophila madeirensis and Drosophila subobscura. Experientia **45:** 310–312.

Perlitz, M., and W. Stephan, 1997 The mean and variance of the number of segregating sites since the last hitchhiking event. J. Math. Biol. **36:** 1–23.

Petrov, D. A., and D. L. Hartl, 1998 High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol. Biol. Evol. **15:** 293–302.

Petrov, D. A., E. R. Lozovskaya and D. L. Hartl, 1996 High intrinsic rate of DNA loss in Drosophila. Nature **384:** 346–349.

Pinsker, W., and D. Sperlich, 1984 Cytogenetic mapping of enzyme loci on chromosomes *J* and *U* of *Drosophila subobscura*. Genetics **108:** 913–926.

Powell, J. R., and E. N. Moriyama, 1997 Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94:** 7784–7790.

Przeworski, M., B. Charlesworth and J. D. Wall, 1999 Genealogies and weak purifying selection. Mol. Biol. Evol. **16:** 246–252.

Przeworski, M., R. R. Hudson and A. Di Rienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16:** 296–302.

Robertson, A., 1961 Inbreeding in artificial selection programmes. Genet. Res. **2:** 189–194.

Segarra, C., and M. Aguadé, 1992 Molecular organization of the X chromosome in different species of the obscura group of Drosophila. Genetics **130:** 513–521.

Sharp, P. M., and W.-H. Li, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

Sharp, P. M., and W.-H. Li, 1989 On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28:** 398–402.

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

Slatkin, M., 2000 Balancing selection at closely linked, overdominant loci in a finite population. Genetics **154:** 1367–1378.

Stephan, W., 1994 Effects of genetic recombination and population subdivision on nucleotide sequence variation in *Drosophila ananassae*, pp. 57–66 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

Stephan, W., and C. H. Langley, 1989 Molecular genetic variation

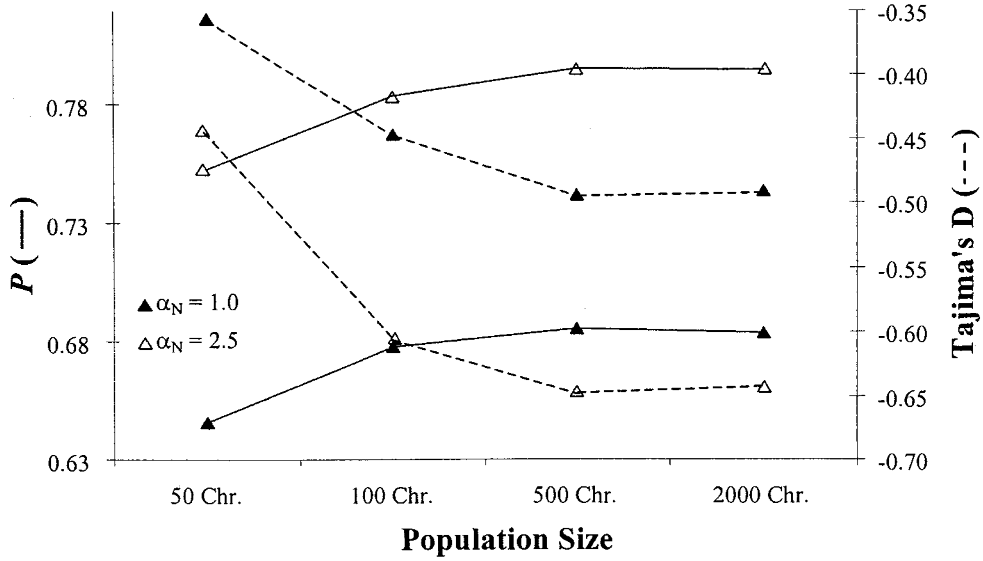

Figure A1.—Effect of the size of simulated populations on evolutionary consequences of interference selection (IS) due to weakly selected mutations. Results are shown for $L = 2500$ completely linked sites. Solid lines depict the frequency of preferred sites (P) and dashed lines depict Tajima's D ($n = 12$). Solid and open triangles show the results for $\alpha_N = 1$ and $\alpha_N = 2.5$, respectively.

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.

Stephan, W., and C. H. Langley, 1998 DNA polymorphism in *Lycopersicon* and crossing over per physical length. Genetics **150:** 1585–1593.

Tachida, H., 2000 Molecular evolution in a multisite nearly neutral mutation model. J. Mol. Evol. **50:** 69–81.

Takano-Shimizu, T., 1999 Local recombination and mutation effects on molecular evolution in Drosophila. Genetics **153:** 1285–1296.

Takano-Shimizu, T., 2001 Local changes in GC/AT substitution biases and in crossover frequencies on Drosophila chromosomes. Mol. Biol. Evol. **18:** 606–619.

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.

Whiting, J. H. Jr., M. D. Philey, J. L. Farmer and D. E. Jeffery, 1989 *In situ* hybridization analysis of chromosomal homologies in *Drosophila melanogaster* and *Drosophila virilis.* Genetics **122:** 99–109.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Wright, S., 1938 Size of population and breeding structure in relation to evolution. Science **87:** 430–431.

Zeng, L.-W., J. M. Comeron, B. Chen and M. Kreitman, 1998 The molecular clock revisited: the rate of synonymous vs. replacement change in Drosophila. Genetica **102/103:** 369–382.

Communicating editor: N. Takahata

APPENDIX: EFFECT OF SMALL NUMBER OF SIMULATED INDIVIDUALS ON ESTIMATES OF IS

The size of the simulated populations is usually very small compared to the actual natural populations due to computational time constraints when a large number of mutable sites and/or recombination events are under study. According to diffusion theory of weak selection, equivalent equilibria are expected for different numbers of individuals (or chromosomes) as long as the products Ns (α) and $N\mu$ (β) are kept constant and $\beta \ll 1$ (Crow and Kimura 1970; Ewens 1979; Li 1987). However, the use of very small populations might alter the complex interactions between linkage, weak selection, and drift that cause IS. Also, as Tachida (2000) pointed out, some sample and population statistics are affected by the study of samples that represent a large fraction of the total population. Thus, the effect of using small simulated populations on population statistics has been studied quantitatively for two extreme cases: under neutrality and when IS due to weak selection on linked selected sites occurs.

Following Gillespie (2000), we can use Tajima's D statistic as a way of reducing several population parameters (or properties of the coalescent) to a single parameter. The study of Tajima's D under neutrality using sample sizes (n) of 10, 20, and 40 sequences when the number of chromosomes in the population ($2N$) ranges between 100 and 2000 shows that the larger the fraction of sampled sequences, the greater the deviation from the neutral expectation [E(Tajima's D) $= 0$]. This deviation generates positive Tajima's D estimates, noticeable for samples involving $\geq$5–10% of the population. As Tachida (2000) indicated, estimates of heterozygosity based on the number of segregating sites (θ) are also underestimated when the sample represents a large fraction of the total population and N is not very large (*i.e.*, <250 diploid individuals).

Figure A1 shows the effect of small populations on the study of IS for the case of $L = 2500$ linked sites under weak selection. Two parameters used to evaluate IS (see text) are depicted: the frequency of preferred codons (P) and Tajima's D. Under the applied selective model (semidominance and multiplicative over sites; see materials and methods), the effects of IS are quantitatively altered by the use of very small populations, causing the tendency to overestimate the reduction of the effectiveness of selection due to IS. In particular, small populations generate sequences with smaller

P and selected mutations segregate at frequencies closer to those expected under neutrality. These trends are less conspicuous, or even absent, when the causes of IS are reduced (*e.g.*, $L < 500$).

Altogether, these results indicate the need for generating large population sizes to obtain a precise picture of the outcome of subtle interactions between drift, multilocus selection, linkage, and mutation. The population size, N, required for studies of IS should represent a compromise between accuracy of results and pragmatic simulation times. Note that the computational time needed to study populations near equilibrium may increase exponentially with N and the failure to allow such a period of time might produce imprecise or biased outcomes. The population size should be adapted to the sample size, mutation rate, number of sites under selection, selection coefficients, and likely the selective scheme to be scrutinized.