Linkage Disequilibrium Patterns Across a Recombination Gradient in African *Drosophila melanogaster*

Peter Andolfatto*,†,1 and Jeffrey D. Wall‡,§

*Department of Zoology, University of Toronto, Toronto, Ontario M5S 3G5, Canada, †Institute of Cell, Animal and Population Biology,
University of Edinburgh, Edinburgh EH9 3JT, United Kingdom, ‡Program in Molecular and Computational Biology,
University of Southern California, Los Angeles, California 90089 and §Department of Human Genetics,
University of Chicago, Chicago, Illinois 60637

ABSTRACT

Previous multilocus surveys of nucleotide polymorphism have documented a genome-wide excess of intralocus linkage disequilibrium (LD) in *Drosophila melanogaster* and *D. simulans* relative to expectations based on estimated mutation and recombination rates and observed levels of diversity. These studies examined patterns of variation from predominantly non-African populations that are thought to have recently expanded their ranges from central Africa. Here, we analyze polymorphism data from a Zimbabwean population of *D. melanogaster*, which is likely to be closer to the standard population model assumptions of a large population with constant size. Unlike previous studies, we find that levels of LD are roughly compatible with expectations based on estimated rates of crossing over. Further, a detailed examination of genes in different recombination environments suggests that markers near the telomere of the X chromosome show considerably less linkage disequilibrium than predicted by rates of crossing over, suggesting appreciable levels of exchange due to gene conversion. Assuming that these populations are near mutation-drift equilibrium, our results are most consistent with a model that posits heterogeneity in levels of exchange due to gene conversion across the X chromosome, with gene conversion being a minor determinant of LD levels in regions of high crossing over. Alternatively, if levels of exchange due to gene conversion are not negligible in regions of high crossing over, our results suggest a marked departure from mutation-drift equilibrium (*i.e.*, toward an excess of LD) in this Zimbabwean population. Our results also have implications for the dynamics of weakly selected mutations in regions of reduced crossing over.

T HE parameters θ $(= 4N_e\mu)$, where N_e is the species effective population size and μ is the mutation rate per generation, and ρ $(= 4N_e r)$, where r is the sex-averaged crossing-over rate per generation, influence the amount and the pattern of genomic sequence variability. Under a Wright-Fisher neutral model, hereafter the standard neutral model, these two parameters are both proportional to N_e. N_e cannot be estimated directly from polymorphism data; instead, it must be estimated indirectly using estimates of θ and μ or estimates of ρ and r. Under the standard neutral model, both methods (*i.e.*, using $\hat{\theta}$ and $\hat{\mu}$ or $\hat{\rho}$ and $\hat{r}$) should yield similar estimates of N_e (Hudson 1987).

Contrary to these expectations, recent studies of human and Drosophila data have revealed that the two methods for estimating N_e yield disparate results. In humans, the estimate of N_e from $\hat{\theta}$ and $\hat{\mu}$, over short physical distances (*i.e.*, within several kilobases), is substantially lower than the estimate from $\hat{\rho}$ and $\hat{r}$. This suggests less intragenic linkage disequilibrium (LD) than expected under the standard neutral model, given

estimated rates of crossing over (Frisse *et al.* 2001; Przeworski and Wall 2001). In *Drosophila melanogaster* and *D. simulans* the opposite pattern is observed: this approach suggests that there is an excess of LD relative to the standard neutral model expectations (Andolfatto and Przeworski 2000; Wall *et al.* 2002). Thus it appears that at least one of the assumptions of the standard neutral model is incorrect for both taxa. Possible explanations include more complex population histories (*e.g.*, changes in effective population size over time, or geographic structure), the widespread effects of natural selection (at the loci studied, or at closely linked loci), or systematic errors in estimating μ, r, or ρ. By determining how various population genetic models differentially affect levels of diversity and levels of LD, we might infer which types of models are most plausible for each species (Wall 2001; Wall *et al.* 2002).

The Drosophila data analyzed thus far have consisted predominantly of non-African samples. Both *D. melanogaster* and *D. simulans* are human commensals; they are thought to have originated in sub-Saharan Africa and only relatively recently colonized other locations (Lachaise *et al.* 1988). Non-African populations may have experienced a recent contraction in population size or a recent burst of selection associated with adaptation to new envi-

ronments (*e.g.*, Aquadro *et al.* 1994; Hamblin and Veu-ille 1999; Andolfatto 2001; Kauer *et al.* 2002; Wall *et al.* 2002). Both of these scenarios are expected to reduce levels of variability in non-African populations relative to African ones, as observed (Begun and Aquadro 1993; Andolfatto 2001; Kauer *et al.* 2002); in addition, either one may have caused an increase in levels of intragenic LD and might explain the observed discrepancy between different estimates of N_e.

Interestingly, patterns of variation in human populations share some similarities with Drosophila (Aquadro *et al.* 2001). Non-African populations have less variability than sub-Saharan African ones (*e.g.*, Frisse *et al.* 2001; Stephens *et al.* 2001), and a similar demographic model has been proposed for humans (*e.g.*, Tishkoff *et al.* 1996): an African origin, a recent population bottleneck in non-African populations, and recent population growth. Unlike the comparable Drosophila studies, human surveys of genetic variation have usually contained a substantial sample from sub-Saharan African populations. It is tempting to speculate that this difference in sampling schemes may explain part of the difference in patterns of LD in the human and Drosophila data. Further suggestive evidence of the sensitivity to sampling scheme comes from studies of different human populations, which find substantially higher levels of LD in non-African populations compared with African ones (Frisse *et al.* 2001).

In addition to the effects of population history on patterns of LD, the nature of the meiotic recombination mechanism itself is an issue. In current models of meiotic recombination (*e.g.*, Szostak *et al.* 1983), all recombination events involve some exchange of segments of DNA between homologous chromosomes. Such exchanges explain the phenomenon of gene conversion (the non-reciprocal transfer of an allele from one homolog to another). Whether or not these exchanges are accompanied by the reciprocal exchange of flanking markers (*i.e.*, crossing over) reflects alternative outcomes of Holliday junction resolution. For clarity, we refer to recombination events without the exchange of flanking markers as "gene conversion," and "crossing over" refers to recombination events with the exchange of flanking markers. Estimates of the recombination rate based on the large-scale comparison of genetic and physical maps ignore gene conversion, which contributes little to the rate of exchange between distant markers. However, at smaller physical scales (*i.e.*, several kilobases), depending on parameters, gene conversion may contribute substantially to the total rate of genetic exchange (Andolfatto and Nordborg 1998). Thus, the interpretation of patterns of LD in the context of demographic models depends to some extent on gene conversion parameters.

Frequent gene conversion in humans may account for inflated estimates of ρ (over small physical scales) relative to large-scale map-based estimates of r in African populations (Frisse *et al.* 2001; Przeworski and Wall 2001). However, incorporating gene conversion would make the observed LD excess in Drosophila even more unusual (Andolfatto and Przeworski 2000; Wall *et al.* 2002). Since gene conversion and demography both affect patterns of LD, estimates of gene conversion parameters from nucleotide polymorphism data may be extremely inaccurate since the true population history is unknown.

In an attempt to distinguish between possible explanations for the discordant estimates of N_e in Drosophila, we analyze sequence polymorphism data at multiple loci in Zimbabwean population samples of *D. melanogaster*. This population is more likely than non-African populations to be closer to equilibrium since it is nearer to *D. melanogaster*'s ancestral range and may not have experienced drastic changes in size or ecology in its recent history.

The physical scale of LD should increase as the rate of recombination decreases. Interestingly, a recent survey of two gene regions at the tip of the X chromosome of *D. melanogaster*, where rates of meiotic crossing over are greatly reduced, suggested that the physical scale of LD is not larger than that in other regions of the genome (Langley *et al.* 2000). Jensen *et al.* (2002) observed a similar pattern on chromosome 4 of *D. melanogaster* and *D. simulans*, which usually lacks crossing over during female meiosis (Hawley *et al.* 1993). These patterns were interpreted as evidence for high levels of gene conversion in these regions of reduced crossing over. Here we explore these claims by quantifying levels of LD at many loci over a broad range of crossing-over rates on the X chromosome. Using these data, we investigate several alternative models of how gene conversion and crossing over may be associated.

METHODS

Previously published data: We consider all sequence polymorphism studies that include a sample size (n) of seven or more *D. melanogaster* sequences from Zimbabwe and 10 or more segregating sites (S). We adopt these minimal size restrictions because estimates of ρ are highly biased and inaccurate when both n and S are small (see justification in results). Fifteen previously published X chromosome data sets fit this size requirement: *su(s)* and *su(w^a)* (Langley *et al.* 2000); *6-phosphogluconate dehydrogenase, vinculin,* and *zeste* (*Pgd, vinc,* and *z*; Andolfatto and Przeworski 2001); *glucose-6-phosphate dehydrogenase* (*G6pd*; Eanes *et al.* 1996); *shaggy* (*sgg*; C. Schlötterer, personal communication); *period/CG2650, 100G10.2, sxy4, frag.3, frag.4,* and *CG3592* (Harr *et al.* 2002); *runt* (Labate *et al.* 1999); and *vermilion* (*v*; Begun and Aquadro 1995). Nine autosomal gene regions fit the size requirement: *Acp26Aa* and *Acp26Ab* (Tsaur *et al.* 1998), which were analyzed as one gene region; *Acp36DE*

(BEGUN *et al.* 2000); *Alcohol dehydrogenase* (S.-C. TSAUR, unpublished data); *Esterase-6* promoter region (*Est6-p*; ODGERS *et al.* 2002); *Hexokinase C* (*Hex-C*; DUVERNELL and EANES 2000); *In(2L)t* proximal breakpoint [*In(2L)t-PBP*; ANDOLFATTO and KREITMAN 2000]; *Phosphogluconate mutase* (*Pgm*; VERRELLI and EANES 2000); *Heat-shock protein 70Bb* (*Hsp70Bb*; MASIDE *et al.* 2002); and *bicoid* (BAINES *et al.* 2002). All loci were surveyed in lines from one or both of two Zimbabwean population samples (Sengwa Wildlife Reserve and Harare, respectively) first described by BEGUN and AQUADRO (1993).

Data collection: We collected nucleotide variability data for eight additional gene regions: *yellow* (*y*, 2017 bp, *n* = 49); *Fasciclin-2* (*Fas2*, 588 bp, *n* = 17); *spaghetti squash* (*sqh*, 572 bp, *n* = 16); *Hyperkinetic* (*Hk*, 563 bp, *n* = 21); *dusky* (*dy*, 567 bp, *n* = 20); *licorne* (*lic*, 615 bp, *n* = 23); *rutabega* (*rut*, 548 bp, *n* = 22), and 1049 bp of the *white* gene (in two regions of 546 and 503 bp, respectively, separated by 3.9 kb, *n* = 16), which we analyzed as one region. We also expanded the *snf1A* data set reported in ANDOLFATTO and PRZEWORSKI (2001). For the *snf1A* gene region, we increased the sample size from 13 to 25 individuals and the total length of the surveyed DNA from 514 to 2189 bp (in four separate regions of 591, 514, 488, and 596 bp, spread over 9.33 kb). These four regions surrounding the *snf1A* gene were analyzed as one region.

Details of the sequencing strategy for each gene region can be found at http://helios.bto.ed.ac.uk/evolgen/andolfatto/zimbabweLD. We PCR amplified each region from genomic DNA of single male flies (one male from each isofemale line), precipitated the products with polyethylene glycol (PEG-8000) to remove primers, and sequenced them on both strands using Big-Dye (Applied Biosystems, Foster City, CA) or DYEnamic ET (Amersham, Arlington Heights, IL) sequencing kits. One of our PCR/sequencing primers overlapped with a 63-bp deletion in region 4 of *snf1A* in individual zs53. For this region, this allele was amplified with flanking primers, cloned into pCR4-TOPO (Invitrogen, San Diego), and three independent clones were sequenced. For all loci surveyed, the lines used either are the same 13 individuals surveyed in ANDOLFATTO and PRZEWORSKI (2001) from the Sengwa Wildlife Reserve, Zimbabwe or included additional alleles from a sample of 12 lines from Harare, Zimbabwe (kindly provided by C.-I Wu) and up to 24 lines collected from Victoria Falls, Zimbabwe (kindly provided by B. Ballard). Aligned and annotated sequences are available in nexus file format at the website http://helios.bto.ed.ac.uk/evolgen/andolfatto/zimbabweLD. A summary of polymorphic variation at each locus can be found at the website http://helios.bto.ed.ac.uk/evolgen/andolfatto/zimbabweLD. APPENDIX A summarizes information about each of the loci used in this study.

Analysis of levels of variability: Our first goal is to estimate the effective population size (*N*e) for each locus

independently. For this purpose, we estimate $\hat{N}_\theta = \hat{\theta}/4\hat{\mu}$ assuming a constant mutation rate, μ (see below). We multiply $\hat{N}_\theta$ for loci on the X chromosome by $\frac{4}{3}$ to make them comparable to estimates for autosomal loci (which makes a number of assumptions, see below). We employ two estimators of θ: $\hat{\theta}_W$, which is based on the number of single nucleotide polymorphisms observed in the sample (WATTERSON 1975); and $\hat{\theta}_\pi$, which is the average pairwise divergence between sampled chromosomes per site (*cf.* LI 1997, p. 238). For both of these methods, we use the total number of silent mutations (including multiply hit sites) and for $\hat{\theta}_\pi$, we also use the Jukes-Cantor correction for multiple hits as implemented in DnaSP version 3.59 (http://www.ub.es/dnasp/).

A second goal of this analysis is to compare joint estimates of N_e from $\hat{\theta}$ ($\hat{N}_\theta$) to joint estimates of N_e from LD (see below). Since a correlation between variability and recombination rates has been well documented in *D. melanogaster* (BEGUN and AQUADRO 1992; AQUADRO *et al.* 1994; ANDOLFATTO and PRZEWORSKI 2001), we limit this analysis to genomic regions with $\hat{r} > 1$ cM/Mb (*i.e.*, those regions least likely to be affected by selection at linked sites). To do this, we calculated averages of $\hat{\theta}_W$ and $\hat{\theta}_\pi$ for silent sites (weighted by the number of silent sites surveyed). This approach has the advantage that $\hat{\theta}_W$ and $\hat{\theta}_\pi$ are unbiased estimates of θ; however, confidence intervals must be obtained by coalescent simulations for each locus. We instead estimated θ ($\hat{\theta}_{ML}$) and confidence intervals using a likelihood-based approach based on Equation 12 of HUDSON (1990) as implemented by WRIGHT *et al.* (2003). For each locus, this approach maximizes the likelihood recursion equation

$$L_n(\theta|S) \propto P_n(S|\theta) = \sum_{i=1}^{S} P_{n-1}(S-i|\theta)\,Q_n(i|\theta), \quad (1)$$

where

$$P_2(S|\theta) = \left(\frac{\theta}{1+\theta}\right)^s \frac{1}{1+\theta}, \quad (2)$$

and

$$Q_n(S|\theta) = \left(\frac{\theta}{\theta+n-1}\right)^s \frac{(n-1)}{(\theta+n-1)}, \quad (3)$$

over a range of θ values, where *n* is the sample size and *S* is the observed number of silent polymorphisms (again including multiple hits). Assuming that each genomic region investigated is evolutionarily independent, likelihoods can be combined to produce a joint likelihood surface for all loci, the maximum of which is $\hat{\theta}_{ML}$. Confidence limits for $\hat{\theta}_{ML}$ are estimated using the standard χ^2 approximation. This approach assumes no recombination within loci, but independence among loci. Simulations confirm that the bias of this estimator is not large and that confidence limits for $\hat{\theta}$ estimated by this method are wider than those in the presence of intralo-

cus recombination (S. WRIGHT, personal communication).

Estimating $\hat{N}_\theta$ from $\hat{\theta}$ requires knowledge of the mutation rate for silent sites. By comparing estimates of $\hat{N}_\theta$, or estimating $\hat{N}_\theta$ from a joint estimate of $\hat{\theta}$, we implicitly assume that the mutation rate at silent sites is constant among loci. We assume that $\mu = 1.5 \times 10^{-8}$/silent site/year and that *D. melanogaster* populations undergo an average of 10 generations/year, yielding a $\hat{\mu} = 1.5 \times 10^{-9}$/silent site/generation. Our estimate of the mutation rate depends on many assumptions (reviewed in ANDOLFATTO and PRZEWORSKI 2000), although several estimates based on independent approaches are in close agreement with the above figure. Further, the true mutation rate is unlikely to be $>3 \times 10^{-9}$/site/generation (ANDOLFATTO and PRZEWORSKI 2000; MCVEAN and VIEIRA 2001). We discuss the sensitivity of our results to assumptions about the mutation rate. Where we compare X-linked and autosomal loci, it is worth noting that we have assumed that the ratio of their effective sizes is as expected under a neutral model with equal effective population sizes for males and females. In fact, the appropriate scaling of the X and autosomes is uncertain given that natural populations experience factors that alter the relative variance in reproductive success for males and females (CHARLESWORTH 2001). For most analyses, we consider X-linked and autosomal loci separately, so this is not an issue.

Linkage disequilibrium analysis: The goal of this analysis is to estimate N_e from levels of LD. We choose to summarize LD by an estimate of the population recombination rate $\rho = 4N_e r$, where N_e is the effective population size, and r is the sex-averaged rate of recombination (*cf.* ANDOLFATTO and PRZEWORSKI 2000; WALL *et al.* 2002). For these analyses, we use all biallelic mutations, including both single-nucleotide polymorphisms (SNPs) and simple insertion-deletion mutations. All multiply hit sites or overlapping mutational events are excluded.

We first employ an estimator, $\hat{\rho}_{W00}$, proposed by WALL (2000). To estimate $\hat{\rho}_{W00}$, we summarize the data using the observed number of distinct haplotypes (H) and the minimum number of inferred recombination events (R_M, *cf.* HUDSON and KAPLAN 1985) and estimate via simulation the value of ρ that maximizes the likelihood of obtaining the observed values of H and R_M (*cf.* WALL 2000). The simulations use a generalization of the methodology of HUDSON (1993), which allows noncontiguous but linked regions to be analyzed. We estimate $\hat{r}$ for each locus on the basis of comparisons of physical and genetic maps as described in CHARLESWORTH (1996) and ANDOLFATTO and PRZEWORSKI (2001). We assume that r is known exactly and estimate $\hat{N}_\rho$ ($=\hat{\rho}/4\hat{r}$) over increments of 3.3×10^5 or smaller, using a minimum of 2×10^5 replicates for each parameter value for each locus (*cf.* WALL *et al.* 2002). $\hat{N}_\rho$ is estimated for each locus separately and also jointly for loci with $\hat{r} > 1$ cM/Mb,

assuming that each locus is evolutionarily independent (*cf.* WALL *et al.* 2002). As for diversity-based estimates, we multiply $\hat{N}_\rho$ for loci on the X chromosome by $\frac{4}{3}$ to make them comparable to estimates for autosomal loci. Approximate 95% confidence intervals for joint estimates $\hat{\rho}_{W00}$ are found by making the standard asymptotic maximum-likelihood assumptions. We caution that there is no evidence that these assumptions are appropriate in this context, but the intervals serve as a useful diagnostic.

Several other summary estimators of ρ have also been proposed (*e.g.*, HUDSON 1987, 2001; HEY and WAKELEY 1997). We have chosen the estimator proposed by HUDSON (2001), $\hat{\rho}_{H01}$, which is a composite-likelihood estimator of ρ based on pairwise LD among all pairs of polymorphic mutations in a sample. This estimator performs similarly to $\hat{\rho}_{W00}$ and is substantially better than other estimators (see WALL 2000 and HUDSON 2001 for details). For one of our X-linked data sets, *frag4*, the estimate of $\hat{\rho}_{H01}$ is extremely large (*i.e.*, $>10,000$). We exclude this locus from correlation tests and thus tests involving $\hat{\rho}_{H01}$ use only 23 of 24 loci.

A considerable fraction of polymorphisms segregating among autosomal genes (43/279) are amino acid replacement substitutions. If these are under weak purifying selection (ANDOLFATTO 2001), they may segregate at lower frequencies on average, potentially affecting estimates of ρ. We estimated heterozygosities for each site [$\hat{\pi} = 2np(1 - p)/(n - 1)$, where n is the sample size and p is the polymorphic site frequency] and found that the mean heterozygosities of replacement (mean $\hat{\pi} = 0.29$) and silent (mean $\hat{\pi} = 0.32$) polymorphisms are not significantly different ($P = 0.25$, Student's *t*-test with unequal variances; $P = 0.27$, two-tailed Mann-Whitney *U*-test). We have thus chosen to include replacement polymorphisms; excluding them would considerably reduce the number of polymorphisms for some loci, possibly resulting in an upward bias in estimates of ρ (see RESULTS). The fraction of replacement polymorphisms among X-linked loci was considerably smaller (9/884) and thus their effect on estimates of ρ, if any, should be negligible. Note that part of the difference in the number of amino acid replacement polymorphisms detected on X-linked *vs.* autosomal genes reflects the fact that about two times more coding DNA was surveyed for the latter. The remainder of the difference probably reflects differences in the efficacy of selection against amino acid polymorphisms on the two chromosomes (see ANDOLFATTO 2001).

Assessing the behavior of estimators under neutrality and alternative models: To test the properties of the two estimators of ρ under the standard neutral model, we run simulations with sample sizes (n) of 7, 13, or 50, with $\rho = 3\theta$. We consider a range of θ values and calculate the average and median values of $\hat{\rho}$ divided by the actual value of ρ. A total of 10^4 replicates were run for $\hat{\rho}_{W00}$ (and 2×10^3 replicates for $\hat{\rho}_{H01}$) under each

parameter combination. Results of similar simulations for $n = 50$ for $\rho = 4\theta$ and $\rho = \theta$ are reported by Hudson (2001). We expect, on the basis of estimates of the neutral mutation rate and rates of crossing over, that ρ/θ will vary considerably across the genome, with $\rho < 3\theta$ for *yellow*, *su(s)*, and *su(w^a)* and $\rho > 3\theta$ for most loci in regions of high rates of crossing over. In addition to the simulations above, we have also investigated the distribution of $\hat{\rho}_{W00}$ for parameters (n and $\hat{\theta}$) that best match those expected for the *yellow* locus under $\rho \approx \theta/4$ and $\rho \approx 3\theta$. These results have been posted at the site http://helios. bto.ed.ac.uk/evolgen/andolfatto/zimbabweLD.

Directional selection at linked sites is known to shape patterns of genetic variation (Hill and Robertson 1966) and is thought to explain the observed positive correlation between levels of diversity and the estimated crossing-over rate (Begun and Aquadro 1992; Aquadro *et al.* 1994). Both linkage to positively selected alleles (*i.e.*, hitchhiking or selective sweeps, *cf.* Maynard Smith and Haigh 1974) or deleterious alleles (*i.e.*, background selection, *cf.* Charlesworth *et al.* 1993) have been advanced as possible explanations.

Background selection due to strongly deleterious mutations can be modeled as a simple reduction in effective population size (Hudson and Kaplan 1994). This assumption is equivalent to varying θ (and ρ) in simulations of the standard neutral model. However, under background selection models that posit more weakly deleterious mutations, this assumption may not be valid (Charlesworth *et al.* 1995; Gordo *et al.* 2002). This situation is beyond the scope of our study and will be a topic of future work.

We do examine the effects of recurrent hitchhiking due to advantageous mutations by running simulations under a model of recurrent, nonoverlapping selective sweeps (Przeworski 2002). The program used was kindly provided by M. Przeworski. We consider a locus of fixed size ($n = 50$, number of sites $= 2000$, $\theta = 15$, $\rho = 45$) and determine how the average and median values of $\hat{N}_\theta$ and $\hat{N}_\rho$, as estimated assuming the standard neutral model, change as a function of the rate of selective sweeps (Λ in Przeworski 2002). We fix the selection coefficient $s = 0.002$ and the species population size $N_e = 4 \times 10^6$. The choice of other parameter values yielded similar results (results not shown).

Incorporating the effects of gene conversion: Estimates of $\hat{N}_\rho$ based on rates of crossing over ($\hat{r}$) may be systematically underestimated because they ignore gene conversion. Gene conversion was implemented into the standard coalescent with recombination (*cf.* Hudson 1990; Wiuf and Hein 2000; Przeworski and Wall 2001). Gene conversion is defined here as exchange of short tracts of DNA between homologous chromosomes with no associated crossing over between flanking markers. Thus, in this implementation, gene conversion and crossing over are modeled as mechanistically independent, and gene conversion exchanges associated with

crossing over are ignored. For all models, we assume the rate of crossing over is $\hat{r}$ and that the distribution of gene conversion tract lengths is geometric with a mean of 352 bp (Hilliker *et al.* 1994).

We consider three models of how gene conversion (GC) and crossing-over (CO) rates may be associated. In the first model (GC1), we assume that two-thirds of recombination events are GC, and thus the rate of GC, $\hat{r}_{GC} = 2\hat{r}$ between adjacent sites, where $\hat{r}$ is the estimated local rate of crossing over for a particular gene region. In a second model (GC2), we assume that the rate of GC is constant across a chromosome; $\hat{r}_{GC} = 2\hat{r}_{MAX}$, where $\hat{r}_{MAX}$ is the highest estimated rate of crossing over on the chromosome ($\hat{r}_{MAX} = 2.27 \times 10^{-8}$/generation between adjacent sites on the X chromosome). In a third model (GC3), we assume that the *total* rate of recombination between adjacent sites is constant across a chromosome and that the rate of gene conversion between adjacent sites is $\hat{r}_{GC} = (\hat{r}_{MAX} - \hat{r})$, where $\hat{r}$ is the estimated rate of crossing over for the locus under study. GC1 and GC2 are equivalent for genomic regions with the highest rates of crossing over and diverge in regions of reduced crossing over. Likewise, GC3 is equivalent to a model with no GC for regions with the highest rates of crossing over. These models are meant to be illustrative rather than quantitative and other possible models of recombination are addressed in the discussion. $\hat{N}_\rho$ is estimated as above (from $\hat{\rho}_{W00}$) from simulations with gene conversion to estimate the likelihoods.

How do gene conversion rates in our models compare to data on gene conversion from Drosophila? On the basis of the data from the genes *maroon-like* and *rosy*, it is thought that the rate of gene conversion (γ, see Andolfatto and Nordborg 1998) is on the order of 10^{-5} events/generation and that 50% or more of recombination events are gene conversions without an associated exchange of flanking markers (Finnerty 1976; Hilliker and Chovnick 1981). Under GC1 and GC2, the sex-averaged GC rate is 1.5×10^{-5}/generation in regions of the genome with the highest rates of crossing over, which is similar to the rate that has been measured at *rosy* and *maroon-like* ($\sim 10^{-5}$, Finnerty 1976; Hilliker and Chovnick 1981). The ratio of GC:CO has been estimated to be ~ 1 at *maroon-like* (Finnerty 1976) and ~ 4 for *rosy* (Hilliker and Chovnick 1981). In GC1 and GC3 models, the ratio of GC to CO rates at the *rosy* locus ($\hat{r} = 0.7$ cM/Mb) would be ~ 2 and corresponds to a GC rate of 6×10^{-6}/generation. In the GC2 model, the GC:CO ratio at *rosy* would be ~ 6. For *maroon-like* ($\hat{r} = 1.36$ cM/Mb), the GC:CO ratio $= 2$ under GC1 (rate $\sim 10^{-5}$/generation) and would be ~ 0.7 under GC3 (rate $\sim 3 \times 10^{-6}$/generation). In the GC2 model, the GC:CO ratio at *maroon-like* would be ~ 3. Thus, our assumed rates of gene conversion, and the extent of association between gene conversion and crossing over, are roughly in agreement with the limited data on rates of gene conversion in Drosophila.

TABLE 1

Bias of $\hat{\rho}$ under the standard neutral model

	$\hat{\rho}_{W00}$			$\hat{\rho}_{H01}$		
θ	$n = 7$	$n = 13$	$n = 50$	$n = 7$	$n = 13$	$n = 50$
2.5	5.2 (0.8)	1.4 (0.8)	1.2 (0.8)	7.1 (1.4)	4.4 (1.2)	2.1 (1.1)
3.5	3.3 (0.6)	1.4 (0.9)	1.1 (0.9)	4.4 (1.3)	2.5 (1.1)	1.4 (1.0)
5.0	2.1 (0.9)	1.3 (1.0)	1.0 (1.0)	3.0 (1.3)	1.7 (1.1)	1.2 (1.0)
7.0	1.6 (0.9)	1.3 (1.0)	1.0 (1.0)	2.1 (1.3)	1.5 (1.1)	1.1 (1.0)
10.0	1.4 (1.0)	1.2 (1.0)	1.0 (1.0)	1.8 (1.2)	1.3 (1.1)	1.1 (1.0)
15.0	1.2 (0.9)	1.1 (1.0)	1.0 (1.0)	1.6 (1.2)	1.3 (1.1)	1.1 (1.0)

Bias was measured as the average value of $\hat{\rho}$ (on the basis of 10^4 replicates for $\hat{\rho}_{W00}$ and 2000 replicates for $\hat{\rho}_{H01}$) divided by the true value of ρ. $\hat{\rho}$ with a value >250 were assigned the value 250. A value of 1 represents an unbiased estimator. In parentheses are the median values of $\hat{\rho}$ divided by the true value of ρ. $\rho = 3\theta$ for all simulations.

RESULTS

Performance of ρ estimators and choice of data: With a choice of two similarly performing ρ estimators, $\hat{\rho}_{W00}$ (WALL 2000) and $\hat{\rho}_{H01}$, (HUDSON 2001), we first set out to investigate how these estimators perform under the standard neutral model in the context of parameters relevant to the available data (*e.g.*, sample size, the number of segregating sites, etc.) In Table 1, we summarize the performance of the two estimators as a function of the number of alleles sampled (n) and the amount of variation (θ, proportional to the number of segregating sites). Both estimators show considerable upward bias when the sample size is small and/or θ is small; however, the median of $\hat{\rho}_{W00}$ is generally biased downward under these conditions. Overall, $\hat{\rho}_{H01}$ appears to be more biased than $\hat{\rho}_{W00}$; this is particularly pronounced when the number of segregating sites is small (even for large sample sizes). On the basis of these simulations, we conclude that $\hat{\rho}_{W00}$ should be used cautiously for $7 \leq n \leq 13$ unless θ is in the vicinity of 10 or larger and probably not at all for samples of <7 alleles. Since $\hat{\rho}_{H01}$ seems to be

particularly sensitive to small θ, it should be used cautiously when there are <13 alleles unless $\theta \geq 10$ and probably not at all if $\theta < 5$.

We have used these results as a guide to select data for our study (see METHODS) and accordingly we have limited the data sets considered here to those that have $n \geq 7$ and $S > 10$ (APPENDIX A). Only 4 of our selected data sets have $n = 7$ and for each of these $\hat{\theta}_W > 8$ ($\hat{\theta}_W$ is calculated from S_{tot}; see APPENDIX A). The 10 data sets that have $7 < n \leq 13$ have $\hat{\theta}_W \geq 6$. Possible effects of biases of the estimators on our results are noted in the analyses and discussion that follow.

Figure 1 compares estimates of $\hat{\rho}_{W00}$ and $\hat{\rho}_{H01}$ for all loci fitting the above restrictions. $\hat{\rho}_{W00}$ and $\hat{\rho}_{H01}$ are strongly positively correlated ($R = 0.78$, $P < 0.001$, $n = 33$, Spearman rank correlation test, two-tailed). However, $\hat{\rho}_{H01}$ are also systematically larger than $\hat{\rho}_{W00}$, as might be expected given the results in Table 1. APPENDIX B (http://helios.bto.ed.ac.uk/evolgen/andolfatto/zimbabweLD) summarizes correlations among n, $\hat{r}$, $\hat{\theta}_W$ (per locus), and estimates of $\hat{\rho}$ (per locus) for the X chromosome data.

Differences in sample size and the number of segregating sites are not the only possible causes of the systematic difference between estimators. Regions of reduced recombination in *D. melanogaster* have lower levels of variability due to hitchhiking and/or background selection (AQUADRO *et al.* 1994; CHARLESWORTH 1996). If background selection can be approximated as a simple reduction in effective population size, the expected behavior of ρ estimators will be similar to that under the neutral model (*cf.* Table 1). However, the observation that these regions also have more low-frequency polymorphisms suggests that recurrent selective sweeps may also influence polymorphism patterns in these regions (ANDOLFATTO and PRZEWORSKI 2001).

Since one of our objectives is to compare $\hat{N}_\rho$ and $\hat{N}_\theta$ estimates across a gradient of crossing-over rates, we investigated the properties of the two ρ estimators under a recurrent hitchhiking model (Table 2). More frequent hitchhiking is expected to decrease levels of variability at
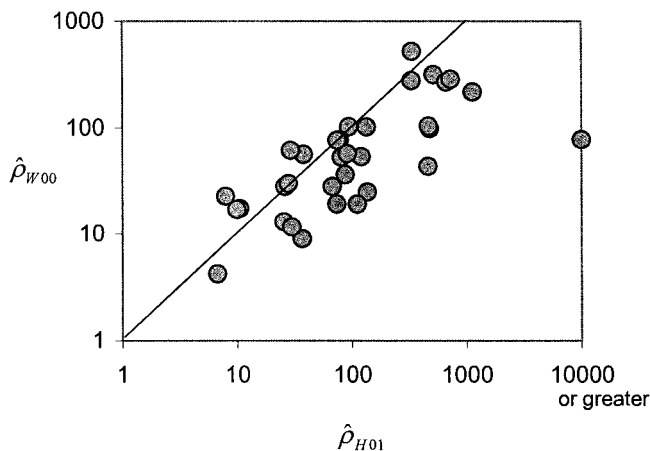

FIGURE 1.—Comparison of estimates of $\hat{\rho}$ (per locus) for the 24 X-linked and 9 autosomal polymorphism data sets used in this study.

Effect of selective sweeps on estimates of N_θ and N_ρ

Rate of sweeps (Λ)[a]	Relative $E(\hat{N}_\theta)$[b]	Relative mean[c] (median[d])	
		$\hat{N}_\rho$ from $\hat{\rho}_{W00}$	$\hat{N}_\rho$ from $\hat{\rho}_{H01}$
0	1.00	1.00 (1.00)	1.00 (1.00)
0.00001	0.68	0.72 (0.72)	0.88 (0.85)
0.00002	0.54	0.57 (0.54)	0.83 (0.75)
0.00005	0.37	0.34 (0.27)	0.92 (0.62)
0.0001	0.26	0.20 (0.14)	1.09 (0.57)
0.0002	0.18	0.12 (0.03)	1.43 (0.56)

Estimates are based on 2000 replicates for each Λ; $n = 50$, number of sites = 2000, $\theta = 15$, $\rho = 45$, $s = 0.002$, $N_e = 4 \times 10^6$. $\hat{N}$ for simulated data are estimated assuming the standard neutral model (*cf.* WALL *et al.* 2002).

[a] The rate of fixation of advantageous mutations per base pair per $4N_e$ generations.
[b] The ratio of $E(\hat{N}_\theta)$ based on $\hat{\theta}_W$ (see METHODS) at a given value of Λ to $E(\hat{N}_\theta)$ at $\Lambda = 0$.
[c] The ratio of $E(\hat{N}_\rho)$ at a given value of Λ to $E(\hat{N}_\rho)$ at $\Lambda = 0$.
[d] The ratio of the median $(\hat{N}_\rho)$ at a given value of Λ to the median $(\hat{N}_\rho)$ at $\Lambda = 0$.

closely linked sites (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989) and thus decrease $\hat{N}_\theta$. As can be seen in Table 2, the two ρ estimators respond differently to increasing rates of hitchhiking. $\hat{\rho}_{W00}$ is strongly positively correlated with $\hat{\theta}$ and generally yields $\hat{N}_\rho$ that are slightly smaller than $\hat{N}_\theta$. In contrast, the median $\hat{\rho}_{H01}$ is only weakly correlated with estimates of $\hat{\theta}$, and $\hat{N}_\rho$ from $\hat{\rho}_{H01}$ are typically much larger than estimates of $\hat{N}_\theta$; this difference is larger as the rate of selective sweeps increases.

Further investigation of the behavior of these ρ estimators under alternative models is beyond the scope of this study. In the analyses that follow, we focus on $\hat{\rho}_{W00}$ for several reasons. First, $\hat{\rho}_{W00}$ allows us to easily combine information from multiple loci in a likelihood framework. Second, $\hat{\rho}_{W00}$ is not severely biased upward under parameters that best match our available data (*i.e.*, most are small samples with few segregating sites); in fact, the median estimate is often below the true ρ. Finally, using $\hat{\rho}_{W00}$ is not likely to result in $\hat{N}_\rho > \hat{N}_\theta$ in regions of reduced crossing over if recurrent hitchhiking is reducing levels of variability in these regions.

Comparing levels of diversity and linkage disequilibrium: In a neutral panmictic population, we expect that $\hat{N}_\rho \approx \hat{N}_\theta$. Table 3 lists estimates of $\hat{N}$ for each data set and these are plotted in cytological order in Figure 2. With the exception of the tip of the X chromosome (the first 10 points cover cytological bands 1B1 to 3B3 on the X chromosome), values of $\hat{N}_\rho$ based on $\hat{\rho}_{W00}$ are slightly lower than the values of $\hat{N}_\theta$(13/14 on the X have $\hat{N}_\rho < \hat{N}_\theta$). For $\hat{N}_\rho$ based on $\hat{\rho}_{H01}$, the pattern is less striking, particularly for the X (7/14 have $\hat{N}_\rho < \hat{N}_\theta$; data not shown); however, the pattern is equally strong for the two estimators on the autosomes (9/9 loci and 8/9 loci have $\hat{N}_\rho < \hat{N}_\theta$, based on $\hat{\rho}_{W00}$ and $\hat{\rho}_{H01}$, respectively). It is difficult to draw solid conclusions from this type of analysis because different sample sizes and numbers of segregating sites for each locus (and different averages for

these loci on the X and autosomes) complicate the interpretation of comparisons.

Further, due to the inherent stochasticity of the evolutionary process, individual estimates of $\hat{N}_\theta$ and $\hat{N}_\rho$ are not particularly accurate. We can achieve greater precision by estimating the relative likelihoods for different $\hat{N}$ values using multiple loci jointly. Figure 3a shows the joint-likelihood curve of $\hat{N}_\rho$ (based on $\hat{\rho}_{W00}$) for the X-linked and autosomal loci with $\hat{r} > 1$ cM/Mb (*i.e.*, loci in areas of high recombination that are least likely to be affected by selection at linked sites). The maximum-likelihood estimate of $\hat{N}_\rho$ for the X chromosome is $\hat{N}_{\rho X} = 1.7$ million, with an $\sim$95% confidence interval of 1.1–2.3 million. The corresponding estimate of $\hat{N}_{\theta X}$ is 4.0 million (based on the $\hat{\theta}_{ML}$), with a 95% confidence interval of 3.1–5.1 million. For the autosomes, $\hat{N}_{\rho A} = 0.4$ million, with an $\sim$95% confidence interval of 0.2–0.9 million. The corresponding estimate of $\hat{N}_{\theta A}$ is 1.8 million with a 95% confidence interval of 1.3–2.9 million.

Several interesting observations emerge from these results. First, the estimate of $\hat{N}_\theta$ for the X chromosome is larger than that for the autosomes. This trend has been noted before on the basis of levels of nucleotide (ANDOLFATTO 2001) and microsatellite variability (KAUER *et al.* 2002). On the basis of the confidence intervals for $\hat{\theta}$, we can reject the standard model with equal population sizes for males and females (*i.e.*, an X:autosome ratio of 3/4). Several factors may contribute to this pattern, including a large variance in reproductive success for males (*i.e.*, sexual selection), background selection, and autosome-specific inversion polymorphisms (CHARLESWORTH 1996, 2001; ANDOLFATTO 2001; KAUER *et al.* 2002). Interestingly, the difference in effective sizes is even more pronounced for estimates of N_e based on linkage disequilibrium; $\hat{N}_{\theta X}/\hat{N}_{\theta A} = 2.2$ whereas $\hat{N}_{\rho X}/\hat{N}_{\rho A} = 4.1$. More LD on the autosomes (reflected in the higher $\hat{N}_{\rho X}/\hat{N}_{\rho A}$ ratio) is not predicted under either background selection or sexual selection models and there-

TABLE 3

Estimates of N_e (in millions) for each locus

	X chromosome				Autosomes		
Locus	$\hat{N}_\theta$ ($\hat{\theta}_w$)	$\hat{N}_\theta$ ($\hat{\theta}_\pi$)	$\hat{N}_\rho$ ($\hat{\rho}_{W00}$)	Locus	$\hat{N}_\theta$ ($\hat{\theta}_w$)	$\hat{N}_\theta$ ($\hat{\theta}_\pi$)	$\hat{N}_\rho$ ($\hat{\rho}_{W00}$)
yellow	0.40	0.15	1.5	*Acp26Aa/Ab*	1.1	1.6	0.6
su(s)	0.63	0.44	6.3	*In(2L)t-PBP*	1.5	1.5	0.3
su(w^a)	1.2	0.91	4.7	*Adh*	1.3	3.3	0.2
snf1a	0.85	0.85	4.7	*Acp36DE*	3.3	3.5	0.4
Pgd	1.1	1.1	3.7	*Hex-C*	2.9	1.8	0.8
vinc	1.3	1.3	2.3	*Esterase-6*	2.7	4.5	0.5
zeste	2.4	2.4	2.0	*Pgm*	1.1	1.2	0.6
shaggy	2.5	2.5	10.7	*bicoid*	0.52	1.2	0.5
period	4.8	4.2	10.7	*Hsp70Bb*	1.0	0.85	0.2
100G10.2	4.0	4.0	8.0				
syx4	1.6	1.6	0.3				
frag3	2.6	2.5	3.7				
white	4.1	2.7	1.0				
frag4	2.2	1.9	1.3				
frag7	3.8	3.2	0.7				
Fasciclin-2	2.9	2.7	1.3				
spaghetti squash	3.1	2.8	0.3				
hyperkinetic	8.8	9.2	2.7				
vermillion	4.5	4.7	3.7				
dusky	6.5	7.5	2.0				
licorne	2.8	2.0	0.7				
rutabega	1.4	0.9	0.7				
g6pd (Zw)	4.7	4.7	4.0				
runt	3.8	3.7	0.7				
High recombination[a]	4.0	—	1.7		1.9	—	0.4
	(4.0)	(3.9)	—		(1.8)	(1.6)	—

[a] Joint maximum-likelihood estimates are based on loci with $\hat{r} \geq 1$ cM/Mb. For $\hat{N}_\theta$ ($\hat{\theta}_W$) and $\hat{N}_\theta$ ($\hat{\theta}_\pi$), estimates based on the weighted average of $\hat{\theta}_W$ and $\hat{\theta}_\pi$, respectively, are given in parentheses.

fore points to possible effects of inversion polymorphisms. Inversions are both much more common and at higher frequencies on the autosomes of *D. melanogaster* (Lemeunier and Aulard 1992). Thus inversion poly-

FIGURE 2.—Point estimates of N based on linkage disequilibrium ($\hat{\rho}_{W00}$, open circles) and levels of diversity ($\hat{\theta}_\pi$, shaded circles). A total of 24 X-linked and 9 autosomal loci are plotted in order of cytological position and in order of appearance in APPENDIX A and Table 3). The assumed mutation rate is 1.5×10^{-9}/site/generation.

morphisms may be reducing variability on the autosomes by increasing the effects of linked selection (Andolfatto 2001). In addition, the associated reduction in the effective rate of recombination (crossing over is suppressed in inversion heterozygotes) and rapid changes in inversion frequency (Andolfatto *et al.* 1999) can both increase levels of LD on the autosomes relative to levels of diversity. Additional data from autosomal genes may allow us to distinguish among alternative hypotheses.

Interestingly, joint estimates of $\hat{N}_\rho$ and $\hat{N}_\theta$ are significantly different from each other on both the X and the autosomes, given our assumed mutation rate ($\hat{\mu} = 1.5 \times 10^{-9}$/generation). This suggests that Zimbabwe populations, despite our *a priori* prediction, show an excess of LD relative to expectations under the neutral model. However, while confidence intervals for $\hat{N}_\theta$ are overestimated (since we have assumed no intragenic recombination), a trivial explanation for the discrepancy between $\hat{N}_\theta$ and $\hat{N}_\rho$ in Zimbabwean *D. melanogaster* might be error in the estimated mutation rate (*e.g.*, $\hat{N}_\theta$ would be half as large if we assume the true mutation rate is twofold higher; see METHODS). Given that the mutation rate is uncertain, the apparent LD excess in this population
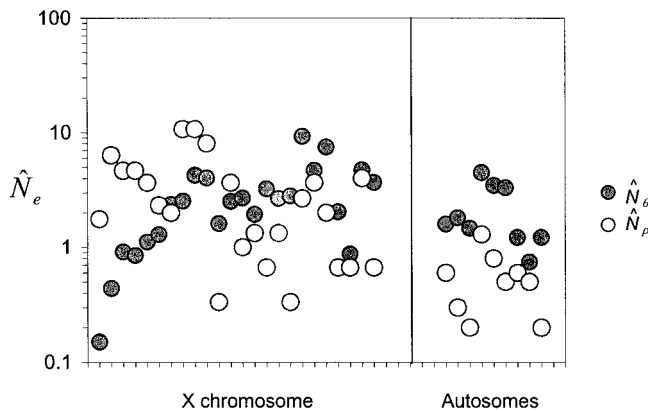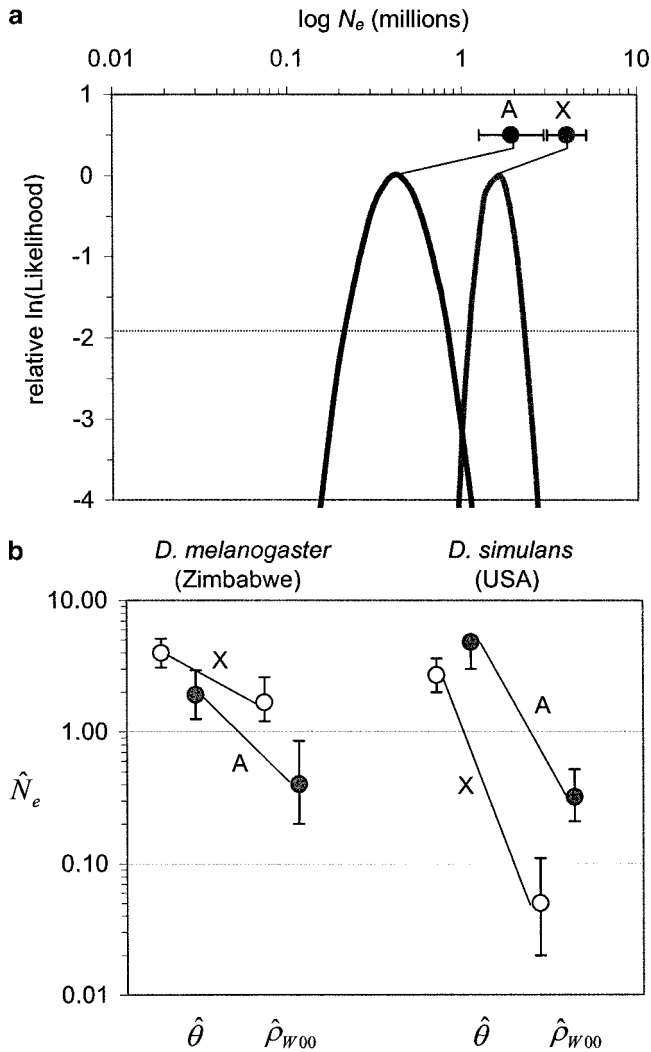
a

b

FIGURE 3.—(a) Joint maximum-likelihood estimates of $\hat{N}_\rho$ (curves) and $\hat{N}_\theta$ (points with ~95% confidence intervals). The assumed mutation rate is 1.5×10^{-9}/generation. Note that confidence intervals for $\hat{N}_\theta$ assume no recombination. The dotted line indicates the approximate 95% confidence bounds using the standard chi-square approximation. (b) Comparison of joint-likelihood estimates of $\hat{N}_\theta$ and $\hat{N}_\rho$ in Zimbabwean *D. melanogaster* and a California population of *D. simulans* (BEGUN and WHITLEY 2000; WALL *et al.* 2002). Approximate 95% confidence interval bars on estimates are indicated.

based on estimated rates of crossing over is not entirely convincing.

Joint-likelihood estimates of $\hat{N}_\rho$ and $\hat{N}_\theta$ for 15 X-linked and 14 autosomal loci from the California population of *D. simulans* are shown for comparison (Figure 3b). It is interesting to note in Figure 3b that ratios of $\hat{N}_\rho/\hat{N}_\theta$ for both the X and the autosomes are much larger for the Zimbabwean population of *D. melanogaster* (0.4 and 0.2, respectively) than for the California population of *D. simulans* (0.02 and 0.07, respectively; *cf.* WALL *et al.* 2002). While uncertainty in the mutation rate (~2-fold) may explain the discrepancy between $\hat{N}_\rho$ and $\hat{N}_\theta$ in Zimbabwean *D. melanogaster*, it is unlikely to explain

the >10-fold difference between these estimates of N in the *D. simulans* population in Figure 3b. This suggests that something is fundamentally different in the evolutionary history of these two populations or species. In particular, the Zimbabwe population of *D. melanogaster* may be closer to mutation-drift equilibrium, while the California population of *D. simulans* clearly is not.

Behavior of estimates of ρ as a function of *r*: A positive correlation between $\hat{\rho}$ and $\hat{r}$ is expected for two reasons: first, $\hat{\rho}$ (an estimate of $4N_e r$) should increase proportionally with $\hat{r}$; second, the gradient in $\hat{r}$ strongly positively covaries with empirical estimates of diversity ($\hat{\theta}$) in Drosophila (BEGUN and AQUADRO 1992; AQUADRO *et al.* 1994; ANDOLFATTO and PRZEWORSKI 2001) and, hence, with the apparent effective population size of the genomic region ($\hat{N}_\theta$, which under the standard neutral model is inversely proportional to the expected linkage disequilibrium for a given r). Here, we focus on data from the X chromosome because we lack data for autosomal loci in regions of reduced crossing over. For the 24 gene regions on the X chromosome, $\hat{\rho}_{W00}$ (per site) is positively correlated with $\hat{r}$ in our data set [Figure 4a; $\hat{\rho}_{W00}$, $R = 0.44$, $P < 0.05$, 24 loci; $\hat{\rho}_{H01}$, $R = 0.43$, $P < 0.05$, 23 loci (see METHODS); Spearman rank correlation test, two-tailed]. As expected on the basis of previous results, $\hat{\theta}$ and $\hat{N}_\theta$ for the 24 X-linked loci considered here also show a strong positive correlation with $\hat{r}$ (Figure 4b; based on $\hat{\theta}_\pi$, $R = 0.62$, $P = 0.001$, 24 loci; as above). However, a striking aspect of the results presented in Figure 2 is that estimates of $\hat{N}_\rho$ are rather large near the telomere of the X chromosome (*i.e.*, loci to the far left of the graph), despite the fact that the crossing-over rate and levels of diversity are considerably reduced relative to other regions on the X (APPENDIX A and Table 3). Under the standard neutral model, a measure that should be independent of the effective population size is ρ/θ ($= 4N_e r/4N_e \mu$), which we estimate as $\hat{\rho}/\hat{\theta}$ (HUDSON 1987; ANDOLFATTO and PRZEWORSKI 2000). Since we assume that the mutation rate is constant across loci, we expect estimates of $\hat{\rho}/\hat{\theta}$ to positively covary with $\hat{r}$. We detect no such correlation (Figure 4c; $\hat{\rho}_{W00}/\hat{\theta}_\pi$, $R = 0.01$, 24 loci, $P > 0.05$; $\hat{\rho}_{H01}/\hat{\theta}_\pi$, $R = 0.13$, 23 loci, $P > 0.05$; as above).

The insensitivity of $\hat{\rho}/\hat{\theta}$ to changes in $\hat{r}$ (Figure 4c) suggests that the correlation between $\hat{\rho}$ and $\hat{r}$ (Figure 4a) is driven by the relationship between $\hat{\theta}$ and $\hat{r}$ (Figure 4b). In fact, $\hat{N}_\rho$ and $\hat{r}$ are significantly *negatively* correlated [Figure 4d; $\hat{N}_\rho$ ($\hat{\rho}_{W00}$), $R = -0.51$, 24 loci, $P = 0.01$; $\hat{N}_\rho$ ($\hat{\rho}_{H01}$), $R = -0.27$, 23 loci, $P > 0.05$; as above], the opposite trend to that expected on the basis of the relationship between $\hat{\theta}$ and $\hat{r}$. In other words, there is less diversity in regions of low crossing over (Figure 4b), but also less LD than expected given levels of diversity and estimated rates of crossing over. This can also be seen in Figure 4e, which plots $\hat{N}_\rho/\hat{N}_\theta$ (estimated from $\hat{\rho}_{W00}$ and $\hat{\theta}_\pi$) as a function of $\hat{r}$. Under the standard neutral model, N_ρ/N_θ should be ≈ 1 and independent
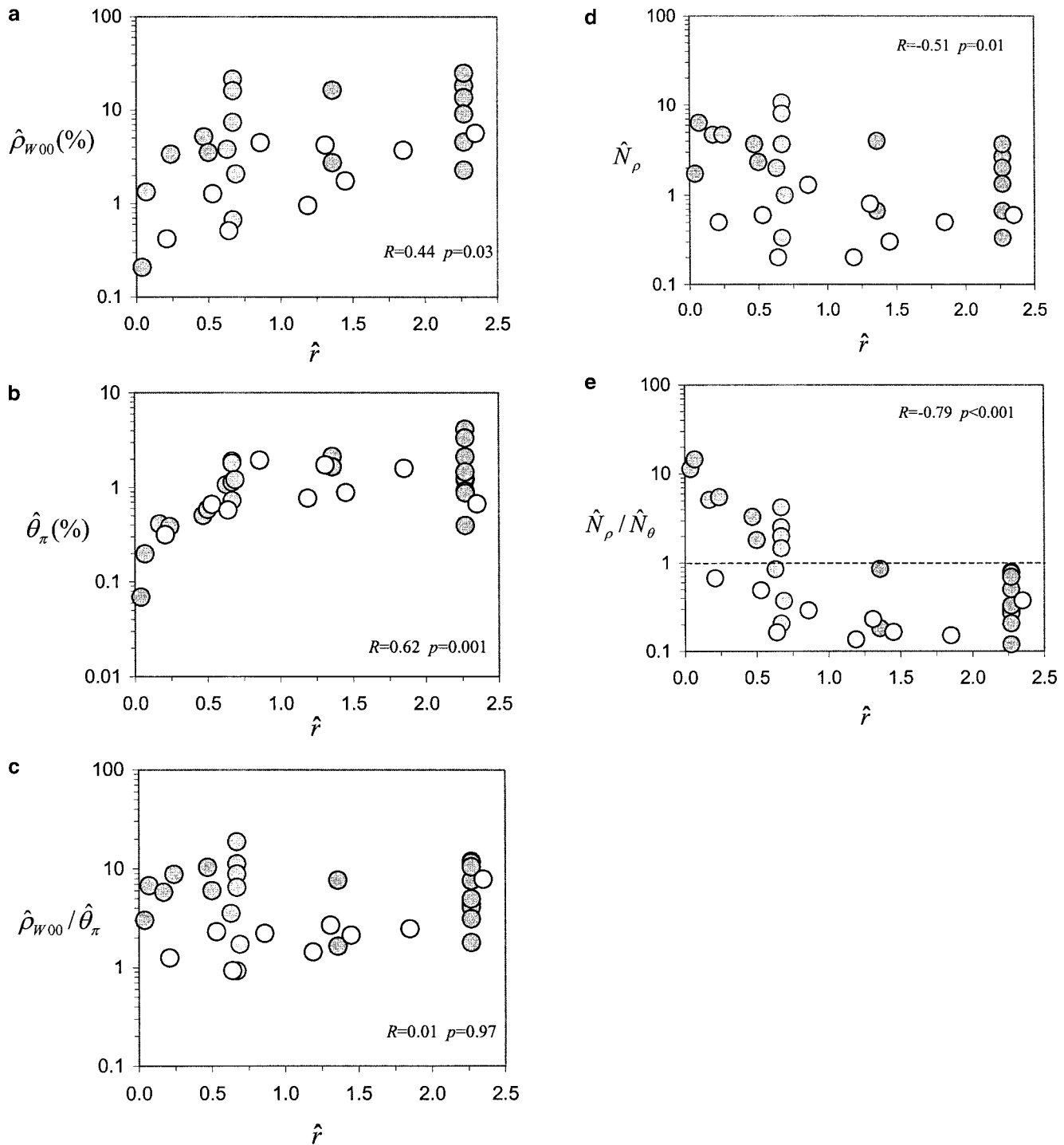
FIGURE 4.—Relationships between (a) $\hat{\rho}_{W00}$ per 100 bp, (b) $\hat{\theta}_{\pi}$ per 100 silent sites, (c) $\hat{\rho}_{W00}/\hat{\theta}_{\pi}$, (d) $\hat{N}_{\rho}$ ($\hat{\rho}_{W00}$), and (e) $\hat{N}_{\rho}$ ($\hat{\rho}_{W00}$)/ $\hat{N}_{\theta}$ ($\hat{\theta}_{\pi}$) vs. sex-averaged $\hat{r}$. Estimates for 24 X-linked loci are plotted as shaded circles; estimates for 9 autosomal loci are plotted as open circles. Spearman correlation coefficients (R) and two-tailed P values are given for X-linked loci only. The dotted line in e indicates the neutral equilibrium expectation for N_{ρ}/N_{θ}.

of r. While this prediction is roughly true for the 11 X-linked loci with $\hat{r} > 1$ cM/Mb ($\hat{N}_{\rho}/\hat{N}_{\theta} \leq 1$), loci with $\hat{r} < 0.5$ cM/Mb show considerably higher values of $\hat{N}_{\rho}/\hat{N}_{\theta}$. Overall, there is a strong negative correlation between $\hat{N}_{\rho}/\hat{N}_{\theta}$ and $\hat{r}$ (Figure 4e; $R = -0.79$, 24 loci, $P < 0.001$; as above), implying that regions of reduced crossing

over have considerably less LD than predicted by rates of crossing over under the standard neutral model.

Could a bias in $\hat{\rho}_{W00}$ estimates explain the observed correlations? In simulations of the neutral model (Table 1), we noted a slight upward bias in mean $\hat{\rho}$ estimates with decreasing θ. However, our correlation result
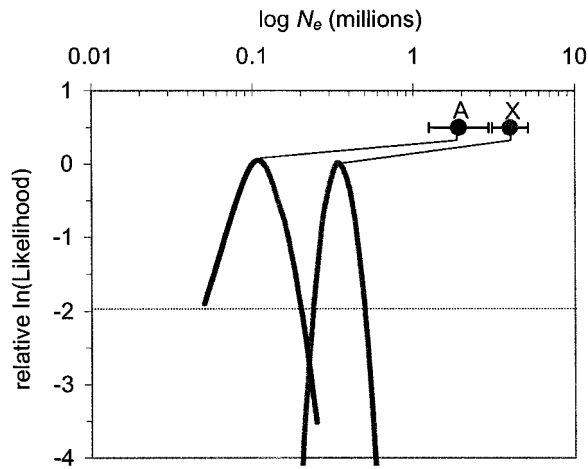
FIGURE 5.—Joint maximum-likelihood estimates of $\hat{N}_\rho$ (curves) and $\hat{N}_\theta$ (points with ~95% confidence intervals) under gene conversion (model GC1, see METHODS). The dotted line indicates the approximate 95% confidence bounds using the standard chi-square approximation.

(which employs a nonparametric test) depends more on median estimates of $\hat{\rho}_{W00}$, which were typically lower than the actual ρ in these simulations (see Table 1). Additional simulations with parameters appropriate for the *yellow* locus ($n = 49$, $S = 18$, $\rho \approx \theta/4$) confirmed that the median $\hat{\rho}_{W00}$ is downward biased (see METHODS). Thus, bias in $\hat{\rho}_{W00}$ is not an explanation for the negative correlation between $\hat{N}_\rho/\hat{N}_\theta$ and $\hat{r}$.

Gene conversion and LD patterns in regions of high crossing over: Our estimates of recombination rate are based on large-scale comparisons of genetic and physical maps that ignore the contribution of gene conversion to the total rate of recombination. At small physical scales (*i.e.*, <1 kb in Drosophila), the contribution of gene conversion to the overall rate of recombination might be substantial (ANDOLFATTO and NORDBORG 1998). The precise relationship between crossing over and gene conversion is not well understood in Drosophila, but the consensus view is that they are associated processes that result from Holliday junction formation and resolution during meiotic recombination (CARPENTER 1984).

Given that gene conversion is a potentially important contributor to the overall rate of recombination, we reexamine the relationship between levels of LD relative to levels of variability (measured as joint $\hat{N}$ for loci with $\hat{r} > 1$ cM/Mb), assuming a model that includes both crossing over and gene conversion (*e.g.*, GC1; see METHODS). In general, we find that, if gene conversion contributes substantially to the overall rate of recombination in regions of high crossing over, there is clear evidence for an excess of LD given levels of diversity in the Zimbabwe data. As an illustration, we present results for gene conversion model GC1 in Figure 5 (this model assumes that two-thirds of recombination events are gene conversions without associated crossing over; see

METHODS). Under this model, $\hat{N}_{\rho X} = 0.33$ million for the X chromosome (~95% confidence interval of 0.24–0.50) and $\hat{N}_{\rho A} = 0.15$ million for the autosomes (~95% confidence interval of 0.04–0.20). Since $\hat{N}_\rho$ and $\hat{N}_\theta$ differ by more than an order of magnitude under this model, this discrepancy is not likely to be explained by uncertainty in the mutation rate (about twofold, see METHODS). Thus, our results suggest either that the level of genetic exchange due to gene conversion is close to negligible in regions of high crossing over or that this Zimbabwean population departs substantially from the predictions of mutation-drift equilibrium in the direction of a excess of LD. Note that estimates of $\hat{N}_\rho$ for *D. simulans* (see Figure 3b) did not include gene conversion, so the claim that the Zimbabwean population is *closer* to a population that is at mutation-drift equilibrium is still a valid one.

DISCUSSION

LD, population history, and gene conversion: Previous studies have found $\hat{N}_\rho \ll \hat{N}_\theta$ for many genes in both *D. melanogaster* and *D. simulans*, which has been interpreted as a genome-wide excess of LD in these two species (ANDOLFATTO and PRZEWORSKI 2000; WALL *et al.* 2002). Here, we find that $\hat{N}_\rho$ (based on rates of crossing over) are much closer to $\hat{N}_\theta$ estimates in Zimbabwean populations of *D. melanogaster* (Figures 2 and 3), suggesting that the latter populations may be closer to expectations under mutation-drift equilibrium (with respect to LD patterns). The marked contrast in LD patterns observed between this and earlier studies (*e.g.*, Figure 3b) indicates that the differences between $\hat{N}_\rho$ and $\hat{N}_\theta$ estimates in non-African populations of *D. melanogaster* and *D. simulans* are not due to systematic errors in the methodology or estimates of r and μ. Rather, these findings point to differences in the history of the populations (or species) studied. In particular, Zimbabwean populations of *D. melanogaster* may have a more stable demographic history than non-African populations, which has led to large differences between these populations in patterns of LD. This may be expected since Zimbabwean *D. melanogaster* populations are closer to the species' ancestral range (LACHAISE *et al.* 1988).

Our results may also shed some light on the importance of gene conversion (*i.e.*, recombination without the exchange of flanking markers) relative to crossing over in Drosophila. Estimates of r based on the large-scale comparisons of genetic and physical maps essentially measure only the crossing-over rate; when there is gene conversion, they may considerably underestimate the actual rate of genetic exchange (ANDOLFATTO and NORDBORG 1998). In contrast to Drosophila, previous studies of human sequence polymorphism data have found that $\hat{N}_\rho$ (based on rates of crossing over) were generally much larger than $\hat{N}_\theta$ (FRISSE *et al.* 2001; PRZEWORSKI and WALL 2001), which was interpreted as evi-

dence that gene conversion may be having a significant impact on LD patterns in humans. If the same situation were true in *D. melanogaster*, we also might expect to generally observe $\hat{N}_\rho \gg \hat{N}_\theta$.

In fact, estimates of $\hat{N}_\rho$ based on rates of crossing over are lower than $\hat{N}_\theta$ in Zimbabwean *D. melanogaster* populations, suggesting that, if these populations are near a mutation-drift equilibrium, levels of exchange due to gene conversion cannot be high in regions of high crossing over. Alternatively, if gene conversion is an important force in regions of high crossing over, we must conclude that these populations depart from equilibrium model expectations. For example, assuming that gene conversion contributes as much to intragenic recombination as crossing over implies that there is actually a marked excess of LD in these populations (Figure 5). Distinguishing between these alternatives is not possible without better estimates of gene conversion parameters in regions of high crossing over (see below).

Uncertainty about the relative importance of gene conversion does not affect our conclusion that population history is an important determinant of patterns of LD in Drosophila. Earlier studies, primarily on non-African Drosophila populations, have ignored the possible impact of gene conversion on LD patterns; including gene conversion would make the observed departures from the standard neutral model even more extreme (Andolfatto and Przeworski 2000; Wall *et al.* 2002). Thus, even if Zimbabwean populations of *D. melanogaster* also depart from the predictions of mutation-drift equilibrium, the discrepancy is smaller than that in previous studies, which focused on non-African populations (see Figure 3b).

Patterns of LD across a recombination gradient: Estimates of the local effective population size based on patterns of diversity ($\hat{N}_\theta$) are strongly positively correlated with estimated rates of crossing over in Drosophila (Begun and Aquadro 1992; Aquadro *et al.* 1994; Andolfatto and Przeworski 2001), as expected under background selection and/or recurrent selective sweep models. It is less clear why we observe a *negative* correlation between $\hat{N}_\rho$ and $\hat{r}$ (Figure 4d) or why $\hat{N}_\rho/\hat{N}_\theta$ are strongly negatively correlated with $\hat{r}$ (Figure 4e).

Could either background selection or recurrent hitchhiking explain the patterns observed in Figure 4, d and e? If background selection due to strongly deleterious mutations can be reasonably well approximated by a reduction in effective population size (Hudson and Kaplan 1994), it can be modeled using the standard neutral model by invoking a smaller local effective population size (*i.e.*, smaller ρ and θ, but with ρ/θ remaining the same) with decreasing *r*. Thus, we expect $\hat{N}_\theta$ and $\hat{N}_\rho$ to be reduced on average by roughly the same amount in regions of lower crossing over. Contrary to this expectation, the ratio $\hat{N}_\rho/\hat{N}_\theta$ *increases* as $\hat{r}$ decreases (Figure 4e). Since the background selection model predicts a strongly positive correlation between $\hat{N}_\rho$ and $\hat{r}$ (and since

the median of $\hat{\rho}_{W00}$ is downward biased with decreasing θ), we can safely reject this model as an explanation for the patterns observed in Figure 4, d and e.

If mutations are only slightly deleterious, background selection may not be well approximated by a simple reduction in effective population size (*e.g.*, Charlesworth *et al.* 1995; McVean and Charlesworth 2000; Tachida 2000; Comeron and Kreitman 2002; Gordo *et al.* 2002). Under such a model, deleterious alleles will be in repulsion disequilibrium (Hill and Robertson 1966), as may be linked neutral variation (Comeron and Kreitman 2002). However, it is not clear how such a model would affect our estimates of $\hat{\rho}$. In particular, the resulting negative skew in the frequency spectrum of linked neutral variants (Tachida 2000; Comeron and Kreitman 2002; Gordo *et al.* 2002) leads to an increase in the number of haplotypes (*H*) per polymorphic mutation, but the minimum number of recombination events (R_M) will decrease. Thus, the combined effect on $\hat{\rho}_{W00}$ is difficult to predict. Likewise, the performance of $\hat{\rho}_{H01}$ under departures from the neutral model has also not been investigated.

Simple recurrent selective sweep models are also known to skew the frequency spectrum of segregating polymorphisms toward an excess of rare variants (*e.g.*, Braverman *et al.* 1995). A significant skew toward rare polymorphisms in regions of reduced crossing over in this Zimbabwean *D. melanogaster* population suggests that some form of hitchhiking may be operating (Andolfatto and Przeworski 2001). We have performed a set of preliminary simulations to examine the effects of recurrent selective sweeps on estimators of ρ (Table 2). As the rate of selective sweeps increases, the level of polymorphism decreases, yielding smaller $\hat{N}_\theta$ estimates on average. However, selective sweeps also affect levels of LD, and the average $\hat{N}_\rho$ value (as estimated by $\hat{\rho}_{W00}$) is reduced by roughly the same amount as the average $\hat{N}_\theta$ value (Table 2). Note that this does not mean that selective sweeps act as a simple reduction in the effective population size, but merely that they tend to decrease $\hat{\rho}_{W00}$ estimates by roughly as much as they decrease θ estimates (though we also note that $\hat{\rho}_{H01}$ responds differently; see results). In summary, it is clear from our simulations (Table 2 and supplementary information available at http://helios.bto.ed.ac.uk/evolgen/andolfatto/zimbabweLD) that a recurrent selective sweep model cannot explain the strong negative correlation between $\hat{N}_\rho/\hat{N}_\theta$ and $\hat{r}$ (Figure 4e).

Evidence for heterogeneity in gene conversion rates across the genome: As discussed earlier, our $\hat{r}$ estimates do not reflect the true rate of genetic exchange when there is gene conversion. In particular, if the relative level of exchange due to gene conversion were higher in regions of reduced crossing over, then $\hat{N}_\rho/\hat{N}_\theta$ would be negatively correlated with $\hat{r}$, as observed in Figure 4e. Langley and colleagues have suggested that such a model may account for the rapid rate of decay of pair-

TABLE 4

Patterns of LD on the X chromosome under four models of gene conversion

Locus	$\hat{r}$ (cM/Mb)	$\hat{N}_\rho/\hat{N}_\theta$[a]			
		No GC	GC1	GC2	GC3
yellow	0.04	11.8	3.3	0.10	0.20
su(s)	0.07	14.4	6.1	0.23	0.68
su(w^a)	0.17	5.1	1.6	0.16	0.38
snf1a	0.24	5.5	1.8	0.35	0.71
Pgd	0.47	3.3	0.91	0.18	0.45
vinculin	0.50	1.8	0.46	0.12	0.30
Highest CO[b]	2.27	0.43	0.08	0.08	0.43

[a] $\hat{N}_\rho$ and $\hat{N}_\theta$ estimates are based on $\hat{\rho}_{W00}$ and $\hat{\theta}_\pi$, respectively.

[b] Average $\hat{N}_\rho/\hat{N}_\theta$ value for X-linked loci with $\hat{r} = 2.27$ cM/Mb.

wise LD with physical distance in regions of reduced crossing over at the tip of the X chromosome (LANGLEY *et al.* 2000). A similarly rapid decay in LD is observed on the fourth chromosome of *D. melanogaster* (JENSEN *et al.* 2002). Since this chromosome lacks crossing over under normal conditions, high levels of gene conversion may also account for lower than expected levels of LD.

Data on gene conversion in Drosophila are sparse. To date, perhaps the best estimation of gene conversion rates in *D. melanogaster* is the detailed study of the *rosy* locus (*e.g.*, HILLIKER and CHOVNICK 1981; HILLIKER *et al.* 1994). The authors estimate that up to 80% of meiotic recombination events may be resolved as gene conversion events at *rosy*, but it is not clear how general these results are. For example, data from the *maroon-like* locus suggest that gene conversions may account for ~50% of recombination events (FINNERTY 1976). Since *rosy* and *maroon-like* are both in regions of intermediate levels of crossing over (*rosy*, $\hat{r} = 0.7$ cM/Mb; *maroon-like*, $\hat{r} = 1.36$ cM/Mb), this degree of association may not generalize to regions of the genome with higher rates of crossing over. This said, the only measurement of gene conversion in a region of high crossing over, though not easily interpreted, suggests that the contribution of gene conversion is not negligible (*rudimentary*; see FINNERTY 1976).

How might GC and CO covary across the genome? To answer this question, we investigated the effect of three possible models of gene conversion on estimates of $\hat{N}_\rho$ across the X chromosome. In the first model (GC1), rates of gene conversion and crossing over positively covary. In the second model (GC2), gene conversion rates are constant across the X chromosome. In the third model (GC3), gene conversion and crossing over negatively covary (see METHODS for details). We measure the "fit" of a gene conversion model on the basis of two features of the data. The first is that, if the total recombination rate (the sum of GC + CO) is

adequately accounted for at each gene, estimates of $\hat{N}_\rho/\hat{N}_\theta$ should be roughly equal in regions of high and low crossing over (*i.e.*, the negative correlation between $\hat{N}_\rho/\hat{N}_\theta$ and $\hat{r}$ should disappear). A second measure of fit is the absolute value of $\hat{N}_\rho/\hat{N}_\theta$. Assuming that this Zimbabwe population is near a mutation-drift equilibrium, we expect $\hat{N}_\rho/\hat{N}_\theta$ to be ~1.

In Table 4, we list estimates of $\hat{N}_\rho/\hat{N}_\theta$ for *yellow, su(s), su(w^a), snf1a, Pgd*, and *vinculin* (see APPENDIX A for details) under these models and compare them to the average estimate in regions of the highest crossing over on the X chromosome ($\hat{r} = 2.27$ cM/Mb). As can be seen, while the model in which GC and CO positively covary (GC1) reduces $\hat{N}_\rho/\hat{N}_\theta$ estimates in regions of reduced crossing over, the negative correlation between $\hat{N}_\rho/\hat{N}_\theta$ and $\hat{r}$ remains. Models GC2 and GC3 are better fits to the data because they yield estimates of $\hat{N}_\rho/\hat{N}_\theta$ that are similar in regions of high and low crossing over. Overall, the model in which gene conversion and crossing-over rates negatively covary (GC3) appears to be the most reasonable explanation because it explains the correlation between $\hat{N}_\rho/\hat{N}_\theta$ and $\hat{r}$ without the need to invoke a large LD excess in this population. However, data on gene conversion from the *rudimentary* locus (reviewed above) suggest that our GC3 model may underestimate the impact of gene conversion on patterns of LD in regions of high crossing over. If this is the case, our results suggest a departure from equilibrium expectations in this population.

The above analysis is intended to be only illustrative as many other models of gene conversion are possible. For example, some form of heterozygosity-dependent gene conversion may be operating (STEPHAN and LANGLEY 1992; LANGLEY *et al.* 2000). If so, regions with reduced crossing over, which tend to have less variability (see Figure 4b), may have higher rates of gene conversion and/or longer gene conversion tracts (see LANGLEY *et al.* 2000 for a discussion). Note that, in our models, we have assumed that the distribution of gene conversion tract lengths is the same in regions of high and low crossing over. A heterozygosity-dependent gene conversion model has not yet been explored enough to make quantitative predictions, but it has the potential to explain the negative $\hat{N}_\rho/\hat{N}_\theta$ since it is qualitatively similar to our GC3 model. Further distinguishing between the models we have presented, and confidently inferring whether Zimbabwean *D. melanogaster* populations are near a mutation-drift equilibrium, will require a more detailed investigation of the gene conversion parameters in regions of high and low crossing over.

Implications for weak selection models and recombination-associated mutation biases: While gene conversion may explain lower than expected LD in regions of reduced crossing over, it does not pose a problem for the interpretation of the strongly positive correlation between $\hat{N}_\theta$ and $\hat{r}$. Models of recurrent hitchhiking (BRAVERMAN *et al.* 1995) or background selection (CHARLESWORTH 1996)

invoke relatively strong selection to account for this correlation (*i.e.*, selection that will affect relatively large chromosomal segments). The effects of these selection models on levels of variability are primarily determined by the crossing over rate, since gene conversion contributes little to the total recombination rate at distances much greater than the average gene conversion tract length (ANDOLFATTO and NORDBORG 1998). However, gene conversion may be relevant to models that posit weaker selection, such as models invoking weak selection on synonymous codon usage or some fraction of amino acid replacement mutations (*e.g.*, LI 1987; MCVEAN and CHARLESWORTH 2000; TACHIDA 2000; COMERON and KREITMAN 2002). Weakly selected mutations are predicted to interfere with the efficacy of selection at closely linked sites, a phenomenon called weak Hill-Robertson interference (wHRi). Our results suggest that gene conversion may limit the magnitude of wHRi, particularly in regions of low crossing over, where wHRi effects are predicted to be the most important.

Finally, we note that a correlation between synonymous site G + C base composition and rates of crossing over has been documented in *D. melanogaster* (KLIMAN and HEY 1993). This pattern has been interpreted as evidence of reduced efficacy of weak selection for preferred (*i.e.*, G or C ending) codons in regions of low crossing over due to Hill-Robertson interference in these regions. This view has recently been challenged on the basis of the observation that G + C content of noncoding DNA also appears to correlate with rates of crossing over (MARAIS *et al.* 2001; MARAIS and PIGANEAU 2002), suggesting a recombination-mediated mutation bias [the biased gene conversion (BGC) hypothesis; BIRDSELL 2002; MARAIS 2003]. Invoking the BGC hypothesis to explain the correlation between G + C content and rates of crossing over presupposes that gene conversion and crossing over positively covary (similar to our model GC1). However, our results based on LD patterns suggest that gene conversion and crossing over are not associated in this way, at least at the tip of the X chromosome in Drosophila. The BGC hypothesis may not explain differences in base composition in regions of high *vs.* low crossing over if rates of gene conversion are constant across the genome or negatively covary with rates of crossing over.

It is worth noting that G + C content of noncoding DNA in telomeric regions of Drosophila appears to be higher than that in centromeric regions (MARAIS *et al.* 2001; MARAIS and PIGANEAU 2002), despite both being regions of reduced crossing over (CHARLESWORTH 1996). Thus, our results may be consistent with the BGC hypothesis if high rates of gene conversion occur in telomeric regions but not in other regions of reduced crossing over. However, noncoding regions on chromosome 4 of *D. melanogaster* (which lacks crossing over) have the lowest G + C content in the genome (MARAIS *et al.* 2001; G. MARAIS, personal communication), yet

polymorphism data from the fourth chromosome suggest considerable levels of meiotic exchange (JENSEN *et al.* 2002). Assuming that all genetic exchange on chromosome 4 is due to gene conversion, we estimate $\hat{\rho}_{W00}/\hat{\theta}$ to be $\sim$20 (on the basis of the data of JENSEN *et al.* 2002 and assuming a gene conversion tract length of 352 bp), which is in close agreement with estimates at the tip of the X chromosome (see Figure 4c). This corresponds to a conversion rate on the order of 10^{-5} events/generation, which is remarkably similar to direct estimates at *rosy* and *maroon-like*. Thus, our results suggesting high rates of gene conversion in regions of reduced crossing over do not appear to be restricted to the tip of the X chromosome. A more detailed investigation of the physical scale of LD in other genomic regions may shed further light on the association of gene conversion with crossing over and its implications for weak selection models and the evolution of base composition.

LITERATURE CITED

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

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

ANDOLFATTO, P., and M. 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.

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

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.

AQUADRO, C. F., V. BAUER DUMONT and F. A. REED, 2001 Genome-wide variation in the human and fruitfly: a comparison. Curr. Opin. Genet. Dev. **11:** 627–634.

BAINES, J. F., Y. CHEN, A. DAS and W. STEPHAN, 2002 DNA sequence variation at a duplicated gene: excess of replacement polymorphism and extensive haplotype structure in the *Drosophila melanogaster bicoid* region. Mol. Biol. Evol. **19:** 989–998.

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

BEGUN, D. J., and C. F. AQUADRO, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365L:** 548–550.

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

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

BEGUN, D. J., P. WHITLEY, B. L. TODD, H. M. WALDRIP-DAIL and A. G. CLARK, 2000 Molecular population genetics of male accessory gland proteins in Drosophila. Genetics **156:** 1879–1888.

BIRDSELL, J. A., 2002 Integrating genomics, bioinformatics and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. **19:** 1181–1197.

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.

CARPENTER, A. T. C., 1984 Meiotic roles of crossing-over and of gene conversion. Cold Spring Harbor Symp. Quant. Biol. **49:** 23–26.

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

CHARLESWORTH, B., 2001 The effect of life history and mode of inheritance on neutral genetic variability. Genet. Res. **77:** 153–166.

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

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

COMERON, J. M., and M. KREITMAN, 2002 Population, evolutionary and genomic consequences of interference selection. Genetics **161:** 389–410.

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

EANES, W. F., M. KIRCHNER, J. YOON, C. H. BIERMANN, I. N. WANG *et al.*, 1996 Historical selection, amino acid polymorphism and lineage-specific divergence at the *G6pd* locus in *Drosophila melanogaster* and *D. simulans*. Genetics **144:** 1027–1041.

FINNERTY, V., 1976 Gene conversion in Drosophila, pp. 331–349 in *The Genetics and Biology of Drosophila*, edited by M. ASHBURNER and E. NOVITSKI. Academic Press, London/New York.

FRISSE, L., R. R. HUDSON, A. BARTOSZEWICZ, J. D. WALL, J. DONFACK *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

GORDO, I., A. NAVARRO and B. CHARLESWORTH, 2002 Muller's ratchet and the pattern of variation at a neutral locus. Genetics **161:** 835–848.

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

HARR, B., M. KAUER and C. SCHLOTTERER, 2002 Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **99:** 12949–12954.

HAWLEY, R. S., K. S. MCKIM and T. ARBEL, 1993 Meiotic segregation in *Drosophila melanogaster* females: molecules, mechanisms and myths. Annu. Rev. Genet. **27:** 288–317.

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

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

HILLIKER, A. J., and A. CHOVNICK, 1981 Further observations of intragenic recombination in *Drosophila melanogaster*. Genet. Res. **38:** 281–296.

HILLIKER, A. J., G. HARAUZ, A. G. REAUME, M. GRAY, S. H. CLARK *et al.*, 1994 Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. Genetics **137:** 1019–1026.

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

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. J. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford.

HUDSON, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Japan Scientific Society, Tokyo.

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

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

HUDSON, R. R., and N. L. KAPLAN, 1994 Gene trees with background selection, pp. 140–153 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. GOLDING. Chapman & Hall, London.

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

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitch-hiking effect" revisited. Genetics **123:** 887–899.

KAUER, M., B. ZANGERL, D. DIERINGER and C. SCHLOTTERER, 2002 Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. Genetics **160:** 247–256.

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

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

LACHAISE, D., L. M. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159–225.

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(wᵃ)* regions of the *Drosophila melanogaster* X chromosome. Genetics **156:** 1837–1852.

LEMEUNIER, F., and S. AULARD, 1992 Inversion polymorphism in *Drosophila melanogaster*, pp. 339–405 in *Drosophila Inversions Polymorphism*, edited by C. B. KRIMBAS and J. R. POWELL. CRC Press, Boca Raton, FL.

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

LI, W. H., 1997 *Molecular Evolution.* Sinauer Press, Sunderland, MA.

MARAIS, G., 2003 Biased gene conversion: implications for genome and sex evolution. Trends Genet. **19:** 330–338.

MARAIS, G., and G. PIGANEAU, 2002 Hill-Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. Mol. Biol. Evol. **19:** 1399–1406.

MARAIS, G., D. MOUCHIROUD and L. DURET, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98:** 5688–5692.

MASIDE, X., C. BARTOLOME and B. CHARLESWORTH, 2002 S-element insertions are associated with the evolution of the *Hsp70* genes in *Drosophila melanogaster*. Curr. Biol. **12:** 1686–1691.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of favorable genes. Genet. Res. **23:** 23–35.

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

ODGERS, W. A., C. F. AQUADRO, C. W. COPPIN, M. J. HEALY and J. G. OAKESHOTT, 2002 Nucleotide polymorphism in the Est6 promoter, which is widespread in derived populations of *Drosophila melanogaster*, changes the level of *Esterase 6* expressed in the male ejaculatory duct. Genetics **162:** 785–797.

PRZEWORSKI, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179–1189.

PRZEWORSKI, M., and J. D. WALL, 2001 Why is there so little intragenic linkage disequilibrium in humans? Genet. Res. **77:** 143–151.

Stephan, W., and C. H. Langley, 1992 Evolutionary consequences of DNA mismatch inhibited repair opportunity. Genetics **132:** 567–574.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.

Szostak, J. W., T. Orr-Weaver, R. J. Rothstein and F. W. Stahl, 1983 The double-strand-break repair model for recombination. Cell **33:** 25–35.

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

Tishkoff, S. A., E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd *et al.*, 1996 Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. Science **271:** 1380–1387.

Tsaur, S. C., C. T. Ting and C.-I Wu, 1998 Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa*, of Drosophila: II. Divergence versus polymorphism. Mol. Biol. Evol. **15:** 1040–1046.

Verrelli, B. C., and W. F. Eanes, 2000 Extensive amino acid polymorphism at the *Pgm* locus is consistent with adaptive protein evolution in *Drosophila melanogaster*. Genetics **156:** 1737–1752.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Wall, J. D., 2001 Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. Curr. Opin. Genet. Dev. **11:** 647–651.

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

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

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

Wright, S. I., B. Lauga and D. Charlesworth, 2003 Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Mol. Ecol. **12:** 1247–1263.

Communicating editor: W. Stephan

APPENDIX A

Summary statistics for each gene region used in this study

Locus	Position[a]	n[b]	Total bp[c]	S_{tot}[d]	Silent bp[e]	S_{SNP}[f]	H[g]	R_M[h]	$\hat{r}$[i] (cM/Mb)
X chromosome									
yellow	1B1	49	2017	18	2013	16	14	1	0.04
su(s)	1B13	50	4216	46	3213	37	37	6	0.07
su(w^a)	1E1	50	2571	50	1955	39	37	7	0.17
snf1a	2A1	25	9330	35	2094.7	32	24	7	0.24
Pgd	2D4	13	540	14	538	12	11	1	0.47
vinc	2E2	13	546	11	544	10	9	1	0.50
zeste	3A3	13	596	24	581	23	12	2	0.63
shaggy	3B1	12	1250	39	770.7	31	12	9	0.67
period	3B2	16	1320	60	866	62	19	16	0.67
100G10.2	3B3	19	1345	58	775.0	57	19	14	0.67
syx4	3B4	16	1339	22	634.5	15	12	1	0.67
frag3	3B5	11	1332	49	1292	45	11	8	0.67
white	3C1	16	5002	45	800.5	49	16	4	0.67
frag4	3C7	7	851	21	845	20	7	3	2.27
frag7	3C7	7	954	39	923	38	7	4	2.27
Fasciclin-2	4B2	17	588	24	498.2	22	15	4	2.27
spaghetti squash	5D6	16	572	28	568	26	12	2	2.27
hyperkinetic	9B7	21	563	64	465	66	19	12	2.27
vermillion	10A1	11	2076	78	1287.9	77	11	21	2.27
dusky	10E1	20	567	55	538	56	17	10	2.27
licorne	11D6	23	615	26	565.7	26	17	3	2.27
rutabega	12F5	22	548	12	443.1	10	13	1	2.27
g6pd (Zw)	18D13	12	1682	34	482.5	31	12	9	1.36
runt	19E2	7	1931	33	791.6	33	7	3	1.36
Autosomal									
Acp26Aa/Ab	26A1	10	1347	20	529.6	10	9	4	2.35
In(2L)t-PBP	34A8	10	665	17	640	16	8	1	1.45
Adh	35B3	10	2007	33	1428.7	31	10	2	1.19
Acp36DE	36E2	7	2268	37	502.3	24	6	6	0.86
Hexokinase C	51F5	11	1365	19	334.5	17	9	4	1.31
Esterase-6	69A1	12	981	38	915	44	12	4	1.85
Pgm	72D8	13	2354	30	1078.8	22	12	3	0.53
bicoid	84A5	25	4020	42	2866.9	34	13	6	0.21
Hsp70Bb	87B14	12	3386	43	1963.4	34	11	3	0.64

[a] Cytological position of each marker.

[b] Sample size.

[c] Total length of the region surveyed in base pairs.

[d] Number of biallelic segregating mutations used for linkage disequilibrium analysis including insertion-deletion polymorphisms, but excluding multiply hit sites.

[e] Total number of silent sites surveyed.

[f] Number of silent segregating mutations used for diversity analysis excluding insertion-deletion, but including multiply hit sites.

[g] The number of distinct haplotypes in a sample.

[h] The minimum number of inferred recombination events (*cf.* HUDSON and KAPLAN 1985).

[i] The sex-averaged estimate of the rate of crossing over.