Nonadaptive Explanations for Signatures of Partial Selective Sweeps in *Drosophila*

J. Michael Macpherson, Josefa González,* Daniela M. Witten,† Jerel C. Davis,* Noah A. Rosenberg,‡ Aaron E. Hirsh,§ and Dmitri A. Petrov**

*Department of Biological Sciences, Stanford University; †Department of Statistics, Stanford University; ‡Department of Human Genetics, University of Michigan, Ann Arbor; and §Department of Ecology & Evolutionary Biology, University of Colorado at Boulder

A beneficial mutation that has nearly but not yet fixed in a population produces a characteristic haplotype configuration, called a partial selective sweep. Whether nonadaptive processes might generate similar haplotype configurations has not been extensively explored. Here, we consider 5 population genetic data sets taken from regions flanking high-frequency transposable elements in North American strains of *Drosophila melanogaster*, each of which appears to be consistent with the expectations of a partial selective sweep. We use coalescent simulations to explore whether incorporation of the species' demographic history, purifying selection against the element, or suppression of recombination caused by the element could generate putatively adaptive haplotype configurations. Whereas most of the data sets would be rejected as nonneutral under the standard neutral null model, only the data set for which there is strong external evidence in support of an adaptive transposition appears to be nonneutral under the more complex null model and in particular when demography is taken into account. High-frequency, derived mutations from a recently bottlenecked population, such as we study here, are of great interest to evolutionary genetics in the context of scans for adaptive events; we discuss the broader implications of our findings in this context.

Introduction

Judging whether a mutation is adaptive is a question of central importance to evolutionary genetics. When an adaptive, or positively selected, mutation sweeps through a population, it leaves behind several traces, including reduced linked neutral variation (Maynard Smith and Haigh 1974; Kaplan et al. 1988, 1989), excessive low- and high-frequency polymorphisms (Tajima 1989a; Fay and Wu (2000), and unusually long haplotypes associated with the beneficial mutant (Sabeti et al. 2002; Kim and Nielsen 2004; Voight et al. 2006); each of these signatures has been used to identify probable instances of positive selection. However, the concurrent operation of nonadaptive evolutionary forces, perhaps most potently the force of demographic change, can mimic one or more of the signatures of adaptation. This greatly complicates the task of inferring the action of positive selection (Kreitman 2000; Andolfatto and Przeworski 2001; Teshima et al. 2006; Thornton et al. 2007). In view of this difficulty, it is essential to specify the null hypothesis accurately to avoid spurious inference of positive selection.

Transposable elements are increasingly being recognized as agents of adaptive change (Brosius 2003; Brookfield 2004, 2005). Although characteristically under negative, or purifying, selection (Montgomery et al. 1987; Nuzhdin 1999; Petrov et al. 2003), several instances of likely adaptive transposition in which transposons confer insecticide resistance to their respective hosts have been discovered in recent years (Daborn et al. 2002; McCollum et al. 2002; Catania et al. 2004; Schlenke and Begun 2004; Aminetzach et al. 2005; Chung et al. 2007). Furthermore, the intriguing possibility is emerging that transposable elements might play a vital role in the evolution of genomic regulatory systems, providing the raw material for the action of regulatory control sequences (Davidson and Britten 1973; Brosius 2003; Bejerano et al. 2006). Because population genetic analyses of long haplotypes are important tools in identifying and characterizing adaptive transposition, it is important to form a suitable null hypothesis for this form of variation.

An accurate null hypothesis for a putatively adaptive transposable element clearly must account for demographic history. In addition, transposon-specific forces might be expected to affect the pattern of polymorphism near an insertion. First, there is considerable support for the notion that transposition into a functional genomic region can have deleterious effects and that chromosomal rearrangements caused by ectopic recombination due to nearby transposons may also be deleterious (Montgomery et al. 1987; Nuzhdin 1999; Petrov et al. 2003). If transposons are often selected against, then using a null model that assumes that they segregate neutrally may be inappropriate. Second, there is evidence to suggest that transposons disrupt nearby recombination in heterozygous individuals (Clark et al. 1986, 1988). Because the signatures of a selective sweep depend critically on recombination events (Maynard Smith and Haigh 1974; Hudson and Kaplan 1988; Przeworski 2002; Kim and Nielsen 2004), recombination suppression might also alter the expected pattern of polymorphism near a transposon.

Here, we employ coalescent simulations to explore how each of these 3 possibilities, namely demography, purifying selection, and recombination suppression, affects our interpretation of polymorphism data from 5 loci that flank transposable elements segregating in *Drosophila melanogaster*. All 5 data sets exhibit polymorphism patterns resembling the pattern expected under a partial selective sweep: few haplotypes and reduced nucleotide diversity linked to the element. One of these elements, a non-LTR retrotransposon from the *Doc* family called *doc1420*, was studied previously by our group (Aminetzach et al. 2005). This element was found to confer resistance to organophosphate pesticides by inserting into a coding exon of the gene *CHKov1* and truncating its protein product. The

Key words: bottleneck, transposable element, coalescent simulation, partial selective sweep.

E-mail: macpher@stanford.edu

Mol. Biol. Evol. 25(6):1025–1042. 2008
doi:10.1093/molbev/msn007
Advance Access publication January 16, 2008

other 4 loci are also non-LTR retrotransposons and are all from the *BS* family. They were selected for sequencing because, like *doc1420*, they are found at intermediate or high frequency in the non-African subpopulation of *D. melanogaster*. In contrast to *doc1420*, there is no evidence to suggest that any of the 4 *BS* elements are positively selected. None of the *BS* elements insert into a region known to be functional; although some possibility exists that one or more of the *BS* elements has experienced positive selection, we regard *doc1420* as a positive control and the *BS* elements as negative controls. Our simulation results support this hypothesis, suggesting that, for elements ascertained at high frequency in a bottlenecked population, a polymorphism pattern consistent with a partial selective sweep is not unusual. When demography is incorporated into the null hypothesis, the *doc1420* element retains a significant signature of positive selection, whereas the *BS* elements largely do not; purifying selection and recombination suppression both tend to strengthen this result. These simulations apply beyond transposable elements because high frequency–derived mutations are of particular interest in evolutionary studies; we consider the implications of our results in the broader context of genomic scans for adaptive variation.

Materials and Methods

Drosophila Strains

All strains studied in all 5 data sets are from non-African populations of *D. melanogaster*. Strains a1, a3, a6, a8, a18, and a20 are from Ann Arbor, MI (courtesy of Greg Gibson). Strains wi1, wi1.5, wi3, wi4, wi9, wi15, wi18, wi31, wi35, wi41, wi45, wi68, wi69, wi77, wi83, wi98, wi137, and wi146 are from Wolfskill Orchard, Davis, CA (courtesy of Sergey Nuzhdin). Strains we4, we7, we10, we11, we13, we17, we21, we25, we28, we33, we37, we44, we 47, we50, we57, we60, we63, we67, we70, we75, we 80, we83, we88, and we91 are from Raleigh, NC (courtesy of Greg Gibson). Strains 5A, 11B, 21A, 20A, 31A, and 33A are from Countryside Winery, Blountville, TN (courtesy of Lev Yampolsky). Strains w2, w7, w9, w11, w22, w29, and w31 are from a (non-African) worldwide collection. Based on the sequenced genome of *D. melanogaster*, we designed primers to amplify the 5′ and 3′ flanking regions of the 4 *BS* insertions. Polymerase Chain Reaction products were sequenced by Genaissance Pharmaceuticals, Inc. (Genaissance, New Haven, CT). Details about the *doc1420* strains are given in Aminetzach et al. (2005); only the North American strains from that study appear here. The FlyBase IDs corresponding to the 4 *BS* insertions are *BS3443*:FBti0019604, *BS3457*:FBti0018879, *BS3618*:FBti0019133, and *BS3730*:FBti0019410.

Recombination rates were obtained from FlyBase (Grumbling and Strelets 2006). In calculating all statistics, gapped sites were ignored and sites with more than 2 alleles were ignored.

Coalescent Simulations

We simulated polymorphism data for a neutrally evolving, recombining locus of some known length in nucleotides and known sample size. This "neutral" locus is linked to a second, "segregating" locus intended to represent the transposable element, which may or may not evolve neutrally. The simulation derives from the coalescent-based model described in Kaplan et al. (1988), Hudson and Kaplan (1988), and Przeworski (2002), in which the recombinant sample genealogy of the neutral locus is generated conditional on the frequency trajectory of a new selected allele at the segregating locus. For a review of this model, see Hein et al. (2005). Briefly, under this model the population is partitioned into 2 subpopulations, and the sample is partitioned into 2 subsamples, defined by which of the 2 alleles at the segregating locus they possess. The simulation proceeds backwards in time until full coalescence of the sample, and during this time recombination may occur between the segregating locus and the neutral locus or within the neutral locus. Mutations are then placed on the sample genealogy according to the infinite sites model. In Kaplan et al. (1988), Hudson and Kaplan (1988), and Przeworski (2002), the focus was on strongly advantageous alleles at the segregating locus, but here, we are interested in the trajectories of neutral or weakly selected alleles at the segregating locus, that is, the trajectories of the transposable elements. Because our interest is in weak selection, we describe simulation methods better suited to the case of a weakly selected variant than the methods in these earlier studies. After introducing the specific demographic histories we are considering, we describe extensions to this model that allow ancestral purifying selection and recombination suppression to be incorporated.

Demographic Models

We consider 3 demographic scenarios for the history of the non-African subpopulation from which our 5 data sets derive. The first is simply a randomly mating population of constant effective size $N_e = 10^6$, corresponding to the standard neutral null hypothesis. The second demographic scenario derives from Thornton and Andolfatto (2006) (henceforth "TA"). In this scenario, the non-African subpopulation emigrates from Africa at a time t_b^{TA}. From that time until a time t_r^{TA}, its size is instantaneously reduced to a fraction f^{TA} its ancestral African size, after which it instantaneously grows to its present size, presumed equal to the ancestral size (fig. 1*a*). We denote the identical prebottleneck and postbottleneck sizes under this scenario N_E^{TA}. Thornton and Andolfatto (2006) obtained estimates for several recombination regimes, and we used their point estimate corresponding to $\rho/\theta = 10$, where $\rho = 4Nr$ is the population recombination rate, as is appropriate to the highly recombining regions in which the 5 loci we study are found. Their point estimate is $t_b^{\mathrm{TA}} = 0.022 \times 4N_E^{\mathrm{TA}}$, $t_r^{\mathrm{TA}} = 0.0042 \times 4N_E^{\mathrm{TA}}$, and $f^{\mathrm{TA}} = 0.029$.

The third demographic scenario derives from Li and Stephan (2006) (henceforth "LS"). The difference between this and the TA bottleneck scenario is that the prebottleneck African population is presumed to have undergone an expansion in size. This model requires 3 parameters beyond f^{LS}, N_E^{LS}, t_b^{LS}, and t_r^{LS}: the respective pre-expansion and postexpansion population sizes N_{A1}^{LS} and N_{A0}^{LS}, and a time of expansion t_e^{LS} (cf., fig. 1*b*). Their point estimate is $f^{\mathrm{LS}} = 2200/1.075 \times 10^6 = 0.002$,

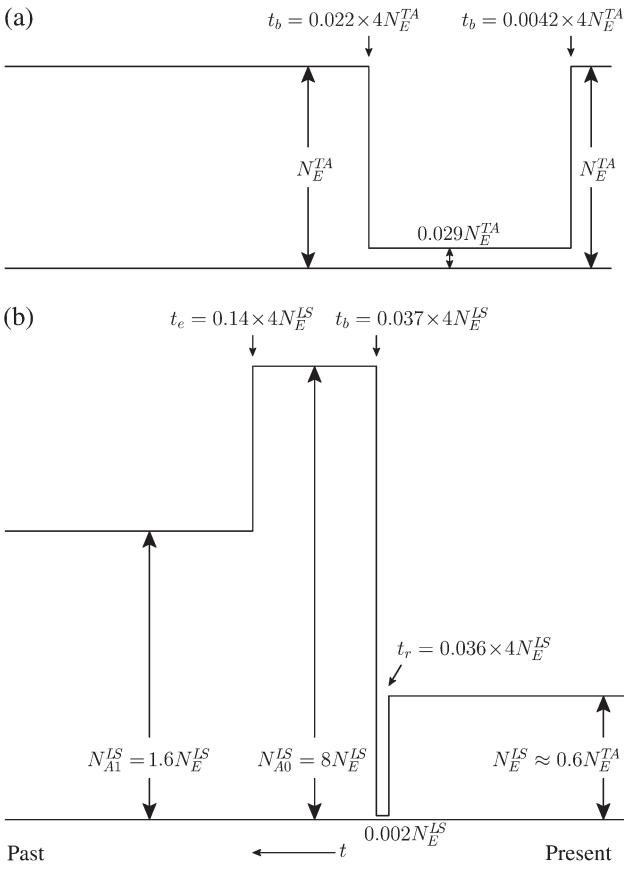

FIG. 1.—The 2 bottleneck scenarios explored. The population size is shown as a function of time. Part (*a*) corresponds to the scenario of Thornton and Andolfatto (2006), and part (*b*) corresponds to the scenario of Li and Stephan (2006).

$t_e^{\mathrm{LS}} = 60000$ y, $t_b^{\mathrm{LS}} = 15800$ y, $t_r^{\mathrm{LS}} = 15460$ y, $N_{A1}^{\mathrm{LS}} = 1.6 \times N_E^{\mathrm{LS}}$, and $N_{A0}^{\mathrm{LS}} = 8 \times N_E^{\mathrm{LS}}$. We converted the time units from years into coalescent units of 4 N_E^{LS} generations, assuming, as do both TA and LS, 10 generations per year, yielding $t_e^{\mathrm{LS}} = 0.14 \times 4N_E^{\mathrm{LS}}$, $t_b^{\mathrm{LS}} = 0.037 \times 4N_E^{\mathrm{LS}}$, $t_r^{\mathrm{LS}} = 0.036 \times 4N_E^{\mathrm{LS}}$. Both bottleneck studies explicitly provide a modern non-African population size, which depends on an assumed mutation rate of $\sim 1.5 \times 10^{-9}$ bp^{-1} gen^{-1}. These are $N_E^{\mathrm{TA}} = 2.4 \times 10^6$ and $N_E^{\mathrm{LS}} = 1.075 \times 10^6$. TA assumed equal male and female effective population sizes, and because the estimates derive from X chromosomal data, they multiplied their estimate of effective population size by 4/3. LS did not perform this multiplication, although they also used X chromosomal data, so the ratio of the non-African population size estimates is $N_E^{\mathrm{LS}} / N_E^{\mathrm{TA}} = 1.075 / \left(\left(3/4\right) \times 2.4\right) = 0.572$. To ease comparison of the 2 bottleneck studies, we use this size ratio to rescale the population size change times for LS in terms of N_E^{TA} rather than N_E^{LS} generations are then: $t_e^{\mathrm{LS}} - 0.083 \times 4N_E^{\mathrm{TA}}$, $t_b^{\mathrm{LS}} = 0.022 \times 4N_E^{\mathrm{TA}}$, $t_r^{\mathrm{LS}} = 0.021 \times 4N_E^{\mathrm{TA}}$. For computational convenience, in our simulations we used $N_E^{\mathrm{TA}} = 10^6$.

Simulation of Transposable Element Trajectories

We use 3 distinct techniques to produce the transposable element frequency trajectories at the segregating locus used in this study: 1) forward simulation by binomial sampling to a prespecified frequency, a technique we introduce here; 2) simple forward simulation by binomial sampling; and 3) reverse simulation using Slatkin's importance sampling method (Slatkin 2001). By "trajectory," we mean the full history of an allele's frequency from the origin of the mutation to the present.

To simulate transposable element trajectories under the demographic scenario of constant population size, we use only the first of these 3 techniques. First, to account for the ascertainment of the element in k of n strains sampled, we sample a frequency $x \in [1/(2N), 1 - 1/(2N)]$ from the equilibrium neutral frequency spectrum $1/(x \log(2N - 1))\, \mathrm{d}x$. If a random draw from Binomial(n, x) equals k, we accept x. If not, we discard x and repeat the procedure until this ascertainment condition is met.

Next, we wish to sample an element frequency trajectory from the distribution of all random trajectories from the Wright–Fisher process which terminate at frequency x. We assume a panmictic, constant-sized population of N diploid individuals, with genotype fitnesses 1 and $1 + s$ for the 2 homozygotes and fitness $1 + hs$ for the heterozygote, for some fixed selection coefficient s and dominance effect h. In this paper, we assume $h = 1/2$. To sample from this distribution, we generate a trajectory by iterating forward by the Wright–Fisher process from initial frequency $1/(2N)$ and continuing until either fixation or loss occurs (Ewens 2004). If the element is lost before frequency x is reached at all, that trajectory is discarded. As there may be multiple occasions at which x is reached, some choice must be made about which of these occasions will represent the end of the trajectory. We truncate the trajectory at the last generation at which frequency x was reached. A demonstration that this procedure generates allele trajectories correctly may be found in the Appendix.

To generate element frequency trajectories for the demographic scenario specified in TA, we use a combination of the truncation procedure and simple binomial sampling. First, we use simple forward iteration to generate the postbottleneck segment of the element trajectory that runs from the beginning of the bottleneck to the present. We assume that the transposable element was at transposition–drift equilibrium when the bottleneck began, and sample a frequency $x \in [1/(2N), 1 - 1/(2N)]$ from $1/(x \log(2N - 1))\, \mathrm{d}x$. This implicitly assumes that the element had entered the population by the time the bottleneck began, discussed further below. The value of N used is the ancestral African population size immediately before the bottleneck begins (cf., fig. 1*a*). We iterate from x forward in time through the bottleneck by binomial sampling until either fixation or loss occurs, or the trajectory reaches the present time. If the element becomes fixed or lost before the present, we discard the trajectory and begin anew. For those elements that are not fixed or lost, which reach the present time at some frequency x', we then account for ascertainment of the element in k of n strains by drawing from Binomial(n, x'); if the drawn value is equal to k, we accept the trajectory and if not the process is begun anew.

The prebottleneck segment of the trajectory, that ends at frequency x, is then generated according to the truncation procedure described above for the constant-sized population. The prebottleneck and postbottleneck segments of

the trajectory are then concatenated together to form the complete trajectory of the transposable element from insertion to the present day.

To generate element trajectories for the demographic scenario specified in LS, we also generate the element trajectory in 2 stages. We generate the postbottleneck segment of the trajectory exactly as described above, except that the bottleneck parameters differ as detailed above and in figure 1. To generate the prebottleneck segment of the trajectory, we use the importance sampling method of Slatkin (2001). This is done because our truncation method assumes constant population size, whereas this bottleneck scenario specifies a population size expansion in the prebottleneck population. We generate proposal trajectories backwards in time, starting from frequency x, that is, the frequency of the postbottleneck segment at the beginning of the bottleneck, according to the reverse process given in Slatkin (2001). The prebottleneck segment of the trajectory thus produced is assigned an importance weight w, which is the ratio of the segment's probability under the reverse process to its probability under the forward process (Slatkin 2001). When the mean and confidence intervals (CIs) are later computed over the distributions of summary statistics of these polymorphism data sets generated using this method, the summary statistics are weighted by these weights. As above, the prebottleneck and postbottleneck segments of the trajectory are concatenated to form a complete trajectory.

Genealogical Simulation

Recombinant genealogies were generated conditional on the transposable element trajectories described above according to the algorithm described in Hudson (1993). For each of the 5 data sets, respectively, genealogies were generated conditional on the local recombination rate, the length of the locus, and the position of the transposable element within the locus. Then, the same number of segregating sites as observed in the respective sample was distributed over the resulting genealogy according to the infinite sites model, as described in Hudson (1993).

Modeling Prebottleneck Purifying Selection

To explore the effect of purifying selection against transposable elements prior to the bottleneck, we generated element trajectories which assume a population selection coefficient of $Ns = -4$ until time t_b, and $Ns = 0$ from time t_b until the present. The choice of $Ns = -4$ stems from a maximum likelihood–based estimate of the selection coefficient, which utilizes *BS* element frequency data from both non-African and African samples (unpublished results). We note that the *doc1420* element we study here is not in the *BS* family and was likely under considerably stronger purifying selection than $Ns = -4$ prior to the bottleneck (Petrov et al. 2003). At the present, we lack sufficient data to obtain a parallel estimate for the *Doc* family and proceed with $Ns = -4$. We generated the postbottleneck segment of each trajectory as described above, except that the frequency x from which the postbottleneck segment begins was drawn from the equilibrium transposition–selection distribution obtained by diffusion approximation with $Ns = -4$ (eq. 5.48, Ewens 2004):

$$\bar{t}(x;Ns) = \begin{cases} \left(1 - \frac{1-\exp(-s)}{1-\exp(-2Ns)}\right)\frac{\exp(2Nsx)-1}{Nsx(1-x)} & x \le 1/(2N) \\ \frac{1-\exp(-s)}{1-\exp(-2Ns)}\ \ \frac{1-\exp(-2Ns(1-x))}{Nsx(1-x)} & x \ge 1/(2N) \end{cases}. \quad (1)$$

Once the postbottleneck trajectory segment was obtained, the prebottleneck segment was generated by the truncation method we introduced here for the TA demographic scenario. For the LS demographic scenario, the prebottleneck trajectory segment was generated using the method of Slatkin (2001) with $Ns = +4$.

Recombination Suppression

Recombination suppression was incorporated into the coalescent simulations by modifying the probabilities of recombination between the TE-bearing (TE) and non-TE–bearing (non-TE) subpopulations. We assume that the sample is so small compared with the population size that 2 chromosomes from the sample never recombine with one another. There are then 4 distinct types of recombination that can occur during the simulation. A TE chromosome from the sample may recombine with a TE chromosome from the population or a non-TE chromosome from the sample may recombine with a non-TE chromosome from the population; these within-class events are the first 2 types of recombination. There are also 2 cross-class types of recombination: a TE chromosome from the sample may recombine with a non-TE chromosome from the population or a non-TE chromosome from the sample may recombine with a TE chromosome from the population. The probability of recombination per generation is then $4Lcn_jN_j$, where n_j and N_j are the sizes of the respective subsample and subpopulation, j indicates whether the TE or non-TE subsample/subpopulation is used, L is the number of nucleotide links in the sample at which recombination may occur, and c is the per-link recombination rate (cf., Przeworski 2002).

We consider 2 models of recombination suppression. In the first model, we assume that recombination does not occur in individuals heterozygous for the transposable element. Here, we simply disallow any recombination in the 2 classes that involve crossing-over between a TE and a non-TE chromosome by setting their probabilities to zero. In the second model, we assume that recombination cannot occur within a radius of fixed size surrounding the element, in individuals heterozygous for the element. To model this, we disallow recombination breakpoints at sites falling inside the restricted region surrounding the element in the calculation of L for each chromosome in the sample. If a cross-class recombination event then does occur, we ensure that the crossover location is chosen uniformly from among the links outside the restricted region.

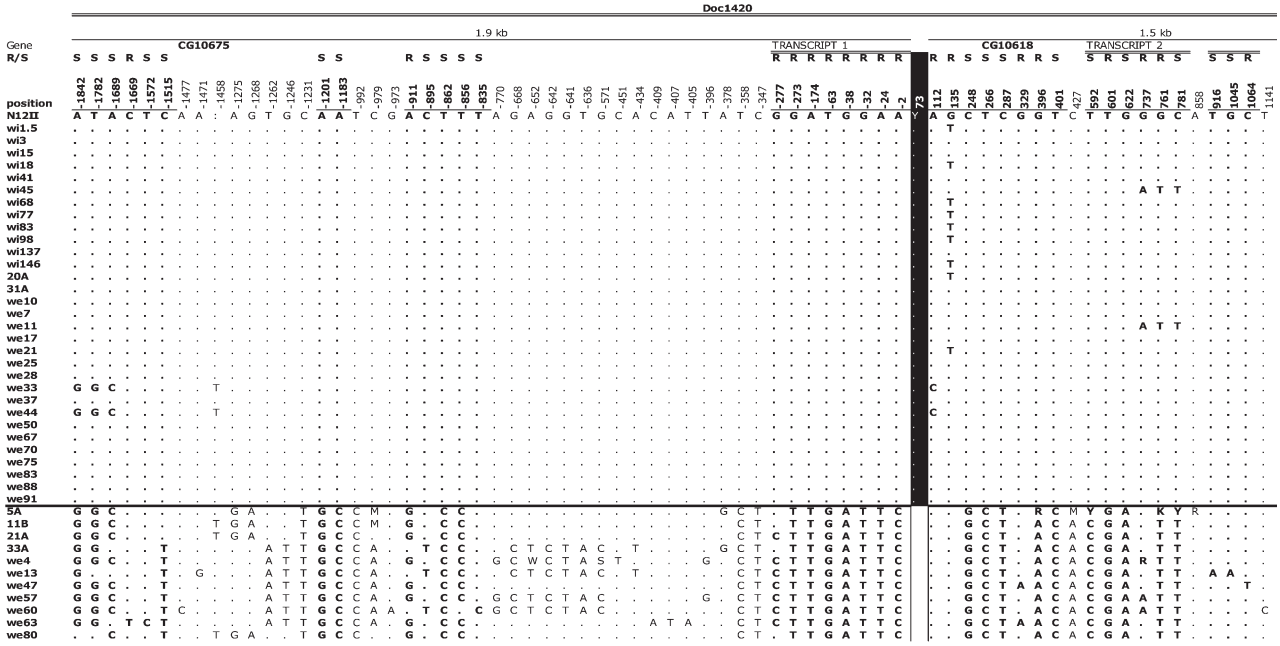

FIG. 2.—Locus *doc1420*. A column is shown at the position of the transposable element insertion. The column is shaded only for those elements that have the insertion. A horizontal rule divides the insertion-bearing sequences from the sequences that do no bear insertions.

Statistical Tests of Neutrality

For each simulated data set, we calculated a number of standard statistics, including θ_W (Watterson 1975) and π (Tajima 1983), Tajima's *D* (Tajima 1989a), the number of haplotypes *H*, and the haplotype diversity h_d (Nei 1987). Each statistic was computed over the entire sample and then separately for both the TE and non-TE subsamples.

Two additional statistics specifically designed to measure differences between the TE and non-TE subsamples were calculated. These include f_{TE}, defined as $\pi_{TE}/(\pi_{TE} + \pi_{non\text{-}TE})$, a measure we introduce to quantify the relative reduction in diversity. The second statistic is a modified version of *iHS* (Voight et al. 2006). As originally defined, *iHS* is the logarithm of $iHH_{ancestral}/iHH_{derived}$, where *iHH* is the integral of haplotype homozygosity, measured from the putatively adaptive focal site, for the subsamples linked, respectively, to the ancestral and derived states of the focal site. This integral is taken over the region delimited by the positions at which the haplotype homozygosity drops to 5% on the left side of the focal site and at which haplotype homozygosity drops to 5% on the right side of the focal site. In some of the 5 data sets, the flanking regions do not extend far enough away from the focal site (here, the transposable element) for the haplotype homozygosity to drop to 5%. We computed the respective *iHH* values to the end of the sequenced region, regardless of the value of haplotype homozygosity there. As in Voight et al. (2006), we standardized *iHS*, according to $iHS = (iHS' - E[iHS'])/\sigma[iHS']$, where iHS' is the unstandardized *iHS* value, and the expectation and standard deviation are computed from a set of simulations run with parameters identical to the pertinent simulation, but under the standard neutral null model.

To determine statistical significance, we performed 1-sided tests, defining the *P* value of each statistic from each distribution as the fraction of simulated values more extreme than the observed value. The null hypothesis for each of *D*, *H*, h_d, f_{TE}, and *iHS* is that they are less than their expectation under neutrality. Because we are performing numerous statistical tests on numerous loci, we must make a correction for multiple comparisons. We use the false discovery rate (FDR) measure, as implemented in the *qvalue* package (Storey and Tibshirani 2003), to account for this and evaluate significance at the 5% and 1% FDR thresholds.

Implementation Notes

The trajectory and genealogy simulation package, called *combesce*, was written in *java*; the trajectory implementation was verified where possible against test cases derived from Slatkin (2001), and the genealogy implementation was verified where possible against Hudson's *ms* Hudson (2002). Population genetic statistics were computed using a package called *pgTools* written by the authors in *python;* their implementation was verified by comparing the output to that of *DnaSP* (Rozas et al. 2003) on constructed and actual data sets. Postprocessing was performed with scripts written in *R* and *python*. The simulations were run on a cluster of 24 dual-processor 2GHz Xeon computers at the Stanford Genome Technology Center. All software is available online at http://petrov.stanford.edu.

Results

Summary of 5 Data Sets

The haplotype configurations of the 5 loci are shown in figures 2–6 and are statistically summarized in table 1. Strains with the insertion are separated from those without by a horizontal black line; the presence of the transposable

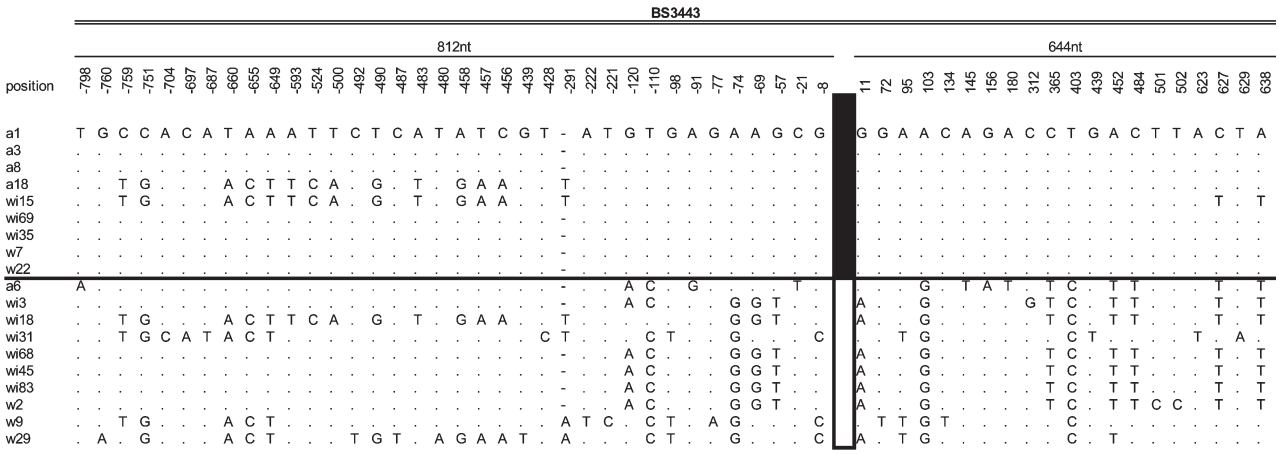

FIG. 3.—Locus *BS3443*. A column is shown at the position of the transposable element insertion. The column is shaded only for those elements that have the insertion. A horizontal rule divides the insertion-bearing sequences from the noninsertion-bearing sequences.

element is indicated by a filled box in the midst of a sequence. Note the visual similarity of the 5 haplotype configurations: each configuration bears a strong resemblance to the pattern expected from a partial selective sweep (Voight et al. 2006). Namely, there tend to be fewer haplotypes and lower nucleotide diversity in the TE subsample than in the non-TE subsample. Recombination events are evident in several of the loci, for example, in locus *BS3443* to the left side and between 222 and 291 nucleotides from the insertion in strains a18 and wi15 (fig. 3).

Standard Neutral Null Model Simulations

We wish to explore how our conclusions regarding the neutrality of 5 polymorphism patterns in figures 2–6 might change as successively more complex null hypotheses are considered. To do so, we compare the distributions of several statistical measures of polymorphism obtained by simulation under a given null hypothesis to the values of these statistics observed from the respective loci. We begin by characterizing the distributions expected under the "standard neutral" null hypothesis (Rosenberg and Nordborg 2002), which posits a panmictic population of constant size $N = 10^6$ diploid individuals. Under this model, the element drifts neutrally from initial frequency $1/(2N)$ to its contemporary frequency, and the pattern of neutral polymorphism in the region flanking the element is simulated conditional on the element's frequency trajectory (Materials and Methods). The results of these simulations are summarized in table 2.

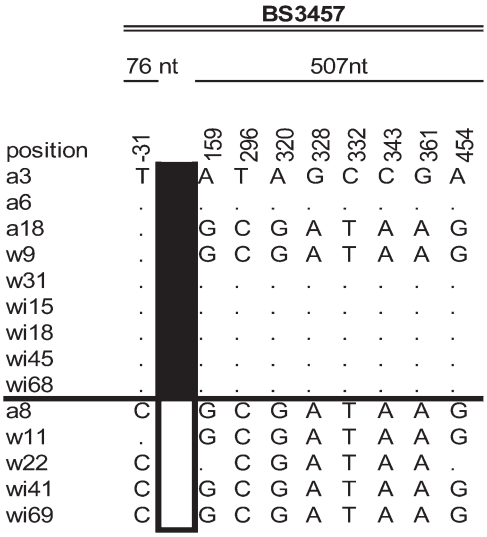

FIG. 4.—Locus *BS3457*. A column is shown at the position of the transposable element insertion. The column is shaded only for those elements that have the insertion. A horizontal rule divides the insertion-bearing sequences from the noninsertion-bearing sequences.

Among the summary statistics presented in table 2, *iHS* is likely to be the most informative in judging whether a putative partial selective sweep is consistent with neutrality. The *iHS* focuses on the expected signature of a partial selective sweep, namely slow decay of homozygosity away from the putatively selected element (Voight et al. 2006; Materials and Methods). Voight et al. (2006) showed by simulation that *iHS* was the most powerful test of neutrality among several related tests by comparison to the alternative hypothesis of a partial selective sweep, for a panmictic population of constant size. To capture the extent of reduction in diversity between the TE and non-TE subsamples, we calculate an additional between-subsample statistic, $f_{TE} = \pi_{TE}/(\pi_{TE} + \pi_{non\text{-}TE})$, the proportion of diversity among the 2 subsamples accounted for by the TE subsample. We present further summaries, including Tajima's D and the number of distinct haplotypes, H, both overall and within each subsample. These statistics may themselves be used as tests of selective neutrality (Nielsen 2005) and are provided to form a more complete picture of the expected pattern of polymorphism than would be seen by *iHS* alone.

Based on *iHS*, among the 5 data sets, only locus *BS3457* is consistent with the standard neutral null hypothesis, after a FDR correction for multiple hypothesis testing is made (Storey and Tibshirani 2003). The *iHS* is significant at the 1% FDR for *doc1420* and 2 of the *BS* elements and at the 5% FDR for a third *BS* element. That the region flanking *BS3457* fails to show significance under *iHS* may be due to its short length, 584 bp, and few segregating sites, 9. The

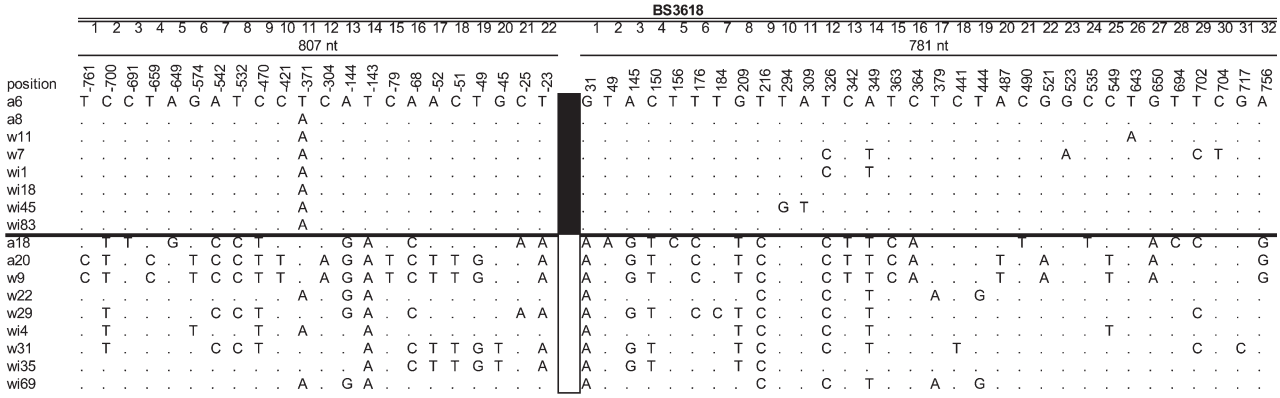

FIG. 5.—Locus *BS3618*. A column is shown at the position of the transposable element insertion. The column is shaded only for those elements that have the insertion. A horizontal rule divides the insertion-bearing sequences from the noninsertion-bearing sequences.

other *BS* loci are ≈1500 bp in length, and the region flanking the *doc1420* insertion has length 3400 bp (table 1). Most of the loci have fewer haplotypes and more low frequency mutations linked to the element than expected under the constant population size null model, as indicated by small values of H_{TE} and negative values of Tajima's D_{TE}. However, only a few significant tests occur within the non-TE subsample, which is consistent with the visual impression from the haplotype configurations of a partial selective sweep. Further, the expected value of f_{TE} is close to 1/2 for all 5 loci, but the observed f_{TE} is substantially lower than this for all loci except *BS3457*. If the *BS* elements are segregating neutrally, this suggests that the constant population size null model is either not correctly predicting a reduction in diversity in the TE subsample or underestimating the variance in f_{TE}.

Evaluated over the entire sample, Tajima's *D* is not significant according to a 1-sided test (H_0: $D > 0$) for any of the loci under the constant population size null model. One of these loci, *BS3457*, has a large and positive Tajima's *D* of 2.57, which clearly would be rejected under this null hypothesis by a 2-sided test. Although the observed *iHS* values of each of the 5 loci are negative, the Tajima's *D* values seem to follow no particular pattern, taking both positive and negative values. After a recent selective sweep, Tajima's *D* is expected to take sharply negative values, driven by excess low- and high-frequency polymorphisms. During a sweep, however, the expectation is less clear because a beneficial allele at intermediate frequency would cause intermediate frequency linked neutral polymorphisms, tending to increase Tajima's *D*. It is thus not necessarily surprising that Tajima's *D* overall is consistent with the constant population size null model.

Lastly, the overall number of haplotypes, *H*, is smaller than expected under the constant population size null model, except at of locus *BS3730*. The number of haplotypes in the non-TE subsample, $H_{non\text{-}TE}$, appears to be consistent with a null hypothesis of constant population size, but H_{TE} is found to be significant in each case for which *H* overall is significant. If we again assume that the *BS* elements segregate neutrally, this collection of results suggests that the constant population size null model either overestimates H_{TE} or underestimates its variance.

The Effects of a Bottleneck

If we were to test for the neutrality of these 5 elements, neutrality would apparently be rejected under the standard

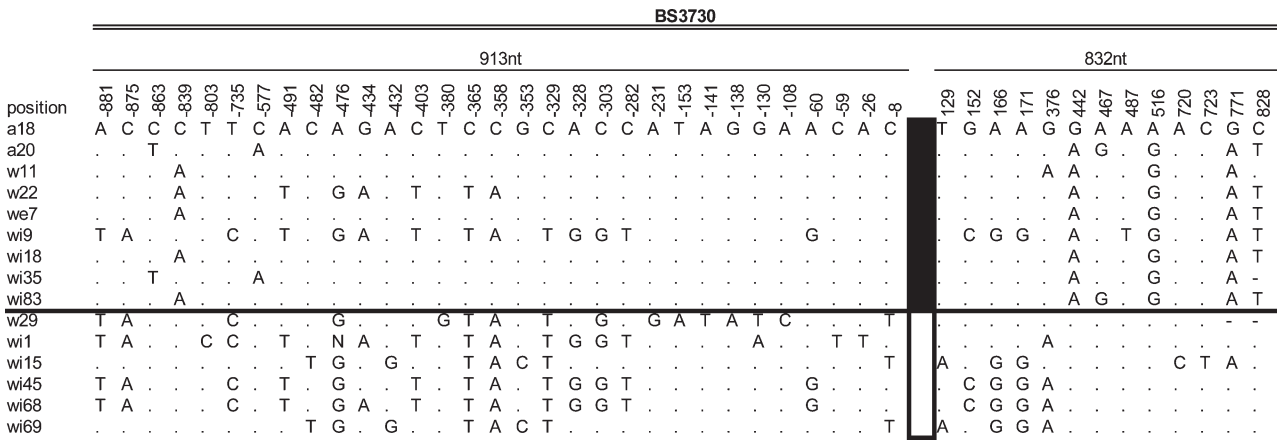

FIG. 6.—Locus *BS3730*. A column is shown at the position of the transposable element insertion. The column is shaded only for those elements that have the insertion. A horizontal rule divides the insertion-bearing sequences from the sequences that do not bear insertions.

Table 1
Summary Statistics for the 5 Loci

		L	*r*	*n*	*S*	π	θ_W	*D*	*H*	h_d
	All			43	64	4.383	4.351	0.026	15	0.769
	TE	3400	3.30	32	7	0.336	0.511	−0.996	4	0.579
doc1420	Non-TE			11	40	3.711	4.017	−0.357	11	1.000
	All			19	55	10.179	10.800	−0.235	11	0.860
	TE	1457	4.12	9	15	3.775	3.788	−0.017	3	0.417
3443	Non-TE			10	53	11.058	12.858	−0.685	8	0.933
	All			14	9	8.091	4.846	2.570	4	0.703
	TE	584	3.47	8	8	5.871	5.283	0.538	2	0.429
3457	Non-TE			6	9	6.164	6.749	−0.516	4	0.800
	All			17	54	9.644	10.052	−0.170	13	0.963
	TE	1589	3.98	8	9	1.641	2.184	−1.219	6	0.893
3618	Non-TE			9	47	11.782	10.883	0.421	7	0.944
	All			15	40	7.818	7.371	0.259	14	0.990
	TE	1669	3.01	9	24	4.194	5.291	−1.029	8	0.972
3730	Non-TE			6	30	8.468	7.872	0.480	6	1.000

NOTE.—The labels "all" denotes all sequences in the sample for a given locus; "TE" denotes the subsample bearing the retrotransposon; and "non-TE" denotes the subsample not bearing the retrotransposon. *L*, length in bp; *r*, recombination rate in cM/Mb; *n*, sample size; *S*, number of segregating sites; *D*, Tajima's *D*; *H*, number of haplotypes; and h_d, haplotype diversity. π and θ_W are given per nucleotide and multiplied by 1,000.

neutral null model. However, a number of studies have shown that polymorphism in *D. melanogaster* does not accord with this model: nucleotide diversity is consistently much lower in non-African populations than in African populations (Begun and Aquadro 1993; Glinka et al. 2003; Haddrill et al. 2005). The favored alternative is a bottleneck associated with an emigration event from the species' ancestral home in Africa 10–15 kya (David and Capy 1988; Lachaise et al. 1988). Such a demographic scenario can yield spurious results in tests of neutrality for recently completed selective sweeps because both selective sweeps and population expansion from a bottleneck produce excess low-frequency polymorphisms (Tajima 1989b). Could a recent bottleneck also produce a pattern resembling a partial selective sweep?

We modified the standard neutral null hypothesis to account for population size change in *D. melanogaster*, by in turn incorporating the maximum likelihood bottleneck parametrizations from each of 2 studies from the recent literature (Li and Stephan 2006; Thornton and Andolfatto 2006). We made one further assumption that the element transposed prior to the beginning of the bottleneck, that is, before the contemporary non-African population had emigrated from Africa. For 3 of the loci, *doc1420*, *BS3618*, and *BS3730*, the element is observed to segregate at low frequency in sub-Saharan African populations (*doc1420*: 2/8, *BS3618*: 7/39, and *BS3730*: 4/40), so this assumption appears to be justified for them. For the other 2 loci, *BS3443* and *BS3457*, we do not observe any elements in the sub-Saharan sample (*BS3443*: 0/34 and *BS3457*: 0/43). We assume nevertheless that the elements were present in the African population at the beginning of the bottleneck for consistency with the other simulations. This assumption is not unreasonable: It is quite possible that these elements are segregating in African populations that we did not sample (Schöfl et al. 2005).

When either of the 2 bottleneck scenarios are incorporated, the significance pattern of the neutrality tests changes substantially (table 2). Where all 5 of the loci, except the short 3457, would have been found significant under the standard neutral null model according to *iHS*, under the scenario of LS only *doc1420* remains strongly significant (table 2). Locus 3618 is the only *BS* element that still appears to be significant according to *iHS*, but only marginally so. The distributions of *iHS* change from the standard neutral case; the variances generally become larger, and the means depart from zero. For the high-frequency elements *doc1420* (74.4%) and *BS3730* (60.0%), the expectation of *iHS* becomes sharply negative, whereas for the intermediate frequency elements *BS3618* (47.0%), *BS3443* (47.4%), and *BS3457* (57.1%), the mean *iHS* is close to zero.

With several marginally significant exceptions, the values of Tajima's *D* and *H* for the *BS* elements overall and within both the TE and non-TE subsamples are within the null hypothesis CIs under the TA bottleneck model (table 2). Where the mean Tajima's *D* was near zero for all 5 elements under the standard neutral null model, it ranges from 1.1 to 1.5 under the bottleneck null model, and its variance increases markedly; in fact the sharply positive Tajima's *D* of 2.57 observed in *BS3457* now falls within the 95% CI. The mean number of haplotypes overall declines under the bottleneck, and the variance in the number of haplotypes increases. Although similar changes in the distribution of the number of haplotypes occur in both the TE and non-TE subsamples, the distributions of Tajima's *D* within the TE and non-TE subsamples are quite different than the distribution of Tajima's *D* overall. Like Tajima's *D* overall, the variance in the subsamples' Tajima's *D* increases substantially, but the means are close to zero, not positive.

Under the alternative bottleneck scenario of LS, the pattern of significance is qualitatively similar to that found under the TA model (table 2). Again, on the basis of *iHS*, we would reject the null hypothesis for locus *doc1420* but would not reject several of the *BS* elements. For the high-frequency elements *doc1420* and *BS3730*, the distributions of *iHS* are shifted toward negative values relative to the standard neutral null model as under the TA model, and the variance of *iHS* becomes larger than under the standard neutral null model. To a lesser extent than under the TA model, the distribution of overall Tajima's *D* shifts

Table 2
FDR-Corrected Significance of Several Neutrality Tests for Each Transposable Element, under the Standard Neutral Null, and the Demographic Models of TA and LS

	D	H	f_{TE}	iHS	D_{TE}	H_{TE}	$D_{non\text{-}TE}$	$H_{non\text{-}TE}$
Constant-sized population								
	0.03	15 ●●	0.08 ●●	−7.78 ●●	−1.00 ●●	4 ●●	−0.36	11
doc1420	0.00 (−0.65, 0.65)	40.77 (38, 43)	0.51 (0.45, 0.57)	0 (−1.94, 2.00)	0.03 (−0.62, 0.72)	30.31 (27, 32)	0.09 (−0.57, 0.76)	10.47 (9, 11)
	−0.24	11 ●●	0.25 ●●	−4.40 ●●	−0.02	3 ●●	−0.69 ●	8
3443	0.02 (−0.61, 0.68)	18.26 (16, 19)	0.49 (0.41, 0.57)	0 (−2.05, 1.98)	0.05 (−0.63, 0.73)	8.64 (7, 9)	0.05 (−0.59, 0.73)	9.63 (8, 10)
	2.57	4 ●●	0.49	−1.57	0.54	2 ●●	−0.52	4
3457	0.04 (−1.17, 1.29)	9.97 (7, 13)	0.50 (0.28, 0.73)	0 (−1.91, 2.05)	0.05 (−1.28, 1.45)	5.92 (4, 8)	0.06 (−1.29, 1.39)	4.77 (3, 6)
	−0.17	13 ●●	0.12 ●●	−5.28 ●●	−1.22 ●●	6 ●	0.42	7 ●
3618	0.02 (−0.66, 0.67)	16.42 (15, 17)	0.49 (0.41, 0.57)	0 (−2.11, 1.91)	0.05 (−0.64, 0.74)	7.71 (6, 8)	0.04 (−0.62, 0.70)	8.71 (7, 9)
	0.26	14	0.33 ●●	−2.31 ●	−1.03 ●●	8	0.48	6
3730	0.02 (−0.73, 0.75)	14.42 (13, 15)	0.50 (0.40, 0.61)	0 (−1.94, 2.11)	0.04 (−0.70, 0.78)	8.62 (7, 9)	0.04 (−0.69, 0.81)	5.80 (5, 6)
Thornton and Andolfatto (2006)								
	0.03	15 ●●	0.08 ●●	−7.78 ●●	−1.00	4 ●●	−0.36	11
doc1420	1.09 (−0.57, 2.38)	32.96 (24, 40)	0.48 (0.26, 0.72)	−3.61 (−5.74, −1.34)	0.31 (−1.76, 2.07)	23.65 (16, 30)	0.29 (−1.50, 1.73)	9.33 (6, 11)
	−0.24 ●	11	0.25	−4.40	−0.02	3 ●	−0.69	8
3443	1.57 (−0.06, 2.76)	14.44 (10, 18)	0.48 (0.15, 0.82)	0.19 (−5.58, 5.94)	−0.04 (−1.89, 1.73)	6.87 (4, 9)	0.05 (−1.90, 1.85)	7.57 (4, 10)
	2.57	4	0.49	−1.57	0.54	2	−0.52	4
3457	1.46 (−0.62, 2.63)	5.21 (3, 8)	0.48 (0, 1)	−0.20 (−2.26, 1.89)	−0.21 (−1.67, 1.89)	2.77 (1, 5)	−0.14 (−1.39, 1.91)	2.55 (1, 5)
	−0.17 ●	13	0.12 ●	−5.28 ●	−1.22	6	0.42	7
3618	1.49 (−0.21, 2.67)	13.29 (9, 17)	0.49 (0.14, 0.82)	0.30 (−5.48, 6.06)	−0.02 (−1.78, 1.75)	6.30 (3, 8)	0.04 (−1.85, 1.81)	7.00 (4, 9)
	0.26	14	0.33	−2.31	−1.03	8	0.48	6
3730	1.41 (−0.29, 2.57)	10.56 (7, 14)	0.49 (0.09, 0.93)	−1.18 (−6.33, 4.21)	−0.17 (−1.91, 1.83)	6.09 (3, 9)	−0.08 (−1.50, 1.66)	4.48 (2, 6)
Li and Stephan (2006)								
	0.03	15 ●●	0.08 ●●	−7.78 ●●	−1.00	4 ●●	−0.36	11
doc1420	0.61 (−0.98, 1.89)	38.27 (32, 42)	0.47 (0.34, 0.58)	−2.11 (−4.66, 0.56)	0.30 (−1.51, 1.72)	28.24 (23, 32)	0.55 (−0.62, 1.53)	10.05 (8, 11)
	−0.24 ●	11 ●●	0.25 ●	−4.40 ●	−0.02	3 ●●	−0.69	8
3443	1.12 (−0.03, 2.19)	17.30 (14, 19)	0.49 (0.30, 0.67)	0.10 (−3.70, 3.83)	0.29 (−1.19, 1.46)	8.22 (6, 9)	0.41 (−1.10, 1.53)	9.09 (7, 10)
	2.57	4 ●	0.49	−1.57	0.54	2	−0.52	4
3457	0.98 (−0.67, 2.41)	7.99 (5, 12)	0.46 (0.07, 0.89)	−0.26 (−2.48, 2.09)	−0.09 (−1.64, 1.73)	4.37 (2, 7)	0.11 (−1.37, 1.73)	3.88 (2, 6)
	−0.17 ●	13 ●	0.12 ●●	−5.28 ●●	−1.22 ●	6	0.42	7
3618	1.04 (−0.13, 2.10)	15.97 (14, 17)	0.49 (0.32, 0.66)	0.13 (−3.06, 3.27)	0.33 (−1.01, 1.40)	7.54 (6, 8)	0.42 (−0.95, 1.47)	8.44 (7, 9)
	0.26	14	0.33	−2.31	−1.03	8	0.48	6
3730	0.93 (−0.40, 2.07)	13.52 (11, 15)	0.47(0.24, 0.69)	−0.71 (−4.65, 2.82)	0.10 (−1.56, 1.43)	7.98 (6, 9)	0.27 (−1.04, 1.41)	5.54 (4, 6)

NOTE.—Within a statistic/locus cell, the upper left number is the observed value of the statistic for that locus. In the upper right, the notation "●●" indicates significance at the 1% FDR, and the notation "●" indicates significance at the 5% FDR but not the 1% FDR. Across the bottom of each cell from left to right are the mean, 2.5% and 97.5% CI limits, calculated as described in Materials and Methods. For the standard neutral null and TA simulations, 5,000 replicates were performed for each statistic/locus pair; for the LS simulations, which rely on the importance sampling method of Slatkin (2001), 50,000 replicates were performed for each statistic/locus pair. Standardization of *iHS* was performed using the standard neutral null simulations.

toward positive values and the number of haplotypes overall decreases relative to the standard neutral null model.

Thus, when we incorporate demographic change, only *doc1420* stands out as particularly unusual among the 5 data sets. Some of this appears to come from the increased variance in *iHS*, but for the high-frequency elements, the expected *iHS* drops sharply, in the direction expected for a selective sweep. We note that the expected f_{TE} stays close to 1/2 for both bottleneck scenarios, although there are fewer significant results, apparently due the increased variance in f_{TE}.

The Effects of Purifying Selection

There is much evidence to suggest that transposable elements are subject to purifying selection, for a number of reasons, including the deleterious effects of insertion, ectopic recombination, and misexpression (Montgomery et al. 1987; Nuzhdin 1999; Petrov et al. 2003). If transposable elements were under purifying selection in Africa, then those elements we observe in North America today would have segregated at lower frequencies at the time the bottleneck began and thus would have entered the bottleneck at lower frequencies on average than if they evolved neutrally. This relative youth might be expected to change the genealogies of neutral variants linked to the element. To explore this effect, we extended the null hypothesis for the 2 bottleneck simulations, adding the assumption that the transposable elements were subject to a selection coefficient of $Ns = -4$ prior to the beginning of the bottleneck. The particular value of *Ns* derives from a related study of the entire *BS* family (Material and Methods).

Incorporating purifying selection has a small but consistent effect on the null distributions. The distribution of Tajima's *D* shifts toward negative values relative to the simulations without purifying selection, although the means remain positive (table 3). The distributions of *iHS* consistently shift toward negative values when purifying selection is included (table 3). The magnitude of these shifts is not great, and does not alter our assessment of significance for any of the loci, but we note that the change in *iHS* is in the direction one expects under positive selection.

Table 3
Effects of Purifying Selection on Tajima's *D* and *iHS*

		T & A (2006)	L & S (2006)
Tajima's *D*			
	$Ns = 0$	1.09 (−0.57, 2.38)	0.60 (−0.96, 1.76)
doc1420	$Ns = -4$	1.05 (−0.63, 2.40)	0.54 (−1.19, 1.70)
	$Ns = 0$	1.57 (−0.06, 2.76)	1.15 (−0.14, 2.38)
3443	$Ns = -4$	1.54 (−0.17, 2.73)	1.11 (−0.06, 2.17)
	$Ns = 0$	1.46 (−0.62, 2.63)	0.96 (−0.92, 2.42)
3457	$Ns = -4$	1.42 (−0.75, 2.63)	0.66 (−1.93, 2.34)
	$Ns = 0$	1.49 (−0.21, 2.67)	0.81 (−1.00, 1.93)
3618	$Ns = -4$	1.45 (−0.23, 2.63)	0.99 (−0.72, 2.12)
	$Ns = 0$	1.41 (−0.29, 2.57)	1.14 (−0.35, 2.27)
3730	$Ns = -4$	1.37 (−0.42, 2.58)	0.85 (−0.90, 2.21)
iHS			
	$Ns = 0$	−3.61 (−5.74, −1.34)	−2.11 (−4.66, 0.56)
doc1420	$Ns = -4$	−3.66 (−5.74, −1.42)	−2.37 (−4.79, 0.34)
	$Ns = 0$	0.19 (−5.58, 5.94)	0.10 (−3.70, 3.83)
3443	$Ns = -4$	0.15 (−5.45, 5.83)	−0.07 (−3.87, 3.57)
	$Ns = 0$	−0.20 (−2.26, 1.89)	−0.26 (−2.48, 2.09)
3457	$Ns = -4$	−0.32 (−2.33, 1.77)	−0.50 (−2.59, 1.85)
	$Ns = 0$	0.30 (−5.48, 6.06)	0.13 (−3.06, 3.27)
3618	$Ns = -4$	0.20 (−5.44, 5.92)	0.05 (−3.19, 3.21)
	$Ns = 0$	−1.18 (−6.33, 4.21)	−0.71 (−4.65, 2.82)
3730	$Ns = -4$	−1.31 (−6.54, 3.94)	−0.91 (−4.85, 2.57)

NOTE.—In each cell, the upper row corresponds to a prebottleneck selection strength $Ns = 0$, the lower row to $Ns = -4$. From left to right, the values shown are the weighted mean, 2.5% and 97.5% CI limits. T & A, Thornton and Andolfatto (2006); L & S, Li and Stephan (2006).

The Effects of Recombination Suppression

Recombination is an important force in shaping the pattern of polymorphism. There is evidence to suggest that recombination is suppressed in individuals heterozygous for transposable element insertions: Clark et al. (1986, 1988) found that the presence of an 8.0-kb *B104* insertion reduced by roughly half the recombination rate observed between the insertion site and a nearby marker 3 kb away, at the *rosy* locus in *D. melanogaster*. They also found that the recombination rate was reduced to roughly a quarter its background value, over a different interval, also of 3 kb, that separated 7-kb *calypso* and 8.5-kb *B104* insertions, at the *rosy* locus. If recombination were suppressed in heterozygotes for the 5 polymorphic elements we study here, this might contribute to the unusual haplotype configurations we observe. To explore this possibility, we consider 2 models of recombination suppression and apply these to each of the demographic scenarios we have considered. In the first model, recombination is suppressed completely in heterozygotes over the full length of the flanking region and recombination proceeds at the normal rate in homozygous individuals (Materials and Methods). In the second model, recombination is suppressed completely within a 250-bp radius about the insertion site in heterozygotes and proceeds as usual outside this radius. The radius 250 bp was chosen using rough calculations based on the data from Clark et al. (1986, 1988) and from visual inspection of the locations of crossover events in figures 2–6.

Recombination suppression substantially affects the patterns of polymorphism we observe. For the case in which crossovers are suppressed completely over the full length of each of the 5 loci, the distribution of *iHS* does shift toward negative values, and its variance increases, for all demographic scenarios (fig. 7). Where the null hypothesis was rejected for some of the *BS* loci under normal recombination, the CIs now overlap the observed values of *iHS* for all *BS* loci under the standard neutral null model and the bottleneck null models. For locus *doc1420*, the *iHS* CIs do not overlap the observed value but come closer to the observed value than under normal recombination.

Complete recombination suppression amplifies the trends we have noted in Tajima's *D* and the number of haplotypes, *H*, in the presence of a bottleneck. For all loci, the distributions of Tajima's *D* are shifted toward positive values under a bottleneck, and the magnitude of this shift appears to be much greater than under normal recombination. The number of haplotypes overall is substantially smaller under complete suppression than under normal recombination. Intriguingly, under either bottleneck scenario, the number of haplotypes falls so low that even the seemingly low $H = 15$ in a sample of 43 for *doc1420* does not appear to be unusual.

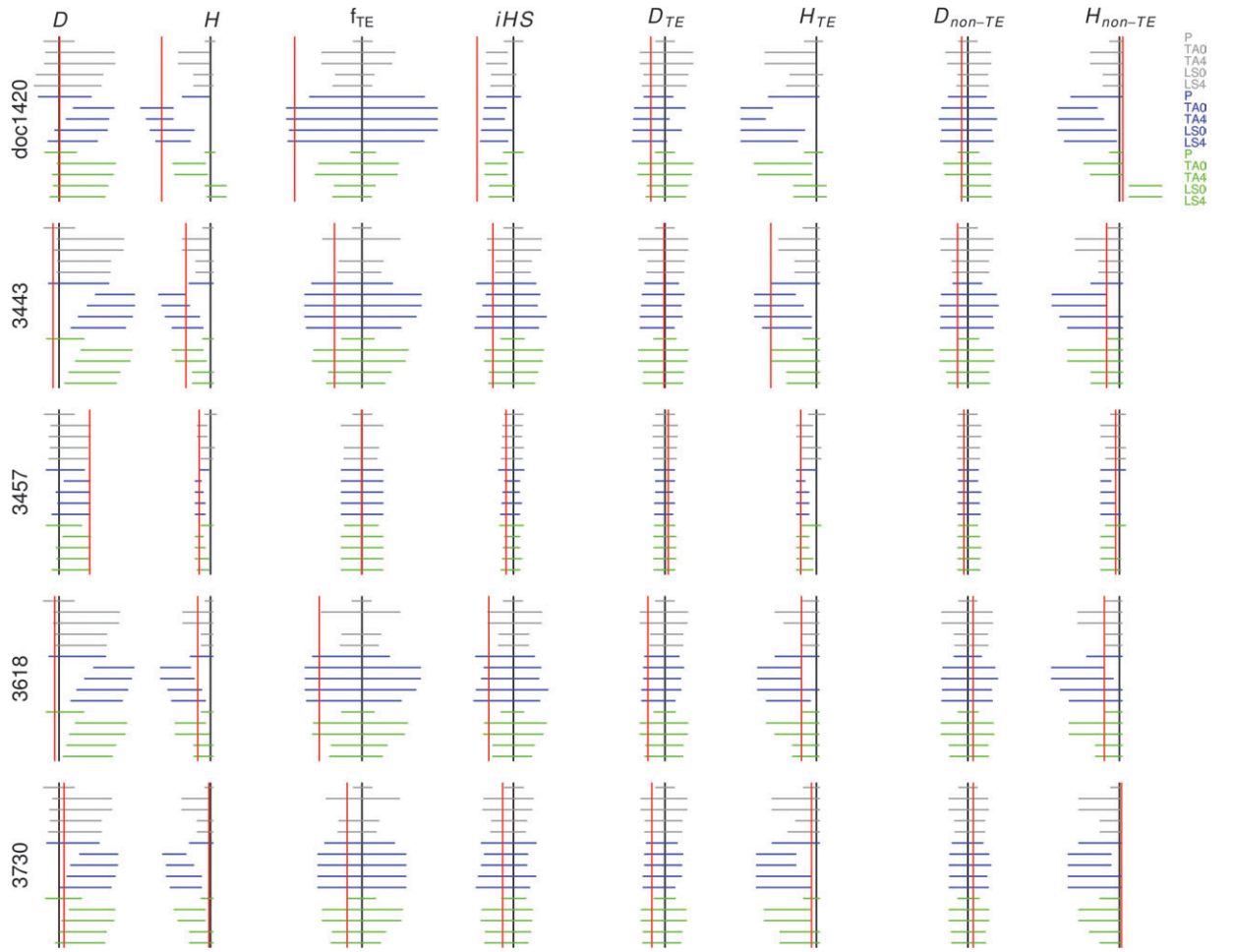

FIG. 7.—Comparison of CIs for Tajima's *D*, *H*, and *iHS* under different levels of recombination suppression. The gray, blue, and green horizontal lines represent the span between the 2.5% and 97.5% CI for the nonsuppressed, fully suppressed, and 250-bp radius of suppression models described in the text. Within each of the 3 colors, the 5 horizontal lines indicate the demographic scenario; from top to bottom, these are 1) standard neutral null, 2) Thornton and Andolfatto (2006) bottleneck scenario with prebottleneck $Ns = 0$, 3) Thornton and Andolfatto (2006) bottleneck scenario with prebottleneck $Ns = -4$, 4) Li and Stephan (2006) bottleneck scenario with prebottleneck $Ns = 0$, and 5) Li and Stephan (2006) bottleneck scenario with prebottleneck $Ns = -4$. The black vertical line drawn through each statistic/locus pair corresponds to the location of the weighted mean value of the statistic under the standard neutral null under normal recombination. The red vertical line drawn through each statistic/locus pair corresponds to the location of the observed value of the statistic for that locus.

Under the second, intermediate form of recombination suppression in which crossovers could not occur within 250 bp of the insertion, we observe distributions of the statistics which are themselves intermediate to those observed under normal recombination and complete suppression (fig. 7). If the *BS* elements serve as a neutral reference, then it appears that the recombination suppression models are more consistent with the observed values of the statistics than is the model with normal recombination. No combination of demographic scenario and recombination model is consistent with all of the *BS* loci, but complete recombination suppression appears to result in an overly positive Tajima's *D* distribution and too few haplotypes. Recombination suppression within a 250-bp radius seems to accord better than the normal recombination model with the number of haplotypes we observe for the *BS* data sets, but the normal recombination model is more consistent with the observed values of Tajima's *D*.

Both models of recombination suppression result in reduced mean f_{TE}, in addition to increased variance (table 4). This effect is seen both for the standard neutral null simulations and for the bottleneck simulations. The reduction is more pronounced when the element is assumed to have experienced purifying selection in its ancestral population. For the highest frequency element, *doc1420*, the expected f_{TE} appears to increase somewhat relative to no recombination suppression for ancestral $Ns = 0$, but if $Ns = -4$ is assumed expected f_{TE} for *doc1420* declines below the nonsuppressed mean. If recombination is suppressed only within a 250-bp radius, the values of f_{TE} are intermediate to those found for complete suppression.

The Effects of Reduction in Bottleneck Intensity

The 3 extensions to the standard neutral null model that we have explored each tend to make the 5 data sets appear less unexpected by comparison to the standard neutral null model. Only for the simulations of complete recombination suppression do we see a substantial reduction in

Table 4
Effects of Recombination Restriction and Purifying Selection on f_{TE} under Several Demographic Scenarios

		SNM	TA	LS
Complete Recombination Suppression				
	$Ns = 0$	0.52 (0.18, 0.90)	0.56 (0.03, 1)	0.53 (0.05, 1)
doc1420	$Ns = -4$		0.45 (0.02, 1)	0.37 (0.04, 0.90)
	$Ns = 0$	0.32 (0.06, 0.75)	0.40 (0, 1)	0.38 (0, 0.96)
3443	$Ns = -4$		0.33 (0, 1)	0.26 (0.01, 0.83)
	$Ns = 0$	0.40 (0, 1)	0.46 (0, 1)	0.45 (0, 1)
3457	$Ns = -4$		0.37 (0, 1)	0.32 (0, 1)
	$Ns = 0$	0.32 (0.06, 0.73)	0.40 (0, 1)	0.38 (0, 0.96)
3618	$Ns = -4$		0.32 (0, 1)	0.26 (0.01, 0.83)
	$Ns = 0$	0.41 (0.08, 0.87)	0.50 (0, 1)	0.46 (0, 1)
3730	$Ns = -4$		0.41 (0, 1)	0.33 (0, 1)
Limited Recombination Suppression				
	$Ns = 0$	0.51 (0.44, 0.59)	0.48 (0.24, 0.74)	0.47 (0.34, 0.59)
doc1420	$Ns = -4$		0.46 (0.21, 0.73)	0.45 (0.33, 0.57)
	$Ns = 0$	0.46 (0.32, 0.59)	0.47 (0.08, 0.89)	0.46 (0.20, 0.74)
3443	$Ns = -4$		0.43 (0.07, 0.86)	0.42 (0.19, 0.69)
	$Ns = 0$	0.45 (0.08, 0.88)	0.49 (0, 1)	0.45 (0, 1)
3457	$Ns = -4$		0.43 (0, 1)	0.37 (0, 1)
	$Ns = 0$	0.46 (0.32, 0.60)	0.47 (0.07, 0.89)	0.46 (0.23, 0.71)
3618	$Ns = -4$		0.44 (0.07, 0.86)	0.43 (0.21, 0.67)
	$Ns = 0$	0.48 (0.33, 0.65)	0.47 (0.05, 0.96)	0.46 (0.17, 0.77)
3730	$Ns = -4$		0.44 (0.05, 0.94)	0.41 (0.15, 0.71)

NOTE.—In each cell, the upper row corresponds to a prebottleneck selection strength $Ns = 0$, the lower row to $Ns = -4$. From left to right, the values shown are the weighted mean, 2.5% and 97.5% CI limits. SNM, standard neutral null model; T & A, Thornton and Andolfatto (2006); and L & S, Li and Stephan (2006). The first half of the table is for complete recombination suppression, and the second half of the table is for suppression within a radius of 250 bp surrounding the transposable element.

diversity in the TE subsample relative to the non-TE subsample, measured by f_{TE}, as we observe for each of the putatively neutral *BS* elements. However, a less intense bottleneck might also yield such a reduction in f_{TE}. If the bottleneck were briefer than those we have considered, and reduced the population size by the same extent, then an element entering the bottleneck at low frequency and found at high frequency at the present day would have less time to traverse between these frequencies, and the trajectory of the element would more closely resemble that expected under positive selection. Our simulations so far have used the bottleneck scenario point estimates from the studies of Thornton and Andolfatto (2006) and Li and Stephan (2006). As both studies note, there is considerable uncertainty around these point estimates, and in particular there is support in both studies for bottlenecks of lesser intensity.

To explore this possibility, we conducted a set of bottleneck simulations in which we varied the duration of the bottleneck, using sample parameters similar to those of the 5 data sets (Materials and Methods; fig. 8). We used a smaller population size $N = 10^5$ to reduce computation time and assumed that the element entered the bottleneck at frequency 5% for consistency across simulations, and to accord with the low element frequencies observed for the *BS* and *Doc* families in Africa (Petrov et al. 2003). As the bottleneck becomes shorter, the mean values of both

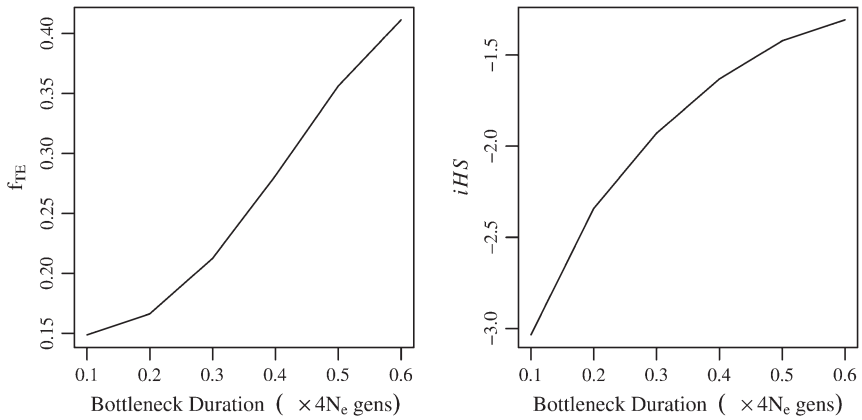

FIG. 8.—Effect of bottleneck duration on f_{TE} and *iHS*. Plotted are the mean values of f_{TE} and *iHS* from a bottleneck scenario similar to that of Thornton and Andolfatto (2006), except that the population continues at its bottlenecked size until the present. We assumed that the frequency of the transposable element at the start of the bottleneck is 5% and that during the bottleneck the population is 0.05 of its normal size. The bottleneck duration is specified in coalescent units of $4N_e$ generations and is the time before the present at which the bottleneck begins. The element is present in 40 of 50 individuals and is centered within a neutrally evolving flanking region of 2000 bp. There are 75 segregating sites, the recombination rate is 3.0 cM/Mb and the prebottleneck population size is $N_e = 10^5$ individuals. Each point is based on 10^4 replicates.

f_{TE} and *iHS* drop precipitously, and for the briefest bottlenecks we consider, namely 0.1–0.3 coalescent units of $4N$ generations, come close to the observed values of these statistics for the *BS* elements. For comparison, 0.62 and 0.39 coalescent units have elapsed by the beginning of the bottleneck scenarios of TA and LS, respectively. The 4 *BS* loci we have studied are too few to attempt to infer the most likely bottleneck scenario based on these data, but these illustrative simulations suggest that a briefer bottleneck could indeed yield the putative partial selective sweep signatures we observe.

Discussion

The initial motivation for this study was to understand why the polymorphism pattern of the putatively adaptive *doc1420* element resembled that of the 4 putatively neutral *BS* elements so closely. Did an adaptive transposition underlie the seeming partial selective sweep at each locus? Alternatively, might these patterns have arisen simply as a consequence of the recent demographic history of *D. melanogaster* or as a result of other nonadaptive departures from the standard neutral model?

Our results suggest that the high-frequency *doc1420* insertion, which appears to confer pesticide resistance (Aminetzach et al. 2005), has a haplotype structure that would still be considered nonneutral when the demographic history of *D. melanogaster* is taken into account. The 4 intermediate or high-frequency *BS* elements we studied, which have haplotype structures resembling *doc1420*, would be rejected as nonneutral on the basis of *iHS* under the assumption of constant population size, but in most cases would not be rejected under the 2 bottleneck models we considered. These conclusions do not change qualitatively if the element is assumed to have been under weak purifying selection in the ancestral population, but *iHS* shifts in the direction expected under positive selection. Intriguingly, recombination suppression in heterozygotes for the insertion tends to differentially reduce diversity in haplotypes linked to the element. Together, these extensions to the null model make it appear unlikely that the *BS* elements are positively selected because the patterns of polymorphism we observe may largely be explained under neutrality. After considering our simulation results in a genealogical context, we discuss further implications of our findings.

Genealogies of Regions Flanking High-Frequency Elements

Under both bottleneck scenarios, the expected Tajima's *D* overall increased sharply, the number of haplotypes declined and, for high-frequency elements, *iHS* decreased. The variance of each of these statistics increased substantially as well. These changes may be understood in a coalescent framework. The bottleneck scenarios imply that a comparatively great length of time, in coalescent units of $4N_e$ generations, elapses between the start of the bottleneck and the present day. Namely, $0.62 \times 4N_e$ generations and $0.39 \times 4N_e$ generations elapse during this period, respectively, under TA and LS. Here, N_e is understood to mean the size of the *D. melanogaster* population at the present time; the value of N_e inferred by TA is $\approx$1.7 times larger than the value inferred by LS. Because the sample overall is expected to have coalesced to 2 individuals by $0.5 \times 4N_e$ generations, we expect that both the TE and non-TE subsamples will often have coalesced to a single individual, retrospectively, by the start of the bottleneck under either of the demographic scenarios. This is borne out, for example, by the observation that for locus *doc1420*, the average sizes of the TE and non-TE subsamples at the start of the bottleneck are 1.21 and 1.13 under TA and 1.57 and 1.25 under LS. These estimates were obtained from 10^3 simulated genealogies, per demographic scenario, corresponding to the *doc1420* dataset. This implies that the characteristic genealogical shape is very different under a bottleneck than under constant population size; the TE and non-TE subsamples coalesce quickly, typically by the beginning of the bottleneck, then find their common ancestor on average at $2N_A + t_p$ generations earlier, where t_p is the age of the element at the time the bottleneck commences and N_A is the ancestral population size for the given demographic scenario. Coalescence cannot occur between the TE and non-TE subsamples until the element has gone from the population, and from this time the expected time to final coalescence is $2N_A$ generations. The characteristic genealogical shape under a bottleneck should then be 2 short subgenealogies, of height $\sim$16000 y, connected by long internal branches, each of much greater height $\sim 2 \times 10^6$ y, leading to the common ancestor of the sample. Thus, we expect the fraction of the total genealogical depth accounted for by the TE and non-TE subgenealogies to be much less under the bottleneck scenarios than under the standard neutral null model. This is the case again for locus *doc1420*, under the standard neutral model this fraction is $65 \pm 14\%$, but under TA, the fraction is $21 \pm 15\%$, and under LS, it is $38 \pm 18\%$.

Given this peculiar genealogical shape, we expect most mutations to fall on the long internal branches. This should produce an excess of intermediate frequency polymorphisms, leading to positive Tajima's *D* values for the sample overall, as we observe for the bottleneck simulations. Because mutations falling on the internal branches will tend to share the same segregation pattern, we should also expect the number of haplotypes overall to be reduced relative to the standard neutral null model (Wall 1999), as is consistent with our simulation results. We would further expect the variance of both statistics to increase relative to the standard neutral null model, because there will be an ensemble of genealogies, in some of which both the TE and non-TE subgenealogies coalesce before the beginning of the bottleneck, and in others several lineages persist beyond the beginning of the bottleneck.

This change in genealogical shape due to the bottleneck also shifts the distribution of *iHS*, our primary measure of positive selection, toward negative values, and apparently to a greater extent for high-frequency elements. Because negative values of *iHS* are a signature of positive selection, such a shift implies that a bottleneck may result in spurious inference of positive selection. We can again use the coalescent framework to understand why *iHS* declines under a bottleneck. Unlike after a selective sweep

(Barton 1998; Kaplan et al. 1988; Przeworski 2002), there is no great discrepancy in the depths of the TE and non-TE subgenealogies because both subsamples are forced to coalesce quickly as a result of the bottleneck. This means that the number of segregating sites should be similar in the TE and non-TE subsamples, as mean values of f_{TE} near 1/2 for all loci under all simulated scenarios attest, and so the reduction in *iHS* must not arise from a simple reduction of variation in the TE subsample.

One plausible explanation is that for high-frequency elements, where by definition the size of the TE subsample is substantially larger than the size of the non-TE subsample, the bottleneck is effectively more severe for the TE subsample than the non-TE subsample as follows. A bottleneck increases linkage disequilibrium (LD) relative to the standard neutral model; the more severe the bottleneck, whether by greater duration or greater reduction in population size, the greater the increase in LD (McVean 2002). If the TE subsample experiences a bottleneck of greater severity than the non-TE subsample, we would expect LD to increase relatively more in the TE subsample. This in turn should depress *iHH* in the TE subsample relative to the non-TE subsample, and thus, *iHS* should decrease. It stands to reason that the TE subsample would experience a more severe bottleneck than the non-TE subsample, simply because its greater number of lineages would be forced to coalesce during roughly the same length of time.

If this is true, we would expect the greatest departures in *iHS* to occur for the highest frequency elements, and we would expect the size of the departure to be attenuated by a reduction in the recombination rate. To test this explanation, we conducted a further set of bottleneck simulations based on element *doc1420*, in which we varied the number of transposable elements in the sample (fig. 9). As we expect, the mean value of *iHS* decreases as the element frequency increases. As the recombination rate is decreased, mean *iHS* still decreases as the element frequency increases, but the magnitude of *iHS* is considerably less than under higher recombination, consistent with our explanation. The reason that *iHS* does not fall to zero for high-frequency elements in the absence of recombination is likely a consequence of the small number of sites which segregate in the 2 subgenealogies; if recombination is held at zero, but the number of segregating sites is doubled, forcing more polymorphisms in the subsamples, the value of *iHS* for the high-frequency elements approaches zero (results not shown).

An element which experiences ancestral purifying selection will be younger than a counterpart that is ancestrally neutral. This implies that the long internal branches will be somewhat shorter for an ancestrally deleterious element than for an ancestrally neutral element because the time from transposition to the beginning of the bottleneck, t_p, is less. Thus, Tajima's *D* overall should be positive, but less positive than for an ancestrally neutral element, as we observe. Because an ancestrally deleterious element must traverse a greater range of frequencies than its ancestrally neutral counterpart to arrive at the same contemporary frequency, the difference in the intensity of the bottleneck between the TE and non-TE subsample should increase, and we would expect *iHS* to become more negative. This is also consistent with our simulation results.

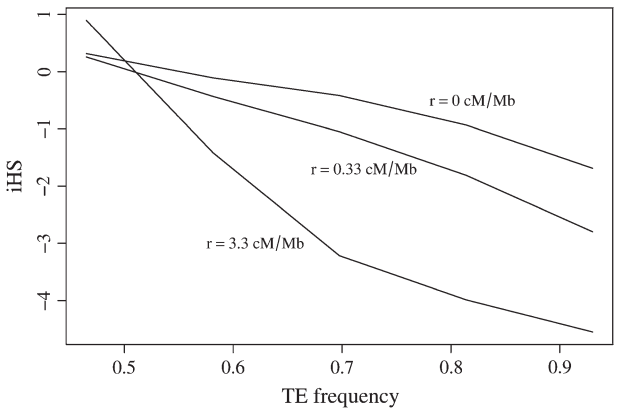

FIG. 9.—Effect of ascertainment frequency and recombination rate on *iHS* under a bottleneck scenario. Each curve consists of the mean *iHS* value for TE subsample sizes 20, 25, 30, 35, and 40, in a sample of size 43, corresponding to locus *doc1420*, under the demographic scenario of Thornton and Andolfatto (2006). The 3 curves shown differ only in the recombination rate as annotated on the plot. All other parameters are identical to those of locus *doc1420*. Each point in each curve is calculated with respect to a null hypothesis of constant population size (Materials and Methods) and the same respective ascertainment frequency and recombination rate; 4,000 replicates were generated for each point, 2,000 for the bottleneck scenario and 2,000 for the null scenario.

Recombination suppression prevents crossovers between the TE and non-TE subpopulations as a source of new haplotypes, which is consistent with the shift toward fewer haplotypes observed in our simulations. The shift we observe toward positive Tajima's *D* with recombination suppression is also expected because there are fewer opportunities to unlink a polymorphism falling on one of the long internal branches typical of our bottlenecked genealogies from the element in the TE subpopulation or from the lack of an element in the non-TE subpopulation. This results in more polymorphisms that segregate in the same proportion as the element, that is, at intermediate frequency, which should increase Tajima's *D*. That *iHS* is somewhat decreased by recombination suppression is consistent with our explanation above for the dependence of *iHS* on recombination. First, there are simply fewer recombination events, which tends to increase *iHS* (cf., fig. 9). Second, the number of between-class recombinations, which are forbidden under the recombination suppression models, depends on the product of the subsample size and the other class' subpopulation size. Because the non-TE subpopulation is much larger than the TE subpopulation while the TE is at low frequency, and under these bottleneck models we have seen that the subsample sizes are similar during this time, between-class crossovers constitute a greater proportion of the crossovers involving the TE subsample than they do for the non-TE subsample. Thus, the number of haplotypes should be reduced in both subsamples, as we observe, but reduced more in the TE subsample than in the non-TE subsample, as we also observe. This asymmetry should tend to reduce *iHS*, consistent with our simulation results.

Last, reducing the intensity of the bottleneck produced a decline in f_{TE}. In these simulations, as the bottleneck becomes shorter, the probability that the 2 subsamples each coalesce to a single individual declines. Because we assumed that the element entered the bottleneck at low

frequency, consistent with the low element frequencies we observed in Africa and the substantial evidence suggesting that transposable elements are characteristically selected against (Charlesworth and Langley 1989), then the TE subsample still tends to coalesce fully before the beginning of the bottleneck. However, as the bottleneck length declines, the non-TE subsample increasingly has not fully coalesced by the beginning of the bottleneck. The resulting depth discrepancy between the TE and non-TE subgenealogies should lead to a reduction in the number of segregating sites in the TE subsample and thus to reduced f_{TE}, consistent with our simulations. The same argument made above for the reduction in *iHS* also applies here; the TE subsample is expected to have higher LD than the non-TE subsample because it experiences a more intense bottleneck, and thus, *iHS* should decline to a greater extent as the bottleneck becomes shorter.

Implications for the Study of Adaptive Transposition

One of the most interesting aspects of transposable elements is their role in the evolution of the genome (Kidwell and Lisch 2001; Brookfield 2004) and, in particular, their role in adaptive evolution (Kazazian 2004). In several cases of possible adaptive insertions in *Drosophila*, population genetic data have been used to evaluate whether the transposition is adaptive (Maside et al. 2001; McCollum et al. 2002; Catania et al. 2004; Schlenke and Begun 2004; Aminetzach et al. 2005). The observation of a high-frequency allele with few linked haplotypes is a population genetic hallmark of positive selection, but we have shown that this is not unexpected under a null model that includes a bottleneck, on the basis of the statistic *iHS*. Furthermore, the highest frequency elements, namely those that might be thought most likely to have experienced positive selection, are those for which the effect of the bottleneck on *iHS* is strongest. The comparatively weak purifying selection we considered here, $Ns = -4$, results in a slight but consistent shift in *iHS* toward negative values. Many families of transposable elements are likely to have experienced considerably stronger purifying selection (Petrov et al. 2003), which implies that shifts in the distribution of *iHS* would be greater for elements from these families. Recombination suppression also shifts the distribution of *iHS* in the direction expected under positive selection. It is thus critical to exercise caution in constructing the null hypothesis for putatively adaptive transpositions.

Implications for Genomic Scans for Positive Selection

The data sets we have explored are also revealing outside the context of adaptive transposition. We have simulated the situation in which a possibly adaptive mutation reaches a high frequency in a recently bottlenecked population, which is the configuration expected under a partial selective sweep (Voight et al. 2006). The chief difference between the null model we consider and that of Voight et al. (2006) is that we focus on mutations which were segregating in the population when the bottleneck began. In their null simulations, Voight et al. (2006) do not condition on the trajectory of the focal site, and thus their null distributions include both mutations that arose during the bottleneck and mutations that were segregating at the beginning of the bottleneck. For their purposes, this practice is completely reasonable because they are interested in selection on both new and standing variation. In our case, we know or may reasonably expect that the elements we observe were segregating in the population at the time the bottleneck began.

The fate of standing variants under positive selection or bottleneck has received much attention in recent years (Orr and Betancourt 2001; Innan and Kim 2004; Hermisson and Pennings 2005; Przeworski et al. 2005; Teshima et al. 2006). Comparatively little is known about the origin of adaptations and, in particular, whether they tend to arise more often as new mutations or as standing variants which become advantageous when the environment changes. In the case of *D. melanogaster*, the population's migration to climates quite different from its ancestral sub-Saharan homeland, and exposure to synthetic pesticides, among other novel chemicals (David and Capy 1988; Lachaise et al. 1988), gives reason to suspect that selection pressures on existing standing variation would have changed as they emigrated. For transposable elements, apart from environmental considerations, the diminution in copy number due to the bottleneck alone is expected to reduce selection pressure from ectopic recombination. In our simulations, we considered the effect on neutral variation linked to neutral standing variants and weakly deleterious standing variants. Our simulations are closely related to those of Innan and Kim (2004) and Teshima et al. (2006), both of which modeled the effect of a recently completed selective sweep on neutral standing variants. Those studies were concerned with characterizing the signature of positive selection on standing variation and assessing whether this signature differed from the signature of positive selection on new variation. They found that the signature of standing variation is much less regular than that of directional selection on a new mutation.

Our results expand on these conclusions in 2 ways. First, we showed that a bottleneck on standing neutral variation yields haplotype configurations similar to those expected under a partial selective sweep. This finding accords with the large body of work demonstrating that bottlenecks can mimic the effects of complete selective sweeps (e.g., Andolfatto and Przeworski 2001; Przeworski 2002; Haddrill et al. 2005). Second, we showed that when the standing variant is under purifying selection, the distributions of several summary statistics commonly used to test for departure from neutrality shift in the direction expected under positive selection. If the selection pressure on a large fraction of standing variation has changed from deleterious or weakly deleterious to neutral in *Drosophila*, or in a different bottlenecked population such as maize or human (Harpending et al. 1998; Wright et al. 2005), and this is not included in a genomic scan for selection based on a bottleneck model in which all alleles are neutral (Nielsen 2005; Thornton et al. 2007), then this may result in a substantial number of false discoveries. Alternatively, in an empirical genomic scan for selection (Thornton et al. 2007), a large number of ancestrally deleterious alleles could shift the distribution of the summary statistics to more conservative values and, assuming that positive selection is rare, result in a substantial number of false negatives.

Implications for Structural Variation

Our finding that recombination suppression in heterozygotes, on its own or in combination with a bottleneck, can also cause spurious inference of positive selection, is of potentially great importance to genomic inference of positive selection. There is strong but limited experimental evidence that recombination is suppressed near transposable element insertions (Clark et al. 1986, 1988). If insertions are capable of reducing recombination, then it is likely that deletions of similar size would also suppress recombination. Although there have been no studies to date examining whether deletions suppress recombination in *Drosophila*, there is abundant evidence showing that polymorphic deletions of a wide range of sizes are common in *Drosophila* (Petrov and Hartl 2000; Blumenstiel et al. 2002). It is also well documented that recombination is suppressed in individuals heterozygous for inversions near the inversion breakpoints (Navarro et al. 2000; Andolfatto et al. 2001).

These 3 forms of structural variation, that is, insertions, deletions, and inversion, are also known to exist in large numbers as polymorphisms in the human population (Iafrate et al. 2004; Tuzun et al. 2005; Feuk et al. 2006). Thus, any scan for the signatures of positive selection that fails to take into account whether the putatively adaptive site is nearby segregating structural variation might result in spurious inference of positive selection. If there truly is recombination suppression at some of the loci we considered, then a possible cause for our observation of an apparent signal of positive selection could be that we used an external, genetic map-based estimate of the local recombination rate that overestimates the rate. This possibility may be somewhat attenuated by using recombination rate estimates obtained directly from the data. More data and further theoretical work are needed to characterize the phenomenon more fully.

Appendix

Here, we demonstrate that the truncation procedure described in the Materials and Methods correctly generates trajectories from the distribution of interest. We would like to simulate from the distribution of P_x, where P_x denotes a random frequency trajectory from the Wright-Fisher process that begins at frequency $1/(2N)$ and ends at frequency x. Suppose that a trajectory from this distribution reaches frequency x with probability p_x, the trajectory eventually returns to frequency x; with probability f_x, it goes to fixation without passing through x again; and with probability e_x, it goes to extinction without passing through x again. Then $p_x + f_x + e_x = 1$. Let $N(P_x)$ be the total number of times that the trajectory hits frequency x given that it hits frequency x at least once. $N(P_x)$ thus follows a geometric distribution:

$$\Pr\left[N(P_x) = n\right] = p_x^{n-1}\left(1 - p_x\right) \qquad \text{A1}$$

Suppose that we simulate random trajectories from the Wright-Fisher process that end at frequency x, according to some yet unspecified procedure that samples from the distribution of a variable Q_x. Further suppose that the only trajectories accepted are those that go to fixation or that go to extinction after hitting frequency x; thus, the procedure amounts to a rule for truncating the trajectory at one of the times it hits x. If Q_x is equal to P_x, which would validate the procedure, it must first be true that $\Pr[N(Q_x) = n] = \Pr[N(P_x) = n]$ for each positive integer n. For each n, it must also be true that Q_x is a random draw from the set of paths that end with frequency x. Because each trajectory is simulated independently, this latter condition must be true.

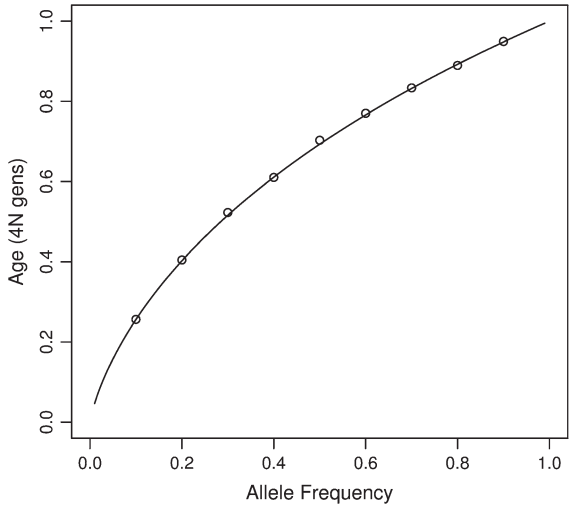

FIG. A1.—Verification of the truncation procedure. Allele trajectories terminating at several frequencies were simulated according to the truncation procedure described in Materials and Methods, using $2N = 10^5$. The mean ages are plotted as circles, in units of $4N$ generations; each point is based on 10^4 replicates. The analytic allele age (cf., Ewens 2004) is superposed as a line.

Consider the procedure that truncates the trajectory at the last time it hits x before fixation or extinction. The probability, $\Pr[N(Q_x) = k]$, that an accepted trajectory has k occasions at which it hits frequency x is the probability that the simulated trajectory has exactly k hits or $p_x^{k-1}(1 - P_x)$. Thus, $\Pr[N(Q_x) = n] = \Pr[N(P_x) = n]$ for all $n > 0$.

That this procedure generates allele trajectories of the correct age is shown in figure A1. The simulations summarized in the figure are for a neutral allele. We also confirmed that the truncation procedure gives the correct allele ages under positive and negative directional selection, in a population of constant size, by comparison to the nonneutral, constant population size entries in table 1 of Slatkin (2001) (results not shown).

Acknowledgments

We thank Marc Feldman for insightful comments on the manuscript. We thank the Stanford Genome Technology Center and particularly Lisa Diamond and Ron Davis, for the use of the computing cluster. J.M.M. is an Howard Hughes Medical Institute predoctoral fellow. J.G. is a Fulbright/Secretaria de Estado de Universidades e Investigacion, Ministerio de Educación y Ciencia postdoctoral fellow. This research was supported in part by National Institutes of Health (NIH) grant GM 28016 to Marcus W. Feldman, and by NIH and National Science Foundation grants to D.A.P.

Literature Cited

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. Science. 309:764–767.

Andolfatto P, Depaulis F, Navarro A. 2001. Inversion polymorphisms and nucleotide variability in *Drosophila*. Genet Res. 77:1–8.

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

Barton NH. 1998. The effect of hitch-hiking on neutral genealogies. Genet Res. 72:123–133.

Begun DJ, Aquadro CF. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature. 365:548–550.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 441:87–90.

Blumenstiel JP, Hartl DL, Lozovsky ER. 2002. Patterns of insertion and deletion in contrasting chromatin domains. Mol Biol Evol. 19:2211–2225.

Brookfield JF. 2004. Evolutionary genetics: mobile DNAs as sources of adaptive change? Curr Biol. 14:344–345.

Brookfield JF. 2005. Evolutionary forces generating sequence homogeneity and heterogeneity within retrotransposon families. Cytogenet Genome Res. 110:383–391.

Brosius J. 2003. The contribution of RNAs and retroposition to evolutionary novelties. Genetica. 118:99–116.

Catania F, Kauer MO, Daborn PJ, Yen JL, Ffrench-Constant RH, Schlotterer C. 2004. World-wide survey of an Accord insertion and its association with DDT resistance in *Drosophila melanogaster*. Mol Ecol. 13:2491–2504.

Charlesworth B, Langley CH. 1989. The population genetics of *Drosophila* transposable elements. Ann Rev Genet. 23:251–287.

Chung H, Bogwitz MR, McCart C, Andrianopoulos A, Ffrench-Constant RH, Batterham P, Daborn PJ. 2007. Cis-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene Cyp6g1. Genetics. 175:1071–1077.

Clark SH, Hilliker AJ, Chovnick A. 1988. Recombination can initiate and terminate at a large number of sites within the *rosy* locus of *Drosophila melanogaster*. Genetics. 118:261–266.

Clark SH, McCarron M, Love C, Chovnick A. 1986. On the identification of the *rosy* locus DNA in *Drosophila melanogaster*: intragenic recombination mapping of mutations associated with insertions and deletions. Genetics. 112:755–767.

Daborn PJ, Yen JL, Bogwitz MR, et al. (13 co-authors). 2002. A single p450 allele associated with insecticide resistance in *Drosophila*. Science. 297:2253–2256.

David JR, Capy P. 1988. Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. 4:106–111.

Davidson EH, Britten RJ. 1973. Organization, transcription, and regulation in the animal genome. Q Rev Biol. 48:565–613.

Ewens WJ. 2004. Mathematical population genetics, 2nd edition. Berlin (Germany): Springer.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155:1405–1413.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet. 7:85–97.

Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics. 165:1269–1278.

Grumbling G, Strelets V. 2006. FlyBase: anatomical data, images and queries. Nucleic Acids Res. 34:D484–D488.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. Genome Biol. 6:R67.

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST. 1998. Genetic traces of ancient demography. Proc Natl Acad Sci USA. 95:1961–1967.

Hein J, Schierup MH, Wiuf C. 2005. Gene genealogies, variation and evolution: a primer in coalescent theory. New York: Oxford University Press.

Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 169:2335–2352.

Hudson RR. 1993. The how and why of generating gene genealogies. In: N. Takahata and A.G. Clark, editors. Mechanisms of molecular evolution: Introduction to molecular paleopopulation biology. Sunderland (MA): Sinauer. p. 23–36.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18: 337–338.

Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. Genetics. 120:831–840.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet. 36:949–951.

Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci USA. 101:10667–10672.

Kaplan NL, Darden T, Hudson RR. 1988. The coalescent process in models with selection. Genetics. 120:819–829.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. Genetics. 123:887–899.

Kazazian HH. 2004. Mobile elements: drivers of genome evolution. Science. 303:1626–1632.

Kidwell MG, Lisch DR. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution. 55:1–24.

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics. 167:1513–1524.

Kreitman M. 2000. Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet. 1:539–559.

Lachaise D, Cariou ML, David JR, Lemeunier F, Tsacas L, Ashburner M. 1988. Historical biogeography of the *Drosophila-melanogaster* species subgroup. Evol Biol. 22:159–225.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2:1580–1589.

Maside X, Bartolome C, Assimacopoulos S, Charlesworth B. 2001. Rates of movement and distribution of transposable elements in *Drosophila melanogaster*: in situ hybridization vs Southern blotting data. Genet Res. 78:121–136.

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

McCollum AM, Ganko EW, Barrass PA, Rodriguez JM, McDonald JF. 2002. Evidence for the adaptive significance of an LTR retrotransposon sequence in a *Drosophila* heterochromatic gene. BMC Evol Biol. 2:5.

McVean GA. 2002. A genealogical interpretation of linkage disequilibrium. Genetics. 162:987–991.

Montgomery E, Charlesworth B, Langley CH. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila* melanogaster. Genet Res. 49:31–41.

Navarro A, Barbadilla A, Ruiz A. 2000. Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in *Drosophila*. Genetics. 155: 685–698.
Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.
Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39:197–218.
Nuzhdin SV. 1999. Sure facts, speculations, and open questions about the evolution of transposable element copy number. Genetica. 107:129–137.
Orr HA, Betancourt AJ. 2001. Haldane's sieve and adaptation from the standing genetic variation. Genetics. 157:875–884.
Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. Mol Biol Evol. 20:880–892.
Petrov DA, Hartl DL. 2000. Pseudogene evolution and natural selection for a compact genome. J Hered. 91:221–227.
Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics. 160:1179–1189.
Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. Evolution Int J Org Evolution. 59:2312–2323.
Rosenberg NA, Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet. 3:380–390.
Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19:2496–2497.
Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature. 419:832–837.
Schlenke TA, Begun DJ. 2004. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proc Natl Acad Sci USA. 101:1626–1631.
Schöfl G, Catania F, Nolte V, Schlötterer C. 2005. African sequence variation accounts for most of the sequence polymorphism in non-African *Drosophila melanogaster*. Genetics. 170:1701–1709.
Slatkin M. 2001. Simulating genealogies of selected alleles in a population of variable size. Genet Res. 78:49–57.
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 100:9440–9445.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437–460.
Tajima F. 1989a. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.
Tajima F. 1989b. The effect of change in population size on DNA polymorphism. Genetics. 123:597–601.
Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16:702–712.
Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics. 172:1607–1619.
Thornton KR, Jensen JD, Becquet C, Andolfatto P. 2007. Progress and prospects in mapping recent selection in the genome. Heredity. 98:380–348.
Tuzun E, Sharp AJ, Bailey JA, et al. (12 co-authors). 2005. Fine-scale structural variation of the human genome. Nat Genet. 37:727–732.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:446–458.
Wall JD. 1999. Recombination and the power of statistical tests of neutrality. Genet Res. 74:65–79.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7:256–276.
Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. 2005. The effects of artificial selection on the maize genome. Science. 308:1310–1314.

Marcy Uyenoyama, Associate Editor

Accepted January 01, 2008