Evidence of Gene Conversion Associated with a Selective Sweep in *Drosophila melanogaster*

Sascha Glinka, David De Lorenzo, and Wolfgang Stephan

Section of Evolutionary Biology, Department of Biology II, Ludwig-Maximilians University, Planegg-Martinsried, Germany

Since *Drosophila melanogaster* colonized Europe from tropical Africa 10 to 15 thousand years ago, it is expected that adaptation has played a major role in this species in recent times. A previously conducted multilocus scan of noncoding DNA sequences on the X chromosome in an ancestral and a derived population of *D. melanogaster* revealed that some loci have been affected by directional selection in the European population. We investigated if the pattern of DNA sequence polymorphism in a region surrounding one of these loci can be explained by a hitchhiking event. We found strong evidence that the studied region around the gene *unc-119* was shaped by a recent selective sweep, including a valley of reduced heterozygosity of 83.4 kb, a skew in the frequency spectrum, and significant linkage disequilibrium on one side of the valley. This region, however, was interrupted by gene conversion events leading to a strong haplotype structure in the center of the valley of reduced variation.

Distinguishing between demography (e.g., bottlenecks) and selection has received much recent attention in population genetics (e.g., Glinka et al. 2003; Orengo and Aguadé 2004; Storz et al. 2004; Ometto et al. 2005; Stajich and Hahn 2005) because both forces can lead to a reduction in diversity (Galtier et al. 2000). Demographic events will affect the whole genome, whereas selective events (e.g., directional selection) will affect only specific loci (Andolfatto 2001).

Genetic hitchhiking of neutral loci linked to rapidly fixed beneficial mutations (Maynard Smith and Haigh 1974) is expected to reduce heterozygosity locally, and the size of the affected region depends on the selection coefficient and the recombination rate (Kaplan et al. 1989; Stephan et al. 1992). The reduction is greatest at the site of the beneficial mutation but decreases with increasing distance from the selected site due to recombination. This results in a valley of reduced nucleotide diversity (Kim and Stephan 2002). In the absence of recombination, variation at linked neutral sites is completely removed but recovers slowly due to newly arising mutations. This leads to an excess of low-frequency variants and a star-shaped genealogy (Braverman et al. 1995). In the presence of recombination, hitchhiking may be incomplete such that the frequencies of neutral loci depend on whether they belong to the same lineage as the beneficial mutation or not. As a result, neutral variation may form a bipartite frequency spectrum. With the knowledge of the ancestral and derived states (using an outgroup), one can distinguish between low- and high-frequency variants (Fay and Wu 2000; Przeworski 2002). The resulting genealogy of surrounding neutral loci is also star-shaped but with long branches between the recombined and the swept lineages (Fay and Wu 2000; Przeworski 2002; Meiklejohn et al. 2004). This topology creates a strong association among alleles due to the long branches in the genealogy. Therefore, the resulting haplotype structure leads to linkage disequilibrium (LD) between polymorphisms at neutral loci (on the side of the selected site), which decreases with increasing distance from the target of selection (Przeworski 2002; Kim and Nielsen 2004; Stephan et al. 2006). These features are unique to genetic hitchhiking (Kim and Stephan 2002) and can therefore be used to distinguish it from background selection, the selection against recurrent deleterious mutations (Charlesworth et al. 1993).

A combination of these features has recently been observed in various studies of *Drosophila*. Evidence for directional selection has been reported for *Drosophila simulans* (Parsch et al. 2001; Quesada et al. 2003; Schlenke and Begun 2004) and *Drosophila melanogaster* (Depaulis et al. 1999; Nurminsky et al. 2001; Mousset et al. 2003; Bauer DuMont and Aquadro 2005; Beisswanger et al. 2006). Both species have extended their range from tropical Africa (south of the Sahara) to the Eurasian continent after the last glaciation 10 to 15 thousand years ago (kya) (David and Capy 1988). Due to these colonization events, the genetic composition of these species is likely to be affected by both demographic and selective processes.

A recent multilocus scan of noncoding DNA sequences on the X chromosome of a putatively ancestral population from Africa (Lake Kariba, Zimbabwe) and a derived population from Europe (Leiden, The Netherlands) of *D. melanogaster* revealed a large number of loci with no variation in the derived population (Glinka et al. 2003; Ometto et al. 2005). Although demography may explain most of the chromosome-wide lack of variation, there is evidence that the reduced polymorphism of some loci cannot be explained by bottlenecks alone (Ometto et al. 2005). One locus with zero polymorphism, a fragment within an intron of gene *unc-119* (fragment *125*; Glinka et al. 2003), is located in a region of intermediate recombination rate, with an estimated 1.926×10^{-8} recombination events per base-pair per generation (rec/bp/gen; Comeron et al. 1999). This locus is about 7 Mb away from the telomere on the X chromosome (see also fig. 1; Glinka et al. 2003). Because a local reduction of variation on a recombining chromosome may be observed by chance (Kim and Stephan 2002), we further investigated if the region surrounding fragment *125* shows a similar pattern, which would support the idea of directional selection. Thus, we screened 17 loci around fragment *125*, delimiting the region of reduced variation in the European population of *D. melanogaster*. In addition, we analyzed the same fragments in a putatively ancestral population of *D. melanogaster* (see above) because evidence is mounting that selective sweeps may have originated in the

Key words: *Drosophila melanogaster*, nucleotide diversity, gene conversion, selective sweep.

E-mail: glinka@zi.biologie.uni-muenchen.de.

Mol. Biol. Evol. 23(10):1869–1878. 2006
doi:10.1093/molbev/msl069
Advance Access publication July 25, 2006

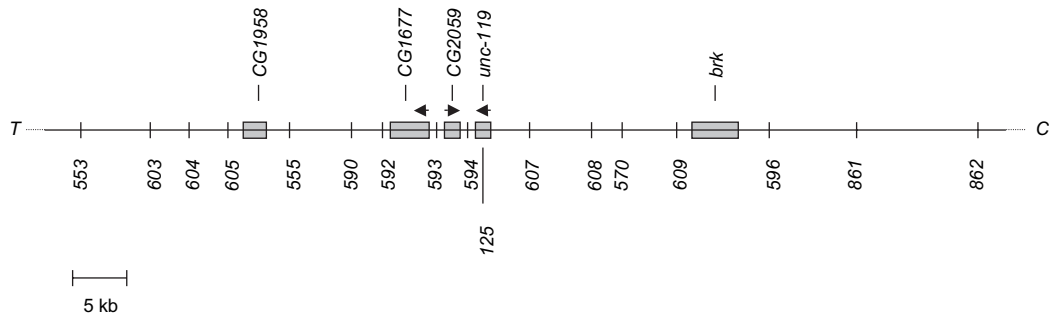

FIG. 1.—Map of the studied region, 7A2–7A5, around fragment *125* on the X chromosome, oriented from the telomere, *T*, to the centromere, *C*. The arrow indicates the direction of transcription of each gene.

ancestral range of *D. melanogaster* (e.g., Beisswanger et al. 2006).

Materials and Methods

Population Samples, Polymerase Chain Reaction Amplification, and DNA Sequencing

For the following analyses, we used 12 highly inbred lines of both a European (Leiden, The Netherlands; kindly provided by A. J. Davis) and an African (Lake Kariba, Zimbabwe; Begun and Aquadro 1993; kindly provided by C. F. Aquadro) *D. melanogaster* population and a single strain of *D. simulans* (Davis, CA; kindly provided by H. A. Orr), as described in Glinka et al. (2003). Following their procedure, we polymerase chain reaction amplified and sequenced (on both strands) 16 more noncoding loci proximal and distal to fragment *125* (European Molecular Biology Laboratory [EMBL] database, http://www.ebi.ac.uk, accession numbers AJ571381–405; Glinka et al. 2003) on the X chromosome (see fig. 1). This was done based on the available DNA sequence of the *D. melanogaster* genome (Flybase 2004, Release 3.2.0, http://www.flybase.org). In addition, we sequenced the coding regions of 3 genes (*CG1677*, *CG2059*, and *unc-119*) and their 5′ flanking regions (fig. 1). The 5′ region of *unc-119* begins 5.7 kb away from the start codon and contains a binding site for the transcription factor Dorsal (Markstein et al. 2002). We aligned only high-quality sequences with the application Seqman of the DNAstar (Madison, WI) package, as described in Glinka et al. (2003). All sequences were deposited in the EMBL database with the accession numbers AM284420–AM284965. The alignments used for the following analyses are available at http://www.zi.biologie.uni-muenchen.de/evol/Downloads.

Sequence Analyses

Standard population genetic analyses were performed using a program kindly provided by H. Li. To investigate the extent of gene conversion in our data set, we used the gene-conversion presence (GCP) test (Song et al. 2006), which is implemented in the program SHRUB-GC and was kindly provided by Y. Song (http://www.csif.cs.ucdavis.edu/~gusfield/). The program of H. Li was also used to conduct coalescent simulations for determining the probabilities of the statistical significance of Tajima's *D* (Tajima 1989), Fay and Wu's *H* (Fay and Wu 2000), and Fu and Li's *D* (Fu and Li 1993), making the assumption of no recombination. The homologous sequences of *D. simulans* were used to determine the derived state of a given site for Fay and Wu's *H* and Fu and Li's *D* and to perform the multilocus Hudson-Kreitman-Aguadé test (Hudson et al. 1987; Kliman et al. 2000). The latter approach is implemented in the program HKA, which was kindly provided by J. Hey (http://www.lifesci.rutgers.edu/~heylab). We assumed no intragenic but free recombination between fragments for the HKA test, and we have not corrected for multiple tests. For those sites where we could not identify a base in the *D. simulans* sequence, we used the corresponding position of the published *Drosophila yakuba* genome (http://flybase.net/blast/). We estimated interspecific divergence and the LD measure Z_{ns} (Kelly 1997) for each locus using the program VariScan (Vilella et al. 2005). The probability associated with LD measure Z_{ns} was calculated using DnaSP 3.99 (Rozas et al. 2003).

Estimation of the Selective Sweep Parameters

To examine the significance of the observed local reduction of genetic variation, we applied a composite likelihood ratio (CLR) test (Kim and Stephan 2002). This test requires independent estimates of the mutational parameter θ and the scaled recombination rate R_n. Because it is difficult to estimate θ by $3N_e\mu$, where N_e is the effective population size and μ the mutation rate, we used the mean (standard error [SE]) of the Watterson estimator (Watterson 1975), θ_W, of 0.0044 and 0.0127 estimated from 105 loci of the European and the African population, respectively (Glinka et al. 2003). Because the value of the European population size is about one-third of that of the African one (which we assumed as 10^6; Przeworski et al. 2001), $N_e = 300{,}000$ is used for the European population. Due to the absence of recombination in male *D. melanogaster*, R_n was estimated by $2N_er$ (Przeworski et al. 2001), where N_e is 300,000 or 10^6 (see above) and r is the per-site recombination rate of 1.926×10^{-8} rec/bp/gen (Comeron et al. 1999). The probability of the initiation per nucleotide, G_n, of a gene conversion event is estimated by $2R_n$ (Andolfatto and Wall 2003). For this test, we used a mean tract length of 352 bp (see Hilliker et al. 1994). The input files used for the CLR test are available at http://www.zi.biologie.uni-muenchen.de/evol/Downloads.

Table 1
Summary of Sequence Data of Each Locus in the Studied Region of the European Sample

Fragment	Position (kb)	*n*	*L*	*S*	π	θ_W	*K*	Z_{ns}	T's *D*	F & W's *H*	F & Li's *D*
553	1	12	375	7	0.0072	0.0062	0.0296	0.2135	0.5807	−0.2474	0.7763
603	6935	12	307	0	0.0000	0.0000	0.0739	n.a.	n.a.	n.a.	n.a.
604	10668	12	305	1	0.0010	0.0011	0.0668	n.a.	−0.1726	0.5033	0.6641
605	14103	12	355	0	0.0000	0.0000	0.0057	n.a.	n.a.	n.a.	n.a.
555	19745	12	550	1	0.0003	0.0006	0.0597	n.a.	−1.0100*	0.3145	−1.3413*
590	25447	12	314	1	0.0005	0.0011	0.0923	n.a.	−1.0100*	0.3145	−1.3413*
592	28443	12	446	0	0.0000	0.0000	0.0634	n.a.	n.a.	n.a.	n.a.
593	33721	12	423	6	0.0053	0.0047	0.1085	0.7600*	0.4444	−0.6574	1.2386
594	36506	12	292	1	0.0006	0.0011	0.0508	n.a.	−1.0100*	0.3145	−1.3413*
125	36938	12	241	0	0.0000	0.0000	0.0711	n.a.	n.a.	n.a.	n.a.
607	41821	12	372	1	0.0004	0.0009	0.0830	n.a.	−1.0100*	0.3145	−1.3413*
608	47395	12	382	0	0.0000	0.0000	0.0306	n.a.	n.a.	n.a.	n.a.
570	50577	12	474	1	0.0011	0.0007	0.0158	n.a.	1.2230	0.3145	0.6641
609	55169	12	300	1	0.0010	0.0011	0.0212	n.a.	−0.1726	0.5033	0.6641
596	63553	12	346	4	0.0043	0.0038	0.0499	0.1818	0.4197	−0.0796	0.2824
861	70771	12	399	7	0.0083	0.0058	0.0369	0.4603	1.5325	−0.2651	1.3653
862	82980	12	379	12	0.0145	0.0105	0.0355	0.7878*	1.4731	−0.5670	1.4550

NOTE.—Position is relative to the first site of the first fragment. *S* is the number of segregating sites in the European *Drosophila melanogaster* sample with its size *n*. *L* represents the number of sites sequenced. *K* is divergence to *Drosophila simulans*, and levels of nucleotide diversity were estimated using π (Tajima 1983) and θ_W (Watterson 1975). Z_{ns} (Kelly 1997) is LD; T's *D*, Tajima's *D* (Tajima 1989); F & W's *H*, Fay and Wu's *H* (Fay and Wu 2000); and F & Li's *D*, Fu and Li's *D* (Fu and Li 1993); and n.a., not available.

Asterisk indicates significance at 0.05 level.

The original approach by Kim and Stephan (2002) incorporates only the spatial distribution of polymorphic sites and the frequency spectrum. We therefore applied the extended version of the maximum likelihood method that uses information of LD as well (Kim and Nielsen 2004). Both methods allow us to evaluate the maximum likelihood estimates for the position of the selected site *X* and the population selection parameter α. We used 1-kb intervals between initial steps for *X* over the entire range of the studied region and calculated the selection coefficient *s* by $\alpha/1.5N_e$ (e.g., Kaplan et al. 1989; Braverman et al. 1995). In the case where the neutral model is rejected in favor of the hitchhiking model by the CLR test (see above), we evaluated the significance of the selective hypothesis by a goodness-of-fit (GOF) test (Jensen et al. 2005). For this test, we compared the GOF values of our data with a distribution generated from 1,000 simulated data sets under a selective scenario. These simulated genealogies were also used to estimate the confidence intervals (CIs) of *X* and *s* (J Jensen, personal communication).

Demographic Modeling of the European Population

To examine if the observed pattern of nucleotide diversity in the European population could also be explained by a bottleneck, we used an extended version (Beisswanger et al. 2006) of a maximum likelihood approach (Ometto et al. 2005) implemented in a coalescent-based program (Ramos-Onsins et al. 2004). Following the model proposed by Galtier et al. (2000), a bottleneck is characterized by its time T_b and strength S_b and the population mutation rate θ. As input parameters, we used different combinations of T_b (i.e., 0.0100–0.0500) and S_b (i.e., 0.340 and 0.400) and the average African θ_W observed in the region of reduced heterozygosity (see also Beisswanger et al. 2006). Then, the probability of our data (i.e., the valley of reduced variation) under the bottleneck scenario was calculated as the fraction of those simulated genealogies (i.e., 100,000) with at most the observed segregating sites in the entire region (provided that fragment *125* was monomorphic). These simulations were conducted with ($r = 1.926 \times 10^{-8}$ rec/bp/gen; see above) and without recombination between fragments.

Results

Region of Reduced Level of Nucleotide Diversity

To investigate levels of nucleotide diversity surrounding fragment *125*, we surveyed a total of 17 loci with an average distance between loci of 4.5 kb in both the European and the African *D. melanogaster* samples (fig. 1, tables 1 and 2). The mean size (SE) of the DNA fragments analyzed (excluding insertions and deletions) varied slightly between the 2 population samples (368 [19] bp to 363 [17] bp for the European and African sample, respectively; see tables 1 and 2), and the entire region in which these 17 fragments are located spans 83.4 kb (tables 1 and 2).

The observed level of nucleotide diversity varies along the studied region within the European and the African sample (tables 1 and 2) and between both population samples (fig. 2). Whereas the nucleotide diversity levels of the flanking loci (fragment *553*, *596*, *861*, and *862*) and a central locus (fragment *593*) of the European sample are similar (see table 1) as the reported mean (SE) values of 0.0046 (0.0005) and 0.0044 (0.0004; for π and θ_W, respectively; see Glinka et al. 2003), the remaining 12 loci show either very low or zero polymorphism (table 1). In contrast, levels of nucleotide diversity are much higher in the African sample (see table 2) and similar to those reported by Glinka et al. (2003) (0.0112 [0.0007] and 0.0127 [0.0007] for π and θ_W, respectively), with the exception of 4 low-variation loci (fragment *605* and *570*, *609* and *596*; fig. 2). As a result, this leaves a small valley of low variation at the centromere-proximal end of the studied region in the African sample, which differs from that of the European sample (fig. 2).

Table 2
Summary of Sequence Data of Each Locus in the Studied Region of the African Sample

Fragment	Position (kb)	*n*	*L*	*S*	θ_W	Z_{ns}	T's *D*	F & W's *H*	F & Li's *D*
553	1	12	373	11	0.0098	0.1866	−0.5116	0.0495	−0.5673
603	6935	12	286	16	0.0185	0.0887	−0.3806	−0.5065	−0.2543
604	10668	12	304	16	0.0174	0.1777	0.4895	−0.0994	0.7092
605	14103	12	348	7	0.0067	0.0083	−0.7658	0.4418	−0.4017
555	19745	12	533	19	0.0118	0.1294	−0.7461	−1.2699	−0.1892
590	25447	12	333	21	0.0209	0.0812	−0.7129	−0.3501	−0.6079
592	28443	12	419	11	0.0087	0.0828	−0.0380	0.2907	−0.4977
593	33721	11	412	24	0.0199	0.1314	−0.5451	−0.4721	−0.4352
594	36506	12	301	10	0.0111	0.0869	−0.0952	0.6251	−0.9861
125	36938	12	240	7	0.0097	0.1149	−0.1854	0.1237	0.7763
607	41821	12	362	26	0.0238	0.1968	−0.7593	−1.4226	0.1592
608	47395	12	378	10	0.0088	0.1347	−0.6098	−0.2969	−0.1493
570	50577	12	474	10	0.0070	0.0649	−0.9014	0.2940	−1.6607
609	55169	12	312	4	0.0042	0.0132	−1.2476	0.6103	−1.4578
596	63553	12	375	9	0.0079	0.0791	−1.6660*	0.6715	−2.1687*
861	70771	12	398	15	0.0125	0.2722	0.1818	−0.9075	−0.0107
862	82980	10	325	18	0.0196	0.1541	−0.7003	0.6048	−0.9916

NOTE.—Position is relative to the first site of the first fragment. *S* is the number of segregating sites in the African *Drosophila melanogaster* sample with its size *n*. *L* represents the number of sites sequenced, and levels of nucleotide diversity were estimated using θ_W (Watterson 1975). Z_{ns} (Kelly 1997) is LD; T's *D*, Tajima's *D* (Tajima 1989); F & W's *H*, Fay and Wu's *H* (Fay and Wu 2000); and F & Li's *D*, Fu and Li's *D* (Fu and Li 1993).

Asterisk indicates significance at 0.05 level.

Taking the interspecific divergence between *D. melanogaster* and *D. simulans* into consideration, the observed low variation for fragment *605* and *570* in both population samples could be explained by a low mutation rate or selective constraints (fig. 2). The opposite might be true for fragment *593* (fig. 2). A higher mutation rate could have led to the observed high peak in nucleotide diversity in the European sample in this fragment. However, the observed level of polymorphism in the European sample results from a distinct haplotype structure, that is, 6 sites are segregating in 3 haplotypes present at different frequencies (fig. 3*A*). In addition, all sites segregating in the European sample are also segregating in the African sample (fig. 3*A* and *B*). Because the genealogy of fragment *593* (i.e., star-shaped, but with long inner branches) differs from those observed in the surrounding fragments (fragments *592* and *594*), we investigated if the observed haplotype pattern in the European sample is caused by a gene conversion event. To be able to partition sequences into derived and ancestral, we used, in addition to fragment *593*, its adjacent fragments *592* and *594* of both population samples. The GCP test (Song et al. 2006) on this data set revealed 2 conversion events in this region of tract length less than 100 bp. The haplotype observed in the European *D. melanogaster* line 12 can be explained by a gene conversion event between sites 33,990 and 34,069, where its tract was donated by the African *D. melanogaster* line 377 and the flanking part by an ancestor of all other European lines (see fig. 3*A* and *B*). A second gene conversion event between the European *D. melanogaster* lines 12, 16, and 18 (fig. 3*A*) and the African *D. melanogaster* line 84 (fig. 3*B*) was detected by the GCP test between sites 34,029 and 34,036. The flanking

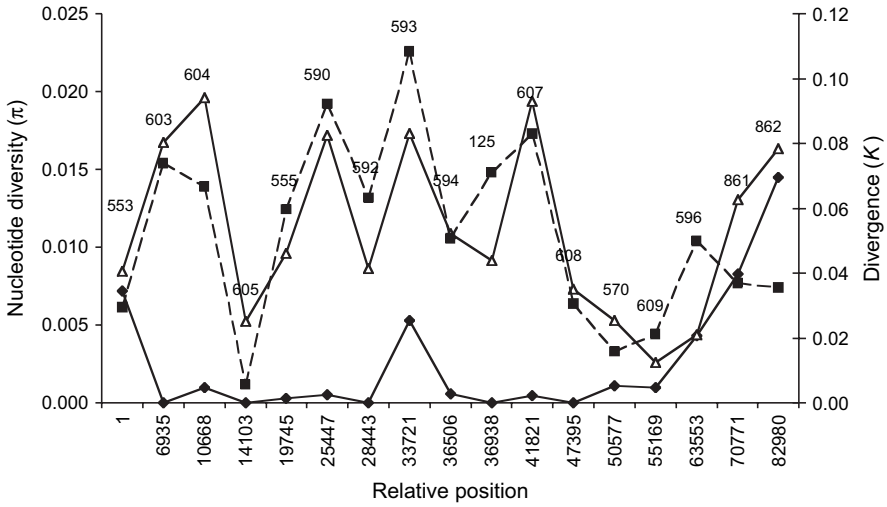

FIG. 2.—Nucleotide diversity (π; Tajima 1983) of the European and African sample and divergence (*K*) against the relative position of each fragment (in bp; see tables 1 and 2). Solid black lines and diamonds correspond to the European, solid black lines and open triangles to the African sample, and dashed black lines and squares to divergence.

A

Fragment	553	604	555	590	593	594	607	570	609	596	861	862
Line		Relative position										
	1122223	7	9	6	9000000	5	9	8	3	5666	9000001	0122222222222223
	8090186	5	6	6	9233466	2	6	1	8	8279	2366791	5001244555576
	1990219	3	0	6	0916552	1	7	2	9404		1423201	0339205123112
D. yakuba					A							- - -
D. simulans	CTGCAAT	C	A	T	AGAT-G	G	G	A	T	GAGG	TCCTGGG	CAGC- - - AGAAC
D. melanogaster 01	AC...CA	.	.	.	G..C..	.	.	.	.	..AT	..GC..A	.C.GGTA....T
D. melanogaster 02	AC.T.C.	.	.	.	G..C..	.	.	.	.	..AT	.....A	GCA.ACGGTGT.
D. melanogaster 11	..A.G..	T	.	.	G..C..	.	.	T	.	....	CTGCAA.	GCA.ACGGTGT.
D. melanogaster 12	.....C.	.	.	C	....TA	.	.	.	A	.C..	..GC..A	...GGTA....T
D. melanogaster 13	AC.T.C.	.	.	.	G..C..	T	A	.	.	..AT	.....A	GCA.ACGGTGT.
D. melanogaster 14	AC...C.	.	.	.	G..C..	.	.	.	.	....	.....A	.C.GGTA....T
D. melanogaster 15	AC.T.C.	.	.	.	G..C..	.	.	T	.	T.A.	..GC..A	.C.GGTA....T
D. melanogaster 16	.....C.	.	.	.	.TC.TA	.	.	.	.	..AT	.TGCAA.	.C.GGTA....T
D. melanogaster 17	...T.C.	.	T	.	G..C..	.	.	T	A	..AT	C.....A	.C.GGTA....T
D. melanogaster 18	.....C.	.	.	.	.TC.TA	.	.	.	.	..AT	.TGCAA.	.C.GGTA....T
D. melanogaster 19	AC.T.C.	.	.	.	G..C..	.	.	T	.	T.A.	..GC..A	.C.GGTA....T
D. melanogaster 20	..A.G..	T	.	.	G..C..	.	.	T	.	....	CTGCAA.	GCA.ACGGTGT.

B

Fragment	593	596
Line	Relative position	
	3333333333333333333333333	6666666666
	3333334444444444444444444	3333333333
	8899990000000000000000111	5566688889
	3924790011233446678803 3	7707945924
	6763903423916057597368674	1420486455
D. yakuba	-AA	- -
D. simulans	GGCGGAGGAGGAT---GATTATCA	C-GGGA-AT
D. melanogaster 82	...A.GAA.ATC.CTTA...G.A.	.CA...A..
D. melanogaster 84	...A..AA..TC.CTTA...G.A.	.C.AT.A..
D. melanogaster 95	A..A..AA.ATC.CTTA...G..G	.C....A..
D. melanogaster 131	...A.G....T..CTT.......G	.C....A..
D. melanogaster 145	..AA..AA..T..CTTA...G.A.	.C....A..
D. melanogaster 157	...A.....T..TT.A...G.A.	.A...GAC.
D. melanogaster 186	...AAG.....CC.T.T..G.A.	.C....G..
D. melanogaster 191	.A.AAG.....CTTA.A...A.	.C....A..
D. melanogaster 229	-----------------------	.C....A.C
D. melanogaster 377	..............CTTA...G.A.	TC....A..
D. melanogaster 384	...A..AAT.T..CTTA..CGA.G	.C..T.A..
D. melanogaster 398	...A.....T..CTTA.....A.	.C....A..

FIG. 3.—Alignment of polymorphic sites observed in 12 lines of the European (*A*) and the African (*B*) *Drosophila melanogaster* sample for each fragment. The relative position (see table 1) and the derived state inferred from *D. melanogaster*/*Drosophila simulans* or *Drosophila yakuba* (see Materials and Methods) comparisons are given for each polymorphic site. Dashes indicate 1-bp deletion.

part of the observed haplotype in lines 16 and 18 came from line 12, whereas line 84 donated the tract sequences (see Discussion).

Departure from Standard Neutral Model

To evaluate the significance of the observed reduction in nucleotide diversity in both populations, we used the multilocus HKA test (Kliman et al. 2000). This test takes the observed differences in mutation rates in our data sets into account by comparing intraspecific diversity and interspecific divergence (Hudson et al. 1987). Although we detected a significant deviation from the neutral expectation for the studied region in the European sample ($\chi^2 = 62.015$, $P = 0.0001$), no departure from the neutral equilibrium model was found for the African sample ($\chi^2 = 10.679$, $P = 0.8289$). Further evidence for a departure from neutrality can be obtained by examining the frequency spectrum. A skew toward low-frequency variation can be measured by Tajima's *D* statistic (Tajima 1989). Under the standard neutral model, Tajima's *D* is expected to be zero. Out of 10 fragments with some variation in the European sample, we

observed 4 fragments (i.e., *555*, *590*, *594*, and *607*; see table 1) with D values significantly less than zero ($P < 0.05$), indicating an excess of singletons (table 1). In comparison, we found only 1 out of 17 loci (i.e., fragment *609*; see table 2) with a significantly negative D value in the African sample. If this skew in the frequency spectrum is due to new mutations, then these singletons should represent derived variants. This can be examined by Fu and Li's D statistic (Fu and Li 1993), which uses an outgroup to identify the state of a mutation. In this statistic, the number of mutations observed in internal and external branches is compared with the expectations under neutrality (Fu and Li 1993). The same fragments that showed a departure from neutrality by Tajima's D statistic also deviated from neutrality for Fu and Li's D statistic in both population samples (see tables 1 and 2). Support for a hitchhiking event can also be obtained from Fay and Wu's H statistic (Fay and Wu 2000). This statistic measures the skew toward high-frequency derived variants. However, we observed a deviation from neutrality in the H statistic neither in the European nor in the African sample (tables 1 and 2). Given the strong haplotype structure in fragment *593* in the European sample, we would expect to find LD among the alleles as well. Using the assumption of no recombination, the Z_{ns} value of 0.76 is significantly higher than expected under neutrality ($P = 0.049$; table 1). In addition, whereas the Z_{ns} values of 3 of the 4 loci flanking the region of reduced variation (see above; table 1) are comparable with the chromosome-wide average (Ometto et al. 2005), one is significantly higher ($P < 0.05$). In contrast, we observed no LD in any fragment of the African sample (table 2). The latter result agrees with those observed for the entire X chromosome (see Ometto et al. 2005).

Estimation of Selective Sweep Parameters

The observed valley of variation and the skew in the frequency spectrum provide strong evidence for the recent occurrence of a selective sweep in the European population (see table 1 and fig. 2). In addition, we observed similar signs of selection in the African population, however, to a different degree (see table 2 and fig. 2). Because we have independent estimates of the effective population size, the mutational parameter θ, and the recombination rate for both populations (see Materials and Methods), we can apply a composite maximum likelihood approach (Kim and Stephan 2002; Kim and Nielsen 2004) to simultaneously test for a hitchhiking event and to estimate the location of the beneficial mutation and the strength of selection using all loci together. Under the assumption of a randomly mating population of constant size and given the estimates of parameters used for the simulations, our data of the European population fit significantly better a hitchhiking than a neutral model using the CLR test proposed by Kim and Stephan (2002; $P < 0.0001$). Furthermore, the estimated strength of selection, s, is 0.0038, and the estimated position of the selected site, X, is 22,147 (i.e., near fragment *590*; fig. 3*A*). The test proposed by Kim and Nielsen (2004), which includes information about LD, however, did not reject neutrality in favor of a hitchhiking model for the European population ($P = 0.1250$). This can be explained by the one-sided LD structure in our region, which is different from the one outlined in Kim and Nielsen (2004; see Discussion). When we applied the CLR test (Kim and Stephan 2002) to the data of the African population, we also observed a significantly better fit to a hitchhiking than a neutral model ($P = 0.0300$). The strength of selection, s, is 0.0016, and the position of the selected site, X, is 56,808 (i.e., corresponding to fragment *609*). We did not apply the Kim and Nielsen test (2004) to this data set because we did not observe LD in the African sample (see table 2). To investigate if the observed polymorphism pattern of both populations is caused by a selective sweep alone or by demographic events (i.e., population structure or a recent bottleneck), we applied the GOF test (Jensen et al. 2005) to both data sets.

Neither the polymorphism pattern of the European nor that of the African population can solely be explained by demographic events ($P = 0.171$ and $P = 0.326$ for the European and the African population sample, respectively). The 95% CIs for X of the European and African population are (2,650, 73,151) and (17,901, 76,183), and for s (0.0004, 0.0110) and (0.0001, 0.0028), respectively (see Materials and Methods). The large CIs for X and s may be due to partial sequencing of the region around *unc-119* used in this part of the analysis (J Jensen, personal communication). Finally, we note that the observed P value of the GOF test for the European population is rather low, which is consistent with either a selective sweep or a recent, severe bottleneck (see Jensen et al. 2005).

Demographic Modeling of the European Population

Because the observed P value of the GOF test for the European population is rather low, we investigated if the reduced variation in the studied region could be the result of a population bottleneck. To do this, we applied an extended version of a maximum likelihood approach (Ometto et al. 2005; Beisswanger et al. 2006) implemented in a coalescent-based program (Ramos-Onsins et al. 2004). We simulated various bottleneck scenarios by changing the time (i.e., $T_b = 0.0100$–0.0500) and strength (i.e., $S_b = 0.340$ and 0.400 for each T_b value) (see table 2). These values were chosen from the parameter range used in Ometto et al. (2005). They reflect bottlenecks starting between 3,000 and 15,000 years ago (assuming 10 generations per year and T_b measured in $3N_e$ generations). In addition, we investigated the effect of recombination between loci (i.e., Method I vs. Method II; see below). However, because this method considers only crossing-over but not gene conversion, we excluded fragment *593* from the analysis. In the case of no recombination between loci (i.e., Method I), the probability of observing at most 7 segregating sites between fragments *603–609* (located in the valley of reduced variation) in the European population is low for all examined bottleneck scenarios (i.e., conditioned on the observation of at most 37 segregating sites in the entire region where fragment *125* was monomorphic; table 3). However, when we assumed some recombination (i.e., 1.926×10^{-8} rec/bp/gen) between loci (i.e., Method II) and conditioned on the same assumptions as in Method I, the polymorphism data can be explained by young bottlenecks ($T_b < 0.02$;

Table 3
Summary of Bottleneck Simulations

T_b	S_b	Method I	Method II
0.0100	0.340	<0.0041	0.2418
	0.400	<0.0049	0.3258
0.0125	0.340	<0.0042	0.1763
	0.400	<0.0044	0.2280
0.0200	0.340	<0.0044	0.0397
	0.400	<0.0045	0.0499
0.0267	0.340	<0.0041	0.0073
	0.400	<0.0045	0.0115
0.0500	0.340	<0.0038	<0.0010
	0.400	<0.0028	<0.0006

NOTE.—T_b represents the age (measured in $3N_e$ generations) and S_b the strength of the bottleneck. Methods I and II are explained in the text (see also Beisswanger et al. 2006).

see table 3). However, such young bottlenecks seem to be unrealistic (Ometto et al. 2005).

Localization of Potential Beneficial Mutation

The predicted site of the beneficial mutation of the European population is located between gene *CG1958* and a cluster of 3 genes, *CG1677*, *CG2059*, and *unc-119*, which are located −7.5 kb and 6.7, 12.3, and 14.6 kb away from the predicted site (fig. 1). For the African population, however, the predicted site is located −13.4 kb from the 5′-region of gene *unc-119* (see Materials and Methods) and within gene *brk* (fig. 1). Because the potential target site of selection is likely to be found in a regulatory or coding region and because the mutation (divergence) rate in the fragments near the 2 genes *CG1958* and *brk* (i.e., *605* and *609* respectively; fig. 2) is low, we focused our investigation on the gene cluster and its 5′-regions. This may partly alleviate the problems arising from the large confidence interval of *X* reported above. We sequenced the 5′ flanking and the coding regions of all 3 genes in the European and African *D. melanogaster* samples and in the *D. simulans* strain. In the 5′ region of the genes *CG1677* and *unc-119* (which are 514- and 401-bp long, respectively), we found neither length differences nor substantial sequence divergence between the European and the African samples (supplemental figs. 1 and 3, Supplementary Material online). However, in the 5′ region of gene *CG2059* (with a length of 504 bp), we observed substantial divergence at 3 sites and a similar haplotype structure as in fragment *593* in the European sample, which extends until (relative) position 34,237 (indicating the centromere-proximal end of the first gene conversion event, where the African *D. melanogaster* line 377 donated its tract to European line 12; see supplemental fig. 2, Supplementary Material online). Further visual inspection of the sequences revealed 1 fixed replacement site in a derived state in *CG1677* and 2 in *CG2059* and 1 fixed replacement site in the ancestral state in *CG1677* and *unc-119* in the European sample (see supplemental figs. 1, 2, and 3, Supplementary Material online).

Discussion

Our study provides evidence that beneficial mutations were recently fixed in the *unc-119* region of a European and an African *D. melanogaster* population, causing the observed valley of reduced variation. In addition, we detected gene conversion events leading to an unusual haplotype pattern in the center of this valley in the European population.

Evidence for a Selective Sweep

Our results suggest that the observed reduction in nucleotide diversity was caused by recent selective sweeps in the European and African *D. melanogaster* populations (similar to the observations of Beisswanger et al. 2006). The origin of the sweep in the European population, however, remains unknown and may not be related to the sweep observed in the African population. The CIs of the predicted sites of selection of both populations are overlapping, which may suggest an African origin of the European sweep. However, they are too large to delimit the genomic positions of the beneficial mutation(s) to sufficiently small regions (see Results).

A candidate for the selected mutation may be found in the 5′ region of gene *CG2059* and within the *CG1677* and the *CG2059* genes. Although we observed 3 fixed substitutions in the 5′ region of gene *CG2059*, it is unlikely that these sites altered the *cis*-regulation of this gene because their distance to each other (i.e., 31 and 38 bp; see supplemental fig. 2, Supplementary Material online) exceeds the value of 14 bases estimated for the mean conservation length of such elements (Richards et al. 2005). Therefore, if the European sweep has not originated in Africa, we propose that the candidates for the selective target are the replacement substitutions occurring in the *CG1677* and the *CG2059* genes about 6.8 kb and 12.3 kb away from the predicted sweep center. Because these mutations are also present in the African *D. melanogaster* sample (see supplemental figs. 1, 2, and 3, Supplementary Material online), it is most likely that the sweep occurred from the standing variation in the ancestral population when *D. melanogaster* colonized Europe 10 to 15 kya (i.e., soft sweep).

Our analyses are based on a model of a sweep associated with a single new mutant. However, this does not affect the analyses because both modes of selection cause the same evolutionary trajectory as long as the frequency of the favored allele is very low at the beginning of the sweep (Innan and Kim 2004; Hermisson and Pennings 2005; Przeworski et al. 2005).

Gene Conversion Associated with the Selective Sweep

Our analysis suggests that 2 gene conversion events associated with the selective sweep are responsible for the strong haplotype structure observed in fragment *593*. Given the observed valley of nucleotide diversity, the following hypothetical scenario can explain the haplotype structure observed in fragment *593* and in the 5′ region of gene *CG2059*. Consider neutral variants linked to a selected mutation going to fixation. In the later phase of the sweep, the first gene conversion event led to the haplotype structure observed in the European lineage 12. Given the frequency of the haplotype observed in the European lineages 16 and 18, the second gene conversion event must

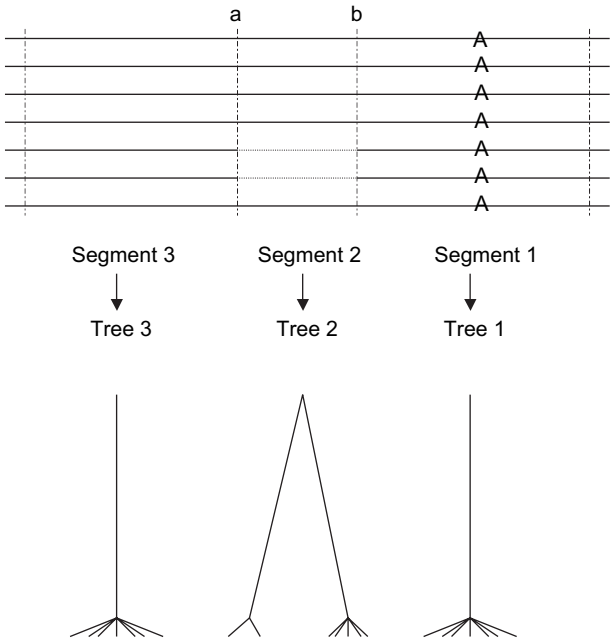

FIG. 4.—An example of DNA sequences (horizontal lines) and the genealogical structure resulting from a recent selective sweep with gene conversion (after fig. 7 in Kim and Nielsen 2004). Solid lines represent sequences originally linked to the beneficial mutation A. Dashed lines represent "recombinant" sequences originally linked to the unfavored allele a but recombined via gene conversion with A during the selective phase. Break points of gene conversion are labeled as a and b. Segments between break points are defined as segment 1, 2, and 3, and the coalescent tree is given below for each segment.

have happened before the favored allele was fixed in the European population. Similarly, Meiklejohn et al. (2004) observed a potential gene conversion tract in which a stretch of ancestral variants was present in an otherwise derived haplotype associated with a selective sweep in the *janus* region of *D. simulans*. However, in this case only a single chromosome showed evidence for gene conversion, suggesting that the conversion event occurred relatively late in the sweep.

A similar pattern of nucleotide diversity has been reported from a natural population of *D. melanogaster* due to a break point of the common cosmopolitan inversion *In(2L)t* (Andolfatto et al. 1999). Although this inversion is probably recent (Andolfatto et al. 1999) and has reached high frequency in a population from the Ivory Coast (Bénassi et al. 1993), a sweep of the *Suppressor of Hairless* gene, *Su(H)*, occurred independently of the inversion in that population (Depaulis et al. 1999; Mousset et al. 2003). However, no chromosomal rearrangement on the X chromosome has been observed in any of the European lines used in this study (Ometto et al. 2005). This reflects the rarity of inversions on the X chromosome in *D. melanogaster*, possibly due to their potential deleterious effect in hemizygous males (Coyne et al. 1991). Only 2 studies reported inversion polymorphism on the X chromosome in natural population of *D. melanogaster* (Das and Singh 1991; Aulard et al. 2002).

If a crossing-over event would have caused the strong haplotype structure observed in fragment *593* and the fixation of the beneficial mutation occurred very quickly, one would expect to find high LD on both sides of the beneficial mutation due to the mutations on the long inner branches (see fig. 7 in Kim and Nielsen 2004). However, LD is expected to decrease quickly due to the increase of recombination break points on both sides of the beneficial mutation, leading eventually to genealogies as expected under neutrality (Kim and Nielsen 2004). When we consider only gene conversion, however, the expected LD pattern is different. Assuming that the gene conversion event happened only on one side of the beneficial mutation A (fig. 4), a genealogy with long inner branches responsible for the high observed LD (segment 2) is surrounded by star-like genealogies (segment 1 and 3). This is due to the relatively short track length of a gene conversion (with a mean of 352 bp; Hilliker et al. 1994). However, if in addition, a crossing-over event happened at some distance to either side of the beneficial mutation during the selective sweep, genealogies will be found as described by Kim and Nielsen (2004; see above). The predicted spatial pattern of LD, which was not present in our study, was detected by Kim and Nielsen (2004) in the sequencing data of a Californian *D. simulans* population (Schlenke and Begun 2004).

The results of our study indicate that the signature of a selective sweep may be obscured by gene conversion events occurring during the course of the sweep. Previous statistical methods that consider only LD caused by reciprocal recombination (Kim and Nielsen 2004) may thus overlook potential sweep regions. A more detailed analysis of the location and length of stretches of high LD may lead to better detection of sweep regions and more accurate mapping of beneficial nucleotide substitutions.

Supplementary Material

Supplementary figures 1, 2, and 3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Y. Kim for advice on his test, L. Ometto for the use of his bottleneck program, Y. Song for comments and advice on his gene conversion test, J. Jensen for the interpretation of our results on CIs, A. Wilken for excellent technical assistance, and J. Parsch, P. S. Pennings, and a reviewer for comments on the manuscript. This work was funded by the Deutsche Forschungsgemeinschaft (STE 325/6) and the Volkswagenstiftung (I/78815).

Literature Cited

Andolfatto P. 2001. Adaptive hitchhiking effects on genome variability. Curr Opin Genet Dev 11:635–41.

Andolfatto P, Wall JD. 2003. Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. Genetics 165:1289–305.

Andolfatto P, Wall JD, Kreitman M. 1999. Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics 153:1297–311.

Aulard S, David JR, Lemeunier F. 2002. Chromosomal inversion polymorphism in Afrotropical populations of *Drosophila melanogaster*. Genet Res 79:49–63.

Bauer DuMont V, Aquadro CF. 2005. Multiple signatures of positive selection downstream of notch on the X chromosome in *Drosophila melanogaster*. Genetics 171:639–53.

Begun DJ, Aquadro CF. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365:548–50.

Beisswanger S, Stephan W, De Lorenzo D. 2006. Evidence for a selective sweep in the wapl region of *Drosophila melanogaster*. Genetics 172:265–74.

Bénassi V, Aulard S, Mazeau S, Veuille M. 1993. Molecular variation of *Adh* and *P6* genes in an African population of *Drosophila melanogaster* and its relation to chromosomal inversions. Genetics 134:789–99.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783–96.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–303.

Comeron JM, Kreitman M, Aguadé M. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics 151:239–49.

Coyne JA, Aulard S, Berry A. 1991. Lack of underdominance in a naturally occurring pericentric inversion in *Drosophila melanogaster* and its implications for chromosome evolution. Genetics 129:791–802.

Das A, Singh BN. 1991. Genetic differentiation and inversion clines in Indian natural populations of *Drosophila melanogaster*. Genome 34:618–25.

David JR, Capy P. 1988. Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet 4:106–11.

Depaulis F, Brazier L, Veuille M. 1999. Selective sweep at the *Drosophila melanogaster Suppressor of Hairless* locus and its association with the In(2L)t inversion polymorphism. Genetics 152:1017–24.

Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405–13.

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

Galtier N, Depaulis F, Barton NH. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155:981–7.

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 165:1269–78.

Hermisson J, Pennings PS. 2005. Soft sweep: molecular population genetics of adaptation from the standing variation. Genetics 169:2335–52.

Hilliker AJ, Harauz G, Reaume AG, Gray M, Clark SH, Chovnick A. 1994. Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. Genetics 137:1019–26.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–9.

Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci USA 101:10667–72.

Jensen JD, Kim Y, Bauer DuMont V, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401–10.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. Genetics 123:887–99.

Kelly JK. 1997. A test of neutrality based on interlocus associations. Genetics 146:1197–206.

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513–24.

Kim Y, Stephan W. 2002. Detecting the local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–77.

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, McCarter J, Wakeley J, Hey J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics 156:1913–31.

Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. Proc Natl Acad Sci USA 99:763–8.

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

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652–4.

Meiklejohn CD, Kim Y, Hartl DL, Parsch J. 2004. Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. Genetics 168:265–79.

Mousset S, Brazier L, Cariou M-L, Chartois F, Depaulis F, Veuille M. 2003. Evidence of a high rate of selective sweeps in African *Drosophila melanogaster*. Genetics 163: 599–609.

Nurminsky D, De Aguiar D, Bustamante D, Hartl DL. 2001. Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. Science 291:128–30.

Ometto L, Glinka S, De Lorenzo D, Stephan W. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* from a chromosome-wide DNA polymorphism study. Mol Biol Evol 22:2119–30.

Orengo DJ, Aguadé M. 2004. Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. Genetics 167:1759–66.

Parsch J, Meiklejohn CD, Hartl DL. 2001. Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. Genetics 159:647–57.

Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics 160:1179–89.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. Evolution 59: 2311–23.

Przeworski M, Wall JD, Andolfatto P. 2001. Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. Mol Biol Evol 18:291–8.

Quesada H, Ramirez UE, Rozas J, Aguadé M. 2003. Large-scale adaptive hitchhiking upon high recombination in *Drosophila simulans*. Genetics 165:895–900.

Ramos-Onsins SE, Stranger BE, Mitchell-Olds T, Aquadé M. 2004. Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics 166:373–88.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, et al. (52 co-authors). 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res 15:1–18.

Rozas J, Sánchez-Del Barrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–7.

Schlenke TA, Begun DJ. 2004. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proc Natl Acad Sci USA 101:1626–31.

Song YS, Ding Z, Gusfield D, Langley CH, Wu Y. 2006. Algorithms to distinguish the role of gene-conversion from

single-crossover recombination in the derivation of SNP sequences in population. Proceedings of Recomb2006. Available from: http://recomb06.dei.unipd.it/. Accessed 2006 Apr 4.
Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human. Mol Biol Evol 22:63–73.
Stephan W, Song YS, Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172:2647–63.
Stephan W, Wiehe T, Lenz MW. 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor Popul Biol 41:237–54.
Storz JF, Payseur BA, Nachman MW. 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol Biol Evol 21:1800–11.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–60.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–95.
Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. 2005. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 21:2791–3.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–76.

Jody Hey, Associate Editor

Accepted June 23, 2006