Population Genetic Approaches to Detect Natural Selection in *Drosophila melanogaster*

Dissertation
der Fakultät für Biologie der
Ludwig-Maximilians-Universität München

vorgelegt von
Sascha Glinka
aus Heilbronn

03.02.2005

1. Gutachter: Prof. Dr. Wolfgang Stephan
2. Gutachter: Prof. Dr. John Parsch

Tag der mündlichen Prüfung: 03.05.2005

SUMMARY

This thesis intends to detect evidence of Darwinian selection and ultimately identify genes and substitutions that were involved in adaptation. The model organism *Drosophila melanogaster* was chosen as the study object, since the availability of the genome sequence and its evolutionary history allows us to investigate ancestral and derived populations.

To identify the footprints of natural selection, the first objective was to locate genomic regions subject to selection. Such footprints involve a reduction in genetic variation along a recombining chromosome caused by a fixation of a beneficial mutation (*i.e*., so-called "selective sweep") in a population under study. Single nucleotide polymorphisms of non-coding regions (*i.e*., 105 fragments) of the X chromosome in a putatively ancestral population of *D. melanogaster* from Zimbabwe were surveyed and compared to a derived European population in the first chapter. In contrast to the European population, evidence of selection was weak in the African population, but a strong signature of a population size expansion was observed. To examine the impact of demography and selection more deeply, an analysis of an enlarged DNA sequencing data set (*i.e*., 253 fragments) of the African population is presented in chapter two. A clear signature of a recent size expansion was observed and the time estimated of the expansion is 15,000 years before present, which was probably caused by drastic climatic changes. The enlarged data set revealed, in addition, that recombination is mutagenic in *D. melanogaster*.

In the second part of this thesis, candidate regions of selective sweeps detected in the genome scan in both populations of *D. melanogaster* were investigated. In chapter three, a more detailed analysis of the region comprising an observed local reduction in variation in one X-linked fragment in the derived European population revealed significant evidence of recent Darwinian selection. The target of selection was attributed to three replacement sites leading to amino acid changes in two predicted genes, *CG1677* and *CG2059*. In contrast, a lower number of haplotypes and a trend for low haplotype diversity suggesting the recent action of a selective sweep was examined in chapter four in the ancestral *D. melanogaster* population. An enlarged DNA sequencing data set revealed another feature unique to a selective sweep, namely the decay in haplotype structure. The target of selection was localized at the 5' region of gene *CG4661*.

In the third part of this thesis, the genetic variation of *D. melanogaster* populations from Southeast Asia were examined to provide first insights into these derived populations and the groundwork for future studies. Since no population genetic approach was done in natural *D. melanogaster* populations from this region, inversions were used as genetic markers. Other than a high frequency of the four common cosmopolitan inversions, there were neither signs for genetic differentiation between populations nor for natural selection. These findings can best be explained by a homogeneous habitat and a joint history of these populations revealing the existence of a panmictic population on Sundaland ~18,000 years ago.

LIST OF ABBREVIATIONS

ACE	adjusted coefficient of exchange
BKK	Bangkok
bp	base pair(s)
CEB	Cebu
CI	confidence interval
cM	centimorgan
CNX	Chiang Mai
CRE	*cis*-regulatory element
HCO	Holocene climatic optimum
HG	hunter-gatherer
HKA	Hudson–Kreitman–Aguadé
IN	Inversions
kb	kilobase(s)
KK	Kota Kinabalu
KL	Kuala Lumpur
kya	thousand years ago
LD	linkage disequilibrium
LGM	last glacial maximum
LR	likelihood ratio
Mb	megabase(s)
MK	McDonald–Kreitman
MRCA	most recent common ancestor
rec/bp/gen	recombination events per base pair per generation
SE	standard error
SNPs	single nucleotide polymorphisms
ST	standard

INTRODUCTION

Owing to this struggle for life, variations, however slight and from whatever cause proceeding, if they be in any degree profitable to the individuals of a species, in their infinitely complex relations to other organic beings and to their physical conditions of life, will tend to the preservation of such individuals, and will generally be inherited by the offspring. The offspring, also, will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive. I have called this principle, by which each slight variation, if useful, is preserved, by the term natural selection.

Charles Darwin, 1859. The Origin of Species

The significance of heritable variation in natural populations has been a central question in evolutionary biology since Darwin introduced his theory of natural selection (Darwin 1859). Many recent studies have indicated that there is extensive polymorphism in protein and DNA sequences within species. However, in most cases, the relationship between molecular variants and organismal fitness is unknown. One of the first theories developed to explain the observed genetic variation was the neutral theory of molecular evolution (Kimura 1983). According to this theory, mutations observed in a population are selectively neutral and therefore have no effect on the carrier's fitness. This theory relies on the assumption that the majority of arising mutations is strongly deleterious and quickly eliminated from the population, while the frequency of those which remain in the population is determined only by random genetic drift.

Under these circumstances, the evolutionary process changes allele frequencies by chance due to random sampling of gametes at each generation, whereby the sampling process is only influenced by the individuals that take part in reproduction in a given generation (*i.e.*, effective population size; Graur and Li 1999, p. 39). Eventually, if the sampling process continues for long periods of time, the allele frequency reaches either 0 (*i.e.*, extinction) or 1 (*i.e.*, fixation). However, at any given time, some loci will posses alleles at intermediate frequencies making these polymorphic loci. Thus, under the neutral theory, the level of genetic variation within a population is determined by its effective population size and the rate of newly arising mutations (*i.e.*, neutral mutation rate; Kimura 1983). Due to the nature of random genetic drift, most of the differences found between two species (*i.e.*, divergence) can be assumed to have accumulated at the same rate as new mutations arise. The level of divergence between species is therefore determined by the neutral mutation rate and the time of the splitting of the species from their common ancestor.

During the past 15 years, studies of genetic variation on *Drosophila* and other species resulted in observations inconsistent with the predictions of the neutral model (*e.g.*, Aguadé *et al*. 1989; Stephan and Langley 1989; Begun and Aquadro 1992). Most of these studies found a strong correlation between the local meiotic rate of recombination and levels of nucleotide diversity, which in a neutral framework could only be explained by a higher mutation rate in regions of high recombination (Kimura 1983). However, levels of divergence between closely related species were not affected by recombination (Berry *et al*. 1991; Begun and Aquadro 1992), as would be expected under the neutral theory. Two alternative models involving natural selection were proposed to explain the observed reduction in variability:

the selective sweep model (Maynard Smith and Haigh 1974) and the background selection model (Charlesworth *et al*. 1993, 1995). The selective sweep model predicts that genetic variation at neutral sites is suddenly wiped out due to genetic linkage (*i.e*., correlation of genealogical histories among nucleotide sites) to a rapidly fixed beneficial mutation. This so-called "hitchhiking" of neutral alleles with the selected allele persist in the population unless recombination between them breaks down the association. Therefore, the size of the hitchhiked region depends on the strength of selection (*i.e*., selective advantage of the beneficial mutation) and the local rate of recombination (Kaplan *et al*. 1989; Stephan *et al.* 1992). If selection coefficients are similar across the genome, repeated episodes of hitchhiking (*i.e*., recurrent selective sweeps) will affect loci in regions of low recombination more severely than loci in regions of normal recombination (Kaplan *et al*. 1989; Braverman *et al*. 1995; Gillespie 2000).

In contrast, the background selection model explains the observed correlation between levels of variation and recombination by purifying selection against strongly deleterious mutations (Charlesworth *et al.* 1993, 1995). In this model, a neutral allele will persist in the population only if it finds itself on a deleterious-mutation-free chromosome (or segment of a chromosome), either when it first arises in the population or when it is no longer linked to such a deleterious mutation by recombination (Charlesworth *et al.* 1993, 1995; Hudson and Kaplan 1994). If the average selection coefficients and deleterious mutation rates are the same in different regions of the genome, the rate of recombination will determine the extent of the reduction in neutral diversity (*i.e*., the extent to which neutral alleles can escape from background selection). Therefore, background selection will greatly reduce genetic variation only in regions of low recombination, while genetic hitchhiking is expected to leave a footprint along a recombining chromosome (*i.e*., regions with intermediate to high recombination rates; Kim and Stephan 2002). In these regions, the greatest impact on genetic variation will be at the site of selection, but it will weaken with increasing distance from the selected site thereby producing a valley of reduced variation.

The neutral theory can explain genetic variation in both nonfunctional (*i.e*., most intergenic regions, introns, and degenerate positions of codons) and functional regions (*i.e*., 5' flanking regions, exons, 3' flanking regions) of the genome. Since the replacement of one nucleotide by another (*i.e*., substitution; Graur and Li 1999, p. 5) in nonfunctional regions has no effect on protein synthesis, these silent (or synonymous in the case of a degenerate codon position; Graur and Li 1999, p.

5) mutations can be modeled as selectively neutral mutations (KREITMAN 2000). Selection, however, can cause a reduction in variation in these regions if they are genetically linked with a variant under selection (see above). Similarly, in functional regions, substitutions leading to a change in an amino acid (*i.e.*, nonsynonymous change or replacement; GRAUR and LI 1999, p. 5) may persist in a population, if they are neutral. Therefore, the hypothesis of selective neutrality can be used as a null hypothesis against which to test for evidence of directional selection (KREITMAN 2000).

The basis of most tests for selection is the standard neutral model, where the applied test compares some feature of observed polymorphism data with that expected under this neutral model, which incorporates a mutation and a reproduction model (*e.g.*, KREITMAN 2000). The former model assumes that new mutations occur at sites that were previously monomorphic (*i.e.*, infinite-sites model; KIMURA 1969, 1971; WATTERSON 1975), whereas the latter model assumes a population of constant and finite size N_e with random mating, no population structure, and non-overlapping generations that is at mutation-drift equilibrium (*i.e.,* Wright-Fisher model; FISHER 1930; WRIGHT 1931). Taking both models together, neutrality can be modeled assuming selectively neutral mutations arising in a diploid population with size N_e with probability μ per generation (*i.e.*, infinite-sites-neutral-equilibrium model; *e.g.*, KREITMAN 2000). Thus, under the standard neutral model, the expected nucleotide variation for a diploid is given by the population mutation parameter, θ, which can be estimated by $4N_e\mu$.

Under neutrality, genome regions that evolve at high rates (*i.e.*, show a high θ) should also exhibit high levels of divergence. The fit of the correlation between polymorphism and divergence to the neutral model can be evaluated by the Hudson–Kreitman–Aguadé test (HKA; HUDSON *et al*. 1987), which is a goodness-of-fit test to both quantities. In this test, the comparison to a reference (*i.e.*, neutrally evolving) locus from the same population sample is used to infer whether an observed reduction in heterozygosity in another locus is due to a lower mutation rate or to a selective event. Another commonly used test that makes use of divergence data is the McDonald–Kreitman test (MK; MCDONALD and KREITMAN 1991). Here, divergence is measured as the number of monomorphic sites within both species but differing between species (*i.e.*, fixed differences). The MK tests the neutral prediction that the ratio of nonsynonymous to synonymous fixed differences between species is the same as the ratio of nonsynonymous to synonymous polymorphisms within species.

Other test statistics are based on comparing different estimates of the parameter θ. Under neutrality, the difference between these estimators has an expected value of zero. TAJIMA (1989a), for example, proposed a test statistic, *D*, where the two estimators θ_W (WATTERSON 1975) and π (TAJIMA 1983) derived from the total number of segregating variant sites in a sample (*i.e.*, number of segregating sites) and from the average probability that two nucleotides will differ between two randomly chosen sequences (*i.e.*, the average number of pairwise differences), respectively, are compared. If there has been no recombination event between a neutral and the selected site during the sweep phase, hitchhiking is complete and all variation is removed from the neutral locus. Due to new mutations accumulating in a population subsequent to a hitchhiking event, variants will be first present at low frequencies due to the lack of time before they can drift to intermediate or high frequencies. Thus, recent hitchhiking events produce a skew in the distribution of nucleotide polymorphism frequencies in a population sample (*i.e.*, frequency spectrum) towards low frequency variants (BRAVERMAN *et al*. 1995). Since θ_W is most sensitive to rare variation, whereas π is most sensitive to intermediate frequency variation, the *D* statistic will be negative (TAJIMA 1989a). Similarly, FU and LI (1993) proposed the *D** statistic which can detect a skew in the frequency spectrum towards low frequency variants by examining the difference between the estimators θ_W and θ_η, where θ_η is estimated from the number of singletons. In addition, the statistic *D** uses an outgroup to distinguish between a recent mutation on a short external branch and an ancient mutation inherited by all but one member of the sample.

In the presence of recombination, hitchhiking is incomplete (*i.e.*, partial selective sweep) when a neutral locus was linked to the selected one only partially during the sweep phase and the frequency of a neutral variant depends on whether it belongs to the same lineage as the advantageous mutation or not (FAY and WU 2000). Thus, subsequent to a strong hitchhiking event, neutral variation is found at either high or low frequencies and thus forms a bipartite frequency spectrum. The statistic *H* (FAY and WU 2000) compares θ_H, which measures an excess of high frequency variants, to π, and an outgroup is used to distinguish between high and low frequency derived variants.

However, violations of the assumptions of the standard neutral model, in particularly by demographic events, could lead to misinterpretation of the applied statistics (*e.g.*, KREITMAN 2000; ANDOLFATTO 2001a). For example, a skew in the frequency spectrum towards low frequency variants is also expected by population expansion or after a strong bottleneck. In both cases, the statistics *D* and *D** would be negative reflecting

a false hitchhiking event (*e.g.*, TAJIMA 1989b), whereas variation is only reduced in the latter case. To disentangle demographic from selective events one needs to use data from multiple loci sampled from the same population, since demographic events would affect the whole genome, whereas selection acts only locally (ANDOLFATTO 2001a).

Both demographic and selective events influence allelic configurations of multiple markers (*i.e.*, haplotypes) that are present on a single chromosome of a given individual sampled from a population. During the spread of an advantageous mutation through a population, a haplotype of very tightly linked neutral variants will increase in frequency until fixation. However, with increasing distance from the selected site, more alleles will escape complete hitchhiking due to recombination (see above) leading to an increase in the number of haplotypes or haplotype diversity (*i.e.*, decay in haplotype structure; DEPAULIS *et al*. 2005). Therefore, a strong haplotype pattern may be present if the rate of recombination is low enough so that there is no recombination within a sequence surveyed but high enough so that variation remains segregating subsequent to hitchhiking. This could lead to linkage disequilibrium (LD) between sites, and the degree of LD between two alleles can be measured by their correlation coefficient r^2 (ARDLIE *et al*. 2002). Considering a sample of sequences, all pairwise comparisons of *S* segregating sites (see above) can then be summarized through the measure Z_{nS} (KELLY 1997) by averaging over their correlation coefficients.

A number of haplotype tests have been developed to detect a high frequency haplotype or a lack of haplotype diversity that may occur during or subsequent to a hitchhiking event. HUDSON *et al*. (1994) developed a test, H_P, to determine the probability of observing a given number of segregating sites, *S*, or fewer in a subset of sequences from a sample. DEPAULIS and VEUILLE (1998) proposed two tests conditioned on *S*, *K*- and the *H*-haplotype tests, which are based on the distribution of the haplotype number, *K*, and the haplotype diversity, *H*. For both statistics, low values could result from structuring of polymorphic sites into few haplotypes due to selective events, such as incomplete hitchhiking or hitchhiking with partial linkage. However, low values of these statistics could also result from demographic events, such as population substructure and recent bottlenecks (DEPAULIS and VEUILLE 1998). In contrast, high values can result from either an old complete hitchhiking event without recombination or population expansion (DEPAULIS and VEUILLE 1998).

To evaluate the significance of an observed level of nucleotide diversity with respect to various models (*i.e*., hitchhiking *vs*. neutral model), one can use the coalescence approach. Coalescence is the merging of ancestral lineages going backwards in time until the most recent commom ancestor (MRCA) of a particular set of sequences has been found (Kingman 1982; Hudson 1990, 1993). Here, all existing copies of a particular site must be related to each other and to a MRCA through a genealogical tree. Polymorphism is due to mutations that occurred along the branches of this tree, and the frequency of each sequence variant is determined by the fraction of branches that inherits the variant (Rosenberg and Nordborg 2002). Therefore, the pattern of polymorphism reflects both the history of the coalescence of lineages, which give rise to the tree, and the mutational history. In comparison to a genealogy under the neutral model, hitchhiking will lead to a star-like genealogy due to the relatively young variants in low frequencies. To examine if the observed data fits better to a neutral than to a selection model, the coalescent can be used as a simulation tool (*e.g*., Rosenberg and Nordborg 2002). Here, the distribution of a given test statistic obtained from many possible simulated neutral data sets can be compared to the value estimated from the real data set. If patterns that are characteristic of the actual data are rarely seen in the simulations, the null hypotheses favoring, for example, the neutral model can be rejected.

To detect evidence of positive directional selection (*i.e*., Darwinian selection) and ultimately identify genes and substitutions that were involved in adaptation are the main goals of this thesis. Since beneficial mutations occur infrequently (*e.g*., every 1250 generations; Stephan 1997) and cause relatively small differences in fitness (Orr and Coyne 1992) thus limiting laboratory experiments due to their relatively short time scales, I used natural populations of the model organism *D. melanogaster* to accomplish these goals. The availability of the genome sequence of this species (Adams *et al*. 2000) has allowed me to screen large genomic regions to search for footprints of natural selection (see above). In addition, because it is widely accepted that *D. melanogaster* originated in sub-Saharan part of the African mainland and extended its range towards Europe and Asia 10 to 15 thousand years ago (kya; David and Capy 1988), not only can patterns of past selective sweeps in the ancestral population be observed, but also adaptation to newly colonized habitats in temperate and tropical zones of *D. melanogaster* can be examined.

This thesis is structured in three parts. In the first part, I implemented a genome scan of variation to search for genomic regions, which have been shaped recently by selective sweeps. To do this, I surveyed single nucleotide polymorphisms (SNPs)

in non-coding regions of the X chromosome in a putatively ancestral population of *D. melanogaster* from Zimbabwe (described in chapter one). Sequencing data gathered from a population sample of 12 isofemale lines (*i.e.*, each line was established by one inseminated female) mainly from genomic regions of intermediate to high recombination rates allowed me to detect local signatures of directional selection, which can be distinguished from other selective forces (*i.e.*, background selection) and chromosome-wide features of demographic events. The observed genetic variation in the ancestral population is compared to a derived *D. melanogaster* population from Europe to highlight differences of the evolutionary history of both populations.

To examine the impact of demography and selection in the ancestral population of *D. melanogaster* more deeply, I extended this multi-locus scan using the same population sample to increase the density of analyzed non-coding regions in chapter two. Since patterns of nucleotide variation may be influenced by inversions (*i.e.*, portions of the chromosome whose gene order is reversed relative to the standard reference orientation; Sturtevant 1917) and most species of the genus *Drosophila* are polymorphic for inversions (Krimbas and Powell 1992), I examined all isofemale lines used for any chromosomal rearrangements (although inversions are rare on the X chromosome in natural populations of *D. melanogaster*; Krimbas and Powell 1992). The enlarged data set allowed me to investigate the observed correlation between nucleotide variation and recombination rates in terms of its selective origin. This is important, since Hellmann *et al*. (2003) found strong evidence of a neutral explanation for this observation in humans. Furthermore, I was able to provide additional insights into the evolutionary history of this species by disentangling the observed genetic patterns shaped by the genomic-wide effects of demography from the locus-specific effects of natural selection.

In the second part of my thesis, I focused on the analysis of candidate regions of selective sweeps detected in the genome scan in both populations of *D. melanogaster*. In chapter three, I examined if an observed local reduction in variation on the X chromosome in the European *D. melanogaster* population identified in chapter one is caused by Darwinian selection. Such a result provides strong evidence for the adaptation process of this species to temperate zones. To accomplish this goal, I gathered more X-linked sequencing data delimiting the region of reduced variation of the European population. Detailed analysis of this region by various maximum-likelihood approaches (Kim and Stephan 2002; Kim and Nielsen 2004; Ometto, unpublished) allowed me to examine if the observed pattern is caused by positive

directional selection, random genetic drift or demographic processes (*i.e.*, population bottlenecks).

In chapter four, I investigated an observed haplotype structure in six adjacent X-linked loci in the African population of the genome scan outlined in chapter one. These loci showed an overall deficit in the number of haplotypes and haplotype diversity suggesting that this pattern was shaped by directional selection in the ancestral population. In contrast to chapter three, where a reduction in variation was examined, I focused in this chapter on another expected feature unique to a selective sweep, namely the decay in haplotype structure. I analyzed an enlarged data set and applied a newly developed maximum-likelihood approach (Mousset *et al.*, submitted) to evaluate if the observed haplotype structure is better explained by a neutral or a selective sweep model.

In the third part of my thesis, I examined derived populations of *D. melanogaster* from Southeast Asia. This geographical region is particularly interesting because the ecological conditions differ from those present in Europe and evidence of a Far Eastern *D. melanogaster* race supports the hypothesis that this species might have colonized the Southeast Asian region earlier than Europe (Lachaise and Silvain 2004). Since no population genetic approach has been undertaken of natural populations in Southeast Asia, I focused on the analysis of inversion polymorphisms. Inversions are important genetic markers for the genus *Drosophila* (*e.g.*, Krimbas and Powell 1992). Therefore, I examined the major autosomal arms and the X chromosome of five Southeast Asian *D. melanogaster* population samples for chromosomal rearrangements. This analysis provides the first insights into these derived population samples of *D. melanogaster* and the groundwork for future studies of similar nature as done for the European population.

Part I: Genome Scan of Variation

CHAPTER 1

Demography and Natural Selection Have Shaped Genetic Variation in *Drosophila melanogaster*: A Multi-locus Approach

1.1 INTRODUCTION

In the past decade, evidence that natural selection plays a key role in shaping genome-wide patterns of variability in *Drosophila* has been mounting (Aquadro 1997). However, it remains a challenge to discern selection from other forces, particularly demographic factors. Only recently, studies have begun to address this problem by consistently sampling populations and using multiple loci (Begun and Whitley 2000). The rationale of this approach is that demographic processes affect the entire genome in a similar way, whereas selective forces tend to leave locus-specific footprints that are detectable in a genome-wide survey.

Drosophila melanogaster, originating from sub-Saharan Africa, is believed to have expanded its range after the last glaciation (*i.e.*, ~10–15 kya; David and Capy 1988; Lachaise *et al.* 1988). During this habitat expansion, demographic processes (such as bottlenecks and subsequent population size increases) would be expected to have occurred. In addition, selective events are likely to have played an important role in the adaptation of this species to its new environments.

To distinguish demographic and selective processes important for the recent adaptations of *D. melanogaster*, we compared a putatively ancestral population from Africa (Zimbabwe) with a derived population from Europe (The Netherlands). Since a whole-genome scan of DNA sequence variation is currently not feasible, we used a multi-locus approach. The availability of the genomic sequence of *D. melanogaster* made this approach possible. To be able to discern different selective regimes, we focused on chromosomal regions of normal recombination (Kim and Stephan 2002). Furthermore, we used sequence variation rather than microsatellites (Harr *et al.* 2002) for the following reasons. One of our long-term goals is to estimate the rate of advantageous substitutions in the recent past of *D. melanogaster.* Advantageous substitutions causing sweeps that have occurred no longer than ~0.1N_e (effective population size) generations ago can be detected with sufficiently high power using SNPs (Kim and Stephan 2000; Przeworski 2002). For *D. melanogaster*, 0.1N_e generations correspond to ~10,000 to 15,000 years. This window of time matches

very well the colonization of Europe by *D. melanogaster*. Thus, the use of DNA sequence variation should enable us to detect most of the sweeps that have occurred during this colonization period and hence to obtain a reliable estimate of the rate of advantageous substitutions. In contrast, with microsatellites that mutate faster than nucleotides we may be able to observe only the very recent sweeps. Since this is the first screen of DNA sequence variation in *D. melanogaster*, we concentrated on the X chromosome.

1.2 MATERIALS AND METHODS

1.2.1 Population Samples

D. melanogaster data were collected from 24 highly inbred lines derived from two populations: 12 lines from Africa (Lake Kariba, Zimbabwe; Begun and Aquadro 1993) and 12 lines from a European population (Leiden, The Netherlands). The Zimbabwe lines were kindly provided by C. F. Aquadro, and the European ones were provided by A. J. Davis. Furthermore, a single *D. simulans* inbred strain (Davis, CA, USA; kindly provided by H. A. Orr) was used for interspecific comparisons.

1.2.2 PCR Amplification and DNA Sequencing

On the basis of the available DNA sequence of the *D. melanogaster* genome (Flybase 2000, Release 2, http://www.flybase.org), we amplified and sequenced 105 fragments of non-coding DNA (from 63 introns and 42 intergenic regions), randomly distributed across the entire euchromatic portion of the X chromosome. Most fragments are located in regions of intermediate to high recombination rates. However, 11 fragments are from the telomeric region exhibiting low recombination rates, *i.e.*, distal to the *white* locus (see Appendix 1.1 and 1.2). We amplified and sequenced the homologous 105 fragments in a single strain of *D. simulans*.

We extracted genomic DNA from 10 females of each inbred line using the PUREGENE™ DNA Isolation Kit (Gentra Systems, Minneapolis, MN, USA). The PCR products were then purified with EXOSAP-IT (USB, Cleveland, OH, USA). Sequencing reactions were performed for both strands according to the protocol of the DYEnamic ET terminator cycle sequencing kit (Amersham Biosciences, Buckinghamshire, UK) and run on a MegaBACE 1000 automated capillary sequencer (Amersham Biosciences). Analysis of the data was done using the software Cimarron 3.12 (Amersham Biosciences) for lane tracking and base calling. Only good-quality sequences (MegaBACE quality score of at least 95 of 100) were aligned and checked manually with the application Seqman of the DNAstar (Madison, WI, USA) package. Singletons were confirmed by reamplification and resequencing. The sequences

were deposited in the EMBL database (for accession numbers, see Appendix 1.1–1.3).

1.2.3 Statistical Analysis

Basic population genetic parameters were estimated with the program DnaSP 3.98 (Rozas and Rozas 1999). Levels of nucleotide diversity were estimated using π (Tajima 1983) and θ_W (Watterson 1975). For this analysis, we considered the total number of mutations rather than the number of segregating sites, because in a few instances we observed three different nucleotides segregating at the same position.

To test the neutral equilibrium model, we employed the multi-locus HKA and Tajima's *D* tests (Hudson *et al.* 1987; Tajima 1989a). Both tests were done using the program HKA, kindly provided by J. Hey (http://lifesci.rutgers.edu/~heylab), in which the test statistics were compared with the distributions generated from 10,000 coalescent simulations (Kliman *et al.* 2000).

In addition, we used the following statistics: the number of haplotypes, *K*, and the haplotype diversity, *H* (Depaulis and Veuille 1998), and, for the African population, Fay and Wu's *H* (Fay and Wu 2000). These statistics were calculated with the program DnaSP 3.98 (Rozas and Rozas 1999). We generated the empirical distributions of these statistics for each fragment using coalescent simulations (10,000 iterations; Hudson 1990, 1993), conditioned on the number of segregating sites, *S* (Depaulis *et al.* 2001), and a population recombination rate, *R* (programs are available from S.M.). Since in *D. melanogaster* there is no recombination in males, the population recombination rate, *R*, was estimated by $2N_e c$, where *c* is the female recombination rate per fragment per generation (Przeworski *et al.* 2001). N_e was assumed to be 10^6 (Li *et al.* 1999), and, for each fragment, *c* was estimated by multiplying the per-site-recombination-rate, *r* (see below), by its length, *lth*.

1.2.4 Recombination Rate

We estimated *r* [expressed in recombination events per base pair per generation (rec/bp/gen)] for each fragment as follows. We used a computer program of Comeron *et al.* (1999) to obtain an estimate of the recombination rate for each fragment. This algorithm follows the method of Kliman and Hey (1993). We compared our results to two other estimators of the recombination rate: the adjusted coefficient of exchange (ACE; Begun and Aquadro 1992) and the procedure proposed by Charlesworth (1996).

For the latter method, we used the absolute position of each fragment to calculate physical distances. The estimate of the recombination rate is therefore expressed in centimorgans per megabase (cM/Mb) instead of centimorgans per band (cM/band; see CHARLESWORTH 1996). We divided the X chromosome into two regions containing all of our 105 fragments: (I) the distal-*white* region (0.2–2.45 Mb, 0.02–1.5 cM), and (II) the proximal-*white* region (2.45–16.89 Mb, 1.5–56.7 cM). Following CHARLESWORTH (1996), the *white* locus (2.45 Mb, 1.5 cM) was chosen as a transition point between region I and region II.

1.2.5 Demographic Modeling of the European Population

Because extant European *D. melanogaster* are believed to be derived from an ancestral African population (DAVID and CAPY 1988), we tested the observed data against simple demographic null models: (i) a constant-population-size model and (ii) a population-size-bottleneck model with subsequent expansion (WALL *et al*. 2002; LAZZARO and CLARK 2003). In the latter model, we simulated a population of initial effective size N_0, crashing T_b generations ago to size N_b. After T_m generations, the population was allowed to grow exponentially to the current effective population size, N_c.

The following parameters had to be specified for each fragment: the mutational parameter, θ (estimated from data); the sample size, *n*; and fragment length, *lth*. Constant-population-size models were tested using the observed average θ_W value of the European population, while the bottleneck models were conditioned on the observed average θ_W value of the African population (*i.e*., the value of the hypothetical ancestral population). Our simple models assumed no intragenic recombination but did assume free recombination between fragments. We used several combinations of values of N_b, N_c/N_0 and T_b. T_m was adjusted to obtain a total number of segregating sites in a simulation close to the observed value of 737. For each fragment, 10,000 genealogies were simulated using the program “ms” (HUDSON 2002) under the demographic models mentioned above. The probability of observing exactly F_c = 13 fragments with no polymorphism in our simulation (see RESULTS) was then calculated as the proportion of simulated samples with exactly 13 fragments with no polymorphic sites. This probability was used in a two-tailed likelihood-ratio test as a likelihood of our observation; when the probability $<10^{-4}$, we used 10^{-4} as a conservative overestimate of this value.

1.3 RESULTS

DNA sequences for 105 X chromosome fragments were obtained from 10–12 lines of an African and a European population of *D. melanogaster* (with an average of 11.9 lines per sample). The size of the fragments varied between 240 and 781 bp (excluding insertions and deletions) with a mean (SE) of 517 bp (11 bp). The total region from which these fragments derive spans ~14 Mb. This results in an average distance between adjacent fragments of ~140 kilobases (kb; Figure 1.1).

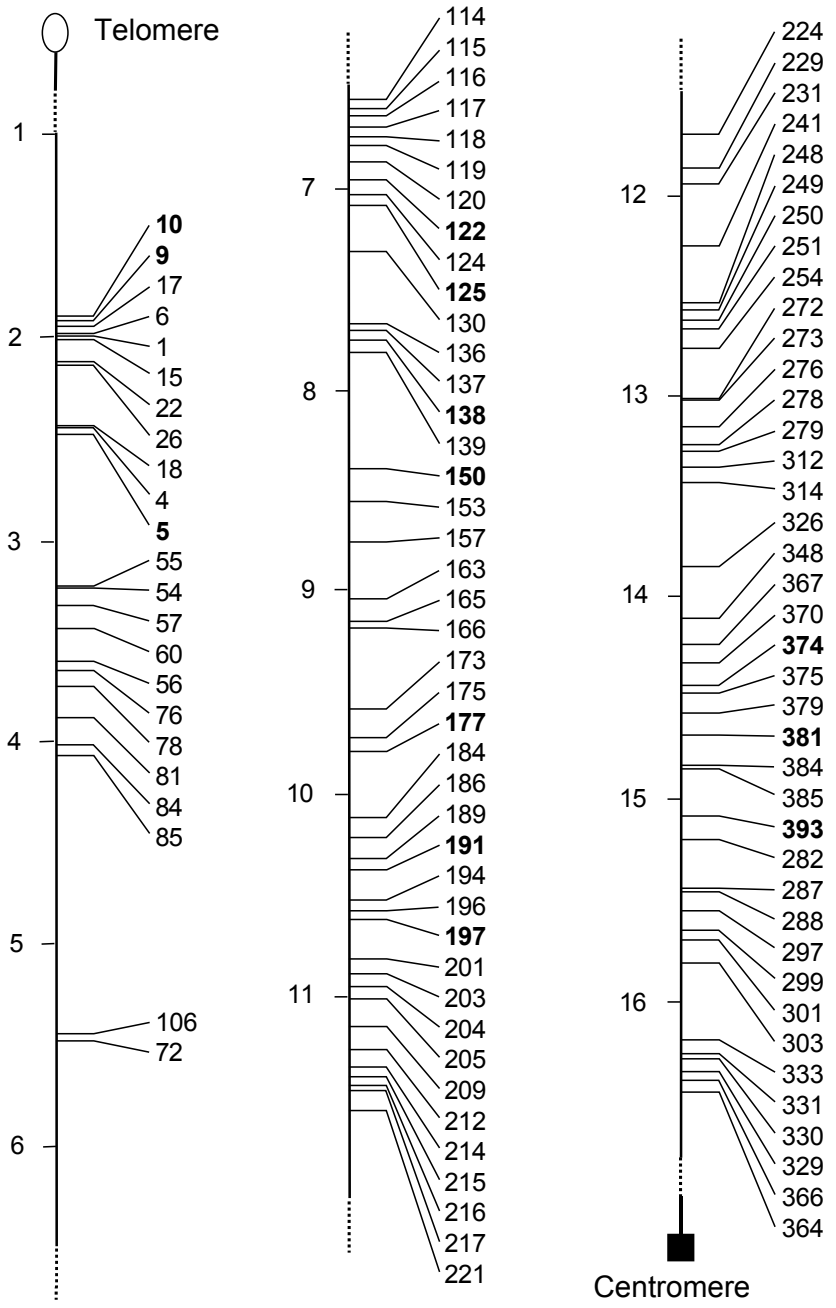


FIGURE 1.1 Distribution of the sequenced fragments along the X chromosome. Fragments are shown by their absolute position (distances in Mb from the telomere). Fragments with no polymorphism in the European sample are in boldface type.

There are several large gaps in our genome scan (Figure 1.1), in which we could not recover a sufficient number of sequences (*i.e.*, at least 10 per sample and the sequence of the *D. simulans* line). The majority of fragments (103) are located in two segments (between coordinates 1.9 and 4.1 Mb, and between 6.5 and 16.4 Mb from the telomere, respectively), thus spanning a region of 12 Mb with an average distance of 119 kb between fragments. The region between these two segments appears to contain a high density of repetitive DNA (for instance microsatellites; Harr *et al*. 2002) that may have caused problems with PCR and sequencing. The details are being investigated.

In both *D. melanogaster* samples, intergenic regions and introns did not produce significantly different results when analyzed separately (results not shown) and are therefore pooled in the following analyses.

1.3.1 Polymorphism Patterns in the African Population

Figures 1.2, a–c, and Appendix 1.1 and 1.3 provide a summary of the polymorphism and divergence data. Of the 54,944 sites sequenced (excluding insertions and deletions), 2057 are polymorphic. The mean of θ_W (SE) is 0.0127 (0.0007), which is higher than the average value of 0.0071 reported for non-coding regions on the *D. melanogaster* X chromosome (Moriyama and Powell 1996), but lower than the average value of 0.0257 estimated for synonymous X-linked sites for African populations from diverse geographic localities (Andolfatto 2001b). For π, the result is similar: 0.0112 (0.0007) to 0.0074 (Moriyama and Powell 1996) and 0.0242 (Andolfatto 2001b).

We tested our data for compatibility with the neutral equilibrium model. The HKA test is used to determine whether the levels of intraspecific polymorphism and interspecific divergence at our set of fragments are consistent with the equilibrium model (Hudson *et al*. 1987). A multi-locus version of the original HKA test was applied to all 105 fragments in the African sample (Figure 1.3a). No significant departure from the equilibrium model was detected (X^2 = 93.31, P = 0.765; Appendix 1.3).

We also calculated the Tajima's D statistic for each fragment and tested whether the observed average across fragments was consistent with the equilibrium model by estimating the critical values of this distribution from coalescent simulations (see materials and methods). In these simulations, we assumed no intragenic recombination (but free recombination between fragments). The African population shows a negative average value (SE) of Tajima's D of –0.578 (0.058). None of the

10,000 simulated samples of 105 fragments had a more extreme average value of *D*. This suggests that our data depart from the neutral equilibrium model. In fact, most of the fragments have negative *D* values (sign test, two-tailed, $P < 0.001$; Figure 1.2d).

To further investigate the pattern of variation in the African sample, we focused on two statistics, the number of haplotypes, *K*, and the haplotype diversity, *H* (Depaulis and Veuille 1998). Low values of these statistics indicate that there are too few haplotypes in the sample due to demographic (*e.g.*, population substructure and/or weak bottlenecks) and/or selective events (*e.g.*, incomplete hitchhiking; Depaulis and Veuille 1998). On the other side, high values can result from population expansion or old complete hitchhiking events (Depaulis and Veuille 1998). Because recombination tends to increase both statistics, we used the estimated recombination rate (Comeron *et al.* 1999; see materials and methods) for each fragment in the coalescent simulations. Assuming that this recombination rate is correct, we can perform a two-tailed test. Under neutrality, we expect an equal proportion of the observed values to be lower and higher than the simulated median.

We found that the observed haplotype diversity, *H*, was higher than the simulated median in 78 of the 105 fragments; this proportion is significantly larger than expected (sign test, two-tailed, $P < 0.001$; Appendix 1.1). For the number of haplotypes, *K*, a significant trend toward a higher number was also observed (sign test, two-tailed, $P = 0.03$; Appendix 1.1). High values of haplotype diversity and large numbers of haplotypes can result from a star-like genealogy due to population expansion or complete hitchhiking events (Depaulis and Veuille 1998).

Assuming that recurrent complete selective sweeps occur along a recombining chromosome, we expected to detect the footprints of partial sweeps as well. We thus examined whether there is evidence for partial hitchhiking events using the *K*- and *H*-haplotype tests (Depaulis and Veuille 1998) and Fay and Wu's *H* test (Fay and Wu 2000). Since we were exploring possible departures of these statistics at their lower bounds, we used the conservative assumption of zero recombination (Depaulis and Veuille 1998). For the 105 fragments, we observed only one significant Fay and Wu's *H* value (one-tailed, $P = 0.03$).

These results, together with the observations from the HKA test, argue against a model of recurrent selective sweeps (Braverman *et al.* 1995) as an explanation of the

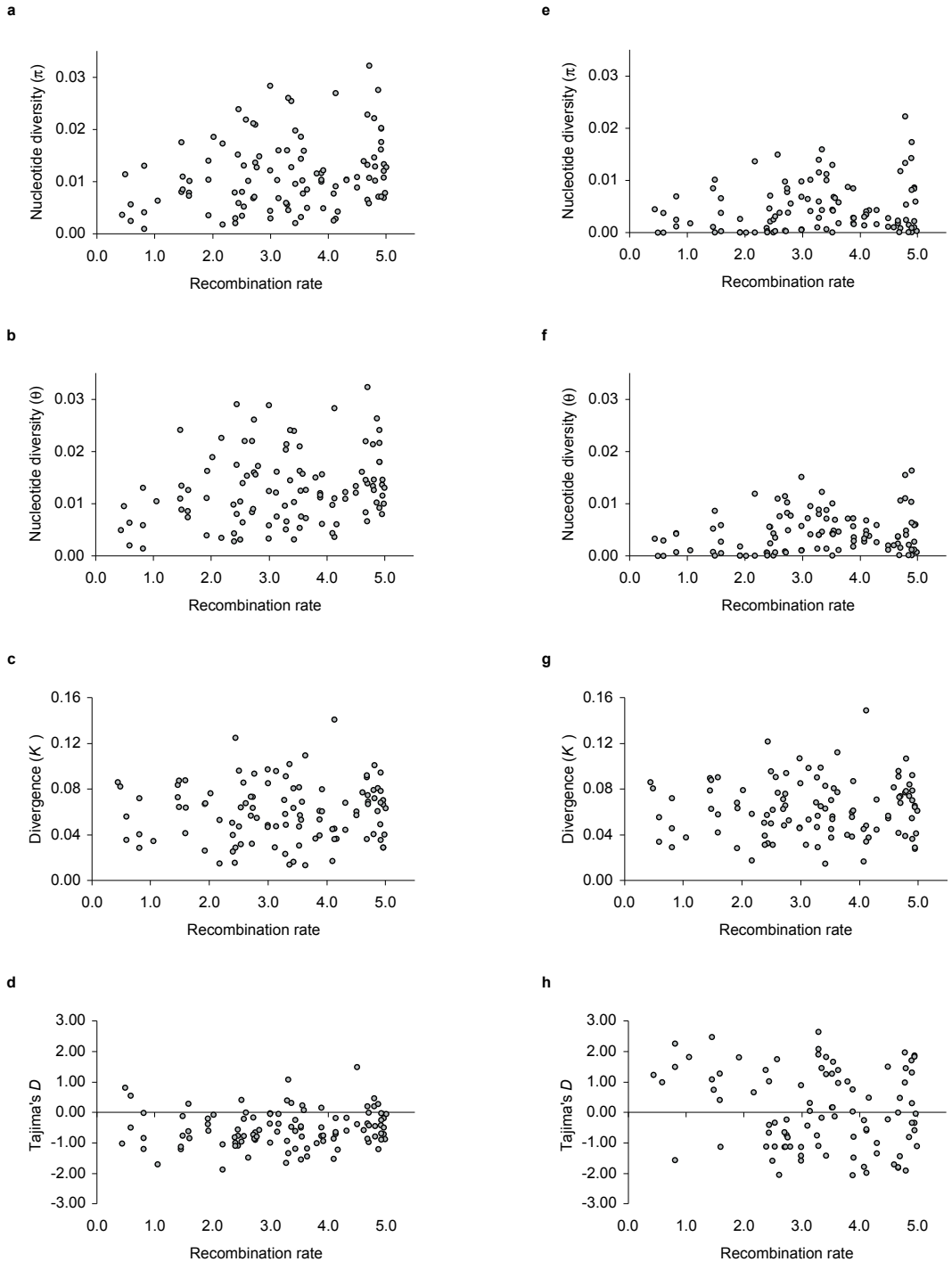


FIGURE 1.2 Nucleotide diversity π and θ_W, divergence *K*, and Tajima's *D vs*. recombination rate. (a–d) African population. (e–h) European population. Recombination rate is expressed in rec/bp/gen x 10^8 (COMERON *et al*. 1999).

chromosome-wide excess of singletons observed in the African population. It appears that this pattern of polymorphism has most likely been shaped by demography.

Is there any evidence for a signature of selection in the African population? Using two-tailed tests, we found a (weak) positive correlation between recombination rate and nucleotide variation (as measured by π and θ_W; see Figures 1.2, a and b): for π, Pearson's $R = 0.246$, $P < 0.02$, Spearman's $R = 0.237$, $P < 0.02$; for θ_W, Pearson's $R = 0.237$, $P < 0.02$, Spearman's $R = 0.234$, $P < 0.02$. If this observation were due to a lower neutral mutation rate in regions of reduced recombination, then these regions should also be less diverged. However, we found no correlation between recombination rate and levels of divergence (Pearson's $R = 0.003$, $P > 0.10$, Spearman's $R = 0.028$, $P > 0.10$; Figure 1.2c). If we consider only fragments above a certain recombination rate (for example, 2×10^{-8} rec/bp/gen, which corresponds to our previously defined region II; see MATERIALS AND METHODS), thus including 94 loci, then the correlation between recombination rate and polymorphism disappears (for π, Pearson's $R = 0.158$, $P > 0.10$; for θ_W, Pearson's $R = 0.115$, $P > 0.20$). These conclusions hold for all three measures of recombination rates (see MATERIALS AND METHODS), except that the (weak) correlation between nucleotide diversity and ACE was still found when the 11 fragments located in regions of low recombination were excluded (Pearson's $R = 0.203$, $P < 0.05$, and Pearson's $R = 0.199$, $P < 0.05$ for π and θ_W, respectively). This suggests that the strong positive correlation between recombination rates and nucleotide diversity reported in previous studies is attributable mainly to loci in low recombination regions (BEGUN and AQUADRO 1992; AQUADRO *et al*. 1994; ANDOLFATTO and PRZEWORSKI 2001).

1.3.2 Polymorphism Patterns in the European Population

A summary of the polymorphism and divergence data is shown in Figures 1.2, e–g. Of the 55,150 sites sequenced, 737 are polymorphic. The number of segregating sites and estimates of nucleotide diversity for each fragment are shown in Appendix 1.2. The means (SE) of π and θ_W across the X chromosome are 0.0046 (0.0005) and 0.0044 (0.0004), respectively.

In Figures 1.2, e and f, the estimates of π and θ_W are plotted against the recombination rate. We observed no significant correlation between nucleotide diversity and any of the three estimates of the recombination rate (see MATERIALS AND METHODS). With regard to the first of these recombination rate estimates, the results of the correlation analysis are as follows (two-tailed tests). Pearson's $R = 0.150$ and 0.180 with $P > 0.12$ and $P > 0.06$ for π and θ_W, respectively; Spearman's $R = 0.137$ and 0.183 with $P > 0.16$ and $P > 0.06$. Also, no correlation between recombination rate and divergence was observed (Figure 1.2g; Pearson's $R = 0.035$, $P > 0.73$, Spearman's $R = 0.021$, $P > 0.82$). These results contradict to some extent our findings in the

African sample, where a weak positive correlation between recombination rate and levels of variation was detected. Since this correlation has been proposed to be an effect of selection (Maynard Smith and Haigh 1974; Charlesworth 1996), it may indicate that selection in the European population is not as strong as in the African population, perhaps due to interfering demographic processes.

Tajima's (1989a) test was applied to the European sample as described in materials and methods. The observed average of Tajima's *D* (SE) across fragments is 0.045 (0.574). The average value is not significantly different from zero, but the standard error is ($P < 0.0001$). Does this mean that the European population is in equilibrium with regard to demographic and selective forces? Several lines of evidence speak against this hypothesis. Although the mean of Tajima's statistic is close to zero, for 11 fragments the data are not compatible with the neutral equilibrium model. The Tajima test (in its single-locus version; Tajima 1989a) revealed seven fragments with significantly negative *D* values and four with positive ones. Inspection of the data shows that Tajima's *D* is negative in the fragments exhibiting a rare haplotype with many singletons or strongly positive when most of the variants are organized in a few common haplotypes (Figure 1.2h). As a result of this, it appears that the mean of *D* across fragments does not differ from zero.

Using the same approach as for the African population sample, we computed the distribution of the *H*- and *K*-haplotype statistics (Depaulis and Veuille 1998) and recorded the proportion of observed values that were lower and higher than the simulated median. The observed *H* values were lower than the simulated median for 83 fragments; this proportion is higher than expected (sign test, two-tailed, $P < 0.0001$). For *K*, the trend toward fewer haplotypes was also significant (sign test, two-tailed, $P < 0.005$). In agreement with this observation, we found 13 fragments with a significantly low value of *K* or *H*, using the conservative assumption of no recombination in one-tailed *K* or *H* tests. These observations are consistent with the occurrence of bottlenecks and/or selective events in the recent past.

To further investigate whether the data deviate from the neutral equilibrium model, we used the multi-locus version of the HKA test (see materials and methods). A significant departure of the data from this model was detected ($X^2 = 238.28$, $P = 0.0016$). Figure 1.3b shows the contributions of each fragment to the summary statistic (see Appendix 1.3). Furthermore, Figure 1.3b depicts whether the observed polymorphism and divergence values are lower or higher than expected. The HKA test was repeated with the exclusion of just those fragments with the strongest

departures from expectation. The value of the overall test statistic dropped below the critical value at which the test was no longer significant, if 24 fragments with the largest contributions were removed (data not shown; 12 of these fragments show an excess of polymorphism, and 12 a deficiency of polymorphism; see Figure 1.3b).

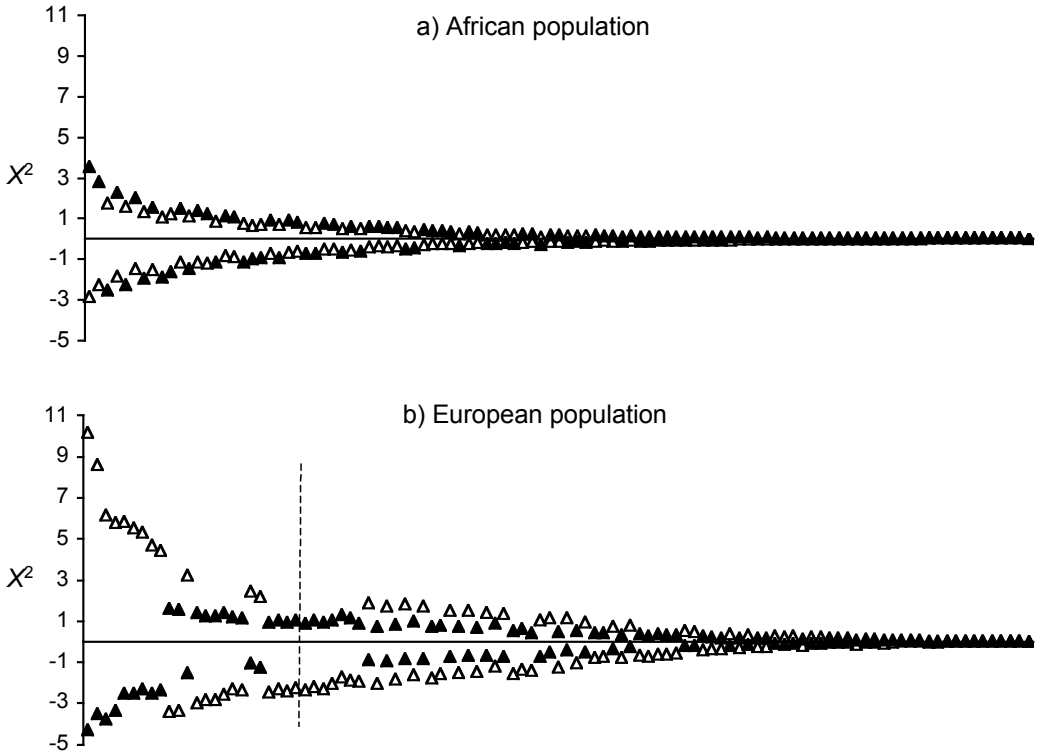


FIGURE 1.3 Contribution of each fragment to multi-locus HKA statistic. (a) African population and (b) European population. For each fragment, the contributions to the overall test statistic by the polymorphism (△) and divergence (▲) data are shown. Values above (below) the x-axis indicate a larger (smaller) contribution than expected. Fragments are ranked along the x-axis according to their total contribution to the test statistic (including polymorphism and divergence components). When the 24 fragments at the left of the vertical (dashed) line were excluded from the test (for the European sample), the value of the overall test statistic dropped below the critical value.

Note that some of these low-polymorphism fragments contribute to the overall test statistic to a very similar degree as the ones following at higher ranks; *i.e.*, between the fragments at rank 20 and at rank 30 the per-fragment contribution differs <0.5. All these fragments have values of $\theta_W \leq 0.0011$.

Next we analyze the fragments exhibiting low levels of variation. In our survey, 13 fragments had no polymorphic sites at all (Figure 1.1 and Appendix 1.2). Furthermore, 12 low-variation fragments have been identified by the HKA test, including 8 of the non-polymorphic fragments and 4 with extremely reduced nucleotide variability ($\theta_W \leq 0.0007$).

TABLE 1.1
Demographic modeling of the European population

Model	Model parameters				$\overline{F}_c$	$P(F_c \le 13)$	$P(F_c = 13)$	$P(F_c \ge 13)$	Average $\overline{D}$	$P(\overline{D} \ge 0.045)$
	T_b	N_b	T_m	N_c/N_o						
Constant	–	–	–	–	1.26	1	$<10^{-4}$	$<10^{-4}$	–0.077	0.0847
Bot 1	100,000	1,000	3,600	0.5	2.60	1	$<10^{-4}$	$<10^{-4}$	–0.967	$<10^{-4}$
Bot 2	100,000	1,000	7,500	0.25	0.60	1	$<10^{-4}$	$<10^{-4}$	–1.050	$<10^{-4}$
Bot 3	100,000	500	1,750	0.5	2.50	1	$<10^{-4}$	$<10^{-4}$	–0.955	$<10^{-4}$
Bot 4	100,000	500	4,150	0.25	0.55	1	$<10^{-4}$	$<10^{-4}$	–1.049	$<10^{-4}$
Bot 5	50,000	1,000	2,900	0.5	9.14	0.9336	0.0512[a]	0.1176	–0.672	$<10^{-4}$
Bot 6	50,000	1,000	4,400	0.25	3.13	1	$<10^{-4}$	$<10^{-4}$	–1.028	$<10^{-4}$
Bot 7	50,000	500	1,500	0.5	9.08	0.9314	0.0484[a]	0.1167	–0.712	$<10^{-4}$
Bot 8	50,000	500	2,250	0.25	2.94	1	$<10^{-4}$	$<10^{-4}$	–1.049	$<10^{-4}$
Bot 9	25,000	1,000	2,750	0.5	22.40	0.0132	0.0070	0.9938	–0.355	$<10^{-4}$
Bot 10	25,000	1,000	3,850	0.25	12.51	0.6333	0.1153[a]	0.4820	–0.790	$<10^{-4}$
Bot 11	25,000	500	1,300	0.5	20.21	0.0440	0.0210	0.9770	–0.335	0.0013
Bot 12	25,000	500	2,000	0.25	11.56	0.7407	0.1093[a]	0.3696	–0.850	$<10^{-4}$

The models are denoted as follows: Constant, constant population size without recombination; Bot 1-12, bottleneck models without recombination for 12 different sets of values of T_b, N_b, and N_c/N_0. A severe bottleneck of size N_b was introduced T_b generations ago in a population of initial size N_0 and maintained for T_m generations. After that time, the population was allowed to grow exponentially to the current population size N_c. $N_0 = 10^6$ was assumed. The value of the population mutation parameter was 0.0127, which is equal to the observed average value of θ_W for the African sample. For the constant-size simulations, the corresponding θ_W value of the European sample was used. The values of T_m were chosen such that the simulated and observed total numbers of segregating sites across all 105 fragments are in close agreement. F_c is the number of fragments with no variation; $P(F_c \le 13)$, $P(F_c = 13)$ and $P(F_c \ge 13)$ are the probabilities of obtaining at most, exactly, or at least 13 fragments with no polymorphism, respectively; Average $\overline{D}$ is the value of Tajima's *D* across all fragments averaged over all 10,000 simulation runs, and $P(\overline{D} \ge 0.045)$ is the probability of observing a value of Tajima's $\overline{D}$ across fragments equal or larger than the value observed in the European sample.

[a] Likelihood ratio test, two-tailed, $P < 0.05$ (*i.e.*, the respective bottleneck model fits better the observation of $F_c = 13$ than the Constant).

We first concentrate our analysis on the set of fragments with zero polymorphisms. We used coalescent simulations to test the hypothesis that simple demographic null models (see MATERIALS AND METHODS) can explain our observation of 13 fragments with zero polymorphisms. These are a neutral model of constant population size and various bottleneck models (Table 1.1). Since the European population is believed to be derived from Africa (DAVID and CAPY 1988; ANDOLFATTO 2001b), the pre-bottleneck effective population size, N_0, is assumed to be equal to the effective size of the Zimbabwe population (*i.e.*, ~10^6). Different values of N_c for the European population (between 0.25 and 0.5 N_c) — accounting for the fact that the observed θ_W value in the European population is about one-third of the estimate of the African population — were assumed. Severe bottlenecks were introduced mimicking the founding of the European *D. melanogaster* population. The values of the parameters (describing the time of occurrence, severity, and duration of a bottleneck) were chosen such that the current simulated population has about the same number of segregating sites as observed.

Among the models tested, a likelihood-ratio two-tailed test shows that some models fit the observation of 13 fragments with no polymorphism better than the neutral (constant population size) model [*e.g.*, bottleneck (Bot) 10, G = 14.1, P = 0.014, see Table 1.1]. Appreciable probabilities of getting at least 13 fragments with no polymorphic sites were obtained only for parameter values of the bottleneck model in which the effective population size recovered to its current size in a relatively short time period (~0.1N_c generations). Other more realistic scenarios, in which the European population was founded 10–15 kya, corresponding to > ~100,000 generations (DAVID and CAPY 1988; LACHAISE *et al.* 1988), and grew more slowly to its current effective size, appear to be inconsistent with our observation of 13 fragments with no polymorphism.

Further evidence against a simple model of population founding followed by expansion is provided by the last two columns of Table 1.1. First, the average value of Tajima's D is negative in all simulations of the bottleneck model. Second, very few simulation runs produced values of Tajima's D greater than the observed value (across fragments).

1.3.3 Comparison of the African and European Populations

The European population shows lower levels of variation than the African one (see above). These differences are statistically significant (Wilcoxon matched-pairs signed-ranks test, two-tailed, $P < 0.0001$ for both π and θ_W). As evident from the

larger difference in the means of θ_W (relative to those of π), the African population harbors more rare variants than the European one. This is also suggested by the significantly negative average value of Tajima's *D* for the African population, whereas, in the European population, average *D* is close to zero.

A large proportion (65%) of the polymorphisms in the European population are also present in the African one (comprising ~23% of the variation found in the African population). This result supports the African origin of the European population. Nonetheless, both populations are considerably differentiated: average F_{ST} (SE; Hudson *et al.* 1992) across fragments is 0.293 (0.017; see Appendix 1.3).

Because of suggestions in the literature of differential migration patterns of neutral and selected loci (Cavalli-Sforza 1966; Lewontin and Krakauer 1973), we have investigated differentiation across fragments in more detail. However, instead of using the F_{ST} approach (which was questioned by many authors, *e.g.*, Nei and Maruyama 1975 and Robertson 1975), we asked (more directly) whether derived variants that are fixed in the European sample are in high frequency in the African sample. If this were the case, the colonization history of Europe by African *D. melanogaster* may be explained by a combination of demographic processes and genetic drift, without invoking selection.

For each fragment, we recorded the frequency of the derived variants in the African and the European samples. A variant was classified as ancestral when present also in *D. simulans*; when neither of the two *D. melanogaster* variants was found in *D. simulans*, the segregating site was not considered. A total of 1974 segregating sites were classified, including shared polymorphisms, population-specific polymorphisms and fixed differences. The fragments were partitioned into two groups: (i) those with very low polymorphism [using the HKA test, this was defined in two ways (see above); independent of this definition, however, this group contained the fragments with zero polymorphisms] and (ii) the rest of the fragments. Our results of the HKA test suggest classifying a fragment as a low-variation fragment if (a) $\theta_W \leq 0.0007$ (21 fragments), or if (b) $\theta_W \leq 0.0011$ (29 fragments).

Figure 1.4 compares the relative number of segregating sites for each frequency class for the low-variation fragments defined by criterion b; criterion a gave similar results (data not shown). In this analysis, a total of 260 segregating sites with the variant fixed in the European population sample have been used (53 and 72 in the low-variation fragments for a and b, respectively). In the fragments with low variation,

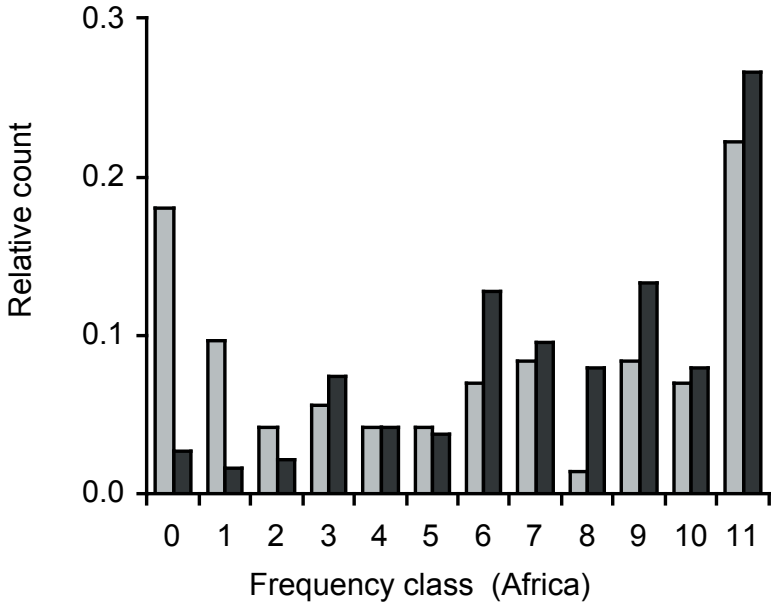


FIGURE 1.4 Relative count of derived variants that are fixed in the European sample against their frequency in the African sample. The count of variants across frequency classes is normalized to one. Shaded bars denote variants found in low-variation fragments; solid bars denote variants in the rest of the fragments. Low-variation fragments are defined by criterion b (see 1.3.3).

there is an excess of derived variants that are fixed in the European sample and rare in the African population. The difference between the low-variation fragments and the rest of the fragments is highly significant. If all 12 frequency classes are considered separately, $\chi^2 = 28.72$, d.f. = 11, $P = 0.0025$, and $\chi^2 = 32.39$, d.f. = 11, $P = 0.0007$ for a and b, respectively; if the low-frequency classes “0” and “1” are lumped together into a single category, leaving all the other classes as the second category, $\chi^2 = 25.19$, d.f. = 1, $P < 0.0001$, and $\chi^2 = 26.42$, d.f. = 1, $P < 0.0001$ for a and b, respectively.

A neutral model, in which the European variants were “sampled” from the African pool and, after colonization, reached high frequency by drift, cannot explain the observed excess of derived variants that are fixed in the low-variation fragments of the European sample and in low frequency in Africa. This observation is consistent with the result that the European population is significantly more diverged from *D. simulans* than the African population (Wilcoxon matched-pairs signed-ranks test, two-tailed, $P < 0.001$).

1.4 DISCUSSION

Our genomic scan of X-linked variation in an African and a European *D. melanogaster* population provides evidence for the impact of demography and natural selection in the recent past during which this species expanded its range. The main features of our data are discussed below.

1.4.1 Demography

Our findings that levels of polymorphism are higher in the African population and that the majority of the sites segregating in the European population are also polymorphic in the African sample confirm previous results (Begun and Aquadro 1993, 1995; Andolfatto 2001b). Furthermore, our results are consistent with the hypothesis that *D. melanogaster* originated in sub-Saharan Africa before spreading to the rest of the world (David and Capy 1988; Lachaise *et al.* 1988).

A surprising observation, however, was that the African population shows a signature of a recent population size expansion, *i.e*., a significant excess of singletons at a chromosome-wide level. The reason of this population size expansion remains unclear. Since we found only very little evidence for selective adaptations in the African population (see below), the population size increase does not appear to mirror a change of or an expansion to a new habitat.

The demographic processes that have occurred in the European population are more complex. Our observation that a large number of loci have strongly positive and negative *D* values (although the mean of Tajima's *D* across loci is close to zero) argues against the simple explanation that the European population is in equilibrium. It is more likely that several different confounding processes have occurred during the habitat expansion of *D. melanogaster*, thus producing a mean value of *D* close to zero with a significantly higher-than-expected variance. Since some fragments show a significant haplotype structure (see results and Appendix 1.2), admixture following different colonization events may have shaped the observed pattern of polymorphism (in addition to the occurrence of a bottleneck). This scenario should lead to positive *D* values. The observed mean of Tajima's *D* of ~0 may therefore be explained by counteracting demographic and selective effects (*i.e*., population size expansion following colonization and positive directional selection due to local adaptation, both producing negative *D* values).

1.4.2 Selection

The influence of demographic factors on the patterns of variation poses a problem for detecting possible footprints of selection. However, at least to some extent, this difficulty was overcome by our multi-locus approach using a large number of fragments. As discussed above, it allowed us to get insights into demographic forces that shaped the standing variation in both populations. However, since the level of polymorphism across all fragments is on average relatively high, it was also possible to search for fragments with low variation that may be footprints of recent positive directional selection (selective sweeps).

In the highly variable African population, we did not find clear evidence for positive selection. Although we employed a series of neutrality tests (including the HKA test, Depaulis and Veuille's haplotype tests, and Fay and Wu's *H* test), only one test was significant in one fragment. This observation is surprising. It may, however, not generally hold for African populations, as Mousset *et al*. (2003) found footprints of positive selection in a West African population.

Under a recurrent hitchhiking model, average Tajima's *D* value is expected to be negative due to a skew in the frequency spectrum toward an excess of rare variants (Braverman *et al.* 1995). We have observed this skew toward rare variants leading to an average negative Tajima's *D*. However, in contrast to Andolfatto and Przeworski (2001), who found a positive correlation between Tajima's *D* and recombination rates on a genome-wide scale (as expected under recurrent hitchhiking), we could not detect such a correlation on the X chromosome. The only signature of selection we observed in our sample was a (weak) correlation between recombination rate and levels of nucleotide diversity.

The data from the European population show two salient features: (i) a large number of fragments with zero or low levels of variation was identified, and (ii) a significant excess of derived variants was found at the low-variation loci (relative to the rest of the fragments) that are fixed in the European sample but rare in the African population. Both observations are difficult to explain without invoking positive natural selection. First, demographic modeling suggests that our observation of 13 fragments with zero variation is not consistent with a neutral equilibrium model or a neutral model of population founding followed by expansion. To explain our second finding, an evolutionary force needs to be postulated that brings newly arisen or rare African variants into high frequency in Europe in genomic regions of low variation (but not in the rest of the genome examined). It is difficult to imagine that any evolutionary

force other than locus-specific positive directional selection is able to simultaneously produce both features i and ii. These results are consistent with the hypothesis that the European population has experienced frequent selective sweeps in the recent past during its adaptation to new habitats.

CHAPTER 2

New Insights Into the Evolutionary History of *Drosophila melanogaster* Using an Enlarged Multi-locus Data Set

2.1 INTRODUCTION

Understanding a species evolutionary history is important for the interpretation of patterns of genetic variation observed within its populations. Substantial difference in variation has been found between African and non-African populations of the cosmopolitan species *Drosophila melanogaster* (Begun and Aquadro 1993) reflecting its out-of-Africa expansion 10–15 kya (David and Capy 1988). This range expansion into new habitats was probably accompanied by adaptive and demographic processes (*i.e*., founder events), as recently observed in various population genetic studies (Harr *et al*. 2002; Glinka *et al*. 2003; Orengo and Aguadé 2004). In contrast, populations from the ancestral range of *D. melanogaster* (*i.e*., central Africa; David and Capy 1988) should exhibit variation closer to the neutral equilibrium model due to the long evolutionary history of these populations (Andolfatto and Przeworski 2001). However, a departure from this model was found in ancestral populations (Andolfatto and Przeworski 2001; Andolfatto and Wall 2003; Glinka *et al*. 2003) suggesting that *D. melanogaster* has been faced selective and demographic processes in its ancestral range in the recent past.

In *D*. *melanogaster*, the rate of crossing over varies substantially across the genome and it has been taken as a strong predictor of levels of nucleotide variability of linked loci (Aquadro *et al*. 1994). Natural selection has been suggested to explain the observed positive correlation between the rate of crossing over and levels of diversity observed in African (Andolfatto and Przeworski 2001; Glinka *et al*. 2003) and non-African populations (Begun and Aquadro 1992; Aquadro *et al*. 1994), since this relationship was absent with the levels of divergence. Both the fixation of strongly beneficial mutations (“selective sweep”; Maynard Smith and Haigh 1974) and the selection against recurrent deleterious mutations (“background selection”; Charlesworth *et al*. 1993) reduce variation at linked loci, whereby the effect is stronger in regions of low recombination (Charlesworth 1996; Gillespie 1997). In contrast to the selective explanation, a positive correlation between recombination

rate and divergence was recently reported in humans (Hellmann *et al*. 2003). This observation emphasizes that recombination itself is mutagenic and that the correlation between crossing over rates and nucleotide diversity may therefore be of purely neutral nature — at least in humans (Hellmann *et al*. 2003).

To elucidate the relative contributions of both modes of selection in shaping the positive correlation between genetic variation and recombination, recent studies have focused on other features of the data. A hitchhiking event alters the frequency spectrum of mutations towards an excess of rare (Tajima 1989a; Fu and Li 1993; Braverman *et al*. 1995) and high-frequency derived alleles (Fay and Wu 2000), changes the distribution of haplotypes (Depaulis and Veuille 1998) and increases linkage disequilibrium (Kelly 1997; Kim and Nielsen 2004). However, demographic events, such as population size expansion or strong bottlenecks, produce also a genome-wide excess of rare variants (Tajima 1989b) thereby mimicking a false hitchhiking event.

A recently implemented multi-locus scan of X-linked non-coding DNA reported a genome-wide excess of singletons in a Zimbabwean *D. melanogaster* population (Glinka *et al*. 2003). This observation has been interpreted as a signature of a recent population size expansion, because evidence for a selective origin was missing (Glinka *et al*. 2003). Since this observation contradicts the selective findings by others (*e.g*., Andolfatto and Przeworski 2001), we extended this multi-locus scan by generating more non-coding DNA sequencing data of the same population sample to disentangle genomic patterns shaped by selective, demographic and other evolutionary processes. In addition, although X-linked inversions are rare in natural populations of *D. melanogaster* (Krimbas and Powell 1992) we screened the studied chromosomes for rearrangements, since their potential impact on levels of nucleotide variation (*i.e*., skew in the frequency spectrum) in population genetic studies might not be negligible (Andolfatto *et al*. 2001). In agreement with Glinka *et al*. (2003), we found a clear signature of a recent size expansion in the ancestral population, which was not influenced by chromosomal inversions. In addition, the large data set of our study revealed a significant correlation between the level of divergence and the rate of recombination, suggesting that recombination is mutagenic in *D. melanogaster*.

2.2 MATERIALS AND METHODS

2.2.1 Population Samples

We used 12 highly inbred lines from an African population of *D. melanogaster* (Lake Kariba, Zimbabwe, Begun and Aquadro 1993; kindly provided by C. F. Aquadro) and

one inbred line of the sister species *D. simulans* (Davis, CA, USA; kindly provided by H. A. Orr) for interspecific comparisons (see GLINKA *et al.* 2003).

2.2.2 Cytological Analyses

Several males of each *D. melanogaster* line were crossed to the same number of virgin Canton-S females, which are homozygous for the standard chromosome sequence. To maximize the detection of a heterozygote inversion in any of the studied isofemale lines, we used five F_1 third-instar larvae from these crosses (maintained at 18 °C) for the preparation of the salivary gland. To stain the chromosomes, we applied the lacto-acetic orcein method and observed the polytene chromosomes using an inverted compound microscope. The banding patterns were designated according to the standard maps of LEFEVRE (1976).

2.2.3 PCR Amplification and DNA Sequencing

According to the approach implemented by GLINKA *et al*. (2003), we designed 153 new primer pairs in non-coding DNA on the X chromosome using the available DNA sequence of the *D. melanogaster* genome (Flybase 2004, Release 3.2, http://www.flybase.org). DNA sequence data were generated and checked manually with the application Seqman of the DNAstar (Madison, WI, USA) package as described in GLINKA *et al*. (2003). For some fragments, we designed new primer pairs for a successful amplification and sequencing in *D. simulans*.

In addition, we included DNA sequence data available from the African population with the homologous sequence of *D. simulans* analyzed by GLINKA *et al*. (2003). Since the analyses in GLINKA *et al.* (2003) were based on an older version of the *D. melanogaster* genome (Release 2), we checked the location of the previously analyzed 105 fragments. The updated annotation revealed that five fragments are located in putative coding regions (*i.e*., fragment *4*, *6*, *9*, *15*, *303*). Therefore, we excluded them from the following analyses.

2.2.4 Statistical Analyses

Various summary statistics, nucleotide diversity, and tests of neutrality (including their associated probabilities) were calculated using a program kindly provided by H. Li (li@zi.biologie.uni-muenchen.de). We used 10,000 coalescent simulations to determine the statistical significance for Tajima's *D* (TAJIMA 1989a) and Fay and Wu's *H* (FAY and WU 2000) conditioned on the population mutation rate parameter, θ_F, of each fragment. We estimated θ_F by multiplying the per-site mutation rate estimator, θ_W (WATTERSON 1975), of each fragment by its length, *lth*. The LD measure Z_{nS} (KELLY

1997) and interspecific divergence were estimated by the program VariScan (Vilella *et al*., submitted). Since we had the homologous sequences of *D. simulans* we could determine the state (ancestral or derived) of an observed variant allowing us to use haploid-phased sequence data to calculate Z_{nS} for all pairs of polymorphic sites for each fragment. To determine the probabilities associated with the Z_{nS} values, we generated an empirical distribution for each fragment by a coalescent-based program (Ramos-Onsins *et al.* 2004) conditioned on θ_F and the population recombination rate, *R*. We estimated *R* by $2N_ec$, where *c* is the female recombination rate per fragment per generation, and N_e was assumed to be 10^6 (Li *et al.* 1999). The female recombination rate, *c*, was estimated by multiplying the per-site recombination rate, *r* (see below), by *lth* (see Glinka *et al.* 2003).

Departure from the neutral equilibrium model was investigated by the multi-locus HKA (Hudson *et al.* 1987) and Tajima's *D* (Tajima 1989a) tests with the program HKA kindly provided by J. Hey (http://lifesci.rutgers.edu/heylab). We used a distribution generated from 10,000 coalescent simulations for comparison with each of the test statistics (Kliman *et al*. 2000). In addition, we used the number of haplotypes, *K*, and the haplotype diversity, *H* (Depaulis and Veuille 1998), as described in Glinka *et al*. (2003). The empirical distributions of both statistics were generated using 10,000 coalescent simulations conditioned on the number of segregating sites, *S*, and *R* (see Glinka *et al*. 2003).

We estimated *r* (expressed in rec/bp/gen) for each fragment (see Glinka *et al*. 2003) by applying the method of Comeron *et al*. (1999). We compared our results to a recombination rate estimator following the procedure proposed by Charlesworth (1996). We modified this method as described in Glinka *et al*. (2003) and used the recombination rate estimated for the *white* locus as a threshold distinguishing regions of low (region I) and regions of normal to high recombination (region II). This approach differs slightly from the spatial partitioning of the X chromosome in distal- and proximal-*white* recombination regions used in Glinka *et al*. (2003).

2.2.5 Demographic Modeling of the African Population

Since evidence was found of a recent size expansion in the African *D. melanogaster* population in Glinka *et al.* (2003) we sought to extract information about their population history. We were particularly interested in the time, *t* (in years), when the population started to increase and the strength, ρ, of the expansion. To estimate these parameters, we applied a maximum-likelihood method proposed by Weiss and von Haeseler (1998), which is based on two summary statistics: the mean pair-

wise sequence difference, *K*, and the number of variable positions in a sample of DNA sequences, *S*. In their approach, the estimation of the mutation parameters (see below) is decoupled from the analysis of a specific population sample (Weiss and von Haeseler 1998). However, instead of modeling the mutation process, we estimated these parameters (base frequencies, π_A, π_C, π_G, π_T; the transition/transversion parameter, κ; the pyrimidine/purine transition parameter, ξ; see Weiss and von Haeseler 1998) directly from the sequencing data for each fragment using a program kindly provided by H. Li (li@zi.biologie.uni-muenchen.de) and assumed that each site evolves according to an exponential waiting time.

The reproduction process is based on a Wright-Fisher population at equilibrium starting to grow (or decrease) exponentially at a certain time in the past, N_0, to the current population size, N_c (Weiss and von Haeseler 1998). According to their approach, the evolution of a sample of sequences is characterized by the population mutation parameter, θ, the time when the population size started to change, τ, and the ratio of the current and initial population size, ρ (Weiss and von Haeseler 1998). The parameter θ is defined by $3N_0\mu$, where μ is the mutation rate per fragment and generation, and τ is measured in units of $1/\mu$. The most probable population history of each fragment is then defined by the set of parameters that maximizes the likelihood L (θ, τ, $\rho|k, s$; Weiss and von Haeseler 1998). The likelihood value of a given parameter set of each fragment is determined through 10,000 coalescent-based computer simulations without recombination (see Weiss and von Haeseler 1998) by a program “iphula” kindly provided by G. Weiss. To maximize L (t, $\rho|k, s$) of the data from all fragments, we assumed free recombination between fragments and maximized the likelihood over all simulated θ-values of each fragment for a certain (t, ρ). Here, t is estimated by $3N_c\tau/\rho\theta$ where we assumed a N_c of 10^6 (see above) and ten generations per year. The confidence intervals (95% CI) for these estimates are obtained by the standard MAX-2 rule (*e.g.*, Kaplan and Weir 1995).

2.3 RESULTS

2.3.1 Chromosomal Analysis

We gathered sequencing data from 253 fragments with an average distance between fragments of 67,109 bp (Figure 2.1). The length of fragments ranged from 199 to 781 bp with a mean (SE) of 510.4 (6.9) bp (see Appendix 2.1). We sequenced a total of 129,133 nucleotide sites (excluding insertions and deletions), of which 4,922 are polymorphic. Interestingly, over half of the observed polymorphic sites are at low frequency (*i.e.*, singletons). However, the observed polymorphism cannot result from loci being associated to inversions since we did not identify an inversion on

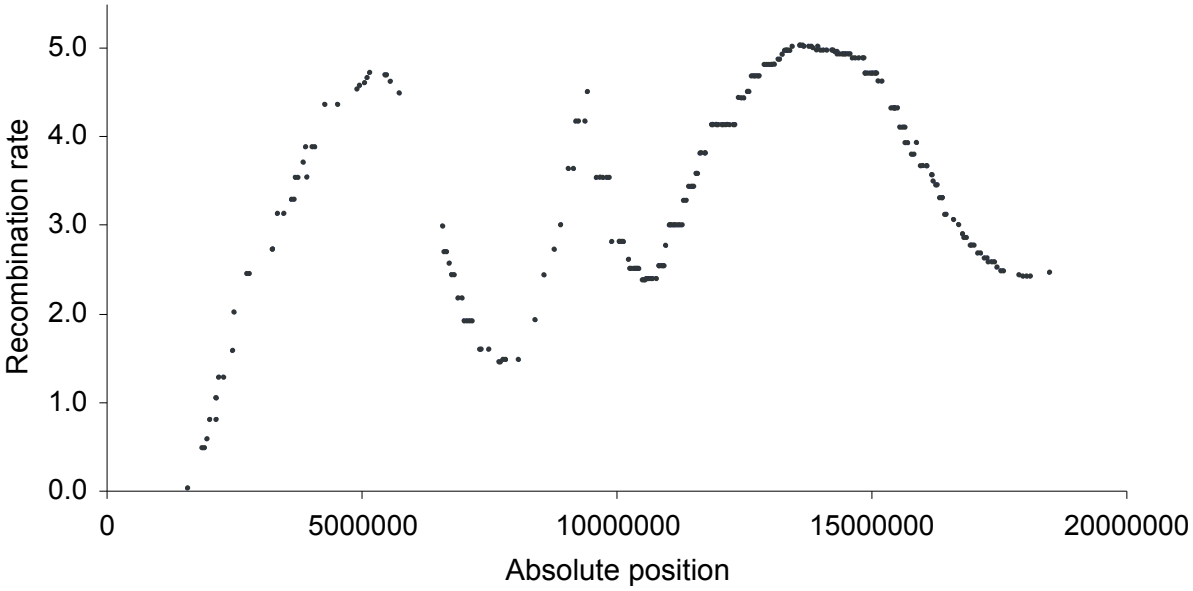


FIGURE 2.1 Location of the 253 sequenced fragments against the recombination rate (rec/bp/gen x 10^{-8}, COMERON *et al*. 1999) across the X chromosome. Fragments are shown by their absolute position (distances in Mb from the telomere).

the X chromosome in any of the analyzed lines (data not shown). Since intergenic regions and introns did not produce significantly different results we pooled them for the following analyses.

2.3.2 Diversity and Divergence

A summary of the polymorphism and divergence data of all analyzed fragments is provided in Figure 2.2, a–c, and Appendix 2.1 and 2.2. The mean levels of diversity (SE) of 253 loci were 0.0114 (0.0004) for π (TAJIMA 1983) and 0.0131 (0.0004) for θ_W (WATTERSON 1975). When we related levels of nucleotide diversity to recombination rate (Figure 2.2, a and b; after COMERON *et al*. 1999), we observed a significantly positive correlation (for π, Spearman's R = 0.140, P = 0.026; for θ_W, Spearman's R = 0.147, P = 0.020). If the observed correlation of nucleotide polymorphism and recombination rate were a function of the mutation rate, then we would expect that regions of high recombination should also be more diverged. Indeed, a weak correlation between recombination rate and divergence is present in the 232 fragments (Spearman's R = 0.127, P = 0.054; Figure 2.2c) from which we obtained the homologous *D. simulans* sequence. To investigate this pattern more closely, we divided the data set into fragments of regions of low (region I) and normal to high recombination rates (region II) according to our previously defined transition point (see MATERIALS AND METHODS). This approach results in a threshold value of 2 x 10^{-8} rec/bp/gen (Figure 2.1; COMERON *et al*. 1999) leaving 23 fragments for region I.

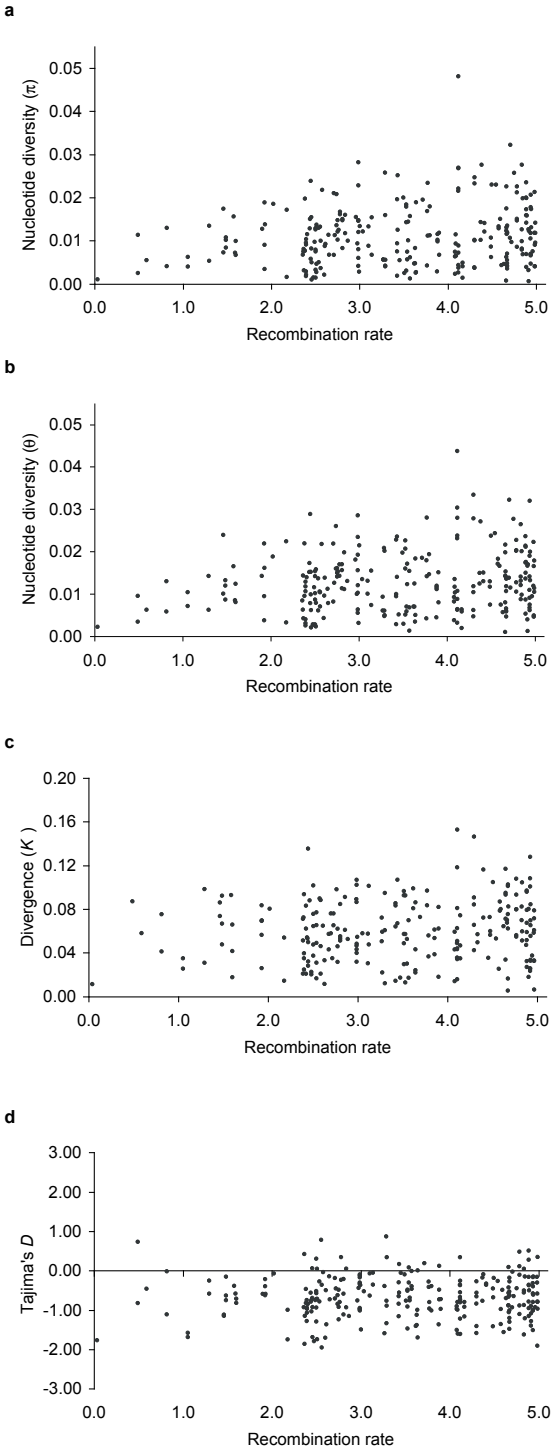


FIGURE 2.2 Levels of nucleotide diversity π and θ_W, diversity *K*, and Tajima's *D* *vs*. recombination rate (a–d, rec/bp/gen x 10^{-8}, COMERON *et al*. 1999).

The correlation between levels of nucleotide diversity and recombination rates still exists in region I (for π, Spearman's $R = 0.459$, $P = 0.024$; for θ_W, Spearman's $R = 0.510$, $P = 0.011$), but weakens in region II (for π, Spearman's $R = 0.114$, $P = 0.087$; for θ_W, Spearman's $R = 0.116$, $P = 0.080$). The opposite is seen when we relate levels of divergence with recombination rates. A significant correlation is observed in region II (Spearman's $R = 0.142$, $P = 0.040$), whereas no association was found in

region I (Spearman's $R = 0.085$, $P = 0.702$). These observations hold for the second measure of recombination rate (see MATERIALS AND METHODS), except that levels of nucleotide diversity and recombination rate are not correlated in region I (for π, Spearman's $R = 0.491$, $P = 0.125$; for θ_W, Spearman's $R = 0.527$, $P = 0.096$). This suggests that a selective effect can still be found in fragments located in regions of low recombination, although recombination is mutagenic in *D. melanogaster*.

The effects of the neutral mutation rate on intraspecific polymorphism and interspecific divergence can also be used to test if the African population departs from neutral equilibrium. Under neutrality, genome regions that evolve at high rates within a species (*i.e.*, corresponding to high π and θ_W) should also show high levels of divergence between species (HUDSON *et al.* 1987). In our analyses, no departure from the equilibrium model was detected using the multi-locus HKA test ($X^2 = 184.15$, $P = 0.990$).

2.3.3 Haplotype Structure and LD

To investigate the haplotype structure in the African sample, we used the number of haplotypes, *K*, and haplotype diversity, *H*, as done in GLINKA *et al*. (2003). Under neutrality, the proportion of the observed values of *K* and *H* lower and higher than the simulated median should be equal based on the estimated recombination rate (see MATERIALS AND METHODS; GLINKA *et al*. 2003). We observed a significantly larger proportion of fragments with higher values than expected (198 and 206 for *K* and *H*, respectively; Appendix 2.1) in both statistics (sign test, two-tailed, $P < 0.001$ and $P < 0.001$ for *K* and *H*, respectively). High values in these statistics can result from a star-like genealogy due to population expansion or an old complete sweep (DEPAULIS and VEUILLE 1998).

Since we found strong evidence that most loci deviate from a neutral genealogy in favor of a star-like genealogy, we would expect to find less LD across loci in our African population sample. We used the haploid-phased known data set of 232 fragments (see MATERIALS AND METHODS) for this analysis. To assess whether the observed Z_{nS} values are consistent with a neutral scenario, we performed neutral coalescent simulations with recombination. This assumption is conservative since recombination decreases LD measure Z_{nS} (KELLY 1997). We observed 80 of 232 loci with a significantly lower Z_{nS} value than expected under neutrality (one tailed, $P < 0.05$; Appendix 2.1).

Although we observed an overall excess of haplotypes in the studied data set, we asked whether a structuring of polymorphic sites into few haplotypes, as expected under partial selective sweeps (Depaulis and Veuille 1998), is present in some fragments. To do this, we applied the *K*- and *H*-haplotype tests (Depaulis and Veuille 1998) using a conservative assumption of zero recombination (see Glinka *et al.* 2003). Of 253 loci, we observed one significant value of the *H* statistic (one tailed, $P < 0.05$).

2.3.4 Patterns of Polymorphism and Frequency Spectrum

The observed star-like genealogy due to hitchhiking events and/or population expansion in the current data set also produces a skew in the frequency spectrum. Most fragments (225; Appendix 2.1) have a negative Tajima's *D* value (sign test, two-tailed, $P < 0.0001$; Figure 2.2d), of which 21 show a significant departure from neutrality ($P < 0.05$; Appendix 2.1). In addition, we tested whether the observed mean of the Tajima's *D* was consistent with the average estimated from the distribution of neutral coalescent simulations. None of the 10,000 simulated samples of 253 fragments had a more extreme value than the observed average Tajima's *D* (SE) of –0.6081 (0.0333). In addition, the observed variance of 0.292 is significantly smaller than expected under neutrality ($P < 0.0001$). Moreover, we did not observe a correlation between the statistic *D* and the recombination rate across the studied region ($P > 0.05$; Figure 2.2d), as would be predicted by the recurrent hitchhiking model (Braverman *et al*. 1995; Andolfatto and Przeworski 2001). This observation holds for all defined regions (see above) and both measures of recombination rates (data not shown).

A skew in the frequency spectrum towards high frequency variants, however, can result from hitchhiking with recombination. This can be investigated by the Fay and Wu's *H* test, which requires information about the state of a given variant (Fay and Wu 2000). Using the conservative assumption of zero recombination we observed five fragments with a significant Fay and Wu's *H* value (one-tailed, $P < 0.05$).

2.3.5 Demographic Modeling of the African Population

Taken together, these results suggest that population expansion is the most plausible explanation for the observed chromosome-wide excess of singletons in the African *D. melanogaster* population. To gain insight into the demographic history of the ancestral *D. melanogaster* population, we used the maximum-likelihood method proposed by Weiss and von Haeseler (1998). Since fragments in regions of low recombination are more affected by the impact of selection (see above), we

excluded them from the analysis. In addition, we excluded another 21 fragments due to undefined values of the parameter κ or ξ. For the conducted simulations, we estimated the input parameters (see MATERIALS AND METHODS) for the remaining 208 fragments (see Appendix 2.3) and used 301 different combinations of the parameters θ, τ and ρ (*i.e.*, parameter range: θ, 1–19; τ, 0–3.0; ρ, 1–1,000). An expansion model (*i.e.*, $\rho > 1$) better explained the data in all fragments (see Appendix 2.3), consistent with the observed variation of an overall negative Tajima's *D* in the studied data set. When we multiplied the maximum likelihood over all θ of a common *t* and ρ for all fragments, the maximum-likelihood estimates (95% CI) for *t* and ρ are 15,000 (0-30,000) years and 5 (1-1,000), respectively, indicating a relatively recent size expansion (see Figure 2.3).

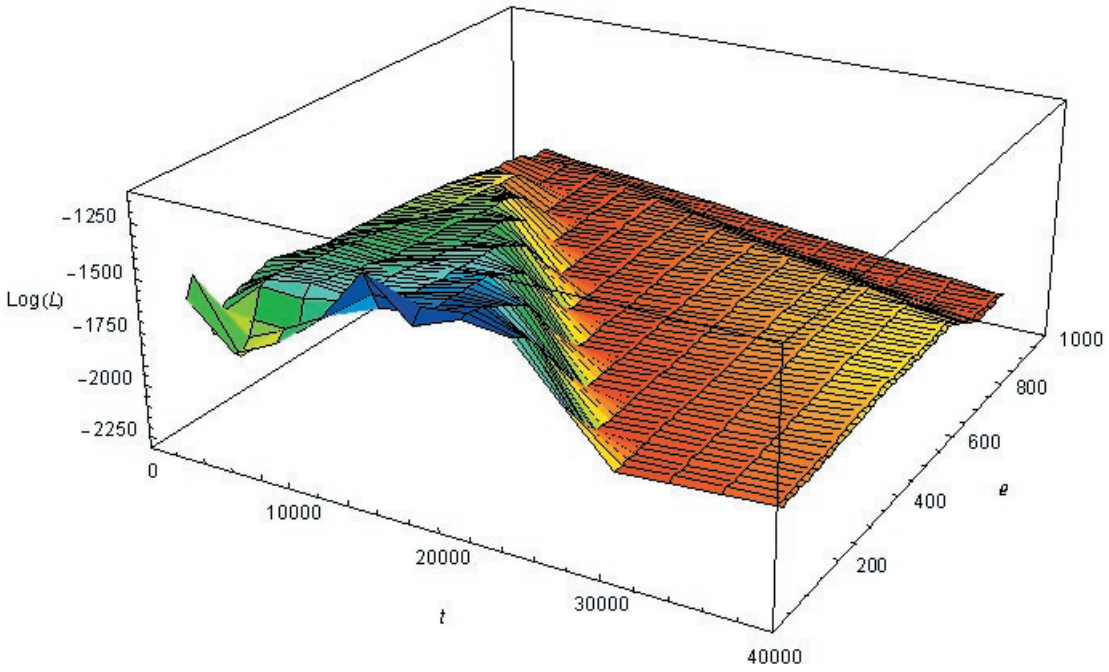


FIGURE 2.3 Log-likelihood surface plot obtained for various combinations for *t* and ρ (see text for explanations) of 208 fragments. Interpolation for missing data points was done using a triangular function (WICKHAM-JONES 1994).

2.4 DISCUSSION

The results of this multi-locus analysis of X-linked non-coding DNA are in close agreement with the findings reported by GLINKA *et al*. (2003). The only signature of selection was found by a positive correlation between recombination rate and levels of nucleotide diversity in regions of low recombination, whereas in other regions recombination itself has shaped the genomic pattern of the studied fragments. In addition, the observation of a genome-wide excess of low frequency variants

suggests that this pattern reflects an expansion of the ancestral population in the recent past.

2.4.1 Diversity and Divergence

Besides the selective explanation for the correlation between levels of nucleotide diversity and recombination rate (*i.e*., Aquadro *et al.* 1994) a neutral one has to be taken into consideration for *D. melanogaster*. As recently observed for humans (Hellmann *et al.* 2003), divergence is correlated to recombination rate in the defined region II whereas the observed association is lacking in region I. In contrast, levels of nucleotide diversity are strongly correlated to recombination rate in region I, but not in region II. These findings suggest on the one hand that recombination is mutagenic in *D. melanogaster* where the effect on the genomic pattern increases with increasing recombination rate. On the other hand, a selective explanation is still valid for the observed genomic pattern in regions of low recombination because recombination is one factor that determines the size of the region affected by selection (Kaplan *et al*. 1989; Stephan *et al*. 1992). A similar trend is seen between synonymous site divergence with *D. simulans* and recombination rates in a survey of 254 genes in *D. melanogaster* (Betancourt and Presgraves 2002) indicating the mutagenic effect of recombination.

However, neither directional nor purifying selection models can fully explain our results. The hitchhiking model (Maynard Smith and Haigh 1974) predicts an excess of rare alleles at linked neutral sites (Braverman *et al.* 1995), while the background selection model (Charlesworth *et al*. 1995) does not, as long as the population is large and the deleterious mutation rate is not extremely high (Hudson and Kaplan 1994; Charlesworth *et al*. 1995). Furthermore an expected association between the meiotic rate of crossing over and Tajima's *D* is predicted by the recurrent selective sweep model (Braverman *et al.* 1995). We did find a skew in the frequency spectrum, but did not find a correlation between Tajima's *D* and rate of crossing over. The latter observation is in strong contrast to the findings reported by Andolfatto and Przeworski (2001). In addition, assuming a large effective population size due to the long evolutionary history of this ancestral population (David and Capy 1988), background selection cannot explain the observed overall excess of low frequency variants. However, a recurrent selective sweep model also predicts a U-shaped distribution of mutations in the frequency spectrum (Ewens 1979; Kim, unpublished data). This is due to the fact that a population recovering from the last hitchhiking event should have an excess of low frequency variants, whereby the next hitchhiking event is expected to sweep the existing low frequency variation to high or lower

frequencies (Fay and Wu 2000). Using the 23 fragments located in regions of low recombination, we note a pronounced U-shape in the frequency distribution of derived alleles (Figure 2.4), suggesting that positive selection has indeed shaped this pattern. This observation further confirms the low power of the Fay and Wu's *H* test (Fay and Wu 2000) to detect old selective sweeps (Przeworski 2002), because high-frequency alleles drift to fixation shortly after the fixation of the advantageous mutation and therefore no longer contribute to polymorphism (Kim and Stephan 2002).

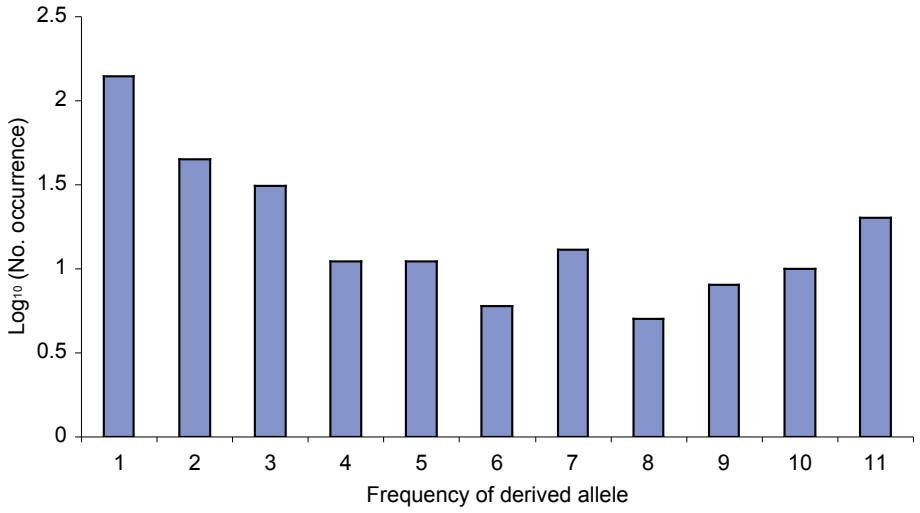


Figure 2.4 Frequency spectrum of derived alleles of fragments in regions of low recombination (region I).

2.4.2 Demographic Expansion

The observation of a chromosome-wide excess of singletons in this study is in close agreement with the results reported by Glinka *et al*. (2003). Since we found no evidence for selection by various neutrality tests applied, this chromosome-wide pattern may be due to demographic processes. Staijch and Hahn (2005) have recently shown that also admixed population samples lead to a negative Tajima's *D*. However there is strong evidence for sexual isolation between *D. melanogaster* from Zimbabwe and those from other geographic locations (Wu *et al*. 1995). In addition, evidence of admixture of the X chromosome by alleles of non-African ancestry of the rural *D. melanogaster* population in Zimbabwe is missing (Kauer *et al*. 2003). Therefore, the overall excess of singletons clearly shows that the ancestral population has recently been expanding. This size expansion cannot be due to the colonization of new habitats, since we found no sign of adaptive processes (see also Glinka *et al*. 2003). Instead the history of this species has been influenced dramatically by

climatic changes in the past 20,000 years, which has been a key determinant of the distribution of animal and plant species around the world (*e.g.*, Hewitt 2000).

During this time, the Earth's climate has undergone a transition from glacial to interglacial conditions resulting in large biotic responses including migrations of individual taxa and rearrangements of vegetation (Web III and Bartlein 1992). The last glacial maximum (LGM) in the late Pleistocene (18 to 21 kya) was dry and arid, leading to a reduction of rain forests in favor to an extension of deserts and a mosaic of savannas and open forests on the African continent (De Vivo and Carmignotto 2004). After the LGM, however, the climate changed substantially towards warmer and moisture conditions and rains returned 12 kya in the Holocene (Grove 1993). In East Africa, the Holocene climatic optimum (HCO) occurred between 12,000–10,000 to 4,000–3,000 years, in which forests began to expand (Maley 1993) and dense savanna was covering most of that region (De Vivo and Carmignotto 2004). Assuming that *D. melanogaster* was a forest-dwelling species during that time, stable forest habitats in Central Africa (De Vivo and Carmignotto 2004; Lachaise and Silvain 2004) could have served as a refuge during the LGM, from which the ancestral population expanded during the HCO. In contrast to this hypothesis, the reduction of forest during the LGM could also explain the wild-to-domestic habitat shift in *D. melanogaster*, since some of the hunter-gatherers (HG) were already sedentary (Lachaise and Silvain 2004). However, any sign of a recent expansion is missing for HG's (Excoffier and Schneider 1999). Furthermore, human populations in Africa show signals of Pleistocene expansions at around 70 kya (Excoffier and Schneider 1999). This time estimate is substantially different from the estimate of our study. Therefore, one can postulate that *D. melanogaster* expanded its range as a wild forest-dwelling species since the time when forests extended their ranges (see above), fitting well the estimated time of expansion of the ancestral population from Zimbabwe. Besides the compatibility of our sequencing data to an expansion model with various parameters, a constant size model fits equally well. However, the growth phase is unlikely to have started earlier than 30 kya. The inclusion of additional parameters (*i.e.*, patterns of segregating sites; Wakeley and Hey 1997) may result in sharper estimates and conditioning the likelihoods on the entire information in the data would reduce the variability of the estimates (Griffiths and Tavaré 1994). In addition, since recombination is known to reduce the variance of the distributions of *S* and *K* (Hudson 1990; Wall 1999), allowing recombination within each fragment might have led to sharper estimates as well.

However, the evidence of a recent size expansion of the ancestral *D. melanogaster* population fits well with the observed levels of LD in our study. Since LD is primarily governed by recombination (*i.e*., recombination erodes LD over time; *e.g*., Brown *et al*. 2004), the observed deficit in LD can be explained by more recombination in polymorphism data than expected under the equilibrium model (Przeworski and Wall 2001). Thus, an overestimate of the population recombination rate, R (*i.e*., estimated by $2N_ec$; see materials and methods), of the X chromosome (Przeworski *et al*. 2001) can result by a violation of the assumption of a constant population size (*i.e*., higher effective population size N_e) or by recombination events beside crossing-over. Gene conversion is likely to play an important role in breaking down allelic associations over short distances (Frisse *et al*. 2001) and high levels of gene conversion were reported from the fourth chromosome of *D. melanogaster*, leading to lower than expected levels of LD (Jensen *et al*. 2002). In contrast, Andolfatto and Wall (2003) reported an excess of LD in the Zimbabwean *D. melanogaster* population only if this population is close to the mutation-drift equilibrium. The discrepancy with our study may be explained by the underlying assumptions of a constant mutation rate and/or constant population size (Andolfatto and Wall 2003). In conclusion, both population size expansion and a sufficiently high rate of gene conversion may have led to the observed deficit in LD of this study.

Part II: Analysis of Candidate Sweep Regions

CHAPTER 3

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

3.1 INTRODUCTION

The level of genetic variation along a recombining chromosome can be influenced greatly by the evolutionary history of the population under study. In particular, the distinction between demography and selection has received much recent attention (*e.g.*, Glinka *et al*. 2003; Orengo and Aguadé 2004; Stajich and Hahn 2004; Storz *et al.* 2004), because both forces can lead to a reduction in diversity (Galtier *et al.* 2000). However, demographic events (*e.g.*, bottlenecks) will affect the whole genome, whereas selective events (*e.g.*, directional selection) will affect only specific loci (Andolfatto 2001a).

Genetic hitchhiking of neutral loci linked to rapidly fixed beneficial mutations ("selective sweep"; 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 weakens 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 results in an excess of low frequency variants and a star-shaped genealogy (Braverman *et al*. 1995). In the presence of recombination, hitchhiking is incomplete and 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 forms a bipartite frequency spectrum and with the knowledge of the ancestral and derived states (by using an outgroup) one can distinguish between low- and high-frequency variants (Fay and Wu 2000; Przeworski 2002). The resulting genealogy 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 LD between

polymorphisms in neutral loci, which weakens with increasing distance from the selected site (Przeworski 2002; Kim and Nielsen 2004). 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 *D. simulans* (Parsch *et al.* 2001; Meiklejohn *et al*. 2004; Schlenke and Begun 2004) and *D. melanogaster* (Depaulis *et al.* 1999; Nurminski *et al*. 2001; Harr *et al.* 2003; Mousset *et al*. 2003). Both species are human commensals and they may have extended their range from tropical Africa (south of the Sahara) to the Eurasian continent after the last glaciation 10–15 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 multi-locus scan of non-coding 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). This observation has been taken as evidence for natural selection, since all loci showed deviations from neutrality and a bottleneck model could not explain the number of loci with zero polymorphism (Glinka *et al.* 2003). One locus with zero polymorphism (fragment *125*; Glinka *et al*. 2003), is located in a region of intermediate recombination rate, with an estimated 1.926 x 10^{-8} rec/bp/gen (Comeron *et al.* 1999). This locus lies about 7 Mb away from the telomere on the X chromosome (see also Figure 1; Glinka *et al.* 2003). Because a local reduction of variation on a recombining chromosome could 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. To do this, we screened 14 loci around fragment *125*, delimiting the region of reduced variation in the European population of *D. melanogaster*. The observed valley of reduced heterozygosity comprising a region of 63.9 kb suggests a recent selective sweep. A striking peak in variation in the center of this valley accompanied by an unusual haplotype structure proposes that a non-reciprocal recombination event (“gene conversion”) was associated with the sweep. We localized three potential beneficial mutations resulting in an amino acid change in two nearby genes.

3.2 MATERIALS AND METHODS

3.2.1 Population Samples, PCR Amplification and DNA Sequencing

For the following analyses we used the 12 inbred lines of the European *D. melanogaster* population (Leiden, The Netherlands; kindly provided by A. J. Davis) and a single strain of *D. simulans* (Davis, CA, USA; kindly provided by H. A. Orr) as described in Glinka *et al*. (2003). Following their procedure, we PCR amplified and sequenced (both strands) 14 more non-coding loci proximal and distal to fragment *125* (EMBL database, http://www.ebi.ac.uk, accession numbers AJ571382–93; Glinka *et al*. 2003) on the X chromosome (see Figure 3.1) on the basis of 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 three genes (*CG1677*, *CG2059* and *unc-119*) and their 5' flanking regions (Figure 3.1). 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). Since the African *D. melanogaster* population did not show a deviation from neutrality in this region (Glinka *et al*. 2003), we gathered sequence data from the coding and the 5' regions of these genes from 12 inbred lines (Lake Kariba, Zimbabwe; Begun and Aquadro 1993; kindly provided by C. F. Aquadro) for selective comparisons. In addition, we sequenced one locus (fragment *593*) using the same African lines to clarify the origin of a gene conversion event observed in the European population. We aligned only high-quality sequences with the application Seqman of the DNAstar (Madison, WI, USA) package as described in Glinka *et al*. (2003).

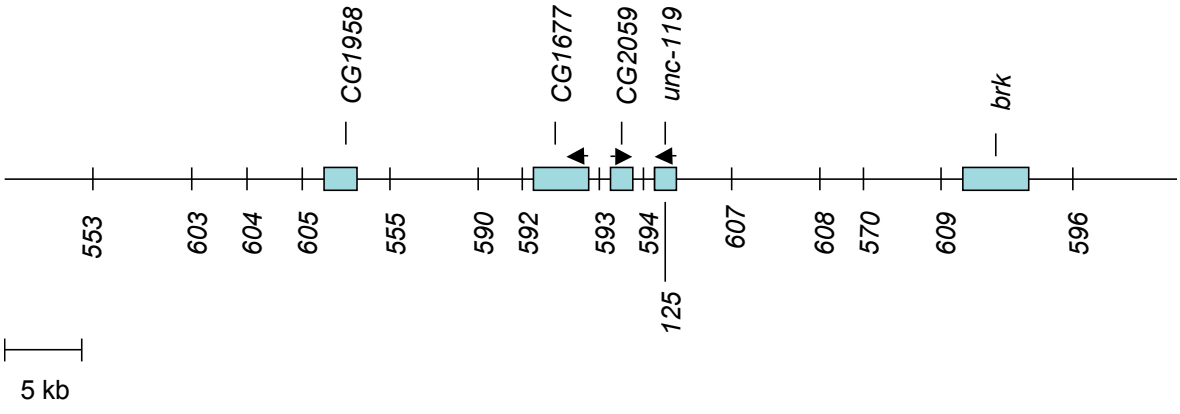


Figure 3.1 Map of the studied region around fragment 125 on the X chromosome. The arrow indicates the direction of transcription of each gene.

3.2.2 Sequence Analyses

Standard population genetic analyses and coalescent simulations to determine the probabilities of the statistical significance of Tajima's *D* (TAJIMA 1989a), Fay and Wu's *H* (FAY and WU 2000) and Fu and Li's *D* (FU and LI 1993) were performed using a program kindly provided by H. Li (li@zi.biologie.uni-muenchen.de). The homologous sequences of *D. simulans* were used to determine the derived state of a given site for Fay and Wu's *H*, Fu and Li's *D*, and to perform the multi-locus HKA test (HUDSON *et al.* 1987). The latter approach is implemented in the program HKA, which was kindly provided by J. Hey (http://lifesci.rutgers.edu/heylab). In addition, the homologous sequences of the coding regions (see above) of *D. simulans* were used to perform the MK test (MCDONALD and KREITMAN 1991) using DnaSP 3.99 (ROZAS *et al.* 2003). We estimated interspecific divergence and the LD measure Z_{nS} (KELLY 1997) for each locus using the program VariScan (VILELLA *et al*., submitted). The probability associated with LD measure Z_{nS} was calculated using DnaSP 3.99 (ROZAS *et al*. 2003).

3.2.3 Estimation of the Selective Sweep Parameters

To examine the significance of the observed local reduction of genetic variation, we applied a composite maximum-likelihood method (KIM and STEPHAN 2002). In this test, the likelihood ratio (LR) of the data under the hitchhiking model is compared to the distribution of LR obtained from 10,000 generated data sets under the neutral model (KIM and STEPHAN 2002). This test requires independent estimates of the mutational parameter, θ, and the scaled recombination rate, R_n. Since 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 (SE) of the Watterson estimator (WATTERSON 1975), θ_W, of 0.0044 (0.0004) estimated from 105 loci of the European population (GLINKA *et al*. 2003). Since this value is about 1/3 of the mean value (SE) of 0.0127 (0.0007) of the African *D. melanogaster* population (GLINKA *et al*. 2003) and given the assumed effective population size of 10^6 for *D. melanogaster* (LI *et al*. 1999), a N_e of 330,000 is assumed for the European population. Due to the absence of recombination in male *D. melanogaster* (PRZEWORSKI *et al*. 2001), R_n was estimated by $2N_er$, where N_e is 330,000 (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 (HILLIKER *et al*. 1994).

Since this approach incorporates only the spatial distribution of polymorphic sites and the frequency spectrum (KIM and STEPHAN 2002), we applied the extended

version of the described maximum-likelihood method, which uses information of LD as well (Kim and Nielsen 2004). Both methods allow us to evaluate the maximum composite-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) where N_e is 330,000 (see above).

3.2.4 Demographic Modeling of the European population

To examine if the observed pattern of nucleotide diversity could also be explained by a bottleneck we used a maximum-likelihood approach (Ometto, unpublished) implemented in a coalescent-based program (Ramos-Onsins *et al*. 2004). Following the model proposed by Galtier *et al.* (2000), a bottleneck is fully characterized by its time, T_b, and strength, S_b, and the population mutation rate, θ. Since a bottleneck affects the entire chromosome equally (Andolfatto 2001a; Galtier *et al*. 2000), we estimated the probability, P_b, of observing equal or less segregating sites given the maximum-likelihood estimates of T_b and S_b calculated from the observations of 105 loci for the European population (Ometto, unpublished) for each locus in our study. Since we performed 10,000 coalescent simulations, the probability is then given by the mean of P_b for each locus. In addition, we asked whether the total number of observed segregating sites, *S*, in our region could be explained by a bottleneck with the estimated parameters. Given these parameters, we generated 10,000 genealogies for each fragment and the probability of observing up to *S* segregating sites was calculated as the proportion of the simulated samples with less than or exactly *S*. In both cases, we used the mean θ_W of 0.0127 of the putatively ancestral African population (Glinka *et al*. 2003).

3.3 RESULTS

3.3.1 Region of Reduced Level of Nucleotide Diversity

We surveyed a total of 15 loci with an average distance between loci of 4.5 kb in the European *D. melanogaster* population (Figure 3.1 and Table 3.1). The length of the DNA fragments analyzed varied between 271 and 542 bp (excluding insertions and deletions; Table 3.1), resulting in a mean size (SE) of 365 (21) bp. Thus, the entire region from which these 15 loci derive spans 63.9 kb (Table 3.1).

The observed level of nucleotide diversity varies along the studied region (Table 3.1). The estimated π (Tajima 1983) and θ_W for both flanking loci (fragment *553* and *596*) and a central locus (fragment *593*) are on a similar scale (see Table 3.1) as the

TABLE 3.1

Summary of sequence data of each fragment of the studied region

Fragment	Position (kb)	*n*	*lth*	*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	0.2232	0.7763
603	6935	12	307	0	0.0000	0.0000	0.0777	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	1.3080
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	0.0861	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

Position is relative to the first site of the first fragment. *S* is the number of segregating sites in the European *D. melanogaster* sample with its size, *n*. *lth* represents the number of sites sequenced. *K* is divergence to *D. simulans* and levels of nucleotide diversity were estimated using π (TAJIMA 1983) and θ_W (WATTERSON 1975). Z_{nS} (KELLY 1997) is linkage disequilibrium, and T's *D*, F & W's *H* and F & Li's *D* are Tajima's *D* (TAJIMA 1989a), Fay and Wu's *H* (FAY and WU 2000) and Fu and Li's *D* (FU and LI 1993), respectively. * indicates significance at 0.05 level and n.a. is not applicable.

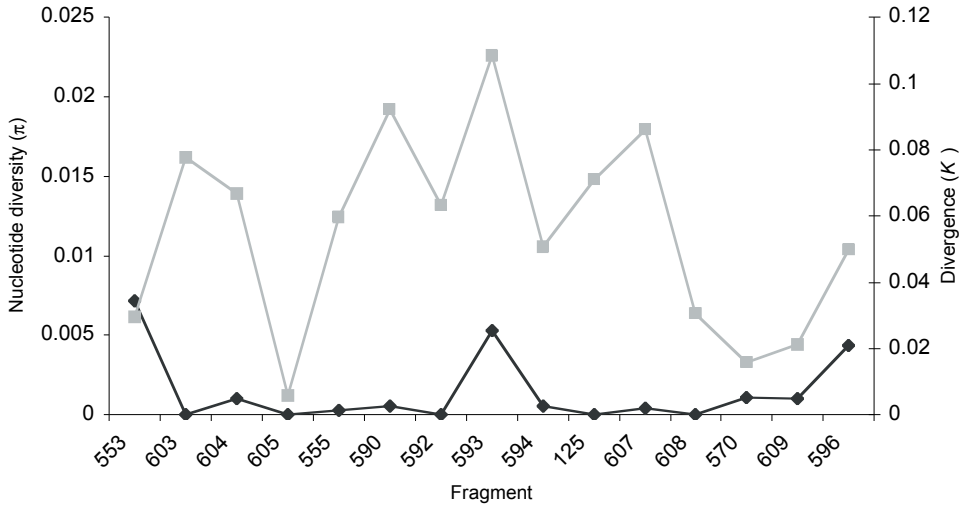


FIGURE 3.2 Nucleotide diversity (black) and divergence (gray) against the relative position of each fragment.

reported mean (SE) values of 0.0046 (0.0005) and 0.0044 (0.0004), respectively, for the European population (GLINKA *et al*. 2003). In the remaining 12 loci, however, we observed either very low or zero polymorphism (Table 3.1) resulting in a valley of reduced level of variation (Figure 3.2).

Taking the interspecific divergence between *D. melanogaster* and *D. simulans* into consideration, the observed low variation in two loci, fragment *605* and *570*, could be explained by a low mutation rate (Figure 3.2). The opposite might be true for fragment *593* (Figure 3.2). A higher mutation rate could have led to the observed high peak in nucleotide diversity in this fragment. However, the observed level of polymorphism results from a distinct haplotype structure due to six sites segregating in three haplotypes present at different frequencies (nine, two and one; Figure 3.3).

Fragment	*553*							*604*	*555*	*590*	*593*							*594*	*607*	*570*	*609*	*596*			
Line	Relative position																								
								1	1	2	3	3	3	3	3	3		3	4	5	5	6	6	6	6
								0	9	5	3	4	4	4	4	4		6	1	0	5	3	3	3	3
		1	1	2	2	2	3	7	9	6	9	0	0	0	0	0		5	9	8	3	5	6	6	6
	8	0	9	0	1	8	6	5	6	3	9	2	3	3	4	6		2	5	1	8	8	2	7	9
	1	9	9	0	2	1	7	3	0	7	0	9	1	6	6	5		2	6	7	2	9	4	0	4
D. simulans	C	T	G	C	A	A	T	C	A	T	A	G	A	T	-	G		G	G	A	T	G	A	G	G
D. melanogaster 01	A	C	.	.	.	C	A	.	.	.	G	.	.	C	.	.		.	.	.	.	.	.	A	T
D. melanogaster 02	A	C	.	T	.	C	.	.	.	.	G	.	.	C	.	.		.	.	.	.	.	.	A	T
D. melanogaster 11	.	.	A	.	G	.	.	T	.	.	G	.	.	C	.	.		.	.	T	.	.	.	.	.
D. melanogaster 12	.	.	.	.	.	C	.	.	.	C	.	.	.	.	T	A		.	.	.	A	.	C	.	.
D. melanogaster 13	A	C	.	T	.	C	.	.	.	.	G	.	.	C	.	.		T	A	.	.	.	.	A	T
D. melanogaster 14	A	C	.	.	.	C	.	.	.	.	G	.	.	C	.	.		.	.	.	.	.	.	.	.
D. melanogaster 15	A	C	.	T	.	C	.	.	.	.	G	.	.	C	.	.		.	.	T	.	T	.	A	.
D. melanogaster 16	.	.	.	.	.	C	.	.	.	.	.	T	C	.	T	A		.	.	.	.	.	.	A	T
D. melanogaster 17	.	.	.	T	.	C	.	.	T	.	G	.	.	C	.	.		.	.	T	A	.	.	A	T
D. melanogaster 18	.	.	.	.	.	C	.	.	.	.	.	T	C	.	T	A		.	.	.	.	.	.	A	T
D. melanogaster 19	A	C	.	T	.	C	.	.	.	.	G	.	.	C	.	.		.	.	T	.	T	.	A	.
D. melanogaster 20	.	.	A	.	G	.	.	T	.	.	G	.	.	C	.	.		.	.	T	.	.	.	.	.

FIGURE 3.3 Alignment of polymorphic sites observed in 12 lines of the European *D. melanogaster* population for each fragment. The relative position (see Table 3.1) and the derived state inferred from *D. melanogaster*/*D. simulans* comparisons is given for each polymorphic site. At the site for which the derived state could not be determined due to an insertion/deletion (–, 1bp) difference between species, the base with the higher frequency in the African population (*i.e.*, fragment *593*; data not shown) was assumed to be ancestral.

Since this peak in polymorphism deviates from the predicted valley of heterozygosity (KIM and STEPHAN 2002), we investigated the historical relationship of the observed haplotypes by analyzing DNA sequence polymorphism in a putatively ancestral *D. melanogaster* population from Africa (see MATERIALS AND METHODS). All sites segregating in the European population are also segregating in the African population, whereas the most frequent haplotype observed in the European is at low frequency in the African population (data not shown). Since the observed pattern in this locus is not present in the neighboring loci in the European population, the most parsimonious explanation of the observed haplotype structure is a gene conversion event (see DISCUSSION).

3.3.2 Departure from Standard Neutral Model

To evaluate the significance of the observed reduction in nucleotide diversity, we used the multi-locus HKA test (Kliman *et al*. 2000). This test takes the observed differences in mutation rates in our data set into account by comparing intraspecific diversity and interspecific divergence (Hudson *et al.* 1987). We detected a significant deviation from the neutral expectation for the studied region in the European population (X^2 = 38.112, P = 0.0005). Further evidence for a departure from neutrality can be gained by examining the frequency spectrum. A skew towards low frequency variation can be measured by Tajima's *D* statistic (Tajima 1989a). Under neutrality, Tajima's *D* is expected to be zero. Out of ten fragments with some variation, we observed four fragments with *D* values significantly (P < 0.05) less than zero, indicating an excess of singletons (Table 3.1). If this skew in the frequency spectrum is due to new mutations, as expected under a hitchhiking model (see Braverman *et al*. 1995) 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 to 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 (Table 3.1). Support for a hitchhiking event in the European *D. melanogaster* population also can be gained from Fay and Wu's *H* statistic (Fay and Wu 2000). This statistic measures the skew towards high frequency derived variants. However, we observed no deviation from neutrality in the *H* statistic in any of the fragments (Table 3.1). Given the strong haplotype structure in fragment *593*, we would expect to find linkage disequilibrium among the alleles as well. Using a conservative assumption of no recombination, the Z_{nS} value of 0.7600 is significantly higher than expected under neutrality (P = 0.049; Table 1). In contrast, the Z_{nS} values of the two terminal loci (fragment *553* and *596;* see Table 3.1) are not significant (P > 0.05). This result, together with the observations of nucleotide diversity and the results of other neutrality tests (see Table 3.1), suggests that our survey covered the entire region subject to a selective sweep.

3.3.3 Estimation of Selective Sweep Parameters

The observed valley of variation, the skew in the frequency spectrum, and the observed LD in fragment *593* provide strong evidence for the recent occurrence of a selective sweep. Since we have independent estimates of the effective population size, the mutational parameter θ, and the recombination rate (see materials and methods), we can implement 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. Given the estimates of parameters used for the simulations, our data fit significantly better to a hitchhiking than to a neutral model using the composite maximum-likelihood test proposed by KIM and STEPHAN (2002; $P < 0.0001$). Furthermore, the strength of selection, s, is 0.0038 and the estimated position of the selected site, x, is 22,625. The test proposed by KIM and NIELSEN (2004), which includes information about LD, however, did not reject neutrality in favor of a hitchhiking model ($P = 0.3440$). This can be explained by the one-sided LD structure in our region, which is different to the one outlined in KIM and NIELSEN (2004; see DISCUSSION).

3.3.4 Demographic Modeling of the European Population

Since *D. melanogaster* colonized Europe 10–15 kya (DAVID and CAPY 1988), the reduced variation in the studied region could also be the result of a population bottleneck. Given the observed number of segregating sites and the corresponding θ_W value of each of the 105 loci of the European and African *D. melanogaster* population (GLINKA *et al*. 2003), the maximum likelihood estimates for T_b and S_b are 0.0125 and 0.3755, respectively (OMETTO, unpublished results). These estimates allow us to investigate if the observed number of segregating sites for each locus can be explained by a bottleneck. None of the 15 loci showed a significant departure from the bottleneck model when tested individually ($P > 0.05$). However, considering that a selective sweep affects the entire studied region, we can compare the expected number of segregating sites across fragments with the observed total number S. The total number of segregating sites observed in our region is significantly smaller than expected under a bottleneck model ($P < 0.0001$).

3.3.5 Localization of Potential Beneficial Mutation

The predicted site of the beneficial mutation is located between gene *CG1958* and a cluster of three genes, *CG1677*, *CG2059* and *unc-119*, which are located –14.7 kb and 6.4, 11.9 and 14.2 kb away from the predicted site (Figure 3.1). Assuming that the potential target site of selection is likely to be found in a regulatory or coding region, we concentrated our efforts on these neighboring genes. However, since we observed a low mutation rate in a nearby fragment (*605*) of gene *CG1958* (Figure 3.2), we focused our investigation on this gene cluster. Here, we observed a significant deviation from neutrality by fragments surrounding the gene cluster (*555* to *607*; Table 3.1). We sequenced the 5' flanking and the coding regions of all three genes in the European and African *D. melanogaster* population and in the *D.*

TABLE 3.2

Observations of each gene for McDonald and Kreitman test (McDonald and Kreitman 1991)

	CG1677		*CG2059*		*unc-119*	
	Fixed	Polymorphic	Fixed	Polymorphic	Fixed	Polymorphic
Replacement	71	2	12	2	7	1
Synonymous	118	3	37	0	14	0

simulans strain. In the 5' region of the genes *CG1677* and *unc-119* (*i.e.*, 514 and 401 bp in length, respectively), we found neither length differences nor substantial sequence divergence between the European and the African population (Appendix 3.1 and 3.3). However, in the 5' region of gene *CG2059* (*i.e.*, 504 bp in length) we observed a similar haplotype structure as found in fragment *593* in the European *D. melanogaster* population, which extends its pattern until the relative position of 34,166 indicating one end of the gene conversion event (Appendix 3.2). In addition, three sites are fixed in the European but in low frequency in the African population (Appendix 3.2). In contrast, a comparison of the European population with the *D. simulans* strain revealed a higher number of fixed substitutions than polymorphisms within the European population in all three coding regions (Table 3.2). Under neutrality, the ratio of replacement to synonymous fixed differences between species is expected to be the same as the ratio of replacement to synonymous polymorphisms within species (*i.e.*, MK test; McDonald and Kreitman 1991). Although the previous analyses provided strong support for directional selection, this test of neutral protein evolution shows no deviation from expectations for all three genes (Fisher's exact test, $P > 0.05$), possibly due to the low level of polymorphism (Table 3.2). However, visual inspections of the sequences revealed one fixed replacement site in a derived state in *CG1677* and two in *CG2059* and one fixed replacement site in the ancestral state in *CG1677* and *unc-119* in the European population, and these substitutions are in low frequency in the African population (Appendix 3.1–3).

3.4 DISCUSSION

Our study provides strong evidence that a beneficial mutation, which arose in a very short time scale, recently went to fixation in a European *D. melanogaster* population. This process, known as a selective sweep (Maynard-Smith and Haigh 1974), has removed variation at linked neutral sites over a region comprising 63.9 kb. Furthermore, one locus linked to the potential selected site showed significant LD indicating a gene conversion event during the selective sweep phase.

3.4.1 Evidence for Selective Sweep

Our results support that the observed reduction in nucleotide diversity was caused by a recent selective sweep. Although the observed polymorphism in each fragment can be explained by a bottleneck, the polymorphism data of the entire region fits significantly better to the selective sweep model. The estimated selection coefficient in our study is of similar magnitude as reported by others (Harr *et al.* 2002; Schlenke and Begun 2004), and given the local recombination rate, the expected size of the sweep of 71.0 kb (Stephan *et al.* 1992) is similar to the observed size of 63.9 kb. The candidate for the selected mutation may be found in the 5' region of gene *CG2059* and in the *CG1677* and the *CG2059* genes. Although we observed three fixed substitutions in the 5' region of gene *CG2059*, it is unlikely that these sites are within the same *cis*-regulatory element (CRE), because their distance from each other exceeds the value of 14 bases estimated for the mean conservation length of such elements (Richards *et al*. 2005). However, it is possible that each variant occurs in a separate CRE or that one or more of these variants is not involved in *cis* regulation, but is linked to selected variants. Therefore, we propose that the candidates for the selective target are the replacement substitutions occurring in the *CG1677* and the *CG2059* genes about 6.4 kb and 11.9 kb away from the predicted sweep center. Since these mutations are fixed in the European but are in low frequency in the African *D. melanogaster* population, we postulate that they became favored when *D. melanogaster* colonized Europe 10–15 kya. The lack of variation or the presence of derived variants in low frequency in most of the loci studied argues for the short time frame described above. In addition, this explains the low level of polymorphism in each analyzed coding region.

The observation of low frequency derived variants is consistent with a complete selective sweep (*e.g*., Kim and Stephan 2002). Evidence of a recent selective sweep has been reported from other regions in *Drosophila* by various studies. Schlenke and Begun (2004) identified a transposable element insertion as the beneficial mutation on the chromosome 2R in a Californian population of *D. simulans*. Nurminsky *et al*. (2001) found strong support, that the newly-formed *Sdic* gene (Nurminsky *et al*. 1998) on the X chromosome has undergone one or more recent selective sweeps in *D. melanogaster*. Harr *et al.* (2002) observed three sweep regions in non-African *D. melanogaster* populations of which one potential location of a selected site was mapped to the *syx4* gene on the X chromosome. Meiklejohn *et al*. (2004) localized the target of selection to a 1.5 kb region surrounding *janusB*, a previously identified hitchhiking region (Parsch *et al.* 2001), in which the selected allele has not gone to

fixation yet. However, most of these studies could not identify the specific site of the beneficial mutation.

3.4.2 Gene Conversion Associated with Selective Sweep

We propose that a gene conversion event associated with the selective sweep is responsible for the strong haplotype structure observed in fragment *593* and in the 5' region of gene *CG2059*. Given the observed valley of nucleotide diversity, the following hypothetical scenario can explain the observed haplotype pattern: consider neutral loci linked to a selected site going to fixation. Suppose a lineage associated with the unfavored allele recombines non-reciprocally by donating its genetic information to a lineage associated with the favored allele in the sweep phase. The result of this gene conversion event is the observation of two distinct haplotypes in the population and its frequency in the population depend on the time of the gene conversion during the sweep phase. Meiklejohn *et al*. (2004) observed a potential gene conversion tract in which a stretch of ancestral variants were 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 on nucleotide diversity has been reported from a natural population of *D. melanogaster* due to a breakpoint 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 on 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, personal communication). 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 two 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 given that the fixation of the beneficial mutation occurred very quickly, then one would expect to find high LD on both sides of the beneficial mutation due to the mutations on the long inner branches (after Figure 7; Kim and Nielsen 2004).

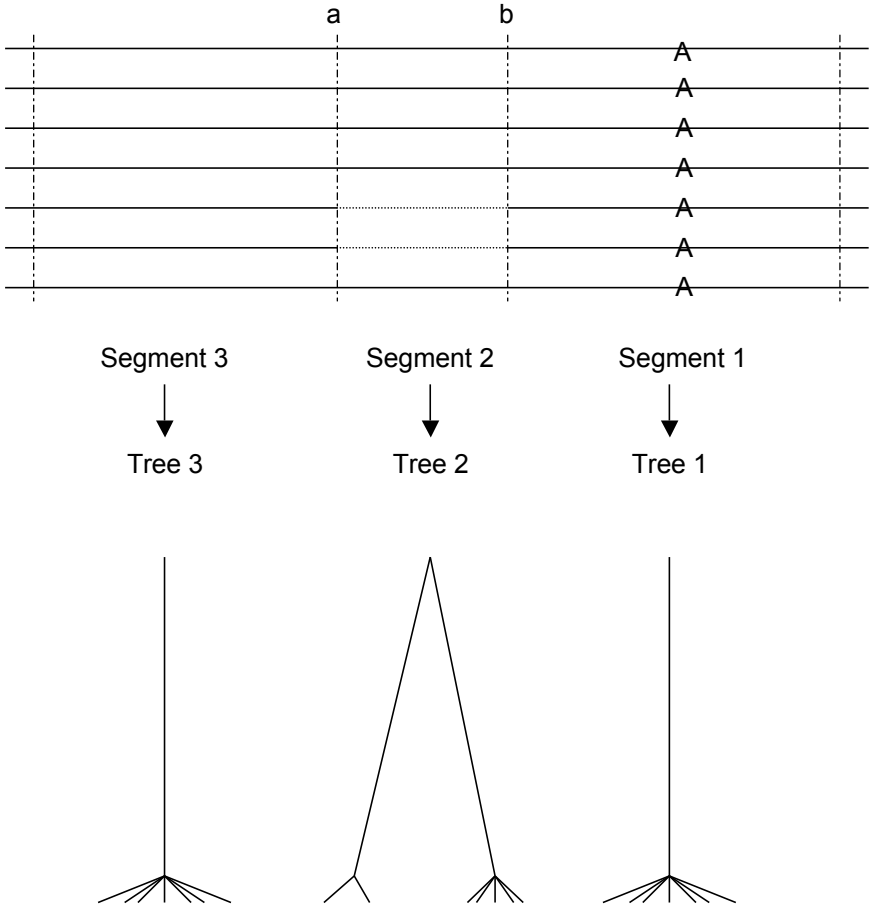


FIGURE 3.4 An example of DNA sequences (horizontal lines) and the genealogical structure resulting from a recent selective sweep with gene conversion (after Figure. 7; KIM and NIELSEN 2004). Solid lines represent sequences originally linked to the beneficial mutation A. Dashed lines represent "recombinant" sequences originally linked to the unfavored allele a, but recombined via gene conversion with A during the selective phase. Breakpoints of gene conversion are labeled as a and b. Segments between breakpoints are defined as segment 1, 2 and 3 and the coalescent tree is given below for each segment.

However, LD is expected to decrease quickly due to the increase of recombination breakpoints 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 (Figure 3.4), 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.

CHAPTER 4

The Detection of Recent Positive Selection in Ancestral *Drosophila melanogaster* from Haplotype Structure

4.1 INTRODUCTION

The rapid fixation of a beneficial mutation typically sweeps neutral variation around the selected allele (Maynard Smith and Haigh 1974). This process alters the frequency spectrum of mutations (Tajima 1989a; Fu and Li 1993; Braverman *et al*. 1995; Fay and Wu 2000), changes the spatial distribution of polymorphic sites (Kim and Stephan 2002), and increases linkage disequilibrium (Kelly 1997; Kim and Nielsen 2004), thus creating a haplotype structure by shifting certain haplotypes to high frequencies (Hudson *et al*. 1994; Depaulis and Veuille 1998; Andolfatto *et al*. 1999).

These effects are strongest on neutral alleles closest to the target of selection and weaken with increasing distance due to recombination (Kim and Stephan 2002). Therefore, a homogeneous haplotype of very tightly linked neutral alleles is found close to the fixed selected site, whereas two or more haplotypes may be found with increasing distance from the site under selection, which have recombined onto the advantageous chromosome and thus escaped extinction (Fay and Wu 2000; Kim and Nielsen 2004). This results in an increase in number of alleles that escaped complete hitchhiking with increasing distance from the selected site and thus leading not only to a valley of reduced genetic variation (Kim and Stephan 2002), but also to a decay in haplotype structure (Fay and Wu 2000; Przeworski 2002; Sabeti *et al*. 2002; Kim and Nielsen 2004).

Since it has been shown that the effect on linkage disequilibrium extends to a wider area than that determined by the lack of polymorphism, neutrality tests based on haplotypes (Hudson *et al*. 1994; Depaulis *et al*. 2001) retain a higher power to detect a departure from neutrality (Depaulis *et al*. 2003, 2005; Mousset *et al*. 2004) than those focusing on the marginal allele frequencies (Fay and Wu 2000). This advantage was recently implemented in a maximum-likelihood approach to detect positive selection in multi-locus haplotype data, based on a signature of haplotype structure (Mousset *et al*., submitted).

An analysis of polymorphism of non-coding sequencing data of a multi-locus scan showed six adjacent loci with a low number of haplotypes and a trend for low haplotype diversity in a ~500 kb region within the 11D1 chromosomal region in a putatively ancestral *Drosophila melanogaster* population from Africa (Lake Kariba, Zimbabwe; Glinka *et al*. 2003). Low values of these statistics are expected either from demographic (*e.g*., population substructure and/or bottlenecks) or from selective events (*e.g*., partial hitchhiking; Depaulis and Veuille 1998). Although neither the observed low values, nor this clustering was significant, we further investigated if this pattern could be due to a hitchhiking event since it was found only locally and not over the entire X chromosome. We screened 10 loci within the observed cluster for a distinct haplotype structure comprising a region of 56.7 kb. Although we did not find a homogeneous haplotype, an observed decay in haplotype structure suggests that a recent selective sweep has shaped this pattern. The target site of positive selection has been mapped to the 5' flanking region of the gene *CG4661*.

4.2 MATERIALS AND METHODS

4.2.1 Population Samples, PCR Amplification and DNA Sequencing

We PCR amplified and sequenced (both strands) nine more X chromosomal non-coding loci between 12.60 and 12.66 Mb on the basis of the available DNA sequence of *D. melanogaster* genome (Flybase 2004, Release 3.2.0, http://www.flybase.org) in 11 African inbred lines (Lake Kariba, Zimbabwe; Begun and Aquadro 1993; kindly provided by C. F. Aquadro) following the procedure as described in Glinka *et al*. (2003). These loci are located within the clustering of the six loci investigated previously (see Figure 4.1). We used only high-qualtiy DNA sequence data, which were aligned and checked manually with the application Seqman of the DNAstar (Madison, WI, USA) package as described in Glinka *et al*. (2003). For the following analyses, we included sequences of the same isofemale lines of one locus of the previously analyzed cluster (see above) of the African population (fragment *250*; EMBL database, http://www.ebi.ac.uk, accession numbers AJ569935-38, 40, 42-47; Glinka *et al*. 2003).

4.2.2 Sequence Data Analyses

To investigate a potential decay in the frequency of haplotypes due to positive directional selection, we applied a maximum-likelihood approach (Mousset *et al*., submitted), which uses multi-locus haplotype data to estimate the selection parameters τ (*i.e*., time since the fixation of the selected mutation), α (*i.e*., the strength of selection), and *x* (*i.e*., the location of the selected locus). Based on coalescent simulations with recombination of linked loci, this test computes full

likelihoods conditioned on the number of segregating sites and the frequency of the major haplotype (*i.e*., the size of the largest subset that shows no differences in a sample of sequences). The likelihoods of two evolutionary models are assessed: a Wright-Fisher neutral model with constant effective population size, N_e, and a positive selection model with three parameters (see above), where the frequency of the selected mutation evolves deterministically. These likelihoods are then compared using a standard likelihood ratio test (*e.g*., Sokal and Rohlf 2001, p. 689).

We calculated the Watterson estimator, θ_W (Watterson 1975), of the mutational parameter, θ, and the input parameters for this maximum-likelihood method (*i.e*., for each locus: length and number of segregating sites) using the program DnaSp 3.99 (Rozas *et al*. 2003). For every locus, the frequency of the major haplotype was visually assessed and the coalescent simulations to determine the probability associated with this frequency (Hudson *et al*. 1994) were performed with the "allelix" software (Depaulis *et al*. 2001) with a conservative assumption of no recombination. The recombination parameter, *R*, for the 11D1 chromosomal region was estimated by $2N_e r$, where N_e for the African *D. melanogaster* was assumed to be 10^6 (Li *et al*. 1999) and the per-site-recombination rate, *r*, was estimated to be about 4.5×10^{-8} rec/bp/gen using the method of Comeron *et al*. (1999). The selection coefficient, *s*,

TABLE 4.1

Summary of sequence data of each fragment of the studied region

Fragment	Position (kb)	*n*	*lth*	*S*	θ_W	H_{obs}	*P*
546	0	11	286	9	0.0107	5	0.420
581	5159	11	419	17	0.0136	2	0.998
582	9055	11	546	26	0.0163	1	1.000
583	15840	11	394	4	0.0035	4	0.967
585	24433	11	501	25	0.0170	4	0.260
586	30883	11	528	35	0.0226	4	0.169
250	32123	11	593	24	0.0144	5	0.103
588	43738	11	349	25	0.0245	2	0.987
589	50377	11	473	19	0.0137	2	0.996
576	56745	11	299	7	0.0080	5	0.518

Position is relative to the first site of the first fragment. *S* is the number of segregating sites in the African *D. melanogaster* sample with its size, *n*. *lth* represents the number of sites sequenced. Levels of nucleotide diversity were estimated using θ_W (Watterson 1975). H_{obs} is the observed frequency of the major haplotype and *P* is the probability of the one-tailed single-locus haplotype test (Hudson *et al*. 1994) using 10,000 coalescent simulations and a conservative assumption of no recombination.

of the beneficial mutation is estimated by $\alpha/3N_e$ (e.g., KAPLAN *et al*. 1989; BRAVERMAN *et al*. 1995).

4.3 RESULTS

We screened 10 loci in total over a 56.7 kb region with a mean distance between loci of 6.31 kb in the African *D. melanogaster* population (Figure 4.1 and Table 4.1). The size of the loci ranged between 286 and 593 bp (excluding insertions and deletions; Table 4.1), with an average size of 439 bp. Of the 4,388 sites sequenced, 191 are polymorphic resulting in a range of segregating sites between four and 35 across loci (Table 4.1).

The size of the major haplotype class varies between one and five per fragment (Table 4.1). Although part of the variation in the size of the major haplotype class can be explained by varying mutation rate between fragments, the consistent higher number between fragment *583* and *250* (Table 4.1) can only be explained by a non-neutral distribution of segregating sites over the existing haplotypes. In other words, although the estimated θ_W increases from fragment *583* to *250* (see Table 4.1) the

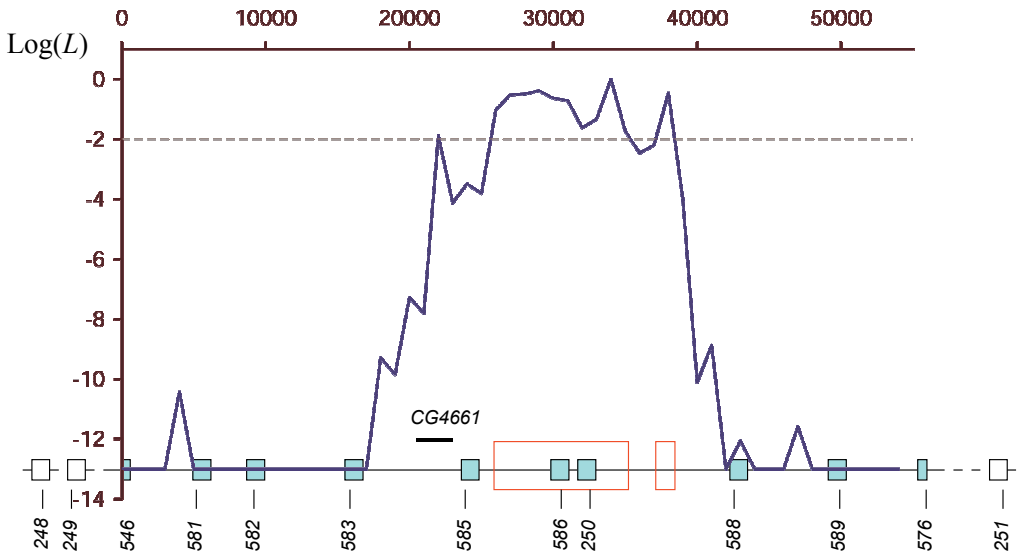


FIGURE 4.1 Log-likelihood plot of the selection model given the data set of the studied region (see Table 4.1) located within previously investigated loci (*i.e*., fragment *248*, *249*, *251*; GLINKA *et al*. 2003). Log-likelihoods of each data point along the studied region (interval of 1kb) are shown for the maximum-likelihood values of $\hat{\alpha}$ = 4000 and $\hat{\tau}$ = 0.0001 and the confidence interval (see text for explanation) for $\hat{x}$ = 33,000 are given at the 5' flanking region of the gene *CG4661* (boxed area).

observed polymorphic sites are distributed over a small fraction of the observed haplotypes.

To examine if positive selection has shaped this haplotype structure, we applied the multi-locus maximum likelihood test (Mousset *et al*., submitted) using different combination of α and τ (*i.e.*, parameter range: α, 1,000–5,000; τ, 0.0001–0.05). We sampled every 1 kb along the studied region and compared the fit to the data of the neutral to the positive selection model. We found that the selection fits the data significantly better than the neutral model ($P < 10^{-10}$) and the maximum likelihood estimates of the selection parameters α, τ, and *x* are 4000, 0.0001, and 33000, respectively (Figure 4.1), leading to a strength of selection, *s*, of 0.0013. Applying the standard MAX – 2 rule (see for instance Kaplan and Weir 1995) the 95% CI for *x* includes sites located between 26 to 38 kb downstream of fragment *546*, at the 5' flanking region of the gene *CG4661* (boxed area in Figure 4.1).

A single-locus neutrality test (Hudson *et al.* 1994) was applied to the same data set. However, in contrast to the multi-locus test (see above) no departure from the standard neutral model (*i.e*., for each locus) was detected ($P > 0.05$; Table 4.1).

4.4 DISCUSSION

This study provides strong evidence that a recent selective sweep with moderate strength has shaped the observed haplotype pattern, *i.e*. the decay of frequency in haplotypes from the selected site. Moreover, the target of selection has been estimated to be on a site in the 5' flanking region of the gene *CG4661*.

In comparison to the selection parameters *x* and α, an underestimation of the parameter τ is likely due to the definition of the major haplotype (Mousset *et al*., submitted). Very recent selective events are more likely to lead to a haplotype structure with a large major haplotype class, whereas for older selective events this pattern will be obscured by new mutations and recombination events. However, although the maximum-likelihood estimates for *x* and α appear to be accurate, they may be sensitive to other parameters used in the coalescent simulations (Mousset *et al*., submitted). Since the effects of positive selection on genetically linked loci depend on the ratio of the recombination parameter *C* between two loci and α (Przeworksi 2002), and underestimation (overestimation) of *R* would lead to an underestimation (overestimation) of α (Mousset *et al*., submitted). Moreover, the observed haplotype structure could also be caused by demographic events (*i.e*., bottlenecks), which are likely to increase haplotype structuring over the entire genome (Galtier *et al*. 2000).

However, since the effects on neutral sites are higher around the target of positive selection and the non-neutral model of the test used assumes positive selection with spatial parameters (*x* and *R*), it is unlikely that the observed haplotype pattern was shaped by demographic events such as bottlenecks. Therefore, one could postulate that natural selection is favoring a mutation, which alters a regulatory element located at the 5' flanking region of gene *CG4661*. The estimated selection coefficient of the beneficial mutation of this study is similar to those reported by other studies (Harr *et al*. 2002; Schlenke and Begun 2004; Chapter 3, this thesis). The fact that we did not observe a locus with a homogeneous haplotype indicates that the loci used in this study are in some recombination distance from the fixed selected site. Similar observations have been made for the region comprising the gene *rp49* and the paralogous *janus* genes in *D. simulans* (Quesada *et al*. 2003).

This study has shown that positive directional selection has shaped the genetic variation in the ancestral *D. melanogaster* population. The use of a multi-locus haplotype test enabled us to gain evidence of a recent selective sweep, which would have been absent when we would have applied the haplotype data to single-locus haplotype tests. This clearly demonstrates the power of this multi-locus test, which future studies may take therefore into consideration. Moreover, since this likelihood approach enabled us to infer confidence intervals of this estimation, further research project may aim to characterize the target of selection within the estimated interval.

Part III: Genetic Variation of Derived Southeast Asian *Drosophila melanogaster*

CHAPTER 5

High Frequencies of Common Cosmopolitan Inversions in Southeast Asian *Drosophila melanogaster*

5.1 INTRODUCTION

Chromosomal polymorphism has been described for various natural populations of *Drosophila* mainly due to paracentric inversions (for review, Dobzhansky 1970; Sperlich and Pfriem 1986). In several species, observed geographic, seasonal, and altitudinal clines of inversion frequencies have been associated with climatic variables, suggesting that natural selection is operating on inversions (Krimbas and Powell 1992).

D. melanogaster, a cosmopolitan and domestic species, shows a high degree of chromosomal polymorphism in natural populations around the world (Lemeunier and Aulard 1992). According to their geographical distribution and abundance, inversions have been classified into four types: common cosmopolitan, rare cosmopolitan, recurrent endemic, and unique endemic (Ashburner and Lemeunier 1976; Mettler *et al*. 1977). Geographic variation in inversion frequencies has been reported for the common cosmopolitans (*In(2L)t, In(2R)NS, In(3L)P*, and *In(3R)P*) from different natural populations, including North American (Mettler *et al*. 1977), Japanese (Inoue and Watanabe 1979), Australasian (Knibb *et al*. 1981), Indian (Das and Singh 1991; Singh and Das 1992a), and Afrotropical ones (Aulard *et al*. 2002). These findings, together with the nonrandom association between linked and unlinked common cosmopolitan inversions observed in some natural populations of *D. melanogaster* (Knibb *et al*. 1981; Singh and Das 1991) provide strong evidence for the adaptive nature of these inversions.

D. melanogaster originated in the African mainland south of the Sahara and extended its range towards northern and eastern directions 10–15 kya (David and Capy 1988). Due to these colonization events, populations of this species, classified as being “ancient”, can be found nowadays in Europe and Asia (David and Capy 1988). Although studies on inversion polymorphisms are well documented in various regions of Asia (see above), there has been no report of chromosomal inversion polymorphisms of Southeast Asian *D. melanogaster*. Here we present the first analysis of chromosomal rearrangements from five natural *D. melanogaster* population samples of Thailand,

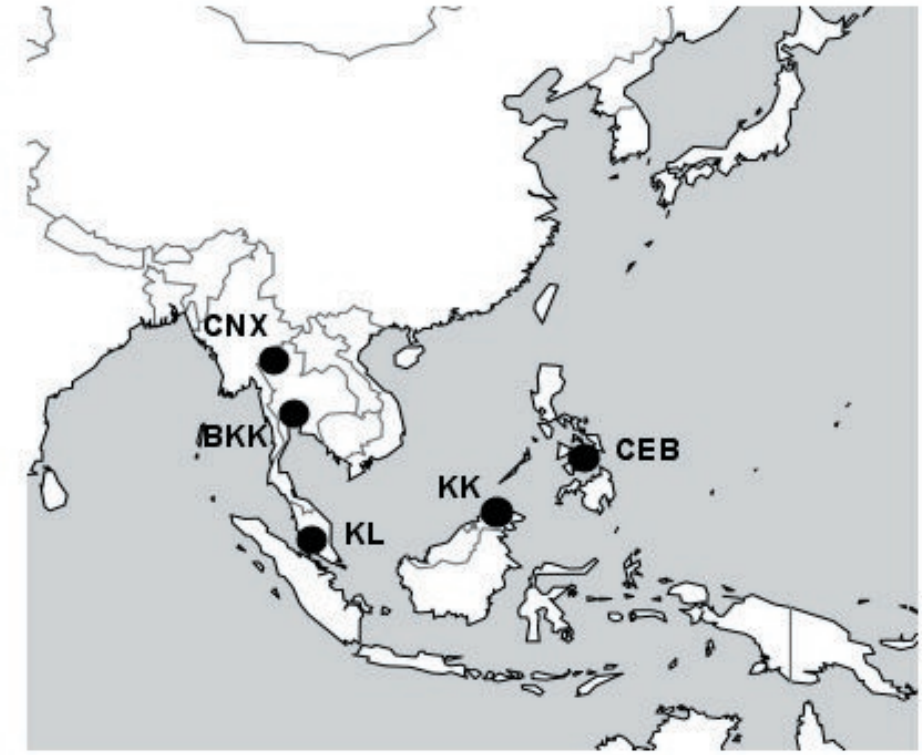


FIGURE 5.1 Geographic locations of the *D. melanogaster* samples collected in Southeast Asia. The abbreviations used are explained in Table 5.1.

Malaysia, and the Philippines. Our chromosomal analyses could not find any sign of natural selection acting on the common cosmopolitan inversions in these population samples, possibly because in the late Pleistocene (~18 kya), the sampling locations belonged to a single landmass ("Sundaland"). Furthermore, in comparison to populations from Africa, Asia and Australia, the Southeast Asian *D. melanogaster* revealed a relatively high level of inversion polymorphism.

5.2 MATERIALS AND METHODS

We sampled wild *D. melanogaster* from five geographically distant locations in Southeast Asia in October 2002. The details of the sampling locations are shown in Figure 5.1 and described with the geographic coordinates in Table 5.1. The abbreviated names of the sampling locations (as given in Table 5.1) are used throughout the chapter. The entire survey area spans a latitudinal range of 15.37° and a longitudinal range of 24.96° (Table 5.1). To collect the flies, we used an insect net in fruit and vegetable markets. Inseminated females from these collections were used to establish isofemale lines. These lines were kept under lab conditions for 18 month (around 38 generations) before we started the experiments. We used a F_1 third-instar larva from each isofemale line (maintained at 18 °C) for salivary

gland chromosome preparation. The lacto-acetic orcein method was applied to stain the chromosomes. The polytene chromosomes were observed using an inverted compound microscope and the banding patterns were designated according to the standard maps of LEFEVRE (1976). The karyotypic information of each larva was confirmed by examining four additional larvae from the same isofemale line.

5.3 RESULTS

5.3.1 Chromosomal Analyses and Inversion Frequencies

We identified a total of four paracentric inversions, each on both arms of chromosome 2 and 3, present at all sampling locations. According to their breakpoints, these are the four common cosmopolitian inversions: *In(2L)t, In(2R)NS, In(3L)P* and *In(3R)P* (see Appendix 5.1). We could neither detect rare cosmopolitans, recurrent or unique endemic inversions on the autosomes, nor did we identify inversions on the X chromosome.

The frequency of the common cosmopolitan inversions varies between sampling locations, most pronounced for *In(3R)P* with a maximum observed difference of 17.27% (Table 5.1). These results are consistent with the classification scheme of inversions, where common cosmopolitans are those that occur in many populations, often with a frequency greater than 5% (METTLER *et al*. 1977). Averaging over all five

TABLE 5.1

Sampling location, frequencies of the four cosmopolitan inversions, and the mean heterozygosity in five Southeast Asian population samples of *D. melanogaster*

Sampling location	Abbreviated sampling location	Latitude (°N)	Longitude (°E)	Total no. of isofemale lines examined	Common cosmopolitan (in %)				Total no. of inverted chromosomes	Mean no. of heterozygous inversions per individual
					In(2L)t	*In(2R)NS*	*In(3L)P*	*In(3R)P*		
Chiang Mai	CNX	18:45	98:58	25	18.00	18.00	10.00	18.00	32	1.28
Bangkok	BKK	13:05	100:29	21	16.67	14.29	14.29	26.19	30	1.43
Cebu	CEB	10:18	123:54	19	18.42	13.16	7.89	13.16	18	1.05
Kota Kinabalu	KK	05:56	116:03	23	17.39	15.22	15.22	30.43	36	1.57
Kuala Lumpur	KL	03:08	101:42	23	23.91	15.22	4.35	17.39	28	1.22
Average					18.88	15.18	10.35	21.02		

sampling locations (see Table 5.1), the highest mean frequency is found for *In(3R)P* (21.02%) followed by *In(2L)t* (18.88%), *In(2R)NS* (15.18%) and *In(3L)P* (10.35%). We observed the common cosmopolitan inversions only in their heterozygous karyotypes in all isofemale lines examined. Given the sample size of isofemale lines, the observed total number of inverted chromosomes ranged between 18 (CEB) and 36 (KK) among sampling locations, which results in a mean number of heterozygous inversions per individual of 1.05 (CEB) to 1.57 (KK; Table 5.1).

5.3.2 Genetic Differentiation and Geographic Variation

Comparison of the observed and expected numbers of different karyotypes for all common cosmopolitan inversions revealed that all population samples are in Hardy-Weinberg equilibrium, except for inversion *In(3R)P* in KK where the deviation from equilibrium is due to the observed significant excess of heterozygotes (χ^2 = 4.40, d.f. = 1, P = 0.043). Based on these results we went on with the analysis and estimated the amount of genetic differentiation among populations by the genetic identity index, I, and the genetic distance index, D (Nei 1978). The values for I are above 0.99400 for all pair-wise comparisons, and the highest estimated value of D is 0.00511 (Table 5.2). This indicates that although the samples were taken from geographically distant sampling locations, the Southeast Asian population samples are highly homogeneous. In addition, we could not detect a correlation between D and geographic distance (Spearman's R = -0.309, P = 0.385; data not shown).

TABLE 5.2

Genetic identity, *I* (above the diagonal), and genetic distance, *D* (below the diagonal; Nei 1978), between all different pairs of five Southeast Asian population samples of *D. melanogaster*

Sampling locations	CEB	CNX	BKK	KK	KL
CEB	******	0.99973	0.99869	0.99491	0.99793
CNX	0.00027	******	0.99929	0.99872	0.99947
BKK	0.00131	0.00071	******	0.99988	0.99841
KK	0.00511	0.00128	0.00012	******	0.99778
KL	0.00207	0.00053	0.00159	0.00222	******

Changes in inversion frequencies across populations can also be investigated by their relationship with latitude and longitude. We used angularly transformed

TABLE 5.3

P*-value of Pearsons correlation test of angularly transformed frequencies of the four common cosmopolitan inversions with latitude and longitude in five Southeast Asian population samples of *D. melanogaster

	In(2L)t	*In(2R)NS*	*In(3L)P*	*In(3R)P*
Latitude	0.275	0.461	0.580	0.829
Longitude	0.748	0.240	0.895	0.812

inversion frequencies from the five population samples under study to perform simple correlation analyses. Across the Southeast Asian population samples, we did neither find a significant latitudinal nor a longitudinal cline in inversion frequencies (Table 5.3).

Since we observed a common cosmopolitan inversion on each autosomal arm, we further investigated if nonrandom associations between these inversions exist in any of the population samples. We analyzed intra- and interchromosomal association for different pairs of linked [*In(2L)t–In(2R)NS* and *In(3L)P–In(3R)P*] and unlinked [*In(2L)t–In(3L)P*, *In(2L)t–In(3R)P*, *In(2R)NS–In(3L)P*, and *In(2R)NS–In(3R)P*] inversions. We used four karyotypic combinations (ST/ST–ST/ST, ST/ST–ST/IN, ST/IN–ST/ST, ST/IN–ST/IN, where ST designates the standard karyotype and IN an inversion type) for all four common cosmopolitan inversions, since we did not observe homozygote genotypes in any of the population samples. In all cases, the

TABLE 5.4

P*-value of one-tailed Fishers exact test for different intra- and interchromosomal combinations (see text for explanation) in five Southeast Asian population samples of *D. melanogaster

Sampling	Intrachromosomal		Interchromosomal			
location	*In(2L)t/In(2R)NS*	*In(3L)P/In(3R)P*	*In(2L)t/In(3L)P*	*In(2L)t/In(3R)P*	*In(2R)NS/In(3L)P*	*In(2R)NS/In(3R)P*
CNX	0.938	1.000	0.610	0.407	0.230	0.736
CEB	0.932	1.000	0.773	0.634	1.000	0.084
BKK	0.945	0.732	0.686	0.562	0.576	0.367
KK	0.467	0.761	0.974	0.069	0.725	0.124
KL	0.778	1.000	0.739	0.611	0.526	0.810

observed number of each pair fitted well with the expectations indicating no evidence for nonrandom association (Table 5.4).

5.4 DISCUSSION

This study represents the first broad-scale analysis of Southeast Asian *D. melanogaster* inversion polymorphism. Our analyses of five population samples provide new population genetic insights into these ancient populations of *D. melanogaster*.

5.4.1 Inversions and Their Frequencies in Southeast Asia

Several hundred of chromosomal rearrangements have been reported for wild *D. melanogaster* populations from various parts of the world (Lemeunier and Aulard 1992). We identified a total of four different inversions, namely the common cosmopolitans. Other studies reported a higher number of different paracentric inversions in various natural populations of *D. melanogaster* (*e.g.,* Knibb *et al*. 1981; Das and Singh 1991; Aulard *et al*. 2002) due to identified rare cosmopolitans and endemic rearrangements. Of the six classical rare cosmopolitans (Lemeunier and Aulard 1992), some have been described in India (Das and Singh 1991; Singh and Das 1992a), Africa (Aulard *et al*. 2002), and Australasia (Knibb *et al.* 1981). The total identified number of unique inversions exceeds 500 from various populations of *D. melanogaster* (for summary*,* Aulard *et al*. 2002). However, within Asia, the number of unique inversions ranged between 18 in India (Singh and Das 1991; Singh and Das 1992a), 54 in Korea (Choi 1977; Choi *et al*. 1984) and 163 in Japan (Inoue 1988). Given the time frame between the fly collection and the actual experiments in our study, rare inversions might have been lost over time due to drift in the laboratory because of their low initial frequency (Singh and Das 1992b). X-linked inversions by themselves are rare in nature because of their potential deleterious effect in hemizygous males (Coyne *et al.* 1991). So far, only two studies reported inversion polymorphism on the X chromosome in natural population of *D. melanogaster* in Africa (Aulard *et al*. 2002) and India (Das and Singh 1991).

Frequencies of the common cosmopolitan inversions have been reported from various natural populations of *D. melanogaster*, including Austalasia (Knibb *et al*. 1981), India (Das and Singh 1991; Singh and Das 1992a), Japan (Inoue and Watanabe 1979; Inoue *et al.* 1984), Korea (Choi 1977; Choi *et al*. 1984) and Africa (Aulard *et al*. 2002). We were interested in whether the observed frequencies of the common cosmopolitans in the Southeast Asian population samples were on a similar scale to those reported from populations of the surrounding regions. Since *D. melanogaster*

originated in the African mainland south of the Sahara (David and Capy 1988), we included the putatively ancestral population samples from the Afrotropical region (Aulard *et al*. 2002) in the analysis. A one-factorial ANOVA on angularly transformed inversion frequencies revealed a significantly higher mean frequency (F = 3.080, d.f. = 4, 95, P = 0.020) in Asian and Australasian populations for the common cosmopolitan inversion *In(2R)NS* in comparison with the frequency observed from the Afrotropical region (Figure 5.2). The Korean population was not included in the analysis due to the sample size of one. It is known that the frequencies of common cosmopolitan inversions decline or become eliminated under laboratory conditions in mass cultures over time in *D. melanogaster* populations (Singh and Das 1992b; Inoue 1979). However, this effect was found to be relatively small in isofemale lines (Knibb *et al*. 1981; Singh and Das 1992b). Therefore, our results on inversion frequencies of the common cosmopolitans are probably biased downwards, and could even be higher in nature.

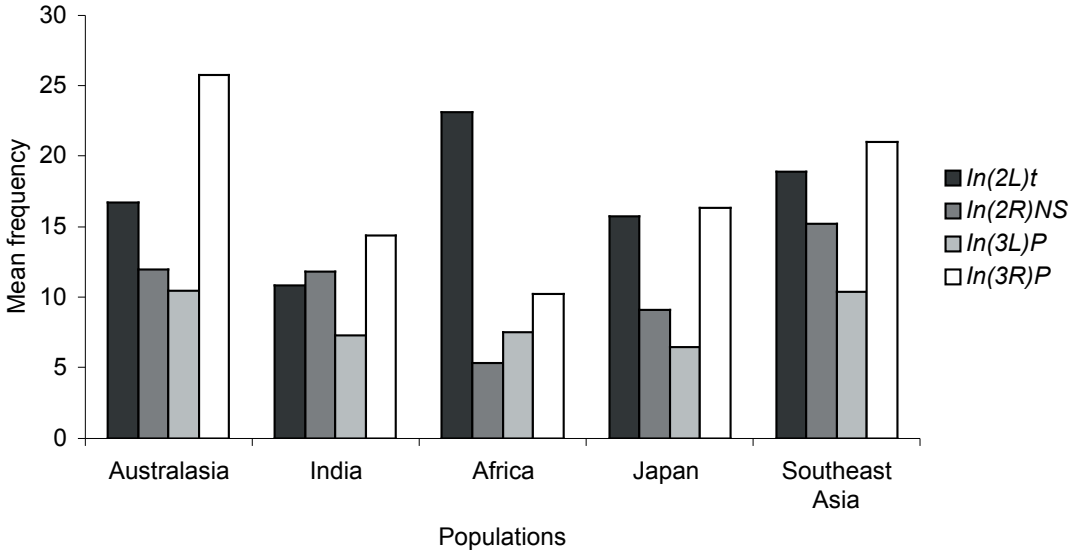


Figure 5.2 Mean frequencies (in %) of the four common cosmopolitan inversions observed in Australasia (Knibb *et al*. 1981), India (Das and Singh 1991), Japan (Inoue *et al*. 1984), Africa (Aulard *et al*. 2002) and Southeast Asia (this study).

5.4.2 Genetic Differentiation and Geographic Variation

Natural populations of *D. melanogaster* are not only characterized by a high degree of inversion polymorphism, but also by strong genetic differentiation due to latitudinal, longitudinal, altidudinal, seasonal, and temporal variation (Lemeunier *et al*. 1986). Our results do not support these observations. In particular, we did not find a latitudinal or longitudinal trend for the common cosmopolitans. This lack of clines can be explained by a homogeneous habitat throughout the sampling

range of this study. Moreover, in the late Pleistocene (~18 kya), the sampling locations were connected by a single large landmass ("Sundaland"; Van-Wright 1990) because the sea level was ~120 meters below the present level (Shackleton 2000). An originally established panmictic population on Sundaland could have led to the homogeneous Southeast Asian population samples found in our study. The very low observed genetic distance between all pairs of population samples also supports this hypothesis. Two hypothetical scenarios may explain the origin of the panmictic *D. melanogaster* population on Sundaland: First, Southeast Asia was colonized by ancient *D. melanogaster* during that time (David and Capy 1988). This implies a coordinated time frame between the arrival of *D. melanogaster* in Southeast Asia and the rising sea level after the glaciation period. Second, *D. melanogaster* colonized Southeast Asia before the last glaciation period allowing a establishment of the population on Sundaland before or when sea levels dropped down 18 kya. Evidence of a Far Eastern race of *D. melanogaster* supports this scenario (David and Capy 1988; Lachaise and Silvain 2004). This race exhibits several distinct morphological and physiological properties (David *et al*. 1976; Lemeunier *et al*. 1986) and Far Eastern populations are strongly diverged in comparison to other *D. melanogaster* populations (Solignac 2004). These differences suggest a long separate evolutionary history from the ancestral populations (David and Capy 1988; Lachaise and Silvain 2004). More interestingly, in a recent multi-locus DNA sequence study in worldwide samples of *D. ananassae*, the Southeast Asian samples (particularly the samples from Sundaland) were found to be ancestral (Das *et al*. 2004). Since both *Drosophila* species are human commensals, a parallel pattern with human diversity and migration in Southeast Asia could also have led to the observed pattern (see also Das *et al*. 2004).

Latitudinal clines, where inversion frequency decreases with increasing distance from the equator, have been reported from studies undertaken in North America (Mettler *et al*. 1977), Australasia (Knibb *et al*. 1981), Japan (Inoue and Watanabe 1979; Inoue *et al*. 1984) and India (Das and Singh 1991; Singh and Das 1992a). The precise association between frequencies and climatic variables varies between different inversions in the same region and also for a particular inversion between different regions (Knibb 1982). In contrast to our findings from Southeast Asia, the observed latitudinal clines may therefore be explained by the greater distance from the equator leading to more heterogeneous habitats. This might also be the reason for the disappearance of the observed latitudinal cline in Japan when the populations sampled from four southern islands are excluded (Inoue and Watanabe 1979; Inoue *et al.* 1984). In addition, no latitudinal cline was observed from *D. melanogaster*

populations from the western United States (Voelker *et al*. 1977) and the Afrotropical region (Aulard *et al*. 2002). For the former study, the latitudinal range (35.4-44.0°N; Voelker *et al*. 1977) was probably too narrow to show differences in the habitat of the three western United States populations. Although the latter study covered a latitudinal range between 14°N and 21°S (Aulard *et al*. 2002), the net distances from the equator were probably too small to show a cline of inversion frequencies, similar to our observations of the Southeast Asian *D. melanogaster* populations.

Longitudinal trends on frequencies have been reported of some cosmopolitan inversions for Afrotropical (Aulard *et al*. 2002), Australasia (Knibb 1982) and Japanese (Inoue and Watanabe 1979; Inoue *et al*. 1984) populations. If the observed longitudinal effect on inversion polymorphism in Afrotropical populations (Aulard *et al*. 2002) can be attributed to heterogeneous habitats alone, or if the observed West-East differentiation has also a historical component, as suggested by a molecular study (Bénassi and Veuille 1995), remains to be answered. In contrast to the Afrotropical region, the inversion frequencies observed in Australasia increase eastwardly, but a correlation between longitude and ecological parameters was not as strong as found for latitude (Knibb 1982). In the case of the Japanese population, the longitudinal trend does not exist anymore when the samples taken from the four southern islands are excluded (Inoue and Watanabe 1979; Inoue *et al*. 1984). This observation strengthens our hypothesis, that although the five Southeast Asian population samples in our study were collected from geographically distant areas, they seem to have been living in a rather homogeneous habitat.

Singh and Das (1992b) categorized *D. melanogaster* as being a species with flexible types of inversion polymorphisms, which show changes in their gene pool composition in space and time. However, although the temperature conditions were consistent in the laboratory over 18 months, inversion frequencies might not have changed due to the “new environment” the species were facing, since, as discussed before, the effects are relatively small in isofemale lines (Knibb *et al*. 1981; Singh and Das 1992b).

5.4.3 Association between Inversions

It is widely known that nonrandom association of linked and unlinked inversions results from selection involving epistatic interaction in *D. melanogaster* (*e.g.,* Knibb *et al*. 1981; Singh and Das 1991). However, we did not observe this association between any of the possible pairs of inversion karyotypes in the populations analyzed. Linkage disequilibrium has been reported for linked inversions, including populations from

Japan (Inoue and Watanabe 1979), Australasia (Knibb *et al*. 1981), India (Singh and Das 1991) and Korea (Choi 1977). Nonrandom association of unlinked inversions has only been reported from Australasia (Knibb *et al*. 1981) and India (Singh and Das 1991). In the absence of selection, recombination would break down linkage disequilibrium over time (Singh and Das 1991). Interchromosomal interaction might be breaking apart if the adaptive value has no fitness advantage anymore in a given environment (Singh and Das 1991). Whether or not our maintenance conditions affected the analysis of the association between inversions is not known. Singh and Das (1992b) suggested that in isofemale lines only random genetic drift is responsible for the differences in inversion frequencies between the initial and the actual analyzed population. Since the cosmopolitan inversions are present in high frequencies throughout the Southeast Asian population samples, preexisting nonrandom association between linked and/or unlinked inversions should be still observable.

In conclusion, our study provides evidence for a unique pattern of the chromosomal polymorphism in Southeast Asian populations of *D. melanogaster*. Since this is the first-ever population genetic study on such a broad population range in this region and considering the fact that these constitute the ancient populations of *D. melanogaster* (David and Capy 1988), more studies preferably with molecular markers would be necessary to reveal the detailed population history of this species in Southeast Asia.

It is also noteworthy to mention here that the mainland of Southeast Asia has served as a pool of genetic diversity among Asian humans (Ballinger *et al*. 2000; Su *et al*. 2000). A detail population genetic study of Southeast Asian *D. melanogaster* populations which cohabit with *D. ananassae* throughout the entire distribution range could thus throw light on the diversity pattern in Southeast Asian fauna in general and help understanding the population history of *D. melanogaster* in particular.

CONCLUSION

The DNA sequence analyses performed in this thesis revealed significant evidence of Darwinian selection on a molecular level in the model organism *D. melanogaster*. Sampling of ancestral and derived populations of this species highlighted not only genetic patterns shaped by natural selection, but also by demography, and provided new insights into the evolutionary history of this species. In addition, evidence that other evolutionary forces, such as recombination, contribute to the observed level of DNA sequence variation in this species was found.

This thesis has shown that Darwinian selection has shaped genetic patterns in both examined *D. melanogaster* populations, in the putatively ancestral one from Africa (Zimbabwe) and in a derived population from Europe (The Netherlands). The detection of a potential target site in the 5' flanking region of gene *CG4661* in the Zimbabwean population may motivate future research to identify the specific beneficial mutation and its effect on the phenotype. Since most regulatory elements (*i.e.*, enhancers or promoters) are located in the 5' flanking region of genes, mutations in these regions may alter the genes' level of transcription (Wray *et al*. 2003). This can be investigated by measuring the genes' expression level using real-time PCR, microarray techniques and mutagenesis combined with germline transformation.

In contrast to the ancestral population, several replacement sites were identified as potential beneficial mutations in the derived *D. melanogaster* population. These mutations change the primary amino acid sequence of the encoded protein and therefore can lead to phenotypic variation. Since each replacement site affects the structure of the protein, site-directed mutagenesis may allow one to identify the effect of each sequence variant on the phenotype. In addition, further sequencing of the 5' flanking region of the genes *CG1677* and *CG2059* may highlight regulatory mutations. An effect of a potential regulatory change in expression may first be investigated by comparing the expression level of each gene between the derived population and the Zimbabwean lineages carrying the ancestral variant. Real-time PCR and microarray techniques can be used to measure the expression level and observed differences may stimulate a further search for the beneficial mutation in the 5' flanking or in the 5' and 3' untranslated regions of these genes in the derived population. These investigations may reveal the role of the beneficial mutation in the adaptation to the newly colonized temporal habitat of Europe.

Such evidence would be consistent with the hypothesis that *D. melanogaster* originated from Africa and expanded its range to the rest of the world after the last glaciation 10 to 15 kya (David and Capy 1988). Given the number of shared haplotypes between African and non-African samples (Baudry *et al*. 2004) and the observation that most of the variants in non-African populations are shared with the ancestral populations (Glinka *et al*. 2003; Baudry *et al*. 2004), non-African populations are likely to have originated from East Africa. Although more data are needed to confirm this hypothesis, it is in good agreement with the observations of this thesis. The Zimbabwean population was found to have expanded its size ~15 kya due to improving climatic conditions on the African continent, which eventually may have led to the colonization of the Eurasian continent. Therefore, the range expansion of *D. melanogaster* has occurred without the help of man and, following this hypothesis, the wild-to-domestic habitat shift happened with the rise of agriculture after the Neolithic revolution (David and Capy 1988; Lachaise and Silvain 2004). However, the hypothesis presented in this thesis of a Far Eastern race having colonized Southeast Asia before the last glacial maximum contradicts the Neolithic habitat shift of *D. melanogaster*.

Archaeological data suggest that modern humans moved along the coast, rather than through the interior of Africa, allowing them to cross the southern part of the Red Sea 65 kya (Stringer 2000), when sea level was low due to a glacial maximum (Webb III and Bartlein 1992). After crossing the Arabian peninsula, modern humans entered Asia between 40 and 60 kya (Cavalli-Sforza and Feldman 2003) followed by a northward (reaching China and Japan) and southward (through Malaysia and Indonesia) migration that coincided with the receding glaciers in these regions (Stringer 2000). *D. melanogaster* could have accompanied (or followed) these human migrations, but only if the wild-to-domestic behavior shift had occurred during that time (Lachaise and Silvain 2004). A more detailed analysis of DNA sequence polymorphism in Southeast Asian *D. melanogaster* populations might shed light into this ongoing controversial discussion. A comparison of the genetic pattern of Southeast Asian populations to other derived populations (*e.g.*, European) may reveal differences between these two potential colonization events. In particular, adaptive events in Southeast Asian *D. melanogaster* populations should be older according to the hypothesis of an earlier colonization of this region in comparison to the European continent.

Although natural selection and demography have substantially contributed to the amount of observed genetic variation in the Zimbabwean *D. melanogaster* population,

this thesis has also shown that recombination by itself is mutagenic and therefore influences genetic variation in this species. This result was surprising, since no other study of genetic variation in *Drosophila* has reported it before. However, it remains to be investigated if the level of divergence observed between *D. melanogaster* and *D. simulans* reflects only the substitutions accumulated after the species split or, in addition, the diversity present in the ancestral species. In other words, divergence could increase with the recombination rate, because diversity increased with the recombination rate in the ancestral species due to variation-reducing selection (see also Hellmann *et al*. 2003). To evaluate this possibility, *D. melanogaster* could be compared to *D. yakuba* since the common ancestor of both species is estimated to have lived about 10 million years ago (Powell 1997; Begun and Langley 2003) and therefore the ancestral polymorphism of the ancestor species should have a minor effect on levels of divergence. The availability of the genome sequence of *D. yakuba* will facilitate this task (see Begun and Langley 2003).

LITERATURE CITED

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

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

Andolfatto, P., 2001a Adaptive hitchhiking effects on genome variability. Curr. Opin. Genet. Dev. **11:** 635–641.

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

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

Andolfatto, P., and J. D. Wall, 2003 Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. Genetics **165:** 1289–1305.

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

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

Aquadro, C. F., 1997 Insights into the evolutionary process from patterns of DNA sequence variability. Curr. Opin. Genet. Dev. **7:** 835–840.

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

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nature Rev. **3:** 299–309.

Ashburner, M., and F. Lemeunier, 1975 Relationship within the *melanogaster* species subgroup of the genus *Drosophila* (Sophophora). I. Inversion polymorphism in *Drosophila melanogaster* and *Drosophila simulans*. Proc. R. Soc. Lond. B. Biol. Sci. **193:** 137–157.

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

Ballinger, S. W., T. G. Schurr, A. Torroni, Y. Y. Gan, J. A. Hodge *et al*., 2000 Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient Mongoloid migrations. Genetics **130:** 139–152.

Baudry, E., B. Viginier and M. Veuille, 2004 Non-African populations of *Drosophila melanogaster* have a unique origin. Mol. Biol. Evol. **21:** 1482–1491.

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

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

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

Begun, D. J., and C. H. Langley, 2003 Proposal for the sequencing of *Drosophila yakuba* and *D. simulans*. White Paper to NHGRI.

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

Bénassi, V., and M. Veuille, 1995 Comparative population structuring of molecular and allozyme variation of *Drosophila melanogaster Adh* between Europe, West Africa and East Africa. Genet. Res. **65:** 95–103.

Bénassi, V., S. Aulard, S. Mazeau and M. Veuille, 1993 Molecular variation of *Adh* and *P6* genes in an African population of *Drosophila melanogaster* and its relation to chromosomal inversions. Genetics **134:** 789–799.

Berry, A. J., J. W. Ajioka and M. Kreitman, 1991 Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

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

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

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley and D. B. Neale, 2004 Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc. Natl. Acad. Sci. USA **101:** 15255–15260.

Cavalli-Sforza, L. L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. B. Biol. Sci. **164**: 362–379.

Cavalli-Sforza, L. L., and M. W. Feldman, 2003 The application of molecular genetic approaches to the study of human evolution. Nat. Genet. Suppl. **33:** 266–275.

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

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

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

Choi, Y., 1977 Chromosomal polymorphism in a Korean natural population of *Drosophila melanogaster*. Genetica **47:** 155–160.

Choi, Y., Y. M. Ha and S. K. Kim, 1984 Further studies on chromosomal inversion polymorphisms in a natural population of *Drosophila melanogaster*. Korean J. Genet. **6:** 81–90.

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

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

Darwin, C., 1859 *The Origin of Species by Means of Natural Selection*. John Murray, London.

Das, A., and B. N. Singh, 1991 Genetic differentiation and inversion clines in Indian natural populations of *Drosophila melanogaster*. Genome **34:** 618–625.

Das, A., S. Mohanty and W. Stephan, 2004 Inferring the population structure and demography of *Drosophila ananassae* from multilocus data. Genetics **168:** 1975–1985.

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

David, J. R., C. Bocquet and E. Pla, 1976 New results on the genetic characteristics of the Far East race of *Drosophila melanogaster*. Genet. Res. **28:** 253–260.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15:** 1788–1790.

Depaulis, F., L. Brazier and M. Veuille, 1999 Selective sweep at the *Drosophila melanogaster Suppressor of Hairless* locus and its association with the *In(2L)t* inversion polymorphism. Genetics **152:** 1017–1024.

Depaulis, F., S. Mousset and M. Veuille, 2001 Haplotype tests using coalescent simulations conditional on the number of segregating sites. Mol. Biol. Evol. **18:** 1136–1138.

Depaulis, F., S. Mousset and M. Veuille, 2003 Power of neutrality tests to detect bottlenecks and hitchhiking. J. Mol. Evol. **57 Suppl 1:** S190–200.

Depaulis, F., S. Mousset and M. Veuille, 2005 Detecting selective sweeps with haplotype tests, *in Selective Sweep,* edited by D. Nurminsky. Landes Biosciences, Georgetown, TX. In press.

De Vivo, M., and A. P. Carmignotto, 2004 Holocene vegetation change and the mammal faunas of South America and Africa. J. Biogeogr. **31:** 943–957.

Dobzhansky, T., 1970 *Genetics of the Evolutionary Process*. Columbia University Press, New York.

Ewens, W. J., 1979 *Mathematical population genetics*. Springer-Verlag, New York.

Excoffier, L., and S. Schneider, 1999 Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc. Natl. Acad. Sci. USA **96:** 10597–10602.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155**: 1405–1413.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon, Oxford.

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

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

Galtier, N., F. Depaulis and N. H. Barton, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics **155:** 981–987.

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

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

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

Graur, D., and W.-H. Li, 1999 *Fundamentals of Molecular Evolution*. Sinauer, Sunderland, MA.

Griffiths, R. C., and S. Tavaré, 1994 Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

Grove, A. T., 1993 Africa's climate in the Holocene, pp. 32–42 in *The Archaeology of Africa. Food, Metals and Towns*, edited by T. Shaw, P. Sinclair, B. Andah and A. Okpoko. Routledge, London and New York.

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

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Pääbo and M. Przeworski, 2003 A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. **72:** 1527–1535.

Hewitt, G., 2000 The genetic legacy of the Quaternary ice ages. Nature **405:** 907–913.

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

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyma and J. Antonovics. Oxford University Press, New York.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*, edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland, MA.

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

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

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

Hudson, R. R., M. Slatkin and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the *superoxide dismutase (sod)* region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

Inoue, Y., 1979 The fate of polymorphic inversions of *Drosophila melanogaster* transferred to laboratory conditions. Japan J. Genet. **54:** 83–96.

Inoue, Y., 1988 Chromosomal mutation in *Drosophila melanogaster* and *Drosophila simulans*. Mutat. Res. **197:** 85–92.

Inoue, Y., and T. K. Watanabe, 1979 Inversion polymorphisms in Japanese natural populations of *Drosophila melanogaster*. Japan J. Genet. **54:** 69–82.

Inoue, Y., T. Watanabe, T. K. Watanabe, 1984 Evolutionary change of the chromosomal polymorphism in *Drosophila melanogaster* populations. Evolution **38:** 753–765.

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

Kaplan, N. L, and B. S. Weir, 1995 Are moment bounds on the recombination fraction between a marker and a disease locus too good to be true? Allelic association mapping revisited for simple genetic diseases in the Finnish population. Am. J. Hum. Genet. **57:** 1486–1498.

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

Kauer, M., D. Dieringer and C. Schlötterer, 2003 Nonneutral admixture of immigrant genotypes in African *Drosophila melanogaster* populations from Zimbabwe. Mol. Biol. Evol. **20:** 1329–1337.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

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

KIM, Y., and W. STEPHAN, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

KIM, Y., and R. NIELSEN, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167:** 1513–1524.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

KIMURA, M., 1971 Theoretical foundation of population genetics at the molecular level. Theor. Popul. Biol. **2:** 174–208.

KIMURA, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press, New York.

KINGMAN, J. F., 1982 The coalescent. Stochast. Proc. Appl. **13:** 235–248.

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

KLIMAN, R. M., P. ANDOLFATTO, J. A. COYNE, F. DEPAULIS, M. KREITMAN *et al.*, 2000 The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics **156:** 1913–1931.

KNIBB, W. R., 1982 Chromosome inversion polymorphisms in *Drosophila melanogaster*. II. Geographic clines and climatic associations in Australasia, North America and Asia. Genetica **58:** 213–221.

KNIBB, W. R., J. G. OAKESHOTT and J. B. GIBSON, 1981 Chromosome inversion polymorphisms in *Drosophila melanogaster*. I. Latitudinal clines and associations between inversions in Australasian populations. Genetics **98:** 833–847.

KREITMAN, M., 2000 Methods to detect selection in populations with applications to the human. Annu. Rev. Genomics Hum. Genet. **01:** 539–559.

KRIMBAS, C. B., and J. R. POWELL, 1992 *Drosophila Inversions Polymorphism*. CRC Press, Boca Raton, FL.

LACHAISE, D., and J.-F. SILVAIN, 2004 How two Afrotropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster-D. simulans* palaeogeographic riddle. Genetica **120:** 17–39.

Lachaise, D., M. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup, pp. 159–225 in *Evolutionary Biology*, edited by M. K. Hecht, B. Wallace and G. T. Prance. Plenum, New York.

Lazzaro, B. P., and A. G. Clark, 2003 Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster*. Mol. Biol. Evol. **20**: 914–923.

Lefevre, G., 1976 A photographic representations and interpretation of the polytene chromosomes of *Drosophila melanogaster* salivary glands, pp. 32–66 in *The Genetics and Biology of Drosophila*, edited by M. Ashburner and E. Novitski. Academic Press, New York.

Lemeunier, F., and S. Aulard, 1992 Inversion polymorphism in *Drosophila melanogaster*, pp. 339–405 in *Drosophila Inversion Polymorphism*, edited by C. B. Krimbas and J. R. Powell. CRC Press, Boca Raton, FL.

Lemeunier, F., J. R. David, L. Tsacas and M. Ashburner, 1986 The *melanogaster* species group, pp. 147–256 in *The Genetics and Biology of Drosophila*, edited by M. Ashburner, H. L. Carson and J. N. Thompson. Academic Press, New York.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of genes frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74**: 175–195.

Li, Y. J., Y. Satta and N. Takahata, 1999 Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes Genet. Syst. **74:** 117–127.

Markstein, M., P. Markstein, V. Markstein and M. S. Levine, 2002 Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. Proc. Natl. Acad. Sci. USA **99:** 763–768.

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

Maley, J., 1993 The climatic and vegetational history of the equatorial regions of Africa during the upper Quaternary, pp. 43–52 in *The Archaeology of Africa. Food, Metals and Towns*, edited by T. Shaw, P. Sinclair, B. Andah and A. Okpoko. Routledge, London and New York.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351:** 652–654.

Meiklejohn, C. D., Y. Kim, D. L. Hartl and J. Parsch, 2004 Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. Genetics **168:** 265–279.

Mettler, L. E., R. A. Voelker and T. Mukai, 1977 Inversion clines in natural populations of *Drosophila melanogaster*. Genetics **87:** 169–176.

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

Mousset, S., N. Derome and M. Veuille, 2004 A test of neutrality and constant population size based on the mismatch distribution. Mol. Biol. Evol. **21:** 724–731.

Mousset, S., S. Glinka and W. Stephan A maximum likelihood neutrality test based on multilocus haplotype data. Submitted.

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

Nei, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89:** 583–590.

Nei, M., and T. Maruyama, 1975 Lewontin-Krakauer test for neutral genes. Genetics **80:** 395.

Nurminsky, D., D. De Aguiar, D. Bustamante and D. L. Hartl, 2001 Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. Science **291:** 128–130.

Nurminsky, D., M. V. Nurminskaya, D. De Aguiar and D. L. Hartl, 1998 Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature **396:** 572–575.

Orengo, J. D., and M. Aguadé, 2004 Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: multi-locus pattern of variation and distance to coding regions. Genetics **167:** 1759–1766.

Orr, H. A., and J. A. Coyne, 1992 The genetics of adaptation: a reassessment. Am. Nat. **140:** 725–742.

Parsch, J., C. D. Meiklejohn and D. L. Hartl, 2001 Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. Genetics **159:** 647–657.

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

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

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

Przeworski, M., J. D. Wall and P. Andolfatto, 2001 Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 291–298.

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

Ramos-Onsins, S. E., B. E. Stranger, T. Mitchell-Olds and M. Aguadé, 2004 Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics **166:** 373–388.

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

Robertson, L. S., 1975 Gene frequency distributions as a test for selective neutrality. Genetics **81**: 775–785.

Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. **3:** 380–390.

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

Rozas, J., J. C. Sánchez-Del Barrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al*., 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Schlenke, T. A., and D. J. Begun, 2004 Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **101:**1626–1631.

Shackelton, N. J., 2000 The 100,000 year ice-age cycle identified and found to lag temperature, carbon dioxide and orbital eccentricity. Science **289:** 1897–1902.

Singh, B. N., and A. Das, 1991 Epistatic interaction between unlinked inversions in Indian natural populations of *Drosophila melanogaster*. Genet. Sel. Evol. **23:** 371–383.

Singh, B. N., and A. Das, 1992a Further evidence for latitudinal inversion clines in natural populations of *Drosophila melanogaster* from India. J. Hered. **83:** 227–230.

Singh, B. N., and A. Das, 1992b Changes of inversion polymorphism in laboratory populations of *Drosophila melanogaster*. Z. zool. Syst. Evolut.-forsch. **30:** 268–280.

Sokal, R. R. and F. J. Rohlf, 2001 *Biometry*. W.H. Freeman and Co, New York.

Solignac, M., 2004 Mitochondrial DNA in the *Drosophila melanogaster* complex. Genetica **120:** 41–50.

Sperlich, D., and P. Pfriem, 1986 Chromosomal polymorphism in natural and experimental populations, pp. 257–309 in *The Genetics and Biology of Drosophila*, edited by M. Ashburner, H. L. Carson and J. N. Thompson. Academic Press, New York.

Stajich, J. E., and M. W. Hahn, 2005 Disentangling the effects of demography and selection in human. Mol. Biol. Evol. **22:** 63–73.

Stephan, W., 1997 Mathematical model of the hitchhiking effect, and its application to DNA polymorphism data, pp. 29–45 in *Advances in Mathematical Population Dynamics: Molecules, Cells and Man*, edited by O. Arino, D. Axelrod and M. Kimmel. World Scientific, London.

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

Stephan, W., T. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.

Storz, J. F., B. A. Payseur and M. W. Nachman, 2004 Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol. Biol. Evol. **21:** 1800–1811.

Stringer, C., 2000 Coasting out of Africa. Nature **405:** 24–27.

Sturtevant, A. H., 1917 Genetic factors affecting the strength of linkage in *Drosophila*. Proc. Natl. Acad. Sci. USA **3:** 555.

Su, B., L. Jin, P. Underhill, J. Martinson, N. Saha *et al*., 2000 Polynesian origins: insights from the Y chromosome. Proc, Natl, Acad, Sci, USA **97:** 8225–8228.

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.

Van-Wright, R. I., 1990 The Philippines – key to the biogeography of Wallacea, pp. 19–34 in *Insects and the Rain Forests of South East Asia (Wallacea)*, edited by W. J. Knight and J. D. Holloway. Royal Entomological Society, London.

Vilella, A. J., A. Blanco-Garcia, S. Hutter and J. Rozas Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Submitted.

Voelker, R. A., T. Muka and F. M. Johnson, 1977 Genetic variation in populations of *Drosophila melanogaster* from the western United States. Genetica **47:** 143–148.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–80.

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

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

Webb, T. III, and P. J. Bartlein, 1992 Global changes during the last 3 million years: climatic controls and biotic responses. Annu. Rev. Ecol. Syst. **23:** 141–173.

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.

Wickham-Jones, T., 1994 *Mathematica Graphics: Techniques and Applications*. TELOS/Springer-Verlag, New York.

WRAY, G. A, M. W. HAHN, E. ABOUHEIF, J. P. BALHOFF, M. PIZER *et al.*, 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20:** 1377–1419.

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

WU, C.-I., H. HOLLOCHER, D. J. BEGUN, C. F. AQUADRO, Y. XU *et al.*, 1995 Sexual isolation in *Drosophila melanogaster*: a possible case of incipient speciation. Proc. Natl. Acad. Sci. USA **92:** 2519–2523.

APPENDIX

APPENDIX 1.1 Nucleotide diversity estimates and test statistics for the African population

Sequences in EMBL database (http://www.ebi.ac.uk), accession numbers AJ568984-AJ571588 (complete set). Fragments are ordered from the telomere to the centromere; for each one, the following information is given:

r is the recombination rate expressed in rec/bp/gen x 10^{-8};

Type indicates if the fragment belongs to intergenic region (IR) or to an intron (In);

Absolute position is in base pairs, from the telomere;

n is the number of lines sequenced;

lth is the number of sites studied (excluding insertions and deletions polymorphism);

S is the number of segregating sites;

π is the nucleotide diversity (TAJIMA 1983);

θ_W is the WATTERSON (1975) estimate of nucleotide diversity;

Tajima's *D* test statistic (TAJIMA 1989a);

for the *H* and *K* haplotype statistics (DEPAULIS and VEUILLE 1998), it is indicated whether the observation is lower (–) or higher (+) than the simulated median across the sample (see text);

* $P < 0.05$;

** $P < 0.01$;

† one-tailed test without recombination, $P < 0.05$.

Fragment	*r*	Type	Abs. posit.	*n*	*lth*	*S*	π	θ_W	Tajima's *D*	*H*	*K*
10	0.486	In	1899930	12	346	10	0.0115	0.0096	0.8179	+	–
9	0.585	IR	1929751	12	323	2	0.0024	0.0021	0.5542	+	–
17	0.585	In	1946108	12	781	15	0.0056	0.0064	–0.4954	+	+
6	0.436	IR	1988709	12	402	6	0.0036	0.0049	–1.0217	–	–
1	0.811	In	2004307	12	380	15	0.0130	0.0131	–0.0108	+	+
15	0.811	In	2010026	12	462	2	0.0010	0.0014	–0.8497	–	–
22	0.811	In	2129973	12	618	11	0.0042	0.0059	–1.2012	+	+
26	1.051	IR	2140729	12	570	18	0.0064	0.0105	–1.7107	+	+
18	1.587	In	2448658	12	502	13	0.0073	0.0086	–0.6279	+	+
4	1.587	IR	2455342	12	359	8	0.0079	0.0074	0.2777	+	+
5	2.019	In	2486993	12	245	14	0.0186	0.0189	–0.0693	+	+
55	2.738	IR	3235896	12	661	32	0.0137	0.0160	–0.6605	+	–
54	2.738	IR	3238859	12	418	33	0.0209	0.0261	–0.9036	+	+
57	3.138	IR	3333268	11	547	12	0.0069	0.0075	–0.3726	+	+
60	3.138	IR	3448557	12	615	30	0.0160	0.0162	–0.0457	+	+

56	3.290	In	3603702	12	325	5	0.0056	0.0051	0.3955		–	–
76	3.290	IR	3653297	12	538	33	0.0161	0.0203	–0.9475		+	+
78	3.549	IR	3727323	12	612	23	0.0102	0.0124	–0.7996		+	+
81	3.883	IR	3879576	12	561	19	0.0116	0.0112	0.1555		+	+
84	3.883	IR	4018352	11	596	21	0.0100	0.0120	–0.7761		+	+
85	3.883	IR	4069979	12	510	18	0.0103	0.0117	–0.5030		+	+
106	4.707	In	5441948	12	404	17	0.0108	0.0139	–0.9873		+	–
72	4.707	IR	5482021	12	379	37	0.0323	0.0323	–0.0092		–	–
114	2.997	In	6567455	12	300	3	0.0029	0.0033	–0.3785		+	–
115	2.710	In	6613211	12	398	10	0.0068	0.0083	–0.7445		+	+
116	2.710	In	6649164	12	512	34	0.0211	0.0220	–0.1771		+	+
117	2.579	IR	6703197	10	553	33	0.0219	0.0220	–0.0128		+	+
118	2.447	IR	6752435	12	540	13	0.0059	0.0080	–1.1050		–	–
119	2.447	In	6797217	12	297	26	0.0239	0.0290	–0.7798		+	+
120	2.178	In	6874455	12	469	32	0.0173	0.0226	–1.0541		+	+
122	2.178	IR	6964795	12	576	6	0.0017	0.0035	–1.8942	*	–	–
124	1.926	In	7041579	12	762	9	0.0035	0.0039	–0.4065		+	+
125	1.926	In	7092312	12	240	8	0.0104	0.0110	–0.2248		+	+
130	1.601	IR	7319723	12	553	21	0.0101	0.0126	–0.8639		+	+
136	1.461	In	7679367	12	371	27	0.0176	0.0241	–1.2139		+	+
137	1.461	In	7710260	12	454	15	0.0080	0.0109	–1.1372		+	+
138	1.486	In	7758526	12	338	9	0.0085	0.0088	–0.1382		+	+
139	1.486	In	7819831	12	347	14	0.0109	0.0134	–0.7805		+	+
150	1.930	In	8393030	12	305	15	0.0140	0.0163	–0.6133		+	+
153	2.441	In	8562010	12	475	25	0.0152	0.0174	–0.5739		+	+
157	2.725	IR	8763089	12	555	15	0.0071	0.0090	–0.8883		+	+
163	3.638	In	9040189	12	630	24	0.0085	0.0126	–1.4619		+	+
165	3.638	In	9149621	12	277	6	0.0049	0.0072	–1.1962		–	–
166	4.175	IR	9185460	12	606	11	0.0042	0.0060	–1.2359		–	–
173	3.536	IR	9587511	12	498	8	0.0032	0.0053	–1.5723		–	–
175	3.536	In	9724676	12	602	38	0.0186	0.0209	–0.5118		+	+
177	3.536	In	9798952	12	409	20	0.0144	0.0162	–0.4834		+	+
184	2.813	In	10123327	12	424	22	0.0149	0.0172	–0.5959		+	+
186	2.620	In	10222506	12	497	23	0.0101	0.0153	–1.5140		+	+

189	2.509	IR	10326177	12	541	17	0.0081	0.0104	–0.9757	+	+
191	2.509	In	10381498	12	432	4	0.0034	0.0031	0.3854	+	–
194	2.391	IR	10530546	12	578	17	0.0078	0.0097	–0.8472	+	+
196	2.402	In	10588263	12	596	5	0.0020	0.0028	–1.0213	–	–
197	2.402	In	10626957	12	547	7	0.0030	0.0042	–1.1251	+	+
201	2.545	IR	10820867	12	677	13	0.0052	0.0064	–0.8068	+	+
203	2.545	In	10897598	12	573	24	0.0131	0.0139	–0.2417	+	+
204	2.773	IR	10959318	12	533	25	0.0127	0.0155	–0.8185	+	+
205	3.000	In	11017914	12	645	24	0.0122	0.0123	–0.0553	+	+
209	3.000	In	11153913	12	483	42	0.0283	0.0288	–0.0791	–	–
212	3.000	In	11271565	12	688	12	0.0044	0.0058	–1.0133	+	+
214	3.282	In	11359194	12	588	17	0.0059	0.0096	–1.6764	–	–
215	3.440	In	11405680	11	568	17	0.0096	0.0102	–0.2697	+	–
216	3.440	In	11450047	12	603	42	0.0198	0.0239	–0.6458	+	–
217	3.440	In	11475426	12	537	5	0.0021	0.0031	–1.2237	–	–
221	3.588	In	11571743	12	380	18	0.0159	0.0157	0.0510	–	–
224	3.813	In	11717141	12	597	27	0.0115	0.0150	–1.0395	+	+
229	4.138	IR	11890206	12	422	14	0.0091	0.0110	–0.7386	+	+
231	4.138	In	11963993	11	520	43	0.0269	0.0282	–0.2180	+	+
241	4.138	In	12268167	12	568	6	0.0029	0.0035	–0.6727	–	–
248	4.512	In	12550032	11	656	23	0.0109	0.0120	–0.4222	–	–
249	4.512	IR	12582821	12	549	22	0.0089	0.0133	1.4600	+	–
250	4.689	IR	12633850	12	593	26	0.0132	0.0145	–0.3948	–	–
251	4.689	In	12677383	12	438	29	0.0228	0.0219	0.1715	–	–
254	4.689	IR	12791901	12	399	10	0.0065	0.0083	–0.8951	–	–
272	4.812	IR	13021801	12	500	20	0.0146	0.0133	0.4311	+	+
273	4.812	In	13027807	12	420	27	0.0221	0.0213	0.1737	+	+
276	4.877	In	13158630	12	326	10	0.0071	0.0102	–1.2528	–	–
278	4.925	IR	13244701	12	610	33	0.0161	0.0179	–0.4650	+	+
279	4.974	In	13277791	12	658	27	0.0133	0.0136	–0.0917	+	+
312	4.974	In	13354548	12	632	15	0.0069	0.0079	–0.5478	+	+
314	5.019	IR	13431378	12	538	21	0.0127	0.0129	–0.0861	+	+
326	5.001	IR	13850289	12	605	18	0.0078	0.0099	–0.9240	+	+
348	4.979	IR	14104733	12	571	25	0.0120	0.0145	–0.7696	+	+

367	4.979	IR	14233745	12	582	20	0.0108	0.0114	–0.2422	+	+
370	4.934	In	14324141	12	507	32	0.0201	0.0216	–0.3021	+	–
374	4.934	In	14435431	12	544	15	0.0071	0.0091	–0.9407	+	+
375	4.934	In	14470360	12	631	34	0.0176	0.0178	–0.0735	+	–
379	4.934	In	14573482	11	568	40	0.0202	0.0240	–0.7498	+	–
381	4.883	In	14674206	11	443	34	0.0275	0.0262	0.2230	+	+
384	4.833	In	14829173	12	502	19	0.0101	0.0125	–0.8458	–	–
385	4.833	In	14842173	12	525	23	0.0129	0.0145	–0.4998	+	–
393	4.718	In	15078750	12	559	11	0.0058	0.0065	–0.4743	+	+
282	4.624	In	15191599	12	682	33	0.0139	0.0160	–0.6092	+	–
287	4.330	IR	15434546	12	572	21	0.0104	0.0122	–0.6334	+	+
288	4.330	In	15445188	12	521	17	0.0103	0.0108	–0.2165	–	–
297	4.108	IR	15544018	12	630	8	0.0026	0.0042	–1.5723	–	+
299	4.108	In	15635880	12	618	18	0.0077	0.0096	–0.9018	+	+
301	3.928	In	15704066	12	556	26	0.0121	0.0155	–0.9762	+	+
303	3.928	IR	15802862	12	608	11	0.0049	0.0060	–0.8032	–	–
333	3.571	In	16183518	12	582	13	0.0077	0.0074	0.1772	–	–
331	3.372	In	16256960	12	552	24	0.0127	0.0144	–0.5129	+	–
330	3.372	IR	16278449	12	597	41	0.0254	0.0239	0.2744	+	+
329	3.318	IR	16336583	10	600	11	0.0045	0.0065	–1.3813	–	–
366	3.318	IR	16378516	11	610	38	0.0260	0.0213	1.0392	–	–
364	3.129	IR	16437205	12	606	22	0.0102	0.0120	–0.6695	+	+

APPENDIX 1.2 Nucleotide diversity estimates and test statistics for the European population

Sequences in EMBL database (http://www.ebi.ac.uk), accession numbers AJ568984-AJ571588 (complete set). Fragments are ordered from the telomere to the centromere; for each one, the following information is given:

r is the recombination rate expressed in rec/bp/gen x 10^{-8};

Type indicates if the fragment belongs to intergenic region (IR) or to an intron (In);

Absolute position is in base pairs, from the telomere;

n is the number of lines sequenced;

lth is the number of sites studied (excluding insertions and deletions polymorphism);

S is the number of segregating sites;

π is the nucleotide diversity (TAJIMA 1983);

θ_W is the WATTERSON (1975) estimate of nucleotide diversity;

Tajima's *D* test statistic (TAJIMA 1989a);

for the *H* and *K* haplotype statistics (DEPAULIS and VEUILLE 1998), it is indicated whether the observation is lower (–) or higher (+) than the simulated median across the sample (see text);

* $P < 0.05$;

** $P < 0.01$;

† one-tailed test without recombination, $P < 0.05$;

n.a. not applicable.

Fragment	*r*	Type	Abs. posit.	*n*	*lth*	*S*	π	θ_W	Tajima's *D*	*H*	*K*
10	0.486	In	1899930	12	348	0	0.0000	0.0000	n.a.	+	+
9	0.585	IR	1929751	12	326	0	0.0000	0.0000	n.a.	+	+
17	0.585	In	1946108	12	773	7	0.0038	0.0030	0.9980	–	–
6	0.436	IR	1988709	12	402	4	0.0045	0.0033	1.2302	+	–
1	0.811	In	2004307	12	381	5	0.0070	0.0044	2.2509 *	–	–
15	0.811	In	2010026	12	461	1	0.0012	0.0007	1.4862	+	+
22	0.811	In	2129973	12	630	8	0.0026	0.0042	–1.5723	–	–
26	1.051	IR	2140729	12	589	2	0.0018	0.0011	1.8244	+	+
18	1.587	In	2448658	12	500	9	0.0066	0.0060	0.3983	– †	–
4	1.587	IR	2455342	12	359	3	0.0038	0.0028	1.2725	+	–
5	2.019	In	2486993	12	248	0	0.0000	0.0000	n.a.	+	+
55	2.738	IR	3235896	11	660	16	0.0078	0.0083	–0.2440	–	–
54	2.738	IR	3238859	12	418	13	0.0085	0.0103	–0.7472	– †	–
57	3.138	IR	3333268	12	565	7	0.0042	0.0041	0.0515	–	–
60	3.138	IR	3448557	12	627	18	0.0102	0.0095	0.3059	–	–
56	3.290	In	3603702	12	325	4	0.0028	0.0041	–1.1032	–	–
76	3.290	IR	3653297	12	540	13	0.0115	0.0080	1.8917	–	–
78	3.549	IR	3727323	12	616	9	0.0068	0.0048	1.6572	–	–

81	3.883	IR	3879576	12	568	3	0.0018	0.0018	0.0217	+	+
84	3.883	IR	4018352	12	596	13	0.0085	0.0072	0.7437	–	–
85	3.883	IR	4069979	12	641	11	0.0029	0.0057	–2.0666 **	– †	– †
106	4.707	In	5441948	12	405	13	0.0118	0.0106	0.4753	–	–
72	4.707	IR	5482021	12	418	2	0.0008	0.0016	–1.4514	–	–
114	2.997	In	6567455	12	299	1	0.0006	0.0011	–1.1405	+	–
115	2.710	In	6613211	12	401	1	0.0004	0.0008	–1.1405	+	–
116	2.710	In	6649164	12	548	19	0.0098	0.0115	–0.6561	–	–
117	2.579	IR	6703197	10	583	18	0.0150	0.0109	1.7394	–	–
118	2.447	IR	6752435	12	554	4	0.0021	0.0024	–0.4192	+	+
119	2.447	In	6797217	12	297	5	0.0071	0.0056	1.0027	+	+
120	2.178	In	6874455	12	447	16	0.0136	0.0119	0.6460	–	–
122	2.178	IR	6964795	11	589	0	0.0000	0.0000	n.a.	+	+
124	1.926	In	7041579	12	763	4	0.0026	0.0017	1.7935	–	–
125	1.926	In	7092312	12	240	0	0.0000	0.0000	n.a.	+	+
130	1.601	IR	7319723	12	597	1	0.0003	0.0006	–1.1405	+	–
136	1.461	In	7679367	12	386	6	0.0085	0.0052	2.4683 **	–	– †
137	1.461	In	7710260	12	462	1	0.0011	0.0007	1.0659	+	+
138	1.486	In	7758526	12	346	0	0.0000	0.0000	n.a.	+	+
139	1.486	In	7819831	12	347	9	0.0101	0.0086	0.7285	–	–
150	1.930	In	8393030	12	328	0	0.0000	0.0000	n.a.	+	+
153	2.441	In	8562010	12	477	8	0.0046	0.0056	–0.6816	–	–
157	2.725	IR	8763089	12	580	1	0.0003	0.0006	–1.1405	+	–
163	3.638	In	9040189	12	641	9	0.0057	0.0047	0.9555	–	–
165	3.638	In	9149621	12	297	1	0.0018	0.0011	1.3811	+	+
166	4.175	IR	9185460	12	603	7	0.0043	0.0038	0.4608	–	–
173	3.536	IR	9587511	12	496	15	0.0130	0.0100	1.2596	–	–
175	3.536	In	9724676	12	622	8	0.0044	0.0043	0.1407	–	–
177	3.536	In	9798952	12	409	0	0.0000	0.0000	n.a.	+	+
184	2.813	In	10123327	12	431	10	0.0056	0.0077	–1.1398	–	–
186	2.620	In	10222506	12	482	11	0.0038	0.0076	–2.0666 **	– †	– †
189	2.509	IR	10326177	12	551	7	0.0025	0.0042	–1.6112	–	–
191	2.509	In	10381498	12	432	0	0.0000	0.0000	n.a.	+	+
194	2.391	IR	10530546	12	580	1	0.0009	0.0006	1.3811	+	+

196	2.402	In	10588263	12	623	1	0.0003	0.0005	–1.1405	+	–
197	2.402	In	10626957	12	547	0	0.0000	0.0000	n.a.	+	+
201	2.545	IR	10820867	12	679	7	0.0031	0.0034	–0.3578	+	–
203	2.545	In	10897598	12	574	1	0.0003	0.0006	–1.1405	+	–
204	2.773	IR	10959318	12	544	8	0.0038	0.0049	–0.8415	–	–
205	3.000	In	11017914	12	648	11	0.0068	0.0056	0.8756	–	–
209	3.000	In	11153913	11	499	22	0.0098	0.0151	–1.6012	–	–
212	3.000	In	11271565	12	690	2	0.0005	0.0010	–1.4514	+	–
214	3.282	In	11359194	11	521	2	0.0010	0.0013	–0.7782	+	–
215	3.440	In	11405680	12	566	12	0.0100	0.0070	1.7898	–	–
216	3.440	In	11450047	12	609	16	0.0112	0.0087	1.2263	–	–
217	3.440	In	11475426	12	506	2	0.0007	0.0013	–1.4514	–	–
221	3.588	In	11571743	12	386	8	0.0066	0.0069	–0.1563	– †	–
224	3.813	In	11717141	12	609	13	0.0087	0.0071	0.9822	–	–
229	4.138	IR	11890206	12	444	9	0.0034	0.0067	–2.0161 *	–	–
231	4.138	In	11963993	12	562	8	0.0040	0.0047	–0.6131	–	–
241	4.138	In	12268167	12	568	7	0.0035	0.0041	–0.5880	–	–
248	4.512	In	12550032	12	676	4	0.0028	0.0020	1.4716	+	+
249	4.512	IR	12582821	12	584	2	0.0010	0.0011	–0.2481	–	–
250	4.689	IR	12633850	11	584	6	0.0019	0.0035	–1.8506 *	–	–
251	4.689	In	12677383	12	452	5	0.0018	0.0037	–1.8309 *	– †	– †
254	4.689	IR	12791901	12	443	3	0.0022	0.0022	–0.0283	–	–
272	4.812	IR	13021801	12	486	16	0.0133	0.0109	0.9547	–	–
273	4.812	In	13027807	12	430	20	0.0222	0.0154	1.9384	–	– †
276	4.877	In	13158630	12	326	2	0.0014	0.0020	–0.8497	+	–
278	4.925	IR	13244701	12	612	19	0.0142	0.0103	1.6837	–	–
279	4.974	In	13277791	12	664	12	0.0086	0.0060	1.8379	–	–
312	4.974	In	13354548	12	637	5	0.0022	0.0026	–0.6165	–	–
314	5.019	IR	13431378	12	551	1	0.0003	0.0006	–1.1405	+	–
326	5.001	IR	13850289	12	611	11	0.0059	0.0060	–0.0763	–	–
348	4.979	IR	14104733	12	571	10	0.0083	0.0058	1.7967	–	–
367	4.979	IR	14233745	12	595	2	0.0010	0.0011	–0.3818	+	–
370	4.934	In	14324141	12	570	28	0.0173	0.0163	0.2715	–	–
374	4.934	In	14435431	12	607	0	0.0000	0.0000	n.a.	+	+

375	4.934	In	14470360	12	633	2	0.0009	0.0011	–0.3818	+	–
379	4.934	In	14573482	12	584	11	0.0082	0.0062	1.2737	–	–
381	4.883	In	14674206	12	429	0	0.0000	0.0000	n.a.	+	+
384	4.833	In	14829173	12	494	7	0.0024	0.0047	–1.9437 *	– †	– †
385	4.833	In	14842173	12	515	6	0.0053	0.0039	1.4213	–	–
393	4.718	In	15078750	12	560	0	0.0000	0.0000	n.a.	+	+
282	4.624	In	15191599	12	700	4	0.0010	0.0019	–1.7469	– †	–
287	4.330	IR	15434546	12	573	10	0.0043	0.0058	–1.0457	–	–
288	4.330	In	15445188	12	519	4	0.0016	0.0026	–1.3848	–	–
297	4.108	IR	15544018	12	628	5	0.0013	0.0026	–1.8309 *	+	–
299	4.108	In	15635880	12	617	6	0.0030	0.0032	–0.2946	–	–
301	3.928	In	15704066	12	573	6	0.0027	0.0035	–0.8472	–	–
303	3.928	IR	15802862	12	608	5	0.0016	0.0027	–1.5273	– †	–
333	3.571	In	16183518	12	578	7	0.0041	0.0040	0.1283	–	–
331	3.372	In	16256960	12	592	8	0.0042	0.0045	–0.2248	–	–
330	3.372	IR	16278449	12	576	21	0.0159	0.0121	1.4120	–	–
329	3.318	IR	16336583	12	599	7	0.0059	0.0039	2.0467 *	–	–
366	3.318	IR	16378516	12	610	16	0.0139	0.0087	2.5968 **	– †	– †
364	3.129	IR	16437205	12	604	13	0.0063	0.0071	–0.4788	– †	–

APPENDIX 1.3 Divergence between *D. melanogaster* populations and *D. simulans*

Sequences in EMBL database (http://www.ebi.ac.uk), accession numbers AJ568984-AJ571588 (complete set). Fragments are ordered from the telomere to the centromere; for each one, the following information is given:

Absolute position is in base pairs, from the telomere;

n is the number of lines sequenced in each population [Afr (African) and Eur (European)];

S is the number of segregating sites;

D_s is the number of fixed differences between *D. melanogaster* and *D. simulans*;

K is the divergence between *D. melanogaster* and *D. simulans*;

F_{ST} is the fixation index between the African and European populations (HUDSON *et al*. 1992);

the total contribution to the multi-locus HKA test is also shown (HUDSON *et al*. 1987; see text).

		n		*S*		D_s		*K*			HKA contr.	
Fragment	Abs. posit.	Afr	Eur	Afr	Eur	Afr	Eur	Afr	Eur	F_{ST}	Afr	Eur
10	1899930	12	12	10	0	22	26	0.0826	0.0808	0.5893	4.06	0.95
9	1929751	12	12	2	0	11	11	0.0356	0.0340	0.2121	2.07	2.44
17	1946108	12	12	15	7	39	40	0.0561	0.0554	0.2797	0.32	1.44
6	1988709	12	12	6	4	33	33	0.0863	0.0863	0.1021	0.95	4.10
1	2004307	12	12	15	5	23	25	0.0719	0.0719	0.2442	0.09	0.09
15	2010026	12	12	2	1	13	13	0.0287	0.0293	0.1299	1.00	3.05
22	2129973	12	12	11	8	22	25	0.0405	0.0456	0.3858	0.18	0.62
26	2140729	12	12	18	2	16	21	0.0346	0.0376	0.1053	1.10	0.42
18	2448658	12	12	13	9	38	40	0.0878	0.0904	0.5391	0.04	1.97
4	2455342	12	12	8	3	13	14	0.0417	0.0421	0.1977	0.03	0.12
5	2486993	12	12	14	0	13	18	0.0766	0.0793	0.5025	3.21	0.17
55	3235896	12	11	32	16	33	35	0.0635	0.0655	0.3099	2.09	0.16
54	3238859	12	12	33	13	23	26	0.0734	0.0760	0.3147	2.27	1.27
57	3333268	11	12	12	7	23	26	0.0475	0.0530	0.4301	0.02	0.38
60	3448557	12	12	30	18	49	55	0.0959	0.0986	0.1620	0.38	0.22
56	3603702	12	12	5	4	6	8	0.0231	0.0291	0.3980	1.07	0.01
76	3653297	12	12	33	13	23	25	0.0584	0.0568	0.1425	2.52	1.27
78	3727323	12	12	23	9	28	31	0.0568	0.0562	0.0922	0.08	0.00
81	3879576	12	12	19	3	18	20	0.0394	0.0384	0.3116	0.48	0.23
84	4018352	11	12	21	13	26	26	0.0534	0.0556	0.3002	2.11	0.00
85	4069979	12	12	18	11	25	33	0.0606	0.0596	0.3627	0.29	0.05
106	5441948	12	12	17	13	23	24	0.0668	0.0677	0.1092	2.79	0.01
72	5482021	12	12	37	2	14	27	0.0691	0.0735	0.2648	1.98	4.24

114	6567455	12	12	3	1	24	24	0.0851	0.0848	0.1039	2.50	5.05
115	6613211	12	12	10	1	26	26	0.0733	0.0713	0.1364	2.85	1.58
116	6649164	12	12	34	19	17	25	0.0568	0.0627	0.1390	6.80	2.87
117	6703197	9	10	33	18	36	45	0.0853	0.0905	0.1156	1.50	0.22
118	6752435	12	12	13	4	13	17	0.0285	0.0322	0.3812	0.00	0.18
119	6797217	12	12	26	5	26	28	0.1245	0.1214	0.2178	0.58	0.02
120	6874455	12	12	32	16	17	19	0.0525	0.0579	0.2197	8.06	2.57
122	6964795	12	11	6	0	8	10	0.0148	0.0172	0.7857	1.86	0.00
124	7041579	12	12	9	4	16	19	0.0260	0.0282	0.3092	0.04	0.25
125	7092312	12	12	8	0	13	16	0.0667	0.0678	0.6809	2.78	0.14
130	7319723	12	12	21	1	30	33	0.0636	0.0575	0.6772	3.37	0.05
136	7679367	12	12	27	6	20	26	0.0725	0.0784	0.2753	0.04	0.89
137	7710260	12	12	15	1	30	36	0.0835	0.0894	0.1442	3.97	0.90
138	7758526	12	12	9	0	15	17	0.0642	0.0623	0.3831	3.51	0.62
139	7819831	12	12	14	9	24	24	0.0872	0.0875	0.0775	0.47	0.37
150	8393030	12	12	15	0	15	20	0.0671	0.0627	0.1760	3.28	0.16
153	8562010	12	12	25	8	17	24	0.0152	0.0568	0.4447	0.25	1.18
157	8763089	12	12	15	1	7	15	0.0320	0.0479	0.6727	3.01	0.36
163	9040189	12	12	24	9	43	43	0.0130	0.0771	0.2607	0.14	0.51
165	9149621	12	12	6	1	26	30	0.1094	0.1120	0.3316	3.25	3.48
166	9185460	12	12	11	7	20	20	0.0362	0.0374	0.1493	0.39	0.33
173	9587511	12	12	8	15	13	13	0.0305	0.0370	0.3618	12.13	0.12
175	9724676	12	12	38	8	17	22	0.0470	0.0445	0.2549	0.37	3.31
177	9798952	12	12	20	0	27	27	0.0812	0.0700	0.4106	4.09	0.03
184	10123327	12	12	22	10	13	17	0.0542	0.0522	0.2388	2.64	1.66
186	10222506	12	12	23	11	30	30	0.0671	0.0764	0.5516	0.69	0.00
189	10326177	12	12	17	7	22	25	0.0470	0.0494	0.3434	0.03	0.01
191	10381498	12	12	4	0	40	41	0.0957	0.0949	0.1984	4.98	6.33
194	10530546	12	12	17	1	17	20	0.0502	0.0502	0.2010	2.99	0.03
196	10588263	12	12	5	1	14	18	0.0250	0.0308	0.5943	1.68	1.15
197	10626957	12	12	7	0	20	20	0.0393	0.0387	0.0992	3.36	1.60
201	10820867	12	12	13	7	18	18	0.0313	0.0306	0.4618	0.70	0.01
203	10897598	12	12	24	1	26	31	0.0635	0.0616	0.6586	3.52	0.02
204	10959318	12	12	25	8	41	46	0.0930	0.0935	0.1437	0.40	0.25

205	11017914	12	12	24	11	23	25	0.0480	0.0458	0.2204	1.12	0.18
209	11153913	12	11	42	22	34	43	0.0967	0.1063	0.4226	3.44	0.78
212	11271565	12	12	12	2	28	29	0.0465	0.0454	0.3519	2.12	1.20
214	11359194	12	11	17	2	34	33	0.0701	0.0677	0.1037	2.33	0.96
215	11405680	11	12	17	12	37	38	0.0802	0.0825	0.3993	0.14	0.98
216	11450047	12	12	42	16	26	30	0.0629	0.0622	0.1243	2.70	1.68
217	11475426	12	12	5	2	8	7	0.0159	0.0143	–0.0227	0.03	0.09
221	11571743	12	12	18	8	20	24	0.0681	0.0799	0.4366	0.23	0.03
224	11717141	12	12	27	13	15	18	0.0360	0.0392	0.0760	4.79	2.02
229	11890206	12	12	14	9	12	13	0.0354	0.0333	0.1226	3.47	0.39
231	11963993	11	12	43	8	57	73	0.1404	0.1482	0.5450	2.11	0.03
241	12268167	12	12	6	7	23	24	0.0449	0.0473	0.5716	0.08	2.53
248	12550032	11	12	23	4	31	32	0.0597	0.0539	0.2011	1.04	0.02
249	12582821	12	12	22	2	25	30	0.0565	0.0561	0.1585	2.27	0.01
250	12633850	12	11	26	6	46	51	0.0897	0.0904	0.6850	1.15	0.35
251	12677383	12	12	29	5	27	36	0.0913	0.0953	0.2703	0.85	0.29
254	12791901	12	12	10	3	12	17	0.0356	0.0408	0.5000	0.20	0.01
272	13021801	12	12	20	16	14	12	0.0399	0.0385	0.0563	14.46	1.09
273	13027807	12	12	27	20	20	20	0.0785	0.0766	0.1297	9.15	0.58
276	13158630	12	12	10	2	18	21	0.0803	0.0833	0.4516	1.53	0.68
278	13244701	12	12	33	19	19	24	0.0482	0.0538	0.1241	7.17	1.91
279	13277791	12	12	27	12	10	12	0.0280	0.0270	0.1725	7.66	3.86
312	13354548	12	12	15	5	14	15	0.0282	0.0280	0.3191	0.24	0.29
314	13431378	12	12	21	1	28	31	0.0626	0.0605	0.2897	3.30	0.02
326	13850289	12	12	18	11	21	22	0.0395	0.0406	0.1784	2.16	0.03
348	14104733	12	12	25	10	28	29	0.0654	0.0646	0.1694	0.29	0.02
367	14233745	12	12	20	2	32	37	0.0694	0.0635	0.4698	2.68	0.16
370	14324141	12	12	32	28	28	33	0.0775	0.0779	0.1453	9.97	0.43
374	14435431	12	12	15	0	15	21	0.0345	0.0356	0.2727	3.39	0.19
375	14470360	12	12	34	2	32	41	0.0671	0.0695	0.5831	3.09	0.31
379	14573482	11	12	40	11	41	48	0.0934	0.0914	0.2030	0.03	0.22
381	14674206	11	12	34	0	14	28	0.0598	0.0729	0.5099	4.37	2.94
384	14829173	12	12	19	7	42	43	0.0999	0.1062	0.5035	0.68	1.25
385	14842173	12	12	23	6	30	35	0.0708	0.0774	0.1489	0.39	0.01

393	15078750	12	12	11	0	39	40	0.0739	0.0719	0.2255	4.94	2.62
282	15191599	12	12	33	4	40	51	0.0765	0.0806	0.4207	2.63	0.00
287	15434546	12	12	21	10	18	20	0.0435	0.0435	0.3200	1.68	0.39
288	15445188	12	12	17	4	28	34	0.0674	0.0701	0.3636	1.17	0.27
297	15544018	12	12	8	5	9	9	0.0160	0.0155	0.0133	1.56	0.03
299	15635880	12	12	18	6	22	24	0.0439	0.0440	0.1654	0.00	0.00
301	15704066	12	12	26	6	33	41	0.0793	0.0863	0.1332	0.96	0.03
303	15802862	12	12	11	5	34	34	0.0599	0.0604	0.2310	0.57	1.96
333	16183518	12	12	13	7	26	28	0.0525	0.0528	0.0327	0.00	0.57
331	16256960	12	12	24	8	29	33	0.0130	0.0641	0.3885	0.01	0.00
330	16278449	12	12	41	21	43	45	0.1009	0.0979	0.1057	1.78	0.02
329	16336583	10	12	11	7	24	23	0.0478	0.0457	0.2477	0.07	0.73
366	16378516	11	12	38	16	41	47	0.0905	0.0895	0.2631	0.54	0.16
364	16437205	12	12	22	13	13	13	0.0283	0.0305	0.2656	8.40	1.60

APPENDIX 2.1 Nucleotide diversity estimates and test statistics for the African population

Fragments are ordered from the telomere to the centromere; for each one, the following information is given:

r is the recombination rate expressed in rec/bp/gen x 10^{-8};

Type indicates if the fragment belongs to intergenic region (IR) or to an intron (In);

Absolute position is in base pairs, from the telomere;

n is the number of lines sequenced;

lth is the number of sites studied (excluding insertions and deletions polymorphism);

S is the number of segregating sites;

π is the nucleotide diversity (TAJIMA 1983);

θ_W is the WATTERSON (1975) estimate of nucleotide diversity;

Tajima's *D* test statistic (TAJIMA 1989a);

for the *H* and *K* haplotype statistics (DEPAULIS and VEUILLE 1998), it is indicated whether the observation is lower (–) or higher (+) than the simulated median across the sample (see text);

Z_{nS} (KELLY 1997) is linkage disequilibrium;

* $P < 0.05$;

n.a. not applicable.

Fragment	*r*	Type	Abs. posit.	*n*	*lth*	*S*	π	θ_W	Tajima's *D*		*H*	*K*	Z_{nS}	
419	0.039	IR	1571015	12	606	5	0.0014	0.0027	–1.6545	*	–	–	0.3388	
12	0.486	In	1863523	12	379	4	0.0026	0.0035	–0.8126		+	–	n.a.	
10	0.486	In	1899930	12	346	10	0.0115	0.0096	0.7453		+	–	0.2395	
17	0.585	In	1946108	12	781	15	0.0056	0.0064	–0.4532		+	+	0.1467	
1	0.811	In	2004307	12	380	15	0.0130	0.0131	–0.0099		+	+	0.1924	
22	0.811	In	2129973	12	618	11	0.0042	0.0059	–1.0957		+	+	0.0617	*
25	1.051	IR	2137479	12	595	13	0.0041	0.0072	–1.6769	*	+	+	0.0593	*
26	1.051	IR	2140729	12	570	18	0.0064	0.0105	–1.5670	*	+	+	0.0693	*
32	1.291	IR	2188201	12	626	12	0.0054	0.0063	–0.5741		+	+	0.0923	
38	1.291	In	2270372	10	394	16	0.0135	0.0144	–0.2393		+	+	0.1623	
18	1.587	In	2448658	12	502	13	0.0073	0.0086	–0.5736		+	+	0.1358	*
5	2.019	In	2486993	12	245	14	0.0186	0.0189	–0.0634		+	+	0.1125	
45	2.450	IR	2740398	11	480	14	0.0083	0.0010	–0.6850		+	+	0.1583	*
46	2.450	IR	2781118	12	586	27	0.0155	0.0153	0.0760		+	+	0.1079	*
55	2.738	IR	3235896	12	661	32	0.0137	0.0160	–0.6072		+	–	0.1367	
54	2.738	IR	3238859	12	418	33	0.0209	0.0261	–0.8308		+	+	0.1042	
57	3.137	In	3333268	11	547	12	0.0068	0.0075	–0.3387		+	+	0.0800	
60	3.137	In	3448557	12	615	29	0.0155	0.0156	–0.0313		+	+	0.1170	
56	3.290	In	3603702	12	325	5	0.0056	0.0051	0.3574		–	–	0.3151	*

76	3.290	IR	3653297	12	538	33	0.0161	0.0203	–0.8711		+	+	0.1326	
77	3.549	In	3680710	12	556	29	0.0145	0.0173	–0.6693		+	+	0.1791	*
78	3.549	IR	3727323	12	612	23	0.0102	0.0124	–0.7337		+	+	0.1031	*
80	3.716	IR	3839129	12	568	32	0.0196	0.0187	0.2114		+	+	0.1116	
81	3.883	IR	3879576	12	561	19	0.0116	0.0112	0.1425		+	+	0.1101	*
462	3.549	In	3918527	12	668	7	0.0028	0.0035	–0.6497		+	+	0.0149	
84	3.883	IR	4018352	11	596	21	0.0010	0.0120	–0.7087		+	+	0.1037	
85	3.883	IR	4069979	12	510	18	0.0103	0.0117	–0.4607		+	+	0.1314	*
66	4.369	In	4260258	12	352	16	0.0145	0.0151	–0.1546		+	+	0.1057	
67	4.369	In	4512472	12	633	24	0.0102	0.0126	–0.7507		+	+	0.1183	
90	4.545	In	4896054	12	419	31	0.0231	0.0245	–0.2354		+	+	0.1126	
91	4.578	IR	4952503	10	471	23	0.0134	0.0173	–0.9523		+	+	0.1384	*
93	4.611	In	5034343	12	391	9	0.0045	0.0076	–1.5345	*	+	+	0.0579	*
94	4.667	IR	5091045	10	505	22	0.0117	0.0154	–1.0309		+	+	0.1128	
95	4.723	IR	5136167	11	560	19	0.0098	0.0116	–0.6210		+	+	0.1111	*
106	4.707	In	5441948	12	404	17	0.0108	0.0139	–0.9040		+	–	0.0931	
72	4.707	IR	5482021	12	379	37	0.0323	0.0323	–0.0085		–	–	0.4095	
73	4.634	IR	5555609	12	574	10	0.0047	0.0058	–0.6784		+	+	0.1032	
109	4.492	In	5730972	11	582	13	0.0064	0.0076	–0.6542		+	+	0.1094	
114	2.997	In	6567455	12	300	3	0.0029	0.0033	–0.3397		+	–	0.0182	*
115	2.710	In	6613211	12	398	10	0.0068	0.0083	–0.6784		+	+	0.0365	
116	2.710	In	6649164	12	512	34	0.0211	0.0220	–0.1628		+	+	0.1697	*
117	2.579	IR	6703197	9	553	33	0.0219	0.0220	–0.0116		+	+	0.1491	*
118	2.447	IR	6752435	12	540	13	0.0059	0.0080	–1.0095		–	–	0.0833	*
119	2.447	In	6797217	12	297	26	0.0239	0.0290	–0.7160		+	+	0.1136	
120	2.178	In	6874455	12	469	32	0.0173	0.0226	–0.9690		+	+	0.1332	
122	2.178	IR	6964795	12	576	6	0.0017	0.0035	–1.7158	*	–	–	0.0744	
502	1.926	IR	6991699	12	537	24	0.0131	0.0148	–0.4707		+	+	0.0947	
124	1.926	In	7041579	12	762	9	0.0035	0.0039	–0.3640		+	+	0.1057	
125	1.926	In	7092312	12	240	7	0.0092	0.0097	–0.1854		+	+	0.1101	*
126	1.926	In	7143267	12	585	39	0.0190	0.0221	–0.5911		+	+	0.1028	*
530	1.601	IR	7316457	12	505	26	0.0159	0.0170	–0.2687		+	+	0.1520	
130	1.601	IR	7319723	12	553	21	0.0101	0.0126	–0.7922		+	+	0.0907	
133	1.601	In	7469487	11	621	15	0.0069	0.0082	–0.6834		+	+	0.1433	*

136	1.461	In	7679367	12	371	27	0.0176	0.0241	–1.1149		+	+	0.1228	*
137	1.461	In	7710260	12	453	14	0.0074	0.0102	–1.0832		+	+	0.1058	
138	1.486	In	7758526	12	338	9	0.0085	0.0088	–0.1258		+	+	0.1209	
139	1.486	In	7819831	12	347	14	0.0109	0.0134	–0.7135		+	+	0.1343	*
143	1.486	In	8068685	12	519	19	0.0103	0.0121	–0.6012		+	+	0.1161	*
150	1.930	In	8393030	12	305	15	0.0140	0.0163	–0.5610		+	+	0.0968	*
153	2.441	In	8562010	12	475	25	0.0152	0.0174	–0.5269		+	+	0.1009	
157	2.725	In	8763089	12	310	12	0.0114	0.0128	–0.4133		+	+	0.1310	*
160	3.009	In	8897903	12	199	13	0.0135	0.0216	–1.4590	*	–	–	n.a.	
163	3.638	In	9040189	12	630	24	0.0085	0.0126	–1.3418		+	+	0.0927	
165	3.638	In	9149621	12	277	6	0.0049	0.0072	–1.0835		–	–	0.2331	
166	4.175	IR	9185460	12	606	11	0.0042	0.0060	–1.1272		–	–	0.1106	
167	4.175	In	9228779	12	607	12	0.0050	0.0065	–0.8812		+	+	0.1101	
169	4.175	In	9367972	12	308	2	0.0015	0.0022	–0.7584		–	–	0.0182	*
170	4.508	IR	9409076	10	517	35	0.0231	0.0239	–0.1482		+	+	0.1228	
173	3.536	In	9587511	12	498	8	0.0032	0.0053	–1.4292	*	–	–	0.0647	*
446	3.536	In	9660426	12	262	16	0.0203	0.0202	0.0149		+	+	0.1512	
175	3.536	In	9724676	12	603	38	0.0185	0.0209	–0.4708		+	+	0.0904	
177	3.536	In	9798952	12	409	20	0.0144	0.0162	–0.4431		+	+	0.1324	*
178	3.536	IR	9839303	12	493	34	0.0180	0.0228	–0.8739		+	+	0.1015	
179	2.813	In	9887176	12	545	26	0.0152	0.0158	–0.1677		+	+	0.1395	*
182	2.813	In	10046886	12	458	16	0.0084	0.0116	–1.0699		+	+	n.a.	
464	2.813	IR	10051437	12	548	29	0.0170	0.0175	–0.1224		+	+	0.1294	
465	2.813	In	10091962	12	449	20	0.0164	0.0148	0.4597		+	+	0.2395	
184	2.813	In	10123327	12	424	22	0.0149	0.0172	–0.5466		+	+	0.1003	*
186	2.620	In	10222506	12	497	21	0.0094	0.0140	–1.3205		+	+	0.0782	
187	2.509	IR	10250800	12	522	14	0.0077	0.0089	–0.5350		+	+	0.1230	*
188	2.509	In	10274029	12	491	4	0.0016	0.0027	–1.2476		+	+	0.0132	*
189	2.509	In	10326177	12	541	17	0.0081	0.0104	–0.8933		+	+	0.0952	
190	2.509	IR	10344630	11	525	24	0.0131	0.0156	–0.6845		+	+	n.a.	
191	2.509	In	10381498	12	432	4	0.0034	0.0031	0.3472		+	–	0.4444	*
470	2.509	IR	10407332	11	606	6	0.0018	0.0034	–1.6686	*	–	–	0.1420	
192	2.509	IR	10432451	10	418	17	0.0128	0.0144	–0.4745		+	+	n.a.	
472	2.391	IR	10499785	12	553	15	0.0071	0.0090	–0.8126		+	+	0.1549	

194	2.391	In	10530546	12	578	17	0.0078	0.0097	–0.7757		+	+	0.0936	
195	2.391	In	10553628	12	508	34	0.0199	0.0222	–0.4260		–	–	0.1688	
473	2.397	IR	10572196	12	530	11	0.0078	0.0069	0.5303		+	+	0.1696	
196	2.402	In	10588263	12	596	5	0.0020	0.0028	–0.9229		–	–	0.0372	*
197	2.402	In	10626957	12	547	7	0.0030	0.0042	–1.0211		+	+	0.0083	
198	2.402	In	10672053	12	662	24	0.0097	0.0120	–0.7818		+	+	n.a.	
475	2.402	IR	10692810	12	619	9	0.0035	0.0048	–1.0274		+	+	0.2281	
743	2.402	IR	10764628	11	298	13	0.0084	0.0149	–1.7307	*	+	+	0.1134	*
201	2.545	IR	10820867	12	677	13	0.0051	0.0064	–0.7371		+	+	0.1330	*
477	2.545	IR	10833367	12	647	32	0.0140	0.0164	–0.6012		+	+	0.1331	*
480	2.545	IR	10881373	12	626	16	0.0050	0.0085	–1.6236	*	+	+	0.0548	
203	2.545	In	10897598	12	573	24	0.0131	0.0139	–0.2218		+	+	0.0935	*
532	2.545	IR	10915182	12	457	4	0.0031	0.0029	0.1660		–	–	0.5152	
204	2.773	IR	10959318	12	533	25	0.0127	0.0155	–0.7514		+	+	0.1009	*
205	3.000	In	11017914	12	645	24	0.0122	0.0123	–0.0507		+	+	0.1003	
483	3.000	IR	11040319	12	656	37	0.0158	0.0187	–0.6387		+	+	0.1198	*
206	3.000	IR	11058202	12	527	31	0.0150	0.0195	–0.9628		+	+	0.1321	*
207	3.000	IR	11087062	12	490	35	0.0229	0.0237	–0.1249		+	+	0.1116	
208	3.000	IR	11114429	12	500	10	0.0062	0.0066	–0.2496		+	+	0.1011	
209	3.000	In	11153913	12	483	42	0.0283	0.0288	–0.0728		–	–	0.1478	
210	3.000	In	11190222	12	661	21	0.0082	0.0105	–0.8978		+	+	n.a.	
211	3.000	In	11227737	12	585	28	0.0147	0.0159	–0.2890		+	+	0.1231	*
212	3.000	In	11271565	12	688	12	0.0044	0.0058	–0.9251		–	–	0.0681	
213	3.282	In	11307249	12	576	11	0.0058	0.0063	–0.3379		+	+	0.1314	
214	3.282	In	11359194	12	588	17	0.0059	0.0096	–1.5350	*	–	–	0.1860	
215	3.440	In	11405680	11	568	17	0.0096	0.0102	–0.2459		+	–	0.1872	*
216	3.440	In	11450047	12	603	42	0.0198	0.0231	–0.5944		+	–	0.1183	
217	3.440	In	11475426	12	537	5	0.0021	0.0031	–1.1058		–	–	0.0860	
218	3.440	In	11492521	12	341	10	0.0077	0.0097	–0.7814		+	+	0.1062	
219	3.588	In	11542518	11	577	12	0.0064	0.0071	–0.4088		–	+	n.a.	
220	3.588	In	11562312	12	411	2	0.0014	0.0016	–0.3407		–	–	0.0303	
221	3.588	In	11571743	12	380	18	0.0159	0.0157	0.0468		–	–	0.1706	*
222	3.813	IR	11614496	10	504	28	0.0181	0.0196	–0.3364		+	+	n.a.	
488	3.813	IR	11642234	10	593	31	0.0172	0.0185	–0.3029		+	+	0.1168	*

224	3.813	In	11717141	12	599	27	0.0115	0.0149	–0.9547		+	+	0.0873	
660	3.813	In	11733290	10	368	5	0.0046	0.0048	–0.1594		+	+	n.a.	
228	4.138	IR	11846104	11	408	8	0.0039	0.0067	–1.5509	*	–	–	n.a.	
492	4.138	IR	11872761	12	649	24	0.0096	0.0122	–0.8751		+	+	0.1067	
229	4.138	IR	11890206	12	422	14	0.0091	0.0110	–0.6752		+	+	0.1319	*
493	4.138	IR	11949243	12	612	26	0.0118	0.0141	–0.6727		+	+	0.1545	*
231	4.138	In	11963993	11	520	43	0.0269	0.0282	–0.1998		+	+	0.1215	
232	4.138	IR	11986343	12	546	15	0.0074	0.0091	–0.7407		–	+	0.1518	
233	4.138	In	12043219	12	441	32	0.0223	0.0240	–0.3047		–	–	0.1625	
235	4.138	In	12080905	12	507	10	0.0045	0.0065	–1.1587		+	+	0.1182	
237	4.138	IR	12134325	12	497	46	0.0271	0.0306	–0.4918		+	+	n.a.	
447	4.138	IR	12164349	12	579	19	0.0068	0.0109	–1.4898	*	+	+	0.1080	
239	4.138	In	12202152	11	310	40	0.0482	0.0441	0.4054		+	+	0.1463	
241	4.138	In	12268167	12	568	6	0.0029	0.0035	–0.6094		–	–	0.1376	
242	4.138	IR	12309945	12	467	33	0.0217	0.0234	–0.3008		+	+	0.1473	
721	4.436	IR	12387758	12	335	28	0.0279	0.0277	0.0273		+	+	0.3415	
245	4.436	In	12439737	12	432	20	0.0143	0.0153	–0.2681		+	+	0.2061	
246	4.436	In	12490225	12	448	18	0.0123	0.0133	–0.2983		+	+	n.a.	
248	4.512	In	12550100	11	656	23	0.0109	0.0120	–0.3857		–	–	0.1582	
249	4.512	IR	12582821	12	549	22	0.0089	0.0133	–1.3479		+	–	0.1141	
250	4.689	IR	12633850	12	593	26	0.0132	0.0145	–0.3625		–	–	0.2315	
251	4.689	In	12677383	12	438	29	0.0228	0.0219	0.1576		–	–	0.2403	
252	4.689	In	12709653	12	428	26	0.0177	0.0201	–0.4996		–	+	0.1626	
253	4.689	IR	12752462	11	467	20	0.0125	0.0146	–0.6008		+	+	0.1601	
254	4.689	IR	12791901	12	399	10	0.0065	0.0083	–0.8157		–	–	0.0920	
258	4.812	In	12886899	12	417	29	0.0228	0.0230	–0.0443		+	+	0.1429	
259	4.812	In	12938544	12	289	18	0.0153	0.0206	–1.0392		–	–	0.1909	
260	4.812	In	12978603	10	554	17	0.0087	0.0108	–0.8177		+	+	0.1531	
272	4.812	IR	13022057	12	506	20	0.0149	0.0131	0.5426		+	+	0.2831	
273	4.812	In	13027807	12	420	26	0.0214	0.0205	0.1787		+	+	0.1431	
722	4.812	In	13090086	12	305	26	0.0260	0.0282	–0.3264		+	+	0.1301	
276	4.877	In	13158630	12	326	10	0.0071	0.0102	–1.1416		–	–	0.0937	*
277	4.877	IR	13194383	12	599	37	0.0201	0.0205	–0.0808		+	+	0.1017	
278	4.925	In	13244701	12	610	33	0.0161	0.0179	–0.4275		+	+	0.1292	

279	4.974	In	13277791	12	658	27	0.0133	0.0136	–0.0842		+	+	0.1354	
280	4.974	In	13311226	12	294	18	0.0175	0.0203	–0.5419		+	+	0.1516	
311	4.974	IR	13315520	12	215	21	0.0209	0.0323	–1.4350		+	+	0.1203	
450	4.974	IR	13323368	12	663	35	0.0126	0.0175	–1.1652		+	+	0.1160	
312	4.974	In	13354548	12	632	15	0.0069	0.0079	–0.5011		+	+	0.1100	
313	4.974	In	13394848	12	456	9	0.0049	0.0065	–0.9147		+	+	0.0961	
314	5.019	In	13431378	12	565	21	0.0121	0.0123	–0.0789		+	+	0.1290	
318	5.026	In	13574274	10	325	10	0.0120	0.0109	0.4258		+	+	0.1726	*
319	5.026	In	13596773	12	489	27	0.0143	0.0183	–0.8851		+	+	0.0850	
320	5.026	In	13636642	11	433	15	0.0097	0.0118	–0.7121		+	+	0.1210	
321	5.025	IR	13668913	12	559	9	0.0043	0.0053	–0.7081		+	+	0.2528	
323	5.023	In	13757364	12	372	19	0.0092	0.0169	–1.8181	*	+	+	0.1228	*
325	5.023	IR	13810837	12	528	8	0.0043	0.0050	–0.4950		+	+	0.0000	*
326	5.001	IR	13850289	12	605	18	0.0078	0.0099	–0.8464		+	+	0.0755	*
745	4.979	IR	13923319	11	371	2	0.0001	0.0018	–1.2691	*	–	–	0.0100	
342	5.023	IR	13933460	11	527	35	0.0215	0.0227	–0.2160		+	+	n.a.	
344	4.979	In	13982020	11	510	29	0.0180	0.0194	–0.3153		+	+	0.1326	
346	4.979	IR	14039423	10	492	10	0.0070	0.0072	–0.0772		+	+	0.1603	
348	4.979	In	14104733	12	571	25	0.0120	0.0145	–0.7065		+	+	0.1192	
350	4.979	IR	14210996	12	452	13	0.0072	0.0095	–0.9278		+	+	0.1293	*
367	4.979	IR	14233745	12	582	20	0.0108	0.0114	–0.2221		+	+	0.0968	
368	4.957	IR	14266922	12	505	26	0.0129	0.0170	–0.9974		+	+	0.1266	*
369	4.957	IR	14303687	12	675	9	0.0035	0.0044	–0.8020		+	+	0.0631	
370	4.934	In	14324141	12	507	33	0.0201	0.0216	–0.2778		+	–	0.1209	
371	4.934	In	14355845	11	532	24	0.0144	0.0154	–0.2664		+	+	0.1228	
373	4.934	In	14423857	11	531	9	0.0042	0.0058	–1.0591		+	+	0.1797	
374	4.934	In	14435431	12	544	15	0.0071	0.0091	–0.8605		+	+	0.1360	*
375	4.934	In	14470360	12	631	34	0.0176	0.0178	–0.0676		+	–	0.1073	
376	4.934	In	14502591	10	259	16	0.0238	0.0218	0.3709		+	+	0.1537	
378	4.934	In	14535555	12	518	22	0.0161	0.0141	0.6004		+	+	0.1109	*
379	4.934	In	14573482	11	568	40	0.0202	0.0240	–0.6871		+	–	0.1116	*
380	4.883	In	14612076	12	504	14	0.0082	0.0092	–0.4076		+	+	n.a.	
381	4.883	In	14674206	11	444	35	0.0278	0.0269	0.1415		+	+	0.1161	*
382	4.883	In	14735732	12	487	20	0.0085	0.0136	–1.5210	*	+	+	0.0625	

384	4.883	In	14829173	12	502	19	0.0101	0.0125	–0.7751		–	–	0.2156	
385	4.883	In	14842173	12	525	23	0.0129	0.0145	–0.4586		+	–	n.a.	
386	4.718	In	14857134	12	426	18	0.0122	0.0140	–0.5216		+	+	0.1353	
387	4.718	IR	14873817	12	577	25	0.0137	0.0143	–0.1750		+	+	0.1162	*
388	4.718	IR	14914630	10	610	11	0.0045	0.0064	–1.2017		+	+	0.0877	
389	4.718	IR	14966327	12	518	8	0.0038	0.0051	–0.9102		+	+	n.a.	
391	4.718	IR	14996648	12	562	12	0.0050	0.0071	–1.1005		+	+	n.a.	
390	4.718	IR	15025983	12	537	20	0.0121	0.0123	–0.0839		+	+	n.a.	
392	4.718	In	15057388	12	444	15	0.0073	0.0112	–1.3756		–	+	0.1713	
393	4.718	In	15078750	12	559	11	0.0058	0.0065	–0.4327		+	+	0.1229	
534	4.718	IR	15088818	12	410	2	0.0011	0.0016	–0.7584		–	–	0.0182	
394	4.624	In	15120263	12	582	33	0.0131	0.0188	–1.2513		+	+	0.1198	*
282	4.624	In	15191599	12	683	35	0.0153	0.0170	–0.4190		+	–	0.1177	
285	4.330	IR	15376408	12	494	42	0.0249	0.0282	–0.4891		+	+	n.a.	
286	4.330	IR	15412298	10	325	31	0.0235	0.0337	–1.3331		+	+	0.2855	
287	4.330	IR	15434546	12	572	21	0.0104	0.0122	–0.5808		+	+	0.1146	
288	4.330	In	15445188	12	521	17	0.0103	0.0108	–0.1982		–	–	0.1716	
295	4.330	IR	15453249	12	557	11	0.0039	0.0065	–1.5061	*	–	–	0.7521	*
296	4.330	In	15508294	12	587	9	0.0040	0.0051	–0.7644		+	+	0.0623	*
297	4.108	IR	15544018	12	630	8	0.0026	0.0042	–1.4292		–	+	0.0448	
298	4.108	IR	15598274	12	502	13	0.0066	0.0086	–0.9005		+	+	0.2307	*
299	4.108	In	15635880	12	618	18	0.0077	0.0096	–0.8261		+	+	0.0938	*
294	3.928	IR	15651395	12	595	15	0.0070	0.0083	–0.6449		+	+	0.1461	
301	3.928	In	15704066	12	556	26	0.0121	0.0155	–0.8964		+	+	0.1032	
306	3.798	IR	15776305	12	595	51	0.0236	0.0284	–0.7195		+	+	0.1057	
304	3.798	In	15815328	12	501	22	0.0114	0.0145	–0.8840		+	+	0.1967	*
725	3.928	IR	15873424	12	392	6	0.0032	0.0051	–1.2416		–	+	0.0860	*
307	3.667	In	15956388	12	513	29	0.0191	0.0187	0.0794		+	+	0.1417	
726	3.667	IR	16006314	12	434	10	0.0050	0.0076	–1.2788		+	+	0.0714	
310	3.667	In	16062419	12	491	8	0.0030	0.0054	–1.6161	*	–	+	0.1471	
336	3.667	In	16078174	12	600	7	0.0029	0.0039	–0.9050		+	+	0.0932	
334	3.571	IR	16162769	12	527	10	0.0059	0.0063	–0.2153		+	+	0.1379	
333	3.571	In	16183518	12	582	13	0.0077	0.0074	0.1618		–	–	0.2078	
451	3.494	In	16209944	12	638	11	0.0058	0.0057	0.0567		+	+	0.2385	*

331	3.460	In	16256960	12	552	24	0.0127	0.0144	–0.4707		+	–	0.1447	
330	3.460	IR	16278449	12	567	41	0.0254	0.0239	0.2525		+	+	0.1020	
329	3.318	IR	16336583	10	600	11	0.0045	0.0065	–1.2480		–	–	0.1035	*
328	3.318	In	16376075	10	545	8	0.0043	0.0052	–0.6792		+	+	0.0423	
366	3.318	IR	16378516	11	610	38	0.0260	0.0213	0.9520		–	–	0.3669	*
364	3.129	In	16437205	12	600	19	0.0091	0.0105	–0.5433		+	+	0.0890	
363	3.129	In	16457525	12	613	25	0.0136	0.0135	0.0196		–	+	0.2129	*
359	3.065	IR	16601328	12	525	22	0.0124	0.0139	–0.4370		+	+	0.0762	
402	3.008	In	16690855	12	600	19	0.0096	0.0105	–0.3308		+	+	0.0966	*
405	2.900	IR	16771183	10	627	24	0.0102	0.0135	–1.0497		+	+	0.1188	
406	2.867	In	16814813	12	656	23	0.0120	0.0116	0.1483		+	+	0.1715	*
407	2.867	IR	16840713	12	571	35	0.0163	0.0203	–0.8275		+	+	0.1186	
410	2.782	In	16934040	12	600	28	0.0124	0.0155	–0.8207		+	+	0.1521	*
411	2.782	IR	16965129	12	545	24	0.0105	0.0146	–1.1318		+	+	0.0617	
422	2.782	In	16995416	11	578	28	0.0150	0.0165	–0.3840		+	+	0.1468	*
727	2.684	IR	17067830	12	450	16	0.0117	0.0118	–0.0190		+	+	0.1461	
424	2.684	IR	17135792	12	634	16	0.0071	0.0084	–0.6179		+	+	0.0850	*
728	2.638	In	17207833	12	270	4	0.0025	0.0049	–1.5738	*	–	–	0.3388	
426	2.638	IR	17260748	12	658	20	0.0091	0.0101	–0.3879		+	+	0.0808	*
428	2.591	IR	17291478	12	606	20	0.0103	0.0109	–0.2497		+	+	0.1028	
729	2.591	IR	17349173	10	443	14	0.0136	0.0112	0.8997		+	+	0.3616	*
430	2.591	In	17399361	12	659	15	0.0040	0.0075	–1.8428	*	+	+	0.1078	
730	2.527	IR	17447527	12	218	8	0.0101	0.0122	–0.6195		+	+	n.a.	
431	2.487	In	17526093	12	509	4	0.0013	0.0026	–1.5738	*	–	–	0.1963	
432	2.487	IR	17569204	11	508	10	0.0053	0.0067	–0.8182		+	+	0.1117	
436	2.436	IR	17886135	11	378	16	0.0118	0.0145	–0.7361		+	+	0.1350	*
438	2.424	IR	17968026	12	570	23	0.0107	0.0134	–0.8227		+	+	0.1164	
439	2.424	IR	18036090	12	546	6	0.0024	0.0036	–1.1626		+	–	0.0368	
440	2.424	IR	18104832	12	594	20	0.0094	0.0112	–0.6366		+	+	0.2628	*
444	2.467	IR	18482497	11	567	26	0.0131	0.0157	–0.6742		+	+	0.1251	

APPENDIX 2.2 Divergence between African *D. melanogaster* population and *D. simulans*

Fragments are ordered from the telomere to the centromere; for each one, the following information is given:

Absolute position is in base pairs, from the telomere;

n is the number of lines sequenced;

lth is the number of sites studied (excluding insertions and deletions polymorphism);

S is the number of segregating sites;

D_s is the number of fixed differences between *D. melanogaster* and *D. simulans*;

K is the divergence between *D. melanogaster* and *D. simulans*;

Fragment	Abs. posit.	*n*	*lth*	*S*	D_s	*K*
419	1571015	12	557	4	10	0.0115
10	1899930	12	321	10	32	0.0875
17	1946108	12	775	15	54	0.0583
1	2004307	12	379	14	37	0.0756
22	2129973	12	591	9	31	0.0416
25	2137479	12	595	13	26	0.0258
26	2140729	12	566	17	33	0.0354
32	2188201	12	624	11	28	0.0312
38	2270372	10	391	16	46	0.0988
18	2448658	12	487	12	49	0.0934
5	2486993	12	224	13	25	0.0808
45	2740398	11	477	12	25	0.0329
46	2781118	12	571	24	45	0.0506
55	3235896	12	647	31	62	0.0663
54	3238859	12	409	30	51	0.0772
57	3333268	11	540	11	34	0.0490
60	3448557	12	604	28	75	0.1026
56	3603702	12	321	5	11	0.0235
76	3653297	12	527	32	54	0.0607
77	3680710	12	517	27	64	0.0942
78	3727323	12	594	22	49	0.0591
80	3839129	12	514	29	52	0.0748
81	3879576	12	537	16	33	0.0405
462	3918527	12	636	3	12	0.0148
84	4018352	11	570	18	41	0.0554

85	4069979	12	480	16	41	0.0632
66	4260258	12	322	16	28	0.0592
67	4512472	12	583	23	47	0.0519
90	4896054	12	391	26	56	0.1068
91	4952503	10	462	22	41	0.0555
93	5034343	12	390	9	17	0.0279
94	5091045	10	452	20	43	0.0668
95	5136167	11	488	17	51	0.0851
106	5441948	12	394	16	38	0.0700
72	5482021	12	357	33	43	0.0725
73	5555609	12	562	11	52	0.0822
109	5730972	11	572	14	31	0.0378
114	6567455	12	285	3	26	0.0903
115	6613211	12	366	6	32	0.0772
116	6649164	12	502	33	49	0.0590
117	6703197	9	538	31	65	0.0905
118	6752435	12	538	13	24	0.0291
119	6797217	12	245	21	47	0.1361
120	6874455	12	465	32	46	0.0544
122	6964795	12	576	6	14	0.0149
502	6991699	12	508	23	44	0.0569
124	7041579	12	762	8	24	0.0265
125	7092312	12	236	8	21	0.0699
126	7143267	12	570	38	72	0.0841
530	7316457	12	505	26	41	0.0419
130	7319723	12	540	19	47	0.0664
133	7469487	11	609	14	20	0.0183
136	7679367	12	350	26	43	0.0740
137	7710260	12	405	12	42	0.0865
138	7758526	12	265	7	21	0.0671
139	7819831	12	322	14	37	0.0927
143	8068685	12	495	19	37	0.0481
150	8393030	12	298	13	28	0.0703
153	8562010	12	455	20	35	0.0537

157	8763089	12	295	9	17	0.0362
163	9040189	12	593	22	64	0.0825
165	9149621	12	251	4	26	0.1008
166	9185460	12	599	11	30	0.0371
167	9228779	12	607	12	38	0.0488
169	9367972	12	279	2	22	0.0763
170	9409076	10	493	32	53	0.0750
173	9587511	12	498	8	21	0.0311
446	9660426	12	235	13	31	0.0943
175	9724676	12	573	32	55	0.0590
177	9798952	12	386	18	48	0.0981
178	9839303	12	476	33	47	0.0503
179	9887176	12	522	27	46	0.0524
464	10051437	12	540	27	47	0.0562
465	10091962	12	415	21	47	0.0859
184	10123327	12	407	21	34	0.0543
186	10222506	12	492	20	45	0.0658
187	10250800	12	499	13	40	0.0654
188	10274029	12	488	4	39	0.0763
189	10326177	12	518	16	37	0.0485
191	10381498	12	432	4	43	0.1024
470	10407332	11	603	6	28	0.0384
472	10499785	12	552	14	23	0.0216
194	10530546	12	407	13	29	0.0519
195	10553628	12	464	24	47	0.0718
473	10572196	12	501	11	42	0.0730
196	10588263	12	590	5	19	0.0254
197	10626957	12	517	6	24	0.0404
475	10692810	12	390	8	40	0.0944
743	10764628	11	291	13	21	0.0355
201	10820867	12	677	13	31	0.0319
477	10833367	12	636	29	52	0.0500
480	10881373	12	331	10	33	0.0779
203	10897598	12	513	21	45	0.0664

532	10915182	12	449	4	10	0.0177
204	10959318	12	515	23	64	0.0993
205	11017914	12	599	23	42	0.0496
483	11040319	12	640	42	75	0.0878
206	11058202	12	467	20	31	0.0413
207	11087062	12	438	29	63	0.1081
208	11114429	12	428	10	26	0.0445
209	11153913	12	474	39	70	0.1035
211	11227737	12	502	20	30	0.0331
212	11271565	12	642	9	37	0.0480
213	11307249	12	493	9	31	0.0521
214	11359194	12	523	15	49	0.0736
215	11405680	11	525	16	52	0.0848
216	11450047	12	566	38	63	0.0657
217	11475426	12	531	4	12	0.0160
218	11492521	12	345	10	37	0.0870
220	11562312	12	394	2	9	0.0188
221	11571743	12	361	16	36	0.0714
488	11642234	10	566	29	51	0.0576
224	11717141	12	571	25	40	0.0386
492	11872761	12	574	18	40	0.0455
229	11890206	12	407	12	24	0.0363
493	11949243	12	554	21	40	0.0510
231	11963993	11	474	38	91	0.1545
232	11986343	12	536	15	39	0.0528
233	12043219	12	428	32	54	0.0831
235	12080905	12	487	10	16	0.0178
447	12164349	12	548	14	32	0.0379
239	12202152	11	144	18	26	0.1203
241	12268167	12	553	6	29	0.0463
242	12309945	12	455	31	48	0.0646
721	12387758	12	335	25	38	0.0745
245	12439737	12	279	19	45	0.1185
248	12550100	11	612	21	52	0.0622

249	12582821	12	503	18	43	0.0587
250	12633850	12	588	25	70	0.0955
251	12677383	12	418	26	52	0.0974
252	12709653	12	405	23	43	0.0737
253	12752462	11	401	15	55	0.1192
254	12791901	12	389	6	18	0.0364
258	12886899	12	404	28	49	0.0826
259	12938544	12	285	15	35	0.0959
260	12978603	10	531	17	41	0.0587
272	13022057	12	272	13	24	0.0576
273	13027807	12	362	24	44	0.0824
722	13090086	12	291	23	43	0.1102
276	13158630	12	273	8	26	0.0849
277	13194383	12	503	34	56	0.0725
278	13244701	12	574	28	47	0.0498
279	13277791	12	613	23	32	0.0285
280	13311226	12	289	17	39	0.1034
311	13315520	12	213	19	41	0.1304
450	13323368	12	458	28	69	0.1110
312	13354548	12	627	15	29	0.0287
313	13394848	12	410	3	13	0.0265
314	13431378	12	409	16	36	0.0613
318	13574274	10	298	10	29	0.0818
319	13596773	12	288	10	18	0.0362
320	13636642	11	433	13	39	0.0748
321	13668913	12	535	11	14	0.0096
323	13757364	12	363	13	33	0.0639
325	13810837	12	177	1	7	0.0357
326	13850289	12	601	18	38	0.0406
745	13923319	11	368	2	15	0.0367
344	13982020	11	489	27	58	0.0872
346	14039423	10	357	10	27	0.0585
348	14104733	12	529	23	50	0.0684
350	14210996	12	450	13	27	0.0358

367	14233745	12	575	20	52	0.0728
368	14266922	12	329	18	36	0.0829
369	14303687	12	661	7	19	0.0210
370	14324141	12	483	30	57	0.0818
371	14355845	11	512	24	44	0.0616
373	14423857	11	527	9	22	0.0299
374	14435431	12	541	15	29	0.0353
375	14470360	12	612	32	61	0.0703
376	14502591	10	245	15	32	0.0984
378	14535555	12	454	16	35	0.0546
379	14573482	11	555	38	76	0.0997
381	14674206	11	366	27	40	0.0624
382	14735732	12	471	19	38	0.0541
384	14829173	12	459	18	58	0.1072
386	14857134	12	375	15	47	0.1050
387	14873817	12	574	24	49	0.0649
388	14914630	10	608	11	21	0.0192
392	15057388	12	185	7	22	0.1037
393	15078750	12	555	11	49	0.0778
534	15088818	12	405	2	5	0.0081
394	15120263	12	486	24	58	0.0893
282	15191599	12	631	27	67	0.0808
286	15412298	10	274	25	54	0.1484
287	15434546	12	535	18	35	0.0448
288	15445188	12	500	15	42	0.0678
295	15453249	12	424	8	41	0.0928
296	15508294	12	578	9	54	0.0866
297	15544018	12	613	8	17	0.0162
298	15598274	12	189	4	14	0.0611
299	15635880	12	590	17	39	0.0453
294	15651395	12	577	13	44	0.0640
301	15704066	12	494	17	48	0.0838
306	15776305	12	531	39	69	0.0853
304	15815328	12	464	20	57	0.0987

725	15873424	12	314	4	9	0.0202
307	15956388	12	474	25	52	0.0879
726	16006314	12	434	10	19	0.0238
310	16062419	12	413	7	22	0.0449
336	16078174	12	594	7	20	0.0261
334	16162769	12	441	8	20	0.0318
333	16183518	12	575	13	39	0.0544
451	16209944	12	622	7	17	0.0189
331	16256960	12	532	23	51	0.0685
330	16278449	12	517	34	72	0.1084
329	16336583	10	392	7	28	0.0622
328	16376075	10	502	6	12	0.0139
366	16378516	11	605	38	76	0.0964
364	16437205	12	458	15	27	0.0320
363	16457525	12	581	21	45	0.0591
359	16601328	12	362	11	27	0.0591
402	16690855	12	563	18	44	0.0561
405	16771183	10	599	22	47	0.0538
406	16814813	12	656	23	55	0.0622
407	16840713	12	503	28	64	0.0923
410	16934040	12	587	27	55	0.0631
411	16965129	12	457	7	21	0.0330
422	16995416	11	539	25	43	0.0472
727	17067830	12	411	13	29	0.0582
424	17135792	12	612	15	37	0.0442
728	17207833	12	270	4	7	0.0124
426	17260748	12	618	15	31	0.0341
428	17291478	12	582	19	40	0.0473
729	17349173	10	332	10	34	0.0891
430	17399361	12	659	15	27	0.0219
431	17526093	12	465	6	15	0.0211
432	17569204	11	498	8	18	0.0234
436	17886135	11	376	16	33	0.0549
438	17968026	12	524	24	48	0.0628

439	18036090	12	546	5	16	0.0217
440	18104832	12	589	19	60	0.0837
444	18482497	11	555	24	62	0.0888

APPENDIX 2.3 Demographic modeling of the African population

Fragments are ordered from the telomere to the centromere; for each one, the following information is given:

Absolute position is in base pairs, from the telomere;

π_A, π_C, π_G, π_T are the base frequencies;

κ is the transition/transvertion paramerter;

ξ is the pyrimidin/purine transition parameter;

the maximum-likelihood estimate of the ratio between the current and initial population size, ρ;

the maximum-likelihood estimate of the time when the population size started to change, τ;

the maximum-likelihood estimate of the population mutation parameter, θ.

Fragment	Abs. posit.	π_A	π_C	π_G	π_T	κ	ξ	ρ	τ	θ
5	2486993	0.3226	0.1935	0.1935	0.2903	1.0000	1.3333	5	1.0	4
45	2740398	0.2897	0.2133	0.2797	0.2173	1.0000	0.7500	50	2.0	1
46	2781118	0.3271	0.1806	0.2147	0.2777	0.9286	0.6250	10	0.5	7
55	3235896	0.3724	0.1727	0.2147	0.2402	1.2857	0.3846	10	2.0	7
54	3238859	0.3365	0.1635	0.1825	0.3175	0.9412	0.7778	50	2.5	7
57	3333268	0.2521	0.1758	0.2836	0.2886	3.0000	0.8000	10	3.0	1
60	3448557	0.2803	0.1799	0.1911	0.3487	1.0714	0.6667	100	1.0	7
56	3603702	0.3446	0.1477	0.1785	0.3292	4.0000	0.3333	5	1.0	1
76	3653297	0.2482	0.2169	0.2353	0.2996	2.3000	0.9167	50	3.0	4
77	3680710	0.3672	0.1408	0.1533	0.3387	1.2308	0.6000	500	1.0	7
78	3727323	0.3118	0.2439	0.2197	0.2246	0.3529	1.0000	10	2.0	4
80	3839129	0.2556	0.1572	0.2383	0.3489	1.1333	1.1250	500	0.5	13
81	3879576	0.2905	0.2095	0.2271	0.2729	0.9000	3.5000	100	0.0	7
84	4018352	0.2433	0.1879	0.2416	0.3272	1.3333	2.0000	50	1.5	4
85	4069979	0.3230	0.2314	0.1615	0.2842	1.0000	0.8000	5	2.5	4
66	4260258	0.3562	0.2087	0.1552	0.2799	0.7778	2.5000	10	1.0	4
67	4512472	0.2568	0.2021	0.1885	0.3526	1.1818	1.6000	10	3.0	4
90	4896054	0.3326	0.1780	0.1780	0.3115	1.2143	0.8889	1000	0.5	10
91	4952503	0.2774	0.1739	0.1884	0.3602	1.3000	1.1667	50	2.5	4
93	5034343	0.3195	0.1902	0.2780	0.2122	1.2500	0.2500	1000	1.0	1
94	5091045	0.2852	0.2246	0.2305	0.2598	1.7500	1.0000	500	3.0	1
95	5136167	0.3416	0.1752	0.1752	0.3080	1.3750	0.8333	5	3.0	4
106	5441948	0.3284	0.1753	0.2173	0.2790	0.7000	0.4000	50	3.0	1
72	5482021	0.3278	0.1699	0.1962	0.3062	0.7619	0.7778	5	2.5	10
109	5730972	0.3000	0.1607	0.2393	0.3000	0.4444	1.0000	50	2.0	1

115	6613211	0.3593	0.1608	0.1206	0.3593	2.3333	1.3333	10	1.5	1
116	6649164	0.3367	0.2077	0.1541	0.3015	0.7895	0.3636	500	1.0	10
117	6703197	0.2815	0.2434	0.2036	0.2715	1.5385	1.0000	100	1.5	10
118	6752435	0.3223	0.1731	0.1842	0.3204	1.6000	3.0000	50	2.0	1
119	6797217	0.3199	0.2054	0.1448	0.3300	1.3636	1.5000	10	3.0	4
120	6874455	0.3348	0.1557	0.1706	0.3390	1.1333	1.1250	50	2.5	4
153	8562010	0.3237	0.1888	0.2075	0.2801	0.9231	0.7143	10	2.5	4
157	8763089	0.2714	0.1714	0.2429	0.3143	3.0000	0.5000	10	2.5	1
160	8897903	0.3745	0.2135	0.1873	0.2247	1.6000	1.0000	500	1.5	1
163	9040189	0.2828	0.2094	0.1813	0.3266	1.0000	5.0000	500	3.0	1
166	9185460	0.2401	0.2237	0.3076	0.2286	0.5714	3.0000	100	1.0	1
167	9228779	0.3635	0.1743	0.1743	0.2878	0.5000	0.3333	50	1.5	1
170	9409076	0.2799	0.1783	0.1854	0.3565	1.9167	1.0909	1000	1.0	13
173	9587511	0.2209	0.2871	0.2570	0.2349	1.6667	0.2500	1000	0.5	1
446	9660426	0.3315	0.1826	0.1124	0.3736	1.2857	3.5000	10	0.5	4
175	9724676	0.3024	0.2016	0.1740	0.3220	1.7143	1.1818	5	2.0	10
177	9798952	0.4230	0.2176	0.1002	0.2592	1.8571	2.2500	10	1.5	4
178	9839303	0.3353	0.1727	0.1145	0.3775	1.2667	1.1111	50	3.0	4
179	9887176	0.3159	0.2094	0.2419	0.2329	2.7143	1.1111	1000	0.5	7
182	10046886	0.3013	0.2271	0.2031	0.2686	0.6000	2.0000	50	2.5	1
464	10051437	0.2391	0.2153	0.2865	0.2591	1.2308	1.6667	1000	0.5	10
465	10091962	0.2892	0.2119	0.1302	0.3687	1.0000	1.5000	5	0.0	7
184	10123327	0.3050	0.1583	0.1858	0.3509	1.4444	3.3333	10	2.5	4
186	10222506	0.3307	0.2198	0.1673	0.2821	0.5000	0.7500	500	2.5	1
187	10250800	0.2989	0.1858	0.2222	0.2931	1.3333	1.0000	10	3.0	1
188	10274029	0.2602	0.2195	0.2276	0.2927	1.0000	1.0000	10	0.5	1
189	10326177	0.3194	0.2051	0.1779	0.2976	0.7000	1.3333	50	2.5	1
190	10344630	0.3094	0.2136	0.1860	0.2910	0.6000	0.2857	10	3.0	4
192	10432451	0.3646	0.1463	0.1594	0.3297	1.4286	1.5000	50	3.0	1
472	10499785	0.4028	0.1891	0.1839	0.2242	0.6667	0.2000	50	2.0	1
194	10530546	0.2724	0.1638	0.2034	0.3603	1.8333	1.7500	50	2.5	1
195	10553628	0.3164	0.1660	0.1699	0.3477	1.4286	0.4286	10	2.5	7
473	10572196	0.3279	0.1377	0.1830	0.3515	0.8333	1.5000	5	0.0	4
197	10626957	0.2139	0.2450	0.2925	0.2486	6.0000	2.0000	100	0.5	1

198	10672053	0.2948	0.1600	0.1807	0.3644	1.6667	0.5000	10	3.0	4
475	10692810	0.3021	0.1955	0.1777	0.3247	2.0000	2.0000	50	1.0	1
743	10764628	0.1851	0.2435	0.3117	0.2597	1.1667	0.7500	1000	1.5	1
201	10820867	0.2290	0.2482	0.2349	0.2880	3.3333	1.0000	50	1.5	1
477	10833367	0.2723	0.2405	0.2466	0.2405	1.9091	0.7500	10	3.0	7
480	10881373	0.3014	0.1866	0.1962	0.3158	0.7778	1.3333	1000	1.5	1
203	10897598	0.2199	0.2234	0.2147	0.3421	0.8462	1.2000	5	3.0	4
204	10959318	0.3370	0.1593	0.1996	0.3040	0.9231	0.7143	10	3.0	4
205	11017914	0.2886	0.2222	0.2099	0.2793	0.7143	4.0000	5	1.0	7
483	11040319	0.2991	0.1696	0.1964	0.3348	0.7619	1.6667	100	2.0	7
206	11058202	0.3346	0.2022	0.1930	0.2702	0.8235	1.0000	100	2.5	4
207	11087062	0.3193	0.1928	0.1948	0.2932	1.1875	1.1111	5	2.0	13
208	11114429	0.3347	0.2020	0.1881	0.2752	1.5000	2.0000	10	2.5	1
209	11153913	0.3108	0.1693	0.1614	0.3586	1.2105	1.3000	50	2.0	10
210	11190222	0.2688	0.1667	0.1907	0.3739	0.9091	0.4286	100	3.0	1
211	11227737	0.3504	0.1440	0.1904	0.3152	1.0000	0.4000	1000	1.0	7
212	11271565	0.3135	0.1974	0.2308	0.2583	2.0000	0.6000	50	1.5	1
213	11307249	0.3797	0.1593	0.1932	0.2678	0.8333	4.0000	10	2.5	1
214	11359194	0.3418	0.1582	0.1718	0.3282	1.4286	0.4286	500	1.5	1
215	11405680	0.3345	0.1637	0.1708	0.3310	0.8889	1.6667	5	3.0	4
216	11450047	0.2553	0.2618	0.1301	0.3528	0.8261	1.3750	50	2.0	10
219	11542518	0.2958	0.1903	0.2145	0.2993	5.0000	0.2500	10	3.0	1
221	11571743	0.3707	0.1236	0.1762	0.3295	0.6364	0.7500	500	0.5	4
222	11614496	0.3217	0.2209	0.1570	0.3004	1.1538	0.8750	5	2.5	7
488	11642234	0.3361	0.1913	0.2313	0.2413	0.7222	0.8571	50	2.0	7
224	11717141	0.3233	0.1683	0.2167	0.2917	1.0769	1.0000	500	2.0	4
660	11733290	0.2581	0.2204	0.1882	0.3333	1.5000	0.5000	5	1.0	1
228	11846104	0.3260	0.1471	0.2010	0.3260	0.6000	2.0000	1000	1.0	1
492	11872761	0.2813	0.1804	0.2156	0.3226	1.4000	1.3333	50	2.0	4
229	11890206	0.2886	0.2192	0.2192	0.2729	1.3333	1.6667	50	2.0	1
493	11949243	0.3037	0.1908	0.2242	0.2814	1.1667	0.7500	5	2.5	7
231	11963993	0.3440	0.1578	0.1631	0.3351	0.9545	0.7500	500	1.5	13
232	11986343	0.2899	0.1920	0.1902	0.3279	0.8750	0.1667	50	2.0	1
233	12043219	0.2435	0.1746	0.2004	0.3815	1.1333	1.1250	5	2.0	10

235	12080905	0.2903	0.2030	0.1879	0.3188	2.3333	1.3333	100	1.0	1
237	12134325	0.2962	0.2028	0.2048	0.2962	0.7037	1.1111	10	2.5	10
447	12164349	0.3540	0.2028	0.1821	0.2612	2.8000	0.7500	1000	2.0	1
239	12202152	0.3564	0.1410	0.1383	0.3644	0.9048	1.3750	5	0.5	16
241	12268167	0.3257	0.2218	0.1690	0.2835	1.0000	2.0000	10	0.5	1
242	12309945	0.3440	0.1783	0.2229	0.2548	0.7368	1.0000	50	1.5	10
721	12387758	0.3602	0.1720	0.1532	0.3145	0.7500	1.0000	50	0.0	10
245	12439737	0.3699	0.1462	0.1720	0.3118	2.3333	0.7500	5	2.0	4
246	12490225	0.4098	0.1737	0.1604	0.2561	1.5714	0.5714	5	1.5	4
248	12550100	0.2610	0.1686	0.2126	0.3578	1.0909	0.7143	5	3.0	4
249	12582821	0.3356	0.1695	0.1798	0.3151	1.2000	2.0000	500	2.5	1
250	12633850	0.3643	0.1669	0.1568	0.3120	0.7333	1.7500	5	2.5	7
251	12677383	0.2356	0.2044	0.1644	0.3956	1.2308	1.6667	5	1.5	7
252	12709653	0.3356	0.1464	0.1374	0.3806	1.0000	1.6000	1000	1.0	7
253	12752462	0.3383	0.1370	0.1478	0.3769	0.2500	1.0000	10	2.0	4
254	12791901	0.2464	0.2346	0.2796	0.2393	1.0000	0.2500	50	1.0	1
258	12886899	0.3112	0.2220	0.1899	0.2769	1.6364	1.0000	100	0.5	10
259	12938544	0.3919	0.1554	0.2027	0.2500	1.5714	0.8333	100	2.5	1
260	12978603	0.2797	0.2949	0.1593	0.2661	0.7000	1.3333	50	3.0	1
272	13022057	0.3523	0.1572	0.1174	0.3731	0.8182	1.2500	500	0.0	4
273	13027807	0.3643	0.1207	0.1787	0.3364	1.0000	0.6250	50	0.5	7
722	13090086	0.3553	0.1318	0.1404	0.3725	0.7333	1.2000	10	2.5	4
276	13158630	0.3190	0.2178	0.1902	0.2730	0.4286	0.5000	100	1.0	1
277	13194383	0.3005	0.2299	0.1888	0.2808	1.4667	2.6667	5	2.0	10
278	13244701	0.2909	0.2141	0.2108	0.2843	1.3571	1.1111	10	2.5	7
279	13277791	0.2462	0.2493	0.1261	0.3784	1.0769	2.5000	50	0.5	10
280	13311226	0.3876	0.1598	0.2041	0.2485	0.3846	0.6667	5	3.0	4
311	13315520	0.2186	0.2884	0.2465	0.2465	0.9091	1.5000	1000	2.5	1
450	13323368	0.2685	0.2293	0.2036	0.2986	1.5000	0.9091	1000	2.5	4
312	13354548	0.3135	0.1803	0.2006	0.3056	1.1429	1.0000	10	3.0	1
313	13394848	0.2697	0.1776	0.2193	0.3333	3.5000	0.4000	50	1.0	1
314	13431378	0.2595	0.2180	0.1990	0.3235	1.3333	1.0000	5	2.5	4
318	13574274	0.3180	0.2202	0.1498	0.3119	0.6667	1.0000	5	2.5	1
319	13596773	0.3266	0.2026	0.1588	0.3120	1.2500	0.8750	100	2.0	4

320	13636642	0.2906	0.2244	0.2204	0.2645	1.1429	1.6667	50	2.5	1
321	13668913	0.3078	0.2517	0.1990	0.2415	0.8000	3.0000	10	1.5	1
323	13757364	0.1935	0.2258	0.1989	0.3817	0.3571	1.5000	1000	2.0	1
325	13810837	0.3502	0.1723	0.1854	0.2921	1.6667	1.5000	10	1.5	1
326	13850289	0.1805	0.2195	0.2699	0.3301	2.6000	1.6000	50	3.0	1
342	13933460	0.3290	0.1765	0.1654	0.3290	1.1875	1.3750	5	2.0	10
344	13982020	0.3301	0.1582	0.2324	0.2793	1.2308	2.2000	5	3.0	7
346	14039423	0.2676	0.2294	0.2133	0.2897	0.4286	2.0000	10	2.5	1
348	14104733	0.2745	0.2343	0.1556	0.3357	2.1250	1.1250	10	3.0	4
350	14210996	0.2894	0.1915	0.2574	0.2617	2.2500	0.8000	50	1.5	1
367	14233745	0.3105	0.1990	0.2196	0.2710	0.8182	1.2500	5	2.0	4
368	14266922	0.2667	0.1941	0.2588	0.2804	1.1667	1.8000	50	2.0	4
370	14324141	0.3035	0.2035	0.1807	0.3123	1.2000	1.2500	5	1.5	10
371	14355845	0.2308	0.2739	0.1782	0.3171	2.0000	0.6000	10	1.5	7
373	14423857	0.2863	0.2053	0.2580	0.2505	0.5000	0.5000	50	1.0	1
374	14435431	0.3003	0.2360	0.1898	0.2739	0.8750	2.5000	50	2.0	1
375	14470360	0.3207	0.2480	0.2006	0.2306	0.6190	0.4444	10	0.5	13
376	14502591	0.4470	0.1174	0.1364	0.2992	0.7778	0.7500	100	0.0	7
378	14535555	0.3700	0.1499	0.1727	0.3074	0.4667	2.5000	5	0.0	7
379	14573482	0.3890	0.1480	0.1601	0.3029	1.2222	1.0000	1000	2.5	10
380	14612076	0.2817	0.2183	0.1845	0.3155	1.0000	1.3333	10	3.0	1
381	14674206	0.3462	0.1699	0.1677	0.3161	0.4000	1.5000	500	0.0	13
382	14735732	0.3481	0.2254	0.1268	0.2998	1.2222	0.8333	1000	2.0	1
384	14829173	0.3793	0.1628	0.1322	0.3257	0.9000	0.2857	50	3.0	1
385	14842173	0.2831	0.2559	0.1434	0.3176	2.8333	3.2500	10	2.0	4
386	14857134	0.2980	0.1413	0.1589	0.4018	0.5000	1.0000	5	3.0	4
387	14873817	0.3163	0.1667	0.1735	0.3435	0.7857	0.5714	10	3.0	4
388	14914630	0.3328	0.2159	0.2192	0.2321	0.5714	1.0000	1000	1.0	1
389	14966327	0.2973	0.1969	0.2510	0.2548	1.0000	3.0000	50	1.0	1
391	14996648	0.3065	0.2154	0.2084	0.2697	2.0000	1.0000	100	1.5	1
390	15025983	0.3496	0.1902	0.1630	0.2971	1.0000	0.4286	500	0.5	7
392	15057388	0.3035	0.1850	0.1414	0.3701	0.8750	2.5000	500	1.5	1
393	15078750	0.2768	0.2018	0.1821	0.3393	1.7500	0.4000	10	2.5	1
394	15120263	0.3445	0.1412	0.1916	0.3227	1.5385	1.2222	100	3.0	4

282	15191599	0.2927	0.1879	0.2181	0.3013	0.8421	1.0000	10	2.0	7
285	15376408	0.2891	0.2026	0.1510	0.3573	0.9091	1.2222	10	1.5	13
286	15412298	0.3686	0.0967	0.1692	0.3656	1.5833	0.5833	1000	3.0	4
287	15434546	0.2914	0.2949	0.1972	0.2164	1.1000	0.5714	10	2.0	4
288	15445188	0.2822	0.2188	0.1900	0.3090	0.7000	0.7500	1000	0.5	4
295	15453249	0.3345	0.1734	0.2119	0.2802	0.3750	0.5000	500	1.0	1
296	15508294	0.2385	0.2351	0.2675	0.2589	3.5000	0.4000	10	1.5	1
298	15598274	0.3058	0.1962	0.1923	0.3058	0.8571	0.5000	50	2.0	1
299	15635880	0.2638	0.1796	0.1974	0.3592	5.0000	2.0000	50	3.0	1
294	15651395	0.3108	0.1755	0.1932	0.3205	2.7500	0.5714	50	3.0	1
301	15704066	0.3095	0.1599	0.2177	0.3129	1.3636	1.1429	50	2.0	4
306	15776305	0.2671	0.2096	0.2081	0.3152	0.9615	1.5000	100	2.0	13
304	15815328	0.3774	0.1170	0.1736	0.3321	2.1429	0.6667	50	1.5	4
725	15873424	0.2236	0.2688	0.2085	0.2990	2.0000	1.0000	500	0.5	1
307	15956388	0.3288	0.1923	0.1885	0.2904	2.2222	1.5000	5	1.0	10
726	16006314	0.2818	0.1886	0.2591	0.2705	1.0000	1.5000	500	1.0	1
336	16078174	0.3372	0.1736	0.1636	0.3256	2.5000	1.5000	500	0.5	1
334	16162769	0.3302	0.1973	0.2125	0.2600	1.5000	5.0000	10	2.5	1
331	16256960	0.3138	0.1695	0.2399	0.2768	1.0000	1.0000	10	2.5	4
330	16278449	0.2678	0.2356	0.1847	0.3119	0.7826	1.0000	5	1.0	13
329	16336583	0.2767	0.1833	0.2100	0.3300	0.8333	4.0000	1000	1.0	1
328	16376075	0.3455	0.2121	0.1718	0.2706	1.0000	3.0000	10	1.5	1
366	16378516	0.3041	0.1854	0.2130	0.2976	1.0000	1.1111	50	0.5	10
364	16437205	0.3085	0.1851	0.2484	0.2579	0.9000	0.8000	10	1.5	4
363	16457525	0.3754	0.1650	0.1845	0.2751	1.5000	0.6667	50	0.5	7
359	16601328	0.2648	0.2038	0.2133	0.3181	0.6923	1.2500	5	3.0	4
402	16690855	0.2917	0.2017	0.1717	0.3350	1.7143	0.7143	5	2.5	4
405	16771183	0.3180	0.1844	0.1717	0.3259	1.6667	1.1429	1000	3.0	1
406	16814813	0.2807	0.1108	0.1832	0.4254	0.5333	7.0000	5	0.5	7
407	16840713	0.3754	0.1562	0.1356	0.3328	0.9444	2.4000	10	3.0	7
410	16934040	0.3458	0.2284	0.1664	0.2594	0.8667	0.8571	50	2.0	4
411	16965129	0.2839	0.1777	0.2344	0.3040	0.8462	0.8333	100	3.0	1
422	16995416	0.2837	0.1644	0.2076	0.3443	1.0000	1.0000	100	1.0	7
727	17067830	0.2723	0.2353	0.2092	0.2832	2.2000	1.7500	500	0.0	7

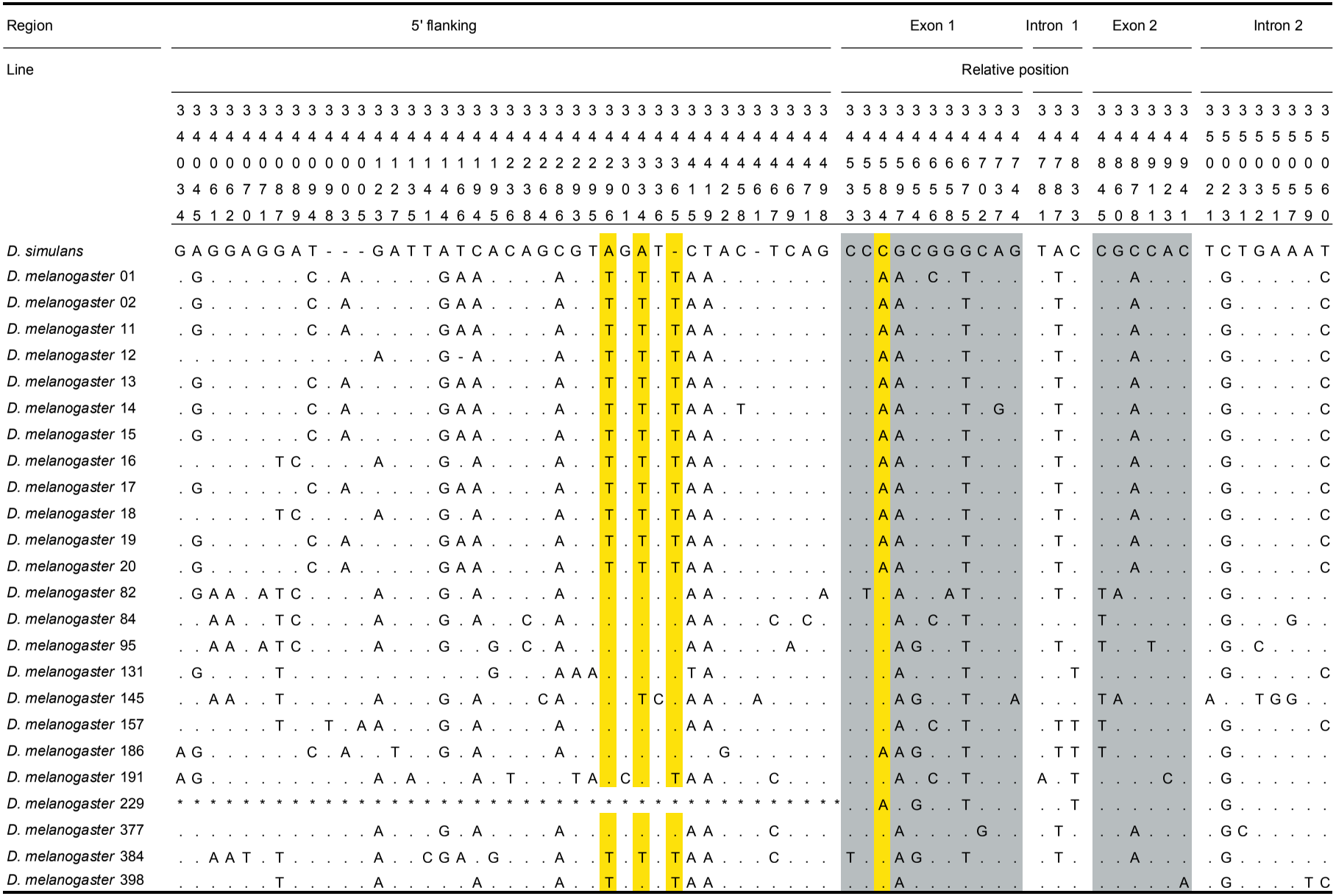
424	17135792	0.3056	0.1806	0.2330	0.2809	1.0000	0.6000	50	2.5	1
426	17260748	0.3294	0.1728	0.2167	0.2811	1.0000	0.2500	5	3.0	4
428	17291478	0.1909	0.2577	0.1860	0.3654	0.5385	6.0000	5	2.0	4
729	17349173	0.2412	0.1881	0.2035	0.3673	0.4000	1.0000	100	0.0	4
430	17399361	0.2045	0.2481	0.2090	0.3383	0.8750	0.7500	1000	1.5	1
730	17447527	0.3453	0.1435	0.1435	0.3677	1.0000	1.0000	10	1.5	1
431	17526093	0.2782	0.2293	0.1917	0.3008	3.0000	0.5000	50	0.5	1
432	17569204	0.3410	0.2267	0.2362	0.1962	4.0000	0.3333	10	2.0	1
436	17886135	0.2228	0.2896	0.2005	0.2871	1.2857	1.2500	50	2.5	1
438	17968026	0.2282	0.3031	0.2631	0.2056	1.0909	1.4000	500	1.5	4
439	18036090	0.2554	0.2283	0.2301	0.2862	0.5000	1.0000	100	0.5	1
440	18104832	0.3356	0.1896	0.1862	0.2886	0.5385	0.1667	5	3.0	4
444	18482497	0.3221	0.1987	0.1683	0.3109	0.7333	0.5714	10	3.0	4

Region	5' flanking	Exon 1	Intron 1	Exon 2	
Line				Relative position	
	3333333	33333333333333333	33333	3333333333333333333333333333333333333222	2
	2222222	22222111111111111	11111	1111111111111100000000000000000000000999	9
	8766665	22100988777666665	55555	5544444322211188877765544433221111000999	8
	1288616	96942996852988669	76432	2096433381183184298097299787207310400811	8
	6342172	94128867903990637	37649	5082485224365163036679610783625950974041	4
D. simulans	TTACCGA	CAGACCCTAGACGCCGC	ATACC	CGGCTGGGC--CCTCACGCCGCCCTCCGCCCCGCTGCCCC	T
D. melanogaster 01	....	.G......T.T..G...	.A.T.	....C.A.T.......T.T...TA...A.T.TA.C.....	G
D. melanogaster 02	A......	.G......T.T..G...	.A.T.	....C.A.........T.T...TA...A.T.TA.C.....	G
D. melanogaster 11	....	.G...A..T.T..G...	.A.T.	....C.A.....G...T.T...TA...A.T.TA.C.....	G
D. melanogaster 12	....	.G......T.T..G...	.A.T.	....C.A.........T.T...TA...A.T.TA.C.....	G
D. melanogaster 13	A......	.G......T.T..G...	.A.T.	....C.A.........T.T...TA...A.T.TA.C.....	G
D. melanogaster 14	....	.G......T.T..G...	.A.T.	....C.A.........T.T...TA...A.T.TA.C.....	G
D. melanogaster 15	....	.G......T.T..G...	.A.T.	....C.A.........T.T...TA...A.T.TA.C.....	G
D. melanogaster 16	....	.G......T.T..G...	.A.T.	....C.A.........T.T...TA...A.T.TA.C.....	G
D. melanogaster 17	....	.G....ACT.T..G...	.A.T.	....C.A.........T.T...TA...A.T.TA.C.....	G
D. melanogaster 18	......T	.G......T.T..G...	.A.T.	....C.A.....G...T.T...TA...A.T.TA.C.....	G
D. melanogaster 19	....	.G......T.T..G...	.A.T.	....C.A.........T.T...TA...A.T.TA.C.....	G
D. melanogaster 20	....	.G...A..T.T..G...	.A.T.	....C.A.....G...T.T...TA...A.T.TA.C.....	G
D. melanogaster 82	....	.G......T........	..G..	T.ATC......T......T....A.....T..A.CT..T.	G
D. melanogaster 84	..G....	.G......T.T..G...	.A.T.	....C.....T.....TAT.TTTA...A.T..AT..TT..	.
D. melanogaster 95	....	.G......T.T..G...	.A.T.	....C....A....A..A.TTTT.A..A.T..A.C....A	.
D. melanogaster 131	....	.G......T.T..G...	...T.	....C.............T...TA...AT.....C....A	G
D. melanogaster 145	.....A.	.G......TA.T..T..	.A.T.	....C....A........T.ATTG....T.....C..T.A	.
D. melanogaster 157	.....A.	.GA.A...TA....T.T	T...T	....C..CTA....A.T.T..TTG...A.T..AT..T...	.
D. melanogaster 186	....	TG.CA...T.T..G...	.A.T.	..A.C...........T.T....G.....T.TA.C....A	G
D. melanogaster 191	.C.....	.G..A...T.T..G...	.A.T.	....C......T..A.T.T.ATT..A.A.TA.ATC....A	G
D. melanogaster 229	....A..	.G..A...T.T..G.T.	...T.	....C.........A...T...TA...A.T..A.C..T.A	G
D. melanogaster 377	.....A.	.G......T.T.AG...	.A.T.	.....A............T...TA...A.T....C.TT.A	G
D. melanogaster 384	...A...	....A........G...	.....	....C........AAG..T..TTG..T..T..ATC.....	G
D. melanogaster 398	.....A.	....A...T.T..G...	.A.T.	.C..C.........A...T..TT....A.T..A.C.....	G

Intron 2	Exon 3	Intron 3	Exon 4	Intron 4	Exon 5
22222	22222	2	222222	222	222
99999	99999	9	999999	999	999
88888	88777	6	555554	444	221
66544	21740	4	997638	732	873
81296	15309	2	432065	456	899
TACTC	CGCAA	C	ACGCGT	GTC	AGG
.....G	...G.	A	..AT..	.C.	.A.
GCT...	.....	A	....T.	..A	.A.
G.....	..A.C	A	.....A	.CA	.A.
G.....	.CAGC	A	.....	.CA	.A.
GCT...	T.A..	A	.....	.CA	.A.
G..A..	..AGC	A	..A...	.CA	.A.
.....	..AGC	A	.....	.CA	.A.
G.....	...GC	A	.....	.C.	.A.
GCT...	T...C	A	.....	.CA	G..
G..A..	..A..	.	.....	.CA	.A.
GCT...	..AGC	A	.....A	.CA	G.A
.....G	...G.	A	..AT..	.C.	.A.

APPENDIX 3.1 Alignment of polymorphic sites observed in 12 lines each of the European and African *D. melanogaster* population according to the direction of transcription (see Figure 3.1) for gene *CG1677.*

The relative position of the 5' flanking and the coding regions to the first site of fragment *553* and the derived state inferred from *D. melanogaster*/*D. simulans* comparisons are given for each polymorphic site. At sites for which the derived state could not be determined due to a third base segregating in *D. simulans* or to an insertion/deletion difference between species, the base with the higher frequency in the African population was assumed to be ancestral. The order of exons (gray) and introns (white) are ascending given the location of the 5' region (white) and potential target sites of selection are highlighted in yellow. -, 1 bp deletion and *, sequence not available.


Region
5' flanking
Exon 1
Intron 1
Exon 2
Intron 2
Line
Relative position
D. simulans GAGGAGGAT---GATTATCACAGCGTAGATCTACTCAG CCCGCGGGCAG TAC CGCCAC TCTGAAAT
D. melanogaster 01 .G.....C.A.....GAA....A..T.T.TAA.....AA.C.T... .T. ..A... .G.....C
D. melanogaster 02 .G.....C.A.....GAA....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 11 .G.....C.A.....GAA....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 12 ..........A...G-A....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 13 .G.....C.A.....GAA....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 14 .G.....C.A.....GAA....A..T.T.TAA.T....AA...T.G. .T. ..A... .G.....C
D. melanogaster 15 .G.....C.A.....GAA....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 16 ......TC....A...G.A....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 17 .G.....C.A.....GAA....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 18 ......TC....A...G.A....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 19 .G.....C.A.....GAA....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 20 .G.....C.A.....GAA....A..T.T.TAA.....AA...T... .T. ..A... .G.....C
D. melanogaster 82 .GAA.ATC....A...G.A....A.......AA.....A.T.A..AT... .T. TA.... .G.....
D. melanogaster 84 ..AA..TC....A...G.A..C.A.......AA...C.C. ...A.C.T... .... T..... .G...G..
D. melanogaster 95 ..AA.ATC....A...G..G.C.A.......AA....A.. ...AG..T... .T. T..T.. .G.C....
D. melanogaster 131 .G....T..........G...AAA.....TA..... ...A...T... ..T ..... .G.....C
D. melanogaster 145 ..AA..T.....A...G.A...CA....TC.AA..A.... ...AG..T..A .... TA.... A..TGG..
D. melanogaster 157 ......T..T.AA...G.A....A.......AA..... ...A.C.T... .TT T..... .G.....C
D. melanogaster 186 AG.....C.A..T..G.A....A.........G..... ..AAG..T... .TT T..... .G.....
D. melanogaster 191 AG.........A.A...A.T...TA.C..TAA...C... ...A.C.T... A.T ....C. .G.....
D. melanogaster 229 ** ..A.G..T... ..T ..... .G.....
D. melanogaster 377 ..........A...G.A....A.....AA...C... ...A....G.. .T. ..A... .GC.....
D. melanogaster 384 ..AAT.T.....A..CGA.G...A..T.T.TAA...C... T..AG..T... .T. ..A... .G.....
D. melanogaster 398 ......T.....A.....A....A..T...TAA..... ...A..... ... .A .G....TC

Exon 3	Intron 3	Exon 4	Intron 4	Exon 5
3333333333	3333333333333	3333	33333333	3333333
5555555555	5555555555555	5555	55555555	5555555
1111122222	3333333333333	4444	55555555	6666666
2478824579	0 2233333455668	4499	2334477	0335112
7854531619	7 3503460255571	6917	8261325	6 9068490
CGGCCGGCCGT	CC-GCTTGGTCT	CGGC	CGAGGACC	TGTCCCT
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	TT....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
A.....A.TAC	T.....G..AT.	T..T	..C..T..	ATCTT..G
...AT...TA.	.T....G.AA..	T..T	.A...T..	.TC....
A.......TAC	T.....G..AT.	T..T	.....T..	ATCT..G.
A.....A.TA.	.T....G.AA..	T..T	...T.T..	ATCT....
A.A..AA.TA.	.T....G.AA.C	T..T	T..T.T..	ATCT.A..
...AT.A.TA.	.T.....A....	.A.T	.A.T.TT.	.TC....
A.A...A.T..	.T.......A..	.A.T	.A...T..	ATCT..G.
A....AA..AC	.T.A..GA.A..	T..T	...T..TT	AT.T....
A.A.....TA.	.T..A.G.A...	T..T	.A.T.TT.	.TCT..G.
.....AA.T..	.T.......A..	T..T	.A.T.TT.	A.C....
A.......TA.	.T..ACG..AT.	T..T	.A..AT..	.TC....
AT...AA.TA.	.T..A.G..AT.	T.AT	.A...T..	.TCT....
A..AT.AGTAC	T.C.....A..	T.A.	.....T..	.TCT..G.

APPENDIX 3.2 Alignment of polymorphic sites observed in 12 lines each of the European and African *D. melanogaster* population according to the direction of transcription (see Figure 3.1) for gene *CG2059*.

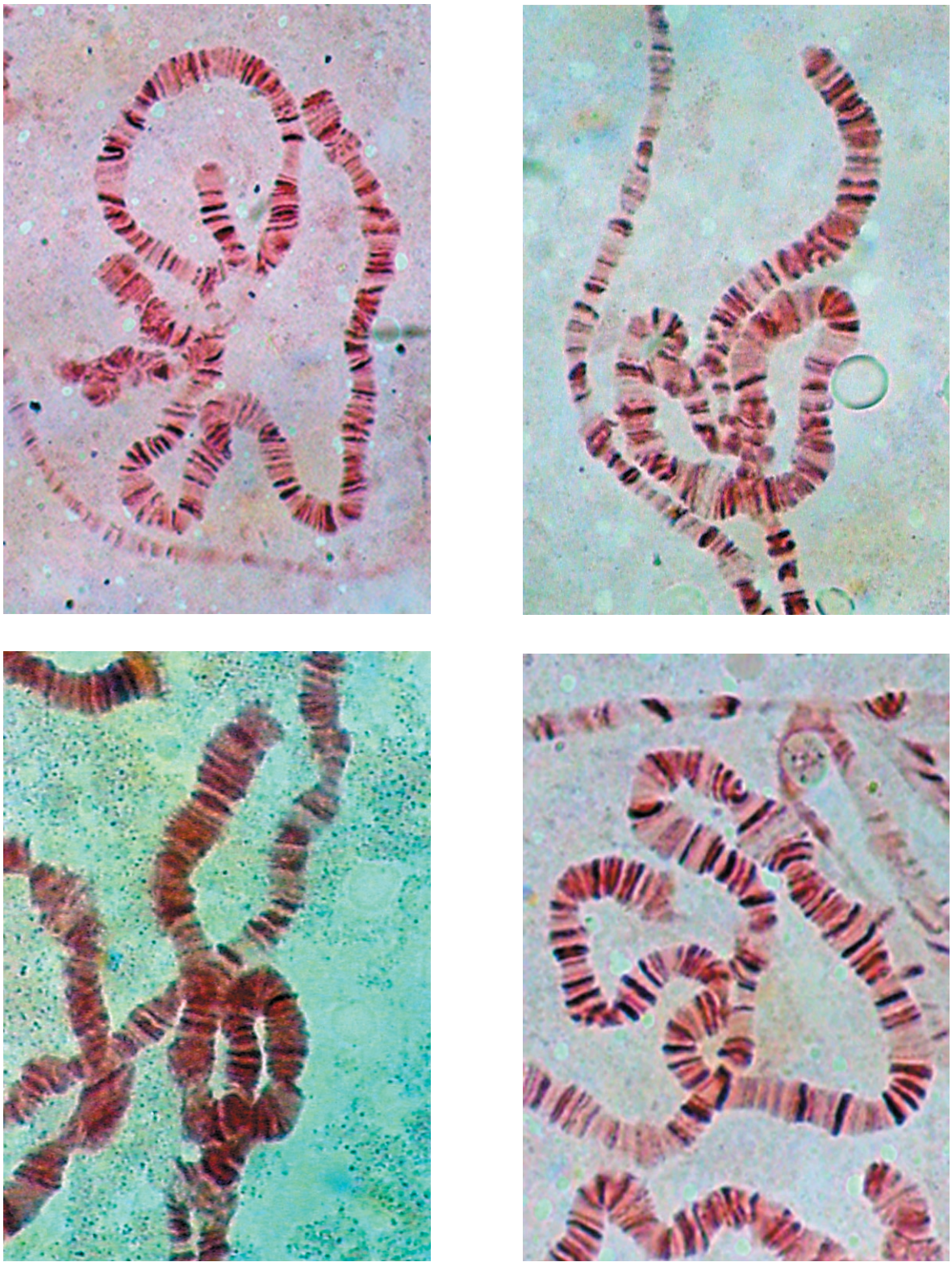
The relative position of the 5' flanking and the coding regions to the first site of fragment *553* and the derived state inferred from *D. melanogaster*/*D. simulans* comparisons are given for each polymorphic site. At sites for which the derived state could not be determined due to a third base segregating in *D. simulans* or to an insertion/deletion difference between species, the base with the higher frequency in the African population was assumed to be ancestral. The order of exons (gray) and introns (white) are ascending given the location of the 5' region (white) and potential target sites of selection are highlighted in yellow. -, 1 bp deletion and *, sequence not available.

Region	5' flanking	Exon 1	Intron 1	Exon 2	Intron 2
Line	Relative position				
	4444	3333333333333333	3333	333	3333333333
	4444	7777777777777777	7777	777	7776666666
	9966	9888888887765555	3333	322	1109999999
	7449	0986420099699751	8554	185	6379922020
	8664	1868622019421045	2546	555	4737030505
D. simulans	TCAC	CGGATTCAGCATAGT	TCCC	TCT	CGTCTGAC
D. melanogaster 01	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 02	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 11	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 12	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 13	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 14	...A	T...A..AC....	....	G..	AT.....A
D. melanogaster 15	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 16	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 17	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 18	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 19	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 20	...A	T.....AC....	....	G..	AT.....A
D. melanogaster 82	...A	.....G.....	....	GT.	A.A..A..
D. melanogaster 84	...A	.....G.....G..	C...	G..	A....A..
D. melanogaster 95	...A	.....G..A...GA.	..T.	GT.	A....A..
D. melanogaster 131	....	..AC.....CG.C	.ATT	G..	A....AG.
D. melanogaster 145	.T.A	.A.....G....G..	....	G..	.....AGT
D. melanogaster 157	...A	.....G.....G..	C.T.	G..	A.A.C...
D. melanogaster 186	...A	.A...G.....CG..	.AT.	G..	A.A..AGT
D. melanogaster 191	...A	.A...G....C.GA.	..T.	GT.	A....A..
D. melanogaster 229	..G.	.....G.....G..	....	...	A....A..
D. melanogaster 377	A..A	.A...GT.....	.AT.	G.C	A....A..
D. melanogaster 384	....	T....G.....G..	CAT.	G..	.....C...
D. melanogaster 398	...A	.....G.....G..	.ATT	G..	A..T.A..

APPENDIX 3.3 Alignment of polymorphic sites observed in 12 lines each of the European and African *D. melanogaster* population according to the direction of transcription (see Figure 3.1) for gene *unc-119*.

The relative position of the 5' flanking (*i.e.*, 5.7 kb away from the start codon; MARKSTEIN *et al*. 2002) and the coding regions to the first site of fragment *553* and the derived state inferred from *D. melanogaster*/*D. simulans* comparisons are given for each polymorphic site. At sites for which the derived state could not be determined due to a third base segregating in *D. simulans* or to an insertion/deletion difference between species, the base with the higher frequency in the African population was assumed to be ancestral. The order of exons (gray) and introns (white) are ascending given the location of the 5' region (white) and potential target sites of selection are highlighted in yellow. -, 1 bp deletion and *, sequence not available.

Appendix 5.1 Images of the common cosmopolitan inversions observed in the Southeast Asian *D. melanogaster* samples.
In(2L)t in the upper left corner;
In(2R)NS in the upper right corner;
In(3L)P in the lower left corner;
In(3R)P in the lower right corner.



EPILOGUE

The research presented in this thesis was conducted by myself, except for the following: In chapter one, I generated and analyzed the data for the African *D. melanogaster* population, calculated and examined the effect of different estimators of recombination rates to the observed level of nucleotide diversity and divergence in both African and European populations, and estimated the F_{ST} value and the proportion of the observed polymorphisms in the European population being present in the African population for half of the analyzed fragments. For the analysis of haplotypes in the African population, Dr. Sylvain Mousset kindly provided a C program for the simulations. In chapter two, Dr. Aparup Das assisted with the experimental inversion analysis and the C program used for the analysis of haplotypes (see above) was also used in this chapter. Dr. Haipeng Li helped to modify the “Iphula” program (kindly provided by Dr. Gunter Weiss) to make it applicable for multi-locus data and kindly provided a Java program to estimate various population genetic parameters and the input parameters used for the demographic modeling of the African population. S. Hutter kindly provided a C program to estimate divergence to *D. simulans* and linkage disequilibrium for the sequenced fragments. The same latter two programs were also used in chapter three. In chapter four, Dr. Sylvain Mousset kindly provided a C program of the maximum likelihood test for the analysis of the potential sweep region in the African population, and Dr. Aparup Das assisted with the experimental inversion analysis in chapter five.

CURRICULUM VITAE

Sascha Glinka
Bahnhofstraße 46
82340 Feldafing
Geburtsdatum: 13.03.1970
Geburtsort: Heilbronn

BILDUNG

2001-2005 Department Biologie II, Ludwig-Maximilians-Universität, München
Promotion ("Population genetic approaches to detect natural selection in *Drosophila melanogaster*")

1995-1998 Albert-Ludwigs-Universität, Freiburg im Breisgau
Diplom in Biologie (Gesamtnote 1,3)

1994-1995 University of Otago, Dunedin, Neuseeland
Postgraduate Diploma in Marine

1991-1994 Albert-Ludwigs-Universität, Freiburg im Breisgau
Vordiplom in Biologie

1990-1991 Arbeiten und Reisen im Ausland

1989-1990 Wehrdienst bei der Bundeswehr

1986-1989 Technisches Gymnasium, Öhringen
Allgemeine Hochschulreife

BERUFSERFAHRUNG

Feb 01-Jun 01 ECOSOFT.NET Deutschland GmbH, Freiburg im Breisgau
SQL-Programmierer,

Okt 00-Dez 00 Institut für Bodenkunde, A.-L.-Uni., Freiburg im Breisgau
Wissenschaftlicher Angestellter

Okt 98-Sep 00 Forstliche Versuchsanstalt Freiburg, Freiburg im Breisgau
Wissenschaftlicher Mitarbeiter

LIST OF PUBLICATIONS

Cooke, J., and S. Glinka, 1999 A comparative analysis of the demography of the Southwest Atlantic and Northwest Atlantic right whale populations from photoidentification of females with calves, document SC/O99/RW 1. International Whaling Commission, Cambridge, England.

Glinka, S., S. Brault and S.D. Kraus, 1999 Population assessment of the North Atlantic right whale (*Eubalaena glacialis*) with the tag-recapture method, abstract of talk presented at the 13th Biennial Conference on the Biology of Marine Mammals, Nov. 29–Dec. 3, Wailea, Maui, Hawaii, U.S.A. Conference Abstracts Volume.

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

Ometto, L., S. Glinka, S. Mousset, W. Stephan and D. De Lorenzo, 2003 A multi-locus survey of *Drosophila melanogaster* X chromosome: Demography and natural selection shaped genetic variation, abstract of poster presented at the 9th Congress of the European Society for Evolutionary Biology, Aug. 18–Aug. 24, Leeds, UK. Conference Abstracts Volume.

Glinka, S., W. Stephan and A. Das, 2005 Homogeneity of common cosmopolitan inversion frequencies in Southeast Asian *Drosophila melanogaster*. J. Genet. In press.

Ometto, L., S. Glinka, D. De Lorenzo and W. Stephan Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Submitted.

ACKNOWLEDGEMENTS

I sincere thank my adviser, Prof. Dr. Wolfgang Stephan, who gave me the opportunity to conduct my PhD in the interesting field of population genetics. His excellent scientific support and advice has particularly contributed to the research presented in this thesis, and his comments on the thesis chapters have improved them considerably.

I also like to thank our leader of the *Drosophila* project, Dr. David De Lorenzo. Not only that he found time to discuss scientific problems, his relaxed attitude has made my life as a PhD a bit easier. It has always been fun working within this project and discussing various research related topics with the "*Drosophila* people", namely Lino Ometto, Steffen Beisswanger and Stefan Hutter. Especially the collaboration with Lino was just great and I must thank him for the comments he made on different chapters of my thesis.

I would like to thank Prof. Dr. John Parsch and Dr. Laura Rose for helpful comments on various chapters of this thesis, and Dr. Peter Pfaffelhuber, who found somehow always time to discuss statistical questions. Especially his suggestions on the statistical part of chapter two has improved it a lot. In this regard, I owe the same gratitude to Dr. Sylvain Mousset for his comments on chapter one and four. I must also thank Dr. Aparup Das, whose expertise in the study of inversion polymorphisms minimized the level of conflict between several *D. melanogaster* larvae and myself. I also like to thank Andreas Buckenmaier and Pleuni Pennings for their help with "Mathematica".

I owe a tremendous amount of gratitude to Anne Wilken, Kawsar Bhuiyan, Bettina Schirrmeister and Gabi Büttner for their help in generating sequences, and Traudl Feldmaier-Fuchs and Anne for providing "tons" of flyfood.

I would also like to thank the entire evolutionary biology group and Dr. Ying Chen, Dr. John Baines and Dr. Sonja Köhler for fruitful discussions and relaxing chats in the lab, in a café or beergarden.

I especially thank my parents, Karin and Rainer Glinka, and my grandma Pauline Dietz for their support and encouragement over the entire years of my career.