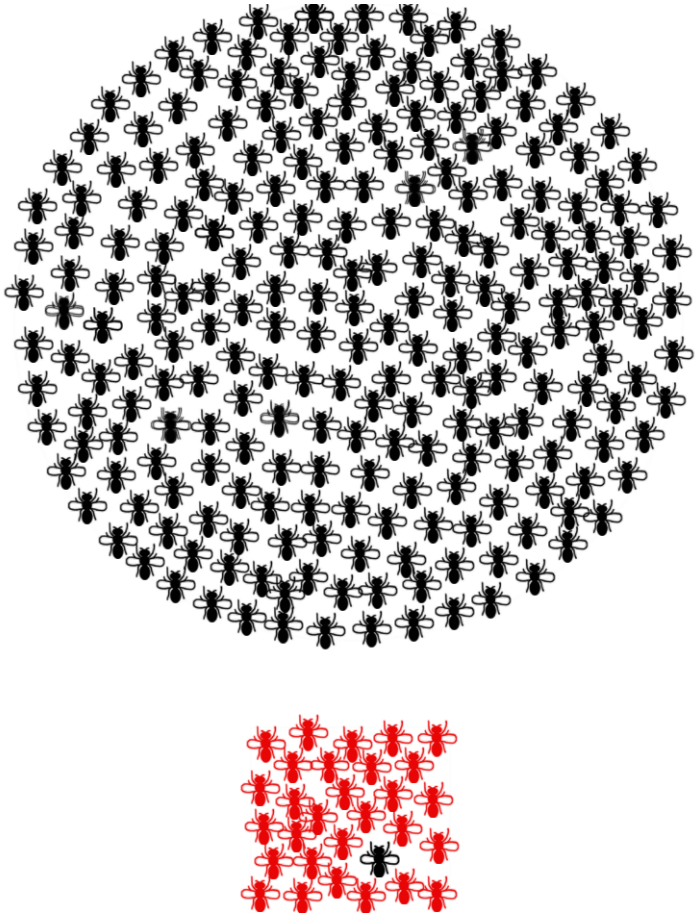
Dissertation zur Erlangung des Doktorgrades
der Naturwissenschaften an der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

The selective and demographic history of *Drosophila melanogaster*



Lino Ometto

aus
Vicenza, Italien

2006

Erklärung

Diese Dissertation wurde im Sinne von §13 Abs. 3 bzw. 4 der Promotionsordnung von Prof. Dr. Wolfgang Stephan betreut.

Ehrenwörtliche Versicherung

Diese Dissertation wurde selbstständig, ohne unerlaubte Hilfe erarbeitet.

1. Gutachter: Prof. Dr. Wolfgang Stephan

2. Gutachter: Prof. Dr. John Parsch

Dissertation eingereicht am: 16.12.2005

Tag der mündliche Prüfung: 17.02.2006

a elisa,
mare

Table of Contents

Summary 1
Note 3
Introduction 5
List of abbreviations 11

1. A survey of DNA variation in the X chromosome of *Drosophila melanogaster* 13
1.1. The impact of demography and natural selection on the genetic variation of *Drosophila melanogaster* 15
1.1.1. Materials and Methods 15
1.1.1.1. Population samples 15
1.1.1.2. PCR amplification and DNA sequencing 16
1.1.1.3. Statistical analysis 16
1.1.1.4. Recombination rate 17
1.1.1.5. Demographic modeling of the European population 17
1.1.2. Results 18
1.1.2.1. Polymorphism patterns in the African population 19
1.1.2.2. Polymorphism patterns in the European population 23
1.1.2.3. Comparison of the African and European populations 27
1.1.3. Discussion 27
1.1.3.1. Demography 27
1.1.3.2. Selection 28
1.2. Inferring the effects of demography and selection on *Drosophila melanogaster* populations **31**
1.2.1. Materials and Methods 31
1.2.1.1. Data collection 31
1.2.1.2. Statistical analysis 32
1.2.1.3. Demographic modeling of the African population 33
1.2.1.4. Demographic modeling of the European population 34
1.2.2. Results 38
1.2.2.1. Polymorphism patterns of the African population 38
1.2.2.2. Polymorphism patterns of the European population 42
1.2.2.3. Estimating the parameters of a simple bottleneck model for the European

population 45
1.2.2.4. Identifying candidate sweep regions in the European population 47
1.2.3. Discussion 52
1.2.3.1. Demographic history of the African population 53
1.2.3.2. Demographic and selection history of the European population 54
1.2.3.3. Estimating the frequency of adaptive substitutions 57
1.2.3.4. Is recombination mutagenic in *D. melanogaster*? 58

2.1. Characterization of a selective sweep in a European population of *Drosophila melanogaster* 59
2.1.1. Materials and Methods 60
2.1.1.1. Data collection and analysis 60
2.1.1.2. Testing the European population against demography 61
2.1.1.3. Gene expression analysis 62
2.1.2. Results 65
2.1.2.1. Nucleotide diversity pattern across the candidate region 65
2.1.2.2. Testing the polymorphism pattern against demography 66
2.1.2.3. Expression of the gene *CG9509* 70
2.1.3. Discussion 71
2.1.3.1. Indications that the valley of reduced variation corresponds to a selective sweep 72
2.1.3.2. Candidate genes associated to the selective sweep 73

3. The effects of neutral and selective forces on the genome evolution of *Drosophila melanogaster* 75
3.1. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions 77
3.1.1. Materials and Methods 78
3.1.1.1. Drosophila data set 78
3.1.1.2. Analysis of insertion and deletion variation 78
3.1.1.3. Modeling of selective constraints 79
3.1.2. Results and Discussion 81
3.1.2.1. Introns and intergenic regions show a similar polymorphic deletion bias 81
3.1.2.2. Insertions have smaller sizes and higher frequencies than deletions 81
3.1.2.3. Estimates of indel and nucleotide sequence variation 84
3.1.2.4. Introns, but not intergenic sequences, are larger in *D. melanogaster* than

in *D. simulans* 85
3.1.2.5. Analysis of selective constraints 87
3.2. Mutational pattern and substitution dynamics in the non-coding DNA of *Drosophila melanogaster* 91
3.2.1. Materials and Methods 92
3.2.1.1. Data collection and analysis 92
3.2.1.2. Statistical analysis 92
3.2.2. Results 93
3.2.2.1 Divergence correlates with recombination rate in intergenic regions 93
3.2.2.2. Differences in size among homologous loci in Drosophila 94
3.2.2.3. In intronic regions, GC content is lower than in intergenic regions and does not correlate with recombination 95
3.2.2.4. Substitution patterns in Drosophila 95
3.2.2.5. Comparing the fixation pattern of *D. melanogaster* and *D. simulans* 99
3.2.2.6. Base composition of indels 99
3.2.2.7. Inferring constraints in conserved intergenic and intronic regions 100
3.2.3. Discussion 102
3.2.3.1. Intergenic and intronic regions have different base composition but similar mutation patterns 102
3.2.3.2. GC content affects nucleotide diversity and the insertion/deletion dynamics 102
3.2.3.3. Replication time and transcription-associated mutation bias have negligible effects on the mutation pattern 104
3.2.3.4. Longer introns are under more constraints 105
3.2.3.5. Evidence for selective constraints in non-coding DNA 105
3.2.3.6. Positive correlation between divergence and recombination rate in intergenic regions 106

Conclusions 107
Literature cited 111
Appendix 125

Curriculum vitae 167
Publications 168
Acknowledgments 171

Summary

A species' evolutionary history is influenced by both neutral and selective processes. The effects that these forces have on genetic variation depend on their relative contributions. It is therefore important to be able to disentangle them. I conducted a comprehensive population genetics analysis of DNA polymorphism in *Drosophila melanogaster*, based on data collected from more than 250 loci spanning the entire X chromosome.

Part of my work was dedicated to unraveling the relative roles of natural selection and demography in the recent history of a European population. First, I found evidence of a large impact of the population-size bottleneck associated with the colonization of Europe by the ancestral sub-Saharan populations. The multi-locus approach was crucial to disentangle neutral and selective forces, since theory predicts that demography has genome-wide effects, whereas selection acts only locally. Hence, I developed a coalescent-based maximum-likelihood method that estimated the population-size bottleneck to be ~4,000–16,000 years old. While this can account for most of the reduction of variation observed in the European sample, I could identify several loci and regions whose polymorphism pattern departs from the expectations under such a demographic scenario. One of these candidate regions was studied further in detail, revealing a pronounced valley of reduced nucleotide variation that is incompatible with a simple bottleneck model. Rather, this finding and the associated skew in the allelic frequency spectrum support the recent action of positive selection. Taken together, these results suggest that the European population experienced numerous episodes of natural selection to adapt to the new environment.

A second goal of my research was to investigate the evolutionary patterns of non-

coding DNA and detect signatures of selective constraint. I found that in this species functional constraints limit the accumulation of nucleotide mutations and of insertion/deletions in both intergenic and intronic regions. In particular, I showed that insertions have smaller sizes and higher frequencies than deletions, supporting the hypothesis that they are selected to compensate for the loss of DNA caused by deletion bias. Analysis of a simple model of selective constraints suggests that the blocks of functional elements located in intergenic sequences are on average larger than those in introns, while the length distribution of relatively unconstrained sequences interspaced between these blocks is similar in the two non-coding regions. Consistently, sequences conserved across species (*i.e.*, free of deletions and/or insertions) have lower variation and divergence compared to the remaining fraction of DNA, supporting the presence of evolutionary constraints in these blocks. Moreover, I show that the base composition of intergenic and intronic regions is shaped by a complex interaction of neutral and non-neutral processes. Remarkably, GC content seems to be an important determinant of genetic diversity.

Note

In this thesis, I present the results of my doctoral research. It is organized in five chapters. The first two (1.1. and 1.2.) focus on the analysis of the polymorphism pattern of a European population of *Drosophila melanogaster*. These chapters also contain complementary data and results relative to the African population that were collected by Sascha Glinka: they have been included to draw and support the conclusions on the selective and demographic history of the European population. In CHAPTER 1.2., I also analyzed the mutational pattern and the linkage disequilibrium decay analysis of the African population (SUBSECTIONS 1.2.2.1. and 1.2.3.1.). Sebastian Ramos-Onsins and Sylvain Mousset helped me considerably during the development of the maximum-likelihood approach to estimate demographic parameters and detect candidate loci for positive selection. S. R.-O. also provided the core of the coalescent program. Finally, the last three chapters are entirely contributed by me, with the exception of two loci previously analyzed by Lena Müller (a former Diploma student).

Although these chapters are complementary one with the other, they are self-contained and can be read separately. The results have contributed to the following papers:

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. (* equally contributed.)

OMETTO, L., W. STEPHAN, and D. DE LORENZO, 2005. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* **169:**1521–1527.

OMETTO, L., S. GLINKA, D. DE LORENZO, and W. STEPHAN, 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Mol. Biol. Evol*. **22:**2119–2130.

Introduction

Natural variation within and between natural populations is closely associated with the differences in fitness among individuals. Darwin was the first to realize that such variation could be the raw material upon which natural selection could operate. Later, the discoveries of heritability by Mendel and of the genetic code integrated the theory of natural selection, making a comprehensive synthesis of natural variation and (molecular) evolution possible. In brief, changes at the molecular level (*i.e.*, mutations) translate into variability among individuals at the phenotypic level. Because DNA variation is heritable, it can be transmitted to the progeny (and thus to the population) proportionally to its relative fitness: alleles that confer a higher fitness to the carrier, *i.e.*, a larger number of descendants, can spread generation after generation to the entire population, until a new and more fit variant appears.

Population geneticists study the forces that shape the variation at the DNA level among the individuals of a population, its causes and its consequences. Nowadays, there is an extensive amount of theoretical work describing the neutral expectations of such variation. Under the standard neutral model, which assumes a panmictic population of infinite and constant size, genetic variation (*i.e.,* the amount of polymorphisms and their frequency distribution in a sample of individuals) depends on the balance between the rate at which new mutations appear, and genetic drift, which drives their fixation or loss. Many "neutrality tests" have been developed to compare neutral expectations to empirical observations. When data are not compatible with neutrality, we can reject one of the assumptions of the neutral model: for example, the demographic history of the population could have comprised a

recent bottleneck or expansion; or other forces beside genetic drift, such as selection, may have shaped the observed mutational pattern.

The role of natural selection as a major force driving molecular evolution has been questioned by the neutral theory of molecular evolution, first proposed by Kimura in the late 1960s (Kimura 1968). This theory states that selection did not influence polymorphism and divergence as we can observe at the molecular level: rather, genetic drift accounts for most, if not all, the observed variation.

A first approach to test this theory was to contrast neutral and empirical observations at a single locus, and assume a correspondence between its demographic and selective history with those of the whole population across the genome. The simplest feature to consider is the level of heterozygosity, or polymorphism. When looking at single nucleotide polymorphism, two statistics are usually calculated: θ, which is based only on the number of segregating (*i.e.,* polymorphic) sites in the sample (Watterson 1975), and π, the average pairwise difference between two sequences (which depends also on the frequency of the alleles; Tajima 1983). Under the standard neutral model, both measures should be equal to $\sim aN_e\mu$, where N_e is the effective population size, μ is the mutation rate and $a = 3$ or 4, for sex chromosomes or autosomes, respectively, of a diploid species. It follows that, if a locus has very low π and θ values, either μ or N_e are expected to be small. If we assume that the population was at equilibrium, the latter inferences must apply to the locus, that is, (i) its specific population size is (or has been in the recent past) extremely reduced, or (ii) the local mutation rate is low. This could in turn have been the result of recent positive selection at the locus, which reduced its effective population size by favoring only a fraction of the individuals, or of selective constraints limiting the accumulation of new mutations. Alternatively, we can test the expected equality of π and θ: if π is larger than θ, then the locus contains too many segregating sites with alleles at intermediate frequency, while the opposite is true when there is an excess of rare alleles. This difference is tested by Tajima's *D*, which compares the frequency distribution of the polymorphic sites with neutral expectations. Similar approaches are used in Fu and Li's *D* and *F* tests (Fu and Li 1993). Other tests rely on the difference between divergence and polymorphism patterns, such as the Hudson-Kreitman-Aguadé test (HKA; Hudson *et al.* 1987) and for protein-coding sequences the McDonald-Kreitman test (McDonald and Kreitman 1991). Again, a departure from neutrality can be ascribed to a

violation of (at least) one of the assumptions of the standard neutral model. Strong negative Tajima's *D* values (Tajima 1983), *i.e.* an excess of mutations segregating at low frequency, may suggest the recent action of positive selection, but population expansion produces the same effect. Therefore, one must be cautious in drawing conclusions without evidence supporting either of the two hypotheses. For example, high polymorphism at the locus would point to an expansion from a population at equilibrium, while low polymorphism would suggest the recovery either from a population bottleneck (which reduced variation) or from the recent action of positive selection. In the latter case, the selected allele and the linked variants go to fixation, a phenomenon known as hitchhiking that removes polymorphism around the selected locus (Maynard Smith and Haigh 1974).

A significant advance in testing the neutral theory was accomplished by using multi-locus data, thus decreasing the chance of having results biased by the low power of the single locus approach. Many studies aimed at the detection of signatures of positive selection in species that are genetically well characterized. Pioneering studies found a positive correlation between the levels of genetic variation and recombination rate in flies (Begun and Aquadro 1992), humans (Nachman *et al*. 1998) and wild tomatoes (Stephan and Langley 1998), as expected if selection affects mainly regions of low recombination due to the stronger association between the target of selection and the linked sites (the lack of a correlation between levels of divergence and recombination, as expected under the neutral model, excluded any mutation bias across the recombination gradient). Two forms of selection contribute to the lower variation in regions of low recombination: (i) background selection (Charlesworth *et al*. 1993), driven by selected mutations that are frequent and strongly deleterious, and (ii) hitchhiking (Maynard Smith and Haigh 1974), which conversely are caused by rare strongly beneficial mutations. In regions of low recombination the two models produce similar, yet distinguishable effects (Stephan *et al.* 1998; Braverman *et al.* 1995; Kim and Stephan 2000; Innan and Stephan 2003). On the other hand, in normal-to-high recombination regions, hitchhiking events leave as characteristic footprint a valley of reduced variation around the selected sites. The width of these valleys depends on the levels of selection and recombination: they are large when selection is strong and recombination low.

Analyzing many loci distributed along a chromosome is an effective way to identify candidate regions (*i.e.*, depressions in DNA polymorphism) where positive selection occurred without any *a priori* knowledge of the action of natural selection. This kind of approach is known as genetic scan of variation (or hitchhiking mapping), and considers multiple loci spaced by about 40–50 kilobases, such that valleys are not missed even in regions of high recombination (Kim and Stephan 2002). Then, a locus showing low levels of genetic variation compared to the surrounding loci may belong to a selective sweep valley. Further detailed analysis of the candidate region is obviously necessary to verify and confirm whether we have indeed a footprint of positive selection, since such valleys do not definitively represent evidence for a selective sweep, because other evolutionary forces, *e.g.,* drift, may be the cause. To help distinguish between neutral and selective effects, one can employ a combination of neutrality tests: *e.g.*, one expects negative Tajima's *D* within the valley and positive values at the borders. Kim and Stephan (2002) recently developed a likelihood ratio test that proved to be a powerful method to distinguish between selection and drift along a recombining chromosome using multi-locus data (see also Jensen *et al.* 2005).

In the present study, we studied the neutral and selective forces that shaped genome variation in *Drosophila melanogaster* by sequencing and analyzing multiple loci distributed along the X chromosome. *D. melanogaster* is an ideal species to look for evidence of positive selection, since its genome is completely sequenced and annotated, making genomic approaches feasible. Most importantly though, this species originated in sub-Saharan Africa and moved to the temperate regions after the last glaciations, in the last 10,000 to 15,000 years ago (David and Capy, 1988). The migration to new habitats was likely accompanied by adaptations to the new biotic and abiotic factors, *e.g.,* different food sources and colder temperatures. Therefore, by comparing a putatively ancestral population from Africa (Zimbabwe) with a derived population from Europe (Netherlands) we have the unique opportunity to look for the traits that have been involved in the process of adaptation. Moreover, since our study covers an (almost) entire chromosome, we can estimate the frequency of favorable substitutions (*i.e.,* of selective sweeps).

We chose to concentrate on the part of the X chromosome with medium-to-high recombination, which spans almost three quarters of its euchromatic portion. Since *D. melanogaster* males are heterozygotes for the sexual chromosomes, any favorable mutation

present in their unique copy of the X chromosome will be visible to natural selection, even when recessive. For this reason, one should expect a faster evolution of this chromosome compared to the autosomes, *i.e.,* more chances to detect selective sweeps. Advantageous substitutions causing sweeps that have occurred no longer than approximately $0.1N_e$ generations ago can be detected with sufficiently high power using single nucleotide polymorphisms (SNPs; Kim and Stephan 2000; Przeworski 2002). For *D. melanogaster*, $0.1N_e$ generations correspond to roughly 10,000 to 15,000 years, a time window that matches the colonization of Europe by this species very well. Thus, the use of DNA sequence variation in multiple loci dispersed along the whole chromosome should enable us to detect most of the sweeps that have occurred during this colonization period. For this aim, we were interested in loci with very low polymorphism, which might be within a valley of variation produced by a selective sweep. In fact, we focus only on non-coding loci, *i.e.,* loci that should be under little constraints and thus evolve, and accumulate variation, neutrally.

The basic questions of this thesis can therefore be summarized as the following:

1) What are the joint effects of the demographic and the selective history of D. melanogaster? (chapter I)

Our multi-locus approach offers an important advantage over the single-locus studies. The colonization of Europe was accompanied by a strong population-size bottleneck, causing a great reduction in heterozygosity in the derived population. This effect creates much "background noise" when looking for loci with low polymorphism. However, while demographic processes affect the entire genome in a similar way, selective forces leave locus-specific footprints that are detectable in our genome-wide survey. As a first approach, we tested whether the empirical data are supported by the sole action of a simple bottleneck. This question was addressed by a detailed analysis of the polymorphism pattern across the X chromosome in the derived European population and by comparing it to that of the putative ancestral African population.

2) Can we identify regions of the genome with a footprint of natural selection? (chapter II)

Once the demographic model is estimated, we have a "neutral model" against which we can test our data. That is, we can test whether the polymorphism present in the analyzed loci is compatible with a simple bottleneck, or if an additional force (selection) must be

invoked. We present a method to disentangle demographic and selective forces across the genome and apply it to our European dataset.

3) Is there evidence for positive selection at a fine scale? (CHAPTER II)

We then tested the power of our methodology by choosing one of the candidate regions identified by the scan. In particular, we focused our attention around a locus that showed a reduction of polymorphism not compatible with demography alone. The additional collected data confirmed that selection is likely to have played a role in this region, producing a characteristic valley of reduced variation.

4) Is non-coding DNA evolving neutrally? (CHAPTER III)

The availability of polymorphism data across the whole chromosome prompted us to look for the effects of weak selection in the genome. The loci sequenced for the scan of variation are in intergenic and intronic regions, and thus are not under such strong purifying selection as coding sequences. Nonetheless, there is evidence that some functional constraints may in fact also be present in non-coding regions. To quantify these selective forces, we analyzed both the nucleotide sequence and the insertion/deletion variation in the African population (because it is closer to the neutral equilibrium). Sequence composition can be important when regulatory elements are present within loci (*e.g.,* transcription-factor bind sites), while at the same time functional units (*e.g.,* exons, whole genes, transcription-factor binding sites, enhancers) have to be correctly spaced to work properly or to avoid interference (*i.e.,* suffer the constraints of the linked sequences; HILL and ROBERTSON 1968).

List of abbreviations

AT→GC	Polymorphic site, or fixed substitution, where A or T mutated to G or C
bp	Base pair
D	Tajima's D
Div	Divergence
Div_{mel}, Div_{sim}	Divergence along the *D. melanogaster* and *D. simulans* lineages, respectively
F_0	Number of loci with no polymorphism
F_{ST}	Differentiation between populations
G	The total number of simulated genealogies
g	A simulated genealogy
GC→AT	Polymorphic site, or fixed substitution, where G or C mutated to A or T
H	Fay and Wu's H
H_{DV}	Depaulis and Veuille's statistic for haplotype diversity
HKA	Hudson-Kreitman-Aguadé test
i	A locus
k	Number of SNPs
k^C	Number of SNPs across C adjacent loci
k^A, k^E	Number of SNPs private to the African and the European samples, respectively
k^{Eall}, k^{Es}	Number of total SNPs and of singletons private to the European sample, respectively
k_{tot}	Total number of SNPs across a region
K_{DV}	Depaulis and Veuille's statistic for haplotype number
kb	Kilobases
L	Length of a locus (in bp)
LD	Linkage disequilibrium
Lik	Likelihood
Mb	Megabases
n	Sample size
N_b	Population size during the bottleneck phase
N_e	Effective population size

N_i	Initial (pre-bottleneck) population size
N_0	Present population size
PDB	Polymorphic deletion bias
Q	Probability for a locus to harbor at most k segregating sites
Q^C	Probability to observe at most k^C segregating sites across C adjacent loci
Q^E	Probability to observe at most k^E and k^A segregating sites
r	Recombination events, per-site per-generation
r^2	Correlation coefficient for a pair of biallelic sites
S_b	Strength of the bottleneck
SNP	Single nucleotide polymorphism (*i.e.*, segregating site)
T_b	Age of the bottleneck
T_e	Time at which the population expanded
T_m	Time spent by the population in the bottleneck phase
T_{tree}	Length of the coalescent tree
T^E_{tree}, T^A_{tree}	Length of the coalescent tree portions after and before T_b, respectively
Z_{nS}	Kelly's linkage disequilibrium measure
Δ	Strength of the population expansion
δ	Maximum non-deleterious fraction of L that can be added by insertions
γ	Maximum non-deleterious fraction of L that can be lost by deletions
θ	Nucleotide diversity based on k
θ_i	Locus-specific mutational parameter
$\overline{\theta}$	Average θ across the African loci
θ^E	θ based on k^E
θ^{Eall}, θ^{Es}	θ based on k^{Eall} and k^{Es}, respectively
π	Nucleotide diversity based on the mean pairwise differences between sequences

1. A survey of DNA variation in the X chromosome of *Drosophila melanogaster*

Inferring a species' demographic history from patterns of genetic variation is essential in a search for adaptive signatures in the genome. Traditionally, DNA variation is compared with the expectations of the neutral theory of molecular evolution (Kimura 1968). While it is tempting to ascribe departures from the neutral equilibrium model to the action of positive selection, caution must be taken due to the possible confounding effects of demography. For example, strong reduction in levels of nucleotide polymorphism may result from hitchhiking associated with positive directional selection (Maynard Smith and Haigh 1974) or from a strong population-size bottleneck. To disentangle demographic and selective forces, it is helpful to employ multi-locus approaches (*e.g.,* in humans, Akey *et al.* 2004; in Drosophila, Begun and Whitley 2000, Harr *et al.* 2002; in Arabidopsis, Schmid *et al.* 2005). The rationale behind these studies is the observation that while demography affects patterns of variation across the entire genome, positive selection acts locally.

The cosmopolitan species *Drosophila melanogaster* is thought to have colonized Europe after the last glaciation about 10,000–15,000 years ago (David and Capy 1988; Lachaise *et al*. 1988). Several studies proposed that this colonization was accompanied by the occurrence of numerous adaptations to the new habitat (Harr *et al.* 2002; Orengo and Aguadé 2004). On the other hand, despite their long evolutionary history, African populations also show a departure from the neutral equilibrium model, suggesting that *D. melanogaster* may have faced recent selective and demographic processes in its ancestral species range (Andolfatto and Przeworski 2001; Andolfatto and Wall 2003).

1.1. The impact of demography and natural selection on the genetic variation of *Drosophila melanogaster*

To assess the role of natural selection in the recent history of *D. melanogaster*, we compared a putatively ancestral population from Africa (Zimbabwe) with a derived population from Europe (Netherlands). Since a whole-genome scan of DNA sequence variation is currently not feasible, we used a multi-locus approach. This allowed us to gather information on the forces, *i.e.,* demography, that shaped genome-wide patterns of genetic variation, and assess whether such pattern is consistent only with a simple demographic scenario or other forces are needed. In particular, we were interested in two things: (i) to what extent a population size bottleneck can explain the levels of variation of the European population; and (ii) if there are footprints of positive selection in this population.

1.1.1. Materials and Methods

1.1.1.1. 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, Netherlands). The Zimbabwe lines were kindly provided by C. F. Aquadro, the European ones by A. J. Davis. Furthermore, a single *D. simulans* inbred strain (Davis, CA; kindly provided by H. A. Orr) was used for interspecific comparisons.

1.1.1.2. PCR amplification and DNA sequencing

Based on the available DNA sequence of the *D. melanogaster* genome (Flybase 2000, Release 2), 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 of these loci are located in regions of intermediate to high recombination rates. However, 11 loci are from the telomeric region exhibiting low recombination rates; *i.e.,* distal to the *white* locus (see below; APPENDIX B). We also amplified and sequenced the homologous 105 fragments in a single strain of *D. simulans*.

We extracted amplified and sequenced genomic DNA from each inbred line according to the protocols give in the APPENDIX C. 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 package (DNAstar Inc., Madison, WI). Singletons were confirmed by re-amplification and re-sequencing.

1.1.1.3. 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 (WATTERSON 1975). For this analysis, we considered the total number of mutations rather than the number of segregating sites, as we have observed in a few instances three different nucleotides segregating at the same position.

To test the neutral equilibrium model, we employed the multi-locus Hudson-Kreitman-Aguadé (HKA) and Tajima's *D* tests (HUDSON *et al.* 1987; TAJIMA 1989). 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_{DV} and the haplotype diversity H_{DV} (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 locus using coalescent simulations (10,000 iterations) (HUDSON 1990, 1993), conditioned on the number of segregating sites (DEPAULIS *et al.* 2001) and the population recombination

rate, R (programs are available from S. Mousset). 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 locus per generation (Stephan *et al*. 1998; Przeworski *et al*. 2001). N_e was assumed to be 10^6 (Li *et al.* 1999), and, for each locus, c was estimated by multiplying the per-site-recombination-rate r (see below) by its length L.

1.1.1.4. Recombination rate

We estimated r (recombination rate per site per generation) for each locus as follows. We used a computer program of Comeron *et al.* (1999) to obtain an estimate of the recombination rate for each locus. 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 locus to calculate physical distances. The estimate of the recombination rate is therefore expressed in units of 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 our loci: (I) the distal-*white* region (0.2 Mb to 2.45 Mb, 0.02 cM to 1.5 cM), and (II) the proximal-*white* region (2.45 Mb to 16.89 Mb, 1.5 cM to 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.

1.1.1.5. 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 model of a population-size-bottleneck with subsequent expansion (Wall *et al*. 2002; Lazzaro and Clark 2003). In the latter model, we simulated a population of initial effective size N_i, 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_o.

The following parameters had to be specified for each locus: the mutational parameter θ (estimated from data), the sample size n and fragment length L. Constant-population-size

models were tested using the observed average θ value of the European population, while the bottleneck models were conditioned on the observed average θ value of the African population (*i.e.,* the value of the hypothetical ancestral population). Our simple models assumed no intragenic recombination but free recombination between loci. We used several combinations of values of N_b, N_0/N_i 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 locus, 10,000 genealogies were simulated using the program ms (HUDSON 2002) under the demographic models mentioned above. The probability of observing exactly F_0 = 13 loci with no polymorphism in our simulation (see RESULTS, SUBSECTION 1.1.2.2.) was then calculated as the proportion of simulated samples with exactly 13 loci with no polymorphic sites. This probability was used in a two-tailed likelihood ratio test as a likelihood of our observation; when the probability was lower than 10^{-4}, we used 10^{-4} as a conservative overestimate of this value.

1.1.2. RESULTS

DNA sequences for 105 X chromosome loci 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 loci are spans approximately 14 Mb. This results in an average distance between adjacent loci of about 140 kb.

There are several large gaps in our genome scan, 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 loci (103) are located in two segments (between coordinates 1.9 Mb and 4.1 Mb, and between 6.5 Mb and 16.4 Mb from the telomere, respectively), thus spanning a region of 12 Mb with an average distance of 119 kb between loci. 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.

1.1.2.1. Polymorphism patterns in the African population

A summary of the polymorphism and divergence data is shown in Figures 1a–c and APPENDIX B (Table B1). Of the 54,944 sites sequenced (excluding insertions and deletions), 2,057 are polymorphic. The mean of θ (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 2001). For π, the result is similar: 0.0112 (0.0007) to 0.0074 (MORIYAMA and POWELL 1996) and 0.0242 (ANDOLFATTO 2001).

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 loci are consistent with the equilibrium model (HUDSON *et al*. 1987). A multi-locus version of the original HKA test was applied to all 105 loci in the African sample (Figure 2a). No significant departure from the equilibrium model was detected (X^2 = 93.31, P = 0.765).

We also calculated the Tajima's D statistic for each locus and tested whether the observed average across loci was consistent with the equilibrium model by estimating the critical values of this distribution from coalescent simulations (see MATERIALS AND METHODS, SUBSECTION 1.1.1.3.). In these simulations, we assumed no intragenic recombination (but free recombination between loci). 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 loci had a more extreme average value of D. This suggests that our data depart from the neutral equilibrium model. In fact, most of the loci have negative D values (Sign test, two-tailed, P < 0.001) (Figure 1d).

To further investigate the pattern of variation in the African sample, we focused on two

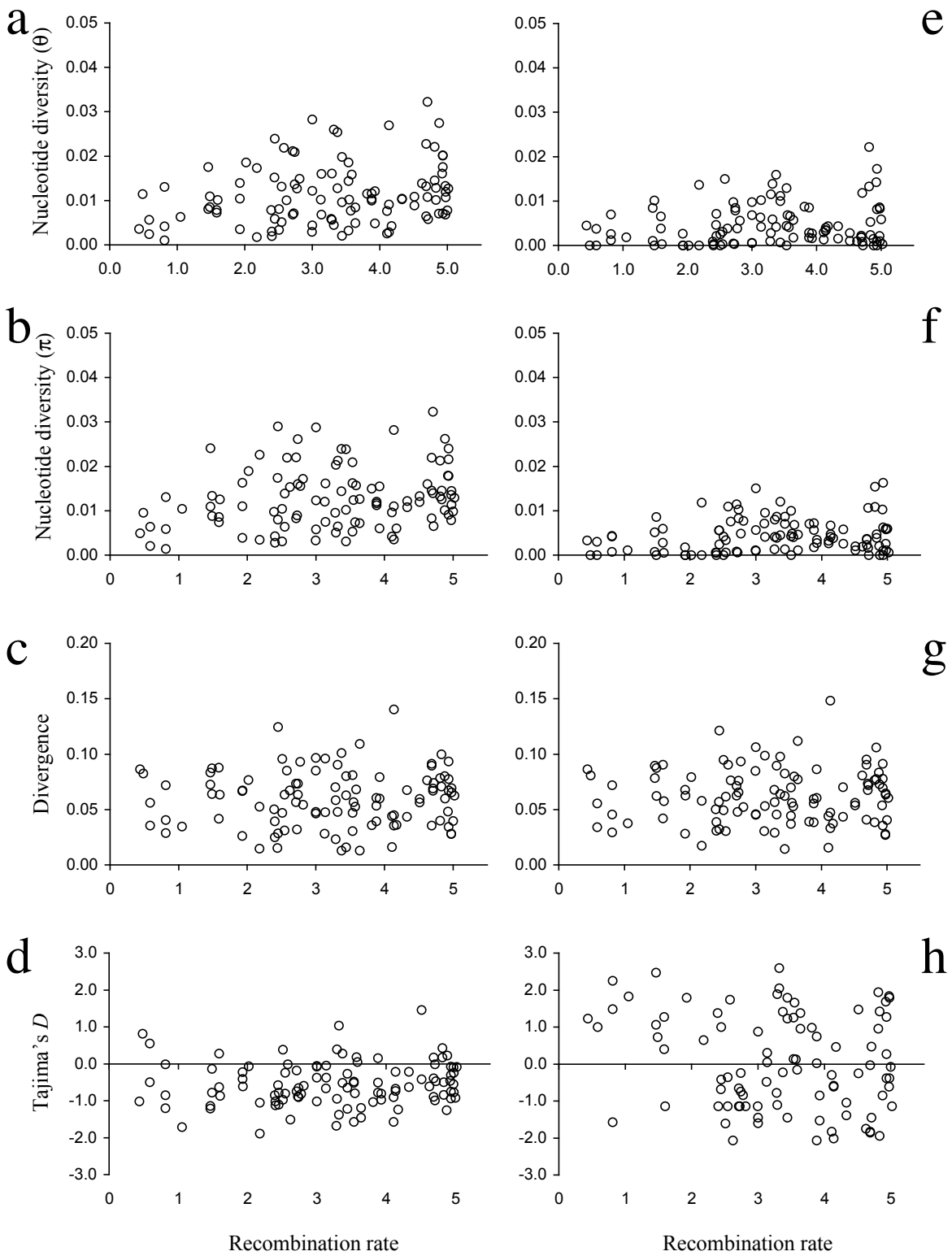


Figure 1. Nucleotide diversity π and θ, divergence, and Tajima's *D* versus recombination rate. Panels a–d refer to the African population, and panels e–h to the European one. Recombination rate is expressed in recombination events per site per generation $\times 10^8$ (Comeron *et al*. 1999).

statistics, the number of haplotypes K_{DV}, and the haplotype diversity H_{DV} (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, SUBSECTION 1.1.1.4.) for each locus 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 being lower and higher than the simulated median.

We found that the observed haplotype diversity H_{DV} was higher than the simulated median in 78 of the 105 loci; this proportion is significantly larger than expected (Sign test, two-tailed, $P < 0.001$). For the number of haplotypes K_{DV}, a significant trend toward a higher number was also observed (Sign test, two-tailed, $P = 0.03$). 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 expect to detect the footprints of partial sweeps as well. We thus examined whether there is evidence for partial hitchhiking events using the K_{DV}- and H_{DV}-haplotype tests (DEPAULIS and VEUILLE 1998), and Fay and Wu's *H* test (FAY and WU 2000). Since we are now 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 loci, 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 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; see Figures 1a and b): for π, Pearson's $R = 0.246$, $P < 0.02$; Spearman's $R = 0.237$, $P < 0.02$; for θ, Pearson's $R = 0.237$, $P < 0.02$;

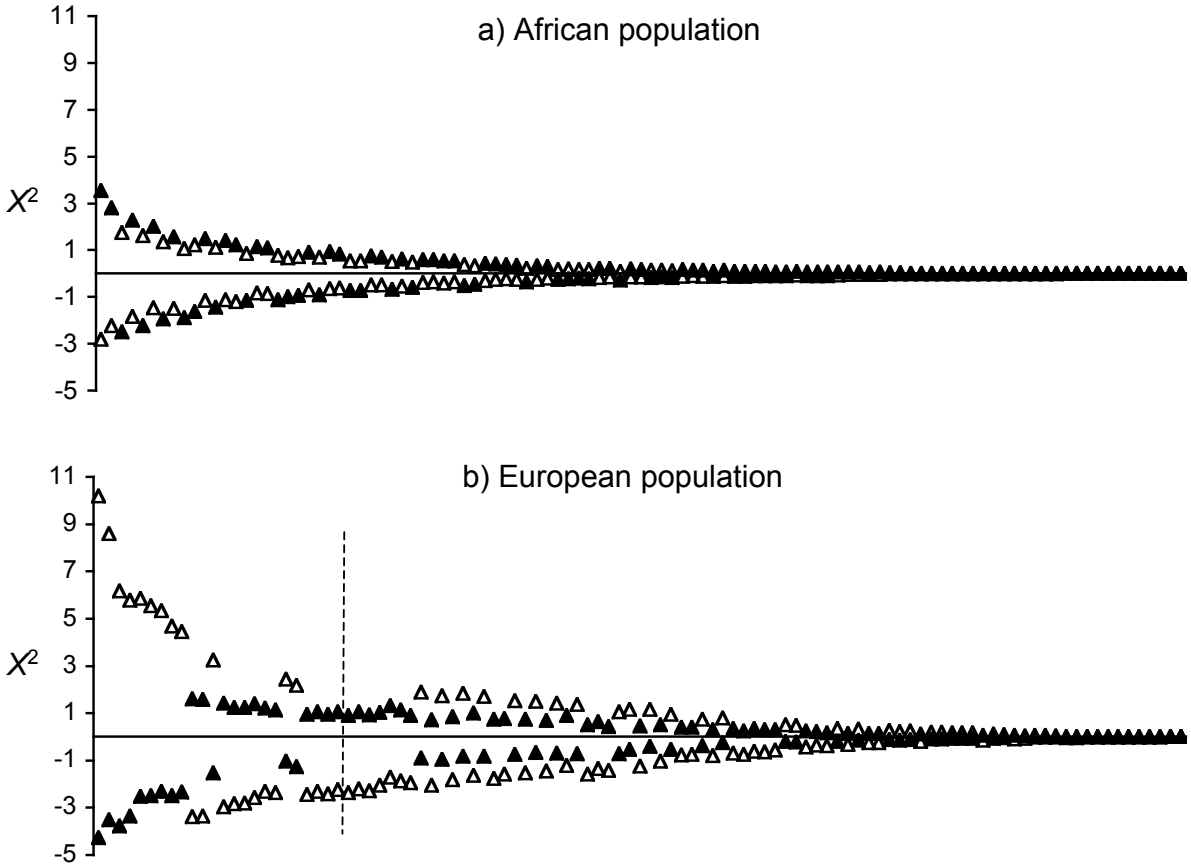


Figure 2. Contribution of each locus to multi-locus HKA statistic. (a) African population, and (b) European population. For each locus, the contributions to the overall test statistic by the polymorphism (empty triangles) and divergence (filled triangles) data are shown. Values above (below) the X-axis indicate a larger (smaller) contribution than expected. Loci are ranked along the X-axis according to their total contribution to the test statistic (including polymorphism and divergence components). When the 24 loci 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.

Spearman's $R = 0.234$, $P < 0.02$. If this observation was 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 1c). If we consider only loci above a certain recombination rate (say, 2×10^{-8} recombination events per base pair per generation, which corresponds to our previously defined region II; see MATERIALS AND METHODS, SUBSECTION 1.1.1.4.), 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: Pearson's $R = 0.115$, $P > 0.20$). These conclusions hold for all three measures of

recombination rates (see MATERIALS AND METHODS, SUBSECTION 1.1.1.4.), except that the (weak) correlation between nucleotide diversity and ACE was still found when the 11 loci 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, respectively). This suggests that the strong positive correlation between recombination rates and nucleotide diversity reported in previous studies is mainly attributable to loci in low recombination regions (BEGUN and AQUADRO 1992; AQUADRO *et al*. 1994; ANDOLFATTO and PRZEWORSKI 2001).

1.1.2.2. Polymorphism patterns in the European population

A summary of the polymorphism and divergence data is shown in Figures 1e–g. Of the 55,150 sites sequenced, 737 are polymorphic. The number of segregating sites and estimates of nucleotide diversity for each locus are shown in APPENDIX B (Table B2). The means (SE) of π and θ across the X chromosome are 0.0046 (0.0005) and 0.0044 (0.0004), respectively.

In Figures 1e–f, the estimates of π and θ are plotted against the recombination rate. We observed no significant correlation between nucleotide diversity and any of the three estimates of the recombination rate (MATERIALS AND METHODS, SUBSECTION 1.1.1.4.). 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 are 0.150 and 0.180 with $P > 0.12$ and $P > 0.06$ for π and θ, respectively; Spearman's R are 0.137 and 0.183 with $P > 0.16$ and $P > 0.06$. Also, no correlation between recombination rate and divergence was observed (Figure 1g; 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 (1989) test was applied to the European sample as described in MATERIALS AND METHODS (SUBSECTION 1.1.1.3.). The observed average of Tajima's D (SE) across loci 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, there are 11 loci for which the data are not compatible with the neutral equilibrium model. The Tajima test (in its single-locus version; TAJIMA 1989) revealed seven loci with significantly negative *D* values and four with positive ones. Inspection of the data shows that often Tajima's *D* is negative in the loci exhibiting a rare haplotype with many singletons, or strongly positive, when most of the variants are organized in a few common haplotypes (Figure 1h). It appears that, as a result of this, the mean of *D* across loci is not different from zero (see also CHAPTER 1.2.).

Using the same approach as for the African population sample, we computed the distribution of the K_{DV} and H_{DV} haplotype statistics (DEPAULIS and VEUILLE 1998) and recorded the proportion of observed values being lower and higher than the simulated median. The observed H_{DV} values were lower than the simulated median for 83 loci; this proportion is higher than expected (Sign test, two-tailed, $P < 0.0001$). For K_{DV}, the trend toward fewer haplotypes was also significant (Sign test, two-tailed, $P < 0.005$). In agreement with this observation, we found 13 loci with a significantly low value of K_{DV} or H_{DV}, using the conservative assumption of no recombination in one-tailed K_{DV}- or H_{DV}-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 (MATERIALS AND METHODS, SUBSECTION 1.1.1.3). A significant departure of the data from this model was detected ($X^2 = 238.28$, $P = 0.0016$). Figure 2b shows the contributions of each locus to the summary statistic. Furthermore, Figure 2b 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 loci with the strongest departures from expectation. The value of the overall test statistic dropped below the critical value where the test was no longer significant, if 24 loci with the largest contributions were removed (data not shown; 12 of these loci show an excess of polymorphism, and 12 a deficiency of polymorphism; see Figure 2b). Note that some of these low-polymorphism loci contribute to the overall test statistic to a very similar degree as the ones following at higher ranks; *i.e.,* between the loci at rank 20 and at rank 30 the per-locus contribution differs less than 0.5. All these loci have values of $\theta \leq 0.0011$.

Next we analyze the loci exhibiting low levels of variation. In our survey, 13 loci had no polymorphic sites at all (APPENDIX B, Table B2). Furthermore, 12 low-variation loci have

been identified by the HKA test, including eight of the non-polymorphic loci and four with extremely reduced nucleotide variability ($\theta \leq 0.0007$).

We first concentrate our analysis on the set of loci with zero polymorphisms. We used coalescent simulations to test the hypothesis that simple demographic null models (see MATERIALS AND METHODS, SUBSECTION 1.1.1.5) can explain our observation of 13 loci with zero polymorphisms. These are a neutral model of constant population size, and various bottleneck models (Table 1). Since the European population is believed to be derived from Africa (DAVID and CAPY 1988; ANDOLFATTO 2001), the pre-bottleneck effective population size (N_i) is assumed to be equal to the effective size of the Zimbabwe population (*i.e.,* ~10^6). Different values of N_0 for the European population (between 0.25 and $0.5N_i$) – accounting for the fact that the observed θ 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 fit better the observation of 13 loci with no polymorphism than the neutral (constant population size) model (*e.g.,* Bot 10, $G = 14.1$, $P = 0.014$, see Table 1). Appreciable probabilities of getting at least 13 loci with no polymorphic sites were only obtained for parameter values of the bottleneck model in which the effective population size recovered to its current size in a relatively short time period (about $0.1N_0$ generations). Other more realistic scenarios, in which the European population was founded 10,000–15,000 years ago corresponding to more than ~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 loci with no polymorphism (but see CHAPTER 1.2.).

Further evidence against a simple model of population founding followed by expansion is provided by the last two columns of Table 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 loci). However, we note that the bottleneck scenarios used in the above simulations are meant only to give a general picture. In fact, we explored only an extremely small set of parameters combinations

Table 1. Demographic modeling of the European population

Model	Model parameters				$\overline{F}_0$	$P(F_0 \leq 13)$	$P(F_0 = 13)$	$P(F_0 \geq 13)$	Avg. $\overline{D}$	$P(\overline{D} \geq 0.045)$
	T_b	N_b	T_m	N_o/N_i						
Constant	–	–	–	–	1.26	1	$< 10^{-4}$	$< 10^{-4}$	–0.077	0.0847
Bot 1	100000	1000	3600	0.5	2.60	1	$< 10^{-4}$	$< 10^{-4}$	–0.967	$< 10^{-4}$
Bot 2	100000	1000	7500	0.25	0.60	1	$< 10^{-4}$	$< 10^{-4}$	–1.050	$< 10^{-4}$
Bot 3	100000	500	1750	0.5	2.50	1	$< 10^{-4}$	$< 10^{-4}$	–0.955	$< 10^{-4}$
Bot 4	100000	500	4150	0.25	0.55	1	$< 10^{-4}$	$< 10^{-4}$	–1.049	$< 10^{-4}$
Bot 5	50000	1000	2900	0.5	9.14	0.9336	0.0512 *	0.1176	–0.672	$< 10^{-4}$
Bot 6	50000	1000	4400	0.25	3.13	1	$< 10^{-4}$	$< 10^{-4}$	–1.028	$< 10^{-4}$
Bot 7	50000	500	1500	0.5	9.08	0.9314	0.0484 *	0.1167	–0.712	$< 10^{-4}$
Bot 8	50000	500	2250	0.25	2.94	1	$< 10^{-4}$	$< 10^{-4}$	–1.049	$< 10^{-4}$
Bot 9	25000	1000	2750	0.5	22.40	0.0132	0.0070	0.9938	–0.355	$< 10^{-4}$
Bot 10	25000	1000	3850	0.25	12.51	0.6333	0.1153 *	0.4820	–0.790	$< 10^{-4}$
Bot 11	25000	500	1300	0.5	20.21	0.0440	0.0210	0.9770	–0.335	0.0013
Bot 12	25000	500	2000	0.25	11.56	0.7407	0.1093 *	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_0/N_i. A severe bottleneck of size N_b was introduced T_b generations ago in a population of initial size N_i and maintained for T_m generations. After that time, the population was allowed to grow exponentially to the current population size N_o. $N_i = 10^6$ was assumed. The value of the population mutation parameter was 0.0127, which is equal to the observed average value of θ for the African sample. For the constant-size simulations, the corresponding θ 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 loci are in close agreement. F_0 is the number of loci with no variation; $P(F_0 \leq 13)$, $P(F_0 = 13)$ and $P(F_0 \geq 13)$ are the probabilities of obtaining at most, exactly or at least 13 loci with no polymorphism, respectively; Avg. $\overline{D}$ is the value of Tajima's D across all loci averaged over all 10,000 simulation runs, and $P(\overline{D} \geq 0.045)$ is the probability of observing a value of Tajima's $\overline{D}$ across loci equal or larger than the value observed in the European sample.

* Likelihood ratio test, two-tailed, $P < 0.05$ (*i.e.*, the respective bottleneck model fits better the observation of $F_0 = 13$ than Constant).

(age, strength…). A substantial step towards the description of the "real" demographic and selective history of the population can be done only estimating the bottleneck parameters. A method for doing so is presented in the next chapter (CHAPTER 1.2.).

1.1.2.3. 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). As evident from the larger difference in the means of θ (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 about 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 loci is 0.293 (0.017). Likely, the bottleneck fixed ancestral segregating variation in the European population: 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, $P < 0.001$).

1.1.3. 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.

1.1.3.1. 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 2001). 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. A detailed analysis of this feature is presented in chapter 1.2.

1.1.3.2. 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 loci. 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 loci is on average relatively high, it was also possible to search for loci 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 locus. 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 shows a large number of loci with zero or low levels of variation. This observation is difficult to explain without invoking positive natural selection, since demographic modeling suggests that our observation of 13 loci with zero variation is not consistent with a neutral equilibrium model or the simulated neutral model of population founding followed by expansion. 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.

1.2. Inferring the effects of demography and selection on *Drosophila melanogaster* populations

As shown in the CHAPTER 1.1., additional forces beside demography shaped the DNA variation pattern observed in the European population, likely selection to adapt to the new environment. On the other hand, we also found clear evidence for non-equilibrium of the ancestral African population, probably due to population expansion.

We enlarged our previous dataset to more than 250 loci distributed across the X chromosome, in both the putatively ancestral African population and the derived European population. To identify genomic regions that may have been recent targets of positive Darwinian selection in the European population it is necessary to disentangle demographic and selective forces. At this aim, we developed a maximum-likelihood approach that estimated the age and strength of a simple bottleneck, and then identified those loci whose polymorphism departs from the expectations under such demographic scenario.

1.2.1. MATERIALS AND METHODS

1.2.1.1. Data collection

D. melanogaster and *D. simulans* sequence data were obtained as described in SUBSECTION 1.1.1.2.

Since chromosomal inversions can alter the polymorphism pattern across the chromosome (ANDOLFATTO *et al*. 2001), before sequencing the additional fragments we verified their absence in both the African and European lines. We made several single-mate crosses between virgin Canton-S females, homozygous for the standard chromosome

arrangement, and males from each of the 24 *D. melanogaster* lines. From each of these crosses, we used salivary gland preparations from five F_1 third-instar larvae (maintained at 18 °C) to maximize the detection of heterozygous inversions. Polytene chromosomes were stained using the lacto-acetic orcein method and examined under an inverted compound microscope.

Sequence data were collected from fragments located in non-coding DNA on the X chromosome of *D. melanogaster* (based on Flybase, Release 3.2, http://flybase.org). We sequenced 158 loci of the African and 163 of the European lines. For 143 loci, we also obtained the homolog in *D. simulans*. DNA sequences were generated as described in SUBSECTION 1.1.1.2. In addition, for the analysis we also used the data from 100 of the fragments described in CHAPTER 1.1., whose location was verified based on the genome Release 3.2 (originally, Release 2 was used). The updated annotation revealed that five of the previously analyzed 105 loci are located in putative coding regions. Therefore, we excluded them from the analysis. As a result, polymorphism (divergence) data was obtained from a total of 253 (232) and 263 (241) fragments for the African and the European populations, respectively. 250 fragments were analyzed in both samples of *D. melanogaster* and 230 of these loci also in *D. simulans*.

1.2.1.2. Statistical analysis

Below we give only methods additional to those described in SUBSECTIONS 1.1.1.3-4. Basic population genetic parameters were estimated by a program provided by H. Li. Nucleotide diversity was estimated using π (TAJIMA 1983) and θ (WATTERSON 1975). The same program was also used to estimate Tajima's *D* (Tajima 1989) and Fay and Wu's *H* (FAY and WU 2000) statistics and to evaluate their statistical significance by 10,000 coalescent simulations. Linkage disequilibrium (LD) measure Z_{nS} (KELLY 1997) and interspecific divergence were estimated by the program VariScan (VILELLA *et al*. 2005). A coalescent-based program (RAMOS-ONSINS *et al*. 2004) was used to determine for each locus the probability associated with the observed Z_{nS} value, conditioned on θ and the population recombination rate (see below). To evaluate the decay of LD, we measured the correlation coefficient r^2 (HILL and ROBERTSON 1968) for each pair of biallelic sites within loci and then plotted the pooled values across loci against the distances between the corresponding pair of sites (singletons,

being non informative, were discarded).

The availability, for most of our loci, of the *D. simulans* homologous sequence, allowed us to polarize the state of our biallelic sites. A variant was considered ancestral if observed in both species, and derived if segregating only in *D. melanogaster*. In the analysis of the European population, we were interested in the fraction of mutations that originated after the bottleneck, which we assumed to equal the single nucleotide polymorphisms (SNPs) exclusive to the European sample. We focused on all these k^{Eall} SNPs, or only the k^{Es} singletons. Then, to denote the polymorphisms that originated after the bottleneck in each locus i, we define $\theta_i^{\mathrm{E}} = \frac{k_i}{L_i}\sum_{j=1}^{n_i-1}\left(\frac{1}{j}\right)^{-1}$, where L_i and n_i are the length and the sample size, respectively. It follows that $\theta_i^{\mathrm{E}} = \theta_i^{\mathrm{Eall}}$ for $k_i = k_i^{\mathrm{Eall}}$, while $\theta_i^{\mathrm{E}} = \theta_i^{\mathrm{Es}}$ for $k_i = k_i^{\mathrm{Es}}$.

1.2.1.3. Demographic modeling of the African population

To investigate the hypothesis of size expansion of the African *D. melanogaster* population, we applied a maximum-likelihood method proposed by Weiss and von Haeseler (1998) which allowed us to extract information about the population history, such as the time, T_e (in years), at which the population started to increase, and the strength of the expansion, Δ, defined as the ratio of the current and the initial effective population size. This method is based on both π and θ, and requires that for each locus the parameters π_A, π_C, π_G, π_T (corresponding to the frequency of each base in the sequence), the transition/transversion parameter, κ, and the pyrimidine/purine transition parameter, ξ, are estimated. This was done with a program kindly provided by H. Li. Since loci in regions of low recombination are more affected by the impact of selection (see results, subsection 1.2.2.1.), we excluded them from the analysis. In addition, we excluded 15 loci due to undefined values of the parameters κ or ξ, leaving 214 loci for analysis. For each locus, the likelihood of a given parameter set was determined through 10,000 coalescent simulations without recombination by the program IPHULA, kindly provided by G. Weiss. To compute the likelihood for all loci, free recombination between loci was assumed. The estimate of T_e was obtained assuming for the current effective population size $N_0 = 10^6$ (Li *et al*. 1999) and ten generations per year. The confidence intervals (95% CI) for these estimates were obtained by the standard MAX–2 rule (*e.g.,* Kaplan and Weir 1995).

1.2.1.4. Demographic modeling of the European population

The population size bottleneck associated with the colonization of Europe likely resulted in a genome-wide reduction of polymorphism that confounds the signature of selective sweeps. However, the detection of a stronger reduction of heterozygosity than expected under a bottleneck model may be attributed to the action of positive selection. To investigate this hypothesis, we examined for each locus *i* the observed number of segregating sites k_i using coalescent simulations of a simple bottleneck model. Candidate loci of selective sweeps can thus be identified as those harboring less genetic variation than expected based on the chromosome-wide reduction of polymorphism due to a bottleneck.

To characterize the bottleneck, we assume that the population drops to a fraction of its current effective population size N_0 at time T_b (measured in units of $3N_0$, since we use X-linked data), and then recovers to N_0. We also assume that the duration of the bottleneck is short such that mutations arising during this time may be ignored. We combine duration and depth of the bottleneck in a compound parameter S_b, describing the strength of the bottleneck (GALTIER *et al*. 2000; see also FAY and WU 1999; APPENDIX A; Figure 3). The parameter S_b can be interpreted as the time that would be necessary to obtain the same number of coalescent events in a neutral constant-size population. Therefore, the total length of the coalescent tree T_{tree} will be a function of both T_b and S_b, *i.e.,* $T_{tree} = T_{tree}(T_b, S_b)$. The estimation of the bottleneck parameters T_b and S_b is done using a maximum-likelihood approach. That is, we consider

$$\mathrm{Lik}_i(T_b, S_b \mid K_i = k_i) \propto P_i(K_i = k_i \mid T_b, S_b, \theta_i) \ , \qquad (1)$$

where Lik_i refers to the likelihood and K_i to the average number of differences between two sequences at locus *i*. We consider loci to be independent (*i.e.,* we assume free interlocus, but no intralocus recombination) and consequently calculate the joint likelihood for the *m* loci by multiplying their individual likelihoods. Then, $\hat{T}_b$ and $\hat{S}_b$ are the maximum-likelihood estimates that maximize the product $\prod_{i=1}^{m} P_i$. Note that the maximum-likelihood estimate applies to the pair ($\hat{T}_b$, $\hat{S}_b$), rather than to $\hat{T}_b$ and $\hat{S}_b$ individually.

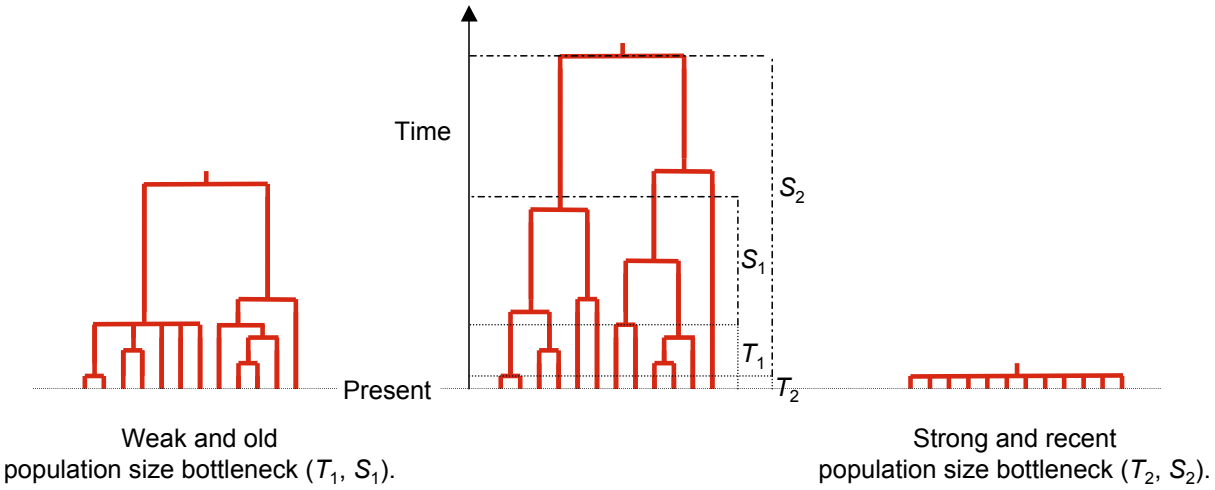


Figure 3. Schematic representation of the demographic model applied to the coalescent simulations. The standard coalescent tree is shown in the center: after a time T_b backward in time, a bottleneck of strength S_b is applied to the tree. The strength S_b corresponds to the time that would be necessary to get the same number of coalescent events under a constant population-size model. The two scenarios illustrate examples of an old (T_1) weak (S_1) bottleneck and a recent (T_2) strong (S_2) bottleneck (modified from GALTIER *et al*. 2000).

To calculate the likelihood of each locus, we simulate G genealogies (see below) and define for each locus i

$$P_i = \frac{1}{G}\sum_{g=1}^{G} P_{i,g} \quad , \qquad (2)$$

where $P_{i,g}$ is the probability of genealogy g for locus i. We used various methods to calculate $P_{i,g}$, each resulting in a different set ($\hat{T}_b$, $\hat{S}_b$). An overview is given in Table 2. In method I, we considered the probability to observe $K_i = k_i$ SNPs in a coalescent tree of length $T_{tree,g}$, based on the Poisson distribution

$$P^{I}_{i,g}(K_i = k_i \mid T_b, S_b, \theta_i) = \frac{(\lambda_{i,g})^{k_i}}{k_i!} e^{-\lambda_{i,g}} \quad , \qquad (3)$$

where $\lambda_{i,g} = L_i \theta_i T_{tree,g}$ and L_i is the length in base pairs of locus i.

In method II, we analyzed a subset w of our m loci, for which we were able to polarize the state of the k_w observed SNPs (see above), where $k_w = \sum_{i \in w} k_i$. We define k_w^A and k_w^E as the polymorphisms in this subset that occurred before and after the colonization of Europe,

Table 2. Demographic modeling of the European population – overview of methods.

Method	$k_i = k_i^A + k_i^E$ [a]	$k_i^E = k_i^{Es}$ [b]	$k_i^E = k_i^{Eall}$ [c]	θ_i and $\bar{\theta}$ [d]
I	no	no	no	no
II^s	yes	yes	no	no
II^{all}	yes	no	yes	no
III^s	yes	yes	no	yes
III^{all}	yes	no	yes	yes

We simulated genealogies under a simple bottleneck model, and estimated its age and strength computing the probability to obtain the k_i SNPs in each locus *i*, based on the Poisson distribution. The following methods were used to model the demographic history of the European population:

[a] Pre- and post-bottleneck SNPs were distinguished.

[b] Only the European private singletons were assumed to have arisen in the post-bottleneck phase.

[c] All European private SNPs were assumed to have arisen in the post-bottleneck phase.

[d] The mutation process of the pre- and post-bottleneck phases were treated separately; in the post-bottleneck phase, the population mutation parameter was set equal to the observed average value of θ for the African sample, $\bar{\theta}$. Otherwise, the corresponding θ value observed at locus *i* in the African sample, θ_i, was used.

respectively (see above), such that $k_w = k_w^{\mathrm{A}} + k_w^{\mathrm{E}}$. We assume that the African sample represents the ancestral population. The k_w^{E} SNPs were either assumed to be identical to k_w^{Es} (method II^s) or to k_w^{Eall} (method II^{all}). The partitioning of the k_w SNPs into k_w^{A} and k_w^{E} SNPs is correlated with the proportion of the time spent by the species in "Africa" and "Europe" and will therefore depend on the demographic history. For a single locus *i* write $k_i = k_i^{\mathrm{E}} + k_i^{\mathrm{A}}$. Then,

$$P_{i,g}^{\mathrm{II}}(K_i^{\mathrm{E}} = k_{\mathrm{i}}^{\mathrm{E}}, K_i^{\mathrm{A}} = k_{\mathrm{i}}^{\mathrm{A}} \mid T_{\mathrm{b}}, S_{\mathrm{b}}, \theta_i) = \frac{(\lambda_{i,g})^{k_i^{\mathrm{E}}+k_i^{\mathrm{A}}}}{(k_i^{\mathrm{E}}+k_i^{\mathrm{A}})!} e^{-\lambda_{i,g}} \times \binom{k_i^{\mathrm{E}}+k_i^{\mathrm{A}}}{k_i^{\mathrm{E}}} \left(\frac{T_{\mathrm{tree},g}^{\mathrm{E}}}{T_{\mathrm{tree},g}}\right)^{k_i^{\mathrm{E}}} \left(\frac{T_{\mathrm{tree},g}^{\mathrm{A}}}{T_{\mathrm{tree},g}}\right)^{k_i^{\mathrm{A}}}, \quad (4)$$

where $T_{\mathrm{tree}}^{\mathrm{A}}$ and T_{tree}^{E} measure the lengths of the coalescent tree portions after and before T_{b} (going backward in time), respectively, and $T_{\mathrm{tree}} = T_{\mathrm{tree}}^{\mathrm{A}} + T_{\mathrm{tree}}^{\mathrm{E}}$.

Finally, a third method was used, where we considered the probability to observe

independently k_w^{A} and k_w^{E} SNPs in the pre- and post-bottleneck phases, respectively,

$$P_{i,g}^{\mathrm{III}}(K_i^{\mathrm{E}} = k_i^{\mathrm{E}}, K_i^{\mathrm{A}} = k_i^{\mathrm{A}} \mid T_{\mathrm{b}}, S_{\mathrm{b}}, \theta_i, \bar{\theta}) = \frac{(\lambda_{i,g}^{\mathrm{E}})^{k_i^{\mathrm{E}}}}{k_i^{\mathrm{E}}!} e^{-\lambda_{i,g}^{\mathrm{E}}} \times \frac{(\lambda_{i,g}^{A})^{k_i^{\mathrm{A}}}}{k_i^{\mathrm{A}}!} e^{-\lambda_{i,g}^{\mathrm{A}}}, \qquad (5)$$

where $\lambda_{i,g}^{\mathrm{E}} = L_i \theta_i T_{\mathrm{tree},g}^{\mathrm{E}}$ and $\lambda_{i,g}^{\mathrm{A}} = L_i \bar{\theta} T_{\mathrm{tree},g}^{\mathrm{A}}$. As before, we considered either $k_w^{\mathrm{E}} = k_w^{\mathrm{Es}}$ (method III$^{\mathrm{s}}$) or $k_w^{\mathrm{E}} = k_w^{\mathrm{Eall}}$ (method III$^{\mathrm{all}}$).

Note that the value of the mutational parameter was estimated in two ways. In the first (for method III only), we considered the observed average value of θ for the African sample, $\bar{\theta}$. In the second (for methods I–III), θ_i is locus specific; *i.e.,* it is the value observed in the African sample at locus *i*. This is reasonable because there is a strong positive correlation between the θ_i values found in the African and in the European samples ($R = 0.527$, $P < 0.0001$).

For each method, we tested more than 600 T_{b} and S_{b} combinations, characterizing severe recent bottlenecks (*i.e.,* $T_{\mathrm{b}} = 0.0005$ and $S_{\mathrm{b}} = 2.0$) to shallow old ones (*i.e.,* $T_{\mathrm{b}} = 0.2$ and $S_{\mathrm{b}} = 0.02$). For each locus and parameter combination, $G = 2{,}500$ genealogies were simulated using a modified version of a program kindly provided by S. E. Ramos-Onsins (Ramos-Onsins *et al*. 2004), which is based on the program 'ms' (Hudson 2002). To evaluate the fit of our model to the data, we used the estimated bottleneck parameter sets to simulate $G = 10{,}000$ genealogies for each locus and calculated for each of the 10,000 resulting samples the average of Tajima's *D* and Kelly's Z_{nS} statistics. For the latter analysis, our simulation methodology did not consent to use simultaneously both θ_i and $\bar{\theta}$. Thus, only θ_i was used as mutation parameter in method III.

To identify candidate loci for positive selection, we used two approaches that tested each locus independently against the expectations under the estimated bottleneck models. In the first one (for methods I and II), $G = 10{,}000$ coalescent simulations were carried out for each locus *i* to compute the probability to harbor at most k_i segregating sites,

$$Q_i = P_i^{\mathrm{I}}\left(K_i \le k_i \mid \hat{T}_{\mathrm{b}}, \hat{S}_{\mathrm{b}}, \theta_i\right) = \frac{1}{G}\sum_{g=0}^{G}\sum_{j=0}^{k_i} P_{i,g}^{\mathrm{I}}(K_i = j \mid \hat{T}_{\mathrm{b}}, \hat{S}_{\mathrm{b}}, \theta_i) \ . \qquad (6)$$

The second one enabled us to use the information on the mutational and demographic history of the sample (method III). G = 10,000 coalescent simulations (for each locus i) were used to calculate the proportion of simulated genealogies, thereby generating at the same time (i) at most k_i^{E} segregating sites in the portion T_{tree}^{E} of the simulated tree with mutational parameter $\overline{\theta}$, and (ii) at most k_i^{A} segregating sites in the portion $T_{\mathrm{tree}}^{\mathrm{A}}$ of the simulated tree with mutational parameter θ_i. We call this probability Q_i^{E}. Then, a locus was considered to lie in a candidate sweep region if it contained fewer polymorphisms than expected under the estimated bottleneck model; *i.e.,* when $Q_i < 0.05$ or $Q_i^{\mathrm{E}} < 0.05$.

Since hitchhiking and bottlenecks affect π more than θ, using k_i in the identification of targets of selection is conservative.

An overview of the computer programs used in the analysis is given in APPENDIX C.

1.2.2. RESULTS

1.2.2.1. Polymorphism patterns of the African population

Before embarking on our sequencing study, we confirmed that all African (and European) X chromosomes are inversion-free. From the African X chromosomes, we gathered sequencing data for 153 fragments. Together with the previously published data for 100 loci (see MATERIALS AND METHODS, SUBSECTION 1.2.1.1.), this increased the density of our scan of the X chromosome considerably. The average distance between adjacent loci for the whole chromosome is now < 70 kb, and for the proximal half < 50 kb. Fragment length ranges from 199 to 781 bp, with a mean (standard error, SE) of 510.4 (6.9) bp. We sequenced a total of 129,133 nucleotide sites (excluding insertions and deletions), of which 4,922 are polymorphic. Over half of the observed polymorphic sites are in low frequency (*i.e.,* singletons). A summary of the polymorphism and divergence data of all analyzed loci is provided in the APPENDIX B (Table B1).

The mean levels of diversity (SE) across the 253 loci are 0.0114 (0.0004) for π and 0.0131 (0.0004) for θ. When levels of nucleotide diversity are plotted against recombination rate (Figure 4a and 1b), we observe a significantly positive correlation (for π, Spearman's

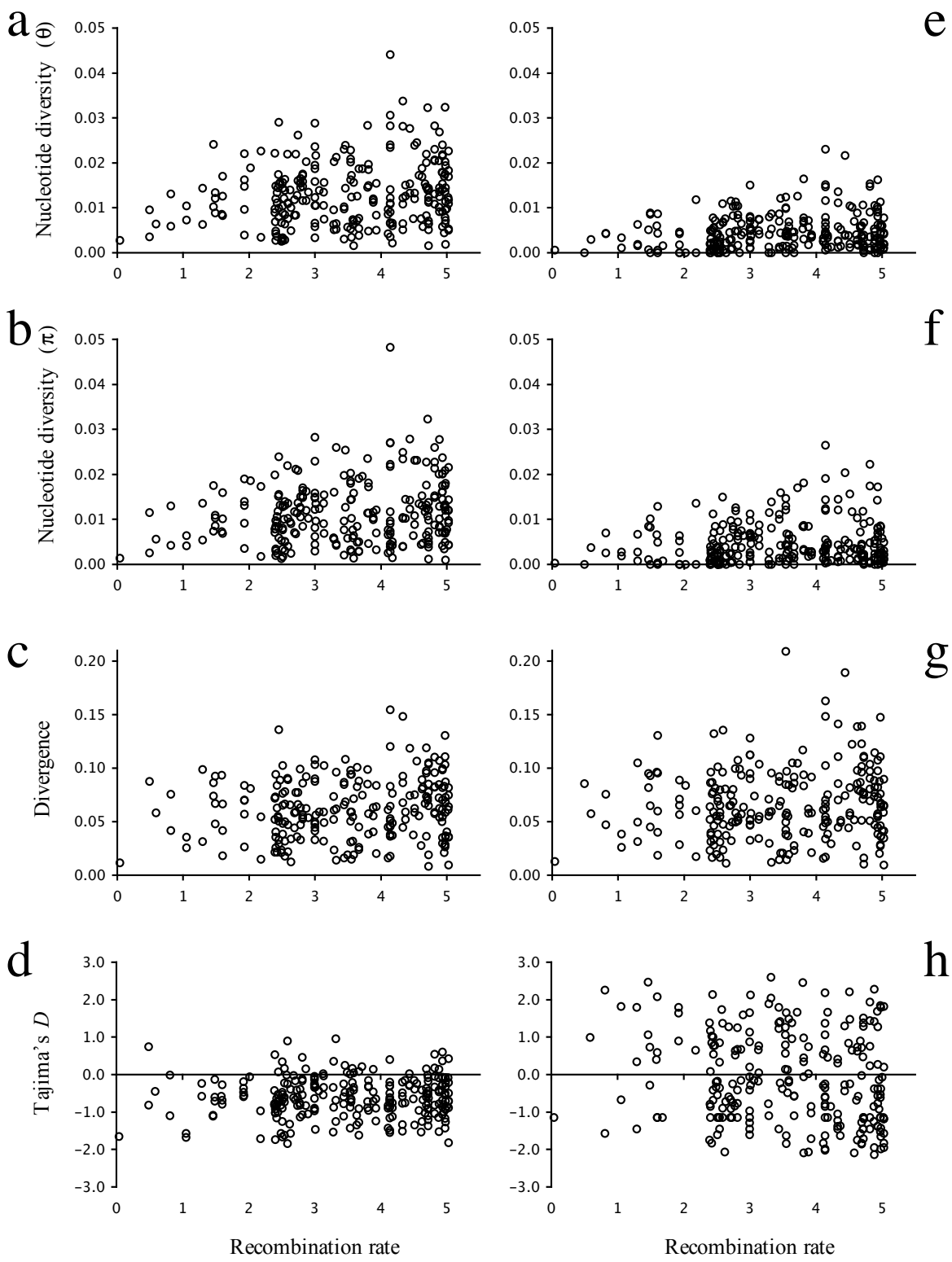


Figure 4. Nucleotide diversity, divergence and Tajima's *D* values versus recombination rate for the African (a–d) and the European (e–h) populations. Recombination rate is expressed in recombination events per site per generation $\times 10^8$ (Comeron *et al*. 1999).

$R = 0.140$, $P = 0.026$; for θ, Spearman's $R = 0.147$, $P = 0.020$; hereafter, all correlations are tested using Spearman's R). Furthermore, a weak correlation between recombination rate and divergence across all sequenced 232 loci is found ($R = 0.127$, $P = 0.054$; Figure 4c). The average divergence (SE) between the African sample and *D. simulans* is 0.0621 (0.0019). To investigate this correlation more closely, we divided the data set into fragments of low (region I) and normal to high recombination rates (region II; MATERIALS AND METHODS, SUBSECTION 1.1.1.4.). This approach, corresponding to a threshold value of 2×10^{-8} recombination events per base pair per generation, yields 24 fragments for region I. The correlation between levels of nucleotide diversity and recombination rates still exists in region I (for π, $R = 0.459$, $P = 0.024$; for θ, $R = 0.510$, $P = 0.011$), but is weaker in region II (for π, $R = 0.114$, $P = 0.087$; for θ, $R = 0.116$, $P = 0.080$). The opposite is seen when we correlate levels of divergence with recombination rates. A significant correlation is observed in region II ($R = 0.142$, $P = 0.040$), whereas none is found in region I ($R = 0.085$, $P = 0.702$). These observations also hold for the second measure of recombination rate (MATERIALS AND METHODS, SUBSECTION 1.1.1.4.), except that levels of nucleotide diversity and recombination rate are not correlated in region I (for π, $R = 0.491$, $P = 0.125$; for θ, $R = 0.527$, $P = 0.096$). The observed correlation between nucleotide diversity and crossing-over rate agrees with the results of several previous studies (*e.g.,* BEGUN and AQUADRO 1992), whereas the correlation between divergence and crossing over came as a surprise and requires further explanation (see Discussion).

Our data on intraspecific polymorphism and interspecific divergence were also used to test if the population departs from neutral equilibrium. Under neutrality, genome regions with high sequence diversity within a species should also show high levels of divergence between species (HUDSON *et al*. 1987). To test if the data depart from this expectation, we used the multi-locus HKA test. No departure from the neutral equilibrium model was detected ($X^2 = 184.15$, $P = 0.990$).

To investigate the haplotype structure in the African sample, we used the number of haplotypes, K_{DV}, and haplotype diversity, H_{DV}, and examined their values across the chromosome as described in SUBSECTION 1.1.1.3. Under neutrality, we expect an equal proportion of the observed values of K_{DV} and H_{DV} being lower and higher than the simulated median. We observed a significant excess of loci with higher values than expected for both statistics (sign test, two-tailed, $P < 0.001$ for both K_{DV} and H_{DV}). High values can result

from a star-like genealogy following population expansion, or from an old complete sweep (Depaulis and Veuille 1998). In the latter case, due to recombination, recurrent selective sweeps across the chromosome are expected to leave footprints of partial hitchhiking. We searched for such evidence using the K_{DV}- and H_{DV}-haplotype tests (Depaulis and Veuille 1998) and Fay and Wu's *H* test (Fay and Wu 2000), with the conservative assumption of zero recombination. We observed only one significant value of the H_{DV} statistic (locus *728*) and another five with a significantly negative Fay and Wu's *H* value (one-tailed, $P < 0.05$), namely loci *276*, *295*, *310*, *392* and *483*. The average *H* value (SE) across all loci was –0.583 (0.155) indicating a significant skew toward high-frequency derived variants ($P = 0.007$).

Star-like genealogies typically produce a skew in the frequency spectrum toward rare variants that can be detected by negative values of Tajima's *D* statistic. Indeed, most loci ($m = 225$) have negative *D* values (sign test, two-tailed, $P < 0.0001$; Figure 4d), 21 of which also depart from neutral equilibrium ($P < 0.05$). The observed average value (SE) of –0.608 (0.033) was compared with the prediction of the standard neutral model using coalescent simulations (subsection 1.1.1.3.). None of 10,000 simulated samples had either a more negative average or a smaller variance than the observed values. Furthermore, we did not find a positive correlation between *D* and recombination rate ($P = 0.426$; Figure 4d), in contrast to the prediction of the recurrent hitchhiking model (Braverman *et al*. 1995; Andolfatto and Przeworski 2001). This observation holds for both measures of recombination rates (data not shown).

We analyzed linkage disequilibrium (LD), using Kelly's statistic Z_{nS} (Kelly 1997). Most of the values are very low, with an average (SE) of 0.139 (0.004). To assess whether they are consistent with the neutral equilibrium scenario, we performed coalescent simulations with recombination. This assumption is conservative, since recombination decreases Z_{nS} (Kelly 1997). We observed 78 of these loci with a Z_{nS} value significantly lower than expected under a neutral equilibrium model (one tailed, $P < 0.05$), pointing to a deficiency of LD across the chromosome. The short-range behavior of LD in the chromosome was studied using the statistic r^2 (Hill and Robertson 1968). Pooling the data from all loci resulted in 13,930 values. The correlation between r^2 values and distance clearly indicates a strong decay of LD with distance, as LD drops ~60% over the average fragment length (*i.e.,* ~500 bp; Figure 5a).

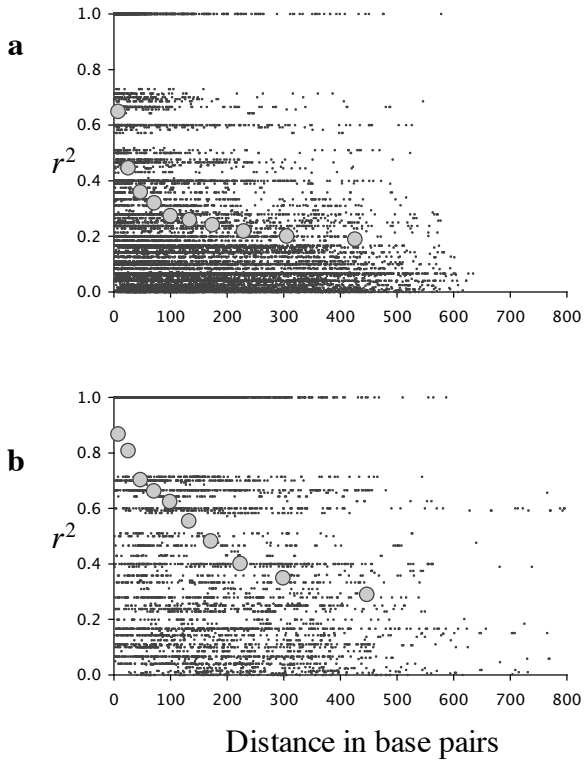


Figure 5. Decay of linkage disequilibrium with distance in the African (a) and the European (b) populations. The squared correlation coefficient of allele frequencies between biallelic sites (r^2) is plotted against distance in base pairs across loci. The average r^2 values for 10 subsets containing an equal number of site pairs (pooled based on the distance) are plotted as filled circles.

Both observations (*i.e.,* a chromosome-wide excess of low-frequency variants and a lack of LD) are consistent with an expansion of the African population (see DISCUSSION).

1.2.2.2. Polymorphism patterns of the European population

We gathered polymorphism data from a total of 263 fragments, spanning 142,135 nucleotide sites (gaps were excluded), 1,925 of which are polymorphic. Number of segregating sites, diversity indices, and basic statistics for each locus are shown in the APPENDIX B (Table B2).

The means (SE) of π and θ across the X chromosome are 0.0047 (0.0003) and 0.0046 (0.0002), respectively. We found 22 loci (18 intronic and 4 intergenic regions) with no polymorphism. When the estimates of nucleotide diversity are plotted against recombination rate, a significant positive correlation is found for θ, but not for π ($R = 0.135$, $P = 0.028$, and $R = 0.079$, $P = 0.201$, respectively; Figure 4e–f). If the lower θ values in regions of reduced recombination were due to a lower mutation rate, they should also be less diverged. Divergence across 241 loci between *D. simulans* and the European population is on average (SE) 0.0666

(0.0021). In contrast to the African sample, it does not correlate with recombination rate (R = 0.100, P = 0.123), even when the data are partitioned into those of recombination regions I and II (results not shown). Furthermore, no significant correlation was found when the ratio between θ and divergence was plotted against recombination rate (R = 0.105, P = 0.102).

Tajima's D was calculated for each locus. We observed an average (SE) of –0.103 (0.079). Coalescent simulations (see SUBSECTION 1.1.1.3.) showed that the average does not deviate from the neutral expectation (P = 0.077), while its standard error is too large (P = 0.0001). We found a strong positive correlation between haplotype diversity and D (R = 0.611, P < 0.0001), reflecting the observation that positive D values are caused by two almost equally frequent haplotypes and negative values by rare divergent haplotypes or by few rare variants (distributed across the entire sample).

Surprisingly, there is a weak but highly significant negative correlation between D and recombination rate (R = –0.188, P = 0.003), whereas no significant correlation was found between haplotype diversity and recombination rate (R = 0.019, P = 0.764). This suggests that the negative correlation between D and recombination rate is due to an excess of rare variants that are distributed over more than one sequence of the sample, rather than being comprised in a single diverged haplotype.

A typical signature of a recent bottleneck is an elevated level of linkage disequilibrium. The average (SE) statistic Z_{nS} across loci is 0.438 (0.018), much higher than the value observed in the African sample and significantly higher than the value at neutral equilibrium (P < 0.0001; 26 loci had significantly higher LD than expected, *i.e.,* P < 0.05; no recombination was assumed). No correlation with the recombination rate was found (data not shown). When the pooled r^2 values across loci are plotted against distance (6162 data points; MATERIALS AND METHODS, SUBSECTION 1.2.1.2.), LD is observed to decay ~60% over the average fragment length (*i.e.,* ~500 bp; Figure 5b).

The multi-locus HKA test showed significant departure from neutrality (X^2 = 483.40, P < 10^{-5}). The statistic was no longer significant if 48 loci (with the largest contributions) were removed (data not shown). The observation that 30 of these show an excess of polymorphism agrees with the frequently observed haplotype structure.

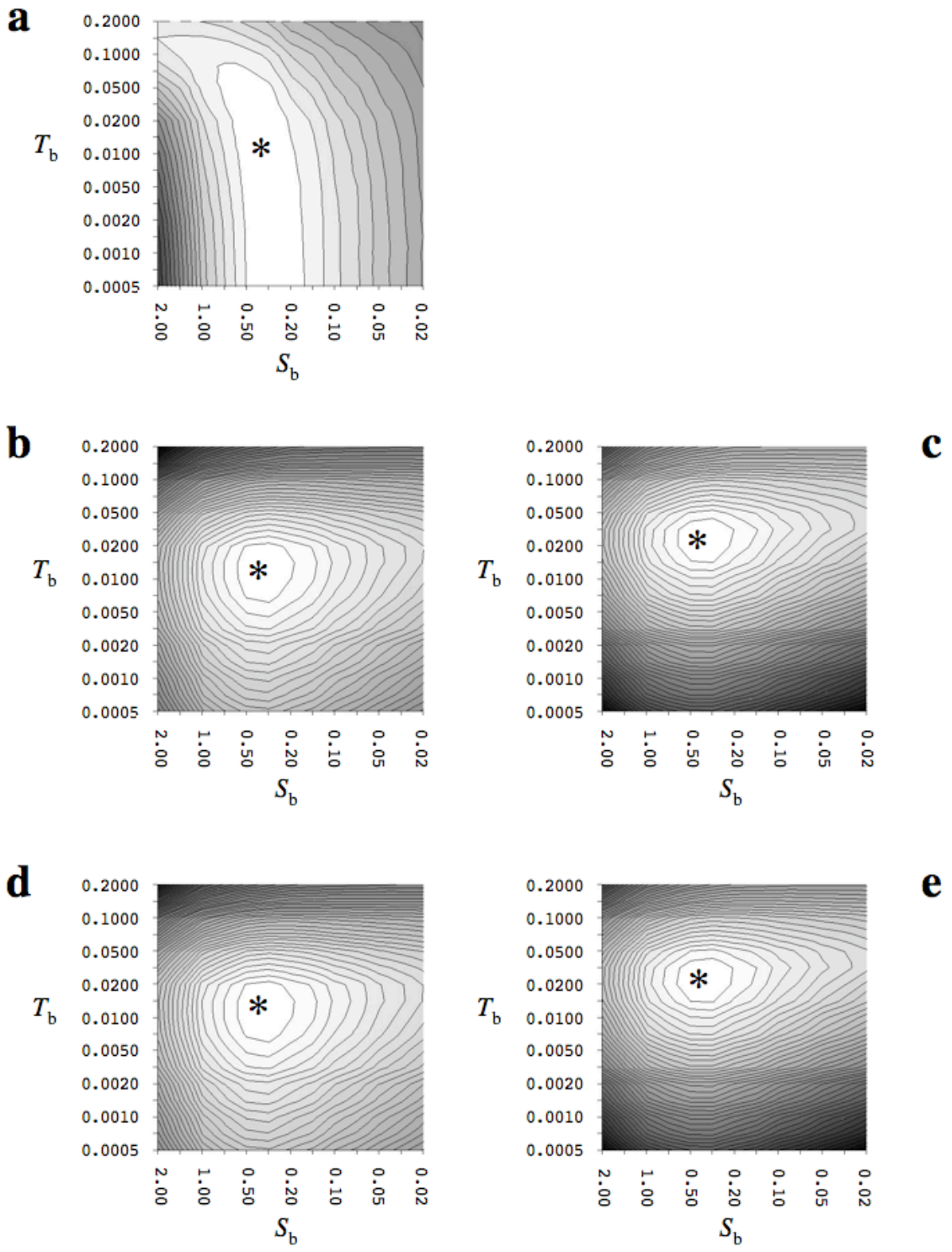


Figure 6. Log-likelihood surfaces for the estimation of the bottleneck associated with the colonization of Europe (the lighter, the larger is the likelihood; contour lines define log-likelihood intervals of 200). Maximum-likelihood estimates of age, T_b (in units of $3N_0$ generations) and strength, S_b, of the bottleneck were obtained exploring the parameter space ranging from severe recent bottlenecks (*i.e.*, $T_b = 0.0005$ and $S_b = 2.0$) to shallow old ones (*i.e.*, $T_b = 0.2$ and $S_b = 0.02$) through coalescent simulation. Asterisks designate the approximate maximum-likelihood parameters' estimates. Plots for method I (a), II (b and c) and III (d and e) are shown (Tables 2 and 3; see MATERIALS AND METHODS, SUBSECTION 1.2.1.4.).

1.2.2.3. Estimating the parameters of a simple bottleneck model for the European population

Several observations reported above indicate the occurrence of bottlenecks during the colonization of Europe by *D. melanogaster*. While a bottleneck results in the loss of heterozygosity at a genome-wide scale, selective sweeps may lead to a further reduction of polymorphism locally in the genome. In order to disentangle the effects of these two forces, we developed a coalescent-based approach that compares the behavior of a single locus against that of all loci on the X chromosome (MATERIALS AND METHODS, SUBSECTION 1.2.1.4.). We assume a simple bottleneck model such that only a single reduction of population size occurred during colonization. The maximum-likelihood estimates of the bottleneck parameters are obtained by five different procedures (Table 2). For method I, we analyzed all loci for which both populations had been sequenced ($m = 250$); for methods II and III, the *D. simulans* homologs were required ($m = 230$).

The likelihood surfaces and the maximum-likelihood estimates for the age, T_b, and the strength, S_b, of the bottleneck obtained by our methods (Table 2) are shown in Figure 6 and Table 3, respectively. Only when the partitioning of segregating sites between pre- and post-bottleneck SNPs is used (methods II and III), a clear maximum becomes apparent (Figure 6b–e versus Figure 6a). Nonetheless, parameters estimated by method I are in close agreement with those estimated by methods II^s and III^s, where only the European private singletons were equated with the SNPs that originated after the bottleneck. On the other hand, methods II^{all} and III^{all} that classified all European private polymorphisms as post-bottleneck ones, point to a more ancient bottleneck (Table 3). Hereafter, bottlenecks will be identified according to the parameter set used; *e.g.,* bottleneck II^{all} refers to the bottleneck simulated with $\hat{T}_b^{II^{all}}$ and $\hat{S}_b^{II^{all}}$.

To evaluate the fit of our model to the data, we compared the observed averages of Tajima's D and Kelly's Z_{nS} across loci to those expected under the estimated bottlenecks. All simulated bottlenecks produced average D and Z_{nS} values close to the observed ones (Table 3). For Z_{nS}, incorporating recombination in our simulations results in an average of ~0.230 (for all methods; data not shown), suggesting that some recombination is needed to have a better fit. However, the empirical averages are statistically compatible only with the (older)

Table 3. Demographic modeling of the European population – results.

		Bottleneck parameters		Polymorphism		Average Tajima's *D*			Average linkage disequilibrium			Candidate loci [g]		
		$\hat{T}_b$ [a]	$\hat{S}_b$ [b]	F_0 [c]	$P(F_0 \geq 20)$	$\overline{D}$ [d]	SE [e]	95% CI [f]	$\overline{Z}_{nS}$ [d]	SE [e]	95% CI [f]	Low poly.	Tajima's *D* (-; +)	Z_{nS} (-; +)
Observed				20		–0.103	0.078		0.430	0.019			80 (51; 29)	35 (9; 26)
Simulation method	I	0.0125	0.350	21.9	0.384	0.251	0.087	0.080–0.421	0.623	0.019	0.577–0.660	4	13 (9; 4)	12 (12; 0)
	II^s	0.0128	0.345	19.4	0.476	0.250	0.087	0.078–0.420	0.618	0.019	0.573–0.654	4	12 (9; 3)	12 (12; 0)
	II^{all}	0.0267	0.400	15.5	0.136	–0.069	0.084	–0.245–0.097	0.506	0.020	0.455–0.543	8	19 (9; 10)	13 (8; 5)
	III^s	0.0122	0.371	18.1	0.358	0.237	0.088	0.053–0.409	0.631	0.020	0.581–0.669	3	---	---
	III^{all}	0.0264	0.380	8.4	< 0.0001	–0.033	0.084	–0.205–0.133	0.510	0.020	0.462–0.547	24	---	---
Shared [h]												3	12 (9; 3)	8 (8; 0)

Bottleneck parameters were estimated according to the methods described in MATERIALS AND METHODS, SUBSECTION 1.2.1.4. (see also Table 2).

[a] Age of the bottleneck, measured in units of $3N_0$ generations.

[b] Strength of the bottleneck.

[c] Number of loci with no variation.

[d] Average across loci, averaged over 10,000 simulation runs.

[e] Standard error across loci, averaged over 10,000 simulation runs.

[f] Confidence interval.

[g] Number of outlier loci. Low polymorphic loci are those with the probability to obtain at most the observed number of segregating sites < 0.05. For methods III^s and III^{all} probability Q^E was used, otherwise Q was used. We also report the total number of significant Tajima's *D* and linkage disequilibrium Z_{nS} values assessed from simulations based on the estimated bottlenecks (in parentheses, the number of significantly negative/low and positive/high values). For the last statistics, a calculation was not feasible for methods III^s and III^{all} (MATERIALS AND METHODS, SUBSECTION 1.2.1.4.). In the "Observed" row, the significant values departing from the standard neutral model are shown.

[h] Number of candidate loci identified by all simulation methods.

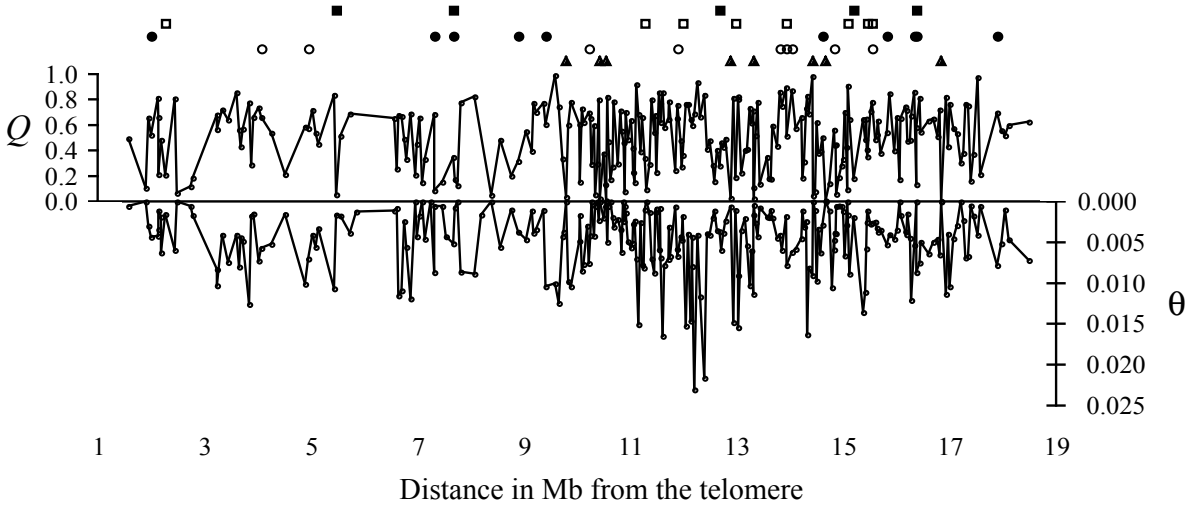


Figure 7. Probability Q to obtain at most the observed number of segregating sites in the European sample for a given locus (given the bottleneck estimated by method II^{all}) against locus position on the X chromosome (based on the *D. melanogaster* genome release 3.2). The corresponding values of nucleotide diversity θ are plotted below the x axis. In the upper part of the figure, the position of outliers ($Q < 0.05$) are denoted by filled triangles; empty and filled circles denote the position of loci with significantly negative or positive Tajima's D values, respectively; empty and filled squares denote significant deficiency or excess of linkage disequilibrium, respectively, as measured by Z_{nS}. Significance was calculated given the bottleneck estimated by method II^{all}.

bottlenecks II^{all} and III^{all}, while they do not lie within the 95 % confidence interval of the other three bottlenecks.

Among the loci used in these simulations, we observed 20 with no polymorphism. For each locus and genealogy, we generated Poisson-distributed mutations according to the different mutation processes (see MATERIALS AND METHODS, SUBSECTION 1.2.1.4.). In each case, an average number of loci with no segregating sites close to 20 was generated, with the exception of those of bottleneck III^{all}, where on average only 8.42 loci harbored no variation (Table 3).

These results indicate a good agreement between our model and the empirical data. In particular, method II^{all} produces estimates of the bottleneck parameters that appear to be consistent with the polymorphism pattern observed in the European sample.

1.2.2.4. Identifying candidate sweep regions in the European population

We attempted to identify loci not compatible with the estimated bottlenecks. Each locus was tested against the model by calculating the probability Q (or Q^E) that it harbors at

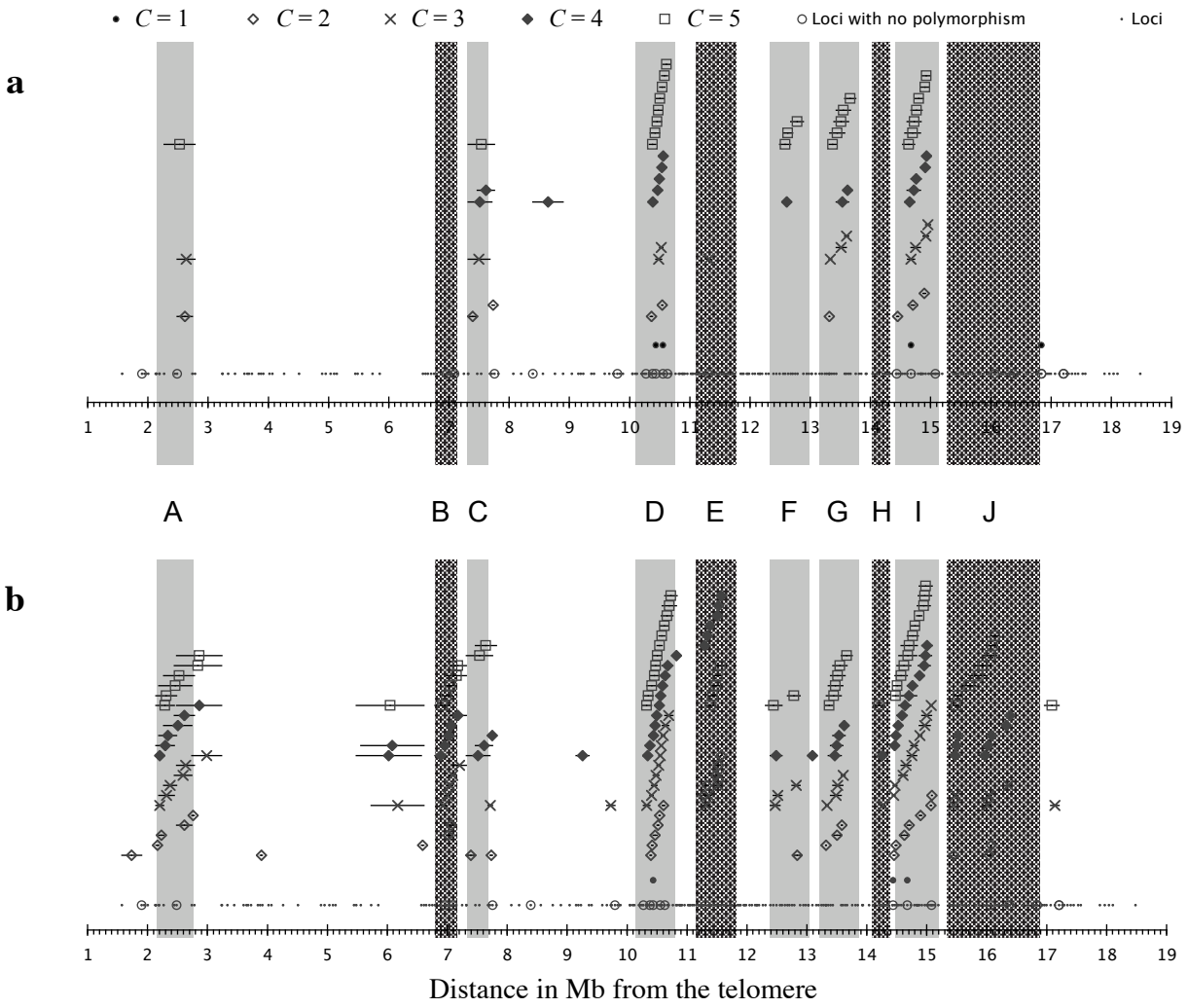


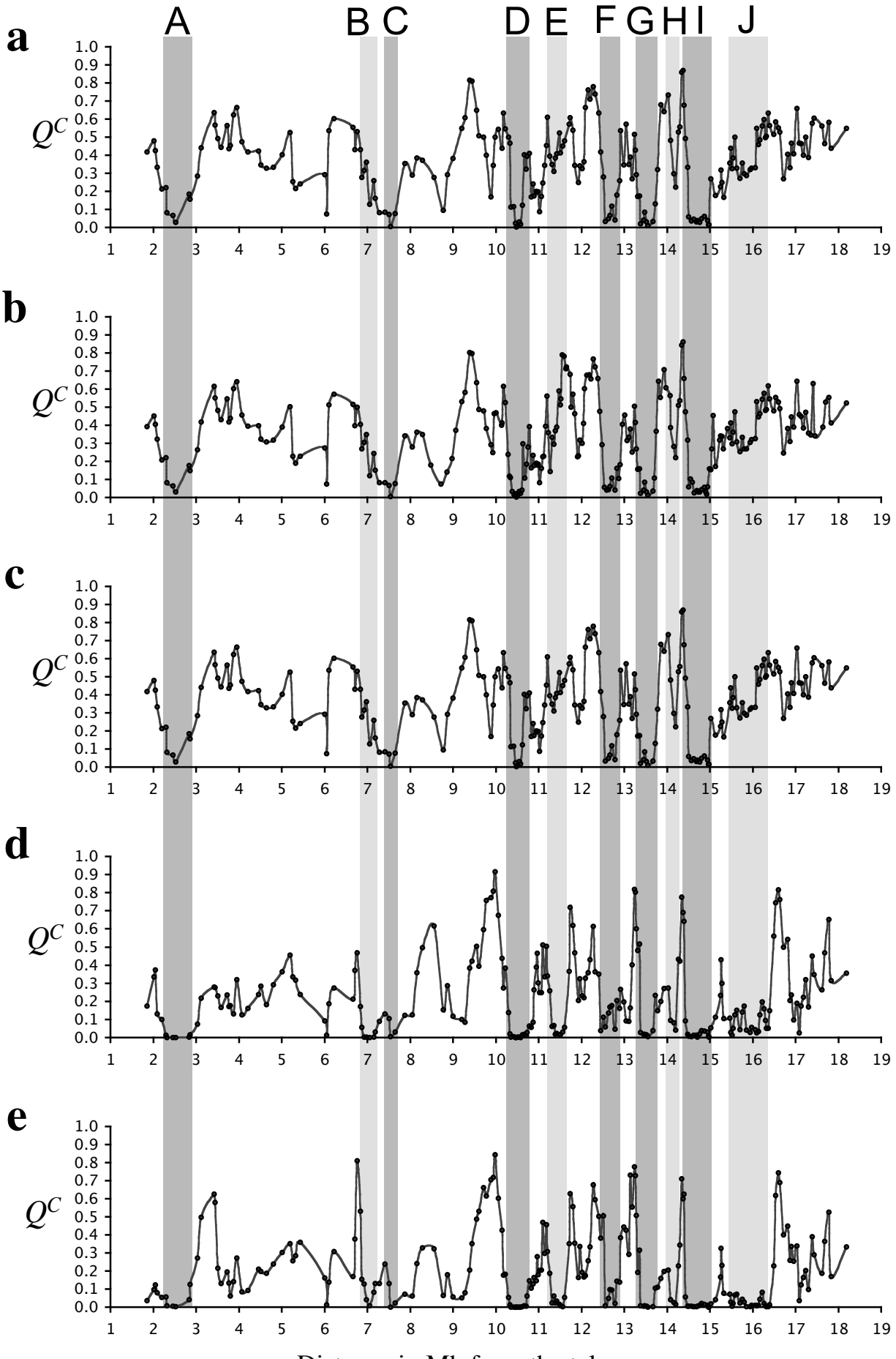
Figure 8. Identification of single loci and sets of *C* consecutive loci with fewer polymorphisms than expected under the simulated bottlenecks [estimated by method II^s (a) and III^s (b)]. Only significant Q^C values are shown, for C = 1–5. Points next to the x axis indicate each locus' position along the X chromosome (based on the *D. melanogaster* genome release 3.2). Horizontal bars are proportional to the extension of the region covered by the *C* consecutive loci. Candidate regions (A–J) for the action of positive selection are shadowed in gray: dark ones identify those supported by all methods (see text for details).

most the observed number of segregating sites (Figure 7; MATERIALS AND METHODS, SUBSECTION 1.2.1.4.). Depending on the estimated age of the bottleneck, 3 to 24 loci cannot be explained by demography alone ($Q < 0.05$ or $Q^E < 0.05$; Table 3; a complete list of the probabilities is given in APPENDIX B, Table B3). We note that, of the 20 loci with no polymorphism, 16 were among the 24 outliers identified by bottleneck III^{all}. These results indicate that demography accounts for most of the chromosome-wide lack of variation, but does not explain the effect alone. This is also seen by pooling all loci. For each bottleneck scenario, the probability to observe at most the observed total number of SNPs was $P < 0.001$.

The probability Q (calculated using method III^{all}) strongly correlates with the decrease in heterozygosity of the European sample relative to the African one (*i.e.,* the ratio between the respective θ values; $R = 0.571$, $P < 0.0001$). This is consistent with our simulation approach, which aimed to identify those loci that lost more of the ancestral variation than expected under a simple bottleneck model (MATERIALS AND METHODS, SUBSECTION 1.2.1.4.).

In contrast, since we did not use the frequency spectrum in defining the outliers, Tajima's D shows no correlation with Q, nor does linkage disequilibrium (*e.g.,* for bottleneck II^{all}, $R = 0.111$, $P = 0.110$, and $R = 0.016$, $P = 0.831$, respectively). The estimated bottleneck parameters were then used to assess the significance of Tajima's D and Z_{nS} by calculating, for each locus i, the proportion of simulations ($G = 10{,}000$) that produced values more extreme than the observed one (*i.e.,* AKEY *et al*. 2004). Since in our simulations we could use only one mutation parameter at the time, this analysis was possible only for methods I, II and II^{all} (MATERIALS AND METHODS, SUBSECTION 1.2.1.4.). Depending on the bottleneck parameters, 12 to 19 loci have too extreme values of Tajima's D and 12 to 13 too extreme values of Z_{nS} ($P < 0.05$; Table 3; Figure 7). Significantly low values of Z_{nS} were usually observed in loci with few singletons. Two loci showed both significantly high Tajima's D and high Z_{nS}. These loci are not associated with the outliers detected by both methods II^{all} and III^{all} (permutation test, 1000 permutations, $P > 0.5$ for both Tajima's D and Z_{nS}).

We also addressed the multiple testing problem by considering the false discovery rate as proposed by STOREY (2002). None of the above findings was still significant after correction. Obviously, this is due to the low power of our single locus approach owing to the limited number of SNPs in each fragment. The relative short sizes of the fragments may contain too little information to recognize the effects of positive selection, particularly in regions where the level of ancestral variation was low (*e.g.*, regions of low recombination rate). To overcome this limitation, we analyzed pooled data by simultaneously analyzing more than one locus. Let Q^C be the proportion of simulation runs generating at most k_j^C segregating sites in j sets of C consecutive loci. For methods III, where the mutational process is different from the other methods, $Q^C = Q^E$, with $k_j^{E,C} = k_i^E$ (MATERIALS AND METHODS, SUBSECTION 1.2.1.4.). For a large number of simulations we assume that Q^C ($K \leq k^C$) converges to Q. This was verified for $C = 1$, when $k_j^C = k_j$ (data not shown). We simulated 10,000 samples under the estimated bottlenecks, treating loci as independent. Then, we used


A
B C
D E
F G H I
J
a
b
c
d
e
Q^C
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Distance in Mb from the telomere

a sliding-window approach and considered, in parallel analyses, $C = 1$ to 5 consecutive loci (with step size = 1). A consequence of this method is a smoother variation of Q^C across the chromosome as C gets larger (compare Figures 7 and 8), which permits a better definition of the regions with less variation than expected.

We considered regions identified by consecutive loci of variable size C with $Q^C < 0.05$ as the most likely candidates for the recent action of positive selection (Figures 8 and 9; APPENDIX B, Table B3). As for the single-locus approach, their number varies with the method used, from 6 to 10 (Figures 8 and 9). Note that one of the candidate loci (with zero polymorphisms) is not found within any of the candidate regions (Figure 8; APPENDIX B, Table B3). The six regions departing from bottlenecks I and II contain only seven of the 22 loci with no polymorphism, and two of the regions (labeled F and G in Figure 9) do not contain any locus with zero SNPs. As expected, they harbor significantly less haplotypes and variation than the rest of the loci (data not shown). With method III, four (method III^s) and five (method III^{all}) more regions are added to the ones above (Figure 9d and e), and they contain six more loci with no polymorphism. Region "J" extends for several hundreds of kb, probably consisting of three distinct sub-regions (Figure 8; APPENDIX B, Table B3). Also in this case variation is significantly lower than in the rest of the chromosome (loci in the other six candidate regions were excluded from the analysis; data not shown). Indeed, visual inspection of the polymorphism pattern revealed that regions identified only by method III contain too few private European SNPs (k^E) compared to the total and/or as expected given the population mutation parameters (MATERIALS AND METHODS, SUBSECTION 1.2.1.4.). We note that region "A" overlaps with the "sweep region 1" described in HARR *et al.* (2002).

We also employed a likelihood ratio test proposed by WRIGHT *et al*. (2005) that compares the likelihood of a single population size bottleneck, with that of two separate bottlenecks affecting a fraction $(1 - f)$ and f of the loci, respectively: the first corresponds to the population

Figure 9. Probability Q^C to obtain at most the observed number of segregating sites at $C = 5$ adjacent loci across the X chromosome in the European sample (based on the *D. melanogaster* genome release 3.2). Each point represents the third (central) of the five consecutive loci. Results for methods I, II^s, II^{all}, III^s, and III^{all} are shown (a, b, c, d, and e, respectively). Regions where we observed consecutive $Q^C < 0.05$ values for $1 \leq C \leq 5$ are highlighted in gray (compare with Figure 8): light gray ones correspond to regions identified only by method III.

size crash accompanying the colonization, and the second mimics the reduction in population size due to the effect of natural selection. The estimation of the bottlenecks' parameters is done simultaneously, considering for each locus *i* the likelihood

$$\mathrm{Lik}_i = (1-f)\times \mathrm{Lik}_i(T_{\mathrm{b}}^1, S_{\mathrm{b}}^1 \mid K_i = k_i) + f \times \mathrm{Lik}_i(T_{\mathrm{b}}^2, S_{\mathrm{b}}^2 \mid K_i = k_i)$$

We simulated two separate bottleneck scenarios using method II$^{\mathrm{all}}$, and found strong statistical support for the presence of two bottlenecks ($\hat{T}_b^1 = 0.0294$ and $\hat{S}_b^1 = 0.346$, and $\hat{T}_b^2 = 0.0063$ and $\hat{S}_b^2 = 2.332$, respectively), with $f \approx 8.7\%$ of the loci experiencing the severe one, mimicking selection ($G = 18.9$, df = 1, $P < 0.0001$). Because in method II$^{\mathrm{all}}$ we analyzed 230 loci, the fraction *f* translates into ~20 loci compatible with the severe bottleneck. For each locus we can calculate the approximate Bayesian posterior probability (Nielsen and Yang 1998) of being in the strong bottleneck using the equation

$$PP_i = \frac{f \times P_i^{II^{all}}(T_{\mathrm{b}}^2, S_{\mathrm{b}}^2 \mid K_i = k_i)}{(1-f)\times P_i^{II^{all}}(T_{\mathrm{b}}^1, S_{\mathrm{b}}^1 \mid K_i = k_i) + f \times P_i^{II^{all}}(T_{\mathrm{b}}^2, S_{\mathrm{b}}^2 \mid K_i = k_i)}$$

While this approach does not exactly specify which locus departs from the "population" bottleneck expectations, it has the advantage of avoiding the problem of multiple testing. Interestingly, we found that the *PP* values highly correlates with the *Q* statistic values, confirming the power of our approach (Spearman's $R = -0.726$, $P < 0.0001$; appendix B, Table B3). For example, the 8 candidate loci identified by method II$^{\mathrm{all}}$ are among the 10 loci with the highest *PP* values. Furthermore, the average *PP* across the loci within the candidate regions is significantly higher than in the rest of the chromosome ($P < 0.0001$).

1.2.3. Discussion

In general, we have found that both demography and natural selection shaped patterns of nucleotide variation on the X chromosome of *D. melanogaster* from Africa and Europe. By developing a method to distinguish between the confounding effects of selection and demography, we were able to localize the gene regions that were the target of selection in the recent past in the derived European population.

1.2.3.1. Demographic history of the African population

An interesting feature of the African sample was the chromosome-wide excess of rare variants. Here we discuss possible explanations of this observation.

First, we can exclude the contribution of chromosomal inversions (Andolfatto *et al*. 2001), as we did not find any on the X chromosome.

Second, mutation bias does not seem to be a valid explanation either. For instance, to investigate the possible contribution of a nucleotide-specific mutation bias (*i.e.,* C/G vs. A/T; *e.g.,* Birdsell 2002), we polarized 3633 alleles, of which 1693 were derived singletons. We observed that derived singletons are more overrepresented among the SNPs with C or G as ancestral state compared to A or T (for singletons $n = 942$ and 751 for C/G and A/T, respectively; for the rest of the SNPs, $n = 999$ and 942, respectively; $\chi^2 = 6.245$, $P = 0.012$). This trend agrees with the study by Kern and Begun (2005), who predicted that G/C to A/T mutations should be at lower frequencies than A/T to G/C mutations due to a recent lineage-specific change in mutation-bias (for singletons $n = 804$ and 455 for C/G and A/T, respectively; for the rest of the SNPs, $n = 840$ and 580, respectively; $\chi^2 = 6.232$, $P = 0.013$). However, both C or G and A or T singletons are in excess when compared to the neutral expectations, calculated as k/a_i where k is the total number of SNPs and $a_i = \sum_{i=1}^{12-1} \frac{1}{i}$ for a sample size of 12 chromosomes (Fu 1995).

Third, since we have previously shown that selection is not a satisfactory explanation for the observed excess of low-frequency variants (for instance, we found no correlation between recombination rate and Tajima's D values, as expected under recurrent selective sweeps; chapter 1.1.), it remains to consider demographic processes. Stajich and Hahn (2005) suggested that admixture can lead to an average negative Tajima's D. This, however, does not seem to apply to rural Zimbabwean populations as used here (Kauer *et al*. 2003). For this reason, the most straightforward explanation for the observed excess of singletons (and the lack of LD) is that the ancestral population has undergone a relatively recent growth in size, expanding either (i) from a population with a long-term constant population size, or (ii) from a severe bottleneck. To explore hypothesis (i), we used an approach proposed by Weiss and von Haeseler (1998) (see materials and methods, subsection 1.2.1.3.). According to this procedure, the maximum-likelihood estimates (95% CI) of the time of expansion, T_e, and the ratio of the present to the past population size, Δ, are 15,000 (0–30,000) years

and 5 (1–1,000), respectively, suggesting a rather recent population-size expansion. While the simple growth model underlying this approach produced estimates that are roughly compatible with the idea of a recent expansion out of Africa (David and Capy 1988; Lachaise *et al*. 1988), some aspects of the data (*e.g.,* the negative average value of the *H* statistic) are not. Hypothesis (ii), proposed by Haddrill *et al*. (2005b), postulates a (slow) population-size growth following a severe old bottleneck (~200,000 years ago). Indeed, the species history may have been strongly influenced by the climatic changes of the past 200,000 years, when three glacial maxima (the last 18,000–21,000 years ago) alternated with warmer and moister periods (Webb and Bartlein 1992; De Vivo and Carmignotto 2004). This scenario of a long-term and slow population growth (interrupted by repeated population size crashes) may also explain why our selective sweep method was unable to find footprints of selection in the African sample. Sweeps that are older than $0.1N_e$ generations cannot reliably be detected by this method (Kim and Stephan 2000). Thus, taken together, our results seem to favor the second scenario.

1.2.3.2. Demographic and selection history of the European population

Our analysis of the European population is based on a thorough approach to distinguish between the confounding effects of selection and demography. This approach consists of two steps: (i) an estimation of the parameters of a simple bottleneck model, and (ii) the identification of loci whose reduction of variation is more extreme than predicted by this bottleneck model. The estimates of the bottleneck parameters obtained in step (i), in particular by method $\mathrm{II}^{\mathrm{all}}$, are consistent with the polymorphism pattern observed in the European sample. Assuming $N_0 = N_e$ for the African population (where $N_0 = 10^6$) and 5–10 generations per year, we find from $T = 3N_0\hat{T}_b$ that the bottleneck occurred between ~3,600 and ~15,800 years ago, depending on the method used to estimate $\hat{T}_b$ (Tables 2 and 3). These values are close to the commonly accepted estimates of 10,000–15,000 years ago as the time of the European colonization (David and Capy 1988; Lachaise *et al*. 1988).

Although a single bottleneck can explain most of the reduction of variation, we identified several candidate loci with less polymorphism than expected. Loci for which the reduction in heterozygosity due to selection is more severe than due to the bottleneck alone are expected to cause an underestimation of the age of the bottleneck. To test this hypothesis,

we removed the 8 loci with $Q < 0.05$ in bottleneck II^{all} (6 of which had no SNPs), and re-estimated the bottleneck parameters. As expected, age and strength point to an older and less severe bottleneck ($\hat{T}_b = 0.0283$ and $\hat{S}_b = 0.335$ *vs.* $\hat{T}_b = 0.0267$ and $\hat{S}_b = 0.400$, respectively). This result does not depend on the low polymorphism of the removed loci. When removing a random set of zero-polymorphism loci (but with $Q > 0.05$), the values of the bottleneck parameter were indistinguishable from those estimated from the complete data set.

In contrast, we estimated a much recent and stronger bottleneck for the 8 outliers alone, consistent with the recent occurrence of positive selection ($\hat{T}_b = 0.0021$ and $\hat{S}_b = 5.0$). This agrees with the signicantly higher fit of two separate bottlenecks than a single one to our data. Namely, we found that the polymorphism pattern across the chromosome is better explained by the occurrence of two bottlenecks, one corresponding to the population crash during the colonization of Europe, and the other, stronger and more recent, corresponding to the effects of natural selection (RESULTS, SUBSECTION 1.2.2.4.).

Selective sweeps are expected to reduce the opportunity of new segregating mutations to accumulate during population expansion (by eliminating those linked to the selected site). Thus, assuming selection acted uniformly over the entire chromosome, we expect more mutations to accumulate in regions of high recombination, where the target of selection and the flanking regions can evolve more independently. Supporting this hypothesis, we found a significant positive correlation between θ^{Es} and recombination rate, which is responsible for the negative correlation observed between Tajima's *D* and recombination rate (RESULTS, SUBSECTION 1.2.2.2.). Furthermore, intronic regions have lower θ^{Eall} ($P = 0.037$) and underwent a more severe drop in genetic diversity than intergenic regions (although not significantly; data not shown), as expected if genes were the main targets of selection, but show less negative *D* values. These findings would contradict those of ORENGO and AGUADÉ (2004), who interpreted a positive correlation between Tajima's *D* values and distance to coding regions (*i.e.,* more negative *D* values for loci close to genes) as signature of recent positive selection in a Spanish population of *D. melanogaster*. The negative correlation between Tajima's *D* and recombination rate can be a consequence of the sampling process associated with the colonization of Europe. This is because the more variation there is in the ancestral African population (which correlates with the recombination rate) the more chances are that a line surviving the bottleneck in the derived population harbors some polymorphisms

that are not shared. However, θ^{Eall} does not correlate with recombination rate, as expected if only the sampling process was responsible for the observed correlation and the correlation between θ^{Es} and recombination rate is significant in intergenic ($R = 0.378$, $P = 0.0001$), but not in intronic regions ($R = 0.007$, $P = 0.941$; this also excludes a possible role of mutation bias across the recombination gradient, see below); moreover, there is no correlation between the African θ and either θ^{Es} nor Tajima's D ($R = –0.063$, $P = 0.353$, and $R = 0.095$, $P = 0.149$, respectively).

When analyzing linked loci, we found several regions where demography alone cannot explain the observed lack of polymorphism, suggesting that the European population experienced multiple episodes of positive selection during its adaptation to the temperate habitat. Remarkably, methods I and II^{all} detected the same candidate regions, despite they estimated bottlenecks of much different age and the latter distinguished between the pre- and post- bottleneck SNPs. Therefore, in our case even when no interspecific comparisons is possible to estimate the frequency of derived alleles, it is possible to detect regions departing from the demographic null model. The six regions with a more robust identification (see RESULTS, SUBSECTION 1.2.2.4.) are not only less polymorphic ($P < 0.0001$), but they contain also less new mutations, as measured by θ^{Es} and θ^{Eall} than the remaining chromosome ($P = 0.001$ and $P < 0.0001$, respectively). This result is in line with the above-mentioned hypothesis that selective sweeps reduced the opportunity for new mutations to accumulate, as observed comparing intronic *vs*. intergenic regions. We also note that many of the loci with no polymorphism were not detected as candidate nor are they located in these candidate regions. This is reasonable, since a bottleneck can easily reduce low ancestral heterozygosity down to zero.

During the bottleneck phase, adjacent loci may have been partially linked and therefore not independent, as we assumed in both the single and multi-locus approaches. Thus, we may have underestimated the age of the bottleneck. The effect of this underestimation is difficult to gauge. On the one hand, it may produce more conservative tests for the individual loci (see CHAPTER 1.1.). On the other hand, ignoring the partial linkage between loci may not be conservative, since low levels of polymorphism at adjacent loci may just reflect their shared ancestry. Non-independence of loci, resulting in pseudo-replicated data, could have also produced smaller confidence interval for the statistics estimated under the bottlenecks

(Table 3). To evaluate the contribution of linkage in our bottleneck estimates, we selected two (partially overlapping) sets of 65 loci with an average distance between adjacent ones of ~250 kb (and average $\theta \approx 0.0046$). Using method II^{all}, we estimated bottleneck parameters comparable with those obtained using the whole data set ($\hat{T}_b$ = 0.0260–0.0275 and $\hat{S}_b$ = 0.398–0.446 *vs.* $\hat{T}_b$ = 0.0267 and $\hat{S}_b$ = 0.400, for 65 and 230 loci, respectively). Furthermore, a population bottleneck explained significantly better the data than a constant population size model in both sets of loci [likelihood ratio test, G = 193.2, df = 1, $P < 10^{-50}$, and G = 105.5, df = 1, $P < 10^{-24}$, for all 230 loci and one set of 65 loci, respectively, using method II^{all}; in the constant population size model, we assumed that the binomial factor of equation (4) equals 1, making it equivalent to method I]. The partial linkage between loci, leading into shared ancestry, might also have influenced our results. However, only among the 24 outliers identified by method III^{all} there are cases of adjacent ones (*i.e.,* three; they are at a minimum distance of 21,362 bp), otherwise they are separated by at least one locus with $Q > 0.05$. The minimum distance between any couple of the 8 candidate loci identified by method II^{all} is 121,177 bp. They are separated by at least 2 loci with $Q > 0.05$ (Figures 7 and 9). Moreover, when we plotted the absolute average difference in Q or θ between a locus and its two neighbors against the average distance or difference in recombination rate to them (as a prediction of association), no correlation is found (data not shown). Therefore, we can be reasonably confident that linkage does not affect our results.

1.2.3.3. Estimating the frequency of adaptive substitutions

A reliable estimate of the number of selective sweeps can be obtained when sampling many loci evenly distributed along the chromosome, so that sweeps have most chances to be detected. Our scan meets these conditions in the centromeric half of the chromosome: between coordinates 10 Mb and 17.6 Mb from the telomere we sequenced 172 loci, spaced on average by ~45 kb. Seven candidate regions are found in this segment, four of which are validated by all methods (Figure 9). If these regions are indeed the result of selective sweeps, we can hypothesize that ~11–18 episodes of positive selection have occurred in the 22 Mb long X chromosome and, neglecting heterochromatic regions that are largely devoid of genes, ~120–200 in the 120 Mb long euchromatic haploid genome. If we take 10,000 years ago as the time point when the colonization started and assume 7 generations per year, this corresponds to one selective sweep per 350–580 generations. For comparison, based on

arguments on the "cost of natural selection", HALDANE (1957) believed that a species cannot tolerate more than one adaptive substitution per 300 generations. An estimate of one every 450–800 generations was reported for Drosophila based on interspecific comparisons of protein-coding sequences (SMITH and EYRE-WALKER 2002; BIERNE and EYRE-WALKER 2004). Although these results are roughly consistent, more work is required to confirm our estimate by examining the individual candidate regions for evidence of selective sweeps.

1.2.3.4. Is recombination mutagenic in *D. melanogaster*?

Another striking observation of our genome scan was the positive correlation between average divergence (of the African sample to *D. simulans*) and recombination rate. The most straightforward explanation of this observation is that recombination is mutagenic in *D. melanogaster*. However, other explanations may also be possible. Although hitchhiking associated with strong positive or negative selection does not increase the rate of neutral molecular evolution, the substitution rate of weakly selected mutations depends on the degree of linkage with strongly selected ones (via the fixation probability; BIRKY and WALSH 1988; MCVEAN and CHARLESWORTH 2000). Thus, depending on the relative magnitude of the rates of slightly advantageous and deleterious mutations, positive or negative correlations between divergence and recombination rate may result (BIRKY and WALSH 1988). Furthermore, the increase of divergence with recombination may simply be due to the relatively recent split of the *D. melanogaster* and *D. simulans* lineages (HELLMANN *et al*. 2003). The latter hypothesis can be tested using *D. yakuba* as outgroup, see CHAPTER 3.2.

2.1. Characterization of a selective sweep in a European population of *Drosophila melanogaster*

Adaptation is one of the major forces driving evolution. When organisms are faced with a novel environment, new mutations conferring an advantage to the carriers will increase in frequency in the population, and eventually go to fixation. This process, coupled with recombination, produces a valley of reduced variation in the DNA flanking the target of selection. The size of the valley depends on the intensity of selection and the local recombination rate (Kaplan *et al*. 1989). This phenomenon, first described by Maynard Smith and Haigh (1974), is known as genetic hitchhiking. An effective way to detect signatures of positive selection is, therefore, to find regions of the genome with a strong reduction in neutral polymorphism (*i.e.*, a "selective sweep"). Then we can test whether neutral processes, *e.g*. genetic drift, are sufficient to explain the observations or if selection must be invoked (Kim and Stephan 2002; Kim and Nielsen 2004; Jensen *et al*. 2005; Nielsen *et al*. 2005).

Other features characterizing a region that experienced the recent action of positive selection are a skew toward rare alleles in its center and haplotype structure at its borders. This is because new, rare mutations will eventually accumulate in regions where the ancestral variation had been depleted, while recombination during the selective phase creates different haplotypes flanking the selected site.

Drosophila melanogaster originated in sub-Saharan Africa and moved to temperate regions only starting 10,000–15,000 years ago (David and Capy 1988; Lachaise *et al*. 1988).

The colonization of the new habitat was likely accompanied by numerous adaptations, but the confounding effect of the associated population bottleneck represents a major obstacle to the detection of the valleys of reduced variation (CHAPTERS 1.1. and 1.2.; KAUER *et al*. 2003; CHAPTER 1.2.). Since selection acts only locally, a way around this problem is to contrast the pattern of single loci with that of the whole genome. In this manner we were able to identify several regions of the X chromosome in a derived European population with lower polymorphism than expected by demographic effects alone (see CHAPTER 1.2.).

In this study, we concentrate in one of these regions (region "I", see Figure 9, CHAPTER 1.2.), centered around fragment *381*. This locus is located in an intron of the gene *Flo-2* and has no polymorphism in the European sample, while normal levels of variation are found in the African sample and the region shows normal levels of divergence to *D. simulans*. Both the region and the single locus showed a significant departure from the expectations based on the assumption of a simple bottleneck (APPENDIX B, Table B3). Therefore this locus represents an ideal candidate to test the power of our methodology in detecting regions consistent with the recent action of positive selection. With this aim, we sequenced and analyzed numerous flanking non-coding regions in both the ancestral and the derived populations and detected a valley of reduced polymorphism that is not compatible with a simple demographic explanation. Intriguingly, this region is close to a gene (*CG9509*) with a significantly different expression profile between African and European flies (MEIKLEJOHN *et al*. 2003), raising the possibility that selection might have occurred to modify (*i.e*., increase, in this case) gene expression in the derived populations. We present a preliminary investigation of this gene's expression in our samples.

2.1.1. Materials and Methods

2.1.1.1. Data collection and analysis

Intraspecific data were collected from highly inbred *D. melanogaster* lines, 12 derived from a European population (Leiden, The Netherlands) and 12 lines from Africa (Lake Kariba, Zimbabwe). For interspecific comparisons, we used a single inbred strain of *D. simulans*.

Primers were designed based on the *D. melanogaster* genome (Flybase, Release 4.2, http://flybase.org). We amplified and sequenced 13 fragments of non-coding DNA, one intergenic region and 12 intronic regions around fragment *381*, as described in SUBSECTION 1.1.1.2. (detailed protocols are presented in APPENDIX C). Sequences were aligned and adjusted by eye when needed using the program Seqman of the DNAstar package (DNAstar, Madison, WI). With the exception of fragment *381*, the homologous *D. simulans* sequences were obtained from its publicly available genome sequence (http://flybase.org/blast).

We calculated nucleotide diversity, measured by π (TAJIMA 1983) and θ (WATTERSON 1975), and Tajima's *D* (TAJIMA 1989) using the program NeutralityTest, kindly provided by H. Li. Divergence and the linkage disequilibrium measure Z_{nS} (KELLY 1997) were estimated by the program DnaSP 4.10 (ROZAS *et al*. 2003).

2.1.1.2. Testing the European population against demography

A population bottleneck, as the one experienced by the European population, can leave a signature in the genome that resembles that of selection. To test the compatibility of our data with this demographic scenario, we performed coalescent simulations using a method that simplifies the bottleneck model to its age, T_b, and its strength, S_b (GALTIER *et al*. 2000; RAMOS-ONSINS *et al*. 2004).

First, we calculated the likelihood that a locus *i* harbors the k_i segregating sites observed in the European sample under a bottleneck using equation 3 of SUBSECTION 1.2.1.4. For each locus, we simulated 10,000 genealogies under a bottleneck defined by the range of T_b and S_b estimated in CHAPTER 1.2. (see Table 3), while the mutation parameter equals the θ value observed in the African sample at locus *i*.

Second, to test if the valley of reduced heterozygosity was caused by the bottleneck, we compared the empirical observations to 1,000,000 genealogies simulated using the same range of T_b and S_b. In this case, the simulations were done either considering complete linkage among loci, or assuming recombination between and within loci. Recombination rate is $R = 2N_e rL$, where $N_e = 10^6$ (LI *et al.* 1999), $r = 4.883\times10^{-8}$ is the average recombination rate per site per generation calculated with the program RecombRate (COMERON *et al.* 1999), and *L* is the total length of the region. The mutation parameter equaled the average θ across loci observed in the African sample. Since we chose to study the present region based on the known absence of polymorphism at locus *381*, we must control for ascertainment bias.

We therefore used only the fraction of genealogies with at most k_{tot} = 87 segregating sites observed across the entire region (making our approach conservative), and for which the fragment *381* was invariant. Then, the probability of our data under the bottleneck scenario can be estimated as the proportion of these simulations generating at most the 3 segregating sites observed across the fragments located within the valley (*i.e*., loci *938*, *940* and *941*).

2.1.1.3. Gene expression analysis

The valley of reduced variation is upstream (~20 kilobases, kb) of the gene *CG9509*, whose relative expression was reported to be significantly higher in males of cosmopolitan origin than in those coming from Zimbabwean strains (Meiklejohn *et al*. 2003). This gene encodes for a protein possibly involved in binding flavin-adenine dinucleotide (FAD), in choline dehydrogenase activity (inferred from sequence or structural similarity), electron transport and mesoderm development (inferred from the expression pattern). It is expressed mainly in larvae and adults (preferentially males; http://genome.med.yale.edu/Lifecycle/gen_query.html).

Here, we analyzed whether the difference in the expression pattern of *CG9509* is also

---

Figure 10. Identification of a valley of reduced variation in the European population of *D. melanogaster*. The position of the 14 analyzed loci is shown in the x axes, relative to that of the first locus. Vertical dotted lines connect the same loci (indicated on top) across the different panels. The nucleotide variation of the European sample, θ_E, is shown in the uppermost panel: note the high correspondent values of polymorphism in the African sample, θ_A and divergence to *D. simulans*, *Div* (mid panel) in correspondence to the valley of reduced variation. Tajima's *D* and haplotype diversity, Hap_{div} across the region are also shown. The region contains 9 genes (one of them only partially) and one transposable element, as shown in the "gene map" panel (pointed tips indicate the direction of transcription). Below, we give the genes' names and their biological and molecular function (when known, or inferred by similarity with characterized genes):
a) *CG9512*: choline dehydrogenase activity; flavin-adenine dinucleotide (FAD) binding; electron transport.
b) *CG9509*: choline dehydrogenase activity; mesoderm development.
c) *CG14406*: unknown.
d) *CG12398*: glucose dehydrogenase (acceptor) activity; FAD binding; electron transport.
e) *CG9504*: oxidoreductase activity; oxidoreductase activity; FAD binding; electron transport.
f) *CG9503*: oxidoreductase activity; oxidoreductase activity.
g) *CG32591*: unknown.
h) Four alternative spliced forms of *Flo-2*: flotillin complex; receptor binding; structural molecule activity; integral to membrane; ectoderm development; cell adhesion; neurogenesis.
i) *CG9009*: long-chain-fatty-acid-CoA ligase activity; fatty acid metabolism.
TE) Transposable element *roo{}132*.

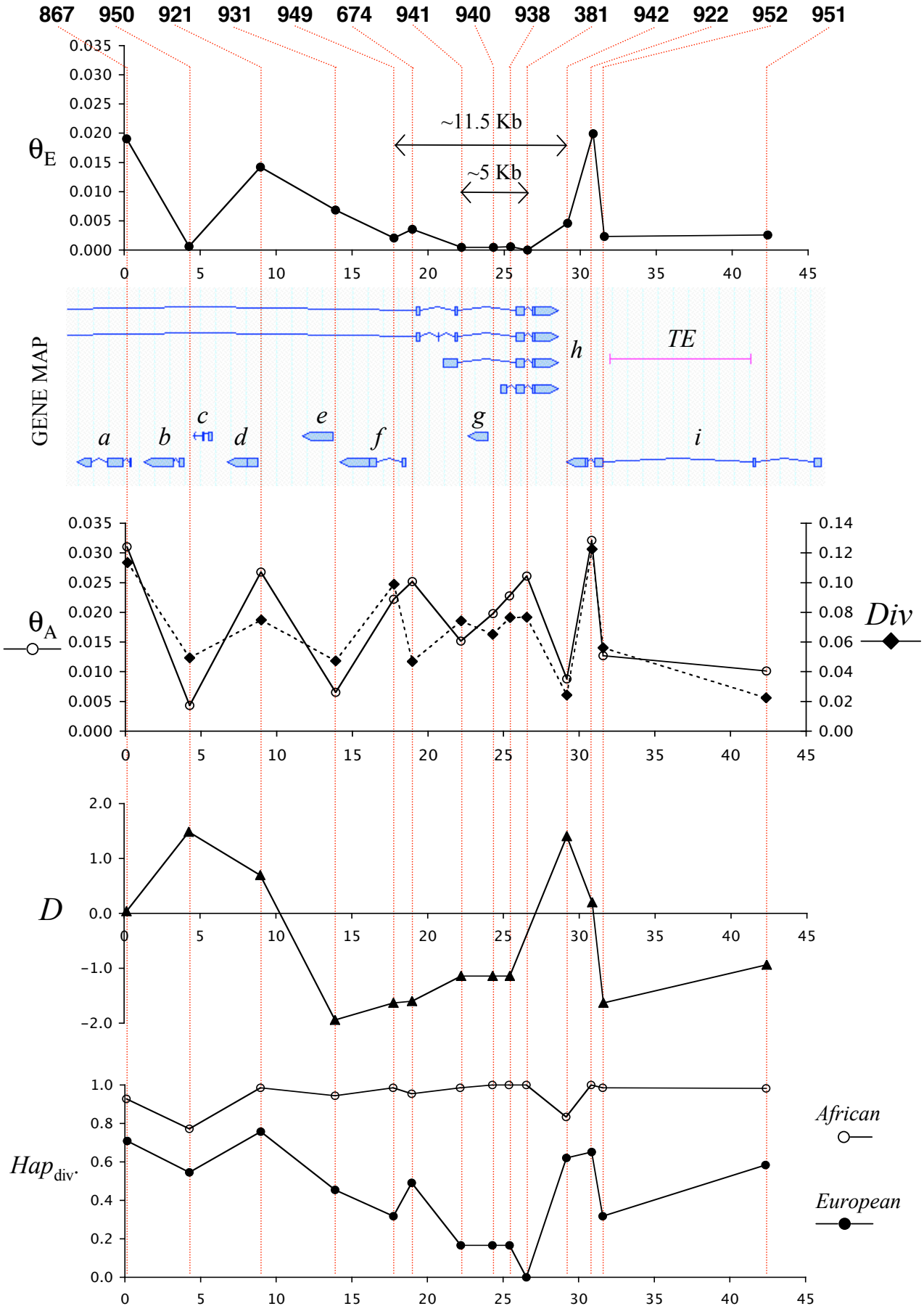

867 950 921 931 949 674 941 940 938 381 942 922 952 951
θE
~11.5 Kb
~5 Kb
GENE MAP
a b c d e f g h i
TE
θA
Div
D
Hapdiv.
African
European

Table 4. Diversity estimates and test statistics for the loci in the candidate region.

			European sample									African sample			
Locus	Abs. Pos. [a]	Relat. Pos. [b]	L [c]	n [d]	k [e]	θ [f]	π [g]	Hap_{div} [h]	D [i]	Div [j]	Z_{nS} [k]	n [d]	k [e]	θ [f]	π [g]
867	14738968	0	305	11	17	0.0190	0.0192	0.709	0.039	0.1135	0.568	11	18	0.0310	0.0357
950	14742954	3986	546	12	1	0.0006	0.0010	0.545	1.486	0.0493	n.a.	12	7	0.0043	0.0033
921	14747744	8776	396	12	17	0.0142	0.0165	0.758	0.694	0.0748	0.798	12	31	0.0268	0.0286
931	14752692	13724	338	12	7	0.0069	0.0035	0.455	–1.944	0.0472	n.a.	9	6	0.0066	0.0058
949	14756474	17506	485	12	3	0.0020	0.0010	0.318	–1.629	0.0989	n.a.	12	31	0.0222	0.0239
674	14757791	18823	290	11	3	0.0035	0.0019	0.491	–1.600	0.0469	n.a.	12	21	0.0252	0.0200
941	14760803	21835	728	12	1	0.0005	0.0002	0.167	–1.141	0.0742	n.a.	12	31	0.0152	0.0146
940	14762907	23939	703	12	1	0.0005	0.0002	0.167	–1.141	0.0651	n.a.	12	43	0.0198	0.0165
938	14764106	25138	564	12	1	0.0006	0.0003	0.167	–1.141	0.0764	n.a.	12	35	0.0228	0.0218
381	14765287	26319	429	12	0	0.0000	0.0000	0.000	n.a.	0.0767	n.a.	11	34	0.0261	0.0268
942	14767970	29002	361	12	5	0.0046	0.0063	0.621	1.408	0.0243	1.000	12	9	0.0088	0.0076
922	14769812	30644	416	12	25	0.0199	0.0208	0.652	0.201	0.1225	0.503	12	35	0.0321	0.0312
952	14770338	31370	431	12	3	0.0023	0.0012	0.318	–1.629	0.0562	n.a.	12	16	0.0127	0.0084
951	14781099	42131	431	9	3	0.0026	0.0019	0.583	–0.936	0.0223	n.a.	11	15	0.0101	0.0078

present in the 24 lines used in the polymorphism survey. We studied the expression in a mixed sample (15 males and 15 females) of 4–5 days old adult flies raised at a constant temperature of 25 °C. We obtained cDNA from the total RNA, extracted with Trizol and chloroform, and followed by isopropanol precipitation (for detailed protocols of this and the following methods, see APPENDIX C). To control for non-homogeneous gene expression across the lines, the quantification of *CG9509* expression was calculated relative to that of an endogenous control gene, the ribosomal protein 49 (*Rp49*; probes by Applied Biosystems). Each line was tested in three replicates (in a single experiment) using RT-RealTime PCR, and the log-average expression of *CG9509* (across replicates) was normalized relative to the corresponding value of *Rp49*.

2.1.2. Results

2.1.2.1. Nucleotide diversity pattern across the candidate region

In our genomic scan of variation across the X chromosome, we identified several loci showing lower than expected polymorphism in the European sample even when bottlenecks were assumed (see CHAPTER 1.2.). To investigate the potential role of selection, we sequenced 13 additional loci around one of them, namely locus *381*. The location of these 14 loci, as

Table 4. (Continues from previous page.)

[a] Absolute position in the chromosome (in base pairs, from the telomere) is according to the *D. melanogaster* genome release 4.2 (http://flybase.org).
[b] Relative position within the studied candidate region.
[c] Length (excluding gaps).
[d] Sample size.
[e] Number of segregating sites.
[f] Nucleotide polymorphism based on the number of segregating sites.
[g] Nucleotide polymorphism, estimated by the average number of pairwise sequence differences in the sample.
[h] Haplotype diversity.
[i] Tajima's *D* (the statistic was not quantifiable at locus *381* due to the lack of polymorphism).
[j] Divergence to *D. simulans*.
[k] Linkage disequilibrium (only informative – non-singletons – sites were used; n.a. = not available).

well as the corresponding nucleotide diversity and test statistics, are shown in Figure 10 and Table 4. In the European sample, the average (standard error, SE) θ across the region is 0.0055 (0.0019), higher than the value observed across the entire X chromosome (θ = 0.0046). The same trend holds in the African sample, with an average (SE) θ of 0.0188 (0.0025) in the region versus an average of 0.0131 on the whole chromosome. These differences are most likely related to the high level of recombination present in the region studied (*i.e.*, 4.883 recombination events per site per generation), since we found a positive correlation between the levels of polymorphism and the recombination rate (see CHAPTER 1.2.).

Despite the high average polymorphism across the region, we observed a total of only 3 segregating sites in four adjacent loci (including locus *381*; Table 4) in the European sample. This is not the result of local constraints, since the corresponding polymorphism in the African sample and that of divergence to *D. simulans* are both within the range of normal values (Figure 10). We also calculated Tajima's *D* (TAJIMA 1989) in the European sample: loci within the valley of reduced polymorphism have a strong skew towards rare mutations (*i.e.*, negative *D* values), while the flanking loci show haplotype structure (*i.e.*, positive *D* values). Interestingly, the low-polymorphism loci have only derived alleles segregating as singletons and private to the European sample, suggesting that these mutations originated recently and after the bottleneck (Table 5).

During a selective sweep, rare alleles go rapidly to fixation by hitchhiking with the favorable allele: visual inspection of the alignments revealed 14 sites where the derived allele was fixed in the European sample while absent (6 cases) or segregating as singleton (8 cases) in the African sample. Interestingly, all these sites (with the exception of one fixed difference at locus *950*) are found in the four loci within the valley of reduced variation.

2.1.2.2. Testing the polymorphism pattern against demography

The bottleneck experienced by the European population considerably reduced the ancestral polymorphism across the whole genome. Therefore, the reduction in heterozygosity observed in our region might simply be the result of the associated stochastic processes, including the effects of genetic drift. To test if the valley of reduced nucleotide diversity can indeed be explained by demography alone, we simulated data sets under various bottleneck models. In particular, we wanted to know whether the number of segregating sites observed across the region is compatible with a bottleneck scenario that can explain the polymorphism

Table 5. Nucleotide mutations in the low-variation fraction of the studied region.

Locus →		931						949			674			941	940	938	942				
Position		13731	13748	13821	13876	13897	13973	14054	17602	18016	18023	18826	18922	19014	22394	24313	25721	29019	29066	29199	29211	29213
	Euro. Cons.	A	A	G	A	T	T	C	T	T	A	C	A	A	G	A	C	T	T	T	C	T
D. melanogaster European sample	Line #1	•	G	•	G	•	G	•	•	•	•	T	•	•	•	•	A	•	G	G	T	A
	Line #2	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
	Line #11	•	•	A	•	A	•	T	•	•	•	•	•	•	A	•	•	•	•	•	•	•
	Line #12	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	G	•	•	•	•
	Line #13	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
	Line #14	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	G	G	T	A
	Line #15	•	•	•	•	•	•	•	•	•	•	•	G	•	•	•	•	•	•	•	•	•
	Line #16	•	•	•	•	•	•	•	•	•	•	–	–	–	•	•	•	•	G	G	T	A
	Line #17	•	•	•	•	•	•	•	•	•	C	•	•	•	•	C	•	•	•	•	•	•
	Line #18	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
	Line #19	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	G	G	T	A
	Line #20	•	•	•	•	•	•	•	C	A	•	•	•	•	•	•	•	•	G	G	T	A
D. melanogaster African sample	Line #82	–	•	•	•	•	•	T	•	•	•	•	T	•	•	•	•	•	G	G	T	A
	Line #84	–	•	•	•	A	•	T	•	•	•	•	T	•	•	•	•	•	•	•	T	A
	Line #131	–	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	G	•	T	A
	Line #95	–	–	–	–	–	–	–	•	•	•	•	T	•	•	•	•	•	•	•	T	A
	Line #131	–	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	G	•	T	A
	Line #145	–	–	–	–	–	–	–	•	•	•	•	T	•	•	•	•	•	G	G	T	A
	Line #157	–	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	A
	Line #186	–	•	•	•	•	•	T	•	•	•	•	T	•	•	•	•	•	•	G	T	A
	Line #191	–	•	•	•	•	•	T	•	•	•	•	T	•	•	•	•	•	G	•	T	A
	Line #229	–	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	G	G	T	A
	Line #377	–	–	–	–	–	–	–	•	•	•	•	T	•	•	•	•	•	•	•	T	A
	Line #384	–	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	G	G	T	A
	Line #398	–	•	•	•	•	•	T	•	•	•	•	T	•	•	•	•	•	•	•	T	A
	D. simulans	A	A	G	A	T	T	T	T	T	A	C	T	A	G	A	C	T	G	G	T	A

Only the sites polymorphic in the European sample are shown: the position is relative to the starting point of the studied region (see also Figure 11). Dots (•) correspond to alleles identical with the consensus sequence for the European sample (Euro. Cons.); in some cases, a line was not sequenced, or a deletion was present (–).

pattern of the whole chromosome (*i.e.*, of the population).

First, we calculated the probability for each locus *i* to harbor at most k_i segregating sites observed in the European sample under two bottlenecks estimated by maximum-likelihood in CHAPTER 1.2. Additional parameter combinations, to explore slightly different bottleneck scenarios, were also employed (Table 6; MATERIALS AND METHODS, SUBSECTION 2.1.1.2.). As

Table 6. Results of the bottleneck simulations – single loci.

Bottleneck model	T_b	S_b	Locus													
			867	*950*	*921*	*931*	*949*	*674*	*941*	*940*	*938*	*381*	*942*	*922*	*952*	*951*
b1	0.0125	0.350	0.738	0.363	0.683	0.911	0.178	0.267	0.070	0.050	0.054	0.023 *	0.696	0.738	0.333	0.447
b2	0.0267	0.400	0.749	0.365	0.694	0.918	0.145	0.26	0.036 *	0.019 *	0.023 *	0.008 *	0.707	0.758	0.346	0.471
b3	0.0100	0.300	0.705	0.327	0.645	0.899	0.143	0.223	0.056	0.042 *	0.046 *	0.021 *	0.659	0.724	0.293	0.405
b4	0.0100	0.400	0.771	0.426	0.727	0.924	0.241	0.336	0.120	0.088	0.100	0.046 *	0.741	0.786	0.398	0.509
b5	0.0200	0.300	0.685	0.295	0.626	0.892	0.105	0.197	0.030 *	0.017 *	0.022 *	0.008 *	0.642	0.694	0.257	0.385
b6	0.0200	0.400	0.759	0.388	0.710	0.922	0.179	0.288	0.059	0.035 *	0.043 *	0.016 *	0.716	0.764	0.360	0.483
b7	0.0300	0.300	0.670	0.266	0.603	0.888	0.074	0.169	0.015 *	0.007 *	0.009 *	0.003 *	0.623	0.676	0.228	0.367
b8	0.0300	0.400	0.743	0.354	0.693	0.917	0.130	0.252	0.029 *	0.014 *	0.018 *	0.006 *	0.696	0.751	0.325	0.468

For each locus i we estimated the probability, Q, to observe at most k_i segregating sites in the European sample (see Table 4) under a bottleneck of age T_b (expressed in units of $3N_e$ generations) and strength S_b. We considered 8 different bottleneck models: b1 and b2 correspond to the bottlenecks I and II^{all} estimated in SUBSECTION 1.2.2.3., while the remaining models describe combinations of close values of age and strength. As mutational parameter, we used for each locus the corresponding θ value observed in the African sample. Asterisks indicate $Q < 0.05$, *i.e.*, less polymorphism than expected under the bottleneck model.

Table 7. Results of the bottleneck simulations – whole region.

Bottleneck model	T_b	S_b	No rec. [a]	Rec. [b]
b1	0.0125	0.350	0.156	0.084
b2	0.0267	0.400	0.008 *	0.007 *
b3	0.0100	0.300	0.245	0.144
b4	0.0100	0.400	0.300	0.209
b5	0.0200	0.300	0.030 *	0.009 *
b6	0.0200	0.400	0.033 *	0.019 *
b7	0.0300	0.300	0.004 *	< 0.003 *
b8	0.0300	0.400	0.003 *	< 0.001 *

For each bottleneck model (see Table 6 for details), we calculated the probability to observe at most $k = 3$ segregating sites in the valley of reduced variation, *i.e.*, in loci *941*, *940* and *938* (see text; Figure 11 and Table 4). The probability was conditioned on the observation of at most $k_{tot} = 87$ segregating sites across the region and none at locus *381*. As mutational parameter, we used the average θ value observed across the region in the African sample. Asterisks indicate probability $P < 0.05$, *i.e.*, less polymorphism than expected under the bottleneck model.

[a] Simulations were conducted assuming complete linkage within and between loci.

[b] Simulations assumed recombination across the whole region, with $r = 4.883 \times 10^{-8}$ (recombination events per site per generation).

expected, the loss of heterozygosity is more probable for recent and strong bottlenecks (*e.g.*, b1 *vs*. b2). Locus *381* and up to three flanking loci depart from the expectations of the simple bottleneck models. These 4 loci, which span over ~5 kb, contain a total of 3 segregating sites in the European sample, while the African sample shows appreciable levels of polymorphism.

In a second approach, we analyzed the pooled data, partitioning the studied region in a valley of marked reduced variation (*i.e.*, the four loci identified in the above analysis) and the remaining 10 loci. We focused on this region because we had preliminary evidence for no variation at locus *381*. Therefore, we must control for ascertainment bias. To do this, we simulated the entire region under the same 8 bottleneck models used above, but

retained only simulations producing zero segregating sites at this locus. Furthermore, only those simulations with at most the total number of observed segregating sites across the entire region (*i.e.*, $k_{tot} = 87$) were further analyzed. Very recent bottlenecks ($T_b < 4{,}000$ years ago, assuming 10 generations per year) can partly explain our data (Table 7). However, the chromosome-wide polymorphism pattern of the European sample is better explained by a much older bottleneck (*i.e.*, model b2, where $T_b \approx 8{,}000$ years ago; CHAPTER 1.2.), suggesting that other forces, other than demography, reduced polymorphism.

2.1.2.3. Expression of the gene *CG9509*

The region analyzed in this study is upstream of the 5' end of the gene *CG9509*, which has been found to differ in expression between cosmopolitan and African flies (MEIKLEJOHN *et al*. 2003). The gene encodes for a protein involved in mesoderm development, and its expression is under the control of the transcription factor Twist (FURLONG *et al*. 2001). The binding sites for Twist, together with those for Dorsal, Suppressor of Hairless and another poorly known transcription factor, are usually organized in enhancers and are under strict organizational constraints (ERIVES and LEVINE 2004). We verified the presence of the binding motifs given in ERIVES and LEVINE (2004) within the valley of reduced variation (plus 3 kb up- and downstream), and found a total of 10 putative binding sites (three for Twist, four for Suppressor of Hairless, two for Dorsal and one for the unknown transcription factor). To be functional, an enhancer needs to have several binding sites close to one another. Interestingly, in the 3' region of locus *381* three motifs are found within less of 600 bp (two for Twist and one for Dorsal). Both a change in the binding motif (which have some compositional flexibility) or in the spacing can affect the functionality of an enhancer.

The expression pattern of the gene *CG9509* is shown in Figure 11. In the African sample, the gene is expressed at higher levels than in the European one, opposite to what was found by MEIKLEJOHN *et al*. (2003), although the difference is not significant ($P = 0.273$). The contradictory results might depend on the use of a mixed female/male sample tested in this study, which could have masked a male specific expression pattern (Meiklejohn and colleagues analyzed only males). If so, the difference in expression had to be subtle to be overcome by that of the females. Interestingly, we observed a higher variance in expression in the European sample (Levene test for homogeneity of variances, F ratio = 3.862, $P = 0.062$): however, re-testing the difference in expression allowing for unequal variances does

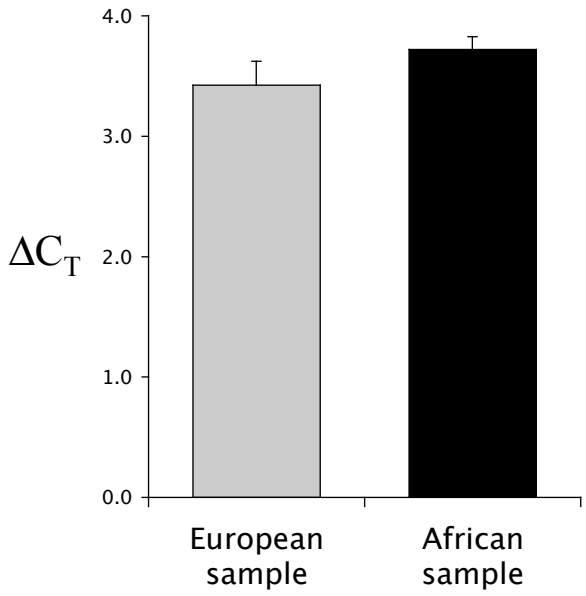


Figure 11. Expression of *CG9509* in the European and African samples. The expression was quantified using RT-RealTime PCR (see text): in brief, the amount of mRNA molecules present in the sample are proportional to the time (*i.e.*, fractional cycles) at which the PCR amplification reaches its maximal speed. The quantification is normalized by considering the difference in time, ΔC_T (expressed in a logarithmic scale), to that of the endogenous control, *Rp49*. Average (SE) ΔC_T values for the African and the European samples are 3.720 (0.105) and 3.425 (0.197), respectively, not significantly different ($P = 0.273$).

not change the results (Welch ANOVA, $P = 0.204$).

2.1.3. Discussion

Despite intensive efforts, identification of a selective sweeps remains scarce. Numerous studies have focused on *D. melanogaster*, facilitated by its fully sequenced genome and extensive characterization of its genes and molecular processes. In particular, the cosmopolitan populations are excellent candidates to look for positive selection, since their recent colonization of novel habitats was likely accompanied by adaptation. The bottleneck associated with such colonization considerably reduced the ancestral polymorphism, as evidenced by numerous studies (chapters 1.2. and 1.1.; Begun and Aquadro 1993; Kauer *et al*. 2002; Baudry *et al*. 2004; Haddrill *et al*. 2005b). This complicates the detection of selective sweeps and the evaluation of the effects of natural selection in genome evolution. On the other hand, advances in statistical and computational methodologies have made genome-wide approaches to disentangle demographic and selective forces possible and revealed the impact of recent adaptive mutations (*e.g.*, chapter 1.2.).

2.1.3.1. Indications that the valley of reduced variation corresponds to a selective sweep

In this study, we analyzed a region of the X chromosome where we found preliminary evidence for incompatibility of the European polymorphism pattern with the sole action of demography (see CHAPTER 1.2.). We detected a valley of reduced variation (4 loci spanning ~5 kb) that cannot be explained by local constraints or by low polymorphism in the ancestor. In fact, this region harbors only 3 segregating sites in the European sample, while it shows high levels of polymorphism in African and normal divergence to *D. simulans*. We used coalescent simulations to test whether the extension of the valley might be explained by a bottleneck. While a recent bottleneck of ~4,000 years of age can account for the little polymorphism across the 4 loci, more realistic scenarios (>8,000 years of age; CHAPTER 1.2.; HADDRILL *et al*. 2005b) reject neutrality. Additional evidence for the action of positive selection comes from the frequency spectrum pattern across the region. Negative Tajima's *D* values are associated with reduced variation at the center of the valley, whereas positive values are found in flanking regions indicating haplotype structure. In the valley, the only segregating mutations are singletons; the derived alleles are private to the European sample, suggesting that they originated after the sweep occurred (but see SCHÖFL *et al*. 2005). Therefore, we can estimate the timing of the loss of polymorphism by finding a bottleneck (mimicking natural selection) that maximize the likelihood that three of the four loci in the valley harbor at most one singleton (*i.e*., loci *938*, *940* and *941*) and one has no variation (*i.e*., *381*). This approach is equivalent to that described in CHAPTER 1.2. to estimate the population size bottleneck. The age of the sweep was estimated to be ~1,500 years ago (assuming 10 generations per year).

The extension of the valley of low polymorphism is much less than those reported in recent studies, where between 50 and 100 kb were found to be (nearly) invariant (SCHLENKE and BEGUN 2004; BEISSWANGER *et al.* 2005; S. Glinka, personal communication). A straightforward explanation for this difference is the 2.5–10 fold higher recombination present in the region compared to that in the sweeps reported in the above studies. The effect of recombination is evident in the gradual increase of the haplotype diversity and of Tajima's *D* with the distance to the valley of reduced variation, and by the opposite behavior of linkage disequilibrium, which decreases accordingly (Table 4).

2.1.3.2. Candidate genes associated to the selective sweep

What was the target of selection? An increase in fitness can be achieved by a non-synonymous change in the coding regions of a gene, altering the functional activity of the coded protein or by a change in the expression profile of a protein. The region of low variation overlaps with only two genes, namely *CG32591* and two exons of *Flo*-2 (this gene spans over ~95 kb; Figure 10). The former has not functionally been characterized, while the latter codes for a structural membrane protein (Flotillin-2) involved in receptor binding and embryogenesis. More detailed sequence analysis of the coding DNA within the region will be required to find a candidate for a substitution under selection. A previous study reported a significantly enhanced expression of the gene *CG9509* in male flies of cosmopolitan origin compared to African ones (Meiklejohn *et al*. 2003). Since the 5' end of studied is located upstream of this gene (~18 kb), there was the possibility that it contained regulatory elements that underwent selection. Preliminary investigations, however, failed to detect any significant difference in expression between our African and European samples. This might be due to the different samples used in our experiment, or due to the fact that the gene is differently expressed only in males or larvae.

Although more analysis is needed to confirm the presence of the selective sweep and to find the target of selection, this study represents an example of the effects of natural selection in *D. melanogaster* (for other examples, see Catania and Schlötterer 2005; Beisswanger *et al*. 2005). Notably, we showed that it is possible to distinguish between the effects of demography and selection, and we confirmed that natural selection was an important force driving the evolution of the European population.

3. The effects of neutral and selective forces on the genome evolution of *Drosophila melanogaster*

Non-coding DNA constitutes a considerable fraction of the genome of eukaryotes. Despite being often referred to as "junk-DNA", there is mounting evidence for its potential functions. Introns can play a role in alternative splicing and exon shuffling (Sharp 1994; Hanke *et al*. 1999) and – in some cases – their pre-mRNA secondary structure can affect gene expression (Chen and Stephan 2003; Hefferon *et al*. 2004). Regulatory elements are present in the immediate 5' neighborhood of genes (*i.e.,* TATA and CG boxes), but they can also modulate gene expression from a greater distance to the target gene (*i.e.,* enhancers and transcription-factor binding sites). Regulatory elements can also reside in introns (*e.g.,* Bergman and Kreitman 2001). Indeed, evidence for selective constraints in non-coding DNA has been found in whole-genome comparisons in Caenorhabditis (*e.g.,* Shabalina and Kondrashov 1999), mammals (*e.g.,* Dermitzakis *et al*. 2002) and Drosophila (Bergman and Kreitman 2001; Halligan *et al*. 2004; Andolfatto 2005). Matrix attachment regions and *cis*-regulatory elements have also been recognized as targets of purifying selection (Ludwig and Kreitman 1995; Glazko *et al.* 2003).

In the following chapters, I thoroughly analyze the insertion/deletion and nucleotide substitution patterns across intronic and intergenic regions, in order to evaluate these selective constraints. In a second step, I study in detail the mutation pattern to infer biases in the substitution process and understand the role of neutral and selective forces in base composition across the genome.

3.1. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions

A recent analysis of polymorphic insertions and deletions in *D. melanogaster* non-coding DNA revealed an overall ratio of deletion-to-insertion events of 1.35 (referred to as polymorphic deletion bias or PDB) (Comeron and Kreitman 2000). The authors hypothesized that this deletion bias must be compensated by selection to maintain minimum intron length and generally favoring longer introns to enhance recombination. The polymorphism data they used to substantiate their claim were from 31 genomic regions (with very different recombination rates), from multiple sources (generated in various labs by restriction mapping, SSCP and DNA sequencing) and multiple sampling locations (with very different sample sizes).

A broad range of PDB estimates are found in the literature. In a survey of sequence length diversity in the *Adh* region of *D. pseudoobscura*, Schaeffer (2002) observed a PDB of 0.83 for all indel types (including repetitive ones such as microsatellites), and of 1.89 for non-repetitive indels (calculated from his Table 1). Similarly, Parsch (2003) reported a ratio of fixed deletions to insertions of 1.66 in a comparison of orthologous introns among species of the *D. melanogaster* subgroup. On the other hand, studies of "dead-on-arrival" non-LTR retrotransposons in Drosophila (Petrov and Hartl 1998; Blumenstiel *et al*. 2002) found deletion-to-insertion ratios ranging from about 4 to 8. The differences among the polymorphic deletion bias estimates are most likely due to different samples, sequences and methods used in these studies. However, disagreements may also derive from the way repetitive indels are treated. Only Schaeffer (2002) distinguished between repetitive and

non-repetitive indels.

In this study, we used nucleotide sequence data from a single population of *D. melanogaster* from Africa to revisit the various hypotheses concerning deletion bias and its consequences. Our data consist of short fragments (introns and intergenic sequences) from regions of normal recombination on the X chromosome. These fragments are of similar length (about 500 bp); *i.e.,* the introns belong to the large size class (> 90 bp; see MOUNT *et al*. 1992, STEPHAN *et al*. 1994). They were previously analyzed for patterns of nucleotide diversity (generally using a sample of 12 chromosomes) and divergence (to a single *D. simulans* line) (CHAPTER 1.1.). This analysis suggested that the African population is close to equilibrium between mutational forces and genetic drift. For these reasons, this sample is particularly suitable for analyzing the selective constraints in introns and intergenic regions (which are expected to fall into the realm of weak selection).

3.1.1. MATERIALS AND METHODS

3.1.1.1. Drosophila data set

To reduce the possible constraints due to the presence of complex transcription-factor binding sites, we use here only the intergenic regions from the original data set that are at least 1 kb away from the 5' UTR of an annotated gene (based on Flybase 3.0 release, retrieved by the Apollo tool; http://flybase.org). Similarly, to avoid potential problems due to the specific location of the fragments within introns (*e.g.,* presence *vs*. absence of splicing elements), we excluded partial introns. The data set meeting the above criteria consists of 22 intergenic regions and 54 introns with an average length (standard error, SE) of 561.1 bp (61.0) and 492.1 bp (128.4), respectively (excluding deletions and insertions).

3.1.1.2. Analysis of insertion and deletion variation

Insertions and deletions segregating in *D. melanogaster* were polarized according to the state observed in *D. simulans*. Only indels for whom the reconstruction of the ancestral state was unambiguous (*i.e.,* those in which one of the two *D. melanogaster* variants was

also present in *D. simulans*) were used in the present study I removed this sentence because these indels are also classifiable as ambiguous, since anyway no variant is found in simulans. Insertions and deletions were classified into two categories (modified from SCHAEFFER 2002): i) non-repetitive and ii) repetitive (duplications, and mononucleotide and microsatellite repeats). Indels containing repeated DNA sequences have been treated separately, as their expansion/contraction dynamics may produce homoplasy and different numbers of repeats may be added (deleted) at the same location in separate events. We follow here SCHAEFFER'S (2002) suggestion, since the discrepancies among the PDB estimates may derive from the definition of indels. Only SCHAEFFER (2002) classified indels in different categories (repetitive and non-repetitive), while COMERON and KREITMAN (2000) grouped complex indels (*i.e.,* repetitive ones) and counted them as one event. Nucleotide and insertion/deletion (indel) diversity π (TAJIMA 1983) and Tajima's *D* (TAJIMA 1989) statistic were estimated using the program NeutralityTest, kindly provided by H. Li (available at http://hgc.sph.uth.tmc.edu/neutrality_test). Divergence was analyzed using DnaSP 4.0 (ROZAS *et al.* 2003).

3.1.1.3. Modeling of selective constraints

To understand how the distribution of selectively constrained regions in intergenic and intronic sequences can relate to the observed pattern of insertions and deletions, we analyzed simple models of sequence constraints. We assume that a sequence consists of subsequences delimited by functionally constrained blocks (*i.e.,* exons, transcription-factor binding sites or regulatory regions). In this way, the model can apply to both introns and intergenic regions. Deletions and insertions are considered neutral if they do not alter the block structure (*i.e.,* if they do not fall into a functionally important region) and, because of their size, if they are meeting the spacing constraints between consecutive blocks (Figure 12). Otherwise, deletions and insertions are subjected to strong purifying selection and thus eliminated from the population very shortly after they appear.

We used an approach similar to that described in PTAK and PETROV (2002) to calculate the following statistics: i) the fraction of deletions and insertions that do not interfere with the functional constraints, ii) the fraction of these deletions and insertions $\leq$ 10 bp, and iii) the resulting deletion-to-insertion ratio. These values were calculated as a function of the length L of a given subsequence and of its maximum (L_{max}) and minimum (L_{min}) lengths

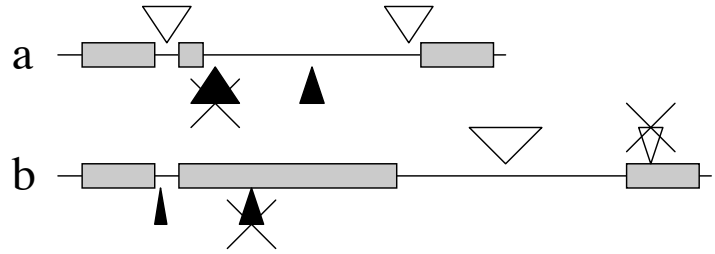


Figure 12. Schematic representation of the model of selective constraints considered in the analysis. Subsequences are delimited by blocks (grey boxes) of coding (exons) or noncoding functional DNA (*e.g.*, regulatory regions or splicing elements). Deletions (filled triangles) are deleterious when they overlap with constrained blocks (crossed-out triangles), while both insertions (open triangles) and deletions may be subjected to purifying selection if they alter spacing constraints (*i.e.*, length of subsequence).

tolerated (reflecting spacing constraints). Then, the fraction of insertions of length S, $f_{\text{ins}}(S)$, that do not interfere with the constraints is

$$f_{ins}(S) = \begin{cases} 1, & \text{if } L + S \le L_{\max} \\ 0, & \text{otherwise.} \end{cases} \qquad (1)$$

Similarly, for deletions we have

$$f_{del}(S) = \begin{cases} \dfrac{L - S + 1}{L}, & \text{if } L - S \ge L_{\min} \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

In order to vary length (spacing) constraints, we define

$L_{\min} = L(1 - \gamma)$ and $L_{\max} = L(1 + \delta)$,

where $0 \le \gamma, \delta < 1$.

It is evident that the smaller L, the fewer indels will be neutral; moreover, the closer $L_{\max}$ and $L_{\min}$ are to L (*i.e.,* the more spacing constraints are present), the higher will be the fraction of small indels.

In applying this model to our data we have to take into account that our fragments may contain subsequences of different lengths, each with possibly specific spacing constraints.

For simplicity, we consider only two length classes of subsequences, "short" and "long" ones, and we compute the indel statistics based on the fraction of short *vs*. long subsequences (thus varying sequence composition). Let F_{short} be the proportion of short sequences in the total sequence ($0 < F_{short} < 1$), and $f_{indel,s}(S)$ and $f_{indel,l}(S)$ the fractions of indels of size S that do not interfere with the constraints of short and long sequences, respectively. The fraction of indels of size S that does not interfere with any sequence constraint is then given as

$$f_{indel}(S) = F_{short} f_{indel,s}(S) + (1 - F_{short}) f_{indel,l}(S),$$

where we substitute for $f_{indel,s}(S)$ and $f_{indel,l}(S)$ the right-hand side of equation (1) and (2) for insertions and deletions, respectively.

The statistics are then computed using equations (1) to (5) of Ptak and Petrov (2002), based on the indel size distributions of Petrov and Hartl (1998). Here we rely on the assumption that the size distributions of deletions and insertions of Petrov and Hartl (1998) are the result of neutral processes. Finally, it should be noted that this analysis refers to the data set as a whole rather than to a single fragment. As Table 8 indicates, the values of PDB across fragments may be rather different.

3.1.2. Results and Discussion

3.1.2.1. Introns and intergenic regions show a similar polymorphic deletion bias

When all indels are considered, the values of PDB are lower than one for both introns and intergenic regions, in agreement with Schaeffer (2002) (Table 8). For the non-repetitive indels we find PDB values of 2.00 and 2.17 for introns and intergenic regions, respectively, in line with Schaeffer (2002). The lower value (1.35) obtained by Comeron and Kreitman (2000) is most likely the result of the way repetitive indels were counted in their study.

3.1.2.2. Insertions have smaller sizes and higher frequencies than deletions

Deletions are significantly larger than insertions (Figure 13 and Table 8). If we exclude

Table 8. Analysis of polymorphic insertions (*ins*) and deletions (*del*) in non-coding DNA of *D. melanogaster*.

		Introns					Intergenic regions				
		n [a]	PDB [b]	Av. size (SE) [c]	Av. freq. (SE) [d]	% ≤ 10 bp [e]	n [a]	PDB [b]	Av. size (SE) [c]	Av. freq. (SE) [d]	% ≤ 10 bp [e]
Non-repetitive DNA indels	*del*	62	2.00 (1.06–2.05)	8.94 (1.13)	0.244 (0.033)	0.73	26 [g]	2.17 (0.62–2.38)	10.00 (1.19)	0.219 (0.044)	0.56
	ins	31 [f]		6.32 (1.54)	0.354 (0.047)	0.81	12		5.33 (2.09)	0.421 (0.103)	0.83
Wilcoxon test	*Z*			–2.122	0.304				–2.823	0.274	
	P			0.034	0.761				0.005	0.784	
All indels	*del*	108	0.92 (0.62–1.91)	6.06 (0.60)	0.268 (0.024)	0.83	41 [g]	0.69 (0.52–1.72)	6.83 (1.00)	0.248 (0.038)	0.71
	ins	118 [f]		3.33 (0.58)	0.382 (0.027)	0.94	59		3.10 (0.52)	0.483 (0.040)	0.95
Wilcoxon test	*Z*			2.988	–1.515				2.975	–2.779	
	P			0.003	0.130				0.003	0.005	

[a] Number of polymorphic events.

[b] Polymorphic Deletion Bias: ratio between the number of observed deletions and insertions. The minimum and maximum values observed per fragment are given in parenthesis. Note that these have been calculated only when at least one insertion and one deletion were available.

[c] Average size in bp. Standard error is given in parenthesis.

[d] Average frequency of the indel event. Standard error is given in parenthesis.

[e] Fraction of indels smaller or equal to 10 bp.

[f] One insertion of 132 bp was excluded.

[g] One deletion of 113 bp was excluded.

very large indels (one insertion in an intergenic region and one deletion in an intron, both > 100 bp), non-repetitive deletions are larger than insertions in both intergenic regions and introns (Wilcoxon test, $P = 0.005$ and $P = 0.034$, respectively; unless indicated, this test is used in all comparisons). Including these two indels, deletions are still significantly larger

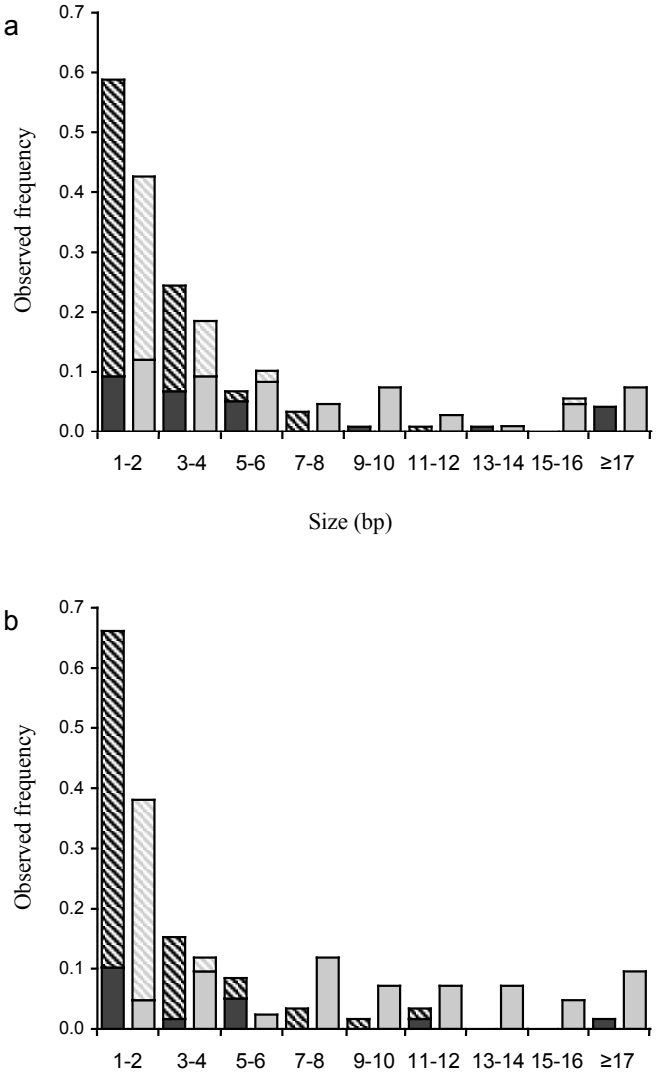


Figure 13. Size distribution of insertions (black bars) and deletions (grey bars) in (a) introns and (b) intergenic regions. The filled portions correspond to non-repetitive indels.

than insertions in intergenic region, but not in introns (data not shown). When repetitive indels are included, the difference is even more significant ($P < 0.005$ for both comparisons).

A consequence of both the higher rate and larger size of deletions is that, in the absence of other forces, a spontaneous loss of DNA should occur. Is this loss compensated? When we average the frequency of each independent indel in the sample, we note that insertions are in higher frequency than deletions (Table 8). In intergenic regions, this difference is significant when all indels are considered ($P = 0.005$). Similarly, in introns, insertions tend to have higher average frequencies than deletions ($P = 0.162$). These results suggest that insertions in both introns and intergenic regions have a higher probability of fixation than deletions, to compensate for the deletion bias by favoring longer regions of non-coding DNA. This agrees

Table 9. Nucleotide and indel diversity in intergenic regions and introns of *D. melanogaster*.

	Nucleotide data			Indel data					
				Non-repetitive indels				All indels	
	π (SE)	Divergence (SE)	Tajima's *D* (SE)	π (SE)	Divergence [a]	Tajima's *D* (Deletions) [a]	Tajima's *D* (Insertions) [a]	π (SE)	Divergence [a]
Intergenic regions	0.010 (0.001)	0.052 (0.005)	–0.744 (0.110)	0.0009 (0.0001)	0.0062	–0.822	–0.297	0.0026 (0.0004)	0.0129
Introns	0.012 (0.001)	0.064 (0.004)	–0.526 (0.065)	0.0011 (0.0002)	0.0082	–0.527	–0.359	0.0027 (0.0003)	0.0132

Unless indicated, the average (SE) across loci is given.
[a] Fragments were lumped before analysis.

with PARSCH (2003), who proposed that large insertions are positively selected to restore the optimal intron length.

3.1.2.3. Estimates of indel and nucleotide sequence variation

We estimated the average indel diversity π and divergence per nucleotide site, considering indels as binary characters of length 1 bp (*i.e.,* presence *vs*. absence of the derived state; for polarization, see above). To estimate divergence, we used the fixed indels observed between the two species. Introns and intergenic regions show similar values for both non-repetitive and all indels, except that divergence is higher in introns than in intergenic regions (Table 9). There are considerable differences in average nucleotide diversity π between introns and intergenic regions. Intergenic regions are less polymorphic and diverged than introns although these differences are not significant (Table 9). This is in line with recent observations by KERN and BEGUN (2005). Furthermore, the frequencies (SE) of derived variants at polymorphic nucleotide sites are significantly higher in introns than in intergenic regions: 0.291 (0.009) and 0.261 (0.013), respectively ($P = 0.02$).

3.1.2.4. Introns, but not intergenic sequences, are larger in *D. melanogaster* than in *D. simulans*

We observed a significant excess of introns that are longer in *D. melanogaster* than in

D. simulans (39 *vs*. 15, $P = 0.0015$; two-tailed sign test); to be conservative, two introns with equal lengths in both species were counted as if they were smaller in *D. melanogaster*. In intergenic regions, however, no difference is found (12 *vs*. 10, $P = 0.832$). Both observations agree with Comeron and Kreitman's (2000) analysis.

The observed differences between introns and intergenic regions may be either due to different mutational patterns or different selective pressures. Indeed, some studies provide evidence of transcription-coupled repair mechanisms and transcription-associated mutations (TAM) that could lead to specific mutational patterns in introns. This effect is well known in bacteria and yeast (Aguilera 2002). In higher eukaryotes, it has only been observed in genes transcribed in mammalian germline cells, where a bias in base composition rather than in substitution rate is observed (Green *et al.* 2003; Comeron 2004). In Drosophila, no evidence has been found for transcription-coupled repair (de Cock *et al.* 1992; Sekelsky *et al.* 2000), although TAM has been recently proposed as a possible cause of compositional bias observed in introns (Kern and Begun 2005).

The following argument suggests, however, that the observed length differences of introns (but not intergenic regions) between *D. melanogaster* and *D. simulans* are probably due to selection rather than mutation. First, introns have a higher (non-repetitive) indel divergence than intergenic regions (Table 9). This means that either more insertions have been fixed in introns of *D. melanogaster* or more deletions in those of *D. simulans*. Second, PDB estimates for introns and intergenic regions are comparable (Table 8). Therefore, something other than mutation must have caused the observed difference in fixed indel divergence between intronic and intergenic sequences.

3.1.2.5. Analysis of selective constraints

The presence of functional elements and/or specific spacing constraints can severely affect polymorphism and divergence patterns. For example, enhancers contain several transcription-factor binding sites separated by spacers with strong length constraints (*e.g.,* Ludwig *et al*. 1998). Furthermore, Ptak and Petrov (2002) suggested that the large difference between PDB observed in introns and in "dead-on-arrival" non-LTR retrotransposons was due to splicing constraints in introns, causing many deletions (particularly the larger ones)

to be deleterious and be removed by purifying selection. Hence, our finding that intergenic regions show a similar PDB value as introns indicates that our intergenic regions may contain a considerable number of regulatory elements under selective constraints. Several putative transcription-factor binding sites were indeed identified using TRANSFAC (Wingender *et al*. 2000) and MatInspector (Quandt *et al.* 1995) tools. Their density (number of hits per

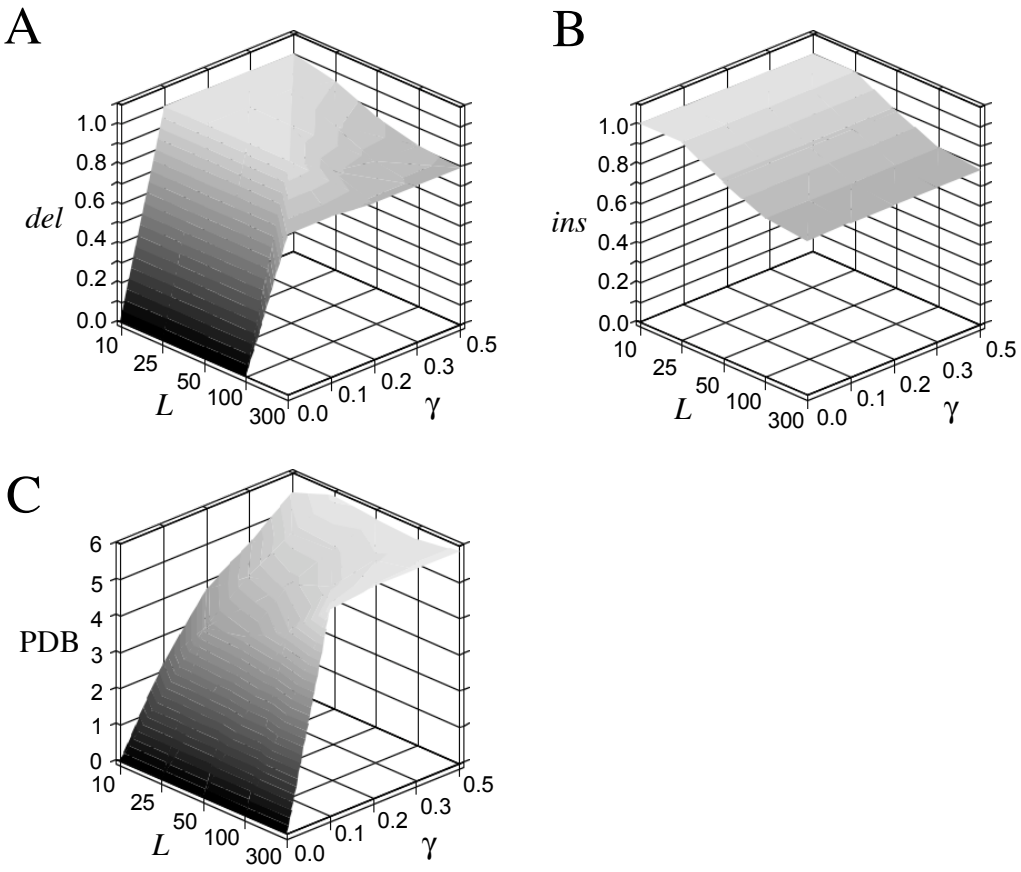


Figure 14. Indel profiles for sequences of length *L* and varying constraints. The fraction of deletions (*del*) and insertions (*ins*) of size ≤ 10 bp are shown (A and B, respectively), as well as the polymorphic deletion bias PDB (C), for sequences of length *L* and varying length constraints. Only a single sequence is considered (rather than a combination of subsequences of different length). The maximum length tolerance is set to $\delta = 0.3$, while the minimum length tolerance has γ values between 0 and 0.5 (see text for details). The fixed value of δ was chosen such that the resulting indel profile was compatible with that observed in our genomic regions. Values of $\delta \geq 0.2$ gave similar results. No single combination of values of the parameters *L* and γ gave theoretical results that were jointly compatible with all three observed indel statistics (*i.e*., values of PDB close to 2, of *ins* around 80 % and of *del* around 70 % for non-repetitive DNA; Table 8), although these observations could be reproduced individually. Long sequences with diverse spacing constraints agree with the value of *ins* (B), but not with the observed PDB and del values (A and C, respectively). Suitable values for PDB are observed for short constrained sequences (C). Low del values are obtained for very strong constraints (where almost all deletions are deleterious), but also in long non-constrained sequences (A). However, in no instance *del* is lower than *ins*. Therefore, this analysis suggests the presence of subsequences of different length and constraints in our fragments.

base pair) does not differ from those of introns (data not shown).

To characterize these constraints and relate them to the observed insertion/deletion pattern, we modeled sequences with a certain proportion of functional non-coding DNA (*e.g.,* exons, regulatory regions; see Figure 12) and calculated the resulting equilibrium deletion and insertion profiles. We assumed that our sequences consist of subsequences delimited by functionally constrained blocks. Preliminary analyses indicated that subsequences of equal (or similar) length are not compatible with our data, independent of the amount of constraints (some examples are provided Figure 14). This suggests the presence of "short" and "long" subsequences with variable length constraints in our fragments.

To model spacing constraints, we considered two contrasting scenarios, in which the short subsequences have either strong (*str*) or relaxed (*rel*) spacing constraints, while only relaxed constraints are present in long subsequences. For the analyses presented here, we assume in the *str* scenario $\delta = 0.1$ and $\gamma = 0$ for the short subsequence, and $\delta = \gamma = 0.3$ for the long subsequence. In the *rel* scenario, $\delta = \gamma = 0.2$ for both subsequences (for the definition of these parameters, see MATERIALS AND METHODS, SUBSECTION 3.1.1.3.). We chose these parameters according to the results reported in Figure 14, in order to obtain theoretical results in close agreement with the observed indel profile. Using $\delta = \gamma \geq 0.2$ in both subsequences or $\gamma = 0$ in the short ones results in indel profiles equivalent to the *rel* and *str* scenarios, respectively.

As shown in Figure 15a, the theoretical results differ according to both sequence composition (*i.e.,* the fraction of short *vs*. long subsequences) and spacing constraints. Depending on whether the short subsequences are under relaxed or strong length constraints, we obtain remarkably contrasting patterns in PDB and the fraction of deletions ≤ 10 bp. When about 85% of the subsequences are short and have strong constraints, we obtain theoretical values close to those observed in both introns and intergenic regions (see Table 8). The indel profiles obtained using short sequences of length ≤ 50 bp and long sequences ≥ 100 bp are similar to those presented. This suggests that the majority of the subsequences in our fragments is indeed short and has strong length constraints.

Our theoretical results provide also evidence that the number of functional elements

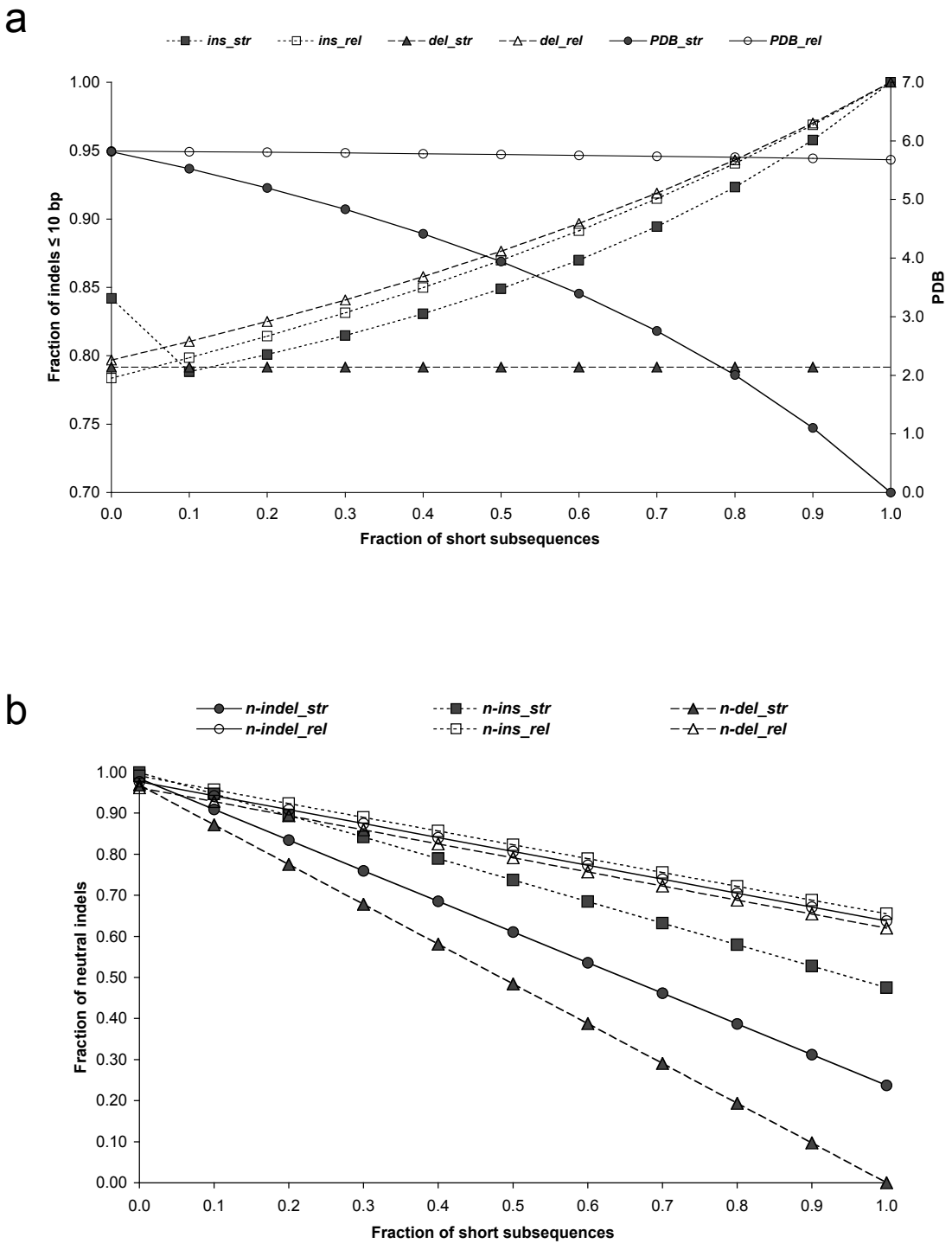


Figure 15. Modeling the insertion and deletion profile in the presence of varying functional constraints. (a) Theoretical results for the fraction of insertions (*ins*) and deletions (*del*) ≤ 10 bp, and the polymorphic deletion bias (PDB). (b) Fraction of insertion (*n-ins*), deletion (*n-del*), deletion and total indel (*n-indel*) events that do not alter functional DNA blocks and spacing constraints. We assume that under neutrality the ratio of deletions to insertions is 6:1, and that there are equal size distributions for insertions and deletions (Petrov and Hartl 1998; Blumenstiel *et al*. 2002). The short and long subsequences have the lengths of 30 bp and 200 bp, respectively, and are subjected to relaxed (*rel*) or strong (*str*) spacing constraints (see text for details).

should not be considered as a direct measure of the amount of constraints. Rather, it is the combined effect of spacing constraints and the proportion of the functional DNA (*i.e.,* the number and spatial extension of the functional elements) that limits the number of neutral mutations (Figure 12). The presence of spacing constraints poses a limit to the number of indels (but not nucleotide substitutions) that can accumulate in the subsequence. Figure 15b gives the proportion of indels that contribute to the polymorphic indel profile, *i.e.,* the expected indel diversity. Since we observed similar indel polymorphism π in intergenic and intronic sequences, spacing constraints seem to be comparable in the two genomic regions.

The low nucleotide sequence diversity and divergence observed in intergenic regions can be understood noticing that the number and spatial extension of functional elements are sources of distinct constraints. In introns, the branch point (which mediates the formation of the lariat structure during splicing) is – strictly defined – only one nucleotide long and defines two subsequences, including a short one of 20–30 bp that is under strong spacing constraints (Mount *et al.* 1992) (*e.g.,* sequence A in Figure 12). On the other hand, a large regulatory element can determine two equivalent subsequences, separated by a large functionally important sequence (*e.g.,* sequence B in Figure 12). While the indel profile is similar in the two situations, the different proportion of functional DNA may affect the number and pattern of nucleotide substitutions and may result in contrasting diversity values. Thus, because our intronic and intergenic regions have similar PDB values and similar fractions of small indels, they may have similar subsequence structures. In contrast, our nucleotide sequence data (Table 9) suggest that intergenic regions host a larger proportion of constrained DNA, *i.e.,* larger functional elements.

Our simple model of sequence constraints is based on the assumption that a subsequence is completely unconstrained, yet delimited by sequence blocks under very strong purifying selection. However, the following observations suggest that this model needs to be used with care. First, we found evidence that compensatory insertions are under weak positive selection to maintain the proper spacing and structure of regulatory elements, which in turn are often negatively affected by the large and numerous deletions. Second, the observed pattern of Tajima's *D* values also suggests that the sequences are under weak selection pressures. *D* is more negative for both single nucleotide polymorphisms and deletions in

intergenic regions than in introns (Table 9). While the observed excess of rare indels and nucleotide variants, leading to an overall negative Tajima's *D*, is likely the result of population expansion (CHAPTERS 1.1. and 1.2.), the more negative value observed for deletions (than for nucleotide variation) may reflect the action of purifying selection. On the other hand, the less negative Tajima's *D* value for insertions is consistent with weak positive selection (discussed above). Notably, this pattern is more pronounced in intergenic regions than introns. The introns analyzed belong to the large size class (MOUNT *et al*. 1992; STEPHAN *et al*. 1994), very different from the small and most common length class of 61 ± 10 bp (YU *et al*. 2002), which on the other hand show little evidence for constraints (HALLIGAN *et al*. 2004: HADDRILL *et al*. 2005a). Additional evidence for the presence of functional constraints in non-coding DNA is presented in the next chapter.

3.2. Mutational pattern and substitution dynamics in the non-coding DNA of *Drosophila melanogaster*

Novel mutations are subject to neutral processes, such as genetic drift and, in some cases, selection, which determine their probability of fixation. The interplay of these forces shapes genome evolution, and it is therefore of great interest to discern between the two. If the background substitution pattern is constant across the genome, differences in base composition among different classes of sequence (*i.e.*, coding regions, introns, intergenic regions) should correspond to variation in selective pressures or in the mutation/fixation pattern. For example, in the first case, this could correspond to their functional role and/or constrained evolution; and in the second case, transcription associated mutation (TAM) can be responsible for differences in mutational pattern between actively transcribed and silent regions of the genome (Aguilera 2002; Comeron 2004).

The variation in mutation rate and selection is supposed to be weak across non-coding regions; hence, a genome-wide approach is essential to evaluate their effect and importance. In particular, studying molecular evolution across a recombination gradient can be effective in detecting the effects of selection. The rationale is that, the less recombination a locus experiences, the more it suffers the interference of selection at linked loci (Hill and Robertson 1968). If, say, there is selection for AT-rich introns, then there should be a positive correlation between the recombination rate and the AT content in these regions. A problem of this approach is that the mutation pattern may in fact be shaped by recombination itself, since (i) recombination can be mutagenic and (ii) there is evidence that the rate of gene conversion, which may be biased towards particular bases, is correlated with recombination (*e.g.*, GC-biased gene conversion). A second (complementary) approach is to compare the polymorphism to the fixation pattern, since weak selection affects only marginally the frequency of a mutation, but more markedly its fixation probability.

In this study, we combine the information of two parallel analyses of nucleotide diversity in intergenic and intronic regions of *Drosophila melanogaster*, with the intent of elucidating neutral and non-neutral forces acting on non-coding DNA. First, we thoroughly investigate the pattern and the dynamics of nucleotide substitution. Second, we expand our previous analysis of the constraints operating in these regions (CHAPTER 3.1.).

3.2.1. Materials and Methods

3.2.1.1. Data collection and analysis

For this study we analyzed all loci sequenced in the African sample (10–12 lines) of *Drosophila melanogaster* reported in CHAPTER 1.2. and for which we could obtain homologues in both *D. simulans* (by sequencing or BLAST search, http://flybase.org/blast) and *D. yakuba* (only by BLAST search). The genomic positions of the loci were based on the *D. melanogaster* genome release 4.2 (http://flybase.org), and only those non overlapping with coding regions or transposable elements were used for the analysis. This left us with 210 loci (out of 232), 116 located in intronic and 94 in intergenic regions. Sequences were aligned using the CLUSTAL W algorithm as implemented in the application MegAlign (DNAstar; Madison, WI), and adjusted by eye when needed. The homologous sequences of *D. simulans* were used to polarize polymorphisms found in *D. melanogaster*: an allele was considered to be ancestral if present in both species. The availability of *D. yakuba* homologues let us also polarize the fixed substitutions between *D. melanogaster* and *D. simulans*; in addition a mutation was considered to have fixed along the *D. melanogaster* lineage if a different allele was observed in both outgroups; along the *D. simulans* lineage if an allele was found only in this species.

3.2.1.2. Statistical analysis

We calculated basic population genetics statistics, such as θ (WATTERSON 1975), π (TAJIMA 1983), divergence and Tajima's *D* (TAJIMA 1989) using the program NetralityTest, kindly provided by H. Li. These statistics were also calculated separately for the "conserved" and "non-conserved" fractions of the alignments. For the intraspecific analysis, a nucleotide

was considered conserved if present in both *D. melanogaster* and *D. simulans* (*i.e*., when not in correspondence to a gap in *D. simulans*). For the interspecific analysis, a nucleotide was considered conserved if present also in the *D. yakuba* homologous sequence (independently of it being substituted in *D. simulans*).

Recombination rate, *r*, (recombination events per site per generation) was estimated for each fragment with the computer program of COMERON *et al*. (1999), which follows the method of KLIMAN and HEY (1993).

3.2.2. RESULTS

In Table 10 we present an overview of nucleotide variation across our loci. The level of polymorphism observed in intergenic regions is comparable to that of intronic regions (Wilcoxon test, $P = 0.898$; hereafter, all comparisons are tested by Wilcoxon test). Also the frequency spectrum, measured by Tajima's *D*, is similar in the two genomic classes ($P = 0.270$). Intergenic regions are, however, less diverged than intronic regions both to *D. simulans* ($P = 0.045$) and to *D. yakuba* ($P = 0.075$). Divergence was also calculated along either the *D. melanogaster* (only one randomly chosen line was used) or the *D. simulans* lineages by polarizing the fixed differences using *D. yakuba* as an outgroup; hereafter these measures are called Div_{mel} and Div_{sim}, respectively. *D. melanogaster* evolves faster than *D. simulans* ($P = 0.0075$), in agreement with the findings of KERN and BEGUN (2005) (Table 10). The difference is significant for intronic regions alone ($P = 0.040$), but not for intergenic ones ($P = 0.084$). The two classes of non-coding DNA do not differ in their rate of evolution (as measured by either Div_{mel} or Div_{sim}), although intronic regions tend to evolve faster (data not shown; see also Table 10). Thus, there is indication for more constraints in intergenic regions, limiting the rate of their evolution.

3.2.2.1 Divergence correlates with recombination rate in intergenic regions

A striking result reported in CHAPTER 1.2. was the significant positive correlation between divergence (to *D. simulans*) and recombination rate. The subset of loci analyzed in the present study shows a marginally significant correlation (Spearman's $R = 0.131$, $P =$

Table 10. DNA variation in non-coding DNA of *Drosophila melanogaster***.**

	Nucleotide diversity (θ)	Divergence to *D. simulans*	Div_{mel}	Div_{sim}	Divergence to *D. yakuba*	Tajima's *D*
Intronic	0.0124 (0.0006)	0.0653 (0.0025)	0.0243 (0.0014)	0.0288 (0.0016)	0.1613 (0.0065)	–0.660 (0.053)
Intergenic	0.0129 (0.0008)	0.0585 (0.0030)	0.0204 (0.0013)	0.0250 (0.0016)	0.1465 (0.0073)	–0.738 (0.065)
All	0.0126 (0.0005)	0.0623 (0.0020)	0.0226 (0.001)	0.0271 (0.0012)	0.1546 (0.0049)	–0.695 (0.041)

Divergence was calculated to *D. simulans*, *D. yakuba* and also along either the *D. melanogaster* (considering only one randomly picked line; Div_{mel}) or the *D. simulans* lineage (Div_{sim}; all values are Jukes-Cantor corrected; see text). Averages (SE) are reported for intronic and intergenic regions separately, as well for the combined dataset.

0.058; hereafter, all correlations are tested using Spearman's correlation). Interestingly, the correlation is significant for intergenic regions ($R = 0.259$, $P = 0.012$), but not for intronic regions alone ($R = 0.036$, $P = 0.705$). However, Div_{mel} does not correlate with recombination in either intronic or intergenic regions ($R = –0.044$, $P = 0.640$, and $R = 0.158$, $P = 0.128$, respectively), while Div_{sim} correlates significantly only in intergenic regions ($R = 0.233$, $P = 0.024$; in intronic regions $R = 0.155$, $P = 0.096$).

Due to the relatively recent split between *D. melanogaster* and *D. simulans*, divergence might still retain the signature of a probable correlation between recombination rate and polymorphism in their ancestor (Hellmann *et al*. 2003). To test this hypothesis, we investigated the relationship between recombination rate and divergence to *D. yakuba*, which is a far more distant outgroup. The correlation disappears for all pooled loci ($R = –0.006$, $P = 0.936$), but in intergenic regions is still present ($R = 0.223$, $P = 0.031$), while in intronic regions it is, although not significantly, slightly negative ($R = –0.174$, $P = 0.062$).

3.2.2.2. Differences in size among homologous loci in Drosophila

In *D. simulans*, loci are shorter compared to their homologues in *D. melanogaster* and *D. yakuba*, also when distinguishing between loci located in intergenic and intronic regions (Table 11). In all cases, the differences among the three Drosophila species are not significant ($P > 0.15$).

Nonetheless, we observed a significant excess of intronic and intergenic regions that are longer in *D. melanogaster* than in *D. simulans*, suggesting that the latter tends to have a more compact genome (87 *vs*. 29, $P < 0.0001$; and 71 *vs*. 23, $P < 0.0001$, respectively; two-

Table 11. Length and base composition of the analyzed intronic and intergenic loci.

	Length		% GC	
	Intronic	Intergenic	Intronic	Intergenic
D. melanogaster	503.2 (11.1)	517.6 (11.3)	0.380 (0.005)	0.415 (0.006)
D. simulans	486.7 (11.0)	505.2 (11.1)	0.390 (0.005)	0.420 (0.006)
D. yakuba	504.3 (16.7)	527.0 (12.1)	0.386 (0.005)	0.415 (0.007)

For each of the three Drosophila species, and non-coding DNA class, we report the average length (SE) expressed in nucleotides, and the average (SE) GC content. In *D. melanogaster*, the averages are calculated across lines (n = 10–12).

tailed sign test; to be conservative, loci with equal lengths in both species were counted in the minor class). Likewise, both classes of loci are longer in *D. yakuba* than in *D. simulans* (68 *vs*. 48, $P = 0.039$; and 60 *vs*. 34, $P = 0.005$, respectively). In contrast, *D. yakuba* and *D. melanogaster* show comparable lengths (49 *vs*. 67, $P = 0.057$; and 49 *vs*. 45, $P = 0.697$, respectively).

3.2.2.3. In intronic regions, GC content is lower than in intergenic regions and does not correlate with recombination

In all three species, intronic regions are less GC rich than intergenic regions ($P < 0.001$; Table 11). Several studies suggest that GC content depends on the rate of recombination, since there is evidence for GC-biased gene conversion. However, intergenic and intronic regions have been sampled in regions of comparable recombination rate (average r is 3.44×10^{-8} for intergenic and 3.57×10^{-8} for intronic regions, respectively, $P = 0.347$). We observed a negative correlation between GC content and recombination, in agreement with Singh *et al*. (2005a), but this holds only in intergenic regions ($R = -0.251$, $P = 0.015$) and not in intronic regions ($R = -0.086$, $P = 0.361$). The trend is opposite to what has been reported for autosomes (Singh *et al*. 2005b).

3.2.2.4. Substitution patterns in Drosophila

To get insights into the causes of the contrasting observations in the two types of non-

coding DNA, we analyzed polymorphism and substitution patterns across our loci.

Mutations were classified according to the ancestral and derived state: *e.g.*, AT→GC refers to polymorphic sites, or fixed substitutions, where A or T mutated to G or C. An overview of the frequencies of the substitutions is given in Table 12.

We first concentrate on the polymorphism pattern. A total of 1920 and 1564 polymorphic sites were polarized in intronic and intergenic regions, respectively.

A selection-driven reduction in GC content in intronic *vs*. intergenic regions and along the recombination gradient can be achieved in two ways: (i) increasing the rate of AT→GC over GC→AT mutations, or (ii) decreasing/increasing the fixation probability of AT→GC / GC→AT, respectively. GC nucleotides show a significantly stronger tendency to mutate to AT than AT to GC, *i.e.* the mutation pressure is extremely asymmetric ($P < 0.0001$, for both intergenic and intronic regions; Table 12). However, neither the fraction of mutations of type AT→GC nor that of GC→AT is correlated with recombination ($P > 0.2$, for both intergenic and intronic regions). This allows us to reject the hypothesis that, in intergenic regions, the large GC content in regions of low recombination results from a recombination-associated bias towards a higher AT→GC rate, unless the subtle effect remained undetected on the polymorphism pattern (due to the weak nature of this selection). If GC are disfavored, AT→GC should segregate at a lower frequency than GC→AT. The frequency spectra of both mutation classes are shown in Figure 16: AT→GC segregate at an average (SE) frequency of 0.291 (0.009), significantly higher than GC→AT, which have an average (SE) frequency

Table 12. Mutational pattern in Drosophila non-coding DNA.

	Polymorphic		Fixed in *D. melanogaster*		Fixed in *D. simulans*	
	AT→GC	GC→AT	AT→GC	GC→AT	AT→GC	GC→AT
Intronic	0.0144 (0.0009)	0.0397 (0.0020)	0.0130 (0.0009)	0.0317 (0.0021)	0.0169 (0.0010)	0.0197 (0.0014)
Intergenic	0.0155 (0.0011)	0.0394 (0.0024)	0.0147 (0.0012)	0.0252 (0.0020)	0.0157 (0.0010)	0.0164 (0.0015)
All	0.0149 (0.0007)	0.0396 (0.0015)	0.0137 (0.0007)	0.0288 (0.0015)	0.0164 (0.0007)	0.0182 (0.0010)

Mutations segregating in *D. melanogaster* were polarized using the *D. simulans* homologous sequence; fixed substitutions between *D. melanogaster* and *D. simulans* were polarized using the *D. yakuba* homologous. We give the average fraction (SE) of A or T nucleotides mutating to G or C (AT→GC) and of G or C nucleotides mutating to A or T (GC→AT): AT→TA and CG→GC, and mutations for which an unambiguous polarization was not possible, were not considered in this analysis.

of 0.256 (0.006) (P = 0.0008). Comparable results are obtained when considering intronic and intergenic regions separately (data not shown). Moreover, in intergenic regions there is a weak but significant positive correlation between AT→GC frequencies and the recombination rate (R = 0.098, P = 0.042). This is unexpected, since it should result in a positive correlation between GC content and recombination, while we found exactly the opposite, *i.e.* a *negative* correlation. A possible bias introduced in the above analysis comes from the overall excess of rare variants in the *D. melanogaster* genome, due to population-size expansion (CHAPTER 1.2.). If we exclude singletons from our analysis, AT→GC and GC→AT mutations segregate at similar average frequencies (P > 0.474, for both intronic and intergenic regions), and mutation frequencies do not correlate with recombination (data not shown). Thus, while there is no clear indication for selection toward GC in intergenic regions, the lack of a difference in the asymmetry of the mutational pattern between intronic and intergenic regions (P > 0.250) suggests that the differences in composition and polymorphism pattern are likely the result of contrasting combinations of selective and neutral forces.

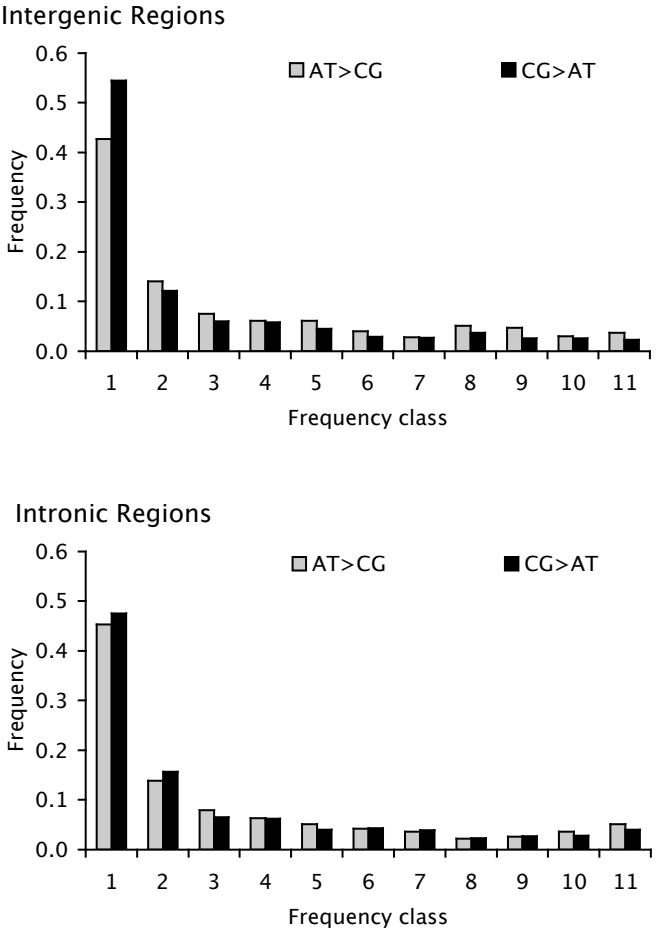


Figure 16. Polymorphism frequency spectrum in non-coding DNA of *D. melanogaster*. The frequencies of AT→GC and GC→AT mutations are shown in grey and black bars, respectively, for both intergenic and intronic regions. Note the contrasting frequency pattern of rare (*i.e.*, frequency class = 1) and common mutations (*i.e.*, frequency class = 11) for AT→GC versus GC→AT.

We begin to investigate this hypothesis by considering GC-biased gene conversion (BGC), *i.e*. a preference to correct a mismatch towards G or C during the repair of double-strand breaks (*e.g*., Birdsell 2002; Galtier *et al*. 2001). This neutral process has been only recently described in *D. melanogaster* (Galtier *et al*. 2005), and it increases the probability of fixation of GC over AT in a way indistinguishable from selection (Nagylaky 1983). Consistent with the BGC model, the AT→GC frequency is significantly higher than that of GC→AT, and it is positively correlated with the recombination rate in intergenic regions (see above). Moreover, while GC→AT are more abundant than AT→GC mutations as singletons, the opposite is true for alleles segregating at high frequency (Figure 16). The difference is significant in intergenic regions, but not in intronic regions ($\chi^2 = 4.09$, $P = 0.043$, and $\chi^2 = 0.99$, $P = 0.320$, respectively). We can speculate that the contrasting sizes of rare and common frequency classes in AT→GC *vs*. GC→AT might be the result of BGC, which, by favoring G and C nucleotides, would decrease the frequency of a mutation of type GC→AT and increase that of type AT→GC.

The availability of the *D. yakuba* genome let us polarize the substitutions fixed between *D. melanogaster* and *D. simulans*, and determine in which of the two lineages the fixation took place. Unless stated, our analysis will focus only on the substitutions that fixed in *D. melanogaster*. A total of 996 and 729 fixed substitutions were polarized in intronic and intergenic regions, respectively. AT→GC have a significantly higher tendency to go to fixation than GC→AT in both intergenic and intronic sequences ($\chi^2 = 9.06$, $P = 0.003$; and $\chi^2 = 8.45$, $P = 0.004$, respectively), in agreement with BGC predictions (Table 12).

There is no reason to think that a neutral process, such as BGC, would affect intergenic and intronic regions in different ways. Therefore, additional forces are necessary to explain the contrasting base composition and polymorphism pattern.

The average frequencies of polymorphisms are not different between the two genomic regions: intergenic regions show a significant excess only of GC→N over AT→N singletons compared to intronic regions ($\chi^2 = 7.92$, $P < 0.005$; N indicates a change to any other base), with no trend to increase the GC content (*i.e*., GC→AT *vs*. AT→GC in the two non-coding classes; $\chi^2 = 0.36$, $P = 0.550$). Again, we analyzed the fixation pattern: no difference in AT enrichment (*i.e*., fixed GC→AT *vs.* AT→GC) between introns and intergenic regions is present ($\chi^2 < 0.01$, $P = 0.999$). Furthermore, there is no correlation between recombination rate and the fraction of substitutions fixing either AT or GC (for both intronic and intergenic

regions; data not shown). Overall, these results fail to detect any bias in the fixation pattern that could explain the difference in base composition between introns and intergenic regions. Curiously, if the analysis of the fixed pattern is done using only a randomly chosen line of *D. melanogaster* (thus "including" polymorphic as fixed differences), intronic regions show a higher ratio of GC→AT over AT→GC fixations than intergenic regions ($\chi^2 = 4.15$, $P = 0.042$), consistent with their larger AT content.

3.2.2.5. Comparing the fixation pattern of *D. melanogaster* and *D. simulans*

The type and frequency of fixed substitutions were determined using alignments consisting of one randomly chosen line of *D. melanogaster* and the homologous sequences from *D. simulans* and *D. yakuba* (MATERIALS AND METHODS, SUBSECTION 3.2.1.1.). The following results are equivalent if all *D. melanogaster* lines were considered (data not shown). We analyzed a total of 1319 and 1044 polarized substitutions fixed in *D. melanogaster* intronic and intergenic regions, respectively; for *D. simulans*, the counts are 1108 and 897, respectively.

The two species fix GC→AT and AT→GC in a complete opposite fashion: while GC→AT are more numerous than AT→GC in *D. melanogaster* (400 *vs*. 345 and 553 *vs*. 390, for intergenic and intronic sequences, respectively), the reverse trend is observed in *D. simulans* (262 *vs*. 376 and 346 *vs*. 485, respectively), and the differences are highly significant ($\chi^2 > 21.95$, $P < 0.0001$).

3.2.2.6. Base composition of indels

In addition to point mutations, a modification in nucleotide composition can be achieved by insertions and deletions. That is, a bias in the indel base composition can affect that of the regions under study. Here, we consider the portion of the alignments that consist of insertions or deletions.

Insertions (*i.e*., the pooled inserted DNA) are more GC rich than deletions in intronic sequences only, both when still segregating (34% *vs*. 26%, $P = 0.005$) and fixed in *D. simulans* (37% *vs*. 31%, $P = 0.054$) or in *D. melanogaster* (38% *vs*. 35%, $P = 0.071$). Interestingly, fixed insertions are more GC rich than those still segregating ($P = 0.001$), suggesting a fixation bias.

Both polymorphic insertions and deletions are less GC rich than the surrounding intronic DNA (which has 38% of GC; $P < 0.0001$, and $P = 0.014$, respectively), while in

intergenic regions this holds for insertions, but not for deletions ($P = 0.001$, and $P = 0.244$, respectively). The differences are, however, no longer significant for the deletions fixed in *D. melanogaster* ($P > 0.700$, for both intergenic and intronic regions), while only in intronic regions fixed insertions increased the average GC content in both Drosophila species ($P < 0.002$). It is noteworthy that, in intronic regions, *D. melanogaster* fixed deletions that are slightly higher in GC content than the remaining sequence (38% *vs*. 39%, P = 0.743), while the opposite is true in *D. simulans* (39% *vs*. 37%, P = 0.011), analogous to the contrasting nucleotide mutation biases observed before. Singh *et al*. (2005b) reported a significantly higher GC content of deletions than the rest of the ancestral sequence in transposable element (TE) dispersed in the *D. melanogaster* autosomes. Our opposite results may be the result of different indel dynamics of TEs or in autosomes, since the indistinguishable GC content of polymorphic *vs*. fixed indel DNA does not support contrasting selective forces in TE *vs*. non-coding DNA.

3.2.2.7. Inferring constraints in conserved intergenic and intronic regions

In the previous chapter, we gave evidence for the presence of constraints in intergenic and intronic sequences, thereby limiting the neutral accumulation of both nucleotide and insertion/deletion (indel) variation (chapter 3.1.). Here, we use the absence of insertions or deletions as a proxy for evolutionary constraints. The rationale is that, if a sequence is not conserved across species, it is less likely to contain functionally important DNA. Hence, we

Table 13. Nucleotide variation in conserved and non-conserved non-coding DNA.

	Nucleotide diversity (θ)		Divergence to *D. simulans*		Tajima's *D*	
	Conserved	Non-conserved [a]	Conserved	Non-conserved [b]	Conserved	Non-conserved
Intronic	0.0125 (0.0006)	0.0145 (0.0017)	0.0607 (0.0024)	0.1163 (0.0078)	–0.636 (0.059)	–0.256 (0.126)
Intergenic	0.0129 (0.0008)	0.0159 (0.0023)	0.0550 (0.0028)	0.1056 (0.0082)	–0.729 (0.060)	–0.267 (0.138)
All	0.0127 (0.0005)	0.0151 (0.0014)	0.0581 (0.0018)	0.1114 (0.0056)	–0.678 (0.042)	–0.261 (0.093)

Average (SE) nucleotide diversity, divergence and Tajima's *D* statistic were calculated separately for the conserved and non-conserved fractions of the analyzed loci. For the intraspecific analysis, a nucleotide was considered conserved if present in both *D. melanogaster* and *D. simulans* (*i.e*., when not in correspondence to a gap in *D. simulans*). For the interspecific analysis, a nucleotide was considered conserved if present also in the *D. yakuba* homologous sequence (independently of it being substituted in *D. simulans*).

[a] θ values above 0.1 were removed (one intronic and 6 intergenic regions); see text.

[b] Divergence values above 0.3 were removed (3 intronic and 4 intergenic regions); see text.

separate the analysis of nucleotide polymorphism and divergence in the alignment stretches where a deletion was present, or absent, in *D. simulans* or *D. yakuba*, respectively (Table 13).

As expected, DNA that is conserved between *D. melanogaster* and *D. simulans* is more variable than that shared in the two species: in intergenic regions average θ (SE) are 0.0294 (0.0064) and 0.0129 (0.0008), respectively ($P = 0.289$); in intronic regions, the values are 0.0174 (0.0034) and 0.0125 (0.0006), respectively ($P = 0.110$). When θ values above 0.1 were removed, to account for the disproportionately high per-site polymorphism in too short stretches (*i.e.*, short deletions in *D. simulans*), the differences are significant in intergenic regions ($P = 0.048$), but not in intronic regions ($P = 0.084$). No significant difference in polymorphism is present between intergenic and intronic DNA (data not shown).

Constraints limiting the accumulation of polymorphism should translate into reduced molecular evolution. Divergence between *D. melanogaster* and *D. simulans* is significantly lower in the DNA conserved across the three species than where *D. yakuba* has a gap, in both intergenic ($P < 0.0001$) and intronic regions ($P < 0.0001$) (divergence values above 0.3 were not considered; see above). Interestingly, while non-conserved DNA has diverged equally in intergenic and intronic regions ($P = 0.492$), the rest of the sequence experienced a marginally significant faster evolution in intronic compared to intergenic regions ($P = 0.061$).

The presence of constraints is also evident by the stronger skew towards rare variants in the conserved DNA, suggesting the action of (weak) purifying selection. Tajima's *D* values are more negative in the conserved than in the non-conserved portions of both intergenic (P = 0.084) and intronic regions (P = 0.015).

In order to infer signatures of selection intensity/efficiency, we correlated θ, divergence and Tajima's *D* with recombination rate. Only divergence in both conserved and non-conserved intergenic DNA shows a significant correlation ($R = 0.238$, $P = 0.021$, and $R = 0.274$, $P = 0.010$, respectively), similar as for the whole locus (see above).

An interesting feature of intronic regions is a significant negative correlation between divergence to *D. simulans* and their length ($R = –0.184$, $P = 0.048$): this suggests that, in introns, constraints increase when length becomes larger, as supported by the contrasting correlations when analyzing conserved and non-conserved portions separately ($R = –0.172$, $P = 0.065$, and $R = –0.074$, $P = 0.437$, respectively).

3.2.3. Discussion

Our analysis of the mutational and compositional patterns of Drosophila DNA revealed important features of the causes and dynamics of natural variation. In particular, we found several lines of evidence suggesting contrasting neutral, and possibly selective, forces in intergenic *vs*. intronic regions.

3.2.3.1. Intergenic and intronic regions have different base composition but similar mutation patterns

Both mutational pattern and diversity analysis point to different forces acting on intergenic and intronic regions. Namely, introns are more AT rich than intergenic regions, and recombination rate affects GC content only in the latter. Nonetheless, despite intensive analysis, we could not find any bias in the mutational pattern to explain these observations. Rather, there is some evidence that GC-biased gene-conversion and a fixation bias tend to increase the GC content, especially in regions of high recombination. However, two effects oppose this process: (i) the mutational rate is asymmetrical and skewed from GC to AT, and (ii) all mutations (and, consequently, θ) decrease in number with increasing GC content ($P < 0.0001$, for both intergenic and intronic regions). In other words, the more AT rich a locus is, the more polymorphic it is, with most of the mutations of type GC→AT. The effect of regional base composition on mutations was recently observed by Morton *et al*. (2005), who reported a positive relationship between AT content and GC→AT pressure in maize. The context dependency of mutations cannot be ascribed to CpG deamination, since in Drosophila methylation is very rare and restricted mainly to non-CpG sequences (Field *et al*. 2004).

3.2.3.2. GC content affects nucleotide diversity and the insertion/deletion dynamics

Intergenic regions are slightly less polymorphic and diverged than intronic regions, and only in intergenic regions is the recombination rate negatively correlated with GC content and nucleotide variation. Therefore, we can hypothesize that GC-biased composition reduces the (speed of) accumulation of new mutations. The partial correlation coefficient between intergenic GC content versus θ (controlling for recombination) is –0.239, while the ones between intergenic GC content versus recombination (controlling for θ) and θ versus

recombination (controlling for intergenic GC content) are –0.237 and 0.181, respectively. Thus, there is an indication that in intergenic regions GC content affects polymorphism. To investigate this finding in more detail, we removed the effect of the recombination rate on θ by considering only their correlation's residuals. The accordingly corrected θ values still correlate with GC content (R = –0.290, P = 0.005), while, as expected, they do not correlate with recombination rate (R = –0.065, P = 0.535). On the other hand, the corrected θ values do not correlate with recombination rate when removing the effect of GC content (R = 0.122, P = 0.240), The above results hold also when divergence is analyzed instead of θ (data not shown), suggesting that GC content is a strong determinant of nucleotide diversity. Additionally, there is a negative correlation between GC content and Tajima's D (R = 0.189, P = 0.068, and R = 0.284, P = 0.002, for intergenic and intronic regions, respectively). Also, divergence to *D. simulans* is negatively correlated with the GC content (P < 0.0001, for both intergenic and intronic regions), in agreement with the findings of Haddrill *et al.* (2005a), who, however, limited their analysis only to introns. This does not hold when Div_{mel} or Div_{sim} are plotted against the interspecific conserved GC content (data not shown).

Alternatively (but not mutually exclusively), the same forces that are shaping polymorphism across the recombination gradient (*i.e.*, selection) are also responsible for the compositional patterns. In agreement with this hypothesis, Begun and Aquadro (1992) and Beisswanger and Hutter (personal communication) reported a positive correlation between levels of variation and recombination also for autosomal loci: in contrast to the X chromosome, autosomes show a (slight) *positive* correlation between GC content and recombination (Marais *et al*. 2001; Marais *et al*. 2003; Singh *et al*. 2005a).

In general, indels are less GC rich than the surrounding sequence, with insertions being overall more GC rich than deletions. These results, and the higher fixation probability of the insertions high in GC content, suggest that AT-rich stretches are more "unstable". In line with this hypothesis, the quantity of inserted and deleted DNA (both segregating or fixed) diminishes as the GC content increases (R < –0.168, P < 0.015, across all four correlations). The indels' overall contribution on the composition depends on their number, frequencies and size. In a previous analysis, we found that deletions are more numerous and longer, but their frequency is lower, than insertions (chapter 3.1.). We calculated the fraction of the *D. melanogaster* sequence affected by segregating insertions or deletions (thus combining the three above variables), and found that there is equilibrium between the two processes (P >

0.6 in both intergenic and intronic regions). In contrast, when considering the fraction of inserted and deleted nucleotides that fixed along the *D. melanogaster* lineage (calculated as the ratio between total indel length and the conserved alignment between *D. melanogaster* and *D. simulans*), a tendency to increase the length is evident ($P = 0.075$ and $P = 0.016$, for intergenic and intronic regions, respectively). We indeed found that both non-coding DNA classes are larger in *D. melanogaster* than in *D. simulans*.

Interestingly, in intergenic regions the fraction of inserted DNA increases with recombination rate ($R = 0.204$, $P = 0.049$, and $R = 0.201$, $P = 0.051$ for segregating and fixed insertions, respectively). Is this a consequence of GC content, *i.e.*, is the higher the GC content limiting the accumulation of indels (see above)? We corrected the values of the fraction of inserted DNA by taking the residuals of their correlation with GC content: no positive correlation with the recombination rate is present (rather, it is slightly negative; $R = –0.183$, $P = 0.078$). Therefore, GC content is also an important determinant of indel variation. Again, the fact that this holds primarily in intergenic regions, and not in introns, confirms that different neutral and/or selective forces act on the different classes of non-coding DNA.

3.2.3.3. Replication time and transcription-associated mutation bias have negligible effects on the mutation pattern

The variation in the regional composition and mutation bias could be a secondary effect of the local replication time, which coupled with a variation in the available nucleotide pool, could produce the observed pattern (see also Morton *et al*. 2005). We correlated GC content and GC→AT/AT→GC mutation pattern with the replication timing reported by Schübeler *et al*. (2002): *i.e.*, each locus was assigned a replication time equal to the average replication time of the flanking genes, weighted by its distance from them. The only significant correlation was between the AT→GC mutation pressure and replication time in intronic regions, with regions replicated early in the cell cycle having more AT mutating to GC ($R = –0.208$, $P = 0.027$). Thus, replication time seems to be only marginally important in controlling mutation dynamics and base composition across the genome.

A possible explanation for the difference in base composition between introns and intergenic regions could be a differential mutation pressure in transcribed *vs*. silent regions, *i.e.*, the so-called transcription associated mutation bias (TAM; de Cock *et al*. 1992; Sekelski

et al. 2000). On the other hand, the difference in the substitution pattern of two species might be the result of a recent mutational shift (*e.g*., Rodriguez-Trellez *et al*. 2000; Takano-Shimizu 2001) or the effect of a population size decline in *D*. *melanogaster*, which interrupted (or reduced in efficiency) a general GC-biased evolutionary process, as proposed by Galtier *et al*. (2005).

3.2.3.4. Longer introns are under more constraints

The negative correlation between length of the intronic loci and divergence agrees with Parsch (2003) and Haddrill *et al*. (2005a), who proposed that long introns (i) host more regulatory elements or (ii) suffer more constraints to limit pre-mRNA secondary structures, thus limiting their rate of evolution. In fact, the AT-biased composition of introns compared to intergenic regions suggest that GC rich regions may be under-represented to avoid strong pre-mRNA secondary structure (G:C bonds being stronger than A:U ones), which would in turn affect splicing efficiency. To maintain high AT content, a balance between the higher tendency of GC to mutate to AT and the GC-fixation bias is needed. We found an excess of GC→AT *vs*. AT→GC fixations in *D*. *melanogaster*, compared to *D*. *simulans*, which could have contributed to the average high AT content of the genome. However, the fact that we observed the same compositional trend in *D. melanogaster* and *D. simulans*, despite the opposite fixation bias in the two species, suggests that intronic DNA is under particular evolutionary forces/constraints.

3.2.3.5. Evidence for selective constraints in non-coding DNA

Our survey of the levels of variation in "conserved" and "non-conserved" non-coding DNA suggested the presence of constraints in the conserved fraction that limit the accumulation of nucleotide polymorphism. These constraints likely correspond to functional elements, such as enhancers, transcription-factors binding sites, sequences involved in pre-mRNA secondary structure, etc. In particular, intergenic regions show a higher degree of conservation, evident by their lower polymorphism and divergence compared to introns and the presence of a significantly lower DNA turnover rate (expressed as the fraction of DNA consisting of fixed insertions and deletions, $P = 0.002$).

An alternative explanation for the difference in variation between conserved and non-conserved regions is that GC content affects variation levels and that segregating indels

are less GC rich than the surrounding region. However, the distinction between conserved and non-conserved regions was based on the absence, or presence, of fixed deletions in the outgroup, respectively. Since we did not observe a significant difference between the GC content in fixed indels and the flanking region, we can eliminate this explanation.

3.2.3.6. Positive correlation between divergence and recombination rate in intergenic regions

An interesting confirmation of this study is the positive correlation between recombination rate and divergence. In particular, we found that this correlation holds both when *D. melanogaster* is compared to *D. simulans* and also when a distant outgroup as *D. yakuba* is used. An explanation might be that the positive correlation between recombination rate and divergence mirrors that between GC content and recombination rate (see above). Alternatively, this finding would suggest that recombination is mutagenic. However, we observed such a correlation only in intergenic regions, which, because they are likely under more constraints than introns, might host more functional elements. Recently, Andolfatto (2005) provided evidence for adaptive evolution in non-coding DNA of the X chromosome of *D. melanogaster*. This finding raises the intriguing possibility that the positive correlation we observed between recombination rate and divergence might be the signature of positive selection in functional regions being more effective in regions of high recombination (Birky and Walsh 1988; Betancourt and Presgraves 2002). Given our results, the correlation would hold only for divergence along the *D. simulans* lineage because of its larger population size compared to *D. melanogaster* (thus increasing selection efficiency) and/or an historical population crash of the latter, which resulted in the fixation of the most numerous deleterious mutations masking the fixation of the adaptive ones. Several lines of evidence suggest indeed that *D. melanogaster* decreased in effective population size, decreasing selection efficiency in synonymous codon bias (Akashi 1996) and on current codon use (Akashi and Schaeffer 1997; McVean and Vieira 2001), and lowering diversity at neutral sites (Moriyama and Powell 1996). Furthermore, recent studies indicate that this species is recovering in population size from an old bottleneck (chapter 1.2.; Haddrill *et al*. 2005b). However, these hypotheses must be taken with caution due to the pervasive effect of base composition on the evolutionary process.

Conclusions

Nowadays, genome-wide studies are becoming popular among researchers, with the available sequencing technology making vast datasets readily available for many population genetic studies. Compared to the single-locus approaches, multi-locus ones have several advantages to detect local features within the genome without *a priori* knowledge. Moreover, the inference of selective and neutral patterns is feasible only through the analysis of considerable datasets, without which we would lack the necessary power.

For example, large-scale analyses of genome sequences have allowed us to understand the basis of base composition variation across the genome (*e.g.*, GC content: Fullerton *et al*. 2001; Marais *et al*. 2003) and the identification of recombination "hot spots" in humans (McVean *et al*. 2004) and chimpanzees (Ptak *et al*. 2005). Besides neutral variation, the most important challenge is to identify and characterize functionally important genetic variation. This can be achieved through microarray-based expression analyses, association studies and quantitative-trait-locus analyses (for a review, see Vasemägi and Primmer 2005). Population genetics studies used multi-locus neutrality tests and likelihood approaches to dissect the effects of natural selection from background neutral patterns. The analysis of multiple coding regions revealed the pervasive weak selection on codon usage (*e.g.*, Hey and Kliman 2002) and was used to estimate the genomic rate of adaptive amino acid substitution (*e.g.*, in Arabidopsis, Bustamante *et al*. 2002; in Drosophila, Bierne and Eyre-Walker 2004; in primates, Fay *et al*. 2001). Finally, genome surveys of nucleotide variation can be used to map targets of natural selection. This methodology, known also as hitchhiking mapping, is based on an expected reduction of polymorphism around such adaptive mutations (Maynard Smith and Haigh 1974), coupled to a skew in the allele frequency spectrum and increased linkage disequilibrium. This approach has been used in Drosophila (see chapters 1.1. and

1.2.; Harr *et al*. 2002), Arabidopsis (Schmid *et al*. 2005), maize (Wright *et al*. 2005), and humans (Payseur *et al*. 2002; Akey *et al*. 2004).

In this thesis, we studied the polymorphism pattern of an ancestral and a derived population of *Drosophila melanogaster* in great detail. By using an efficient combination of classical population genetic approaches, coalescent simulations and large datasets, we were able to get significant insights on the demographic and selective history of this species. In particular, our results and new methodologies could answer a series of fundamental questions.

1) What are the joint effects of the demographic and the selective history of D. melanogaster? (chapter I)

A major problem complicating the detection of valleys of reduced variation are the confounding effects of the population-size bottleneck accompanying the colonization. By sequencing multiple loci (>250) we could decouple the two effects, as demography affects the whole genome, while natural selection acts only locally. Nucleotide polymorphism in the derived European population is much lower than that of the ancestral African population (about one third). The concomitant high levels of linkage disequilibrium in the European sample strongly indicate that this population experienced a recent population-size bottleneck. Is this bottleneck sufficient to explain the high number of loci with no variation? In a preliminary analysis we showed that this was unlikely under most of the realistic demographic scenarios, giving a first indication that the European population most likely suffered novel selective pressures to adapt to the new environment. We therefore moved on and developed a maximum-likelihood method to estimate the age of the bottleneck. This approach made use of the whole dataset and led to the estimation of a bottleneck ~8,000-16,000 years old. We were then in the position to test whether our data are consistent with this demographic model.

2) Can we identify regions of the genome with a footprint of natural selection? (chapter I)

The reduction in size of a population going through a bottleneck is reflected by the loss of most of its ancestral polymorphism. Moreover, the initially small size of the derived population amplified the effects of genetic drift and of stochastic processes (as the contribution

of each founder to the final pool and, in the end, to our sample). This translates into a non-negligible probability to observe a (complete) loss of the ancestral polymorphism simply due to neutral processes. An additional factor to consider is the ancestral locus-specific variation, since the bottleneck could reduce it to near zero. In our coalescent-based approach, we took all these elements into account by (i) simulating our sample under the estimated population bottleneck, and (ii) assuming a locus-specific mutation parameter, estimated from the ancestral population data. As expected, most of the loci with low polymorphism in the derived population can be explained by demography alone. Most importantly however, we detected several loci and regions of the X chromosome with less polymorphism than expected under the bottleneck model. Furthermore, a likelihood ratio test supported the presence of two separate bottlenecks in our dataset (the second mimicking the action of positive selection) to a single one. These findings strongly support the claim that this population experienced numerous adaptive events during the recent past.

3) Is there evidence for positive selection at a fine scale? (CHAPTER II)

One of the main objectives of this research project was to find evidence for positive selection. We focused on one of the candidate regions identified by our maximum-likelihood approach, by sequencing and analyzing 14 loci densely distributed across a ~45 kb region. The polymorphism pattern fits well the expectations of a selective sweep: we found a strong reduction in polymorphism across 4 adjacent loci in the European sample, whereas they have normal values of ancestral polymorphism and of divergence. Only 3 derived singletons are found within the 5 kb-wide valley, producing very negative Tajima's *D* values, while positive values are found in the flanking regions, in agreement with hitchhiking theory. Coalescent simulations showed that the pattern is incompatible with the simple action of a population-size bottleneck. Rather, we estimated a much more recent episode of local decrease in population-size, likely caused by natural selection. This finding confirms the power of our approach to distinguish between the effects of demography and selection, and confirms that natural selection was an important force driving the evolution of the European population.

4) Is non-coding DNA evolving neutrally? (CHAPTER III)

The debate on the whether non-coding DNA has any importance has recently been

settled by the finding of pervasive constraints and, possibly, adaptive significance of this class of DNA. We report the presence of functional constraints in both intergenic and intronic regions, limiting the accumulation of both nucleotide and insertion/deletion mutations. Moreover, we show that the base composition is variable across the genome, and that it influences the mutational pattern and the rate of evolution. These findings have important consequences on the analysis of polymorphism. One important assumption of the genomic scan was that the large majority (if not all) of our loci were evolving neutrally. Our analyses found instead the presence of selective constraints that need to be taken into account. In fact, all simulations used to estimate the bottleneck and test our loci of the European sample relied on locus-specific mutation parameters inferred from the variation in the ancestral population. Hence, any selective constraint that limited the evolution of that locus was correctly accounted for in the simulations.

Literature cited

AGUILERA, A., 2002. The connection between transcription and genomic instability. *EMBO J*. **21:**195–201.

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

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

AKEY, J. M., M. A. EBERLE, M. J. RIEDER, C. S. CARLSON, M. D. SHRIVER, D.A. NICKERSON, and L. KRUGLYAK. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol*. **2:**e286.

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

ANDOLFATTO, P., 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* **437:**1149–1152.

ANDOLFATTO, P. F. DEPAULIS, and A. NAVARRO. 2001. Inversion polymorphisms and nucleotide variability in Drosophila. *Genet. Res*. **77:**1–8.

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.

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.

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

Beisswanger, S., W. Stephan, and D. De Lorenzo, 2005. Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. *Genetics* **doi:**10.1534/genetics.105.049346.

Bergman, C. M., and M. Kreitman, 2001. Analysis of conserved non-coding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res*. **11:**1335–1345.

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.

Bierne, N., and A. Eyre-Walker, 2004. The genomic rate of adaptive amino acid substitutions in Drosophila. *Mol. Biol. Evol*. **21:**1350–1360.

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

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

Blumenstiel, J. P., D. L. Hartl, and E. R. Lozowsky, 2002. Patterns of insertion and deletion in contrasting chromatin domains. *Mol. Biol. Evol.* **19:**2211–2225.

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.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, D. L. Hartl, 2002. The cost of inbreeding in Arabidopsis. *Nature* **416:**531–534.

Catania, F., and C. Schlötterer, 2005. Non-African origin of a local beneficial mutation in D. *melanogaster*. *Mol. Biol. Evol*. **22:**265–272.

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

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

Chen, Y., and W. Stephan, 2003. Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene. *Proc. Natl. Acad. Sci. USA* **100:**11499–11504.

Comeron, J. M., 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167:**1293–1304.

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.

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

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

de Cock, J. G., A. van Hoffen, J. Wijnands, G. Molenaar, P. H. Lohman, and J. C. Eeken, 1992. Repair of UV-induced (6-4)photoproducts measured in individual genes in the *Drosophila* embryonic Kc cell line. *Nucleic Acids Res.* **20:**4789–4793.

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

Dermitzakis, E. T., A. Reymond, R. Lyle, N. Scamuffa, C. Ucla, S. Deutsch, B. J. Stevenson, V. Flegel, P. Bucher, C. V. Jongeneel, and S. E. Antonarakis, 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420:**578–582.

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.

Erives, A., and M. Levine, 2004. Coordinate enhancers share common organizational features in the Drosophila genome. *Proc. Natl. Acad. Sci. USA* **101:**3851–3856.

Fay, J. C., and C. I. Wu, 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Mol. Biol. Evol*. **16:**1003–1005.

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

Fay, J. C., G. J. Wycoff, and C. I. Wu, 2001. Positive and negative selection on the human genome. *Genetics* **158:**1227–1234.

Field, L. M., F. Lyko, M. Mandrioli, and G. Prantera, 2004. DNA methylation in insects. *Insect Mol*. *Biol*. **13:**109–115.

Fu, Y. X., 1995. Statistical properties of segregating sites. *Theor Popul Biol* **48:**172–197.

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

Fullerton, S. M., A. B. Carvalho, and A. G. Clark, 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol*. **18:**1139–1142.

Furlong, E. E. M., E. C. Andersen, B. Null, K. P. White, and M. P. Scott, 2001. Patterns of gene expression during Drosophila mesoderm development. *Science* **293:**1629–1633.

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

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret, 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159:** 907–911.

Galtier, N., E. Bazin, and N. Bierne, 2005. GC-biased segregation of non-coding polymorphisms in Drosophila. *Genetics* **doi:**10.1534/genetics.105.046524.

Glazko, G. V., E. V. Koonin, I. B. Rogozin, and S. A. Shabalina, 2003. A significant fraction of conserved non-coding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.* **19:**119–24.

Green, P., B. Ewing, W. Miller, P. J. Thomas, NISC Comparative Sequencing Program, and E. D. Green, 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet*. **33:**514–517.

Haddrill, P. R., B. Charlesworth, D. L. Halligan, and P. Andolfatto, 2005a. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. *Genome Biology* **6:**R67.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto, 2005b. Multi-locus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* **15:**790–799.

Haldane, J. B. S., 1957. The cost of natural selection. *J. Genet.* **55:**511–524.

Halligan, D. L., A. Eyre-Walker, P. Andolfatto, and P. D. Keightley, 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. *Genome Res*.

14:273–279.

Hanke, J., D. Brett, I. Zastrow, A. Aydin, S. Delbrük, G. Lehmann, F. Luft, J. Reich, and P. Bork, 1999. Alternative splicing of human genes: more the rule than the exception? *Trends Genet*. **15:**389–390.

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.

Hefferon, T. W., J. D. Groman, C. E. Yurk, and G. R. Cutting, 2004. A variable dinucleotide repeat in the *CFTR* gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proc. Natl. Acad. Sci. USA* **101:**3504–3509.

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.

Hey, J., and R. M. Kliman, 2002. Interactions between natural selection, recombination and gene density in the genes of *Drosophila melanogaster*. *Genetics* **160:**595–608.

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

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, NY.

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., 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.

INNAN, H., and W. STEPHAN, 2003. Distinguishing the hitchhiking and background selection models. *Genetics* **165:**2307–2312.

JENSEN, J. D., Y. KIM, V. BAUER DUMONT, C. F. AQUADRO, and C. D. BUSTAMANTE, 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170:**1401–1410.

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

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.

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

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.

KERN, A. D., and D. J. BEGUN, 2005. Patterns of polymorphism and divergence from non-coding sequences of *Drosophila melanogaster* and *D. simulans*: Evidence for non-equilibrium processes. *Mol. Biol. Evol*. **22:**51–62.

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

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.

Kimura, M., 1968. Evolutionary rate at the molecular level. *Nature* **217:**624–626.

Kingman, J. F. C., 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, A. J. Berry, J. McCarter, J. Wakeley, and J. Hey, 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156:**1913–1931.

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

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.

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.

Ludwig, M. Z., and M. Kreitman, 1995. Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*. *Mol. Biol. Evol*. **12:**1002–1011.

Ludwig, M. Z., N. H. Patel, and M. Kreitman, 1998. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* **125:**949–958.

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

Marais, G., D. Mouchiroud, and L. Duret, 2003. Neutral effects of recombination on base composition in Drosophila. *Genet. Res.* **81:**79–87.

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

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

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

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

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly, 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304:**581–584.

Meiklejohn, C. D., J. Parsch, J. M. Ranz, and D. L. Hartl, 2003. Rapid evolution of male-biased gene expression in drosophila. *Proc. Natl. Acad. Sci. USA* **100:**9894–9899.

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

Morton, B. R., I. V. Bi, M. D. McMullen, and B. S. Gaut, 2005. Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* **doi:**10.1534/genetics.105.049916.

Mount, S. M., C. Burks, G. Hertz, G. D. Stormo, O. White, and C. Fields, 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acid Res*. **20:**4255–4262.

Mousset, S., L. Brazier, M.-L. Cariou, F. Chartois, F. Depaulis and, M. Veuille, 2003. Evidence of a high rate of selective sweeps in African *Drosophila melanogaster*.

Genetics **163:**599–609.

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

Nagylaky, T., 1983. Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. USA* **80:**6278–6281.

Nielsen, R., and Z. Yang, 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:**929–936.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A.G. Clark, and C. Bustamante, 2005. Genomic scans for selective sweeps using SNP data. *Genome Res*. **15:**1566–1575.

Orengo, D. J., 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.

Parsch, J., 2004. Selective constraints on intron evolution in Drosophila. *Genetics* **165:**1843–1851.

Payseur, B. A., A. D. Cutter, and M. W. Nachman, 2002. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol*. **19:**1143–1153.

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

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

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.

Ptak, S. E., and D. A. Petrov, 2002. How intron splicing affects the deletion and insertion profile in *Drosophila melanogaster*. *Genetics* **162:**1233–1244.

Ptak, S. E., D. A. Hinds, K. Köhler, B. Nickel, N. Patil, D. G. Ballinger, M. Przeworski,

K. A. Frazer, and S. Pääbo, 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat. Genet*. **37:**429–434.

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

Quandt, K., K. Frech, H. Karas, E. Wingender, and T. Werner, 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23:**4878–4884.

Rodriguez-Trelles, F., R. Tarrio, and F. J. Ayala, 2000. Fluctuation mutation bias and the evolution of base composition in Drosophila. *J. Mol. Evol.* **50:**1–10.

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

Schaeffer, S. W., 2002. Molecular population genetics of sequence length diversity in the *Adh* region of *Drosophila pseudoobscura*. *Genet. Res*. **80:**163–175.

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.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar, and T. Mitchell-Olds, 2005. A multi-locus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169:**1601–1615.

Schöfl, G., F. Catania, V. Nolte, and C. Schlötterer, 2005. African sequence variation accounts for most of the sequence polymorphism in non-African *Drosophila melanogaster*. *Genetics* **170:**1701–1709.

Schübeler, D., D. Scalzo, C. Kooperberg, B. van Steensel, J. Delrow, and M. Groudine, 2002. Genome-wide DNA replication profile for *Drosophila melanogaster*: a link

between transcription and replication timing. *Nat. Genet*. **32:**438–42.

Sekelsky, J. J., M. H. Brodsky, and K. C. Burtis, 2000. DNA repair in *Drosophila*: insights from the *Drosophila* genome sequence. *J. Cell. Biol*. **150:**F31–36.

Shabalina, S. A., and A. Kondrashov, 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* **74:**23–30.

Sharp, P. A., 1994. Split genes and RNA splicing. *Cell* **77:**805–815.

Singh, N. D., J. C. Davis, and D. A. Petrov, 2005a. Codon bias and non-coding GC content correlate negatively with recombination rate on the Drosophila X chromosome. *J. Mol. Evol.* **61:**315–324.

Singh, N. D., J. C. Davis, and D. A. Petrov, 2005b. X-linked genes evolve higher codon bias in Drosophila and Caenorhabditis. *Genetics* **171:**145–155.

Smith, N. G., and A. Eyre-Walker, 2002. Adaptive protein evolution in Drosophila. *Nature* **415:**1022–1024.

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

Stephan, W., V. S. Rodriguez, B. Zhou, and J. Parsch, 1994. Molecular evolution of the metallothionein gene *Mtn* in the *melanogaster* species group: results from *Drosophila ananassae*. *Genetics* **138:**135–143.

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

Stephan, W., L. Xing, D. A. Kirby, and J. M. Braverman, 1998. A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. *Proc. Natl. Acad. Sci. USA* **95:**5649–5654.

Storey, J. D., 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. B* **64:**479–498.

Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105:**437–460.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:**585–595.

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

Vasemägi, A, and C. R. Primmer, 2005. Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Mol. Ecol*. **14:**3623–3642

Vilella, A. J., A. Blanco-Garcia, S. Hutter, and J. Rozas, 2005. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics **21:**2791–2793.

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.

Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer, 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*. **28:**316–319.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley, M. D. McMullen, and B. S. Gaut, 2005. The effects of artificial selection on the maize genome. *Science* **308:**1310–4.

Yu, J., Z. Yang, M. Kibukawa, M. Paddock, D. A. Passey, and C. K. Wong, 2002. Minimal introns are not “junk”. *Genome Res.* **12:**1185–1189.

Appendix A. The coalescent

Coalescent theory is an extremely useful tool employed in population genetics (Kingman 1982; Hudson 1990). The coalescent traces the lineage of a sample of chromosomes (or parts of them) backward in time until their common ancestor, thus describing what is known as a coalescent tree. Obviously, we cannot know the exact tree topology and the timing of each coalescent event (*i.e.*, in which generation the two alleles find their common ancestor), but we have a detailed statistical theory to describe the probability of their occurrence. Typically, the Wright-Fisher model is considered, *i.e.*, the population of interest is panmictic, has constant size and experiences no migration (also, no recombination and selection are assumed). A useful property of the coalescent tree is that time is measured in units of $4N_0$, where N_0 is the present population size (4 is for autosomes in a diploid species). Therefore, it is easy to incorporate population size changes during a coalescent history by changing the absolute time scale of the tree (Figure A1).

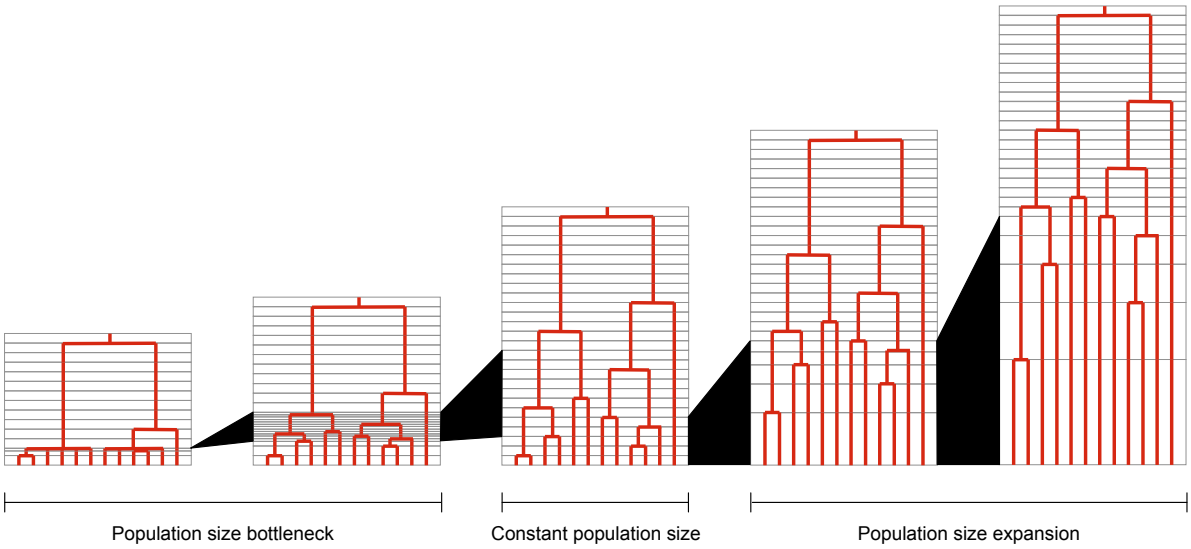


Figure A1. Coalescent trees of a sample of 12 chromosomes. Time goes from present (bottom) to past (top) and is expressed, in units of $4N_0$: horizontal bars delimit fractions of these units. Under the standard neutral model, population size is constant. Changes in population size can be exemplified by "stretching" or "squeezing" time units, as in the cases of population expansion (right panel) and bottlenecks (left panel), respectively. Note that the lengths of internal branches (those that connect nodes) and external branches vary considerably depending on the demographic history.

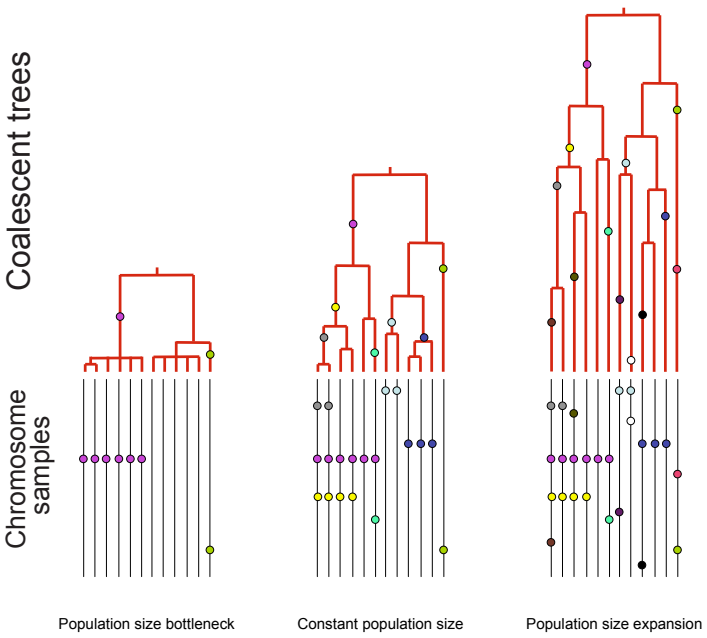


Figure A2. The number of mutations and their frequency distribution in the sample depend on the mutation rate and on the length of the branches of the coalescent tree. In case of population expansion (right tree), terminal branches are much longer than under neutrality (central tree). This results in an excess of low frequency mutations, *i.e.*, more mutations present in only one chromosome than expected. A bottleneck, on the other hand, leaves few chances for post-bottleneck mutations to accumulate (left tree), and in many cases this corresponds to more common (and, in some cases rare) alleles than expected if the population were constant.

A coalescent tree tells us how and when our sample finds its common ancestors, but having a common ancestor does not mean being equal: in fact, a coalescent tree does not tell anything about the differences among the chromosomes (*i.e.*, polymorphism). To integrate the neutral mutation process into the genealogical process, mutations are introduced in the coalescent tree only later, following a Poisson distribution (the infinite-site model is assumed, where a site can be hit by a mutation only once – no back or recurrent mutation). That is, each branch of the tree can host a certain number of mutations, with a probability that depends on its length, which is proportional to time, and the mutation rate (Figure A2).

Because it is so easy to generate coalescent trees, they can be efficiently used to evaluate whether our sample of chromosomes conforms to the standard or to a defined neutral model (*i.e.*, expansion, bottleneck…). The usual approach makes use of computer-generated coalescent trees (with subsequent mutations) that simulate our sample of chromosomes under the neutral model. This will result in virtually infinitely many different random samples, from which we can calculate the probability distribution of the desired test statistic. The empirical value is then compared to this distribution: if it lies outside a defined confidence interval, we can infer that our sample does not conform to the employed neutral model, *i.e.*, it rejects some of its assumptions.

Appendix B. Nucleotide diversity estimates and test statistics

Tables B1 and B2. Nucleotide diversity estimates and basic test statistics for the African and the European populations.

The sequences have been deposited in the EMBL database (http://www.ebi.ac.uk) with accession numbers AJ568984 to AJ571588 and AM000058 to AM003900.

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

- The study where the locus has been used: a = CHAPTER 1.1.; b = CHAPTER 1.2.; c = CHAPTER 3.1.; d = CHAPTER 3.2.;
- Absolute position is in base pairs, from the telomere (based on Flybase, Release 4.2, http://flybase.org);
- Type indicates if the fragment is located in an intergenic region (IG) or in an intron (In); in some instances (cd), the most recent genome annotation revealed that loci previously given as non-coding overlap with putative coding regions or transposable elements (entirely or for a considerable fraction). They have been consequently discarded in the successive analyses (see text);
- The cytological position;
- r is the recombination rate expressed in recombination events per site per generation $\times 10^{-8}$;
- n is the number of lines sequenced;
- L is the number of sites studied (excluding insertions and deletions polymorphism);
- k is the number of segregating sites;
- π is the nucleotide diversity (TAJIMA 1983);
- θ is the WATTERSON (1975) estimate of nucleotide diversity;
- *Ds* is the number of fixed differences between *D. melanogaster* and *D. simulans*;
- *Div* is the divergence between *D. melanogaster* and *D. simulans*;
- Tajima's *D* test statistic (TAJIMA 1989);
- Kelly's Z_{nS} statistic (KELLY 1997).

The tables follow in the next pages.

Table B1. Nucleotide diversity estimates and test statistics for the African population.

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
419	b, c, d	1670429	IG	2B10	0.154	12	606	5	0.0014	0.0027	51	0.0851	–1.655	0.207
12	b, d	1972642	IG	2D2	0.486	12	379	4	0.0026	0.0035	n.a.	n.a.	–0.813	0.192
10	a, b, c, d	2008922	In	2D5	0.486	12	346	10	0.0115	0.0096	54	0.0583	0.745	0.239
9	a	2038743	cd	2E1	0.585	12	323	2	0.0024	0.0021	11	0.0356	0.554	n.a.
17	a, b, c	2055186	In	2E2	0.585	12	781	15	0.0056	0.0064	37	0.0756	–0.453	0.147
6	a	2098059	cd	2F1	0.811	12	402	6	0.0036	0.0049	33	0.0863	–1.022	n.a.
1	a, b, c, d	2113655	In	2F2	0.811	12	380	15	0.0130	0.0131	10	0.0115	–0.010	0.192
15	a	2119374	cd	2F2	0.811	12	462	2	0.0010	0.0014	13	0.0287	–0.850	n.a.
22	a, b, c, d	2239760	In	3A2	1.291	12	618	11	0.0042	0.0059	26	0.0258	–1.096	0.054
25	b, c, d	2247266	IG	3A2	1.291	12	595	13	0.0041	0.0072	33	0.0354	–1.677	0.059
26	a, b, c, d	2250516	IG	3A2	1.291	12	570	18	0.0064	0.0105	28	0.0312	–1.567	0.069
32	b, c, d	2295679	IG	3A3	1.291	12	626	12	0.0054	0.0063	46	0.0988	–0.574	0.092
38	b, d	2378109	In	3A4	1.291	10	394	16	0.0135	0.0144	43	0.0740	–0.239	0.212
18	a, b, c, d	2555495	In	3B2	1.587	12	502	13	0.0073	0.0086	31	0.0416	–0.574	0.116
4	a	2562179	cd	3B2	1.587	12	359	8	0.0079	0.0074	13	0.0417	0.278	n.a.
5	a, b, c, d	2593830	In	3B4	1.587	12	245	14	0.0186	0.0189	32	0.0875	–0.063	0.110
45	b, d	2844745	IG	3C5	2.450	11	480	14	0.0083	0.0100	42	0.0865	–0.685	0.172
46	b, d	2885298	IG	3C5	2.450	12	586	27	0.0155	0.0153	21	0.0671	0.076	0.111
55	a, b, d	3338479	IG	3D4	2.738	12	661	32	0.0137	0.0160	37	0.0481	–0.607	0.131
54	a, b, d	3341441	IG	3D4	2.738	12	418	33	0.0209	0.0261	37	0.0927	–0.831	0.108
57	a, b, d	3436268	IG	3D6	2.738	11	547	12	0.0068	0.0075	47	0.0664	–0.339	0.080
60	a, b, d	3474557	IG	3E1	2.983	12	615	29	0.0155	0.0156	20	0.0183	–0.031	0.109
56	a, b, c, d	3629942	In	3F1	3.290	12	325	5	0.0056	0.0051	49	0.0934	0.357	0.315
76	a, b, c, d	3690386	In	3F3	3.290	12	538	33	0.0161	0.0203	44	0.0569	–0.871	0.129
77	b, d	3717795	IG	3F4	3.290	12	556	29	0.0145	0.0173	28	0.0703	–0.669	0.157
78	a, b, c, d	3764669	IG	3F7	3.290	12	612	23	0.0102	0.0124	25	0.0808	–0.734	0.105
80	b	3876608	IG	4A2	3.549	12	568	32	0.0196	0.0187	46	0.0544	0.211	0.115
81	a, b, c, d	3916988	In	4A4	3.549	12	561	19	0.0116	0.0112	14	0.0149	0.142	0.134
462	b, d	3953777	IG	4A5	3.549	12	668	7	0.0028	0.0035	61	0.0703	–0.650	0.112
84	a, b, c, d	4054401	IG	4B3	3.883	11	596	21	0.0100	0.0120	29	0.0519	–0.709	0.100
85	a, b, c, d	4106003	IG	4B4	3.883	12	510	18	0.0103	0.0117	47	0.0718	–0.461	0.129
66	b, d	4296277	In	4C4	4.369	12	352	16	0.0145	0.0151	41	0.0419	–0.155	0.089
67	b, d	4548130	In	4C14	4.369	12	633	24	0.0102	0.0126	24	0.0265	–0.751	0.125
90	b, d	4935383	IG	4E2	4.611	12	419	31	0.0231	0.0245	23	0.0216	–0.235	0.113

(Continues...)

Table B1. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
91	b, d	4991826	IG	4F1	4.723	10	471	23	0.0134	0.0173	42	0.0730	–0.952	0.138
93	b, d	5073706	In	4F3	4.723	12	391	9	0.0045	0.0076	19	0.0254	–1.535	0.054
94	b, d	5126621	IG	4F4	4.723	10	505	22	0.0117	0.0154	24	0.0404	–1.031	0.116
95	b, d	5171847	IG	4F5	4.723	11	560	19	0.0098	0.0116	40	0.0944	–0.621	0.104
106	a, b, c, d	5478703	In	5A8	4.707	12	404	17	0.0108	0.0139	21	0.0355	–0.904	0.097
72	a, b, d	5518785	IG	5A11	4.707	12	379	37	0.0323	0.0323	21	0.0699	–0.009	0.373
73	b	5592462	IG	5B6	4.634	12	574	10	0.0047	0.0058	72	0.0841	–0.678	0.103
109	b, d	5767488	IG	5C7	4.492	11	582	13	0.0064	0.0076	48	0.0628	–0.654	0.109
114	a, b, c	6605700	cd	6C12	2.997	12	300	3	0.0029	0.0033	16	0.0217	–0.340	0.038
115	a, b, c	6651455	In	6D3	2.710	12	398	10	0.0068	0.0083	60	0.0837	–0.678	0.116
116	a, b, c, d	6687195	In	6D4	2.710	12	512	34	0.0211	0.0220	33	0.0549	–0.163	0.171
117	a, b, c, d	6741220	IG	6E2	2.447	9	553	33	0.0219	0.0220	35	0.0537	–0.012	0.155
118	a, b, c, d	6790458	IG	6E3	2.447	12	540	13	0.0059	0.0080	24	0.0291	–1.010	0.099
119	a, b, c, d	6838119	In	6E4	2.447	12	297	26	0.0239	0.0290	47	0.1361	–0.716	0.109
120	a, b, c, d	6916170	In	6F2	2.178	12	469	32	0.0173	0.0226	25	0.0329	–0.969	0.134
122	a, b, c, d	7006861	IG	7A1	1.926	12	576	6	0.0017	0.0034	45	0.0506	–1.716	0.074
502	b, d	7033935	IG	7A2	1.926	12	537	24	0.0131	0.0148	27	0.0585	–0.471	0.092
124	a, b, c	7083799	cd	7A4	1.926	12	762	9	0.0035	0.0039	62	0.0888	–0.364	0.106
125	a, b, c, d	7134523	In	7B1	1.601	12	240	7	0.0092	0.0097	15	0.0211	–0.185	0.110
126	b, d	7185769	In	7B2	1.601	12	585	39	0.0190	0.0221	18	0.0234	–0.591	0.105
130	a, b, c, d	7362175	IG	7B3	1.601	12	553	21	0.0101	0.0126	40	0.0654	–0.792	0.101
530	b, d	7389987	IG	7B3	1.601	12	505	26	0.0159	0.0170	50	0.0684	–0.269	0.152
133	b	7512346	In	7B6	1.601	11	621	15	0.0069	0.0082	39	0.0763	–0.683	0.118
136	a, b, c, d	7726936	In	7C1	1.461	12	371	27	0.0176	0.0241	37	0.0485	–1.115	0.129
137	a, b, c, d	7759037	In	7C2	1.461	12	453	14	0.0074	0.0102	43	0.1024	–1.083	0.103
138	a, b, c, d	7807461	In	7D1	1.486	12	338	9	0.0085	0.0088	28	0.0384	–0.126	0.081
139	a, b, c	7868565	cd	7D2	1.486	12	347	14	0.0109	0.0134	31	0.0319	–0.714	0.130
143	b, d	8118991	In	7E1	1.680	12	519	19	0.0103	0.0121	45	0.0664	–0.601	0.112
150	a, b, d	8443939	In	7F7	1.930	12	305	15	0.0140	0.0163	52	0.0500	–0.561	0.111
153	a, b, c, d	8613511	In	8A5	2.150	12	475	25	0.0152	0.0174	33	0.0779	–0.527	0.128
157	a, b, d	8815156	IG	8C1	3.009	12	310	12	0.0114	0.0128	10	0.0177	–0.413	0.151
160	b, d	8951179	In	8C9	3.009	12	199	13	0.0135	0.0216	n.a.	n.a.	–1.459	0.487
163	a, b, c, d	9093117	In	8D2	3.638	12	630	24	0.0085	0.0126	65	0.0905	–1.342	0.092
165	a, b, c	9202487	In	8D9	3.638	12	277	6	0.0049	0.0072	40	0.0473	–1.084	0.176
166	a, b, c	9237654	IG	8D12	3.638	12	606	11	0.0042	0.0060	27	0.0219	–1.127	0.129

(Continues...)

Table B1. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
167	b, d	9281023	IG	8E1	4.175	12	607	12	0.0050	0.0065	34	0.0891	–0.881	0.110
169	b, d	9420312	In	8E10	4.175	12	308	2	0.0015	0.0022	45	0.0658	–0.758	0.018
170	b, d	9461583	IG	8F2	4.508	10	517	35	0.0231	0.0239	31	0.0341	–0.148	0.128
173	a, b, d	9639363	IG	9A2	3.536	12	498	8	0.0032	0.0053	7	0.0124	–1.429	0.065
446	b, d	9710908	In	9A2	3.536	12	262	16	0.0203	0.0202	57	0.0818	0.015	0.147
175	a, b, c, d	9774954	In	9A3	3.536	12	603	38	0.0185	0.0209	37	0.0442	–0.471	0.099
177	a, b, c, d	9849981	In	9A4	3.536	12	409	20	0.0144	0.0162	29	0.0582	–0.443	0.116
178	b, d	9890361	IG	9A4	3.536	12	493	34	0.0180	0.0228	32	0.0772	–0.874	0.109
179	b, d	9938187	IG	9B1	2.813	12	545	26	0.0152	0.0158	49	0.0590	–0.168	0.144
182	b, d	10099354	IG	9B5	2.813	12	458	16	0.0084	0.0116	n.a.	n.a.	–1.070	0.217
464	b, d	10104098	IG	9B5	2.813	12	548	29	0.0170	0.0175	32	0.0984	–0.122	0.135
465	b, d	10141953	In	9B6	2.813	12	449	20	0.0164	0.0148	35	0.0546	0.460	0.148
184	a, b, c	10173196	In	9B7	2.813	12	424	22	0.0149	0.0172	17	0.0362	–0.547	0.106
186	a, b, c, d	10272739	In	9C2	2.620	12	497	21	0.0094	0.0140	51	0.0772	–1.321	0.086
187	b, d	10300966	IG	9C4	2.620	12	522	14	0.0077	0.0089	62	0.0663	–0.535	0.110
188	b, d	10324188	IG	9D1	2.509	12	491	4	0.0016	0.0027	64	0.0993	–1.248	0.013
189	a, b, d	10383742	In	9D3	2.509	12	541	17	0.0081	0.0104	55	0.0631	–0.893	0.106
190	b	10402781	In	9D3	2.509	11	525	24	0.0131	0.0156	n.a.	n.a.	–0.685	0.128
191	a, b	10439282	cd	9D3	2.509	12	432	4	0.0034	0.0031	21	0.0330	0.347	0.240
470	b, d	10465155	IG	9D4	2.509	11	606	6	0.0018	0.0034	76	0.0997	–1.669	0.142
192	b	10490302	IG	9D4	2.509	10	418	17	0.0128	0.0144	n.a.	n.a.	–0.475	0.206
472	b, d	10554792	IG	9E1	2.391	12	553	15	0.0071	0.0090	36	0.0829	–0.813	0.140
194	a, b, d	10585619	In	9E1	2.391	12	578	17	0.0078	0.0097	43	0.0472	–0.776	0.139
195	b, d	10596143	In	9E2	2.391	12	508	34	0.0199	0.0222	46	0.0524	–0.426	0.197
473	b, d	10625772	IG	9E10	2.391	12	530	11	0.0078	0.0069	19	0.0210	0.530	0.170
196	a, b, c	10641809	cd	9F2	2.402	12	596	5	0.0020	0.0028	34	0.0543	–0.923	0.037
197	a, b, d	10680736	In	9F4	2.402	12	547	7	0.0030	0.0042	47	0.0562	–1.021	0.042
198	b, d	10725814	In	9F7	2.402	12	662	24	0.0097	0.0120	n.a.	n.a.	–0.782	0.104
475	b	10746693	IG	9F8	2.402	12	619	9	0.0035	0.0048	32	0.0285	–1.027	0.309
743	b, d	10818470	IG	9F13	2.402	11	298	13	0.0084	0.0149	18	0.0362	–1.731	0.113
201	a, b, c, d	10874510	IG	10A2	2.545	12	677	13	0.0051	0.0064	47	0.0859	–0.737	0.133
477	b, d	10887045	IG	10A3	2.545	12	647	32	0.0140	0.0164	39	0.1034	–0.601	0.133
480	b	10935038	cd	10A4	2.545	12	626	16	0.0050	0.0085	41	0.1304	–1.624	0.091
203	a, b, c	10949796	cd	10A4	2.545	12	573	24	0.0131	0.0139	55	0.0622	–0.222	0.106
532	b, d	10968776	IG	10A6	2.545	12	457	4	0.0031	0.0029	27	0.0358	0.166	0.515

(Continues...)

Table B1. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
204	a, b, c, d	11011770	IG	10A9	2.545	12	533	25	0.0127	0.0155	64	0.0923	–0.751	0.095
205	a, b, c, d	11070436	In	10B1	3.000	12	645	24	0.0122	0.0123	47	0.0538	–0.051	0.123
483	b, d	11092779	IG	10B1	3.000	12	656	37	0.0158	0.0187	29	0.0287	–0.639	0.106
206	b, d	11110942	IG	10B2	3.000	12	527	31	0.0150	0.0195	26	0.0903	–0.963	0.103
207	b, d	11140441	IG	10B2	3.000	12	490	35	0.0229	0.0237	42	0.0496	–0.125	0.112
208	b, d	11168267	IG	10B2	3.000	12	500	10	0.0062	0.0066	31	0.0413	–0.250	0.101
209	a, b, c	11207707	In	10B5	3.000	12	483	42	0.0283	0.0288	63	0.1081	–0.073	0.141
210	b, d	11244045	In	10B8	3.000	12	661	21	0.0082	0.0105	n.a.	n.a.	–0.898	0.100
211	b	11281619	cd	10B12	3.000	12	585	28	0.0147	0.0158	26	0.0445	–0.289	0.152
212	a, b, d	11325443	In	10B15	3.000	12	688	12	0.0044	0.0058	70	0.1035	–0.925	0.072
213	b, d	11361859	In	10C2	3.282	12	576	11	0.0058	0.0063	30	0.0331	–0.338	0.138
214	a, b, c, d	11414513	In	10C7	3.282	12	588	17	0.0059	0.0096	37	0.0480	–1.535	0.210
215	a, b, d	11463155	In	10D2	3.440	11	568	17	0.0096	0.0102	75	0.0878	–0.246	0.170
216	a, b, c, d	11507573	In	10D4	3.440	12	603	42	0.0198	0.0231	44	0.0561	–0.594	0.117
217	a, b, c, d	11533051	IG	10D5	3.440	12	537	5	0.0021	0.0031	27	0.0591	–1.106	0.061
218	b, d	11550138	In	10D6	3.440	12	341	10	0.0077	0.0097	45	0.0591	–0.781	0.122
219	b, d	11609468	In	10E2	3.588	11	577	12	0.0064	0.0071	n.a.	n.a.	–0.409	0.192
220	b, d	11629125	In	10E3	3.588	12	411	2	0.0014	0.0016	27	0.0320	–0.341	0.030
221	a, b, c, d	11638396	In	10E4	3.588	12	380	18	0.0159	0.0157	34	0.0490	0.047	0.171
222	b, d	11681093	IG	10F1	3.813	10	504	28	0.0181	0.0196	n.a.	n.a.	–0.336	0.157
488	b, d	11708720	IG	10F2	3.813	10	593	31	0.0172	0.0185	13	0.0265	–0.303	0.123
224	a, b, c, d	11783192	In	10F9	3.813	12	599	27	0.0115	0.0149	75	0.1026	–0.955	0.082
660	b, d	11837057	IG	11A1	4.138	10	368	5	0.0046	0.0048	n.a.	n.a.	–0.159	0.085
228	b, d	11912537	IG	11A3	4.138	11	408	8	0.0039	0.0067	n.a.	n.a.	–1.551	0.302
492	b, d	11939099	IG	11A4	4.138	12	649	24	0.0096	0.0122	69	0.1110	–0.875	0.127
229	a, b, c, d	11956720	IG	11A4	4.138	12	422	14	0.0091	0.0110	31	0.0521	–0.675	0.115
493	b, d	12015826	IG	11A6	4.138	12	612	26	0.0118	0.0141	58	0.0872	–0.673	0.143
231	a, b, c, d	12030561	In	11A6	4.138	11	520	43	0.0269	0.0282	49	0.0736	–0.200	0.122
232	b, d	12052930	IG	11A6	4.138	12	546	15	0.0074	0.0091	11	0.0235	–0.741	0.152
233	b, d	12109811	In	11A7	4.138	12	441	32	0.0223	0.0240	54	0.0607	–0.305	0.167
235	b, d	12147557	In	11A8	4.138	12	507	10	0.0045	0.0065	12	0.0139	–1.159	0.107
237	b, d	12201432	IG	11A9	4.138	12	497	46	0.0271	0.0306	n.a.	n.a.	–0.492	0.120
447	b, d	12231484	IG	11A9	4.138	12	579	19	0.0068	0.0109	44	0.0616	–1.490	0.112
239	b	12269282	In	11A10	4.138	11	310	40	0.0482	0.0441	28	0.0622	0.405	0.247
241	a, b, d	12335618	In	11A12	4.138	12	568	6	0.0029	0.0035	76	0.0964	–0.609	0.138

(Continues...)

Table B1. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
242	b, d	12377504	IG	11B1	4.436	12	467	33	0.0217	0.0234	52	0.0848	–0.301	0.161
721	b, d	12455184	IG	11B4	4.436	12	335	28	0.0279	0.0277	15	0.0367	0.027	0.351
245	b, d	12507177	In	11B9	4.436	12	432	20	0.0143	0.0153	63	0.0657	–0.268	0.142
246	b, d	12557700	In	11B14	4.436	12	448	18	0.0123	0.0133	n.a.	n.a.	–0.298	0.139
248	a, b, c, d	12617534	IG	11C2	4.512	11	656	23	0.0109	0.0120	12	0.0160	–0.386	0.144
249	a, b, c, d	12650222	IG	11C3	4.512	12	549	22	0.0089	0.0133	37	0.0870	–1.348	0.106
250	a, b, c, d	12701253	IG	11D1	4.689	12	593	26	0.0132	0.0145	72	0.1084	–0.363	0.231
251	a, b, c	12744897	cd	11D1	4.689	12	438	29	0.0228	0.0219	51	0.0685	0.158	0.235
252	b, d	12777694	In	11D4	4.689	12	428	26	0.0177	0.0201	17	0.0189	–0.500	0.162
253	b	12820460	IG	11D6	4.689	11	467	20	0.0125	0.0146	21	0.0311	–0.601	0.157
254	a, b, c	12859984	IG	11D8	4.689	12	399	10	0.0065	0.0083	55	0.0590	–0.816	0.163
258	b, d	12955002	In	11E1	4.812	12	417	29	0.0228	0.0230	48	0.0981	–0.044	0.144
259	b, d	13007119	In	11E2	4.812	12	289	18	0.0153	0.0206	47	0.0503	–1.039	0.171
260	b, d	13053066	In	11E4	4.812	10	554	17	0.0087	0.0108	31	0.0943	–0.818	0.153
272	a, b, d	13096432	IG	11E8	4.812	12	506	20	0.0149	0.0131	64	0.0942	0.543	0.238
273	a, b, c, d	13102200	In	11E8	4.812	12	420	26	0.0214	0.0205	49	0.0591	0.179	0.142
722	b, d	13164524	IG	11E11	4.812	12	305	26	0.0260	0.0282	38	0.0406	–0.326	0.135
276	a, b, c, d	13232942	In	11F1	4.877	12	326	10	0.0071	0.0102	12	0.0148	–1.142	0.132
277	b, d	13268022	IG	11F6	4.877	12	599	37	0.0201	0.0205	39	0.0544	–0.081	0.115
278	a, b, d	13318463	In	12A1	4.974	12	610	33	0.0161	0.0179	20	0.0318	–0.428	0.119
279	a, b, c	13351547	In	12A2	4.974	12	658	27	0.0133	0.0136	9	0.0188	–0.084	0.137
280	b, d	13385015	IG	12A4	4.974	12	294	18	0.0175	0.0203	36	0.0714	–0.542	0.148
450	b, d	13397127	IG	12A4	4.974	12	663	35	0.0126	0.0175	22	0.0299	–1.165	0.114
311	b, d	13397557	IG	12A4	4.974	12	215	21	0.0209	0.0323	48	0.0838	–1.435	0.117
312	a, b, c, d	13428303	In	12A7	4.974	12	632	15	0.0069	0.0079	9	0.0202	–0.501	0.110
313	b, d	13468507	In	12A9	4.974	12	456	9	0.0049	0.0065	17	0.0162	–0.915	0.217
314	a, b	13505120	cd	12B2	5.019	12	565	21	0.0121	0.0123	14	0.0611	–0.079	0.130
318	b, d	13647962	In	12C5	5.026	10	325	10	0.0120	0.0109	39	0.0453	0.426	0.173
319	b, d	13670436	In	12C6	5.026	12	489	27	0.0143	0.0183	24	0.0363	–0.885	0.110
320	b, d	13710254	In	12C7	5.026	11	433	15	0.0097	0.0118	91	0.1545	–0.712	0.150
321	b, d	13742548	IG	12D1	5.023	12	559	9	0.0043	0.0053	39	0.0528	–0.708	0.268
323	b, d	13831025	IG	12D2	5.023	12	372	19	0.0092	0.0169	54	0.0831	–1.818	0.286
325	b, d	13895682	IG	12D4	5.023	12	528	8	0.0043	0.0050	16	0.0178	–0.495	0.251
326	a, b, c, d	13935313	IG	12E1	4.979	12	605	18	0.0078	0.0099	26	0.1203	–0.846	0.076
342	b	14017479	IG	12E2	4.979	11	527	35	0.0215	0.0227	n.a.	n.a.	–0.216	0.118

(Continues...)

Table B1. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
344	b, d	14066104	cd	12E5	4.979	11	510	29	0.0180	0.0194	22	0.0763	–0.315	0.141
346	b, d	14129542	IG	12E8	4.979	10	492	10	0.0070	0.0072	54	0.1484	–0.077	0.154
348	a, b, d	14195150	IG	12E8	4.979	12	571	25	0.0120	0.0145	35	0.0448	–0.707	0.124
745	b, d	14219265	IG	12E9	4.979	11	371	2	0.0010	0.0018	39	0.0748	–1.269	0.010
350	b, d	14301478	IG	12E10	4.979	12	452	13	0.0072	0.0095	42	0.0678	–0.928	0.130
367	a, b, c, d	14324184	IG	12E10	4.979	12	582	20	0.0108	0.0114	45	0.1185	–0.222	0.097
368	b, d	14357426	IG	12F1	4.934	12	505	26	0.0129	0.0170	38	0.0745	–0.997	0.130
369	b, d	14394544	IG	12F1	4.934	12	675	9	0.0035	0.0044	31	0.0378	–0.802	0.079
370	a, b, c, d	14414966	In	12F1	4.934	12	507	33	0.0201	0.0216	53	0.0750	–0.278	0.122
371	b, d	14446806	In	12F1	4.934	11	532	24	0.0144	0.0154	52	0.0622	–0.266	0.127
373	b, d	14514836	In	12F2	4.934	11	531	9	0.0042	0.0058	43	0.0587	–1.059	0.144
374	a, b, c, d	14526351	In	12F2	4.934	12	544	15	0.0071	0.0091	56	0.1068	–0.861	0.131
375	a, b, d	14561277	In	12F3	4.934	12	631	34	0.0176	0.0178	41	0.0555	–0.068	0.121
376	b	14593606	In	12F4	4.934	10	259	16	0.0238	0.0218	17	0.0279	0.371	0.150
378	b, d	14626366	In	12F4	4.934	12	518	22	0.0161	0.0141	67	0.0808	0.600	0.165
379	a, b, c, d	14664445	In	12F5	4.934	11	568	40	0.0202	0.0240	58	0.0893	–0.687	0.109
380	b, d	14703175	In	12F6	4.934	12	504	14	0.0082	0.0092	n.a.	n.a.	–0.408	0.101
381	a, b, c	14765244	In	13A1	4.883	11	444	35	0.0278	0.0269	52	0.0822	0.142	0.129
382	b, d	14826789	In	13A5	4.883	12	487	20	0.0085	0.0136	43	0.0668	–1.521	0.077
384	a, b, c, d	14920236	In	13A10	4.883	12	502	19	0.0101	0.0125	70	0.0955	–0.775	0.215
385	a, b, c, d	14933236	In	13A11	4.883	12	525	23	0.0129	0.0145	n.a.	n.a.	–0.459	0.133
386	b, d	14948045	In	13A12	4.883	12	426	18	0.0122	0.0140	52	0.0974	–0.522	0.143
387	b, d	14965085	IG	13B1	4.718	12	577	25	0.0137	0.0143	43	0.0737	–0.175	0.116
388	b, d	15005477	IG	13B1	4.718	10	610	11	0.0045	0.0064	55	0.1192	–1.202	0.088
389	b	15057214	IG	13B3	4.718	12	518	8	0.0038	0.0051	n.a.	n.a.	–0.910	0.104
391	b, d	15087602	IG	13B4	4.718	12	562	12	0.0050	0.0071	n.a.	n.a.	–1.101	0.093
390	b, d	15117024	IG	13B4	4.718	12	537	20	0.0121	0.0123	n.a.	n.a.	–0.084	0.093
392	b	15148546	In	13B6	4.718	12	444	15	0.0073	0.0112	18	0.0364	–1.376	0.147
534	b, d	15152610	IG	13B6	4.718	12	410	2	0.0011	0.0016	52	0.0728	–0.758	0.018
393	a, b, c, d	15169667	In	13B6	4.718	12	559	11	0.0058	0.0065	43	0.0725	–0.433	0.112
394	b, d	15211531	In	13C1	4.624	12	582	33	0.0131	0.0188	38	0.0700	–1.251	0.103
282	a, b, c	15282518	cd	13C3	4.624	12	683	35	0.0153	0.0170	64	0.0825	–0.419	0.124
285	b	15467669	IG	13E3	4.330	12	494	42	0.0249	0.0282	n.a.	n.a.	–0.489	0.144
286	b	15503689	IG	13E4	4.330	10	325	31	0.0235	0.0337	26	0.1008	–1.333	0.280
287	a, b, c, d	15525959	IG	13E7	4.330	12	572	21	0.0104	0.0122	52	0.0879	–0.581	0.114

(Continues...)

Table B1. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
288	a, b, c	15536600	cd	13E8	4.330	12	521	17	0.0103	0.0108	22	0.0449	–0.198	0.168
295	b, d	15544479	IG	13E8	4.330	12	557	11	0.0039	0.0065	19	0.0238	–1.506	0.397
296	b	15599496	In	13E14	4.330	12	587	9	0.0040	0.0051	52	0.0748	–0.764	0.072
297	a, b, c, d	15638469	IG	13F1	4.108	12	630	8	0.0025	0.0042	57	0.0987	–1.429	0.045
298	b	15692843	cd	13F15	4.108	12	502	13	0.0066	0.0086	69	0.0853	–0.901	0.140
299	a, b, c, d	15730456	In	13F18	4.108	12	618	18	0.0076	0.0096	40	0.0386	–0.826	0.091
294	b, d	15746063	In	14A1	3.928	12	595	15	0.0070	0.0083	20	0.0261	–0.645	0.081
301	a, b, c	15798732	cd	14A3	3.928	12	556	26	0.0121	0.0155	51	0.0576	–0.896	0.092
303	a	15893832	cd	14A6	3.928	12	608	11	0.0049	0.0060	34	0.0599	–0.803	n.a.
304	b, d	15906320	In	14A8	3.928	12	501	22	0.0114	0.0145	33	0.0405	–0.884	0.140
725	b, d	15964533	IG	14B1	3.667	12	392	6	0.0032	0.0051	36	0.0613	–1.242	0.103
306	b, d	15995877	IG	14B1	3.667	12	595	51	0.0236	0.0284	41	0.0554	–0.720	0.102
307	b, d	16047623	IG	14B3	3.667	12	513	29	0.0191	0.0187	41	0.0632	0.079	0.164
726	b, d	16096635	IG	14B6	3.667	12	434	10	0.0050	0.0076	33	0.0639	–1.279	0.071
310	b, d	16152786	In	14B9	3.667	12	491	8	0.0030	0.0054	44	0.0640	–1.616	0.148
336	b, d	16168466	In	14B14	3.667	12	600	7	0.0029	0.0039	38	0.0488	–0.905	0.072
334	b, d	16255530	IG	14C4	3.571	12	527	10	0.0059	0.0063	30	0.0371	–0.215	0.119
333	a, b, d	16276272	In	14C6	3.571	12	582	13	0.0077	0.0074	40	0.0510	0.162	0.208
451	b, d	16302764	In	14D1	3.494	12	638	11	0.0058	0.0057	29	0.0353	0.057	0.176
331	a, b	16349660	In	14D4	3.494	12	552	24	0.0127	0.0144	40	0.0455	–0.471	0.163
330	a, b, c	16371094	IG	14E1	3.426	12	567	41	0.0254	0.0239	32	0.0379	0.253	0.109
329	a, b, d	16429444	IG	14F1	3.318	10	600	11	0.0045	0.0065	48	0.0646	–1.248	0.139
328	b, d	16468799	In	14F4	3.318	10	545	8	0.0043	0.0052	29	0.0463	–0.679	0.104
366	a, b, d	16471181	IG	14F4	3.318	11	610	38	0.0260	0.0213	47	0.0519	0.952	0.372
364	a, b, c	16530195	IG	15A1	3.129	12	600	19	0.0091	0.0105	28	0.0592	–0.543	0.111
363	b, d	16550618	IG	15A3	3.129	12	613	25	0.0136	0.0135	54	0.0866	0.020	0.227
359	b, d	16694289	IG	15B1	3.065	12	525	22	0.0124	0.0139	41	0.0928	–0.437	0.083
402	b, d	16783883	IG	15C4	3.008	12	600	19	0.0096	0.0105	47	0.1050	–0.331	0.100
405	b, d	16864287	IG	15D4	2.932	10	627	24	0.0102	0.0135	49	0.0649	–1.050	0.123
406	b, d	16907918	In	15E1	2.867	12	656	23	0.0120	0.0116	21	0.0192	0.148	0.168
407	b, d	16933856	IG	15E3	2.837	12	571	35	0.0163	0.0203	22	0.1037	–0.828	0.126
410	b, d	17027142	In	15F1	2.782	12	600	28	0.0124	0.0155	49	0.0778	–0.821	0.152
411	b, d	17058256	IG	15F4	2.782	12	545	24	0.0105	0.0146	5	0.0081	–1.132	0.232
422	b, d	17088628	IG	15F4	2.782	11	578	28	0.0150	0.0165	49	0.0826	–0.384	0.137
727	b, d	17161040	IG	16A1	2.684	12	450	16	0.0117	0.0118	7	0.0357	–0.019	0.154

(Continues...)

Table B1. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
424	b, d	17229023	IG	16A4	2.684	12	634	16	0.0070	0.0084	35	0.0959	–0.618	0.105
728	b, d	17301040	IG	16A5	2.684	12	270	4	0.0025	0.0049	14	0.0096	–1.574	0.339
426	b, d	17353968	IG	16B1	2.591	12	658	20	0.0091	0.0101	41	0.0587	–0.388	0.098
428	b, d	17384694	IG	16B4	2.591	12	606	20	0.0103	0.0109	24	0.0576	–0.250	0.125
729	b, d	17442422	IG	16B8	2.591	10	443	14	0.0136	0.0112	29	0.0818	0.900	0.251
430	b, d	17492612	IG	16B10	2.591	12	659	15	0.0040	0.0075	44	0.0824	–1.843	0.095
730	b	17541034	IG	16C1	2.527	12	218	8	0.0101	0.0122	n.a.	n.a.	–0.620	0.181
431	b, d	17619496	IG	16D1	2.487	12	509	4	0.0013	0.0026	43	0.1102	–1.574	0.174
432	b, d	17662415	IG	16D4	2.487	11	508	10	0.0053	0.0067	26	0.0849	–0.818	0.079
436	b, d	17980460	IG	16F7	2.436	11	378	16	0.0118	0.0145	56	0.0725	–0.736	0.153
438	b	18064811	IG	17A3	2.424	12	570	23	0.0107	0.0134	40	0.0624	–0.823	0.115
439	b, d	18132877	IG	17A4	2.424	12	546	6	0.0024	0.0036	38	0.0541	–1.163	0.219
440	b, d	18201747	IG	17A7	2.424	12	594	20	0.0094	0.0111	58	0.1072	–0.637	0.227
444	b, d	18579133	IG	17D3	2.467	11	567	26	0.0131	0.0157	47	0.0498	–0.674	0.128

Table B2. Nucleotide diversity estimates and test statistics for the European population.

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
419	b	1670429	IG	2B10	0.154	12	586	1	0.0003	0.0006	25	0.0756	–1.141	n.a.
10	a, b	2008922	In	2D5	0.486	12	348	0	0.0000	0.0000	18	0.0838	n.a.	n.a.
9	a	2038743	cd	2E1	0.585	12	326	0	0.0000	0.0000	11	0.0340	n.a.	n.a.
17	a, b	2055186	In	2E2	0.585	12	773	7	0.0038	0.0030	26	0.0854	0.998	0.353
6	a	2098059	cd	2F1	0.811	12	402	4	0.0045	0.0033	33	0.0863	1.230	n.a.
1	a, b	2113655	In	2F2	0.811	12	381	5	0.0070	0.0043	40	0.0576	2.251	0.739
15	a	2119374	cd	2F2	0.811	12	461	1	0.0012	0.0007	13	0.0293	1.486	n.a.
22	a, b	2239760	In	3A2	1.291	12	630	8	0.0025	0.0042	38	0.0963	–1.572	0.296
25	b	2247266	IG	3A2	1.291	12	588	6	0.0028	0.0034	25	0.0471	–0.673	0.085
26	a, b	2250516	IG	3A2	1.291	12	589	2	0.0018	0.0011	14	0.0262	1.824	0.257
32	b	2295679	IG	3A3	1.291	12	721	4	0.0028	0.0018	21	0.0386	1.793	0.438
33	b	2298889	In	3A3	1.291	11	983	18	0.0067	0.0063	20	0.0315	0.346	0.278
38	b	2378109	In	3A4	1.291	12	432	2	0.0008	0.0015	40	0.0496	–1.451	0.008
18	a, b	2555495	In	3B2	1.587	12	500	9	0.0065	0.0060	41	0.1050	0.398	0.567
4	a	2562179	cd	3B2	1.587	12	359	3	0.0038	0.0028	14	0.0421	1.273	n.a.
5	a, b	2593830	In	3B4	1.587	12	248	0	0.0000	0.0000	16	0.0283	n.a.	n.a.
45	b	2844745	IG	3C5	2.450	12	609	1	0.0007	0.0005	25	0.0465	0.541	n.a.
46	b	2885298	IG	3C5	2.450	12	605	3	0.0020	0.0016	25	0.0801	0.772	0.539
55	a, b	3338479	IG	3D4	2.738	11	660	16	0.0078	0.0083	35	0.0685	–0.244	0.446
54	a, b	3341441	IG	3D4	2.738	12	418	13	0.0085	0.0103	8	0.0297	–0.747	0.476
57	a, b	3436268	IG	3D6	2.738	12	565	7	0.0042	0.0041	26	0.0550	0.052	0.228
60	a, b	3474557	IG	3E1	2.983	12	626	14	0.0087	0.0074	53	0.1036	0.767	0.276
56	a, b	3629942	In	3F1	3.290	12	325	4	0.0028	0.0041	15	0.0491	–1.103	0.515
76	a, b	3690386	In	3F3	3.290	12	540	13	0.0115	0.0080	29	0.0514	1.892	0.411
77	b	3717795	IG	3F4	3.290	10	560	7	0.0025	0.0044	27	0.0773	–1.839	1.000
78	a, b	3764669	IG	3F7	3.290	12	616	9	0.0068	0.0048	42	0.0799	1.657	0.452
80	b	3876608	IG	4A2	3.549	10	618	22	0.0170	0.0126	25	0.0590	1.662	0.503
81	a, b	3916988	In	4A4	3.549	12	568	3	0.0018	0.0017	45	0.0970	0.022	0.050
462	b	3953777	IG	4A5	3.549	12	664	3	0.0008	0.0015	30	0.0584	–1.629	0.339
84	a, b	4054401	IG	4B3	3.883	12	596	13	0.0085	0.0072	42	0.1048	0.744	0.324
85	a, b	4106003	IG	4B4	3.883	12	641	11	0.0029	0.0057	20	0.0394	–2.067	0.820
66	b	4296277	In	4C4	4.369	12	383	6	0.0033	0.0052	26	0.0577	–1.371	0.709
67	b	4548130	In	4C14	4.369	12	654	3	0.0008	0.0015	33	0.0621	–1.629	1.000
90	b	4935383	IG	4E2	4.611	12	426	13	0.0117	0.0101	40	0.1225	0.654	0.161
91	b	4991826	IG	4F1	4.723	12	566	12	0.0035	0.0070	24	0.0515	–2.087	0.835

(Continues...)

Table B2. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
93	b	5073706	In	4F3	4.723	12	407	5	0.0033	0.0041	10	0.0272	–0.684	0.152
94	b	5126621	IG	4F4	4.723	12	593	10	0.0034	0.0056	37	0.0784	–1.573	0.448
95	b	5171847	IG	4F5	4.723	12	706	7	0.0023	0.0033	61	0.1101	–1.176	0.357
106	a, b	5478703	In	5A8	4.707	12	405	13	0.0118	0.0106	15	0.0295	0.475	0.279
72	a, b	5518785	IG	5A11	4.707	12	418	2	0.0008	0.0016	22	0.0709	–1.451	1.000
73	b	5592462	IG	5B6	4.634	12	580	3	0.0014	0.0017	22	0.0442	–0.579	0.364
109	b	5767488	IG	5C7	4.492	12	602	7	0.0047	0.0039	24	0.0900	0.844	0.263
102	b	5886956	In	5D2	4.225	12	534	2	0.0009	0.0012	26	0.0749	–0.850	0.018
114	a, b	6605700	cd	6C12	2.997	12	299	1	0.0006	0.0011	25	0.0654	–1.141	n.a.
115	a, b	6651455	In	6D3	2.710	12	401	1	0.0004	0.0008	45	0.0965	–1.141	n.a.
116	a, b	6687195	In	6D4	2.710	12	548	19	0.0098	0.0115	17	0.0329	–0.656	0.502
117	a, b	6741220	IG	6E2	2.447	10	583	18	0.0150	0.0109	28	0.1325	1.739	0.551
118	a, b	6790458	IG	6E3	2.447	12	554	4	0.0021	0.0024	18	0.0603	–0.419	0.076
119	a, b	6838119	In	6E4	2.447	12	297	5	0.0071	0.0056	10	0.0174	1.003	0.244
120	a, b	6916170	In	6F2	2.178	12	447	16	0.0136	0.0119	19	0.0287	0.646	0.394
122	a, b	7006861	IG	7A1	1.926	11	589	0	0.0000	0.0000	16	0.0711	n.a.	n.a.
502	b	7033935	IG	7A2	1.926	12	544	7	0.0052	0.0043	49	0.0891	0.896	0.566
124	a, b	7083799	cd	7A4	1.926	12	763	4	0.0026	0.0017	51	0.0956	1.793	0.750
125	a, b	7134523	In	7B1	1.601	12	240	0	0.0000	0.0000	33	0.0599	n.a.	n.a.
126	b	7185769	In	7B2	1.601	12	651	9	0.0064	0.0046	11	0.0184	1.637	0.478
128	b	7291059	IG	7B2	1.601	12	622	0	0.0000	0.0000	61	0.1306	n.a.	n.a.
130	a, b	7362175	IG	7B3	1.601	12	597	1	0.0003	0.0006	36	0.0952	–1.141	n.a.
530	b	7389987	IG	7B3	1.601	12	534	14	0.0129	0.0087	24	0.0818	2.078	0.665
133	b	7512346	In	7B6	1.601	12	625	1	0.0003	0.0005	17	0.0650	–1.141	n.a.
134	b	7578727	In	7B7	1.601	12	539	7	0.0049	0.0043	24	0.0930	0.589	0.347
136	a, b	7726936	In	7C1	1.461	12	386	6	0.0085	0.0051	18	0.0453	2.468	1.000
137	a, b	7759037	In	7C2	1.461	12	462	1	0.0010	0.0007	n.a.	n.a.	1.066	n.a.
138	a, b	7807461	In	7D1	1.486	12	346	0	0.0000	0.0000	20	0.0655	n.a.	n.a.
139	a, b	7868565	cd	7D2	1.486	12	347	9	0.0101	0.0086	24	0.0591	0.729	0.395
143	b	8118991	In	7E1	1.680	12	525	14	0.0083	0.0088	15	0.0495	–0.278	0.402
146	b	8251824	In	7E6	1.680	12	206	1	0.0008	0.0016	n.a.	n.a.	–1.141	n.a.
150	a, b	8443939	In	7F7	1.930	12	328	0	0.0000	0.0000	42	0.0814	n.a.	n.a.
153	a, b	8613511	In	8A5	2.150	12	477	8	0.0046	0.0056	26	0.1051	–0.682	0.285
157	a, b	8815156	IG	8C1	3.009	12	330	1	0.0005	0.0010	20	0.0384	–1.141	n.a.
160	b	8951179	In	8C9	3.009	12	268	3	0.0061	0.0037	26	0.0498	2.123	1.000
163	a, b	9093117	In	8D2	3.638	12	641	9	0.0057	0.0046	20	0.0754	0.956	0.481

(Continues...)

Table B2. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
165	a, b	9202487	In	8D9	3.638	12	297	1	0.0018	0.0011	28	0.0818	1.381	n.a.
166	a, b	9237654	IG	8D12	3.638	12	603	7	0.0043	0.0038	13	0.0379	0.461	0.158
167	b	9281023	IG	8E1	4.175	11	607	6	0.0048	0.0034	28	0.0563	1.663	0.623
169	b	9420312	In	8E10	4.175	12	309	1	0.0005	0.0011	23	0.0420	–1.141	n.a.
170	b	9461583	IG	8F2	4.508	12	605	19	0.0157	0.0104	31	0.0849	2.211	0.377
173	a, b	9639363	IG	9A2	3.536	12	496	15	0.0130	0.0100	16	0.0496	1.260	0.285
446	b	9710908	In	9A2	3.536	11	439	16	0.0147	0.0124	20	0.0513	0.794	0.618
175	a, b	9774954	In	9A3	3.536	12	622	8	0.0044	0.0043	n.a.	n.a.	0.141	0.400
176	b	9795700	IG	9A3	3.536	12	627	7	0.0042	0.0037	17	0.0541	0.563	0.196
177	a, b	9849981	In	9A4	3.536	12	409	0	0.0000	0.0000	30	0.0806	n.a.	n.a.
178	b	9890361	IG	9A4	3.536	11	488	14	0.0121	0.0098	32	0.0689	1.058	0.732
179	b	9938187	IG	9B1	2.813	12	540	17	0.0094	0.0104	34	0.0717	–0.415	0.279
182	b	10099354	IG	9B5	2.813	12	483	7	0.0063	0.0048	24	0.0511	1.254	0.540
464	b	10104098	IG	9B5	2.813	12	601	3	0.0020	0.0017	n.a.	n.a.	0.672	0.258
465	b	10141953	In	9B6	2.813	12	545	14	0.0070	0.0085	41	0.1015	–0.767	0.547
184	a, b	10173196	cd	9B7	2.813	12	431	10	0.0056	0.0077	n.a.	n.a.	–1.140	0.358
467	b	10246415	In	9B15	2.813	12	669	6	0.0023	0.0030	20	0.0519	–0.905	0.273
186	a, b	10272739	In	9C2	2.620	12	482	11	0.0038	0.0076	31	0.0642	–2.067	1.000
187	b	10300966	IG	9C4	2.620	12	563	7	0.0039	0.0041	18	0.0315	–0.179	0.329
188	b	10324188	IG	9D1	2.509	12	500	0	0.0000	0.0000	20	0.0397	n.a.	n.a.
189	a, b	10383742	In	9D3	2.509	12	551	7	0.0025	0.0042	n.a.	n.a.	–1.611	0.603
190	b	10402781	In	9D3	2.509	12	597	1	0.0003	0.0006	18	0.0312	–1.141	n.a.
191	a, b	10439282	cd	9D3	2.509	12	432	0	0.0000	0.0000	22	0.0588	n.a.	n.a.
470	b	10465155	IG	9D4	2.509	11	599	4	0.0019	0.0023	32	0.0652	–0.542	0.511
192	b	10490302	IG	9D4	2.509	12	534	0	0.0000	0.0000	46	0.0998	n.a.	n.a.
472	b	10554792	IG	9E1	2.391	12	679	4	0.0010	0.0020	23	0.0473	–1.747	1.000
194	a, b	10585619	In	9E1	2.391	12	580	1	0.0009	0.0006	25	0.0513	1.381	n.a.
195	b	10596143	In	9E2	2.391	12	555	0	0.0000	0.0000	51	0.1126	n.a.	n.a.
473	b	10625772	IG	9E10	2.391	12	535	8	0.0040	0.0050	18	0.0521	–0.796	0.631
196	a, b	10641809	cd	9F2	2.402	12	623	1	0.0003	0.0005	41	0.1125	–1.141	n.a.
197	a, b	10680736	In	9F4	2.402	12	547	0	0.0000	0.0000	n.a.	n.a.	n.a.	n.a.
198	b	10725814	In	9F7	2.402	12	868	5	0.0025	0.0019	13	0.0328	1.171	0.198
475	b	10746693	IG	9F8	2.402	12	636	6	0.0032	0.0031	29	0.0468	0.083	0.394
743	b	10818470	IG	9F13	2.402	12	308	2	0.0015	0.0022	27	0.0556	–0.850	0.455
201	a, b	10874510	IG	10A2	2.545	12	679	7	0.0031	0.0034	32	0.0710	–0.358	0.248
477	b	10887045	IG	10A3	2.545	12	641	12	0.0074	0.0062	38	0.0874	0.845	0.639

(Continues...)

Table B2. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
202	b	10924468	In	10A4	2.545	11	408	0	0.0000	0.0000	30	0.0649	n.a.	n.a.
480	b	10935038	cd	10A4	2.545	12	682	5	0.0026	0.0024	7	0.0144	0.328	0.345
203	a, b	10949796	cd	10A4	2.545	12	574	1	0.0003	0.0006	33	0.0922	–1.141	n.a.
532	b	10968776	IG	10A6	2.545	12	492	2	0.0007	0.0013	n.a.	n.a.	–1.451	1.000
204	a, b	11011770	IG	10A9	2.545	12	544	8	0.0038	0.0049	7	0.0194	–0.841	0.292
205	a, b	11070436	In	10B1	3.000	12	648	11	0.0068	0.0056	24	0.0845	0.876	0.353
483	b	11092779	IG	10B1	3.000	12	716	11	0.0065	0.0051	n.a.	n.a.	1.135	0.430
206	b	11110942	IG	10B2	3.000	12	605	5	0.0018	0.0027	17	0.0403	–1.291	0.339
207	b	11140441	IG	10B2	3.000	12	569	4	0.0021	0.0023	42	0.0939	–0.419	0.040
208	b	11168267	IG	10B2	3.000	12	522	11	0.0069	0.0070	n.a.	n.a.	–0.059	0.140
209	a, b	11207707	In	10B5	3.000	11	499	22	0.0098	0.0151	13	0.0341	–1.601	0.340
210	b	11244045	In	10B8	3.000	12	915	7	0.0027	0.0025	71	0.1627	0.205	0.249
211	b	11281619	cd	10B12	3.000	12	555	13	0.0060	0.0078	29	0.0544	–0.971	0.372
485	b	11304249	IG	10B14	3.000	12	533	13	0.0112	0.0081	26	0.1025	1.653	0.337
212	a, b	11325443	In	10B15	3.000	12	690	2	0.0005	0.0010	5	0.0167	–1.451	0.008
213	b	11361859	In	10C2	3.282	12	587	0	0.0000	0.0000	n.a.	n.a.	n.a.	n.a.
214	a, b	11414513	In	10C7	3.282	11	521	2	0.0010	0.0013	n.a.	n.a.	–0.778	0.022
215	a, b	11463155	In	10D2	3.440	12	566	12	0.0100	0.0070	15	0.1484	1.790	0.524
216	a, b	11507573	In	10D4	3.440	12	609	16	0.0112	0.0087	24	0.0489	1.226	0.329
217	a, b	11533051	IG	10D5	3.440	12	506	2	0.0007	0.0013	23	0.0666	–1.451	1.000
218	b	11550138	In	10D6	3.440	12	390	1	0.0014	0.0008	51	0.1895	1.381	n.a.
219	b	11609468	In	10E2	3.588	12	508	9	0.0056	0.0059	n.a.	n.a.	–0.200	0.793
220	b	11629125	In	10E3	3.588	12	408	1	0.0013	0.0008	32	0.0559	1.486	n.a.
221	a, b	11638396	In	10E4	3.588	12	386	8	0.0066	0.0069	29	0.0583	–0.156	0.610
222	b	11681093	IG	10F1	3.813	12	503	25	0.0181	0.0165	51	0.0964	0.454	0.239
488	b	11708720	IG	10F2	3.813	12	681	16	0.0084	0.0078	36	0.1019	0.337	0.684
224	a, b	11783192	In	10F9	3.813	12	609	13	0.0087	0.0071	55	0.1227	0.982	0.231
225	b	11814097	IG	10F11	3.813	12	597	12	0.0034	0.0067	20	0.0513	–2.087	1.000
660	b	11837057	IG	11A1	4.138	11	443	4	0.0030	0.0031	57	0.1391	–0.054	0.313
228	b	11912537	IG	11A3	4.138	12	518	1	0.0003	0.0006	47	0.0971	–1.141	n.a.
492	b	11939099	IG	11A4	4.138	12	625	11	0.0055	0.0058	22	0.1036	–0.249	0.425
229	a, b	11956720	IG	11A4	4.138	12	444	9	0.0034	0.0067	38	0.0695	–2.016	0.201
493	b	12015826	IG	11A6	4.138	12	621	8	0.0021	0.0043	9	0.0509	–1.983	0.327
231	a, b	12030561	In	11A6	4.138	12	562	8	0.0040	0.0047	20	0.0808	–0.613	0.290
232	b	12052930	IG	11A6	4.138	12	560	3	0.0009	0.0018	21	0.0883	–1.629	0.008
233	b	12109811	In	11A7	4.138	12	392	18	0.0124	0.0152	29	0.0736	–0.791	0.662

(Continues...)

Table B2. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
235	b	12147557	In	11A8	4.138	12	501	6	0.0031	0.0040	24	0.0558	–0.847	0.218
237	b	12201432	IG	11A9	4.138	12	521	23	0.0191	0.0146	12	0.0275	1.370	0.308
447	b	12231484	IG	11A9	4.138	11	549	7	0.0040	0.0044	28	0.0761	–0.353	0.225
238	b	12234442	IG	11A9	4.138	12	545	13	0.0120	0.0079	51	0.0854	2.190	0.866
239	b	12269282	In	11A10	4.138	12	388	27	0.0264	0.0230	n.a.	n.a.	0.659	0.293
241	a, b	12335618	In	11A12	4.138	12	568	7	0.0035	0.0041	44	0.1413	–0.588	0.181
242	b	12377504	IG	11B1	4.436	11	529	18	0.0144	0.0116	19	0.0448	1.064	0.919
721	b	12455184	IG	11B4	4.436	11	379	24	0.0203	0.0216	33	0.0703	–0.271	0.919
245	b	12507177	In	11B9	4.436	12	513	6	0.0034	0.0039	34	0.0582	–0.440	0.515
246	b	12557700	In	11B14	4.436	12	571	7	0.0054	0.0041	45	0.1107	1.305	0.247
248	a, b	12617534	IG	11C2	4.512	12	676	4	0.0028	0.0020	48	0.0854	1.472	0.126
249	a, b	12650222	IG	11C3	4.512	12	584	2	0.0010	0.0011	9	0.0157	–0.248	1.000
250	a, b	12701253	IG	11D1	4.689	11	584	6	0.0019	0.0035	11	0.0612	–1.851	0.406
251	a, b	12744897	cd	11D1	4.689	12	452	5	0.0018	0.0037	24	0.0453	–1.831	1.000
252	b	12777694	In	11D4	4.689	12	555	10	0.0041	0.0060	41	0.0917	–1.272	0.528
253	b	12820460	IG	11D6	4.689	12	517	6	0.0046	0.0038	48	0.1169	0.723	0.376
254	a, b	12859984	IG	11D8	4.689	12	421	3	0.0023	0.0024	35	0.0919	–0.028	0.515
258	b	12955002	In	11E1	4.812	12	534	1	0.0003	0.0006	27	0.0858	–1.141	n.a.
259	b	13007119	In	11E2	4.812	12	313	14	0.0174	0.0148	27	0.1477	0.753	0.568
260	b	13053066	In	11E4	4.812	12	612	2	0.0005	0.0011	14	0.0285	–1.451	0.008
272	a, b	13096432	IG	11E8	4.812	12	257	7	0.0096	0.0090	19	0.0461	0.231	0.308
273	a, b	13102200	In	11E8	4.812	12	430	20	0.0222	0.0154	23	0.0587	1.938	0.614
722	b	13164524	IG	11E11	4.812	10	502	5	0.0048	0.0035	26	0.0898	1.435	0.651
276	a, b	13232942	In	11F1	4.877	12	326	2	0.0014	0.0020	8	0.0346	–0.850	0.018
277	b	13268022	IG	11F6	4.877	12	618	10	0.0037	0.0054	35	0.0645	–1.253	0.709
278	a, b	13318463	In	12A1	4.974	12	612	19	0.0142	0.0103	2	0.0093	1.684	0.657
279	a, b	13351547	In	12A2	4.974	12	664	12	0.0086	0.0060	21	0.0651	1.838	0.602
280	b	13385015	IG	12A4	4.974	12	498	17	0.0073	0.0113	8	0.0411	–1.525	0.325
450	b	13397127	IG	12A4	4.974	12	720	1	0.0002	0.0005	6	0.0120	–1.141	n.a.
311	b	13397557	IG	12A4	4.974	12	204	1	0.0008	0.0016	21	0.0418	–1.141	n.a.
312	a, b	13428303	In	12A7	4.974	12	637	5	0.0022	0.0026	20	0.0594	–0.617	0.319
313	b	13468507	In	12A9	4.974	12	546	7	0.0037	0.0042	1	0.0207	–0.537	0.239
314	a, b	13505120	cd	12B2	5.019	12	445	1	0.0004	0.0007	34	0.0688	–1.141	n.a.
318	b	13647962	In	12C5	5.026	12	351	2	0.0031	0.0019	28	0.0548	1.824	0.714
319	b	13670436	In	12C6	5.026	12	523	3	0.0012	0.0019	14	0.0370	–1.179	0.636
320	b	13710254	In	12C7	5.026	11	644	2	0.0011	0.0011	13	0.0231	0.199	0.120

(Continues...)

Table B2. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
321	b	13742548	IG	12D1	5.023	12	509	3	0.0013	0.0020	n.a.	n.a.	–1.179	0.015
323	b	13831025	IG	12D2	5.023	12	372	5	0.0022	0.0045	17	0.0638	–1.831	0.603
325	b	13895682	IG	12D4	5.023	12	566	7	0.0021	0.0041	29	0.0675	–1.944	1.000
326	a, b	13935313	IG	12E1	4.979	12	611	11	0.0059	0.0060	17	0.0362	–0.076	0.203
342	b	14017479	IG	12E2	4.979	12	551	13	0.0049	0.0078	26	0.0522	–1.567	0.472
346	b	14129542	IG	12E8	4.979	12	428	8	0.0031	0.0062	12	0.0345	–1.983	0.752
348	a, b	14195150	IG	12E8	4.979	12	571	10	0.0083	0.0058	45	0.0953	1.797	0.563
745	b	14219265	IG	12E9	4.979	11	381	2	0.0010	0.0018	25	0.0738	–1.430	0.010
350	b	14301478	IG	12E10	4.979	11	531	7	0.0027	0.0045	37	0.0663	–1.650	0.485
367	a, b	14324184	IG	12E10	4.979	12	595	2	0.0010	0.0011	25	0.0878	–0.382	0.273
368	b	14357426	IG	12F1	4.934	12	530	5	0.0023	0.0031	12	0.0201	–0.920	0.312
369	b	14394544	IG	12F1	4.934	12	678	5	0.0023	0.0024	33	0.0823	–0.144	0.585
370	a, b	14414966	In	12F1	4.934	12	570	28	0.0173	0.0163	25	0.0576	0.271	0.403
371	b	14446806	In	12F1	4.934	12	572	14	0.0070	0.0081	21	0.0445	–0.571	0.437
373	b	14514836	In	12F2	4.934	12	806	22	0.0062	0.0090	22	0.0382	–1.387	0.545
374	a, b	14526351	In	12F2	4.934	12	607	0	0.0000	0.0000	41	0.0729	n.a.	n.a.
375	a, b	14561277	In	12F3	4.934	12	633	2	0.0009	0.0010	29	0.1077	–0.382	0.273
376	b	14593606	In	12F4	4.934	11	351	10	0.0080	0.0097	31	0.0546	–0.766	0.475
378	b	14626366	In	12F4	4.934	12	691	7	0.0022	0.0034	47	0.0975	–1.381	0.607
379	a, b	14664445	In	12F5	4.934	12	584	11	0.0081	0.0062	n.a.	n.a.	1.274	0.255
380	b	14703175	In	12F6	4.934	12	578	5	0.0046	0.0029	28	0.0767	2.285	0.829
381	a, b	14765244	In	13A1	4.883	12	429	0	0.0000	0.0000	25	0.0470	n.a.	n.a.
382	b	14826789	In	13A5	4.883	12	588	2	0.0010	0.0011	n.a.	n.a.	–0.382	0.030
383	b	14883203	In	13A5	4.883	12	473	15	0.0053	0.0105	42	0.1145	–2.133	1.000
384	b	14920236	In	13A10	4.883	12	494	7	0.0024	0.0047	30	0.0822	–1.944	1.000
385	b	14933236	In	13A11	4.883	12	515	6	0.0053	0.0039	43	0.0752	1.421	0.800
386	b	14948045	In	13A12	4.883	12	590	7	0.0022	0.0039	11	0.0162	–1.713	0.153
387	b	14965085	IG	13B1	4.718	12	618	1	0.0003	0.0005	n.a.	n.a.	–1.141	n.a.
388	b	15005477	IG	13B1	4.718	12	693	1	0.0002	0.0005	n.a.	n.a.	–1.141	n.a.
389	b	15057214	IG	13B3	4.718	12	611	1	0.0009	0.0005	n.a.	n.a.	1.381	n.a.
391	b	15087602	IG	13B4	4.718	12	560	2	0.0017	0.0012	16	0.0980	1.290	0.333
390	b	15117024	IG	13B4	4.718	12	548	11	0.0091	0.0066	40	0.0756	1.516	0.370
392	b	15148546	In	13B6	4.718	12	465	4	0.0034	0.0028	62	0.1391	0.627	0.273
534	b	15152610	IG	13B6	4.718	12	405	2	0.0008	0.0016	35	0.0635	–1.451	0.008
393	a, b	15169667	In	13B6	4.718	12	560	0	0.0000	0.0000	28	0.0580	n.a.	n.a.
535	b	15198298	In	13B9	4.718	12	561	10	0.0072	0.0059	51	0.1113	0.893	0.459

(Continues...)

Table B2. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
394	b	15211531	In	13C1	4.624	12	563	15	0.0103	0.0088	31	0.0570	0.736	0.245
282	a, b	15282518	cd	13C3	4.624	12	700	4	0.0010	0.0019	54	0.0990	–1.747	1.000
285	b	15467669	IG	13E3	4.330	10	573	22	0.0145	0.0136	29	0.0585	0.311	0.711
286	b	15503689	IG	13E4	4.330	12	478	16	0.0121	0.0111	46	0.0936	0.411	0.237
287	a, b	15525959	IG	13E7	4.330	12	573	10	0.0043	0.0058	7	0.0128	–1.046	0.448
288	a, b	15536600	cd	13E8	4.330	12	519	4	0.0015	0.0026	21	0.0494	–1.385	0.179
295	b	15544479	IG	13E8	4.330	12	582	2	0.0006	0.0011	26	0.0461	–1.451	0.008
296	b	15599496	In	13E14	4.330	12	615	5	0.0018	0.0027	22	0.0370	–1.224	0.204
297	a, b	15638469	IG	13F1	4.108	12	630	5	0.0013	0.0026	21	0.0477	–1.831	0.008
298	b	15692843	cd	13F15	4.108	12	529	4	0.0029	0.0025	15	0.0234	0.546	0.180
299	a, b	15730456	In	13F18	4.108	12	617	6	0.0030	0.0032	28	0.0555	–0.295	0.172
294	b	15746063	In	14A1	3.928	12	627	7	0.0021	0.0037	11	0.0195	–1.713	0.294
301	a, b	15798732	cd	14A3	3.928	12	573	6	0.0027	0.0035	42	0.0777	–0.847	0.212
303	a	15893832	cd	14A6	3.928	12	608	5	0.0016	0.0027	34	0.0604	–1.5273	n.a.
304	b	15906320	In	14A8	3.928	12	564	9	0.0085	0.0053	35	0.0759	2.462	0.937
725	b	15964533	IG	14B1	3.667	12	331	4	0.0028	0.0040	9	0.0166	–1.023	0.326
307	b	16047623	IG	14B3	3.667	12	506	7	0.0039	0.0046	47	0.0867	–0.562	0.719
726	b	16096635	IG	14B6	3.667	12	384	4	0.0024	0.0034	41	0.0814	–1.103	0.225
310	b	16152786	In	14B9	3.667	12	421	0	0.0000	0.0000	48	0.2090	n.a.	n.a.
336	b	16168466	In	14B14	3.667	12	601	3	0.0014	0.0017	33	0.0702	–0.579	0.079
334	b	16255530	IG	14C4	3.571	12	630	7	0.0032	0.0037	47	0.1108	–0.486	0.170
333	a, b	16276272	In	14C6	3.571	12	578	7	0.0041	0.0040	12	0.0205	0.128	0.255
451	b	16302764	In	14D1	3.494	12	661	3	0.0013	0.0015	9	0.0143	–0.379	0.394
331	a, b	16349660	In	14D4	3.494	12	592	8	0.0042	0.0045	27	0.0506	–0.225	0.491
330	a, b	16371094	IG	14E1	3.426	12	576	21	0.0159	0.0121	45	0.1075	1.412	0.418
329	a, b	16429444	IG	14F1	3.318	12	435	7	0.0081	0.0053	18	0.0323	2.047	0.506
328	b	16468799	In	14F4	3.318	12	546	0	0.0000	0.0000	21	0.0385	n.a.	n.a.
366	a, b	16471181	IG	14F4	3.318	12	610	16	0.0139	0.0087	11	0.0232	2.597	1.000
364	a, b	16530195	IG	15A1	3.129	12	488	11	0.0071	0.0075	34	0.0863	–0.180	0.379
363	b	16550618	IG	15A3	3.129	12	603	9	0.0058	0.0049	34	0.0964	0.667	0.484
359	b	16694289	IG	15B1	3.065	12	624	12	0.0061	0.0064	28	0.0550	–0.148	0.305
402	b	16783883	IG	15C4	3.008	12	604	9	0.0046	0.0049	24	0.0796	–0.283	0.417
405	b	16864287	IG	15D4	2.932	12	652	9	0.0064	0.0046	56	0.0951	1.595	0.221
406	b	16907918	In	15E1	2.867	12	665	13	0.0075	0.0065	51	0.1282	0.669	0.437
407	b	16933856	IG	15E3	2.837	12	618	0	0.0000	0.0000	44	0.0834	n.a.	n.a.
410	b	17027142	In	15F1	2.782	12	644	22	0.0126	0.0113	26	0.0518	0.517	0.206

(Continues...)

Table B2. *(Cont.)*

Locus	Study	Position	Type	Cyto	*r*	*n*	*L*	*k*	π	θ	*Ds*	*Div*	*D*	Z_{nS}
411	b	17058256	IG	15F4	2.782	12	588	7	0.0038	0.0039	32	0.0625	–0.153	0.366
422	b	17088628	IG	15F4	2.782	12	605	19	0.0119	0.0104	25	0.0568	0.640	0.279
727	b	17161040	IG	16A1	2.684	10	472	6	0.0059	0.0045	15	0.0400	1.281	0.396
424	b	17229023	IG	16A4	2.684	12	688	6	0.0024	0.0029	6	0.0165	–0.585	0.417
728	b	17301040	IG	16A5	2.684	12	267	0	0.0000	0.0000	4	0.0105	n.a.	n.a.
426	b	17353968	IG	16B1	2.591	12	732	5	0.0018	0.0023	39	0.0867	–0.718	0.048
428	b	17384694	IG	16B4	2.591	12	626	13	0.0054	0.0069	n.a.	n.a.	–0.896	0.547
729	b	17442422	IG	16B8	2.591	11	460	9	0.0089	0.0067	15	0.0745	1.368	0.643
430	b	17492612	IG	16B10	2.591	12	655	1	0.0003	0.0005	40	0.0903	–1.141	n.a.
730	b	17541034	IG	16C1	2.527	12	189	1	0.0009	0.0018	6	0.0212	–1.141	n.a.
431	b	17619496	IG	16D1	2.487	12	568	7	0.0037	0.0041	9	0.0278	–0.332	0.109
432	b	17662415	IG	16D4	2.487	12	597	1	0.0005	0.0006	24	0.0650	–0.195	n.a.
436	b	17980460	IG	16F7	2.436	9	661	14	0.0113	0.0078	3	0.0114	2.137	0.707
438	b	18064811	IG	17A3	2.424	12	578	9	0.0028	0.0052	42	0.1356	–1.830	0.782
439	b	18132877	IG	17A4	2.424	12	658	2	0.0014	0.0010	n.a.	n.a.	1.022	0.667
440	b	18201747	IG	17A7	2.424	12	638	9	0.0057	0.0047	10	0.0378	0.852	0.727
444	b	18579133	IG	17D3	2.467	12	646	14	0.0085	0.0072	13	0.0364	0.781	0.385

Table B3. Compatibility of the European variation with a simple bottleneck model.

For each locus, we calculated the probability that the observed *k* segregating sites can be explained by a simple bottleneck, as well as the approximate Bayesian posterior probability of being in a strong bottleneck mimicking selection (see CHAPTER 1.2.).

Loci are ordered from the telomere to the centromere; for each one, the following information is given (for additional information, see Table B2):

- θ is the WATTERSON (1975) estimate of nucleotide diversity for the European sample;
- CR-II^{all} and CR-III^{all} specifies whether a locus is within one of the candidate regions identified by method II^{all} (and common to all other methods) or only by method III^{all}, respectively (see CHAPTER 1.2., SUBSECTION 1.2.2.4.): in brief, a locus falls within one of such regions if it is one of 5 consecutive loci with $Q^{C=5} < 0.05$ (*i.e.*, together they contain less segregating sites than expected under the bottleneck estimated by method II^{all} or III^{all}); if the locus is at the sides of such region, it is also required to have low polymorphism. Letters identify the different candidate regions (see Figure 9).
- Q and Q^{E} are the probabilities to harbor at most the *k* observed number of segregating sites under the bottleneck estimated by methods I, II^{s}, and II^{all} (Q), or methods III^{s} and III^{all} (Q^{E}; see CHAPTER 1.2., SUBSECTIONS 1.2.1.4. and 1.2.2.4.);
- *PP* is the approximate Bayesian posterior probability of being in the strong bottleneck mimicking selection, given that the whole dataset is better explained by a combination of two bottlenecks (see CHAPTER 1.2., SUBSECTION 1.2.2.4.).

* Asterisks indicate values of $Q < 0.05$, or $PP > 0.3$. *PP* is the probability for a locus to belong to the strong bottleneck: we chose the arbitrary value of 0.3 to account for this uncertainty, hence distinguishing 31 loci (note that ~20 loci should belong to the strong bottleneck, CHAPTER 1.2., SUBSECTION 1.2.2.4.).
n.a. Not available.

The table follows in the next pages.

Table B3. Compatibility of the European variation with a simple bottleneck model.

Locus	Position	θ	CR-II^{all}	CR-III^{all}	Q^{I}	Q^{IIs}	Q^{IIall}	$Q^{E\text{-}IIIs}$	$Q^{E\text{-}IIIall}$	*PP*
419	1670429	0.0006			0.490	0.489	0.490	0.160	0.118	0.111
10	2008922	0.0000			0.132	0.130	0.109	0.130	0.094	0.472 *
17	2055186	0.0030			0.648	0.642	0.651	0.711	0.585	0.000
1	2113655	0.0043			0.518	0.512	0.529	0.143	0.269	0.000
22	2239760	0.0042			0.804	0.804	0.816	0.633	0.413	0.000
25	2247266	0.0034			0.652	0.646	0.652	0.485	0.447	0.000
26	2250516	0.0011			0.232	0.224	0.209	0.139	0.123	0.004
32	2295679	0.0018			0.476	0.474	0.487	0.150	0.144	0.002
33	2298889	0.0063			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
38	2378109	0.0015	A		0.230	0.224	0.210	0.109	0.033 *	0.005
18	2555495	0.0060	A		0.798	0.795	0.800	0.326	0.524	0.000
5	2593830	0.0000	A		0.090	0.087	0.064	0.056	0.025 *	0.600 *
45	2844745	0.0005	A		0.149	0.147	0.122	0.062	0.066	0.506 *
46	2885298	0.0016	A		0.205	0.196	0.185	0.130	0.145	0.004
55	3338479	0.0083			0.671	0.665	0.680	0.234	0.525	0.000
54	3341441	0.0103			0.556	0.549	0.569	0.879	0.984	0.000
57	3436268	0.0041			0.711	0.702	0.720	0.523	0.447	0.000
60	3474557	0.0074			0.636	0.637	0.643	0.463	0.412	0.001
56	3629942	0.0041			0.849	0.841	0.849	0.245	0.255	0.001
76	3690386	0.0080			0.558	0.544	0.562	0.224	0.380	0.002
77	3717795	0.0044			0.407	0.402	0.429	0.305	0.298	0.001
78	3764669	0.0048			0.563	0.552	0.564	0.330	0.128	0.008
80	3876608	0.0126			0.768	0.764	0.778	0.644	0.595	0.001
81	3916988	0.0017			0.288	0.294	0.285	0.101	0.103	0.003
462	3953777	0.0015			0.648	0.638	0.652	0.192	0.113	0.001
84	4054401	0.0072			0.739	0.728	0.741	0.397	0.462	0.000
85	4106003	0.0057			0.655	0.636	0.666	0.638	0.591	0.000
66	4296277	0.0052			0.524	0.509	0.543	0.187	0.061	0.006
67	4548130	0.0015			0.230	0.222	0.214	0.194	0.098	0.004
90	4935383	0.0101			0.579	0.567	0.591	0.286	0.478	0.002
91	4991826	0.0070			0.578	0.568	0.579	0.523	0.464	0.001
93	5073706	0.0041			0.710	0.704	0.717	0.577	0.402	0.000
94	5126621	0.0056			0.526	0.521	0.545	0.467	0.405	0.000
95	5171847	0.0033			0.443	0.442	0.453	0.371	0.237	0.001
106	5478703	0.0106			0.829	0.826	0.835	0.683	0.753	0.000

(Continues...)

Table B3. *(Cont.)*

Locus	Position	θ	CR-II^{all}	CR-III^{all}	Q^{I}	Q^{IIs}	Q^{IIall}	$Q^{E\text{-}IIIs}$	$Q^{E\text{-}IIIall}$	*PP*
72	5518785	0.0016			0.096	0.093	0.053	0.129	0.119	0.011
73	5592462	0.0017			0.500	0.493	0.509	0.541	0.650	0.001
109	5767488	0.0039			0.673	0.669	0.684	0.415	0.591	0.000
102	5886956	0.0012			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
114	6605700	0.0011			0.644	0.641	0.652	0.103	0.065	0.068
115	6651455	0.0008			0.270	0.265	0.258	0.090	0.074	0.272
116	6687195	0.0115			0.675	0.670	0.685	0.888	0.907	0.000
117	6741220	0.0109			0.644	0.635	0.660	0.530	0.763	0.002
118	6790458	0.0024			0.493	0.486	0.499	0.218	0.185	0.000
119	6838119	0.0056		B	0.322	0.320	0.327	0.309	0.306	0.003
120	6916170	0.0119		B	0.678	0.674	0.683	0.269	0.826	0.000
122	7006861	0.0000		B	0.230	0.223	0.213	0.178	0.137	0.325 *
502	7033935	0.0043		B	0.449	0.430	0.451	0.109	0.110	0.003
124	7083799	0.0017		B	0.644	0.644	0.649	0.145	0.192	0.000
125	7134523	0.0000		B	0.165	0.165	0.146	0.061	0.023 *	0.405 *
126	7185769	0.0046		B	0.324	0.314	0.331	0.287	0.336	0.006
128	7291059	0.0000			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
130	7362175	0.0006	C		0.120	0.114	0.084	0.217	0.179	0.636 *
530	7389987	0.0087	C		0.670	0.666	0.681	0.404	0.798	0.000
133	7512346	0.0005	C		0.176	0.177	0.155	0.165	0.140	0.454 *
134	7578727	0.0043	C		n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
136	7726936	0.0051	C		0.350	0.344	0.355	0.216	0.227	0.003
137	7759037	0.0007	C		0.198	0.190	0.172	0.146	0.100	0.005
138	7807461	0.0000	C		0.145	0.141	0.125	0.066	0.035 *	0.445 *
139	7868565	0.0086			0.769	0.762	0.781	0.702	0.684	0.000
143	8118991	0.0088			0.815	0.812	0.818	0.632	0.437	0.001
146	8251824	0.0016			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
150	8443939	0.0000			0.080	0.075	0.051	0.172	0.104	0.646 *
153	8613511	0.0056			0.480	0.475	0.490	0.677	0.596	0.000
157	8815156	0.0010			0.221	0.210	0.196	0.149	0.086	0.365 *
160	8951179	0.0037			0.319	0.309	0.327	n.a.	n.a.	n.a.
163	9093117	0.0046			0.532	0.536	0.545	0.888	0.674	0.004
165	9202487	0.0011			0.400	0.391	0.397	0.054	0.053	0.155
166	9237654	0.0038			0.773	0.763	0.781	0.756	0.648	0.000
167	9281023	0.0034			0.687	0.681	0.694	0.224	0.132	0.002
169	9420312	0.0011			0.765	0.765	0.771	0.139	0.095	0.044

(Continues...)

Table B3. *(Cont.)*

Locus	Position	θ	CR-II^{all}	CR-III^{all}	Q^{I}	Q^{IIs}	Q^{IIall}	$Q^{E\text{-}IIIs}$	$Q^{E\text{-}IIIall}$	*PP*
170	9461583	0.0104			0.593	0.592	0.607	0.315	0.657	0.001
173	9639363	0.0100			0.990	0.989	0.989	0.665	0.339	0.000
446	9710908	0.0124			0.747	0.739	0.758	0.421	0.188	0.009
175	9774954	0.0043			0.324	0.320	0.331	0.278	0.450	0.002
176	9795700	0.0037			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
177	9849981	0.0000			0.057	0.054	0.034 *	0.072	0.027 *	0.717 *
178	9890361	0.0098			0.588	0.584	0.606	0.803	0.776	0.001
179	9938187	0.0104			0.771	0.778	0.779	0.825	0.804	0.000
182	10099354	0.0048			0.592	0.590	0.610	n.a.	n.a.	n.a.
464	10104098	0.0017			0.175	0.172	0.151	0.622	0.390	0.258
465	10141953	0.0085			0.726	0.718	0.734	0.297	0.314	0.004
184	10173196	0.0077			0.612	0.609	0.621	0.538	0.276	0.003
467	10246415	0.0030			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
186	10272739	0.0076			0.690	0.689	0.706	0.413	0.386	0.000
187	10300966	0.0041			0.639	0.638	0.650	0.330	0.573	0.000
188	10324188	0.0000			0.307	0.302	0.296	0.111	0.050	0.251
189	10383742	0.0042			0.586	0.584	0.594	0.561	0.369	0.001
190	10402781	0.0006			0.089	0.085	0.055	n.a.	n.a.	n.a.
191	10439282	0.0000	D		0.308	0.303	0.297	0.054	0.021 *	0.246
470	10465155	0.0023	D		0.790	0.786	0.801	0.153	0.209	0.000
192	10490302	0.0000	D		0.046 *	0.045 *	0.024 *	0.048 *	0.016 *	0.782 *
472	10554792	0.0020	D		0.369	0.365	0.371	0.179	0.109	0.003
194	10585619	0.0006	D		0.161	0.156	0.138	0.064	0.050	0.501 *
195	10596143	0.0000	D		0.020 *	0.018 *	0.006 *	0.093	0.042 *	0.922 *
473	10625772	0.0050	D		0.809	0.803	0.816	0.301	0.257	0.000
196	10641809	0.0005	D		0.465	0.459	0.469	0.073	0.031 *	0.003
197	10680736	0.0000	D		0.190	0.186	0.172	0.065	0.023 *	0.375 *
198	10725814	0.0019			0.273	0.270	0.270	n.a.	n.a.	n.a.
475	10746693	0.0031			0.778	0.766	0.784	0.492	0.413	0.000
743	10818470	0.0022			0.300	0.296	0.296	0.145	0.130	0.004
201	10874510	0.0034			0.696	0.695	0.714	0.445	0.434	0.000
477	10887045	0.0062			0.542	0.540	0.550	0.167	0.380	0.001
202	10924468	0.0000			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
480	10935038	0.0024			0.461	0.457	0.471	0.377	0.152	0.005
203	10949796	0.0006			0.110	0.108	0.075	0.181	0.156	0.661 *
532	10968776	0.0013			0.698	0.692	0.711	0.442	0.204	0.002

(Continues...)

Table B3. *(Cont.)*

Locus	Position	θ	CR-II^{all}	CR-III^{all}	Q^{I}	Q^{IIs}	Q^{IIall}	$Q^{E\text{-}IIIs}$	$Q^{E\text{-}IIIall}$	*PP*
204	11011770	0.0049			0.469	0.470	0.483	0.521	0.518	0.000
205	11070436	0.0056			0.626	0.618	0.628	0.435	0.363	0.001
483	11092779	0.0051			0.405	0.407	0.42	0.663	0.612	0.002
206	11110942	0.0027			0.242	0.239	0.232	0.171	0.108	0.004
207	11140441	0.0023			0.174	0.168	0.152	0.331	0.243	0.004
208	11168267	0.0070			0.916	0.913	0.918	0.514	0.445	0.000
209	11207707	0.0151			0.664	0.668	0.678	0.987	0.980	0.000
210	11244045	0.0025			0.388	0.379	0.39	n.a.	n.a.	n.a.
211	11281619	0.0078			0.648	0.646	0.659	0.405	0.378	0.001
485	11304249	0.0081			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
212	11325443	0.0010		E	0.345	0.331	0.34	0.286	0.224	0.118
213	11361859	0.0000		E	0.12	0.113	0.094	0.052	0.026 *	0.517 *
214	11414513	0.0013		E	0.289	0.281	0.284	0.113	0.116	0.003
215	11463155	0.0070		E	0.789	0.784	0.803	0.207	0.363	0.001
216	11507573	0.0087		E	0.534	0.526	0.548	0.578	0.653	0.001
217	11533051	0.0013		E	0.669	0.664	0.676	0.144	0.087	0.001
218	11550138	0.0008		E	0.243	0.238	0.232	0.085	0.042 *	0.004
219	11609468	0.0059		E	0.854	0.850	0.857	n.a.	n.a.	n.a.
220	11629125	0.0008		E	0.770	0.768	0.774	0.066	0.028 *	0.003
221	11638396	0.0069			0.613	0.605	0.624	0.248	0.256	0.002
222	11681093	0.0165			0.858	0.859	0.865	n.a.	n.a.	n.a.
488	11708720	0.0078			0.582	0.577	0.592	0.793	0.546	0.005
224	11783192	0.0071			0.643	0.631	0.648	0.406	0.461	0.000
225	11814097	0.0067			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
660	11837057	0.0031			0.777	0.773	0.788	n.a.	n.a.	n.a.
228	11912537	0.0006			0.262	0.256	0.249	n.a.	n.a.	n.a.
492	11939099	0.0058			0.641	0.636	0.647	0.652	0.520	0.000
229	11956720	0.0067			0.745	0.752	0.757	0.662	0.657	0.000
493	12015826	0.0043			0.468	0.457	0.474	0.355	0.368	0.000
231	12030561	0.0047			0.266	0.255	0.263	0.738	0.713	0.004
232	12052930	0.0018			0.356	0.351	0.360	0.133	0.125	0.043
233	12109811	0.0152			0.755	0.750	0.767	0.239	0.076	0.020
235	12147557	0.0040			0.753	0.752	0.762	0.306	0.247	0.000
237	12201432	0.0146			0.627	0.637	0.644	n.a.	n.a.	n.a.
447	12231484	0.0044			0.576	0.574	0.586	0.359	0.306	0.000
238	12234442	0.0079			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

(Continues...)

Table B3. *(Cont.)*

Locus	Position	θ	CR-II^{all}	CR-III^{all}	Q^{I}	Q^{IIs}	Q^{IIall}	$Q^{E\text{-}IIIs}$	$Q^{E\text{-}IIIall}$	*PP*
239	12269282	0.0230			0.678	0.668	0.686	0.248	0.078	0.038
241	12335618	0.0041			0.932	0.932	0.937	0.237	0.230	0.000
242	12377504	0.0116			0.649	0.645	0.655	0.654	0.360	0.004
721	12455184	0.0216			0.827	0.828	0.834	0.431	0.918	0.000
245	12507177	0.0039	F		0.409	0.398	0.411	0.208	0.068	0.009
246	12557700	0.0041	F		0.473	0.466	0.480	n.a.	n.a.	n.a.
248	12617534	0.0020	F		0.294	0.283	0.286	0.237	0.083	0.004
249	12650222	0.0011	F		0.184	0.172	0.155	0.170	0.133	0.007
250	12701253	0.0035	F		0.384	0.389	0.401	0.687	0.560	0.001
251	12744897	0.0037	F		0.286	0.277	0.287	0.237	0.098	0.006
252	12777694	0.0060			0.442	0.445	0.447	0.483	0.401	0.001
253	12820460	0.0038			0.417	0.414	0.428	0.290	0.247	0.001
254	12859984	0.0024			0.489	0.486	0.500	0.186	0.140	0.000
258	12955002	0.0006			0.066	0.057	0.027 *	0.112	0.073	0.855 *
259	13007119	0.0148			0.808	0.806	0.810	0.811	0.915	0.000
260	13053066	0.0011			0.213	0.200	0.199	n.a.	n.a.	n.a.
272	13096432	0.0090			0.794	0.791	0.804	0.419	0.488	0.000
273	13102200	0.0154			0.822	0.822	0.828	0.318	0.330	0.003
722	13164524	0.0035			0.231	0.228	0.227	0.299	0.327	0.004
276	13232942	0.0020			0.402	0.387	0.400	0.155	0.080	0.003
277	13268022	0.0054			0.404	0.396	0.405	0.745	0.719	0.002
278	13318463	0.0103			0.717	0.712	0.723	0.587	0.629	0.000
279	13351547	0.0060			0.616	0.595	0.615	0.579	0.382	0.001
280	13385015	0.0113			0.699	0.702	0.712	0.984	0.945	0.000
450	13397127	0.0005	G		0.060	0.056	0.025 *	0.125	0.070	0.871 *
311	13397557	0.0016	G		0.140	0.132	0.109	0.143	0.121	0.564 *
312	13428303	0.0026	G		0.512	0.511	0.521	0.275	0.103	0.003
313	13468507	0.0042	G		0.772	0.772	0.781	0.681	0.689	0.000
314	13505120	0.0007	G		0.170	0.166	0.145	0.098	0.061	0.484 *
318	13647962	0.0019	G		0.357	0.350	0.355	0.094	0.037 *	0.004
319	13670436	0.0019	G		0.196	0.196	0.178	0.171	0.068	0.005
320	13710254	0.0011	G		0.199	0.191	0.176	0.226	0.150	0.004
321	13742548	0.0020	G		0.589	0.585	0.593	0.417	0.222	0.001
323	13831025	0.0045	G		0.430	0.422	0.438	0.562	0.376	0.003
325	13895682	0.0041			0.849	0.850	0.857	n.a.	n.a.	n.a.
326	13935313	0.0060			0.744	0.737	0.754	0.469	0.470	0.000

(Continues...)

Table B3. *(Cont.)*

Locus	Position	θ	CR-II^{all}	CR-III^{all}	Q^{I}	Q^{IIs}	Q^{IIall}	$Q^{E\text{-}IIIs}$	$Q^{E\text{-}IIIall}$	*PP*
342	14017479	0.0078			0.500	0.490	0.509	n.a.	n.a.	n.a.
346	14129542	0.0062		H	0.869	0.868	0.872	0.412	0.295	0.000
348	14195150	0.0058		H	0.573	0.556	0.582	0.226	0.452	0.000
745	14219265	0.0018		H	0.891	0.888	0.895	0.184	0.158	0.005
350	14301478	0.0045		H	0.645	0.645	0.654	0.521	0.325	0.002
367	14324184	0.0011		H	0.209	0.203	0.185	0.070	0.027 *	0.004
368	14357426	0.0031		H	0.313	0.309	0.310	0.303	0.160	0.002
369	14394544	0.0024		H	0.718	0.713	0.728	0.267	0.164	0.001
370	14414966	0.0163			0.824	0.827	0.836	0.991	0.994	0.000
371	14446806	0.0081			0.681	0.673	0.692	0.517	0.464	0.000
373	14514836	0.0090			0.978	0.978	0.979	0.321	0.319	0.002
374	14526351	0.0000			0.073	0.070	0.050 *	0.048 *	0.009 *	0.651 *
375	14561277	0.0010	I		0.124	0.114	0.080	0.118	0.118	0.006
376	14593606	0.0097	I		0.620	0.603	0.628	0.181	0.078	0.009
378	14626366	0.0034	I		0.380	0.370	0.387	0.431	0.318	0.005
379	14664445	0.0062	I		0.394	0.383	0.400	0.173	0.299	0.004
380	14703175	0.0029	I		0.499	0.492	0.505	n.a.	n.a.	n.a.
381	14765244	0.0000	I		0.024 *	0.021 *	0.007 *	0.048 *	0.018 *	0.907 *
382	14826789	0.0011	I		0.175	0.172	0.148	0.139	0.116	0.006
383	14883203	0.0105	I		n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
384	14920236	0.0047	I		0.549	0.545	0.566	0.182	0.071	0.007
385	14933236	0.0039	I		0.420	0.417	0.435	0.228	0.340	0.001
386	14948045	0.0039	I		0.441	0.430	0.445	0.397	0.378	0.000
387	14965085	0.0005	I		0.096	0.093	0.060	0.111	0.069	0.704 *
388	15005477	0.0005	I		0.208	0.205	0.187	0.110	0.065	0.355 *
389	15057214	0.0005	I		0.287	0.285	0.279	n.a.	n.a.	n.a.
391	15087602	0.0012	I		0.335	0.337	0.348	n.a.	n.a.	n.a.
390	15117024	0.0066	I		0.696	0.694	0.703	n.a.	n.a.	n.a.
392	15148546	0.0028	I		0.412	0.423	0.433	0.171	0.045 *	0.005
534	15152610	0.0016	I		0.899	0.900	0.904	0.198	0.164	0.005
393	15169667	0.0000	I		0.118	0.118	0.096	0.076	0.036 *	0.505 *
535	15198298	0.0059	I		n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
394	15211531	0.0088			0.623	0.630	0.640	0.485	0.546	0.000
282	15282518	0.0019			0.193	0.191	0.182	0.282	0.209	0.003
285	15467669	0.0136			0.629	0.622	0.637	n.a.	n.a.	n.a.
286	15503689	0.0111			0.478	0.472	0.488	0.256	0.436	0.004

(Continues...)

Table B3. *(Cont.)*

Locus	Position	θ	CR-II^{all}	CR-III^{all}	Q^{I}	Q^{IIs}	Q^{IIall}	$Q^{E\text{-}IIIs}$	$Q^{E\text{-}IIIall}$	*PP*
287	15525959	0.0058			0.642	0.642	0.656	0.488	0.477	0.000
288	15536600	0.0026			0.405	0.394	0.399	0.165	0.155	0.009
295	15544479	0.0011		J	0.356	0.359	0.356	0.124	0.096	0.104
296	15599496	0.0027		J	0.700	0.697	0.715	0.259	0.147	0.001
297	15638469	0.0026		J	0.766	0.766	0.772	0.557	0.541	0.000
298	15692843	0.0025		J	0.476	0.469	0.485	0.103	0.143	0.001
299	15730456	0.0032		J	0.506	0.513	0.524	0.361	0.268	0.001
294	15746063	0.0037		J	0.624	0.617	0.631	0.396	0.321	0.000
301	15798732	0.0035		J	0.362	0.363	0.367	0.163	0.051	0.006
304	15906320	0.0053		J	0.528	0.520	0.541	0.240	0.072	0.007
725	15964533	0.0040		J	0.840	0.838	0.847	0.607	0.502	0.000
307	16047623	0.0046		J	0.394	0.373	0.398	0.518	0.263	0.004
726	16096635	0.0034		J	0.651	0.645	0.658	0.105	0.044 *	0.006
310	16152786	0.0000		J	0.195	0.190	0.175	0.056	0.022 *	0.357 *
336	16168466	0.0017		J	0.647	0.640	0.652	0.384	0.274	0.000
334	16255530	0.0037		J	0.733	0.734	0.745	0.361	0.340	0.000
333	16276272	0.0040		J	0.706	0.706	0.711	0.271	0.132	0.002
451	16302764	0.0015		J	0.469	0.454	0.470	0.205	0.142	0.002
331	16349660	0.0045		J	0.469	0.467	0.482	0.471	0.268	0.002
330	16371094	0.0121		J	0.656	0.651	0.669	0.456	0.359	0.003
329	16429444	0.0053		J	0.854	0.850	0.856	0.230	0.065	0.006
328	16468799	0.0000		J	0.154	0.154	0.136	0.087	0.025 *	0.419 *
366	16471181	0.0087			0.574	0.563	0.577	0.251	0.088	0.008
364	16530195	0.0075			0.805	0.795	0.807	0.492	0.273	0.002
363	16550618	0.0049			0.534	0.525	0.542	0.864	0.770	0.000
359	16694289	0.0064			0.623	0.624	0.628	0.660	0.606	0.000
402	16783883	0.0049			0.632	0.631	0.653	0.691	0.680	0.000
405	16864287	0.0046			0.501	0.492	0.512	0.605	0.652	0.000
406	16907918	0.0065			0.712	0.702	0.715	0.510	0.456	0.000
407	16933856	0.0000			0.020 *	0.018 *	0.006 *	0.054	0.017 *	0.922 *
410	17027142	0.0113			0.816	0.810	0.812	0.612	0.593	0.000
411	17058256	0.0039			0.420	0.420	0.423	0.117	0.126	0.003
422	17088628	0.0104			0.749	0.752	0.767	0.592	0.677	0.000
727	17161040	0.0045			0.573	0.551	0.577	0.187	0.198	0.002
424	17229023	0.0029			0.522	0.522	0.530	0.312	0.279	0.001
728	17301040	0.0000			0.312	0.304	0.305	0.060	0.036 *	0.240

(Continues...)

Table B3. *(Cont.)*

Locus	Position	θ	CR-II^{all}	CR-III^{all}	Q^{I}	Q^{IIs}	Q^{IIall}	$Q^{E\text{-}IIIs}$	$Q^{E\text{-}IIIall}$	*PP*
426	17353968	0.0023			0.370	0.368	0.382	0.790	0.628	0.002
428	17384694	0.0069			0.755	0.752	0.759	0.832	0.942	0.000
729	17442422	0.0067			0.736	0.728	0.742	0.415	0.325	0.002
430	17492612	0.0005			0.184	0.180	0.167	0.097	0.073	0.426 *
730	17541034	0.0018			0.379	0.374	0.371	n.a.	n.a.	n.a.
431	17619496	0.0041			0.972	0.970	0.973	0.864	0.785	0.000
432	17662415	0.0006			0.234	0.230	0.213	0.195	0.157	0.337 *
436	17980460	0.0078			0.680	0.674	0.704	0.452	0.224	0.013
438	18064811	0.0052			0.556	0.554	0.563	0.898	0.900	0.001
439	18132877	0.0010			0.520	0.504	0.521	0.153	0.077	0.003
440	18201747	0.0047			0.589	0.583	0.606	0.698	0.550	0.003
444	18579133	0.0072			0.616	0.613	0.632	0.501	0.828	0.001

Appendix C. Methods

DNA extraction and isolation from 10-15 Drosophila

Protocol of the PUREGENE DNA Isolation Kit (Gentra Systems, Minneapolis, MN).

Cell lysis:

1) Chill on ice a 1.5 ml centrifuge tube containing 300 µl of Cell Lysis Solution on ice.
2) Add 10–15 flies (5–15 mg) to the chilled Cell Lysis Solution, remove from ice, and homogenize thoroughly using a disposable pestle. Place sample back on ice until next step.
3) Incubate lysate at 65 °C for 15 minutes.

RNase treatment:

4) Add 1.5 µl RNase "A" Solution (4 mg/ml) to the cell lysate.
5) Mix the sample by inverting the tube 25 times and incubate at 37 °C for 15 minutes.

Protein precipitation:

6) Cool sample to room temperature.
7) Add 100 µl of Protein Precipitation Solution to the cell lysate.
8) Vortex vigorously at high speed for 20 seconds to mix the Protein Precipitation Solution uniformly with the cell lysate.
9) Centrifuge at 13,000–16,000 rpm for 3 minutes. The precipitated proteins and tissue particulates will form a tight pellet. If protein pellet is not tight, repeat step 8 followed by incubation on ice for 5 minutes, then repeat step 9.

DNA precipitation:

10) Pour the supernatant containing the DNA (leaving behind the precipitated protein pellet) into a clean 1.5 ml centrifuge tube containing 300 µl of 100% Isopropanol (2–propanol).
11) Mix the sample by inverting gently 50 times.
12) Centrifuge at 13,000–16,000 rpm for 1 minute.
13) Pour off supernatant and drain tube on clean absorbent paper. Add 300 µl of 70% ethanol and invert tube several times to wash the DNA pellet.
14) Centrifuge at 13,000–16,000 rpm for 1 minute. Carefully pour off the ethanol. Pellet may be loose so pour slowly and watch pellet.
15) Invert and drain the tube on clean absorbent paper and allow to dry at room

temperature for 15 minutes.

DNA hydration:

16) Add 50 µl of DNA Hydration Solution.
17) Allow DNA to rehydrate overnight at room temperature. Alternatively, heat at 65 °C for 1 hour. Tap tube periodically to aid in dispersing the DNA.
18) If particulates are present in the rehydrated DNA sample, centrifuge at 13,000–16,000 rpm for 5–10 minutes and then transfer the supernatant containing the DNA to a clean tube.
19) Store DNA at 2–8 °C (or at –20 °C if not used rapidly).

Standard Polymerase-Chain-Reaction (PCR)

Indicated are volumes and, in parentheses, the concentration; the final volume is of 25 µl.

Distilled Water	16.12 µl
Buffer (10x)	2.50 µl
Magnesium (2nM)	1 µl
dNTP's (0.2 mM of each dNTP)	0.25 µl
Taq-polymerase (5 U/l)	0.13 µl
Forward Primer (10 µM)	2 µl
Reverse Primer (10 µM)	2 µl
DNA Template	1 µl

PCR-run standard program:

Indicated are the temperature and the duration in minutes of each step.

1)	initial denaturation	94 °C	4 minutes
2)	30 amplification cycles, each consisting of:		
	denaturation	94 °C	30 seconds
	annealing	X °C	30 seconds
		X is specific to each primers pair (usually 50–58 °C)	
	extension	72 °C	30 seconds
3)	final extension	72 °C	4 minutes
4)	hold	4 °C	at least 4 minutes

PCR fragments are scored on 1.5% agars gel, and then stored at –20 °C.

PCR purification

PCR product	10 µl
EXOSAP-IT (USB, Cleveland, OH)	1 µl

The reaction takes place at:

1) 37 °C 30 minutes
2) 80 °C 15 minutes
3) 4 °C until storage at –20 °C.

Sequencing reaction and program

This reaction has to be done separately for forward and reverse primers (*i.e.*, for both strands). Protocol of the DYEnamic ET terminator cycle sequencing kit (Amersham Biosciences, Buckinghamshire, UK).

Primer	2 µl
Distilled water	3 µl
Sequencing Mix	4 µl
Purified PCR product	1 µl

Indicated are the temperature and the duration of each reaction step.

1) 26 amplification cycles, each consisting of:
   - denaturation 95 °C 20 seconds
   - annealing 50 °C 15 seconds
   - extension 60 °C 60 seconds
2) hold 4 °C at least 4 minutes
3) Store at –20 °C.

The sequencing product has to be cleaned before proceeding to the run on the sequencer.

The protocol refers to a 96-well plate. Unless stated, volumes to add are for a single well.

1) In a 1.5 ml tube, prepare a solution containing 960 µl of distilled water and 192 µl of Sodium-acetate/EDTA (1/10 Vol.).
2) Add 12 µl of the solution in each well.
3) Add 80 µl of 96% ethanol.
4) Cover the plate with the appropriate adhesive aluminum foil and then vortex.
5) Centrifuge the plate for 30 minutes at 3000 rpm.
6) Pour the supernatant.

7) Short inverted spin for ~30 seconds at 300 rpm (put towel paper to absorb supernatant).
8) Rinse with 150 µl of 70% ethanol.
9) Centrifuge 10 minutes at 3000 rpm.
10) Pour the supernatant.
11) Short inverted spin for ~30 seconds at 300 rpm (put towel paper to absorb supernatant).
12) Let the ethanol evaporate by leaving the plate at room temperature for 5–15 minutes (or until completely dry).
13) The plate can be stored at –20 °C before being run in the sequencer.
14) Before run in the sequencer, add 15 µl of distilled water.
15) Vortex to elute the DNA pellet.
16) Briefly centrifuge.

Sequences have been run on a MegaBACE 1000 automated capillary sequencer (Amersham Biosciences, Buckinghamshire, UK). Analysis of the raw data was done using the software Cimarron 3.12 (Amersham Biosciences, Buckinghamshire, UK) for lane tracking and base calling.

Extraction of total RNA

This protocol is adapted from the Invitrogen Life Technologies Trizol manual, and is available at the Drosophila Genomics Resource Center website (http://dgrc.cgb.indiana.edu/microarrays/downloads.html).

1) Flies must be snap frozen in liquid nitrogen, stored at –80 °C and not allowed to thaw prior to being processed.
2) To 50 mg frozen flies (~30 individuals) in a 1.5 ml microcentrifuge tube add 1 ml Trizol reagent and homogenize immediately with a disposable plastic pestle. Work quickly to avoid RNA degradation.
3) Incubate at room temperature for 5 minutes.
4) Centrifuge at 12,000 rpm for 10 minutes at 4 °C to pellet insoluble debris such as exoskeleton.
5) Transfer the supernatant to a new microcentrifuge tube, taking great care not to take pellet or fat layer.
6) Add 200 µl of Chloroform to each tube.
7) Shake vigorously by hand (do not vortex).
8) Incubate tubes at room temperature for 3 minutes.

9) Centrifuge at 10,000 rpm for 15 minutes at 4 °C.
10) Transfer upper aqueous phase (~0.6 ml) to a fresh RNAse-free microcentrifuge tube.
11) Add 0.5 ml isopropanol
12) Incubate at room temperature for 10 minutes.
13) Centrifuge at 12,000 rpm for 10 minutes at 4 °C.
14) Remove the supernatant and wash the pellet with 1 ml 75% ethanol.
15) Centrifuge at 7,500 rpm for 5 minutes at 4 °C.
16) Remove the supernatant.
17) Centrifuge briefly and carefully remove the last of the supernatant with a micropipette.
18) Air dry for 10 minutes.
19) Resuspend the pellet in 100 µl RNase-free water.
20) Quantify a 1/100 dilution of the RNA on spectrophotometer.
21) Store at –20 °C.

cDNA sythesis

Protocol of the ThermoScript reverse transcriptase (Invitrogen, Carlsbad, CA).

1) Add the following components to a nuclease-free microcentrifuge tube:

Random primers (3 µg/µl)	1 µl
Total RNA	x µl, as to get 5 µg of material.
dNTP mix (10 mM)	2 µl
Distilled water	to 12 µl

2) Incubate mixture at 65 °C for 5 minutes and then place on ice. Collect the contents of the tube by brief centrifugation and add:

cDNA synthesis buffer (5X)	4 µl
DTT (0.1 M)	1 µl
Distilled water	1 µl
ThermoScript RT (15 U/ml)	1 µl

3) Incubate tube at 25 °C for 10 minutes.
4) Mix gently and incubate at 50 °C for 30–60 minutes.
5) Terminate the reaction by heating at 85 °C for 5 minutes.
6) Store at –20 °C.

Relative quantification of gene expression with Reverse-Transcription (RT) Real-Time PCR

The protocol is optimized for use of the TaqMan gene expression assays using the Applied Biosystem 7500 Fast Real-Time PCR system (Applied Biosystems, Foster City, CA).

PCR reaction mix components (20 µl reactions):

TaqMan gene expression assay (20X)	1 µl
cDNA template (~100 µg)	9 µl
TaqMan Universal master mix with AmpErase UNG	10 µl

Thermal cycling conditions (two-step RT-PCR; temperatures and times):

1)	Polymerase activation	95 °C	10 minutes
2)	40 amplification cycles, each consisting of:		
	melting	95 °C	15 seconds
	annealing/extension	60 °C	60 seconds

Analysis of polytenic chromosomal inversions

1) Prepare a clean slide with 2–3 drops of 0.7% NaCl under a stereoscopic (dissecting) microscope.
2) Select a large Drosophila larva and place it on the slide.
3) Extract the salivary glands and keep them constantly moist with saline solution.
4) Add 2–3 drops of aceto-orcein stain and for 10–15 minutes.
5) Put a coverslip on the slide, on top of the salivary glands, and cover with a paper towel.
6) Place the thumb on the paper towel over the coverslip and press down firmly taking care not to allow the coverslip or slide to slip or move.
7) Examine the squashed salivary glands under the microscope. If the chromosomes are not separated and elongated, repeat step 6.

Computer programs

Much of the analysis presented in this thesis has been done using customized programs written in `C`, `C++` or `perl`. Below, I give a short introduction to these programs (Table C1) and how to use them; unless specified, they have been written by myself. All programs are available upon request.

Estimating the bottleneck parameters (CHAPTER 1.2., SUBSECTIONS 1.2.1.4 and 1.2.2.3.)

A schematic step-by-step methodology is shown in Figure C1. The coalescent simulations have been done using a modified program originally written by Sebastian Ramos-Onsins (RAMOS-ONSINS *et al*. 2004). The changes to the original code were done (i) to simplify the bottleneck model such that it is described only by two parameters, *i.e*., its age, T_b, and strength S_b, and (ii) to calculate, for each locus *i* and simulated genealogy, the Poisson distributed probabilities to harbor the k_i observed segregating sites (see also APPENDIX A). An example of input file is shown in Figure C2. Each input file served for a single combination of bottleneck parameters. To estimate *P* according to method I, I simulated 221 parameters combinations across 17 and 13 values of T_b and S_b, respectively with the program '`lik`' (input files were created using '`makefiles`'; for a brief description of all programs, see Table C1). From the 221 output files, each containing the likelihoods for every locus and simulated genealogy for a given parameters combination (Figure C3), I then calculated the average P^{I} across simulations and loci using '`P-lik`'. For method II, I simulated the lengths of the coalescent tree portions after and before T_b using '`gett`'; then, the binomial distribution was estimated with '`binom`', and finally the average P^{II} across simulations and loci was calculated using '`P-tot`'. For method III, I obtained the lengths of the coalescent tree portions after and before T_b using '`gett`', and then '`lik_twophases`' was used to estimate the average P^{III} across simulations and loci. For all three methods, I then identified the parameters set giving the higher probability and repeated the above procedure with a new set of maximum and minimum T_b and S_b values, until finally pinpointing the maximum-likelihood bottleneck parameters (usually, I performed 3 rounds of simulations, *i.e*. simulating a total of $3 \times 221 = 663$ parameters combinations for each method). Once the maximum-likelihood bottleneck was estimated, I verified the fit to the data by simulating the sample under such scenario and estimating Tajima's *D* and linkage disequilibrium measure

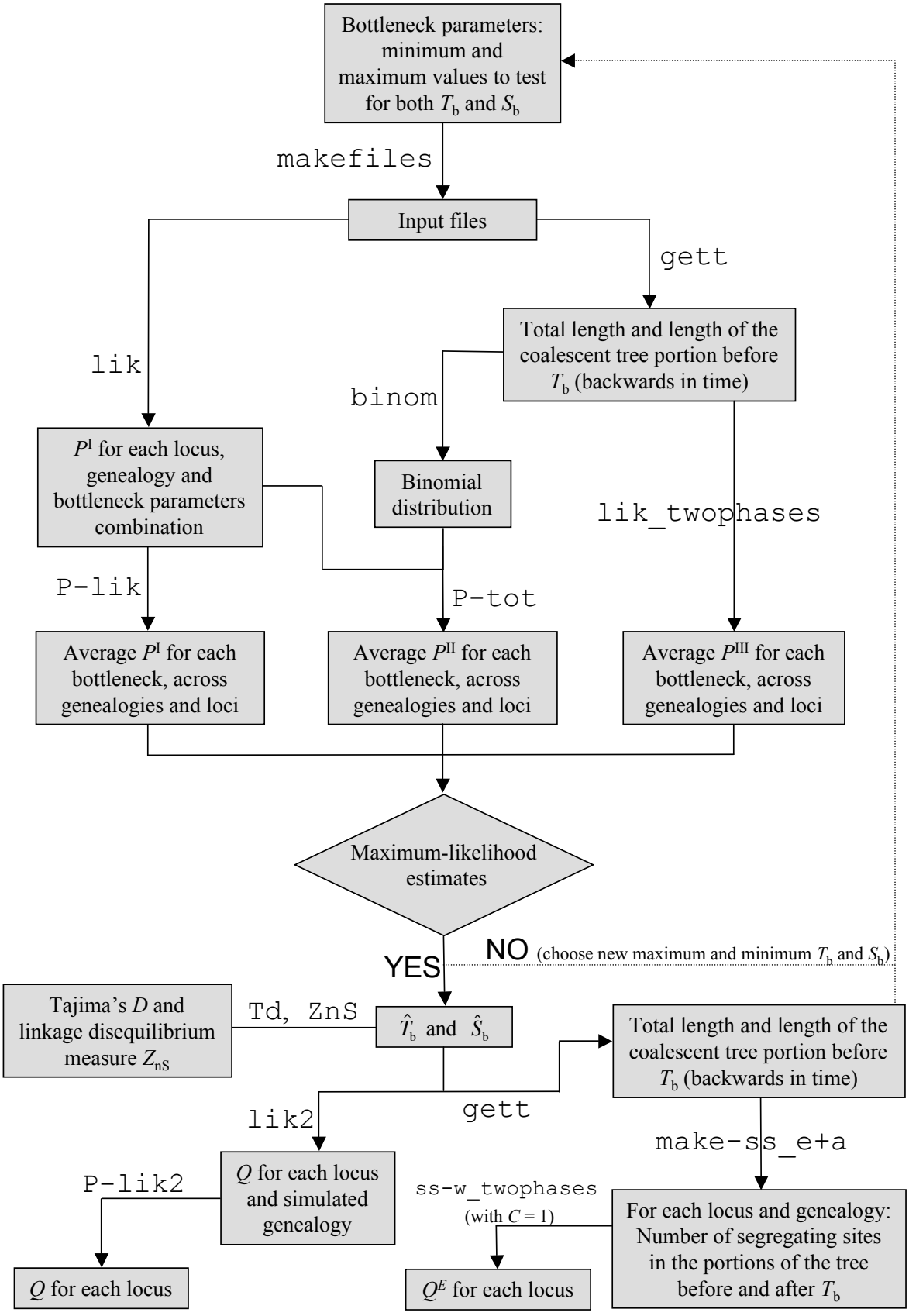


Figure C1. Schematic methodology used to estimate the bottleneck parameters and detect the outlier loci. The name of the computer programs used at each step, and their outputs, are shown. For input parameters and more detailed output, see Table C1.

Z_{nS} using the programs '`td`' and '`ZnS`', respectively.

Detecting the outliers (CHAPTER 1.2., SUBSECTIONS 1.2.1.4 and 1.2.2.4.)

The probabilities Q and Q^E were calculated using the programs (i) '`lik2`' and '`P-lik2`', and (ii) '`gett`', '`make-ss_e+a`' and '`ss-w_twophases`' (with $C = 1$), respectively. In '`make-ss_e+a`', mutations are Poisson distributed based on the mutational pattern and on the length of the coalescent tree portions after and before T_b, obtained using '`gett`' (see also Figure C1).

To detect the groups of C consecutive loci whose polymorphism departed from the bottleneck expectations, I first simulated segregating sites using '`ss`' or '`make-ss_e+a`', and then used a sliding window approach with the programs '`ss-w`' and '`ss-w_twophases`', for methods I-II and method III, respectively.

Testing the polymorphism of a valley of reduced polymorphism against demography (CHAPTER 1.2., SUBSECTIONS 2.1.1.2. and 2.1.2.2.)

I obtained the positions of the simulated mutations across the region using the programs '`positions-r`' and '`positions+r`', depending if recombination within and between loci was neglected or assumed, respectively. Note that in the input file, one has to specify the kind of desired output (see Figure C1).

Then, the programs '`probab-r`' and '`probab+r`' were used to calculate the probability, for the valley of reduced variation, to harbor k segregating sites, given that there are k_{tot} segregating sites across the whole region and the focal locus has k_f segregating sites.

Analysis of the mutational pattern (CHAPTER 3.2.)

In order to polarize the fixed and polymorphic substitutions, to calculate the base composition, and to trim down the alignments to only the "conserved" and "non-conserved" sequences, the files containing the alignments of each locus were parsed using a suite of 17 different `perl` scripts. The import routine had been originally written by R. Piskol and S. Eck (two undergraduate students).

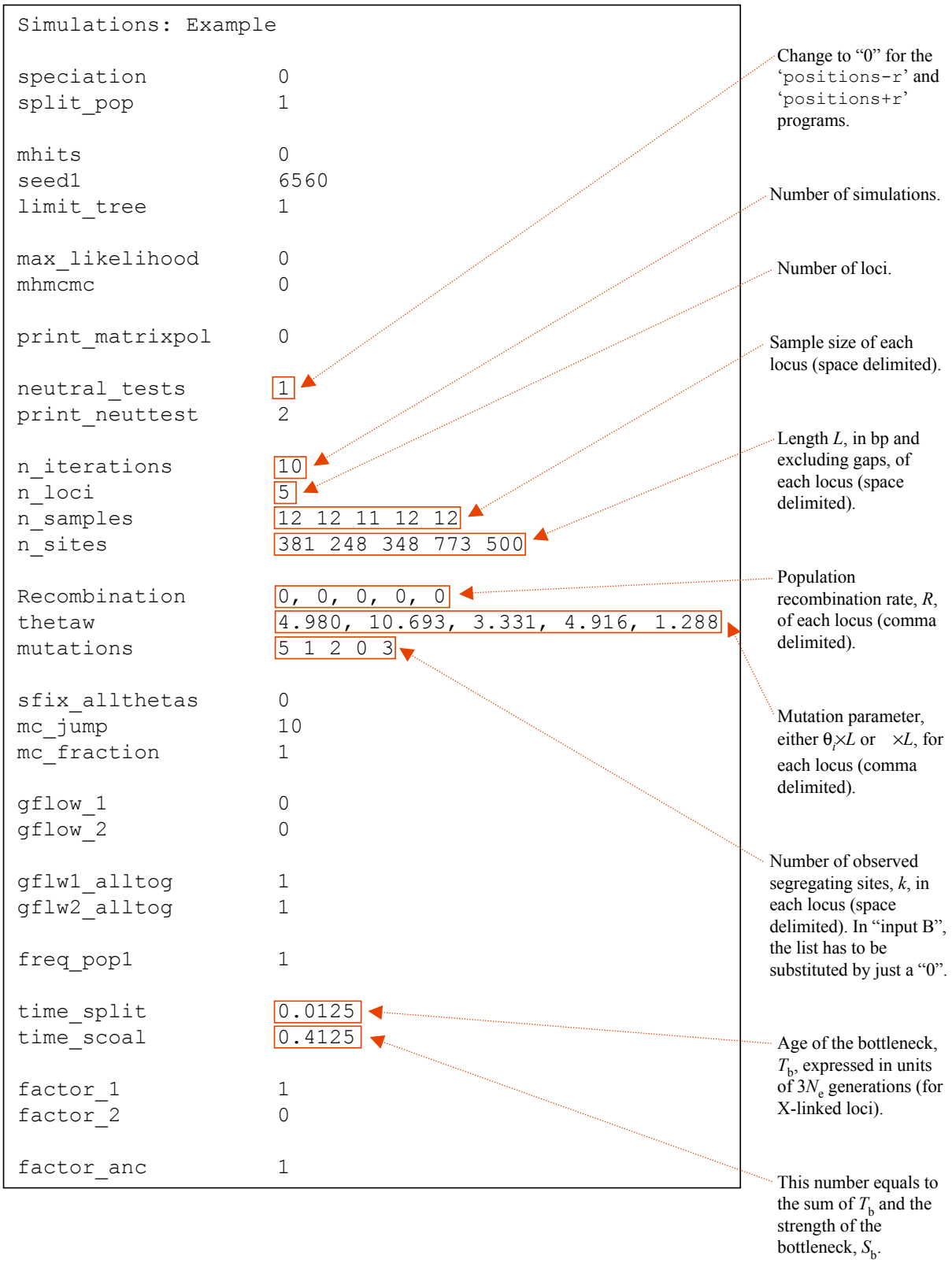


Figure C2. Example of input file for the coalescent simulations. Here, we apply a bottleneck of age T_b = 0.0125 and strength S_b = 0.400 to a sample of 5 loci. The output file is shown in Figure C2.

```
OUTPUT FILE: date Tue Dec 13 15:46:43 2005

Input data from the file: example.txt

Split_pop:              1
Max_likelihood:         0
Mhmcmc:                 0
Neutral_tests:          1
Print_neuttest:         2
Print_matrixpol:        0
Mhits:                  0
N_iterations:           10
Seed1:                  6560
Limit_tree:             1

N_loci:                 5
N_samples:              12 12 11 12 12
Thetaw:                 4.98, 10.693, 3.331, 4.916, 1.288
Mutations:              5 1 2 0 3
N_sites:                381 248 348 773 500

Recombination:          0, 0, 0, 0, 0
Factor_1:               1
Factor_2:               0
Factor_anc:             1
Time_split:             0.0125
Time_scoal:             0.4125
Freq_pop1:              1
Same rate of gflow_1 for every loci
Gflow_1:                0
Same rate of gflow_2 for every loci
Gflow_2:                0



Neutral tests (excluding multiple hits positions)
logP[0]    logP[1]    logP[2]    logP[3]    logP[4]
-1.7790    -8.7898    -1.3282    -2.4481    -1.4970
-2.7110    -10.867    -3.9558    -0.6342    -2.8389
-1.8642    -20.538    -7.6929    -8.3334    -4.5584
-4.3250    -4.5739    -1.4290    -1.4655    -3.5024
-1.8147    -1.1093    -2.8297    -5.6299    -1.9023
-1.7515    -47.521    -1.3648    -6.8311    -2.4692
-1.8510    -26.473    -5.9194    -9.8407    -1.8084
-4.1144    -1.0708    -1.3087    -1.2622    -4.9031
-5.5018    -1.1317    -3.0931    -2.0380    -7.1781
-1.7819    -3.0597    -2.8212    -9.1165    -1.5107
```

Input parameters, as given in the input file (see Figure C1).

Log-likelihoods (P^{I}) values: each column corresponds to a locus, each line to a coalescent simulation.

Figure C3. Example of output file for the coalescent simulations. Here, we applied a bottleneck of age T_b = 0.0125 and strength S_b = 0.400 to a sample of 5 loci (see Figure C1). The output file consists of the log-Likelihood probabilities to harbor *k* segregating sites given the bottleneck, calculated for each locus and simulated genealogy with method I (*i.e.*, P^{I}).

Table C1. Computer programs.

Chapter	Program name	Input	Output	Variable parameters in the code	Language
CHAPTER 1.2.	`binom`	`gett` outputs; k; k^E.	Binomial distribution.	Number of simulated bottlenecks, of genealogies and of loci.	C++
	`gett`	`makefiles` outputs B (Input files B; see Figure C1).	Total length and T_{tree}^A of the simulated coalescent tree, for each locus and genealogy.		C
	`lik`	`makefiles` outputs (Input files A)	P^I for each locus and simulated genealogy.		C
	`lik_twophases`	`gett` outputs; observed k^E and k; θ_i; $\bar{\theta}$; (θ are per locus).	Average P^{III} for each locus and bottleneck, across genealogies.	Number of simulated bottlenecks, of genealogies and of loci.	C++
	`lik2`	Input file B (see Figure C1) with T_b and S_b	Q for each locus and simulated genealogy.		C
	`makefiles`	minimum and maximum values for both T_b and S_b.	221 ready-to-use input files.	Number of T_b and S_b values to explore and all locus-specific parameters (see Figure C1).	C++
	`make-ss_e+a`	`gett` outputs; θ_i; $\bar{\theta}$; (θ are per locus).	Simulated number of segregating sites in the pre- and post- bottleneck phases.	Number of simulated genealogies and of loci.	C++
	`P-lik`	`lik` outputs.	Average P^I for each bottleneck, across genealogies and loci.	Number of simulated bottlenecks, of genealogies and of loci.	C++
	`P-lik2`	`lik2` outputs.	Average Q for each locus, across simulations.	Number of simulated genealogies and of loci.	C++
	`P-tot`	`lik` outputs; `binom` outputs.	Average P^{II} for each bottleneck, across genealogies and loci.	Number of simulated bottlenecks, of genealogies and of loci.	C++
	`ss`	Input file B (see Figure C1) with T_b and S_b.	Simulated number of segregating sites for each bottleneck, genealogy and locus.		C
	`ss-w`	`ss` output file; observed k; loci's positions; loci's identification number.	Average Q^C across simulated samples.	Number of simulated genealogies and of loci.	C++
	`ss-w_twophases`	`make-ss_e+a` output files; observed k^E and k^A; loci's positions; loci's identification number.	Average $Q^{E=C}$ across simulated samples.	Number of simulated genealogies and of loci.	C++
	`td`	Input file B (see Figure C1) with T_b and S_b.	Simulated Tajima's D for each bottleneck, genealogy and locus.		C
	`ZnS`	Input file B (see Figure C1) with T_b and S_b.	Simulated Z_{nS} for each bottleneck, genealogy and locus.		C

(Continues...)

Table C1. *(Cont.)*

Chapter	Program name	Input	Output	Variable parameters in the code	Language
CHAPTER 2.1.	`positions-r`	Input file B (see Figure C1) with T_b and S_b.	Relative position of the simulated mutations within the region; no recombination is assumed.		`C`
	`positions+r`	Input file B (see Figure C1) with T_b and S_b.	Relative position of the simulated mutations within the region; we assume recombination.		`C`
	`probab-r`	`positions-r` output.	Probability to observe k segregating sites in the valley, given that the region has to harbor k_{tot} segregating sites and the focal locus k_f.	Relative start and end of each of the loci within the region (loci are concatenated); number of observed segregating sites in the region and in the focal locus.	`C++`
	`probab+r`	`positions+r` output.	Probability to observe k segregating sites in the valley, given that the region has to harbor k_{tot} segregating sites and the focal locus k_f.	Relative start and end of each of the loci within the region; number of observed segregating sites in the region and in the focal locus.	`C++`
CHAPTER 3.2.	`fixed-i`	A batch of alignments in nexus format.	Ancestral and derived state for each isolated fixed substitution.		`perl`
	`fixed-t`	A batch of alignments in nexus format.	Ancestral and derived state for each fixed substitution; base composition of each line.		`perl`
	`mel-del`	A batch of alignments in nexus format.	Only alignments of deletions segregating in *D. melanogaster*.		`perl`
	`mel-del-fix`	A batch of alignments in nexus format.	Only alignment parts where *D. melanogaster* has fixed deletion.		`perl`
	`mel-ins`	A batch of alignments in nexus format.	Only alignments of insertions segregating in *D. melanogaster*.		`perl`
	`mel-ins-fix`	A batch of alignments in nexus format.	Only alignment parts where *D. melanogaster* has fixed insertion.		`perl`
	`melsim`	A batch of alignments in nexus format.	Monomorphic sites in *D. melanogaster* and *D. simulans*.		`perl`
	`poly-i`	A batch of alignments in nexus format.	Ancestral and derived state for each isolated polymorphism.		`perl`
	`poly-t`	A batch of alignments in nexus format.	Ancestral and derived state for each polymorphism; base composition of each line.		`perl`
	`sg`	A batch of alignments in nexus format.	Only alignment parts where *D. simulans* has gaps.		`perl`

(Continues…)

Table C1. *(Cont.)*

Chapter	Program name	Input	Output	Variable parameters in the code	Language
CHAPTER 3.2.	`sim-del-fix`	A batch of alignments in nexus format.	Only alignment parts where *D. simulans* has fixed deletion.		`perl`
	`sim-ins-fix`	A batch of alignments in nexus format.	Only alignment parts where *D. simulans* has fixed insertion.		`perl`
	`sng`	A batch of alignments in nexus format.	Only alignment parts where *D. simulans* has gaps.		`perl`
	`syg`	A batch of alignments in nexus format.	Only alignment parts where D. simulans and/or *D. yakuba* have gaps.		`perl`
	`syng`	A batch of alignments in nexus format.	Only alignment parts where *D. simulans* and/or *D. yakuba* have gaps.		`perl`
	`yg`	A batch of alignments in nexus format.	Only alignment parts where *D. yakuba* has gaps.		`perl`
	`yng`	A batch of alignments in nexus format.	Only alignment parts where *D. yakuba* has gaps.		`perl`

All programs written specifically for this thesis are shown. Input parameters can be either given in the input file (see Figure C1), or directly by the user when running the program. Some other parameters have to be changed directly in the program code.

Curriculum vitae

Date and place of birth: 7 April 1973, Vicenza (Italy)

Nationality: Italian

Marital status: married

Current work address

Department of Biology II
University of Munich L.M.U.
Grosshadernerstrasse, 2
82152 Planegg-Martinsried
Germany
Tel: +49 (0)89 2180 74 101
Fax: +49 (0)89 2180 74 104
Email: ometto@zi.biologie.uni-muenchen.de

Permanent home address

Via Zanella, 44
36043 Camisano Vicentino (Vicenza)
Italy

Education

2002–present: Ph.D. student, University of Munich L.M.U., Germany.
1999: Laurea Degree in Biological Sciences, University of Padova, Italy.
1996–1997: Maîtrise courses, University of Paris XI, Orsay-sur-Yvette, France.
1992: Scientific High School Graduation, Liceo scientifico "G.B. Quadri", Vicenza, Italy.

Additional research experience

2001–2002: Research assistant, University of Munich, Germany. Supervisor: Dr. B. Nürnberger.
2000–2001: Research assistant, Canterbury University, Christchurch, New Zealand. Supervisor: Dr. B. Waldman.
2000: Research assistant, University of Padova, Italy. Supervisor: Prof. A. Minelli.
1998–1999: Research assistant, University of Padova, Italy. Supervisor: Prof. P. Cardellini.

Publications

Ometto, L., S. Glinka, D. De Lorenzo, and W. Stephan, 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Mol. Biol. Evol.* **22:**2119–2130.

Ometto, L., W. Stephan, and D. De Lorenzo, 2005. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* **169:**1521–1527.

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. (* equally contributed to this work.)

Cardellini, P., and L. Ometto, 2001. Teratogenic and toxic effects of Alcohol Ethoxylate and Alcohol Ethoxy Sulfate surfactants on *Xenopus laevis* embryos and tadpoles. *Ecotox. Environ. Safe.* **48:**170–177.

Abstracts to conferences

Ometto, L., S. Glinka, W. Stephan, and D. De Lorenzo, 2005. Demographic and selective history of *Drosophila melanogaster* inferred by a multilocus scan of DNA variation. *In* Abstract Book of the 1st Congress of the Italian Society for Evolutionary Biology, 24–26 August 2005, Ferrara, Italy. (*Oral presentation*.)

Ometto, L., S. Glinka, L. Müller, W. Stephan, and D. De Lorenzo, 2005. Distinguishing demographic and selective footprints in the X chromosome of *Drosophila melanogaster*.

In Abstract Book of the 10th Congress of the European Society for Evolutionary Biology, 15–20 August 2005, Kraków, Poland. (*Poster presentation*.)

De Lorenzo, D., S. Beisswanger, S. Glinka, S. Hutter, L. Ometto, and W. Stephan, 2005. Demographic and selective history of *D. melanogaster*: a genomic survey. *In* Abstract Book of the 10th Congress of the European Society for Evolutionary Biology, 15–20 August 2005, Kraków, Poland.

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. *In* Abstract Book of the 9th Congress of the European Society for Evolutionary Biology, 18–24 August 2003, Leeds, UK. (*Poster presentation*.)

Ometto, L., and P. Cardellini, 2001. Ecotoxicology of surfactants on *Xenopus laevis* embryos and tadpoles. *In* Abstracts of papers presented at the 9th Society for Research on Amphibians and Reptiles in New Zealand Conference, St Arnaud, Nelson Lakes, New Zealand, 2–4 February 2001. *New Zeal. J. Zool.* **28:**361–372. (*Oral presentation*.)

Acknowledgments

I would like to thank Prof. Wolfgang Stephan for giving me the great opportunity to work on population genetics under his supervision. His love for research and his constant flux of ideas, comments and support have been a precious guidance during my Ph.D. studies. A special thank for having helped me to improve considerably my scientific writing.

Muchas gracias to David De Lorenzo, it was a pleasure to collaborate with him and he always provided useful help in scientific, technical and entomological issues. I especially thank him for letting me enjoy Mediterranean atmosphere at these latitudes and for his friendship.

And of course, dankeschön to all the Drosophila-scanners, Sascha Glinka, Stephan Hutter, Steffen Beisswanger, Nicolas Svetec, Lena Müller and Bettina Schirrmeister for the nice intra- and extra flyroom friendship. A special thank to Sascha for having shared with me intense X-linked days; to Stephan for computer troubleshooting and the Eins-Zwo-Drei… Suffa; and to Steffen for letting me parasite his knowledge on expression analysis.

Huge thanks also to John Parsch, Joachim Hermisson, Pleuni Pennings, Haipeng Li, John Baines, Laura Rose and Peter Pfaffelhuber for their help during various phases of my research. I am particularly indebted to Sebastian Ramos-Onsins and Sylvain Mousset, who made my maximum-likelihood approach becoming real. Of course, life in the lab would have been boring without the lively chats and meals with all of them and with (in order of appearance) Thomas Städler, Kerstin Roselius, Aparup Das, Ying Chen, Tina Hambuch, Ann Arunyawat, Nina Stoletzki, Winfried Hense, Zhi Zhang, Aura Navarro-Quezada, José Álvarez-Castro, Matthias Pröschel, Daven Presgraves, Michael Kopp, Sarah Peter and Lukasz Grzeskowiak: thank you! I must also thank Sonja Köhler, Thomas Alfert, Beate Nürnberger and numerous Bombina tadpoles for having lured me into Germany.

Liters and kilos and plates of thanks to Anne Wilken, Gabi Büttner, Kawsar Bhuyan and Traudl Feldmaier-Fuchs for invaluable help in the lab and in the fly-business, and for patient German courses: so yes, I should have said vielen tausend danke!

This thesis was sponsored by unconditional support from Elisa and my families: grazie mamma, papà, Piero, Marta, Giuliana, Gregorio and Tomaso.

Finally, thanks to flora & fauna: keep evolving!