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

Selection and population structure in *Drosophila melanogaster*

Steffen Beißwanger
aus München

2006

Erklärung

Diese Dissertation wurde im Sinne von § 12 der Promotionsordnung von Herrn Prof. Dr. Wolfgang Stephan betreut. Ich erkläre hiermit, dass die Dissertation nicht einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich anderweitig einer Doktorprüfung ohne Erfolg nicht unterzogen habe.

Ehrenwörtliche Versicherung

Ich versichere hiermit ehrenwörtlich, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt wurde.

München, 12.10.2006

Steffen Beißwanger

Dissertation eingereicht am: 12.10.2006

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

Mündliche Prüfung am: 05.12.2006

Table of Contents

Summary

In this thesis I scrutinized a specific region of the X chromosome of *Drosophila melanogaster* for evidence of positive directional selection. In addition, I analyzed the structure of six Southeast (SE) Asian populations of this species.

In the first chapter, I analyzed a region that showed no polymorphism in a previous scan of the X chromosome in a European *D. melanogaster* population. This region, which I named the *wapl* region, is located on the distal part of the X chromosome, in cytological division 2C10 – 2E1. I observed a 60.5 – kb stretch of DNA encompassing the genes *ph-d*, *ph-p*, *CG3835*, *bcn92*, *Pgd*, *wapl* and *Cyp4d1* that almost completely lacks variation in the European sample. Loci flanking this region show a skewed frequency spectrum at segregating sites, strong haplotype structure, and high levels of linkage disequilibrium. Neutrality tests revealed that these patterns of variation are unlikely under the neutral equilibrium model or simple bottleneck scenarios. In contrast, newly developed likelihood ratio tests suggest that strong positive selection has acted recently on the region under investigation, resulting in a selective sweep. Evidence is presented that this sweep may have originated in an ancestral population in Africa.

In the second chapter, I revisited the center of the *wapl* region analyzed in chapter 1. I concentrated on the African *D. melanogaster* sample, as the valley of reduced variation found in the previous study was much narrower in the African sample than in the European one, which should help to pinpoint the target of selection. About 80% of the valley of reduced nucleotide variation was sequenced. This valley is located between the genes *ph-d* and *Pgd*. I therefore termed this part the *ph-d* – *Pgd* region. The new results confirm previous conclusions about selection having shaped nucleotide variability in this part of the *D. melanogaster* genome. Moreover, by sequencing the center of the selective sweep I was able to establish the

haplotype structure in that region and to infer the historical context of the sweep. Most likely a positively selected substitution occurred at *ph-p* and was fixed before the out-of-Africa expansion of *D. melanogaster*, possibly >30,000 years ago. This substitution might be associated with the specialization of *ph-p* in gene regulation. In addition, the results obtained from the European sample indicate that sequence variation was not affected by demography alone. In fact, it was found that selection affected nucleotide diversity in the *ph-d* – *Pgd* region of the European sample as well. Since heterozygosity across the whole *wapl* region is substantially reduced, I propose that an additional selective sweep occurred at a different site in the European population. This is supported by an analysis regarding the time since the fixation of the (first) beneficial mutation at *ph-p*, which points toward a substitution in *D. melanogaster* before the colonization of Europe.

In chapter 3, I obtained sequence data from six SE Asian samples for ten putatively neutrally evolving X-linked loci. Population genetic parameters were estimated and compared to those previously obtained from the European and the African sample. I observe substantially lower levels of nucleotide diversity in SE Asia than in either Africa or Europe. In particular, samples taken from more peripheral populations (*e.g.* Manila and Cebu, located on the Philippines) show a paucity of haplotypes. Common summary statistics indicate that genetic drift had a significant impact on these populations, which also led to considerable population substructure. One sample, *i.e.* Kuala Lumpur, however, shows rather high levels of heterozygosity among all SE Asian samples and is on average least differentiated from these. This indicates that the Kuala Lumpur population is ancestral to the other SE Asian populations, which is supported by a high amount of shared polymorphic sites. Finally, I revisited the *wapl* region, as analyzed in the first chapter, and find evidence that the selective sweep is older in Kuala Lumpur than in Europe.

Preface

The research I present in this thesis was done by myself, except for the following: the source code for the bottleneck simulations mentioned in chapter 1 was provided by Lino Ometto. The demographic models mentioned in chapter 3 were computed and tested by Haipeng Li. Claus Vogl run the simulations for the analysis regarding the differentation from the migrant pool. For the analysis of shared and private polymorphic sites, I used an algorithm developed with the help of S. Donhauser.

The results from my thesis have contributed to the following publication:

BEISSWANGER, S., W. STEPHAN and D. DE LORENZO, 2006 Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. Genetics **172:** 265-274.

Introduction

Evolution is a fact. However, evolutionary theory developed only slowly over the centuries. The first thoughts about where humans came from and where they are going after the death can be dated back to more than 10,000 years ago (RIEDL 2003). Probably the Neanderthals buried 50,000 ya together with burial objects already dealt with the conundrum of life and death. Yet, in the pre-Christian era, afterlife was mystified in many cultures and mythical creatures were thought to have created all life. A biological view of life and first signs of understanding phylogenetic relationships can be attributed to Aristotle (384 – 322 B.C.E.), who also stated that humans are the result of a principle in nature (RIEDL 2003). In contrast, his tutor Platon was convinced that all things on earth were predetermined. Furthermore, a first perception of selection (in the biological sense) can already be found with Lucrecius (~96 – 55 B.C.E.) in his opus *De natura rerum*, where adaptation was anticipated as well (*ibid.*). Later, in Christian Rome, however, theological belief was imposed and scientific intuition and awareness were at disadvantage. Artists and scientists that did not conform to the zeitgeist had to suffer restrictions. An important major change took place during the age of enlightenment (~17th – 18th century), where liberal ideas were no longer prosecuted and new perspectives became accepted in science. The theory of descent was carried on and further developed by Jean-Baptiste Lamarck (1744 – 1829).

According to ecclesiastic dogmatic not more than ~4,000 years have passed since the creation of earth. However, according to historical traditions many species have already existed for more than 2,000 years without any morphological change. It was therefore not understandable how species emanated from one another in just a little while (*ibid.*). Lamarck and later Erasmus Darwin (1731 – 1802) addressed the

problem of timeline in their works, as can be seen in the following passage of LAMARCK'S "Philosophie Zoologique" (1809):

> *"Mußte ich nicht annehmen, daß die Natur die verschiedenen Organismen nacheinander hervorgebracht habe, fortschreitend von den einfachen bis zu den kompliziertesten, da sich die Organisation in der tierischen Stufenleiter, von den unvollkommensten Tieren an, stufenweise in einer äußerst merkwürdigen Weise kompliziert?"* (aus RIEDL 2003)

Lamarck also proposed an active mode of inheritance (*e.g.* by intensive use of a certain organ or structure), which has been disapproved later. Nevertheless, he correctly addressed the issues of descent and natural variation.

Later, major contributions to evolutionary theory were, of course, made by Charles Darwin (1809 – 1882) and Alfred Russel Wallace (1823 – 1913), who independently from each other gained, but also shared, insights into evolutionary processes. These insights were significantly influenced by world-tours, like Darwin's voyage on the Beagle or Wallace's trip to the Amazonas (see *e.g.* DARWIN 1839, DARWIN 1859). Yet, Darwin was a partial Lamarckian, who recognized the importance of natural variation among individuals (DARWIN 1859). However, it is not entirely clear whether he finally shared Lamarck's view of active acquisition of useful characters (RIEDL 2003) and inheritance of these ("Pangenesis – Theory") or rather abandoned this theory (SCHMITZ 1983). His idea of selection of the fittest individual combined with heritability was appreciated among scientists, heavily rebuted in Victorian England, but eventually promoted by Thomas Huxley (1825 – 1895). Finally, it was Wallace who established Darwinism, although he neglected the important connection between natural variation and heritability (RIEDL 2003).

Ernst Haeckel (1834 – 1919) was the first who established a connection between ontogeny, *i.e.* the history of development of an individual, and phylogeny, *i.e.* the history of development of organic clades. Though his "Biogenetic Fundamental Law", which states that ontogeny recapitulates phylogeny, has been criticized, he recognized this evolutionary relevant relationship for the first time. In addition he contributed the basics to cognitive science by saying that our mind (*i.e.*

psyche) developed from our next closest relative's mind (*ibd.*). This was put forward by Konrad Lorenz, the founder of cognitive science, in the 1960s and 1970s.

Well into the 20th century many facts that verified the theory of descent have accumulated. Evidence for this theory has been found from many fields in biology, including systematics, comparative anatomy (*e.g.* homology *vs.* analogy), palaeontology (fossil record) or embryology. The latter was mainly worked on by Haeckel. The final success of evolutionary theory can, presumably, also be attributed to the *fin du siècle* mood in early 20th century Europe, the beginning of industrialism (period of promoterism) and the continuing effect of Kant's enlightenment.

Interestingly, with the exception of Mendel (1822 – 1884), genetics was more or less absent in evolutionary biology (RIEDL 2003). The most plausible form of inheritance was provided by Darwin's Pangenesis – Theory, which proposed that particles of any structure flow around in the body, are transmitted with the gametes and form new structures in the next generation. The term "genetics" was established by William Bateson (1861 – 1929) and "mutations" were first mentioned by Hugo de Vries (1848 – 1935) in his publication on mutation theory (DE VRIES 1901). The discussion about heritability of acquired characters was finally put to an end by August Weismann (1834 – 1914), a contemporary of Heckel, who pointed out that somatic cells have no influence on gametes (WEISMANN 1885). Acquired characters can therefore be not passed on to the next generation. Another theorem states that the flow of information starts with RNA and, *via* amino acids, ends with proteins and cannot be inverted (CRICK 1958). This was essential in the understanding of molecular genetics and, in consequence, for understanding evolutionary processes.

In the 1940ies the synthetic theory (mainly led by Ernst Mayr, George G. Simpson and Theodosius Dobzhansky) unified systematics, palaeontology and genetics, respectively (RIEDL 2003). In light of this, population dynamics became a subject matter in genetics. The amount of genetic variation, the gene pool, became focus of research (*e.g.* of J.B.S. Haldane, Sewall Wright and R.A. Fisher), since it had to be understood how genetic drift, *i.e.* the stochastic fluctuation of mutations in a population, was acting (*ibid.*).

Genetic drift can have substantial impact on populations, since it removes genetic variation. However, lost variation can be restored by new mutations. But here

again genetic drift interferes, as it affects the probability of survival of new mutations (GILLESPIE 2004). In 1968, Motoo Kimura (1924 – 1994) proposed the "Neutral Theory" of molecular evolution (KIMURA 1968), stating that *"Calculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations involved must be neutral ones"*. This had considerable consequences, since it implied that the interplay of mutation and genetic drift, a stochastic process, is the governing force in evolution, as opposed to Darwinian selection. Likewise, it explains the high level of heterozygosity, the amount of genetic variation, observed in noncoding regions of the genome of a given species, *e.g.* Drosophila. Subsequently, the "standard neutral model" (KIMURA 1983) was established, which assumes that the vast majority of mutations occur at new sites ("infinite sites model", see KIMURA 1971), that they are selectively neutral and that alleles are sampled from a panmictic population at equilibrium. Nowadays, this model often serves as the null model, against which other models (*e.g.* one assuming directional selection) are tested.

As molecular techniques became available, the genomes of many model organisms were systematically screened for signatures of positive selection. The fruitfly, *Drosophila melanogaster* is one such model organism. It is widely used in many areas of biological research as it is easy to keep in the laboratory and has a short generation time of approx. 10 to 14 days. In addition, this species is a human commensal and therefore easy to find everywhere. This enables researchers to study Drosophila behavior, population structure and demography and selective history.

In 1992, David Begun and Charles Aquadro reported that levels of naturally occuring DNA polymorphism correlated with recombination rates in *Drosophila melanogaster* (BEGUN and AQUADRO 1992). According to the neutral theory, this can be explained by an increased mutation rate in regions of high recombination, which should also be reflected in divergence to another species. However, divergence between *D. melanogaster* and its sibling species *D. simulans* was not elevated in regions of high recombination (BEGUN and AQUADRO 1992). They therefore excluded that potential mutagenic effects of recombination are a salient explanation for the observed correlation and, instead, suggested that genetic hitchhiking in regions of restricted recombination caused the observed positive correlation of nucleotide diversity and recombination.

Briefly, the hitchhiking model (MAYNARD SMITH and HAIGH 1974) states that neutral variants, located in the vicinity of a positively selected mutation, can rise in frequency as well, given that they are located on the same chromosome. Eventually they will go to fixation together with the beneficial mutation. Therefore, the region close to the selected site will be devoid of polymorphism. The magnitude of such a valley of reduced heterozygosity depends on the local rate of recombination and on the strength of selection (KAPLAN *et al.* 1989, STEPHAN *et al.* 1992). Consequently, in regions of reduced recombination a very advantagous mutation is supposed to leave a large footprint of selection, since linked neutral variants do not have a great opportunity to escape the selective sweep. In contrast, the signature of selection is assumed to be smaller in regions of normal to high recombination, as the genealogical histories of the selected and a given neutral site are likely to be uncorrelated.

However, an alternative model to explain observed reductions in genetic variability was proposed by Charlesworth *et al.* (1993, 1995). In contrast to the hitchhiking approach, their theory proposes that neutral variants get lost when they are linked to deleterious mutations that are eliminated from the population. Therefore, if the local recombination rate is sufficiently low, a neutral variant cannot escape the removal process. Here as well the strength of selection against deleterious mutations and the recombination rate are important (GORDO and CHARLESWORTH 2001). Yet, for a recombining region the resulting pattern of nucleotide variation under background selection is distinct from the one expected under hitchhiking. For instance, there is no significant skew towards low frequency variants in large populations, as neutral variants can recombine with a mutation free background (CHARLESWORTH *et al.* 1995). Therefore, background selection is not easily detectable using common summary statistics such as TAJIMA'S (1989) *D* or FU and LI'S (1993) *D* (see below). It has been argued that background selection is more effective on autosomes, since deleterious mutations can accumulate and reach considerable frequencies (CHARLESWORTH *et al.* 1995). Furthermore, it has been shown that the hitchhiking model is more likely to explain nucleotide variability in regions of low recombination of *D. melanogaster* (ANDOLFATTO and PRZEWORSKI 2001, INNAN and STEPHAN 2003).

As mentioned before, the standard neutral model is widely used in testing certain hypotheses, *e.g.* that positive selection has acted on a particular genomic region. In the following I will describe the most important neutrality tests, as they are used throughout this thesis. These tests can roughly be grouped in those using intraspecific data only (*i.e.* polymorphism data obtained from DNA sequences) and tests making use of an outgroup, that is they include interspecific comparisons. Among the tests belonging to the first group is TAJIMA'S (1989) *D*. Under the neutral-equilibrium model, the expected nucleotide variation for a diploid is given by $\theta = 4N_e\mu$, where N_e is the effective population size and θ is the mutation rate (KIMURA and CROW 1964). The *D* statistic compares two parameter estimates of the population mutation rate, *i.e.* π (TAJIMA 1983), an estimator based on the average number of pairwise nucleotide differences, and θ_W (WATTERSON 1975), based on the number of segregating sites in a sample. Under neutrality, the difference between π and θ is expected to be zero. However, if a selective sweep completely removes nucleotide variation in the vicinity of a beneficial mutation, then mutations that arise subsequently to the sweep will initially be at very low frequency. In fact, after recent hitchhiking most SNPs will be singletons, producing a star-like phylogeny. Therefore, Tajima's *D* is likely to be negative. In contrast, in the case of two approx. equally selected SNP variants (*i.e.* balancing selection) *D* will probably be positive since two major haplotypes can be observed in a given sample, thus resulting in a high number of pairwise differences. However, results of the *D* statistic have to be interpreted carefully. The reason is that demographic processes will lead to deviations from neutral-equilibrium conditions. For example, a population size expansion typically results in a negative *D* statistic as well, since many segregating sites will be at low frequency (SLATKIN and HUDSON 1991). Likewise, a population size bottleneck or hidden population structure can give a positive *D*.

Another *D* statistic was developed by Yun-Xin Fu and Wen-Hsiung Li in 1993. This test compares the number of mutations on external branches to the total number of mutations in a genealogy (FU and LI 1993). Consequently, both hitchhiking and population size expansion will result in a negative *D*, as both generate an excess of rare variants. The latter also contribute to a high number of observed haplotypes, which is a salient feature of genetic hitchhiking as well as of

size expansion (FU 1997). In contrast, the signature captured by FAY and WU'S *H* (2000) test are high frequency derived variants. The rationale is that in the presence of recombination, hitchhiking will be incomplete for those neutral variants that recombine onto a different genetic background during the sweep phase. Hence, the frequency of neutral variants flanking the selected site depends on whether they are located on the beneficial allele or not. The *H* statistic uses this frequency information by comparing the difference of π and a modified version of π, which is weighted by the number of derived variants (*i.e.* θ_H, FAY and WU 2000). If a high number of derived variants are detected at a given locus, a signature of a selective sweep, *H* will be negative. For both FAY and WU'S (2000) *H* and FU and LI'S (1993) *D* an outgroup is used to distinguish between ancestral and derived variants. However, no further information from the outgroup is incorporated into these test statistics. This also holds for tests of linkage disequilibrium. As mentioned before, during a selective sweep neutral variants can recombine onto another genetic background. The probability for such a situation increases with physical distance to the selected site (KAPLAN *et al.* 1989, KIM and NIELSEN 2004). Distant loci can therefore feature recombinants and a strong haplotype pattern. The length of a region showing such a pattern depends, yet again, on the recombination rate and the length of the selective phase. If the selective phase is very short (as in the case of strong directional selection) there is little opportunity for recombination to take place. Consequently, stretches that show a peculiar pattern of allelic associations tend to be longer than in the case of a weakly selected site located in a region of high recombination (KIM and NIELSEN 2004). The degree of pairwise allelic association is commonly measured by r^2, the correlation coefficient (HILL and ROBERTSON 1968). For a given locus, one can average over all estimates of r^2 to express linkage disequilibrium in terms of Z_{nS} (KELLY 1997) for the whole sequence.

A well known example of a test including both intraspecific data as well as outgroup information is the Hudson-Kreitman-Aguadé (HKA) test (HUDSON *et al.* 1987). This test compares the ratio of polymorphism to divergence at two or more loci. Thus, this test tries to control for differences in mutation rates between loci that might be caused by differences in the level of selective constraint acting in each locus (KREITMAN 2000). If selection is acting on one locus exclusively, then this marker

will have a severely reduced ratio of polymorphism to divergence and, consequently, the HKA test will give a significant result.

One possibility to test the significance of a result obtained by a particular test statistic is to run coalescent simulations (KINGMAN 1982). The coalescent is a population genetic tool, in which ancestral lineages are traced back trough time. Ancestral lineages are the series of genetic ancestors of the samples at a given locus. The history of a sample of size *n* comprises *n*-1 coalescent events, where each coalescent event decreases the number of ancestral lineages by one. The single lineage remaining at the final coalescent event is called the most recent common ancestor (MRCA) of the entire sample (WAKELEY 2006). The coalescent is therefore a stochastic process, where a sample is taken from the present day when there are *n* lineages through a series of steps in which the number of lineages decreases from *n* to *n*-1, then from *n*-1 to *n*-2, *etc*., then finally from two to one (WAKELEY 2006). The coalescent can be modified to allow for recombination and/or selection (see *e.g.* HUDSON 1990 and ROSENBERG and NORDBORG 2002 for a review). Commonly, a large number of coalescent simulations are run under the standard neutral model and a particular summary statistic is computed after each run. Finally, the value estimated from the observed data is compared to the distribution of the statistic obtained from the simulations. If the observed value comes to lie outside of the confidence interval of choice, it is considered significantly different from standard neutral expectations.

As noted above, results obtained from neutrality tests have to be considered carefully and several aspects of the data must be discussed together to argue in favour of directional selection. The knowledge of the demographic history of a given population can help to avoid misinterpretation. Recently, several studies investigating the demographic history and population structure of *D. melanogaster* have been published (*e.g.* GLINKA *et al.* 2003, HADDRILL *et al.* 2005, OMETTO *et al.* 2005, POOL *et al.* 2006). This allows us to examine levels of nucleotide diversity at individual loci, taking the demographic history of the population into account. In the study of GLINKA *et al.* (2003) a large lumber of ~500 bp fragments were analyzed in a European and an African *D. melanogaster* population sample, respectively. They detected a number of loci in the European sample that were devoid of nucleotide variation. In contrast, levels of heterozygosity at these loci were rather high in the

putative ancestral population. This made these loci devoid of polymorphism in the derived sample good candidates for selective sweeps in the course of adaptive evolution.

In the first chapter of this thesis I analyzed patterns of nucleotide variation in a particular genomic region, located on the X chromosome of *D. melanogaster*. I analyzed the pattern of nucleotide variation in fragments flanking one such invariant locus in the European sample. In addition, I studied the corresponding fragments in an African sample and asked whether a simple demographic scenario can sufficiently explain the observed (local) level of heterozygosity in the derived sample.

In chapter 2, I revisited the genomic region mentioned above and analyzed the central part in detail. I was particularly interested in comparing the structure of the region in the African and the European sample. In addition, by sequencing a sizeable contiguous stretch of DNA, I was able to determine the age of that region. Finally I was interested in finding the putative target of selection.

In the last chapter of my thesis I investigated the population structure of SE Asian *D. melanogaster* using SNP data. I observed that all samples are significantly differentiated from one another and that heterozygosity is generally low. This can have substantial effects on evolution. Not only does it imply that genetic drift may play a leading role (founder effect), but also that advantagous mutations can, if they survive, contribute to rapid phenotypic evolution (MAYR 2001). This is because smaller populations are likely to be distributed over more homogenous environments than large populations, and therefore a mutant has a higher chance to be advantagous (OHTA 1972). In large populations, a mutant is hardly advantagous under all environmental conditions. Morevoer, I investigated the historical context of the SE Asian samples in relation to the African and the Europan sample. Lastly, I revisited the genomic region analyzed in chapter 1 to find further support for local adaptation.

List of Abbreviations

bp	Base pair
BP	Before present
CI	Confidence interval
CLR	Composite likelihood ratio
D	Tajima's *D*
h	Number of haplotypes
H	Fay and Wu's *H*
Hd	Haplotype diversity
in	Intron
ir	Intergenic region
HKA	Hudson-Kreitman-Aguadé test
K	Divergence between species
kb	Kilobases
L	bp studied
LD	Linkage disequilibrium
Mb	Megabases
n	Sample size
N_e	Effective population size
r	Local recombination rate
r^2	Correlation coefficient for a pair of biallelic sites
R	Population recombination rate
s	Selection coefficient
S	Number of segregating sites
S_b	Strength of the bottleneck
SE	Standard error
SNP	Single nucleotide polymorphism
T_b	Time since the bottleneck
TFB	Transcription factor binding site
ya	Years ago
θ	Nucleotide diversity based on *S*
π	Nucleotide diversity based on the average pairwise differences

1 Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*

Environmental changes constitute significant challenges to both plant and animal life, since all life history aspects can potentially be affected. At the molecular level, mutations that confer adaptations increase in frequency, whereas those having rather detrimental effects are removed from the population through purifying selection. Besides selection, neutral processes such as random genetic drift, population substructure or population size bottlenecks may also account for substantial changes in allele frequencies. However, these latter processes depend on special demographic conditions (*e.g.* small population size or restricted gene flow between populations) to produce similar effects as selection.

Previous studies provided convincing evidence for local adaptation of *Drosophila melanogaster*, which originated in sub-Saharan Africa and subsequently colonized many parts of the world (LACHAISE *et al.* 1988). They suggested that numerous beneficial mutations were fixed during the habitat expansion after the last glaciation (about 10,000 years ago; DAVID and CAPY 1988) in the process of adaptation to local environments (GLINKA *et al.* 2003, KAUER *et al.* 2003).

At the DNA level, positive Darwinian selection is often associated with a phenomenon known as genetic hitchhiking (MAYNARD SMITH and HAIGH 1974): neutral variants in the proximity of a beneficial mutation rise in frequency as a consequence of selection. The result of this process is largely determined by the effects of recombination and the strength of selection (KAPLAN *et al.* 1989). Thus, in regions of reduced crossing-over (*e.g.* centromeric and telomeric regions of Drosophila), levels of heterozygosity are generally lower than in the middle of chromosome arms (AGUADÉ *et al.* 1989, STEPHAN and LANGLEY 1989, BEGUN and AQUADRO 1992). Similar patterns of reduced variation can also be caused by

background selection (CHARLESWORTH *et al.* 1993), where neutral variants are removed due to linkage to deleterious mutations that are selected against. However, the effects of background selection are primarily limited to regions of restricted recombination (CHARLESWORTH *et al.* 1993).

In a recent study Glinka *et al.* (2003) investigated the evolutionary history of an African and a European *D. melanogaster* population based on X chromosomal data. They identified several loci that are devoid of polymorphic sites within the European population, whereas levels of heterozygosity in the African population and divergence to its congener *D. simulans* appear to be relatively normal. They proposed that some of the loci with reduced levels of variation are not evolving neutrally, *i.e.* they are targets of natural selection. These loci in the European population served as a starting point for further investigation of the adaptation of *D. melanogaster* to temperate zones. One of these loci is a 348-bp fragment within the fifth intron of *wings apart-like (wapl;* denoted "fragment 10" in GLINKA *et al.* 2003), a gene involved in heterochromatin organization and sister chromatid adhesion (VERNI *et al.* 2000); it is located at cytological position 2D5 on the X chromosome. Here we analyze twelve additional fragments in the vicinity of *wapl* in a European and an African sample to examine whether the pattern of variation around this locus is consistent with the recent action of positive selection.

1.1 MATERIAL AND METHODS

1.1.1 Fly strains

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

1.1.2 Molecular methods

We used the publicly available DNA sequence of the *D. melanogaster* genome (THE FLYBASE CONSORTIUM 2003, http://www.flybase.org, Release 3) for primer design. We amplified and sequenced 12 fragments of noncoding DNA from six intergenic regions and six introns around *wapl*. Genomic DNA was isolated from 15 females of each inbred line using the Puregene DNA isolation kit (Gentra Systems, Minneapolis, USA; see Appendix, Protocol B1). Standard PCR (25 µl) contained 1 µl template DNA, 2.5 µl of 10x buffer, 1 µl $MgCl_2$ (2 mM), 0.25 µl of dNTPs (0.2 mM of each dNTP), 2 µl of each primer (10 µM), 16.12 µl destilled water and 0.13 µl *Taq* polymerase (5 U/µl). PCR conditions were as follows: 4 min at 94°C, 30 cycles of 30 s at 94°C, 30 s at primer specific temperatures, 30 s at 72°C, and a final extension step of 4 min at 72°C. Afterwards PCR fragments were scored on 1.5% agarose gels. Following purification of PCR products (using EXOSAP-IT, USB, Cleveland, USA; see Appendix B2), sequencing reactions were conducted for both strands with the DYEnamic ET terminator cycle sequencing kit (Amersham Biosciences, Buckinghamshire, UK; see Appendix B3 and B4). Sequences were run on a MegaBACE 1000 automated capillary sequencer and analyzed using Cimarron 3.12 base calling software (both from Amersham Biosciences). Finally, sequences were aligned, checked manually and assembled into contigs with Seqman (DNAstar, Madison, WI, USA). When *D. simulans* sequences could not be obtained, we used the publicly available DNA sequence of the *D. simulans* genome. In the case of a gap in the *D. simulans* sequence, we used the published *D. yakuba* sequence as outgroup at the corresponding position (http://species.flybase.net/blast/).

1.1.3 Data analysis

Most statistical analyses were performed using DnaSP 4.0 (ROZAS *et al.* 2003). We estimated nucleotide diversity using π (TAJMA 1983) and θ (WATTERSON 1975). Expected numbers of segregating sites were calculated by performing coalescent simulations. Furthermore, we determined the number of haplotypes (*h*), haplotype diversity (*Hd*; NEI 1987) and divergence (*K*) between *D. melanogaster* and *D. simulans*. Linkage disequilibrium (LD) was determined per fragment in terms of Z_{nS} (KELLY 1997), which is the average of r^2 (HILL and ROBERTSON 1968) over all pairwise

comparisons. To test the neutral equilibrium model, we used Tajima's *D* (TAJIMA 1989), Fay and Wu's *H* (FAY and WU 2000) and the multi-locus-HKA statistic (HUDSON *et al.* 1987). The latter was calculated using the program HKA, kindly provided by J. Hey. Significance of the test statistics was assessed by comparing the observed values to those obtained from 10,000 neutral coalescent simulations. Simulated data were generated using the observed θ-values.

1.1.4 Estimation of the parameters of a selective sweep model

We computed the likelihood of a selective sweep model *vs.* the neutral model for our polymorphism data using a recently developed composite likelihood ratio (CLR) test (KIM and STEPHAN 2002). Briefly, in this test the maximum likelihood of observing a given number of derived variants at a polymorphic site under the selective sweep model (L_1 in KIM and STEPHAN 2002) is compared to that expected under the standard neutral model (L_0). L_1 and L_0 are based on the frequency spectrum and the spatial distribution of polymorphic sites where the derived variants occur with given frequencies in a population sample. The resulting likelihood ratio was compared to the cumulative frequency distribution of likelihood ratios obtained from 10,000 simulations of neutral data sets. Significance was determined at the 5% level (one-tailed test). Since levels of heterozygosity were greatly reduced over a considerable stretch in the European sample (see RESULTS), we used a modified version of test A of KIM and STEPHAN (2002), where neutral data sets were generated conditioned on the observed number of segregating sites. Results were evaluated by a recently proposed goodness-of-fit test (GOF, JENSEN *et al.* 2005), where GOF-values obtained from polymorphism data were compared to those estimated from 1,000 data sets simulated under a selection scenario.

In addition, we applied the test of KIM and NIELSEN (2004) to compute the likelihood of a selective sweep model and estimate the strength of selection. In contrast to KIM and STEPHAN (2002), this test takes LD into account. The strength of directional selection required to cause the reductions in nucleotide diversity observed in our data was estimated as $\alpha = 1.5N_e s$, where N_e is the effective population size and s the selection coefficient. For both tests, we estimated the local population recombination rate (R) as $2N_e\rho$, where the recombination rate (per site

per generation) is $\rho = 0.48 \times 10^{-8}$ (following COMERON *et al.* 1999, using the computer program "Recomb-rate", kindly provided by J.M. Comeron). We assumed $N_e = 0.3 \times 10^6$ and $\theta = 0.0044$ for our European sample, and $N_e = 10^6$ and $\theta = 0.0127$ for the African sample (GLINKA *et al.* 2003).

1.1.5 Position of selected site

We estimated the approximate position of the putative selected site using both the composite likelihood ratio approach by KIM and STEPHAN (2002) and the test by KIM and NIELSEN (2004). Input files were prepared with parameter settings (N_e, R, θ and α) as mentioned above. The current frequency of the beneficial allele was set to 1 and, given the observed pattern of variation (see RESULTS), a very recent fixation of the beneficial allele was assumed ($\tau = 0$). Two-locus sampling probability tables under the selective sweep model and the neutral model were kindly provided by Y. Kim (pers. comm.) and R. Hudson (http://home.uchicago.edu/~rhudson1/), respectively.

1.1.6 Demographic modeling of the European population

Since demographic processes, such as a population bottleneck and subsequent expansion, can leave a signature in the genome that resembles that of selection, we tested the likelihood of such a scenario given our data. We used a coalescent-based method (RAMOS-ONSINS *et al.* 2004) that simplifies the bottleneck model to three parameters: θ (population mutation rate), T_b (time of occurrence of the bottleneck) and S_b (strength of the bottleneck; GALTIER *et al.* 2000). The likelihood that the 60.5 – kb reduction in hetereozygosity was caused by a bottleneck was estimated by comparison to 100,000 genealogies (500,000 for Method II, see Table 3) simulated with $\theta = 0.0066$ (the average level of heterozygosity estimated from fragments 4 to 9 in the African sample) and various combinations of T_b and S_b (both measured in units of $3N_e$ generations), chosen across a range of bottleneck times reported by OMETTO *et al.* (2005). The probability of observing a 60.5 – kb region of reduced diversity in the European sample was estimated using only the fraction of genealogies for which either exactly or at most 43 segregating sites were

observed in the entire region and for which the *wapl* fragment was invariant. The probability of our data under the bottleneck situation was then estimated as the proportion of simulations that yielded at most one segregating site in the fragments located in the monomorphic region (*i.e.* loci 4 to 9). The simulations were done both with and without recombination between adjacent fragments (with $\rho = 0.48 \times 10^{-8}$ rec/bp/gen).

1.2 RESULTS

1.2.1 Nucleotide variation around the *wapl* fragment

In order to determine the extent of the low variability region around the *wapl* fragment in the European *D. melanogaster* sample (GLINKA *et al.* 2003), we sequenced 12 new fragments of noncoding DNA (six introns and six intergenic regions with an average length of 492 bp and an average distance between fragments of 9 kb) around the *wapl* locus in the European lines. The entire region encompasses a total of 110 kb (Figure 1a). Of the 12 fragments, only seven were polymorphic (Table 1). With the exception of one singleton in fragment 8, no intraspecific variation could be detected in a region comprising 60.5 kb. The pattern of polymorphism is illustrated in Figure 2a. The remaining fragments showed an average heterozygosity level (θ) of 0.004, consistent with a mean θ-value of 0.0044 for the European population (GLINKA *et al.* 2003).

Furthermore, we analyzed the corresponding nucleotide variation in the African sample. Estimated levels of heterozygosity for 11 fragments are listed in Table 2. The pattern of polymorphism for loci 4 to 8 is illustrated in Figure 2b. For fragment 3 we were not able to obtain sequences. With the exception of two low-variation fragments (*i.e.* loci 6 and 7), the pattern of nucleotide diversity is consistently higher for the African lines (than for the European ones) with an

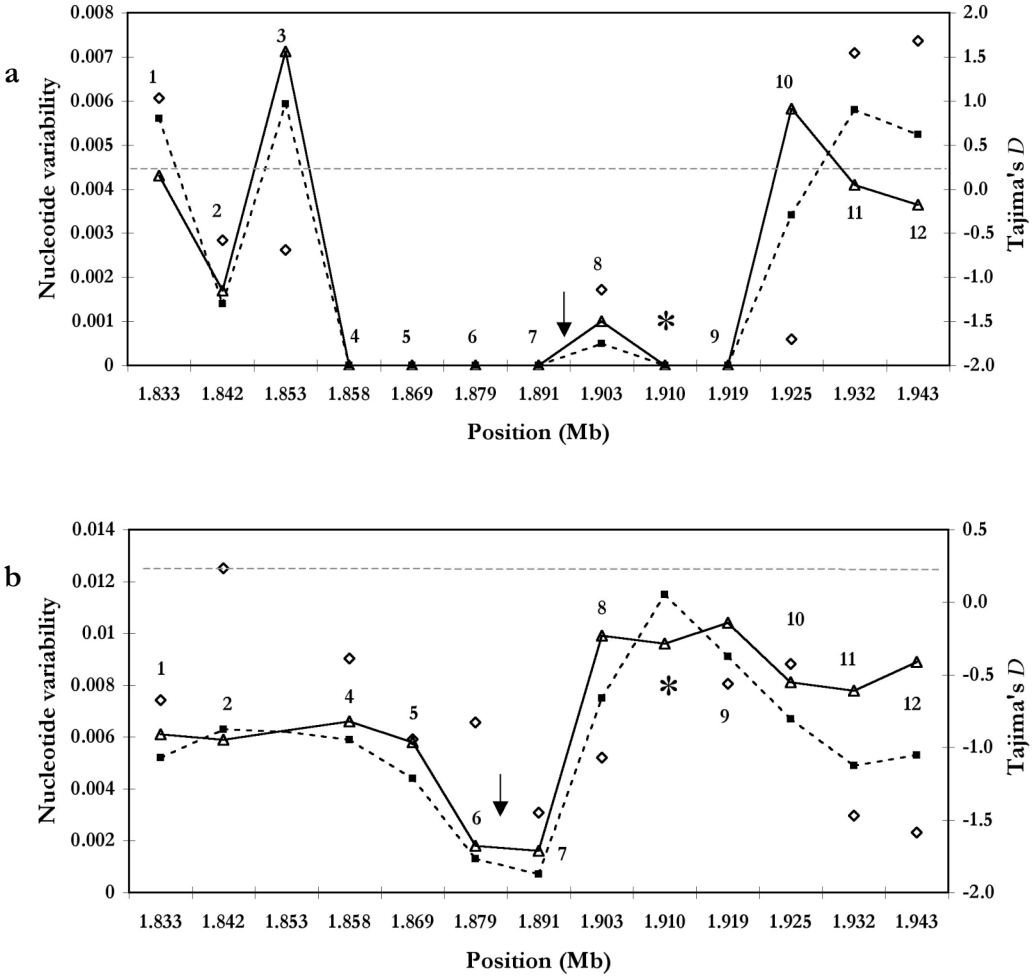


FIGURE 1. Intraspecific variation around the *wapl* fragment. Panel a refers to the European sample, and panel b to the African sample. Solid lines and triangles correspond to θ, black dashed lines and squares show π. Diamonds indicate Tajima's *D*, and the grey dashed lines represent chromosome wide average heterozygosity as reported by GLINKA *et al.* (2003). Arrows indicate estimated positions of the target of selection following tests of KIM and STEPHAN (2002) and KIM and NIELSEN (2004); ✻ indicates the position of the *wapl* fragment. Absolute genomic positions of fragments are given in Mb, according to release 3 of the annotated *D. melanogaster* genome. For fragment 3 of the African lines sequences could not be obtained.

average θ of 0.0065 for fragments 4 to 9 (Figure 1b). However, this value is about 50% lower than the average level of heterozygosity reported for the entire X chromosome in Africa (θ = 0.0127; GLINKA *et al.* 2003).

a

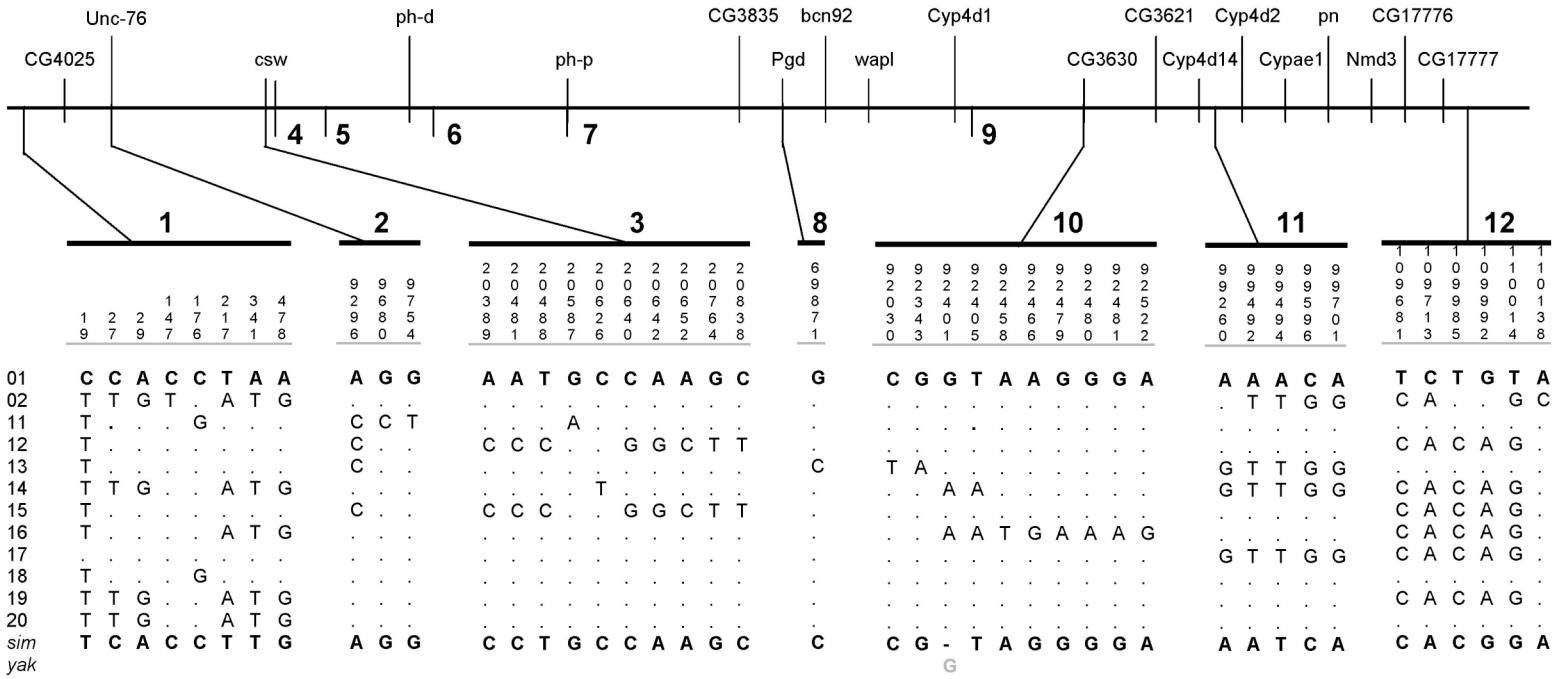


FIGURE 2. Polymorphism data for the 2C10-2E1 region. Panel a refers to the European sample, and panel b to the African one. *D. melanogaster* lines 01 (European sample) and 82 (African sample) are taken as references. For the African sample (panel b) only polymorphism data for fragments 4 to 8 is presented. *sim*, *D. simulans*. *yak*, *D. yakuba*. (-): no sequence data available.

b

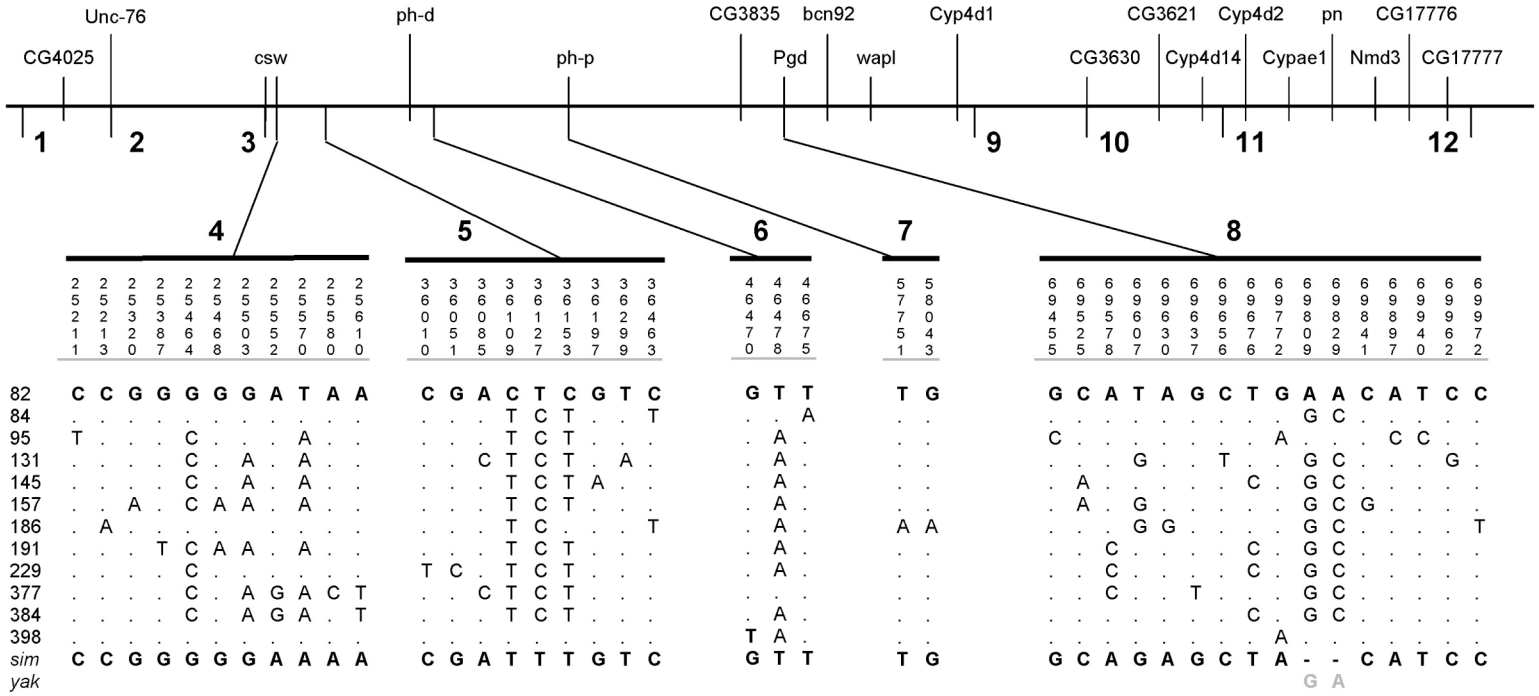


FIGURE 2. (*continued*)

1.2.2 Standard neutrality tests for the European sample

Tajima's *D* (TAJIMA 1989) was highly negative for two distal loci (2 and 3), one proximal fragment (10) directly flanking the invariant region, and the locus showing a singleton (8; Figure 1a). For fragment 10 the observed value is significantly lower than the neutral expectation ($P = 0.038$). Similar to Tajima's *D* statistic, Fay and Wu's *H* (FAY and WU 2000) was negative for three polymorphic fragments ($H = -0.67$, -1.67 and -0.36 for fragments 3, 8 and 10, respectively), indicating an excess of high frequency derived variants. However, these values were not significant. In contrast, we observed positive values of Tajima's *D* for three fragments located more distal to the invariant region (*i.e.* loci 1, 11 and 12). For locus 12 this value is significant ($P < 0.05$).

For each fragment located in the monomorphic region we estimated the probability of observing a locus of length *L* devoid of polymorphisms, given an expected heterozygosity of 0.0044 by comparison with values obtained from 10,000 coalescent simulations of the standard neutral model under the conservative assumption of zero recombination (see HUDSON 1990). As shown in Table 1, all fragments under consideration represent a reduced number of segregating sites and, with the exception of fragment 8 (containing the singleton), this reduction is significant compared to the neutral expectation. In the center of the analyzed region (*i.e.* fragments 4 to 9) that encompasses 2,553 nucleotides, where a valley of reduced variation has been observed, we detected only one segregating site. The probability of this result, under the conservative assumption of no recombination, is significantly low ($P < 0.00001$) and incompatible with the standard neutral model. The possibility of selective constraints or a low regional mutation rate being the cause of the observed reduction in variation can be excluded, since levels of divergence between *D. melanogaster* and its sister species *D. simulans* observed in the 110 – kb region investigated are on average normal (Table 1). Indeed, a multi-locus version of the HKA test (HUDSON *et al.* 1987) revealed a significant departure from neutrality ($X^2 = 30.15$, $P = 0.0015$). This result still holds when the fragment with the largest contribution to the HKA statistic (*i.e.* locus 3) was removed from analysis. Only when, in addition, the next largest contribution (fragment 10) was removed, this result was no longer significant ($P = 0.319$).

TABLE 1. Polymorphism of the European sample and divergence

Fragment	L	K	S_{obs}	θ_{obs}	θ_{exp}	S_{exp}	h	Hd	Z_{nS}
1_{ir}	598	0.13	8	0.0043	0.0011 – 0.0099	2 – 18	6	0.879	0.452
2_{in}	590	0.05	3	0.0017	0.0005 – 0.0101	1 – 18	3	0.530	NA
3_{in}	465	0.04	10	0.0071	0.0007 – 0.0107	1 – 15	4	0.531	1.0*
4_{in}	600	0.06	0	0	0.0011 – 0.0099**	2 – 18	1	0	NA
5_{ir}	465	0.02	0	0	0.0007 – 0.0107*	1 – 15	1	0	NA
6_{ir}	287	0.05	0	0	0.0000 – 0.0115*	0 – 10	1	0	NA
7_{in}	422	0.03	0	0	0.0008 – 0.0102*	1 – 13	1	0	NA
8_{in}	319	0.04	1	0.0010	0.0000 – 0.0114	0 – 11	2	0.167	NA
9_{ir}	460	0.06	0	0	0.0007 – 0.0101*	1 – 14	1	0	NA
10_{in}	569	0.05	10	0.0058	0.0006 – 0.0099	1 – 17	4	0.455	1.0*
11_{ir}	405	0.09	5	0.0041	0.0008 – 0.0106	1 – 13	3	0.530	0.867
12_{ir}	546	0.06	6	0.0036	0.0006 – 0.0103	1 – 17	3	0.621	0.829*

L: number of sites studied, *K*: divergence between *D. melanogaster* and *D. simulans*, S_{obs}: observed number of segregating sites, θ_{obs}: observed heterozygosity, θ_{exp}: expected heterozygosity (95 % confidence intervals), *h*: number of haplotypes, *Hd*: haplotype diversity, Z_{nS}: linkage disequilibrium, ir: intergenic region, in: intron, and NA: not available. The expected number of segregating sites (S_{exp}) is calculated according to TAJIMA (1983) for $n = 12$ lines. * and ** indicate significance at the 0.05 and 0.01 level, respectively.

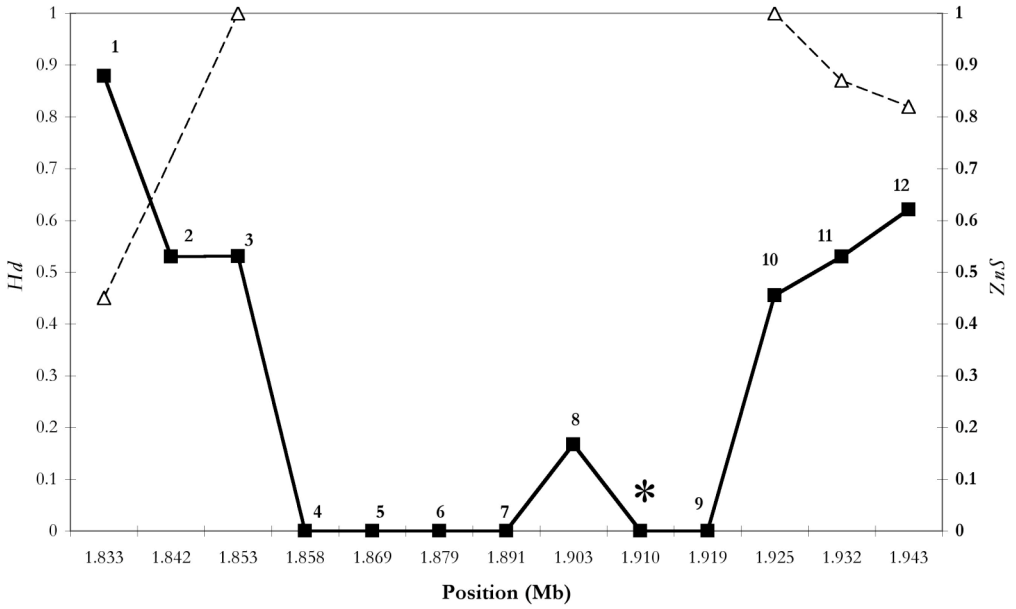


FIGURE 3. Haplotype diversity and linkage disequilibrium in the European sample. Solid black lines and squares denote haplotype diversity (*Hd*). Dashed lines and triangles indicate levels of linkage disequilibrium (Z_{nS}). Z_{nS} was estimated using parsimony informative sites only. Therefore, no value for fragment 2 was obtained. * indicates the position of the *wapl* fragment.

The observed number of haplotypes for the fragments surrounding the invariant region varies from three to six per fragment, with haplotype diversity increasing with distance from the invariant region (Figure 3 and Table 1).

The observed values, however, did not depart significantly from neutral expectations. Yet, two fragments directly flanking the monomorphic region (*i.e.* loci 3 and 10) showed a considerable reduction in haplotype diversity ($P = 0.09$ and 0.05, respectively).

In addition we detected significant LD ($P < 0.05$) within fragments 3, 10, and 12 (Table 1). For fragment 11 the observed value was marginally significant ($P = 0.05$). As expected, LD decays in both directions with distance from the valley of reduced variation. We did not detect LD among adjacent fragments. This may be due to recombination. For instance, between fragments 11 and 12, which are separated by approximately 10 kb, at least one recombination event can be inferred applying the four-gamete rule (HUDSON and KAPLAN 1985).

1.2.3 Standard neutrality tests for the African sample

Tajima's *D* (TAJIMA 1989) estimated from fragments 1 to 12 showed a general trend towards negative values, with the exception of fragment 2 (Figure 1b). However, only the *D*-value estimated from locus 12 was significantly different from neutral expectations ($P = 0.05$). For the same fragment Fay and Wu's *H* (FAY and WU 2000) was significantly negative as well ($H = -9.0$, $P = 0.03$), indicating an excess of high frequency derived variants. *H* was also negative for fragments 6, 8 and 11 ($H = -0.52$, -0.18 and –0.56, respectively). However, these values are not significantly different from zero.

As for the European population, we estimated whether the observed number of polymorphic sites in region 4 to 9 ($S_{obs} = 56$, see Table 2) is significantly different from expectations under the standard neutral model. Given an expected heterozygosity of 0.0127 (GLINKA *et al.* 2003), 41 to 251 segregating sites would be expected. The observation of 56 SNPs is therefore not significantly different from neutral expectations ($P = 0.07$). However, it should be noted that the θ estimates for loci 6 and 7 are significantly low (Table 2). The multi-locus HKA test (HUDSON *et al.* 1987) did not reveal any significant departure from neutrality for our fragments in the African sample ($X^2 = 9.90$, $P = 0.45$).

The observed number of haplotypes varies from two to 11 with generally high haplotype diversities (Table 2). We run coalescent simulations using the estimated recombination rate ($\rho = 0.48 \times 10^{-8}$ rec/bp/gen) to infer significance. This is conservative, since recombination tends to increase both statistics. Of 11 loci analyzed, four show significantly increased numbers of haplotypes. This pattern is in accordance with previous studies reporting a chromosome wide trend towards elevated numbers of haplotypes and high haplotype diversities in African *D. melanogaster* (GLINKA *et al.* 2003, OMETTO *et al.* 2005). With the exception of fragment 1, no significant linkage disequilibrium was detected in the region analyzed (Table 2).

1.2.4 Can the severe reduction in variation observed in the European sample be explained by a population size bottleneck?

To account for the possibility that a population size bottleneck has caused

TABLE 2. Polymorphism of the African sample

Fragment	L	n [a]	S_{obs}	θ_{obs}	h	Hd	Z_{nS}
1_{ir}	557	11	10	0.0061	8	0.927	0.791*
2_{in}	562	12	10	0.0059	11**	0.985**	0.103
3_{in}	NA	NA	NA	NA	NA	NA	NA
4_{in}	551	12	11	0.0066	9*	0.939	0.288
5_{ir}	518	12	9	0.0058	8	0.894	0.253
6_{ir}	563	12	3	0.0018**	4	0.561	NA
7_{in}	515	12	2	0.0013**	2	0.167	NA
8_{in}	534	12	16	0.0099	10*	0.955	0.224
wapl	346	12	10	0.0096	6	0.879	0.239
9_{ir}	493	11	15	0.0104	8	0.927	0.259
10_{in}	462	11	10	0.0074	8*	0.945	0.195
11_{ir}	500	11	11	0.0075	8	0.945	0.127
12_{ir}	475	12	12	0.0084	4	0.636	0.534

[a] number of lines analyzed. For other abbreviations, see Table 1 legend.

the reduction in heterozygosity observed in our European sample, we applied the approach of RAMOS-ONSINS *et al.* (2004), modified by OMETTO *et al.* (2005). That is, based on the parameter estimates of OMETTO *et al.* (2005; see MATERIALS AND METHODS), we simulated a drop in effective population size at various times in the past. For bottleneck times assayed across a range of times, *i.e.* T_b = 0.01, 0.02, 0.03 or 0.05, we simulated genealogies with two different strengths of the bottleneck, as suggested by OMETTO *et al.* (2005). Under the assumption of no recombination between loci the probability of our data to be explained by a simple bottleneck scenario is low (Table 3). However, only *P*-values for reasonably old bottlenecks ($T_b \geq 0.02$, *i.e.* more than 6,000 years ago, assuming 10 generations per year) are significant. Note that according to OMETTO *et al.* (2005) the X-chromosomal nucleotide diversity of the European sample is best described by a combination of $T_b = 0.0267$ and $S_b = 0.400$, *i.e.* a bottleneck that has occurred about 8,000 years ago. If some recombination ($\rho = 0.48 \times 10^{-8}$ rec/bp/gen) was allowed between fragments, the probability of our data is significantly low for all simulated scenarios. When the condition of observing exactly 43 segregating sites was relaxed, our data still remained significant for the older bottleneck scenarios but were only marginally

TABLE 3. Results of bottleneck simulations

T_b [a]	S_b [b]	Method I [c]	Method II [d]	Method III [e]
0.01	0.34	0.18	<0.001	0.09
	0.4	0.21	0.005	0.14
0.0122 [f]	0.371 [f]	0.15	<0.001	0.082
0.0125 [f]	0.350 [f]	0.13	<0.001	0.073
0.0128 [f]	0.345 [f]	0.12	<0.001	0.069
0.02	0.34	0.05	<0.001	0.03
	0.4	0.06	<0.001	0.031
0.0264 [f]	0.380 [f]	0.028	<0.001	0.013
0.0267 [f]	0.400 [f]	0.028	<0.001	0.014
0.03	0.34	0.016	<0.001	0.006
	0.4	0.018	<0.001	0.008
0.05	0.34	0.002	<0.001	0.0007
	0.4	0.003	<0.001	0.0012

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

[b] Strength of the bottleneck.

[c] Probability of observing at most one segregating site in loci 4 to 9 under the assumption of no recombination between fragments, conditioned on the observation of 43 segregating sites in the entire region and zero polymorphism in *wapl*.

[d] Probability of observing at most one segregating site in loci 4 to 9 under the assumption of intergenic recombination (with $\rho = 0.48 \times 10^{-9}$ rec/bp/gen), conditioned on the observation of 43 segregating sites in the entire region and zero polymorphism in *wapl*.

[e] Probability of observing at most one segregating site in loci 4 to 9 under the assumption of intergenic recombination, conditioned on the observation of at most 43 segregating sites in the entire region and zero polymorphism in *wapl*.

[f] From Ometto *et al.* (2005).

significant for more recent bottlenecks (T_b = 0.01 - 0.0128, *i.e.* between 3,000 and 3,840 years ago).

1.2.5 Estimation of selection parameters

We applied the maximum likelihood ratio tests of KIM and STEPHAN (2002) and KIM and NIELSEN (2004) to estimate the significance of the reduction in variation observed in our data under both the standard neutral model and a selection model. Furthermore, we estimated the strength of selection. Since a large fraction of the genomic region under analysis shows highly reduced levels of heterozygosity in the European sample (*i.e.* the putative sweep region), we specified $\theta = 0.0044$ for this analysis as reported by GLINKA *et al.* (2003). Using the KIM and STEPHAN (2002) method we compared the likelihood ratio ($LR_{KS} = L_1/L_0$) to those obtained from 10,000 neutral coalescent simulations. The probability of finding the likelihood ratio obtained from our data ($LR_{KS} = 16.30$) under a neutral scenario is low ($P = 0.037$).

Since polymorphism patterns produced by a selective sweep can be confounded by those resulting from demographic events, *e.g.* population structure or a recent bottleneck, we applied the goodness-of-fit (GOF) test proposed by JENSEN *et al.* (2005). We obtained $\Lambda_{GOF} = 467$ with a Monte Carlo *P*-value estimate of 0.81. Therefore, the significant LR_{KS} value is unlikely to be a false positive, *i.e.* the result of demographic forces alone. This result is supported by the KIM and NIELSEN (2004) test, which also yielded a significantly large likelihood ratio (LR_{KN}) of the selective sweep *vs.* the neutral model ($LR_{KN} = 17.09$, $P = 0.05$ in comparison to 10,000 simulated neutral data sets).

Estimates of the strength of selection ($\alpha = 1.5N_e s$) are 661 and 552 for the KIM and STEPHAN (2002) and the KIM and NIELSEN (2004) method, respectively. Assuming that the effective population size (N_e) of the European *D. melanogaster* population is approximately one third of the African N_e (GLINKA *et al.* 2003), *s* is estimated to be 1.3×10^{-3} and 1.1×10^{-3} using the KIM and STEPHAN (2002) and the KIM and NIELSEN (2004) test, respectively.

Next we consider the African sample. We applied only the KIM and STEPHAN test (2002) to the polymorphism data obtained from the African sample, since estimated levels of LD were rather low (see Table 2). We obtained a LR_{KS} of 21.54, which was significant in comparison to 10,000 neutral data sets ($P = 0.02$). Λ_{GOF} estimated by the GOF test (JENSEN *et al.* 2005) was 912 ($P = 0.87$). As for the European sample, the polymorphism pattern observed in the African sample can

therefore not be explained by simple demographic events alone. The estimate of the strength of selection (α) produced by the test is 2,076, which yields $s = 1.4 \times 10^{-3}$ assuming an effective population size of 10^6.

1.2.6 Position of selected site

Applying the CLR test (KIM and STEPHAN 2002), we estimated the approximate position of the selected site. For the European sample the likelihood ratio is maximized at position 57.7 kb, indicating a target of selection approximately in the middle of the analyzed region. In addition, we used the approach by KIM and NIELSEN (2004). The results obtained by this method indicate that the target of selection is located at 57.9 kb, *i.e.* also within the fourth intron of the gene *ph-p*. Thus, the inclusion of LD results in an estimated position of the target of selection that is shifted slightly downstream, presumably due to the somewhat stronger LD found in the downstream flanking region (compared to the other flanking region). For the African sample we obtained a somewhat different estimate for the target of selection. The KIM and STEPHAN (2002) likelihood ratio is maximized at position 49.8 kb, therefore pointing towards a target of selection closer to fragment 6, which is located between *ph-d* and *ph-p* (see Figure 2).

1.3 DISCUSSION

Previous population genetic studies revealed a general trend towards lower nucleotide diversity in non-African *Drosophila melanogaster* populations (BEGUN and AQUADRO 1993, BEGUN and AQUADRO 1995, SCHLÖTTERER *et al.* 1997, LANGLEY *et al.* 2000, ANDOLFATTO 2001, KAUER *et al.* 2002, GLINKA *et al.* 2003, BAUDRY *et al.* 2004). Current research attempts to reveal the mechanisms that have led to this geographical pattern of genetic variation. Besides the effect of demographic events, such as population bottlenecks, positive directional selection has been hypothesized to substantially contribute to reductions in heterozygosity at individual loci as opposed to the genome wide effects of demography (*e.g.* BEGUN and AQUADRO

1992, HUDSON *et al.* 1994, HARR *et al.* 2002, QUESADA *et al.* 2003, SCHLENKE and BEGUN 2004).

1.3.1 Evidence for a selective sweep in the *wapl* region

Here we describe several lines of evidence suggesting that positive selection has shaped the genetic variation in the *wapl* region of *D. melanogaster*. First, we analyzed a 110 – kb region by sequencing 12 fragments of noncoding DNA in a European *D. melanogaster* sample and detected a stretch of approximately 60 kb that is nearly devoid of nucleotide diversity. That is, all seven loci that were analyzed within that 60 – kb region are monomorphic, with the exception of one fragment containing a singleton. In contrast, levels of heterozygosity in the African sample appear to be relatively normal, but lower than the chromosome wide average reported by earlier studies (GLINKA *et al.* 2003, OMETTO *et al.* 2005). A total of 66 segregating sites was observed in the 60-kb region of the African sample (including the *wapl* fragment of GLINKA *et al.* 2003). A similarly large invariant region (100 kb), as detected in our European sample, has thus far only been found in *D. simulans*, possibly caused by the fixation of a positively selected allele of a cytochrome P450 gene (SCHLENKE and BEGUN 2004).

The pattern of variation we observed in our European data is unlikely under both the neutral equilibrium model and a simple bottleneck scenario in which a single population size reduction occurred ~8,000 years ago (or earlier). A more recent bottleneck (~3,000 to 4,000 years ago) could be sufficient to explain the observed reduction in heterozygosity. However, such a recent bottleneck is unlikely for European Drosophila (LACHAISE *et al.* 1988, HADDRILL *et al.* 2005, OMETTO *et al.* 2005).

It should be noted that in our bottleneck simulations we did not account for the observation that the African population has been undergoing a size expansion (GLINKA *et al.* 2003). To estimate θ from the observed number of segregating sites (see MATERIALS AND METHODS) we assumed a constant population size. Under an expansion model, a higher θ - value would be estimated given the observed number of segregating sites. Therefore, the θ - value used in the bottleneck simulations of the

European population is probably too low. In other words, our method is conservative.

Using the methods of KIM and STEPHAN (2002) and KIM and NIELSEN (2004) we showed that a selective sweep model fits the data significantly better than the neutral equilibrium model. Additional predictions of the selective sweep model were also verified in the data. First, we found a skew in the frequency spectrum of polymorphisms towards rare variants, as indicated by negative values of Tajima's *D* (AGUADÉ *et al.* 1989, HUDSON 1990, BRAVERMAN *et al.* 1995, FAY and WU 2000, PAYSEUR and NACHMAN 2002). This test statistic is notably sensitive to the influx of new mutations that have occurred after a hitchhiking event (FAY and WU 2000). We detected such an excess of low frequency mutations in some of our fragments, consistent with previous studies (*e.g.* LANGLEY *et al.* 2000). Second, we observed an excess of high frequency derived variants (FAY and WU 2000, KIM and STEPHAN 2000). Visual inspection of our data revealed a high frequency of derived variants at three polymorphic loci (3, 8 and 10; Figure 2a). Negative values of Fay and Wu's *H* statistic confirm this observation and are in accordance with the selective sweep hypothesis. Third, under the hitchhiking model strong transient LD is expected between neutral segregating sites located in the vicinity (on one side) of the target of selection (THOMSON 1977, KIM and STEPHAN 2002, KIM and NIELSEN 2004). Consistent with these predictions, we detected strong haplotype structure and high levels of LD in our data. In addition, haplotype diversity increased and LD decayed with distance from the monomorphic region (Figure 3).

Where did the selective sweep detected in the European population originate? Our analysis of the African sample may suggest that the sweep arose in an ancestral African population before the colonization of Europe. A similar trans-population sweep (between Africa and Europe) has been detected by LI and STEPHAN (2005) in a different data set. The hypothesis of a trans-population sweep in the *wapl* region needs to be verified by additional sequencing to establish the complete haplotypes in the region under consideration. An alternative hypothesis is that the sweeps in Africa and Europe are independent, as the estimated positions of the target sites of selection that differ by about 8 kb may seem to indicate (see RESULTS). However, this may simply be a consequence of the fact that the target of

selection is difficult to localize precisely in the European sample due to a lack of variation (see below).

Although most loci in the *wapl* region of the African sample do not show a severe reduction in nucleotide diversity, the π- and θ-values for two fragments located in the center of the region are significantly reduced. This reduction in variation and an associated skew in the frequency spectrum resulted in a significant KIM and STEPHAN (2002) test, which is unlikely to be the sole product of simple demographic forces (JENSEN *et al.* 2005). The observed lack of further statistical evidence for selection may be attributed to the relatively old age of the hitchhiking event. Signatures of directional selection are difficult to identify with our methods if they are much older than ~$0.1N_e$ generations (KIM and STEPHAN 2000, KIM and STEPHAN 2002).

1.3.2 Estimating the strength and target site of selection

Using the methods of KIM and STEPHAN (2002) and KIM and NIELSEN (2004), we estimated the strength of selection and the approximate target site of selection. For the European sample, both methods produced selection coefficients in the order of 10^{-3}, and the target of selection was located approximately 2,500 bp downstream from the center of the monomorphic region (see Figure 1) within the fourth intron of the gene *ph-p*. The method of KIM and NIELSEN (2004) suggested that the beneficial mutation occurred an additional 200 bp downstream from the KIM and STEPHAN (2002) estimate. This result may be explained by the incorporation of LD into the test statistic and the fact that LD appears to be slightly stronger in the fragments located further downstream (*i.e.* fragments 10 to 12) than in those on the other side of the valley of reduced polymorphism. However, it should be noted that is difficult to precisely localize the putative target of selection using composite likelihood ratio tests for technical reasons and, in this case, also due to a strong reduction of variation over a large region. For the African sample, the KIM and STEPHAN test (2002) indicates that the position of the target of selection is 8 kb upstream from the estimate based on the European data. Since the valley of

reduced variation in the African population is much narrower (see Figure 1), this should facilitate pinpointing the target site of selection.

1.3.3 Genes near the target site of selection

The genomic region of reduced variation (in the European sample) from *ph-d* to *Cyp4d1* harbors a relatively high density of genes coding for products with metabolic functions: *CG3835*, putatively involved in carbohydrate metabolism; *Pgd*, involved in the pentose-phosphate-shunt; *bcn92*, with putative oxidoreductase activity, and *Cyp4d1*, a cytochrome P450 gene, putatively involved in steroid metabolism. Genes coding for metabolic enzymes have frequently been suggested to be targets of positive selection (*e.g.* BEGUN and AQUADRO 1994, HUDSON *et al.* 1994, MUTERO *et al.* 1994, EANES 1999, SCHLENKE and BEGUN 2004). It is therefore plausible that the beneficial mutation occured in one of these genes.

1.4 SUMMARY

A previous scan of the X chromosome of a European *Drosophila melanogaster* population revealed evidence for the recent action of positive directional selection at individual loci (GLINKA *et al.* 2003). In this study we analyzed one such region that showed no polymorphism in the genome scan (located in cytological division 2C10-2E1). We detect a 60.5 – kb stretch of DNA encompassing the genes *ph-d*, *ph-p*, *CG3835*, *bcn92*, *Pgd*, *wapl* and *Cyp4d1* that almost completely lacks variation in the European sample. Loci flanking this region show a skewed frequency spectrum at segregating sites, strong haplotype structure, and high levels of linkage disequilibrium. Neutrality tests reveal that these data are unlikely under both the neutral equilibrium model and simple bottleneck scenarios as well. In contrast, newly developed maximum likelihood ratio tests suggest that strong selection has acted recently on the region under investigation, causing a selective sweep. Evidence is presented that this sweep may have originated in an ancestral population in Africa.

2 The *wapl* region revisited

In previous studies researchers have focused on identifying genomic regions that are subject to positive directional selection (*e.g.* HARR *et al.* 2002, BUSTAMANTE *et al.* 2005, OMETTO *et al.* 2005, SCHMID *et al.* 2005). These studies followed different approaches to detect selection: BUSTAMANTE *et al.* (2005) contrasted patterns of human coding sequence polymorphism to divergence to chimpanzees. The authors were interested in the recent molecular evolution of these species. One of their major results was the discovery that certain classes of genes, such as transcription factors (TF), are overrepresented in the proportion of genes that seem to be rapidly evolving. In contrast, housekeeping genes are found to be rather conserved (BUSTAMANTE *et al.* 2005). On the other hand, others used population genetic studies to localize regions containing positively selected loci by comparing nucleotide diversity at multiple loci in samples from derived and ancestral populations. Correcting for demography, *e.g.* population size bottlenecks, candidate regions with severely reduced levels of heterozygosity were identified (*e.g.* SCHLÖTTERER 2002, KAUER *et al.* 2002, KAYSER *et al.* 2003, STORZ *et al.* 2004, OMETTO *et al.* 2005). Subsequent investigations concetrated on single candidate regions that include loci with prior suspicion of non-neutral evolution (HARR *et al.* 2002, SAEZ *et al.* 2003, CATANIA and SCHLÖTTERER 2005, GLINKA *et al.* 2006, OMETTO 2006). In addition there are case studies in which particular regions were analyzed without *a priori* knowledge of the potential mode of evolution (HUDSON *et al.* 1994, NURMINSKY *et al.* 1998, SCHLENKE and BEGUN 2004, POOL *et al.* 2006). Generally, case studies can be classified as follows:

I. Convincing evidence in favour of positive selection is reported, but the target of selection (*i.e.* gene) was not detected (*e.g.* CATANIA and SCHLÖTTERER 2005, GLINKA *et al.* 2006, OMETTO 2006).

II. The target of selection was found, but the beneficial mutation could not be identified (HARR *et al.* 2002, BAUER DUMONT and AQUADRO 2005).
III. Both, the target of selection and the beneficial selection were identified (FFRENCH-CONSTANT *et al.* 1993, ENARD *et al.* 2002, SCHLENKE and BEGUN 2004).

The majority of studies falls into class I. The main reason for this is the difficulty in identifying the putative targets of natural selection. In many investigations the interest in a particular genomic region stems from a previous "hitchhiking mapping" study that identified certain regions with putative non-neutral evolution (see above). These regions are rather large and encompass several hundred kb (OMETTO *et al.* 2005, LI and STEPHAN 2006). Subsequent work typically focuses on one such region and re-examines the assumption of positive selection having shaped that particular region of the genome (*e.g.* GLINKA *et al.* 2006, see also chapter 1). These follow-up studies are performed as miniature mapping studies where small DNA fragments of ~500 bp are screened for reduced nucleotide diversity, increased levels of linkage disequilibrium and other features of a recent selective sweep (see previous chapter). Inference of the location of the target of selection is based on characteristics of the region examined, such as the location of severely reduced heterozygosity and extreme values of common summary statistics (*e.g.* TAJIMA'S (1989) *D*, and FU and LI'S (1993) *D*). Recently developed methods for the detection of non-neutral evolution make use of this information and are able to propose the approximate position of the beneficial site (KIM and STEPHAN 2002, PRZEWORSKI 2003, KIM and NIELSEN 2004). However, estimates of the target of selection are accompanied by large standard deviations (KIM and STEPHAN 2002 and this study). Nevertheless, the approximate localization faciliates the search for the selected site and a candidate gene may be identified, given that a large part of the region of interest has been sequenced and subjected to the afore mentionend tests (J. Jensen pers. comm.). Yet, in most studies the region under investigation has only been sequenced in parts, and the target of selection may not be covered. This further complicates efforts to find the exact location of the beneficial mutation by means of composite likelihood ratio

tests (KIM and STEPHAN 2002). Furthermore the strength of selection may be greatly underestimated (J. Jensen pers. comm.).

In chapter 1 we identified a genomic region, located on the X chromosome of *D. melanogaster*, which most likely has been affected by positive directional selection. We observed that a hitchhiking event involved both an ancestral population (from Zimbabwe) and a derived population (from The Netherlands). However, we were not able to determine how selection has shaped the genomic region of interest regarding the target and the timing of selection. Two different scenarios are plausible: i) the positively selected mutation arose only once, namely in an ancestral African population and was transferred to Europe during the colonization process (*i.e.* a transpopulation sweep) or ii) the sweeps in Europe and Africa were independently caused by different beneficial mutations.

In this study we revisit the *wapl* region and analyze it in detail. Our previous work revealed that levels of heterozygosity are severly reduced between the genes *csw* and *Pgd* in the African sample (see Figures 1 and 2 in the previous chapter). In addition the target of selection was estimated to be located in the vicinity of the *ph-d* – *ph-p* gene complex. Therefore, we chose to concentrate on the *ph-d* to *Pgd* region and sequenced this ~30 kb spanning part in both the African sample and the European sample. We analyzed contiguous DNA segments and established the complete haplotypes in the region under consideration. Our results suggest that a selective sweep occurred at *ph-p* in the African population, but an independent sweep must be proposed for Europe in order to explain levels of nucleotide diversity and haplotype structure in this population.

2.1 MATERIAL AND METHODS

2.1.1 Fly strains

We collected data from 18 highly inbred *D. melanogaster* lines: 12 lines were derived from an African population (Lake Kariba, Zimbabwe) and 6 lines are from Europe (Leiden, The Netherlands; see chapter 1). For interspecific comparisons we

used the publicly available *D. simulans* and *D. yakuba* sequences available at http://www.flybase.org/blast/.

2.1.2 Molecular techniques

Sequences were generated as described in chapter 1, except that sequencing reactions were performed using the ABI BigDye Terminator v1.1 sequencing kit and run on a ABI 3730 DNA Analyzer (Applied Biosystems, Foster City,USA; see Appendix B5). Sequences were aligned and checked by eye using SeqMan (DNAstar, Madison, USA). Since the ~30 kb region under investigation was sequenced in overlapping segments, we assigned the overlapping part to the distal fragment and excised it from the next (proximal) fragment. Sequences were merged manually and finally assembled into contigs using SeqMan and MegAlign (DNAstar, Madison, USA).

2.1.3 Statistical analysis

Basic analyses were done using DnaSP 4.10 (ROZAS *et al.* 2003). Nucleotide diversity was estimated in terms of θ (WATTERSON 1975) and π (TAJIMA 1983). We determined the number of haplotypes (*h*) and haplotype diversity (*Hd*; Nei 1987). Linkage disequilibrium (LD) was estimated for informative sites per fragment as Z_{nS} (KELLY 1997). In addition we estimated r^2 (HILL and ROBERTSON 1968) for all informative sites in the region analyzed using *TASSEL 2.0* (ZHANG *et al.* 2006). Indels and singletons were not included in this analysis. We estimated the minimum number of recombination events in the whole region applying the four-gamete rule (HUDSON and KAPLAN 1985) and performed a sliding window analysis of recombination using the approach detailed in MCVEAN *et al.* (2002). Both methods are implemented in the program LDhat (MCVEAN *et al.* 2002). We tested the neutral equilibrium model using Tajima's *D* (TAJIMA 1989), Fu and Li's *D* (FU and LI 1993), Fu's F_s (FU 1997), and Fay and Wu's *H* (FAY and WU 2000) and the HKA test (HUDSON *et al.* 1987). Significance was tested by 10,000 neutral coalescent simulations under the assumption of equilibrium and zero recombination. Tests were performed one-sided. The folded frequency spectrum was obtained using DnaSP 4.10 (ROZAS *et al.* 2003).

2.1.4 Likelihood analysis of selective sweep

We also analyzed our SNP data using the composite likelihood ratio (CLR) test of KIM and STEPHAN (2002; see chapter 1). This method estimates the likelihood of the selective sweep model compared to the standard neutral model and provides estimates for the strength of selection ($\alpha = 1.5N_e s$) and the position of the selected site ($\hat{X}$). Further information on the CLR test is provided in chapter 1. We tested the signficance of the resulting likelihood ratio by 10,000 Monte Carlo simulations under neutrality, given our fragment structure and the number of observed segregating sites. Since demographic events can produce a local signature similar to that created by a selective sweep, we evaluated the CLR test result by the goodness-of-fit (GOF) test proposed by JENSEN *et al.* (2005), implemented in the CLR algorithm. The GOF test was also used to obtain confidence intervals for $\hat{X}$ by parametric bootstrapping. For likelihood analyses we assumed $N_e = 10^6$ and either $\theta = 0.0067$ or 0.0127 for the African sample and $N_e = 0.3 \times 10^6$ and $\theta = 0.0044$ for the European sample (GLINKA *et al.* 2003).

2.1.5 Pinpointing the target of selection

After the identification of a putative selective sweep and localizing the approximate position of the selected site in the African sample we screenend the proximity for candidate fixations, *i.e.* subsitutions or indels that occurred along the *D. melanogaster* lineage. *D. simulans* and *D. yakuba* were used as outgroups. We also tested whether candidate fixations in intronic and 5'-regions result in regulatory changes (*i.e.* differences in transcription factor binding sites) applying the *MatInspector* tool (CARTHARIUS *et al.* 2005) to sequence data from these three species.

2.1.6 Age of selective sweep

We employed two methods to estimate the age of the selective sweep in the African sample: First, we utilized the rejection-sampling approach of PRZEWORKSI (2003). Briefly, this method generates a posterior distribution for the time since the fixation of a beneficial mutation based on summary statistics from the data: the number of polymorphic sites (S), the number of haplotypes (h) and Tajima's D (TAJIMA 1989). The method assumes that a neutrally evolving region is linked to a

site where a favorable allele deterministically reached fixation in the population at some time (T) in the past (PRZEWORKSI 2003). Furthermore, the neutral locus is presumed to evolve according to the infinite sites model. We ran the algorithm to obtain 2,000 successful matches. The time since the fixation of the selected site was then determined by finding the mode of a histrogram with bin size 0.01, in coalescent units of $3N_e$ generations (assuming $N_e = 10^6$).

Second, we assumed a star phylogeny subsequent to the sweep and computed the time since the sweep as T = number of mutations/(number of sites x sample size x mutation rate per year) (SLATKIN and HUDSON 1991, AYALA *et al.* 2002, BAUDRY *et al.* 2004). We assumed a mutation rate of 1.4 x 10^{-8} rec/bp/gen, which we estimated as $\mu = K/(2t)$, where K is divergence between *D. melanogaster* and *D. simulans*, μ is the mutation rate (in rec/bp/gen) and t is divergence time between the two species. Across the region studied, we empirically estimated $K = 0.063$ and assumed $t = 2.3$ million years (LI *et al.* 1999). Our estimate of μ is therefore in agreement with the chromosome wide average of μ = 1.45 x 10^{-8} rec/bp/gen obtained by LI and STEPHAN (2006).

2.1.7 Analysis of gene expression

The *wapl* region harbors several genes that code for proteins with enzymatic activity, namely *CG3835*, *Pgd*, *bcn92*, and *Cyp4d1*. We tested for expression differences of *CG3835*, *Pgd*, and *Cyp4d1* between our African and European lines by real-time quantitative reverse transcription PCR (qRT-PCR). Flies were raised at a constant temperature of 25°C and all lines were treated equally. We used a mix of 15 male and 15 female 4 – 5 days old flies per line for RNA extraction. Flies were snap frozen in liquid nitrogen and total RNA was extracted following the TRIZOL protocol (Invitrogen, Carlsbad, USA; see Appendix B6). We generated cDNA using random primers and ThermoScript reverse transcriptase (Invitrogen, Carlsbad, USA; see B7). Expression of target genes was measured relative to the expression of an endogenous control (*RpL32*) using TaqMan MGB probes (Applied Biosystems, Foster City). Runs were performed in triplicates per target gene on a 7500 Fast Real-Time PCR system (Applied Biosystems, Foster City) using the standard protocol (see Appendix B8). Amplification efficiency was tested for *Pgd* and *RpL32* by dilution series.

2.2 RESULTS

2.2.1 Polymorphism in the *ph-d – Pgd* region

In the previous chapter we reported evidence for a selective sweep having affected the *wapl* region of both an African and European population of *D. melanogaster*. Neutrality tests and recently proposed CLR tests (KIM and STEPHAN 2002, KIM and NIELSEN 2004) pointed toward a target of selection located in the center of the region. Since levels of nucleotide diversity were highly reduced in the proximity of the gene *ph-p* in the African population (see chapter 1), we sequenced the region between the 3'-flanking part of *ph-d* and the 3'-flanking part of *Pgd* in overlapping segments (Figure 4). This region comprises 31,700 annotated base pairs according to sequence AE0014298 of the *D. melanogaster* genome (release 4.3) available at GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). However, minor parts of the region were not sequenced due to primer malfunction. A list of fragment positions, compared sequence length and individual summaries is available in the Appendix (Table A1). In total we sequenced 53 loci with an overall length of 32,695 bp in 12 African and 6 European *D. melanogaster* lines. We also included fragment "8" from the previous chapter. Since *ph-d* and *ph-p* are paralogous genes (DURA *et al.* 1987, DEATRICK *et al.* 1991) sequence analysis can be exacerbated by false priming of oligonucleotides due to high levels of sequence similiarity. Therefore, we blasted each sequenced fragment against the published *D. melanogaster* genome sequence and only used those fragments that resulted in a blast identity of >99.9%. We identified two segments where primers erroneously annealed to the paralogous locus. These fragments were excluded from subsequent analyses. Excluding sequence overlaps and indels a total of 25,035 bp were used for further investigation. In this section we focus on 14,558 bp of noncoding data, partitioned into 12 loci (see Appendix A1).

As illustrated in Figure 5, levels of noncoding nucleotide variability across the region are heterogenous but low in the African sample, with particularly low heterozygosity in the vicinity of *ph-p*. Average θ across the region is 0.0038. This value is about 40% lower than the average level of heterozygosity reported for the 12 loci extending beyond the *ph-d – Pgd* region (see chapter 1). We observed 173 SNPs

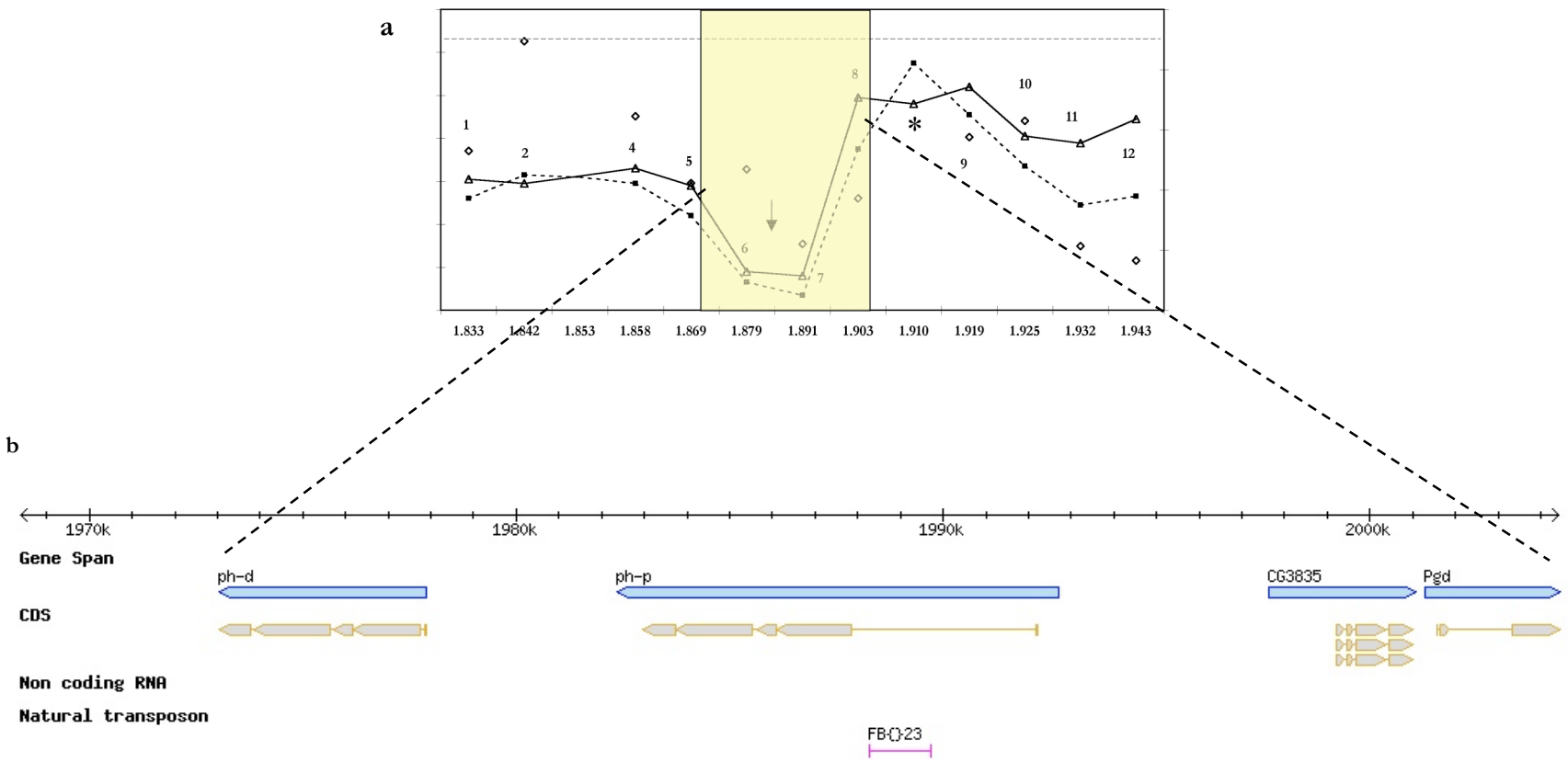


FIGURE 4. *ph-d – Pgd* region. The genes shown in the lower part (b) are located in the middle of the *wapl* sweep region (a), between loci 5 and 9. (a) is taken from chapter 1. The absolute genomic position is shown on top of (b). Gene models are in blue, coding sequences (CDS) in yellow. Alternative transcripts are illustrated below. Pink: annotated transposable element. The direction of transcription is indicated by the tip of the gene model feature. This figure was obtained using the FlyBase Genome Browser (http://www.flybase.org/cgi-bin/gbrowse/dmel/).

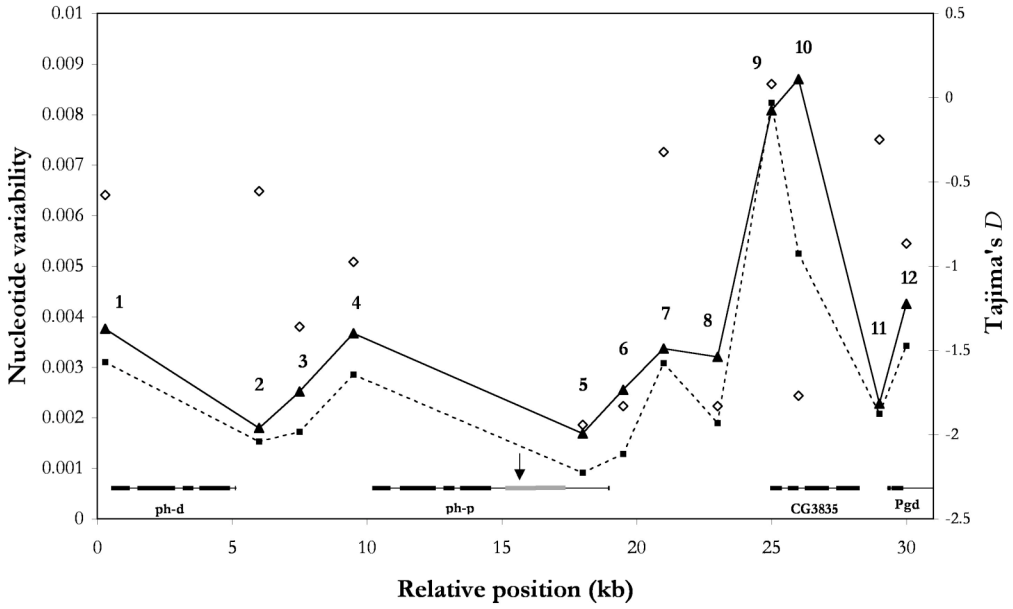


FIGURE 5. Nucleotide variability in the *ph-d – Pgd* region. Solid lines and triangles correspond to θ, dashed lines and squares indicate π. Diamonds correspond to Tajima's *D*. Genes across the region are depicted in the lower part. Solid boxes correspond to exons, lines between boxes indicate introns. The grey box indicates the annotated transposon. The arrow indicates the estimated position of the target of selection following the test of KIM and STEPHAN (2002).

in the African sample, and 201 polymorphic sites considering the joint datasets (see Appendix, Figure A1).

2.2.2 Haplotype structure

The number of haplotypes across the region ranges from 2 to 11 with an average of 8 haplotypes per locus. Haplotype diversity ranges from 0.3 to 0.985 (Table 4). None of the observed values was significantly different from standard neutral expectations. Linkage disequilibrium, as measured by KELLY'S (1997) Z_{nS} statistic, is generally low (average: 0.34) and only one value (locus 11) was significant. However, fragment 11 is short (291 bp) and only harbors two segregating sites, which are separated by 20 bp. For three loci (fragments 1, 5 and 6) no value could be obtained for Z_{nS}, as all polymorphic sites consisted of noninformative sites (*i.e.* singletons).

TABLE 4. Summary statistics for the loci in the *ph-d – Pgd* region (African sample)

Locus	Position [a]	n [b]	L [c]	S_{obs} [d]	h [e]	Hd [f]	θ	π	D [g]	F_S [h]	Z_{nS} [i]	L_{out} [k]	diff [l]	K [m]	Fu & Li D [n]	H [o]
1	1 - 515	12	276	3	4	0.682	0.0038	0.0031	-0.579	-1.048	NA	264	16	0.063	-1.123	0.545
2	5,246 - 6,351	12	1,106	6	8	0.894	0.0018	0.0015	-0.556	-4.678**	0.111	916	52	0.058	-1.072	-0.121
3	6,492 - 8,493	12	1,973	15	11	0.985	0.0025	0.0017	-1.360	-7.482**	0.128	1,350	79	0.059	-1.778*	1.818
4	8,585 - 10,122	12	1,533	17	7	0.879	0.0037	0.0029	-0.976	-0.369	0.456	1,472	101	0.071	-1.557	1.182
5	17,086 - 19,437	12	2,351	12	9	0.909	0.0017	0.0009	-1.942*	-5.487**	NA	2,340	101	0.044	-2.398*	-0.061
6	19,762 - 20,412	12	649	5	5	0.576	0.0026	0.0013	-1.831*	-2.373	NA	631	25	0.040	-2.668*	0.758
7	20,506 - 21,096	12	590	6	7	0.879	0.0034	0.0031	-0.324	-2.864	0.246	579	70	NA	-0.712	0.364
8	21,794 - 24,725	12	2,898	28	11	0.985	0.0032	0.0019	-1.831*	-5.107**	0.256	2,645	248	0.095	-1.857*	0.000
9	24,783 - 25,403	12	615	15	9	0.909	0.0081	0.0082	0.0810	-2.174	0.272	590	36	0.068	-0.429	-0.818
10	25,477 - 26,720	12	990	26	10	0.97	0.0087	0.0053	-1.769*	-3.497*	0.267	912	50	0.059	-2.774*	1.758
11	28,600 - 29,149	12	291	2	2	0.303	0.0023	0.0021	-0.248	1.384	1.000	275	20	0.074	0.947	0.485
12	28,921 - 30,413	11	1,286	16	9	0.945	0.0043	0.0034	-0.867	-3.211*	0.299	1,238	179	NA	-0.982	-0.764

[a] Position of fragment

[b] sample size

[c] number of bp used for computation

[d] number of observed polymorphic sites

[e] number of haplotypes

[f] haplotype diversity

[g] Tajima's *D*

[h] Fu's *F*s

[i] linkage disequilibrium

[k] number of bp used for estimation (with outgroup)

[l] number of fixed differences between species

[m] divergence between species

[n] Fu and Li's *D*

[o] Fay and Wu's *H*

* $P < 0.05$

** $P < 0.01$

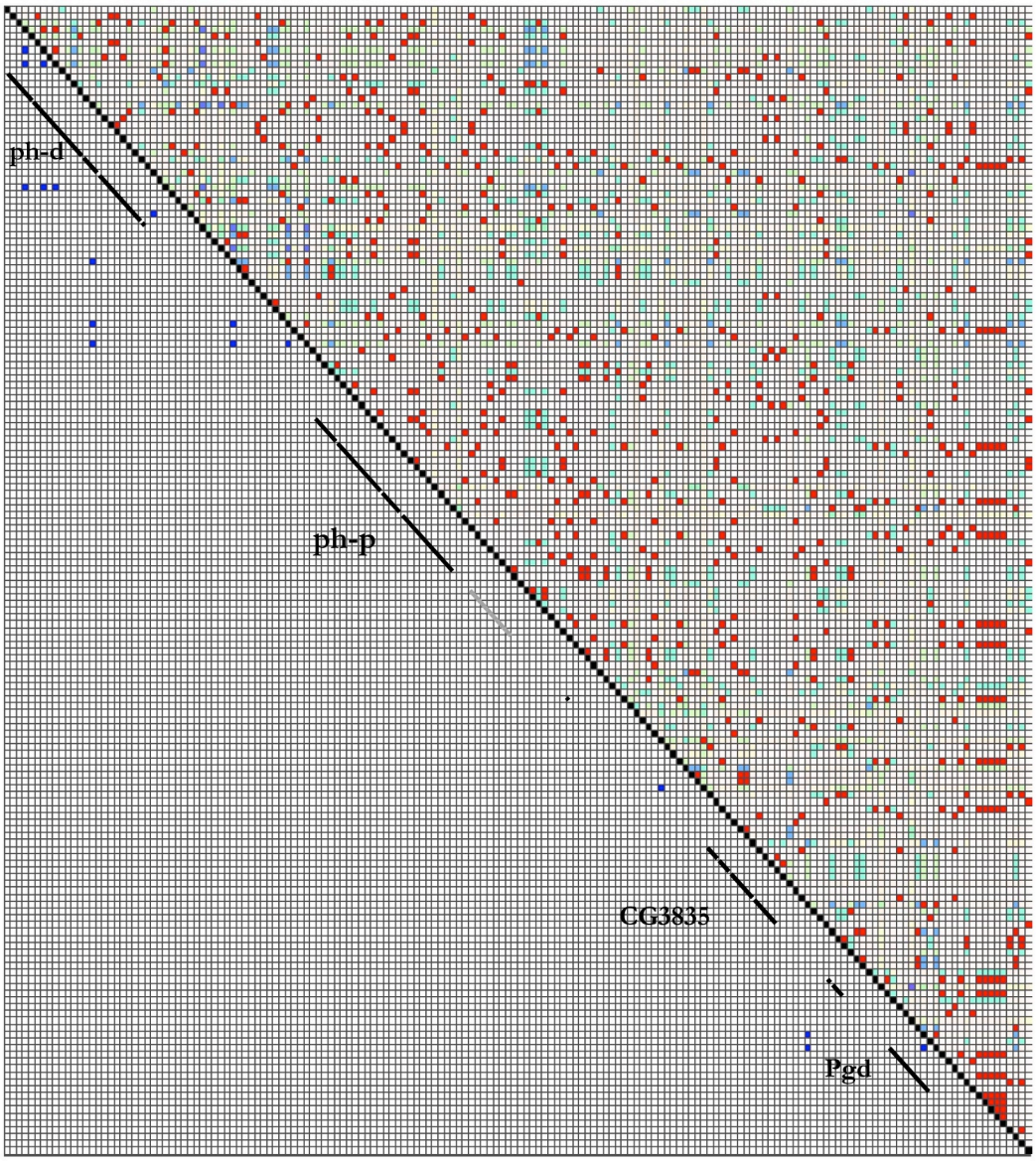


FIGURE 6. Linkage disequilibrium. Upper diagonal: pairwise r^2 , red: $r^2 = 1$, blue: $r^2 = 0.5$, green: $r^2 = 0.3$; Lower diagonal: significance, blue: $P < 0.01$. Exons are indicated by black boxes along the diagonal.

Two of these fragments (5 and 6) are located in the large intron and the 5'-flanking region of *ph-p*, respectively. We also estimated LD in terms of r^2 (HILL and ROBERTSON 1968) for all pairwise comparisons across the region. As depicted in Figure 6, only few significant associations were found. These are confined to the left and the right parts of the *ph-d* – *Pgd* region. Since haplotype diversity estimates were

generally high, we determined the minimum number of recombination events. A minimum of 23 recombination events were identified across the 25 kb investigated.

2.2.3 Standard neutrality tests

TAJIMA'S (1989) *D* was significantly negative ($P < 0.05$) for four loci in the region analyzed (D = -1.94, -1.83, -1.83, and -1.77 for fragments 5, 6, 8, and 10, respectively; Table 4). Additionally, all loci under investigation showed a trend toward a negative Tajima's *D* statistic, with exception of fragment 9 for which D = 0.08. Similarly, FU and LI'S (1993) *D* was significantly different from zero ($P < 0.05$) for five out of 12 loci (D = -1.78, -2.34, -2.67, -1.86, and -2.77 for fragments 4, 5, 6, 8, and 10, respectively). The trend toward negative *D* statistics indicates an excess of low frequency variants, *i.e.* singletons or doubletons. Significantly negative values of FU'S (1997) F_s further support this observation (Table 4). In contrast, we only observed four loci for which FAY and WU'S (2001) *H* statistic was negative (H = -0.12, -0.06, -0.82, and -0.76 for loci 2, 5, 9, and 12, respectively). None of these values was significantly different from standard neutral expectations. This observation indicates that there is no substantial skew toward high-frequency-derived variants across the *ph-d – Pgd* region. We also applied a multi-locus version of the HKA test (HUDSON *et al.* 1987) to our data and obtained a significant result (X^2 = 17.87, P=0.037). In addition, we analyzed our new data with the data from the previous chapter. The result obtained by this method was significant as well (X^2 = 32.79, P = 0.012). This indicates that the *ph-d – Pgd* region is not evolving according to the neutral equilibrium model. Note that we did not detect any deviation from the standard neutral model for the African sample previously (see chapter 1).

2.2.4 Likelihood and strength of selective sweep

We applied the CLR method of KIM and STEPHAN (2002) to our African data to test four different scenarios. First, we tested our new data (*ph-d – Pgd* region), assuming an expected heterozygosity of either $\theta = 0.0067$ (average across the region) or 0.0127 (average θ of the X-chromosome, GLINKA *et al.* 2003). Second, we analyzed our new data with data from chapter 1, which extends beyond the *ph-d – Pgd* region and encompasses ~110 kb. Again, we assumed $\theta = 0.0067$ or 0.0127. As

TABLE 5. Likelihood analysis of selective sweep

	θ = 0.0067					
Parameter	new data [a]	*P*	95% CI	new + old data [b]	*P*	95% CI
LR	77.48	<0.001		63.19	<0.001	
$\hat{\alpha}$	873.69			762.9		
$\hat{X}$	16,139		7,692 - 24,615	16,265		6,872 - 38,823
Λ_{GOF}	1,271.67	>0.9		1,616.80	>0.9	
	θ = 0.0127					
	new data	*P*	95% CI	new + old data	*P*	95% CI
LR	262.34	<0.001		246.81	<0.001	
$\hat{\alpha}$	2,132.90			2,211.67		
$\hat{X}$	15,829		5,761 - 27,892	15,919		583 - 40,720
Λ_{GOF}	1,308.65	>0.9		1,652.40	>0.9	

[a] only data from this chapter were used for analysis.

[b] data from this chapter were analysed in combination with data from chapter 1.

shown in Table 5, we obtained highly significant results for all scenarios tested ($P < 0.001$). Results were robust against simple demographic explanations (GOF *P*-value >0.9). In other words, the significant likelihood ratios are unlikely to be false positives, *i.e.* the result of population size bottlenecks. Estimates of the strengh of selection (α) are 763, 874, 2,133 and 2,212, respectively. Assuming $N_e = 10^6$ (PRZEWORSKI *et al.* 2001), the selection coefficients (s) are 0.0005, 0.0006, 0.0014 or 0.0015, respectively ($s = \alpha/1.5N_e$).

In addition, we applied the CLR test to polymorphism data obtained from the European sample. Given $N_e = 0.3 \times 10^6$ and assuming an expected heterozygosity (θ) of 0.004 (GLINKA *et al.* 2003) the test yielded a likelihood ratio (LR) of 172.75, which was significant in comparison to 10,000 simulated neutral data sets ($P <$

0.001). Furthermore, the GOF test (JENSEN *et al.* 2005) was nonsignificant (Λ_{GOF} = 27.43, P > 0.9). The strength of selection (α) was estimated to be 38,746, which results in s = 0.077.

2.2.5 Target of selection

The CLR test (KIM and STEPHAN 2002) also provides an estimate of the position of the target of selection. For our European sample the estimate of the position of the beneficial allele is 30,648 (95% CI: 209 – 30,483). This location corresponds to the third exon of *Pgd*. In contrast, as shown in Table 5 the likelihood ratio for the African sample is maximized within a narrow window between 15.8 and 16.2 kb, depending on the parameter combinations. Here all estimated positions correspond to the fourth intron of the gene *ph-p*, which is 4,291 bp in length according to release 4.3 of the *D. melanogaster* genome annotation (GRUMBLING *et al.* 2006). Since the 1,445 bp transposable element (TE) FB{}23 is annotated to be located within this intron (position 15,516 to 16,961), we tested all *D. melanogoster* lines and *D. simulans* for the presence of the transposon by diagnostic PCR. Given that the TE is present, the PCR product should be 2,166 bp in length (excluding oligonucleotide positions). However, the diagnostic PCR yielded a fragment of ~700 bp, which is the expected length when the transposon is absent. This result indicates that the TE is present neither in our *D. melanogaster* samples nor in *D. simulans*. Therefore, the *ph-p* intron and the gene itself are 1,445 bp smaller than annotated.

We scrutinized all five *ph-p* exons, the four introns and the 5'-flanking region for fixed differences along the *D. melanogaster* lineage, using *D. simulans* and *D. yakuba* as outgroups. We detected several small indels within exons 3 and 4. However, none of these differentiated *D. melanogaster* from both *D. simulans* and *D. yakuba*. The second exon is most diverse and contains large indels that are private to one of the species. At position 14,863 *D. simulans* features a fragment of 576 bp that is not present in *D. melanogaster* or in *D. yakuba*. At position 14,368 *D. yakuba* has a 64 bp fragment, which is not present in either of the other two species. Sequence comparions are complicated, however, since the third exon and parts of the second and fourth exons are not yet annotated in *D. simulans*. We could not detect any indels within introns 2, 3 and 4. Yet, the first intron harbors several small indels that are

present in only one of the three Drosophila lineages. In contrast, at positions 17,188 and 17,190, *D. yakuba* and *D. simulans* have insertions of 39 and 28 bp, respectively, that are not present in *D. melanogaster*. Between positions 18,669 and 18,933 *D. yakuba* and *D. simulans* sequences are very different from the *D. melanogaster* sequence and feature three fragments of 3, 10, and 36 bp, respectively, that were lost along the *D. melanogaster* lineage. In summary, this indicates that the first intron and the adjacent second exon have a complex evolutionary history. To evaluate the significance of fixed differences observed between the three Drosophila lineages we screened the first intron and the 5'-flanking region for differences in transcription factor binding sites (TFB). We identified 31 TFBs with a matrix score >0.9. The matrix score indicates how well a sequence of interest fits a known TFB sequence. Potential TFBs are depicted in the Appendix (Figure A2). Out of 31 TFB found, 11 are either absent or present in the *D. melanogaster* lineage only, *i.e.* the specificity occurred after the *D. melanogaster* – *D. simulans* split (~2.3 mya, LI *et al.* 1999). All TFB changes detected involve homeotic genes or various zinc finger proteins. In the first intron, between positions 18,657 and 18,677, a zinc finger TFB is predicted for *D. melanogaster*. However, an insertion in *D. yakuba* and *D. simulans* prevents the prediction. Here a different zinc finger splice variant (CF2-II) binding site is proposed instead. In addition, a *dead ringer* gene TFB is predicted for *D. melanogaster* with high sequence similarity (0.99). This gene is a co-factor in embryonic gene expression and posesses a highly conserved DNA binding domain (GREGORY *et al.* 1996, SHANDALA *et al.* 1999). A substitution in *D. yakuba* and *D. simulans* prevents the TFB predicition in these species. Furthermore, a *fushi tarazu* TFB is predicted for the 5'-flanking region of *ph-p*. In contrast, a single substitution in the *D. yakuba* and *D. simulans* lineages results in a *caudal* binding site. Finally, a GAGA factor (GAF) binding site is exclusively predicted for *D. melanogaster* in the *ph-p* 5'-region.

Between the African and the European samples several fixed or nearly fixed differences were identified. At 17 sites, one variant is at high frequency in one sample, but at low frequency (or absent) in the other. Only one of these is a nonsynonymous change, *i.e.* it is associated with an amino acid difference. This G → A transition at position 61,456 (fourth exon of *ph-p*), results in a substitution of threonin by isoleucin. In addition, we found a difference between our two samples in

the *Pgd* 5' – region. More specifically, there are two polymorphic sites that are located between the TATA box and the predicted transcription start point (SCOTT and LUCCHESI 1991). Here, two variants at positions 78,998 and 79,018, respectively, are at high frequency in Europe, but at low frequency in Africa (see Figure A1, Appendix).

2.2.6 Age of selective sweep

We estimated the age of the selected sweep for the African population using two different approaches: First, we applied the rejection-sampling method of PRZEWORSKI (2003) to our data. This method estimates the time since the fixation of a beneficial allele using summaries of the data (*i.e.* the number of SNPs, the number of haplotypes observed, and Tajima's *D*) and the distance of the neutral region of interest to the selected site. We ran the algorithm for three putative neutrally evolving regions: Locus 3 is located 7,500 bp distally from the middle of the first *ph-p* intron, fragment 7 is located 4,200 proximally, and locus 8 is located 5,500 bp proximally. The target of selection was assumed to be located anywhere within this intron of *ph-p*. We obtained the following average age estimates (95% CI): 5,380 (2,920 - 15,550) years for locus 3, 11,500 (6,150 - 74,750) years for fragment 7, and 5,490 (3,050 - 12,400) years for locus 8, respectively (assuming $N_e = 10^6$ and 10 generations per year). Results are summarized in Table 6.

Second, we estimated the time since the fixation of the beneficial allele following SLATKIN and HUDSON (1991) and AYALA *et al.* (2002). This method assumes a star-like phylogeny, *i.e.* a complete sweep and subsequent accumulation of new mutations, where each mutation creates a new allele. Applying the method to fragments 5 and 6, *i.e.* the center of the African sweep region, which harbors 17 singletons in 2,969 bp, we obtained an age estimate of 34,840 years.

We also applied this method to polymorphism data of the Europan sample. Here we only observed seven segregating sites (six singletons and a doubleton) in total. When using all SNP data ($S = 7$ and $L = 25{,}035$ bp) we estimate an age of 3,403 years. However, when we restrict our analysis to the six singletons only, we obain $T = 1{,}745$ years (given $S = 3$ and $L = 20{,}923$ bp).

TABLE 6. Age of the selective sweep (following PRZEWORSKI 2003)

Locus	Position [a]	*L* [b]	Distance [c]	*T* [d]	95% CI [e]
3	6,492 - 9,195	1,973	8,000	5,515	2,839 - 16,334
			7,000	5,015	2,921 - 16,141
			6,000	5,494	2,991 - 14,594
			5,000	5,476	2,945 - 15,135
7	20,506 - 21,096	590	4,000	12,499	6,075 - 77,197
			2,000	10,558	6,217 - 72,297
8	21,794 - 24,725	2,898	6,000	5,476	3,116 - 12,525
			5,000	5,497	3,016 - 12,173
			4,000	5,494	3,056 - 12,505
			3,000	5,487	3,025 - 12,422

[a] location of the fragment subjected to the test.

[b] length in bp used for computation.

[c] distance from the middle of the first *ph-p* intron.

[d] estimate of the time since the fixation of the beneficial allele (in years).

[e] 95% confidence interval for *T* (in years).

2.2.7 Analysis of gene expression

Since our previous results indicate that the target of selection in the European sample might be different from the one estimated for the African sample, we investigated whether significant gene expression differences can be found between these samples. This is reasonable, since it has been observed that differences in gene expression between populations can be caused by positive directional selection (*e.g.* DABORN *et al.* 2002, BETRÁN and LONG 2003, SCHLENKE and BEGUN 2004, HARR e*t al.* 2006). As several enzyme-encoding genes are present in the *wapl* region, we chose to analyze their expression pattern by qRT-PCR. We concentrated on three genes (*CG3835*, *Pgd*, and *Cyp4d1*), as TaqMan probes were available for these. Prior to our analyses we tested whether endogenous control (*RpL32*) and

target genes possess the same amplification efficiencies by regression analyses of a dilution series. This was done to exclude inhibitory effects due to target concentration or secondary structure. We could not detect any inhibitory effects. Target and endogenous control showed the same amplification efficiency ($P < 0.01$). We subsequently tested for expression differences of the endogenous control among samples. No significant expression differences for *RpL32* were detected ($P = 0.51$). This is in accordance with a previous study using *RpL32* as endogeneous control (DABORN *et al.* 2002). We therefore used this gene to normalize expression differences of target genes between lines. Normalization was done to correct for general variation in expression among individuals. Expression patterns are depicted in Figure 7. We observed significant expression differences for *Pgd* ($P < 0.001$) and *CG3835* ($P = 0.03$). For the latter we performed a Welch ANOVA, as variances were significantly different between samples ($P = 0.03$, Levene's test). In contrast, no difference in expression was found for *Cyp4d1* ($P = 0.81$).

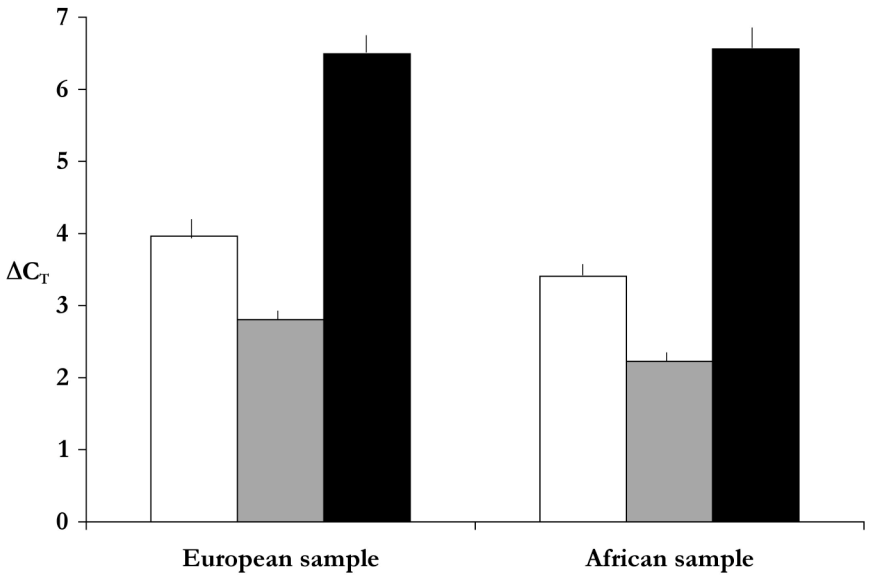


FIGURE 7. Analysis of gene expression. White bars: *CG3835*, grey bars: *Pgd*, black bars: *Cyp4d1*. ΔC_T is the difference in threshold cycles (C_T) for target gene and endogenous control. The C_T denotes the fractional cycle number at which the fluorescence signal passes a defined threshold. The earlier the threshold is passed, the higher is the amount of cDNA. Average (SE) ΔC_T estimates are: 3.96 (0.18), 2.81 (0.08), and 6.49 (0.16) for *CG3835*, *Pgd*, and *Cyp4d1*, respectively, in the European sample. For the African sample we obtained 3.41 (0.10), 2.22 (0.09), and 6.56 (0.21), for *CG3835*, *Pgd*, and *Cyp4d1*, respectively.

2.3 DISCUSSION

In this study we revisited the *wapl* region analyzed in the previous chapter. To investigate the processes that shaped nucleotide variability in this particular region of the genome, we sequenced the valley of reduced variation in our African sample and half of the European sample. Since parts of *ph-d* and *ph-p* are highly conserved between these genes and poly – A/T stretches exacerbated sequence analysis, we excluded 19% of the sequence data of the *ph-d – Pgd* region from further analyses. Our main interest was to evaluate polymorphism in the region to detect further evidence in favor of positive directional selection and to locate the putative target of selection in the African sample. In addition, we asked whether a different target of selection has to be proposed for the European sample.

2.3.1 Positive selection in the *ph-d – Pgd* region

Consistent with our expectation of reduced heterozygosity between the genes *ph-d* and *Pgd* (see previous chapter) we detected nucleotide variation 70% below the average across the X-chromosome (θ = 0.0127, GLINKA *et al.* 2003) in our African sample. In 14,558 bp analyzed, we identified only 173 segregating sites. This number is low, yet not significantly different from neutral equilibrium expectations (P = 0.11). However, we did find evidence for a selective sweep at *ph-p*. First, nucleotide diversity (π) is particularly low in the first intron of *ph-p*, the *ph-p* 5'-flanking region, and the distal part of the *ph-p – CG3835* intergenic region, where low frequency variants predominate (*i.e.* loci 5, 6, and 8, see Appendix, Figure A1). This is supported by significantly negative Tajima's *D* and Fu and Li's *D* at *ph-p*. This pattern is in accordance with genetic hitchhiking at *ph-p* or a locus close to that gene. With the exception of two fragments (loci 5 and 9) we did not observe negative values for Fay and Wu's *H*. However, the power of this statistic to detect selection drops off quickly after the fixation of a favorable allele (PRZEWORSKI 2002). Second, we detected 23 recombination events across the region. None of these were located within *ph-p*. Yet, several recombination breakpoints were identified in flanking regions. We also observed significant LD (in terms of r^2) in fragments flanking *ph-p*. This is in agreement with theoretical predictions of increased LD in flanking regions,

yet reduced levels of LD at the target of selection (KIM and NIELSEN 2004, STEPHAN *et al.* 2006). Third, we obtained CLR test (KIM and STEPHAN 2002) results in favor of the selective sweep model.

As the above mentioned test statistics assume that the population is evolving according to the standard neutral model, deviations from neutral equilibrium expectations can also be the result of demographic forces (*i.e.* population size expansion or a bottleneck). However, this is unlikely to be true for the *ph-d – Pgd* region. First, the genome wide pattern of nucleotide variation in our African sample is not consistent with a recent population bottleneck (HADDRILL *et al.* 2005, OMETTO *et al.* 2005, LI and STEPHAN 2006). With the exception of locally increased LD, some features of our data show signatures of a populations size expansion instead, as revealed by negative Fu's F_s and negative Tajima's *D*. However, loci that show significantly negative Tajima's *D* are concentrated at *ph-p*. In addition, no other low- or intermediate frequency variants are found in an approx. 5 kb-stretch, with the exception of a doubleton at position 17,230. Here, the frequency spectrum deviates the most from standard neutral expectations. Furthermore, at *ph-p* and the 5'-flanking region, the observed frequency spectrum is also different from that obtained from other, putatively neutrally evolving loci in the African sample or that obtained from the *Pgd* intron (Figure 8). In summary, our data clearly reflect the genome wide pattern of population size expansion but certain parts of the region investigated (*i.e.* the *ph-p* region) are unlikely to be shaped by demography alone.

2.3.2 Strength of selection

Previous studies revealed that the strength of selection strongly influences the extent to which nucleotide variability is depleted and, in addition the spatial extent of reduced heterozygosity (*e.g.* KAPLAN *et al.* 1989, STEPHAN *et al.* 1992, KIM and STEPHAN 2002). Conversely, the accuracy of estimating the strength of selection depends on the data subjected to CLR tests (J. Jensen, pers. comm.). If the target of selection and adjacent regions were sequenced, statistical inference can be based on more information and the estimates obtained will have a higher precision. We therefore used two approaches to evaluate our data from the African sample. We ran the CLR test (KIM and STEPHAN 2002) with and without the data gathered

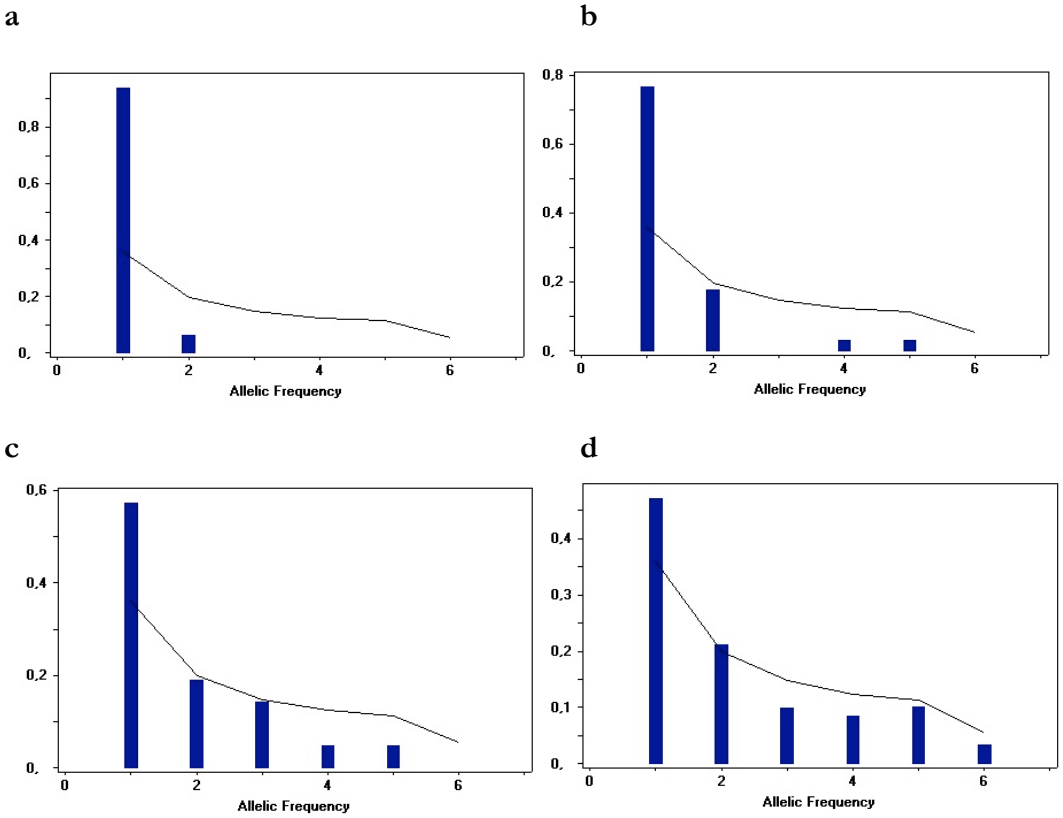


FIGURE 8. Frequency spectra. Panel a: first intron of *ph-p* and *ph-p* 5'-region. Panel b: *ph-p – CG3835* intergenic region. Panel c: *Pgd*, second intron. Panel d: 27 neutral loci (see chapter 3). Bars indicate relative counts per frequency class. Lines denote expected equilibrium frequencies.

previously (chapter 1). However, estimates of the strength of selection only varied marginally between both approaches (s = 1.4x10^{-3} and 1.5x10^{-3} assuming θ = 0.0127 and N_e = 10^6, respectively). This may be explained by the fact that the width of the valley of reduced variation has already been captured, and an additional reduction of variation was not detected in our new dataset. We also detect moderate selection coefficients, in accordance with the size of the region of reduced variation (KAPLAN *et al.* 1989). For the European sample we estimated a much larger selection coefficient (s = 0.077). Given N_e = 0.3x10^6 (GLINKA *et al.* 2003) and ρ = 0.48x10^{-8} rec/bp/generation (COMERON *et al.* 1999) a region of about 120 kb should be affected by a sweep with s = 0.077 (KAPLAN *et al.* 1989). However, we observed depleted nucleotide diversity across only ~60 kb (see chapter 1). This might indicate

that additional processes shaped the *ph-d – Pgd* region in the European *D. melanogaster* population (see below).

2.3.3 Localizing the target of selection

Our sequence data obtained from the African sample indicate that the selected site is located in the proximity of *ph-p*. This is also supported by common summary statistics and the CLR test of KIM and STEPHAN (2002). The *ph-p* locus constitutes one part of the duplicated *ph* gene, and the distally located *ph-d* the other (HODGSON *et al.* 1997). *Ph* is a locus within the Polycomb group (PcG), which contains >20 genes with similar phenotypes that are required for normal segmental specification during Drosophila embryonic development (DURA *et al.* 1987). PcG gene products act as transcriptional repressors, *i.e.* they are involved in gene silencing of homeotic loci (HODGSON *et al.* 1997, BANTIGNIES *et al.* 2003). In contrast, *Trithorax*-group (TrxG) proteins promote maintanance of gene activity (LUND and VAN LOHUIZEN 2004). Recent studies revealed a complex interaction of these groups and suggest that PcG and TrxG genes may function in a concerted fashion (reviewed in LUND and VAN LOHUIZEN 2004). It has been proposed that the *ph* duplication was ancient and the accumulated sequence divergence reflects differences in function (HODGSON *et al.* 1997). We therefore evaluated the intron – exon structure of *ph-p* for fixed differences between the *D. melanogaster* lineage and its sibling species *D. simulans* and *D. yakuba*, which result from positive selection.

Although sequence comparison to *D. simulans* was limited due to sequence annotation, we discovered major dissimilarities between the three sibling species in the first intron and the second exon of *ph-p*. *D. simulans* and *D. yakuba* both harbor an insertion in exon 2, which is not present in their congener *D. melanogaster*. However, nucleotides are inserted at different sites and therefore no uniqueness can be inferred with respect to the *D. melanogaster* lineage. In contrast, we detected several differences in putative transcription factor binding sites in the first intron of *ph-p* and the 5'-flanking region. As large introns and 5'-regions are known to contain *cis*-acting regulatory elements (AYALA *et al.* 2002, HADDRILL *et al.* 2005, KALARI *et al.* 2006) our results might indicate that *ph-p* has acquired specific functions. In the following we

briefly discuss some of the observed differences in TFBs and their potential relevance.

In the 5'- flanking region an A → T substitution in the *D. melanogaster* lineage prevents the prediction of a *caudal* TFB but creates a *fushi tarazu* (*ftz*) binding site instead. *ftz* interacts with the homeotic gene *engrailed*, which in turn regulates *ph* (*i.e*, *ph-d* and *ph-p*; SERRANO *et al.* 1995). In addition, *ftz* is needed for the correct activation of *Antennapedia* and *Bithorax* complex genes (INGHAM and MARTINEZ-ARIAS 1986, DURA *et al.* 1987). The formation of an *ftz* binding site might therefore point toward an interaction of *ftz* and *ph-p* for the concerted regulation (silencing *vs.* activation) of *Antennapedia* and *Bithorax* group genes. Furthermore, in *D. melamogaster* a GAGA factor binding site was identified immediately 5' to *ph-p*. This factor is required for the efficient silencing of homeotic genes (LEHMANN 2004). Within the first intron several TFB differences were found between species. For example, a *dead ringer* binding site was identified in *D. melanogaster* with high probability. This gene harbors a highly conserved DNA-binding domain and interacts with other homeodomain proteins (GREGORY *et al.* 1996). Moreover, within this intron binding sites for different zinc finger isoforms were identified for *D. melanogaster* and *D. yakuba* / *D. simulans*, respectively. In summary, our results may indicate that the observed differences in TFBs result in changes in the regulation of *ph-p*. Consequently, *ph-p* may have gained additional functions along the *D. melanogaster* lineage or is specializing on different functions than *ph-d*, as has been suggested by HODGSON *et al.* (1997). A similar case, where a gene duplicate is involved in a selective sweep, has been reported previously (NURMINSKY *et al.* 1998).

For the European sample we obtained a somewhat different estimate of the position of the target of selection. This is probably due to the fact that the region of substantially reduced nucleotide variation is much greater than in the African sample. This fact may also explain the larger selection coefficient obtained from our European data. However, it is difficult to separate the effects of demography and selection in the ancestor, both of which may have affected heterozygosity in the *wapl* region of European *D. melanogaster*. Therefore, a locus-specific reduction of polymorphism due to directional selection in a derived population is not easy to detect, when selection in the ancestor has already shaped this and other loci in the

vicinity (see HAMBLIN *et al.* 2006). Yet, since the haplotype structure of the *Pgd* region is particularly different between samples and features many nearly fixed differences between Africa and Europe, we investigated whether differences can be observed in the phenotype as well. We detected two genes, *CG3835* and *Pgd*, that are differentially expressed between samples. At one of these genes, *Pgd*, we also found SNP differences in the promotor region between samples. Our preliminary analysis therefore suggests that an additional selective sweep occurred in the European sample, favouring a variant that is at low frequency in the ancestral population. However, further analyses of gene expression patterns are needed to confirm an association of SNP variants and expression patterns. In particular, the effect of *trans* factors, which can affect gene expression as well, cannot be ruled out. However, it has been shown that gene expression is mainly governed by *cis* – located elements in Drosophila (WITTKOPP *et al.* 2004, OSADA *et al.* 2006), maize (STUPAR and SPRINGER 2006), yeast (RONALD *et al.* 2005) and humans (STRANGER *et al.* 2005).

2.3.4 Time since the fixation of the selected site

As the *D. melanogaster* and *D. simulans* lineages separated ~2.3 mya ago (LACHAISE *et al.* 1988, LI *et al.* 1999), we asked when the positively selected mutation was fixed along the *D. melanogaster* lineage. Our results imply that the subsitution occurred in the fairly recent history of *D. melanogaster*. Time estimates based on the method of PRZEWORSKI (2003) range between 2,920 and 74,750 years (all confidence intervals overlap). However, the PRZEWORSKI (2003) approach is sensitive to demography, in particular to population size expansion, as summary statistics such as Tajima's *D* and the number of haplotypes are utilized in the rejection-sampling algorithm (PRZEWORSKI 2003). In addition, it has been suggested that the effective population size of African *D. melanogaster* is $>10^6$ (THORNTON and ANDOLFATTO 2006). Our estimates of sweep age might therefore represent underestimates. Furthermore, following SLATKIN and HUDSON (1991), we obtained a sweep age of 34,840 years. Yet, this approach is affected by demography as well, since the number of new mutations is a critical parameter in this method (see MATERIAL AND METHODS). We therefore suggest that our estimates of sweep age represent minimum approximations. This is supported by our results from the European

sample, which shows joint effects of selection and demography (*i.e.* an out-of-Africa bottleneck ~10,000 – 20,000 years ago, OMETTO *et al.* 2005) in the *ph-d – Pgd* and the larger *wapl* region.

2.4 SUMMARY

In this study we revisited the center of the *wapl* region analyzed in chapter 1. We concentrated on the African *D. melanogaster* sample, as the valley of reduced variation was much narrower than in the European sample, which should help to pinpoint the target of selection. Our new results confirm our previous conclusions about selection having shaped nucleotide variability in this part of the *D. melanogaster* genome. Moreover, by sequencing large parts of the center of the selective sweep we were able to establish the haplotype structure in that region and to infer the historical context of the sweep. We conclude that a positively selected substitution occurred at *ph-p* and was fixed before the out-of-Africa expansion of *D. melanogaster*, possibly >30,000 years ago. This subsitution might be associated with the specialization of *ph-p* in gene regulation. Sequence analysis of *ph-p* in *D. simulans* and subsequent comparion of orthologous and paralogous sequences might further elucidate the mode of evolution at *ph-p*. In addition, our results obtained from the European sample indicate that sequence variation was not affected by demography alone. In fact, we found that selection affected nucleotide diversity in the *ph-d – Pgd* region of the European sample as well. Since heterozygosity across the whole region is substantially reduced, we propose that an additional selective sweep occurred at a different site in Europe. This is supported by our analysis regarding the time since the fixation of the beneficial mutation at *ph-p*, which points toward a substitution in *D. melanogaster* before the colonization of Europe took place.

3 Population structure of Southeast Asian *D. melanogaster*

It is widely accepted that non-African populations of *D. melanogaster* were established about 10,000 years ago (DAVID and CAPY 1988, LACHAISE *et al.* 1988, LACHAISE and SILVAIN 2004). The rationale is that, as a strict human commensal with an African origin, *D. melanogaster* had a chance to colonize non-African habitats only after the last glaciation. The most recent glacial maximum began to recede about 17,000 years before present (BP) and current sea levels were reached *ca.* 6,000 years ago (FAIRBANKS 1989, WEBB and BARTLEIN 1992). Improving climatic conditions and the spread of agriculture throughout Western Europe, which started about 10,000 years ago in the Northern Levantine (also known as the Fertile Crescent, PINHASI *et al.* 2005), probably enabled Drosophila to disperse throughout this region and finally colonize other parts of the world (DAVID and CAPY 1988). Recent estimates of the colonization time of Europe based on population genetic data (*e.g.* GLINKA *et al.* 2003, OMETTO *et al.* 2005) support this view. In addition, genomic candidate regions have been identified that were possibly shaped by positive Darwinian selection enabling local adaptation (HARR *et al.* 2002, OMETTO *et al.* 2005, LI and STEPHAN 2005, POOL *et al.* 2006, LI and STEPHAN 2006).

In previous work we scrutinized a genomic region on the X chromosome in an African and a European *D. melanogaster* population that has most probably been shaped by positive selection (see chapters 1 and 2). Initially we could not determine whether the beneficial mutation occurred just once, namely in the African population, thus causing a transpopulation sweep, or an independent sweep occurred in the European population at a later date. It has recently been argued that demographic forces, such as severe population size bottlenecks, can cause a pattern of local genetic variation that closely resembles that produced by a selective sweep (*e.g.* JENSEN *et al.* 2005). A strong reduction in local genetic diversity in a derived

population could thus be the product of low nucleotide variability in the ancestral population amplified by demography. In chapter 1 we concluded that nucleotide variability in the *wapl* region of the European sample is unlikely to be the result of demography alone. Furthermore, in chapter 2 we reasoned that an independent selective sweep presumably occurred at *Pgd* in the European population. However, it is difficult to determine to what extent our data can sufficiently be explained by demography and which features can be attributed to selection, since both forces shaped nucleotide variability in the *wapl* region.

One possibility to address this issue is the analysis of additional population samples from different geographic regions. The rationale is as follows: If selection occurred only once, namely in the ancestral population (at *ph-p*), no additional footprint of selecton should be visible in the *wapl* region of derived populations. However, if levels of heterozygosity at *Pgd* were also shaped by selection in the European sample, then this signature should not be visible in additional, derived population samples from distinct geographical regions, given that their demographic histories were not significantly different. In addition, it has been proposed that the analysis of highly substructured populations could generally faciliate the identification of targets of selection by reducing the region affected by genetic hitchhiking (SANTIAGO and CABALLERO 2005).

Until now population genetic studies have concentrated on European and North American *D. melanogaster* populations. Asian samples, on the other hand, have thus far mainly been characterized on the phenotypic level (reviewed in LACHAISE and SILVAIN 2004). These studies, focusing on morphological and physiological characters, revealed considerable differences between Asian populations and those from other geographical regions (DAVID *et al.* 1976, LEMEUNIER *et al.* 1986). In addition, they showed that there is substantial morphological variation among Asian populations alone. However, the genetic diversity of these populations has so far remained unexplored in large part. It was only recently that three studies analyzing mtDNA, inversion polymorphism or microsatellite variation, were published (SOLIGNAC 2004, GLINKA *et al.* 2005, and SCHLÖTTERER *et al.* 2006, respectively). Despite this, for further consideration of DNA sequence data and SNP analysis of

candidate loci the standing level of genetic variation and the degree of population differentiation have to be assessed first.

In this study we analyze DNA sequence variation using SNP data from 10 noncoding X-linked loci obtained from six SE Asian population samples. To these data we added previously published data from an African and a European sample (OMETTO *et al.* 2005) to assess population structure among geographical regions. In addition, we investigate the demographic history of the SE Asian samples. Finally, we re-examine the *wapl* sweep region studied in chapter 1.

3.1 MATERIAL AND METHODS

3.1.1 Fly samples

For intraspecific analyses we used *D. melanogaster* flies from Chiang Mai (CNX, Thailand), Bangkok (BKK, Thailand), Kuala Lumpur (KL, Malaysia), Kota Kinabalu (KK, Malaysia), Manila (MNL, Philippines) and Cebu (CEB, Philippines), which were sampled in October 2002 and kindly provided by A. Das. Detailed information on sampling locations are given in the Appendix (Table A2), but are also shown in Figure 12. For interspecific comparisons we used the publicly available DNA sequence of the *D. simulans* genome. Additional sequences from an African (ZBM, Zimbabwe) and a European (EUR, The Netherlands) sample were taken from OMETTO *et al.* (2005) and added to our data.

3.1.2 Molecular techniques

We used a single male fly from each isofemale line for isolation of genomic DNA with the Puregene DNA isolation kit (Gentra Systems, Minneapolis, USA). PCR set-up and conditions were as described in chapter 1. PCR fragments were scored on 1.5% agarose gels and purified using ExoSap-IT (USB, Cleveland, USA). Sequencing reactions were conducted for both strands with the ABI BigDye Terminator v1.1 sequencing kit and read on an ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, USA). Protocols are provided in the Appendix (B). Sequences were aligned and checked manually using SeqMan and finally assembled

into contigs with MegAlign (DNAstar, Madison, USA). In addition, we used SeaView (GALTIER *et al.* 1996) for sequence inspection and formatting.

3.1.3 Analysis of genetic variation

Basic population genetic parameters were estimated using DnaSP 4.10 (ROZAS *et al.* 2003). We estimated nucleotide variability using θ (WATTERSON 1975) and π (TAJIMA 1983). These parameters were estimated for each population separately and subsequently averaged to avoid the Wahlund effect (WAHLUND 1928, HARTL and CLARK 1997). In addition, we determined divergence (*K*) between *D. melanogaster* and its congener *D. simulans*, the number of distinct haplotypes (*h*) and haplotype diversity (*Hd*; NEI 1987). Linkage disequilibrium (LD) was estimated per locus in terms of Z_{nS} (KELLY 1997) using informative sites only (*i.e.* we omitted singletons, indels and sites for which more than two variants were present). Frequency spectrum information was obtained using DnaSP 4.10 (ROZAS *et al.* 2003). To test the standard neutral model, we applied the following statistics: Tajima's *D* (TAJIMA 1989), Fu and Li's *D* (FU and LI 1993), Fay and Wu's *H* (FAY and WU 2000), and the multi-locus-HKA statistic (HUDSON *et al.* 1987). A program to calculate the latter was kindly provided by J. Hey. Furthermore, we used multi-locus coalescent simulations to determine the probababilty of Z_{nS} (KELLY 1997), Tajima's *D* (TAJIMA 1989), *h* and *Hd* (NEI 1987) estimates under neutral equilibrium conditions. Programs used for this analysis were developed by S. Ramos-Onsins and kindly provided by L. Ometto. Simulated data were generated using the observed θ values. Simulations were run under the assumption of no recombination, except for LD where we allowed for intralocus recombination with given rate $R = 2N_e\rho$, where N_e is the effective population size and ρ the local recombination rate as estimated following COMERON *et al.* (1999). This approach is conservative, as recombination tends to reduce LD. Having no prior information on the SE Asian samples, we assumed $N_e = 10^6$ (LI 1999).

3.1.4 Population structure

We tested for population substructure following three approaches. First, we employed a hierarchical locus-by-locus analysis of molecular variance (AMOVA, EXCOFFIER *et al.* 1992) to partition the total variance components into those derived within and among *a priori* defined groups. In addition, we computed pairwise F_{ST} values to estimate genetic divergence among population samples. Significance of fixation indices was tested by permuting haplotypes 10,000 times among populations, since this method does not rely on Hardy-Weinberg assumptions (GOUDET *et al.* 1996, EXCOFFIER *et al.* 2005). For both methods we used ARLEQUIN 3.0 (EXCOFFIER *et al.* 2005). F_{ST} estimates between pairs of populations were used to construct an unrooted neighbor-joining (NJ) tree (SAITOU and NEI 1987) using MEGA 3.1 (KUMAR *et al.* 2004). Population trees were displayed with TreeView (PAGE 1996). Second, we performed a Bayesian clustering and admixture analysis using STRUCTURE (PRITCHARD *et al.* 2000). Briefly, this method uses multi-locus genotypes to infer population structure and simultaniously assign individuals to populations. The model assumes that there are *C* distinct populations, where *C* may be unknown. Furthermore, Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE) are assumed among the unlinkend marker loci (PRITCHARD *et al.* 2000). Individuals are assigned to populations in such a way as to achieve these equilibria (PRITCHARD *et al.* 2000). Each population is characterized by a set of allele frequencies at each locus. The number of clusters is inferred by estimating the probability $P(X|C)$ of the data given a certain uniform prior value of *C* over a number of Markov Chain Monte Carlo (MCMC) iterations. The posterior probabilities $P(C|X)$ can then be calculated following Bayes' rule. Individuals in the sample are assigned to populations probabilistically, regardless of their geographical origin, or jointly to two or more populations if their genotypes indicate that they are admixted (see RANDI *et al.* 2001). We assumed prior values of *C* ranging from 2 to 10. All simulations were done using a burn-in phase of 10^5 iterations, followed by 10^6 iterations of the MCMC procedure. We applied the algorithm five times for two priors of *C* (2 and 4, respectively) to assure homogeneity over runs. All simulations reported were run under the "admixture model", *i.e.* assuming that individuals have

inherited some fraction of their genomes from ancestors in population *c* (PRITCHARD *et al.* 2000). We did not incorporate any other prior population information.

Third, we applied the method of VOGL *et al.* (2003). This approach estimates the degree of population subdivision using an MCMC procedure. For each subpopulation the differentiation from the reconstructed migrant pool, *i.e.* the ancestral population before the split into subpopulations, is denoted by the migration-drift parameter θ_p. In addtion, molecular variation in the migrant gene pool can be estimated, which allows the computation of common summary statistics, such as Tajima's *D* (TAJIMA 1989). Thus, global demographic processes of the entire population over longer time periods can be evaluated.

3.1.5 Demographic analysis

We investigated the historic relationship of the ZBM, EUR, and KL samples following the joint mutation frequency spectrum approach of LI and STEPHAN (2005), which was extended to three populations. We tested two demographic scenarios. The general outline is depicted in Figure 9. In brief, we investigated whether our data is best described by an ancestor that migrated out of Africa at time t_2 and after some time of expansion (t_1) split into two populations (model 2) or whether a situation without a shared phase of expansion is favored (model 1). Therefore, model 1 is a simple case of model 2, which assumes that $t_1 = 0$. The most likely scenario was inferred by a likelihood ratio test.

3.2 RESULTS

We analyzed nucleotide variability of ten X-chromosal loci (*i.e.* fragments 78, 206, 237, 259, 359, 392, 422, 431, 721 and 727) previously reported by GLINKA *et al.* (2003) and OMETTO *et al.* (2005) in six population samples from SE Asia. In addition, we added data from an African and a European sample (OMETTO *et al.*

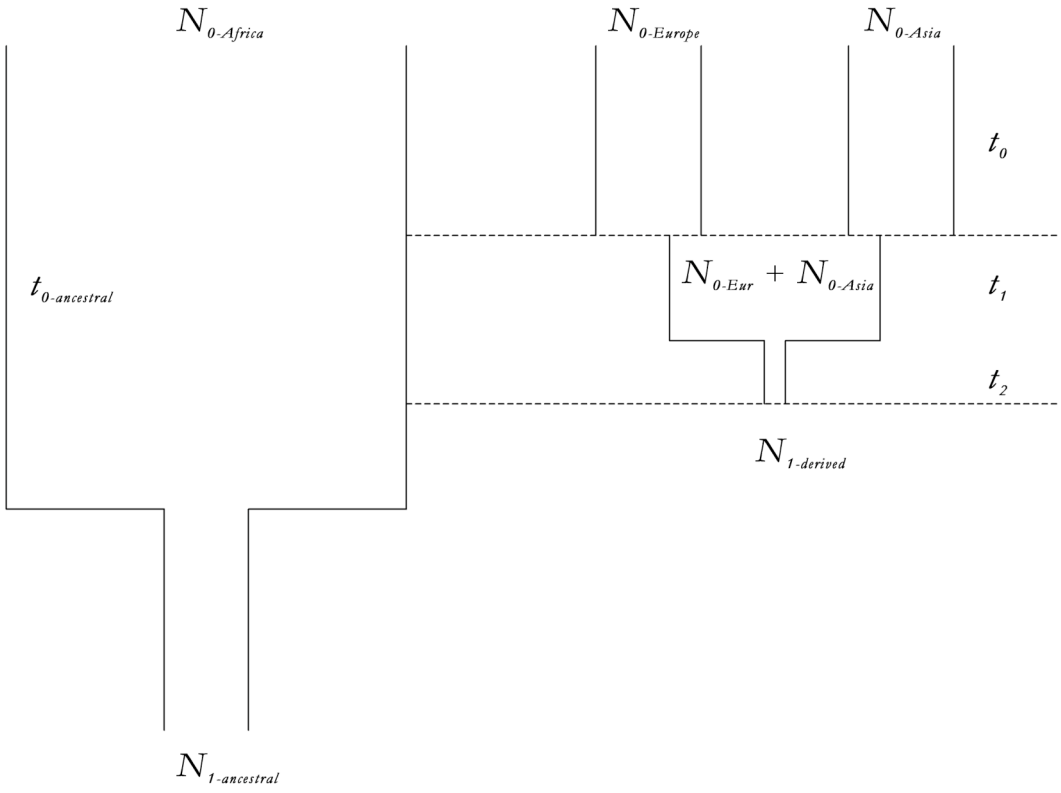


FIGURE 9. Demographic modeling. Model 2 is shown. N_1 indicates effective population size in the past (Africa) or at time t_2 (ancestor of Europe and Asia). N_0 corresponds to the effective population size in the present (Africa, Europe or Asia) or at time t_1 in the past (shared expansion phase of Europe and Asia).

2005) to our analyses. Local recombination rates are moderate to high, ranging from 2.49 to 4.81×10^{-8} rec/bp/generation. Mean distance between fragments is 154 kb. Fragments used in this study are not located in any of the candidate regions for non-neutral evolution identified by OMETTO *et al.* (2005). We therefore assume that they are evolving neutrally. An exception is locus 392, which is located in candidate region "I" (OMETTO *et al.* 2005). However, this fragment is located at the edge of that region and does not show exceptionally reduced variation in a European sample (see OMETTO *et al.* 2005, online suppl.). Furthermore, because this candidate region is rather large (>750kb), we assume that a fragment located far from the center is not directly affected by non-neutral evolution (L. Ometto, pers. comm.).

3.2.1 Nucleotide varitation

Combining all 10 loci, we sequenced ~4,500 bp per isofemale line, with individual fragment length ranging from 295 to 614 bp. We found 278 SNPs in total, of which 110 were singletons, *i.e.* SNPs that occurred only once in the sample. The average number of segregating sites (*S*) ranges from 2.3, observed in Manila, to 22.7 in Africa. Mean *S* across all SE Asian populations is 3.9, which is much lower than the European value (11.7). The average number of haplotypes varies from 1.8 to 3.8 among the SE Asian samples, with haplotype diversities ranging from 0.376 to 0.633. For the European and the African sample, the mean number of observed haplotypes (*h*) is 4.9 and 9.7, respectively, with higher average haplotype diversity (*Hd*) in Africa (0.892 *vs.* 0.727 in the Europe sample). Estimates of *h* and *Hd* in Africa are significantly different from neutral expectations ($P < 0.05$). As evident from Figure 10a, mean heterozygosity (θ) is highest in Africa (0.0172, SE 0.0025), followed by Europe (0.009, SE 0.0022) and SE Asia (0.0028, SE 0.0005). Among the SE Asian samples nucleotide variability is highest in KL, whereas MNL shows a rather low average value (Table 7). Nucleotide diversity measured by π follows the same pattern, but the lowest value is observed in BKK. When averages of π are compared among groups (*i.e.* SE Asia *vs.* Africa or Europe), SE Asia has a significantly lower nucleotide diversity compared to Africa (Wilcoxon rank sums test, $P = 0.002$) and Europe ($P = 0.03$). Furthermore, π is slightly higher than the corresponding θ estimate in most cases, indicating a paucity of low frequency variants. We estimated multi-locus linkage disequilibrium (LD) for all population samples. As evident from Table 7, average LD is significantly low compared to standard neutral expectations in the African sample, consistent with previous results (HADDRILL *et al.* 2005, OMETTO *et al.* 2005). In contrast, the derived populations show medium to high levels of LD. For KK and MNL estimated values are significantly different from neutral equilibrium expectations ($P < 0.05$ for both samples). Parameter estimates for individual loci are provided in the Appendix (Table A3).

TABLE 7. DNA variation and summary statistics (averaged over 10 loci)

Pop	*n*	*L*	*S*	*h*	*Hd*	θ	π	Tajima's *D*	Z_{nS} (inf.)	Z_{nS} (all)	*K*	Fu & Li's *D*	Fay & Wu's *H*
KL	12.6	463	7.8	3.8	0.648	0.0057	0.0059	0.093	0.532	0.383	0.08	0.228	-0.950
KK	11.9	464	2.7	2.6	0.507	0.0024	0.0030	0.855*	0.669	0.669*	0.08	1.026*	-0.885*
BKK	15.1	463	2.3	2.5	0.483	0.0016	0.0021	0.767*	0.546	0.475	0.08	0.594	-0.464
CNX	12.1	464	3.5	2.7	0.504	0.0024	0.0027	0.592	0.588	0.545	0.08	0.473	-0.036
MNL	7.0	463	2.4	1.8	0.376	0.0015	0.0025	1.198*	0.934	0.934*	0.08	1.107*	-0.497
CEB	11.9	461	4.7	3.4	0.544	0.0033	0.0039	0.124	0.625	0.413	0.08	0.421	-0.776
EUR	11.7	463	11.7	4.9	0.727	0.0090	0.0099	0.336	0.482	0.384	0.08	0.499	-0.459
ZBM	11.9	456	22.7	9.7*	0.892*	0.0172	0.0147	-0.793*	0.218	0.171*	0.08	-0.644	-1.955

Pop., population; *n*, sample size; *L*, length in bp; *S*, number of segregating sites; *h*, number of haplotypes; *Hd*, haplotype diversity; *ZnS* (inf.), LD using informative sites only; *ZnS* (all), LD using all polymorphic sites; *K*, divergence. * indicates significance ($P < 0.05$). For individual loci see Table A3 (Appendix).

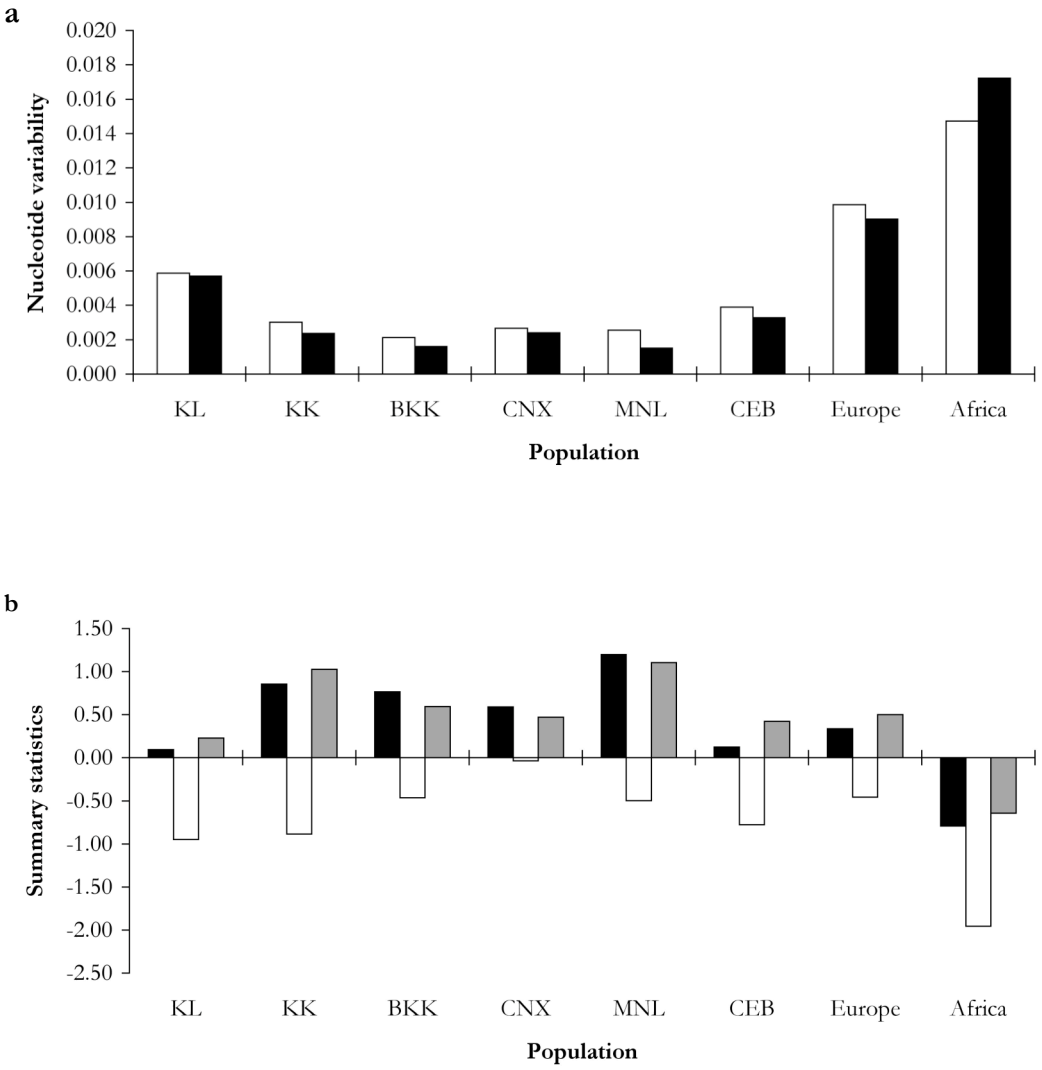


FIGURE 10. Polymorphism and summary statistics. Panel **a** shows average θ (black bars) and π (white bars). Panel **b** illustrates Tajima's *D* (black), Fu and Li's *D* (grey) and Fay and Wu's *H* (white).

Divergence between *D. melanogaster* and its congener *D. simulans* is on average normal (~0.08, SE 0.0002) and homogenous among geographic regions (Table 7). However, the observed value is slightly higher than the estimate provided by OMETTO *et al.* (2005), which might be due to differences in sample size, *i.e.* 10 loci *vs.* >250 fragments analyzed in the aforementioned study.

3.2.2 Neutrality tests

We applied standard neutrality tests to determine whether our data are compatible with the standard neutral model. For testing intraspecific data we used Tajima's *D* (TAJIMA 1989), Fu and Li's *D* (FU and LI 1993) and Fay and Wu's *H* (FAY and WU 2000). As illustrated in Figure 10b, all SE Asian samples show positive values for both Tajima's *D* and Fu and Li's *D* with the exception of KL, which shows a slightly negative Tajima's *D*. For three of our six SE Asian samples (KK, BKK and MNL), Tajima's *D* is significantly different from standard equilibrium expectations. For KK and MNL, Fu and Li's *D* is significantly positive as well ($P < 0.05$). The trend towards positive estimates is paralleled by the European sample. However, values were not statistically significant. In contrast, the African sample generally shows negative values for these statistics, with Tajima's *D* being significantly less than zero ($P = 0.04$). We also analyzed the frequency spectrum of variants at segregating sites. In all samples we observe negative Fay and Wu's *H*, indicating a general skew towards high frequency derived variants. However, only the value obtained from the KK sample was significant in our coalescent simulations ($P = 0.02$). A multi-locus version of the HKA statistic (HUDSON *et al.* 1987) revealed a significant deviation from the standard neutral model for KL ($P = 0.006$) and CNX ($P = 0.02$).

3.2.3 Population structure

We tested for population differentiation following three approaches. First, we grouped our population samples according to geographical location, *i.e.* Africa, Europe and SE Asia. Applying an analysis of molecular variance framework (EXCOFFIER *et al.* 1992) to the SE Asian sample, we detect significant population structure ($F_{ST} = 0.21$, $P < 0.001$). In addition, populations are significantly differentiated when all three groups are analyzed jointly ($F_{ST} = 0.39$, $P < 0.001$). This pattern is supported by a pairwise F_{ST} analysis. All populations are significantly differentiated from one another (s. Table 8), with Europe and Africa showing the lowest degree of differentiation among all comparisons ($F_{ST} = 0.098$). However, even this value is significant according to 10,000 permutations. We observed highest F_{ST}

TABLE 8. Population differentiation (pairwise F_{ST})

	BKK	CEB	CNX	KK	KL	MNL	EUR	ZBM
BKK		+	+	+	+	+	+	+
CEB	0.30		+	+	+	+	+	+
CNX	0.16	0.13		+	+	+	+	+
KK	0.19	0.20	0.13		+	+	+	+
KL	0.19	0.15	0.15	0.15		+	+	+
MNL	0.41	0.27	0.31	0.33	0.23		+	+
EUR	0.38	0.27	0.31	0.34	0.14	0.35		+
ZBM	0.38	0.31	0.32	0.35	0.21	0.32	0.10	

+ denotes significance ($P < 0.05$).

between MNL and BKK (0.41). Among the SE Asian samples MNL generally is most differentiated from the others, with F_{ST} ranging from 0.23 (MNL *vs.* KL) to 0.41 (MNL *vs.* BKK). In contrast, generally low differentiation was found for KL. A distance based NJ population tree illustrates this pattern (see Figure 11).

Second, we applied a recently developed Bayesian clustering approach (PRITCHARD *et al.* 2000). This method estimates the probability of ancestry for each individual from one of the eight samples. Average probability for each sample is outlined in Figure 12. Highest posterior probabilities for the number of distinct clusters (C) were obtained for $C = 7$. As depicted in Figure 12, the KL sample shows the greatest heterogeneity of all SE Asian samples, suggesting a more diverse genetic ancestry compared to flies from other locations. By contrast, MNL forms a distinct and nearly homogenous cluster. The remaining SE Asian samples (*i.e.* BKK, KK, CEB and CNX) share common ancestries in varying degrees. Interestingly, KL shares about ~10% of its ancestry with the European sample, which consists of one major cluster only. Unlike KL, the eastern populations are more differentiated from Europe, which is in accordance with our F_{ST} approach (Table 8). The African sample, however, is most separated from all others, suggesting an ancestry that is different from both the European and the SE Asian samples.

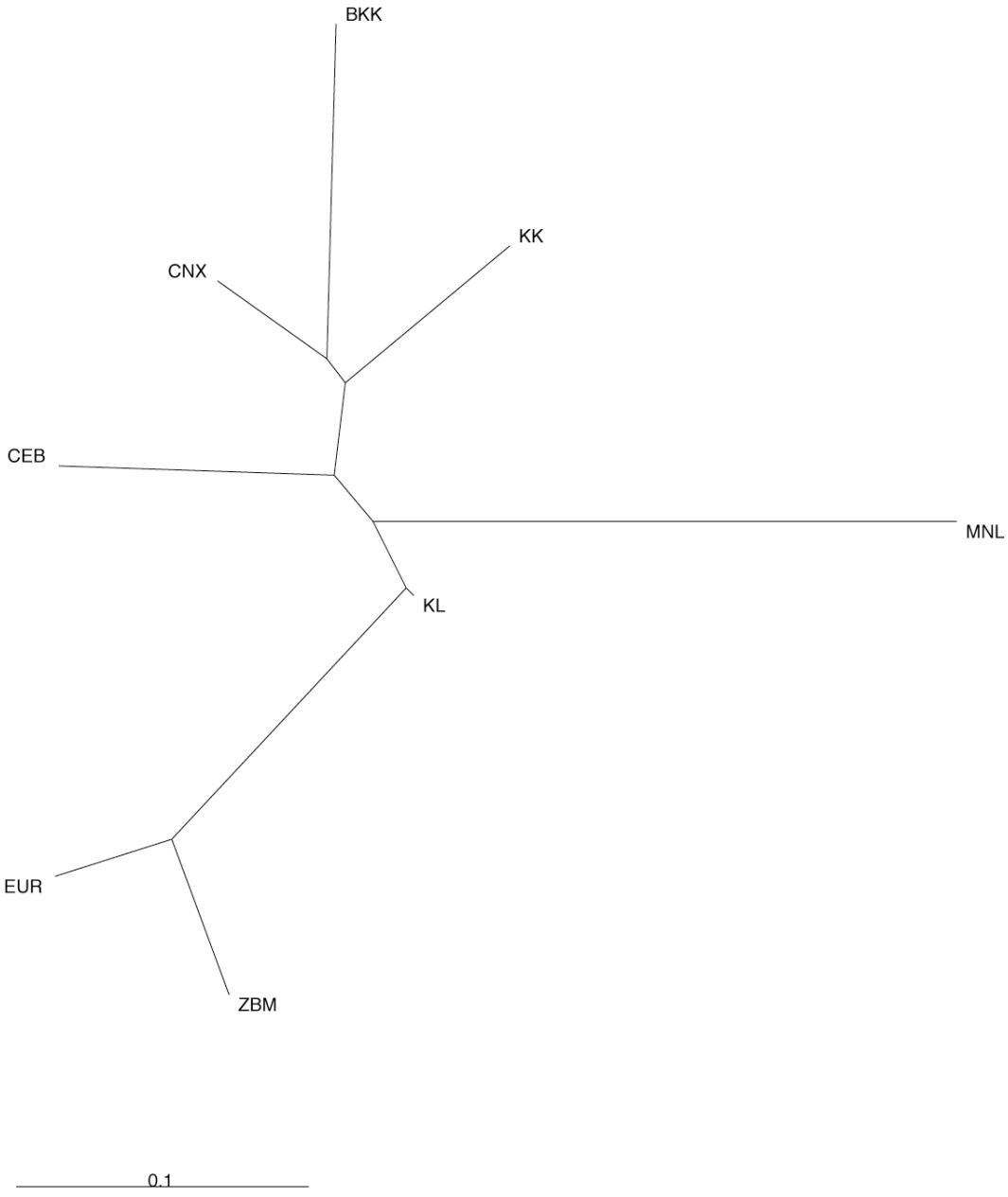


FIGURE 11. Unrooted neighbor-joining (NJ) population tree based on pairwise F_{ST}. Observed genetic relationships agree with those detected using the approach of PRITCHARD *et al.* (2000).

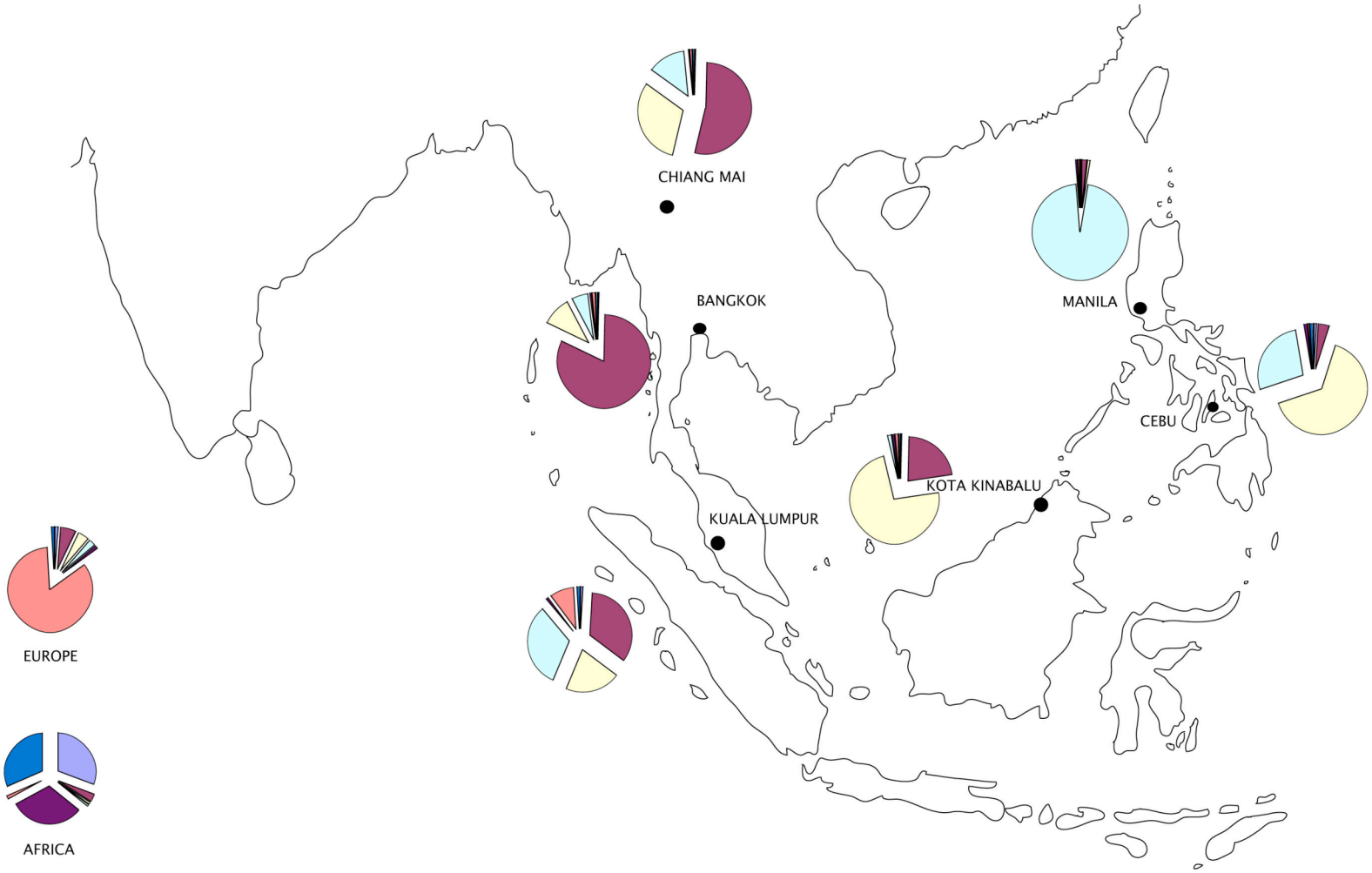


FIGURE 12. Sampling locations and population structure. Pie charts show results of STRUCTURE (PRITCHARD *et al.* 2000).

Third, we evaluated the level of population substructure following the approach of Vogl *et al.* (2003). Average differentiation (θ_p) between all populations is 0.36, which is in agreement with our F_{ST} result. As can be inferred from Table 9, ZBM is least differentiated from the migrant pool (θ_p = 0.007). The European population shows the second lowest value (θ_p = 0.15). In contrast, we obtained a large θ_p for MNL (0.62), indicating strong differentation. The remaining SE Asian populations show θ_p-values ranging from 0.25 (KL) to 0.53 (BKK). In addition, as estimated from the migrant pool, all loci but one have negative Tajima's *D*, with a mean value of -0.98 (see Table 10). This value is close to the one observed for ZBM (-0.79, see Table 7), suggesting a long-term expansion of *D. melanogaster*.

TABLE 9. Differentiation from the migrant pool (according to VOGL *et al.* 2003)

Pop [a]	average	2.5 % [b]	5 %	50 %	95 %	97.5 %
BKK	0.53	0.39	0.41	0.53	0.66	0.68
CEB	0.37	0.26	0.28	0.37	0.48	0.50
CNX	0.44	0.29	0.32	0.43	0.56	0.59
KK	0.49	0.35	0.38	0.49	0.66	0.64
KL	0.25	0.15	0.16	0.24	0.34	0.36
MNL	0.62	0.44	0.47	0.62	0.76	0.78
EUR	0.15	0.08	0.09	0.14	0.21	0.22
ZBM	0.007	0.0009	0.002	0.006	0.017	0.019

[a] Population sample

[b] Confidence limit

FIGURE 12. *(continued from previous page)*

Results for EUR and ZBM are presented in the lower left corner. The different colors correpsond to different genetic clusters.

TABLE 10. Tajima's *D* estimated from the migrant pool (according to VOGL *et al.* 2003)

Locus	average *D*	2 %	5 %	50 %	95 %	97.5 %
78	-0.71	-0.79	-0.78	-0.71	-0.59	-0.57
206	-1.56	-1.68	-1.67	-1.56	-1.49	-1.48
237	0.14	0.07	0.08	0.13	0.19	0.21
259	-0.36	-0.78	-0.70	-0.36	0.03	0.09
359	-1.68	-1.78	-1.76	-1.68	-1.59	-1.55
392	-2.03	-2.07	-2.06	-2.03	-1.99	-1.99
422	-0.73	-0.94	-0.92	-0.75	-0.50	-0.48
431	-1.24	-1.34	-1.33	-1.24	-1.16	-1.15
721	-1.15	-1.40	-1.38	-1.19	-0.73	-0.67
727	-0.52	-0.66	-0.64	-0.51	-0.41	-0.40

Measures of population differentiation (F_{ST} in particular) are influenced by the degree of variation within subpopulations (CHARLESWORTH 1998). This is especially important when demes were established through bottlenecks, which is likely to be the case for non-African *D. melanogaster*. Visual inspection of our data revealed that for some population samples, *e.g.* BKK, high differentiation is due to low within-population variability (π) rather than to high levels of divergence between populations. We therefore investigated the relationship of our population samples by comparing the number of shared and private polymorphisms across populations (see Baudry *et al.* 2006). As shown in Table 11, most segregating sites are shared between ZBM and MEL (88), consistent with our previous observations. EUR and KL share 66 SNPs, followed by ZBM and KL (61). This observation suggests a close relationship of these populations. The remaining SE Asian populations show rather similar levels of shared polymorphism (15 to 35). However, all demes consistently share most SNPs in pairwise comparisons with KL (on average 39), implying a central position of this population. A close association between EUR and KL is also supported by the low number of sites unique to KL when compared to EUR (Table

TABLE 11. Number of shared polymorphic sites among samples

	BKK	CEB	CNX	KK	KL	MNL	EUR	ZBM
BKK								
CEB	15							
CNX	15	32						
KK	17	19	18					
KL	19	34	31	25				
MNL	9	20	17	11	35			
EUR	13	32	21	18	66	16		
ZBM	15	30	25	19	61	19	88	

TABLE 12. Number of segregating sites private to the samples in the rows, when compared to the samples in the column

	BKK	CEB	CNX	KK	KL	MNL	EUR	ZBM
BKK		8	7	6	4	14	10	8
CEB	30		13	26	11	25	13	15
CNX	19	3		17	4	19	10	10
KK	10	9	8		2	16	9	8
KL	60	44	47	53		64	12	17
MNL	15	4	7	13	10		10	5
EUR	104	85	77	99	51	92		29
ZBM	217	202	207	213	171	212	144	

12). In contrast, KL shows more private polymorphic sites when compared to the other SE Asian samples (ranging from 44 to 64), implying that this population harbors most variation. An interesting feature of our data is the observation of a distance effect on the number of private sites. *I.e.* when compared to KL, sampling sites that are located far from KL show more unique SNPs than those sites located close to KL. In addition, some of these private sites are not present in ZBM either, implying that they represent newly derived mutations.

3.2.4 Frequency spectra

To analyze the frequency distribution of segregating sites among the closely related populations ZBM, EUR, and KL in detail, we sequenced additional 17 X-chromosomal loci in the KL sample (fragments 57, 60, 84, 93, 106, 166, 178, 205, 231, 241, 272, 276, 277, 286, 326, 439 and 465) and added them to our data. Corresponding sequences of the African and the European sample were taken from OMETTO *et al.* (2005) and reanalyzed. We first compared average π and θ for the enlarged dataset (27 loci) obtained from the KL sample to our previous dataset (10 loci) to assure consistency. Mean values were not significantly different. This holds for both estimators, π (P = 0.25, Wilcoxon rank sums test) and θ (P = 0.33), respectively. Tajima's D was not affected by dataset enlargement either (P = 0.74). Figure 13 illustrates the frequency spectrum of the African, European and Kuala Lumpur population samples. Consistent with previous results we observed an excess of singletons in the African sample (Figure 13a). In contrast, the samples from Europe and KL show fewer singletons than expected and a surplus of intermediate to high frequency variants (Figures 13b and c).

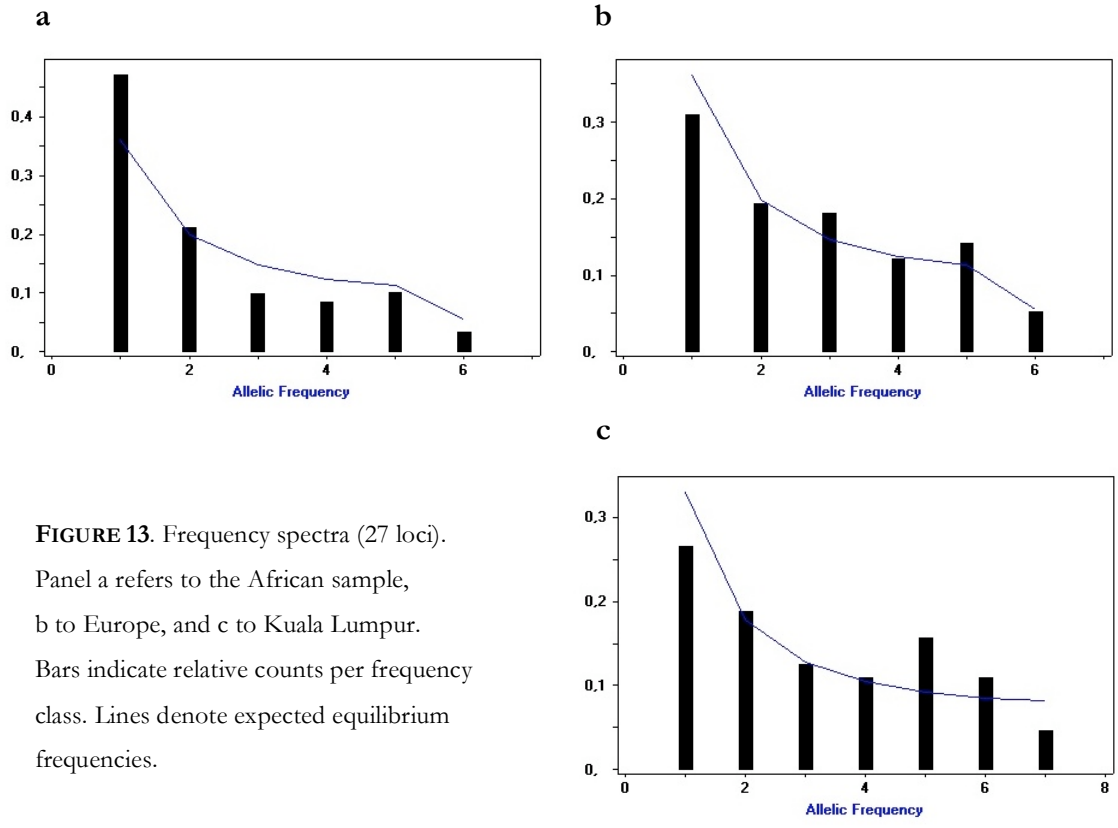


FIGURE 13. Frequency spectra (27 loci). Panel a refers to the African sample, b to Europe, and c to Kuala Lumpur. Bars indicate relative counts per frequency class. Lines denote expected equilibrium frequencies.

3.2.5 Demographic analysis

We investigated the demographic history of our samples from Zimbabwe, Europe, and Kuala Lumpur applying the method of LI and STEPHAN (2005, 2006) to our extended dataset (n = 27 loci). Since it has been argued that non-African *D. melanogaster* have a unique origin (BAUDRY *et al.* 2004), we determined the time since the split from the common ancestor and investigated whether the two derived samples shared a joint phase of population size expansion prior to separation. Results of this analysis are detailed in Table 13. Model 2 fits our data significantly better than model 1 ($P < 0.05$). *I.e.* it is most likely that Europe and Kuala Lumpur had an ancestor that originated from Africa ~20,000 years ago (assuming $N_{Africa0} = 8.6x10^6$, and 10 generations per year; LI and STEPHAN 2006). After a short period of population size expansion (~7,000 years) outside of Africa, this derived population split into two separate populations, namely Europe and Asia ~12,000 years ago.

TABLE 13. Inference of the demographic history of samples from Africa, Europe and Kuala Lumpur

Parameter [a]	Model 1 [b]	95% CI [c]	Model 2 [d]	95% CI
t_0	12,000	10,300 - 17,100	12,000	5,100 - 17,100
t_1	-	-	6,800	170 - 20,500
t_2	3,400	170 - 13,700	1,700	10 - 13,700
$N_{0\text{-}Europe}$	2.86×10^6	2.2×10^6 - 5.7×10^6	2.86×10^6	1.2×10^6 - 5.7×10^6
$N_{0\text{-}Asia}$	191,000	143,000 - 246,000	172,000	143,000 - 215,000
$N_{1\text{-}derived}$	30,600	7,600 - 306,000	15,200	7,600 - 304,000

[a] estimated parameters, as illustrated in Figure 9.

[b] estimates of t in years and N (effective size) for model 1.

[c] 95% confidence interval for a given parameter estimate.

[d] estimates of t and N under model 2 assumptions.

3.2.6 Heterozygosity across the *wapl* region

We chose to analyze the *wapl* region in the KL sample, as this SE Asian population seems to deviate least from neutral equilibrium expectations (see Discussion). All 13 markers studied in the first chapter were sequenced and analyzed in KL. The pattern of nucleotide variation is depicted in Figure 14. For the African and European samples see chapter 1. As evident from Figure 14, heterozygosity in KL is intermediate between Africa and Europe. Yet, compared to the European sample, which shows six monomorphic loci (*i.e.* fragments 4 – 7, 9 and the *wapl* locus), the KL sample features only three fragments without variation. Mean θ in the *wapl* region is significantly different from the average heterozygosity across the 27 neutral loci in KL ($P = 0.008$). In spite of this, a multi-locus HKA test (HUDSON *et al.* 1987) yielded a low but nonsignificant result ($P = 0.12$) for the markers in the *wapl* region. Tajima's *D* (TAJIMA 1989) was significantly negative for only one fragment (*i.e.* locus 10, $D = -1.7$, $P < 0.05$). When compared among the three samples, mean heterozygosity (θ) across all 13 loci is significantly different between Africa and Europe ($P = 0.0005$, Wilcoxon rank sums test), as well as between Africa and KL ($P = 0.0004$). However, average θ estimates of Europe and KL are not significantly different ($P = 0.75$). Divergence between *D. melanogaster* and *D. simulans* is approx. the same for all samples ($P = 0.97$). In addition, we inspected polymorphic fragments 4 to 9, comparing sequence data of all three samples. As shown in Figure A3 (Appendix), haplotype structure is most complex in Africa, as opposed to KL and Europe, where two distinct haplotypes can be observed in most cases. In spite of this, there are two additional aspects in the data that distinguish the two derived populations: First, at some nucleotide positions, individuals from KL and Africa share the same variants (*e.g.* at loci 4 and 8), suggesting that KL has retained ancestral polymorphism. Interestingly, none of these variants can be found in the European sample. Second, several individuals from the KL sample show derived variants, none of these being present in either the European or the African sample. All derived variants are singletons, except for a doubleton in fragment 8 (Figure A3, Appendix). Furthermore, three KL individuals (*i.e.* KL7, KL17 and KL24) feature a 6-bp insertion that is not present in any other sequence analyzed. This may indicate that the selective sweep in KL is older than the one observed in Europe.

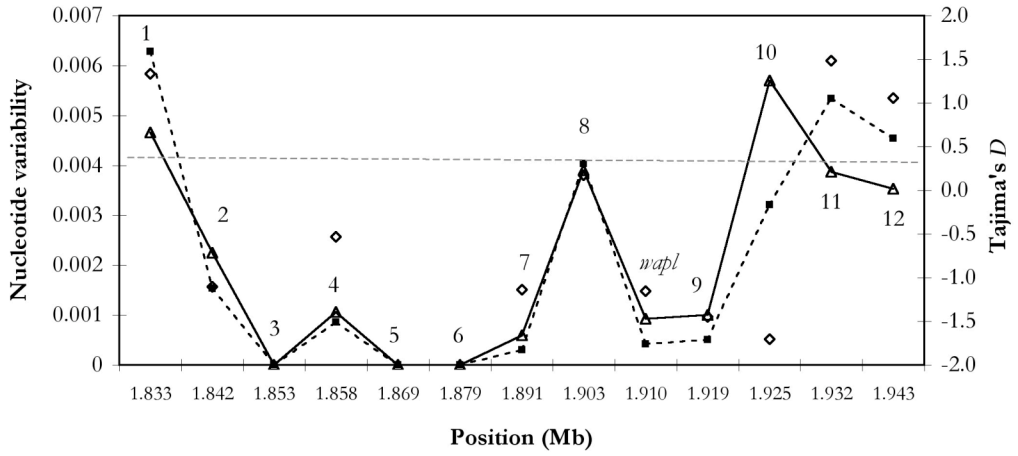


FIGURE 14. *wapl* sweep region in the Kuala Lumpur sample. Solid line and triangles show θ, dashed line and squares denote θ. Diamonds indicate Tajima's *D*. The grey dashed line represents average heterozygosity as estimated from the 27 loci.

3.3 DISCUSSION

3.3.1 Nucleotidy diversity

In this study we analyzed nucleotide variation in six *D. melanogaster* samples from SE Asia and compared parameter estimates to those obtained from two additional populations, *i.e.* a European and an African sample. Sub-saharan populations are generally assumed to be ancestral to non-African populations, which are considered derived (DAVID and CAPY 1988). We analyzed 10 X-chromosomal fragments, putatively evolving neutrally (OMETTO *et al.* 2005), and found low levels of nucleotide diversity in all six SE Asian samples. In comparison, mean heterozygosity in SE Asia is only about 30% of the European and 16% of the African average. In addition, standard summary statistics reveal a deviation from the standard neutral model for three SE Asian samples (*i.e.* BKK, KK and MNL) and the African population. The latter has recently been shown to be expanding (OMETTO *et al.* 2005, LI and STEPHAN 2006), probably after having suffered from population size bottlenecks in the far past (DIERINGER *et al.* 2005, HADDRILL *et al.* 2005). Negative averages of Fu and Li's *D* (FU and LI 1993), Tajima's *D* (TAJIMA 1989), and significantly low levels of linkage disequilibrium support the hyptothesis of a size

expansion. Applying a recently proposed coalescence approach (VOGL *et al.* 2003) we find that *D. melanogaster* is an expanding species on the long-term. However, samples from Europe and SE Asia show a general trend towards positive values of these statistics, a pattern that has been shown to be the signature of a recent population size bottleneck (*e.g.* TAJIMA 1989, BEGUN and AQUADRO 1993, AQUADRO 1997, GLINKA *et al.* 2003). High levels of LD, as observed in our SE Asian data, strengthen this interpretation (BEGUN and AQUADRO 1995). Interestingly, all populations analyzed in this study show a negative, yet nonsignificant, average Fay and Wu's *H* (FAY and WU 2000), indicating a skew towards high frequency derived variants. This trend is strongest in the African population (H = -1.95, ns) and may point toward a high rate of selective sweeps in the ancestral population (MOUSSET *et al.* 2003). Recent work by LI and STEPHAN (2006) supports this hypothesis. Note that locus 392 was reported to show a significantly negative *H* by OMETTO *et al.* (2005). If removed from the test statistic, however, average *H* is still negative (H = -1.82). Despite this, a mean negative *H* can also be caused by population size bottlenecks of recent to intermediate age (DEPAULIS *et al.* 2003, HADDRILL *et al.* 2005). A negative average *H*, as observed in our SE Asian data (H = -0.63 across all six samples), could therefore be explained by demographic history, *i.e.* population size bottlenecks in the recent past. This is in accordance with other features of our data, *e.g.* low levels of heterozygosity, elevated LD and positive *D* statistics. One of our SE Asian samples, namely Kuala Lumpur, shows a pattern that is slightly different from the others: estimates of heterozygosity (θ) and nucleotide diversity (π) are highest of all SE Asian samples. Furthermore, LD is comparable to the European average, and summary statistics, such as Tajima's *D* and Fu and Li's *D*, deviate least from neutral equilibrium expectations. A multi-locus HKA test (HUDSON *et al.* 1987), however, revealed a significant result for KL. Visual inspection of our data indicates that two loci, *i.e.* 237 and 422, are responsible for the deviation from neutral equilibrium expectations. When these loci are removed, the HKA test is no longer significant (P = 0.1). The observed deviation is caused by an excess of polymorphism at these loci. Sequence comparison of KL, Europe and Africa reveals that most SNPs existing in KL (at 237 and 422) are also present in Africa. At locus 237, however, KL features three SNPs that are not present in ZBM. Yet, all polymorphisms are shared with

EUR. This suggests that the excess of polymorphism is not due to a recent African immigrant, but represents variation that stems from the common ancestor of EUR and KL. Also, since most SNPs are in fairly low frequency in EUR, a recent immigrant from Europe seems to be unlikely. In addition, in KL we observed a unique segregating site at locus 422, which is neither present in EUR nor in ZBM. We therefore argue that the deviation from neutral expectations, as revealed by the multi-locus HKA test (HUDSON *et al.* 1987), is due to generally high heterozygosity at loci 237 and 422, respectively. Subsequent bottlenecks leading to more peripheral populations (*e.g.* MNL), in contrast, had a more severe impact affecting fragments 237 and 422 as well. In addition, we performed locus specific HKA-tests for KL. We obtained non-significant results for the loci mentioned above (237: $X^2 = 2.35$, $P = 0.12$; 422: $X^2 = 4.27$, $P = 0.07$). Analysis of a larger number of fragments in KL might help to eliminate variance effects on polymorphism caused by demography.

3.3.2 Population differentiation

When grouping samples according to their geographical region, *i.e.* Africa, SE Asia and Europe, we find that populations are significantly differentiated from each other. We analyzed the SE Asian group in detail and detected significant substructure among all six samples. Interestingly, KL is on average least differentiated from the remaining SE Asian samples, suggesting a central role of this population. This is supported by a recent microsatellite study, which revealed that KL is the least differentiated population compared to all other SE Asian samples they studied (SCHLÖTTERER *et al.* 2006). In addition, in our data all SE Asian samples consistently share most of their segregating sites with KL. MNL, located in the Northeast, shows on average high levels of differentiation in pairwise comparisons. This pattern is further supported by a Bayesian clustering analysis (PRITCHARD *et al.* 2000), which separated MNL from the other SE Asian populations. Furthermore, KL shares an admixed ancestry, suggesting that this population is ancestral to the others. Applying a recently proposed coalescence approach (VOGL *et al.* 2003) we find that, among all SE Asian samples, KL appears to be least differentiated from the migrant pool. Subsequent colonization of more distant regions (*e.g.* MNL or CNX) was most likely accompanied by severe reductions in population size and could have led to the

observed patterns of nucleotide variation in these peripheral populations. Interestingly, when compared to KL and EUR, individual SE Asian samples show unique polymorphic sites, *i.e.* SNPs that are not present in the other samples. This may indicate that population structure was not established very recently.

3.3.3 When was SE Asia colonized?

Within the last 250,000 years SE Asia formed a large single landmass ('Sunda shelf') for a long time period, with sea levels below 120 m (VORIS 2000). Yet, sea levels were at their maximum lows for only relatively short phases, but were at or below intermediate levels for more than half of the time within the last 150,000 years. At a sea level of 40 m below present-day levels, large parts of SE Asia, such as the Malay Peninsula, Sumatra and Borneo were still connected by land bridges (VORIS 2000), which enabled human migration between different geographical regions and general floral and faunal exchange (KLEIN 1999, M. Krings, pers. comm.). The most recent glacial maximum began to recede about 17,000 years before present and current sea levels were reached *ca.* 6,000 years ago (FAIRBANKS 1989). The last phase where sea levels were 40 m below present-day levels lasted from about 17,000 to 9,000 years before present (INGER and VORIS 2001). During that time period humans were therefore able to migrate across different regions. However, later occuring water barriers must still have been conquerable, as humans managed to cross water barriers, such as the gap between Indonesian islands and Australia (~80 km), at much earlier times (>40,000 ya; KLEIN 1999, G. Grupe pers. comm.). As reported by DAS *et al.* (2004), a sister species of *D. melanogaster*, namely *D. ananassae*, originated in SE Asia and central populations retained genetic homogeneity because of unrestricted migration before the formation of present geographic structures. The genetic differentiation pattern observed in our data suggest a different situation for *D. melanogaster*. Since we found low levels of polymorphism and significant population substructure among all SE Asian samples, a long-term established population, as reported for *D. ananassae*, seems rather unlikely. Instead, we propose that extant *D. melanogaster* populations were established from peninsular Malaysia along with the increase of sea levels (or shortly before). Out of all SE Asian samples analyzed, KL consistently shares most polymophism with EUR and only

features few private sites, which supports a common ancestry. This hypothesis is in accordance with previous results of BAUDRY *et al.* (2004). In addition, our demographic analysis suggests that European and SE Asian *D. melanogaster* shared a common ancestor until ~12,000 years ago. Prior to separation, the ancestral population experienced a short period (~7,000 years) of population size expansion outside of Africa. Some of the observed polymorphism shared between SE Asia and Europe, but not present in Africa, might date back to this episode of common ancestry.

The foundation of an SE Asian *D. melanogaster* population about 10,000 – 20,000 years ago is supported by recent anthropological data (note that *D. melanogaster* is a human commensal). Evidence for Early Holocene changes in land use and likely food production is reported for SE Asia as early as 8,000 – 7,000 years ago (KEALHOFER 2002). An agro-ecosystem transformation, *i.e.* the establishment of irreversible landscape transformations in the course of intensive land use, can be dated to 9,000 – 7,000 years before now (KEALHOFER 2002). Considering the onset of agriculture, however, there is considerable variation between regions, which was also dependent on past climate and vegetation changes. *E.g.* in the Late Pleistocene and Early Holocene, woodland savanna of peninsular Thailand turned into Holocene forests. In the Northeast of Thailand, open grassland turned into dense seasonal forest (KEALHOFER 2002). This probably has led to fragmentations of faunal populations dependent on open habitats. Our results, which suggest a foundation of the KL population ~12,000 years ago are in accord with these data and suggest subsequent fragmentation of *D. melanogaster* into local populations, which was accompanied by genetic drift. Limited gene flow between closely located populations, caused by physical barriers such as large bodies of water or high mountain ranges might have enforced differentiation. The population in Chiang Mai, situated in the Northwest of Thailand, is separated from its nearest neighbor population in Bangkok by ~580 km (air path) and several mountain ranges with altitudes >1,000 m. Kota Kinabalu on the other hand, is separated from KL by the South China Sea (~1,600 km). Since these populations are significantly differentiated from each other, geographical barriers, as described above, might well explain our observations.

3.3.4 Historical context of the *wapl* region

In chapter 1 we reported evidence for a selective sweep in the *wapl* region of *D. melanogaster*. This hypothesis was supported by a closer investigation of the core area around the putative target of selection (*i.e. ph-p*, see chapter 2). In this study we scrutinized the *wapl* region in an additional sample from SE Asia and found a pattern of nucleotide variation that is different from the one observed in the European sample. We observed a higher level of heterozygosity in KL than in EUR and, in addition, fewer monomorphic loci. This may indicate that the sweep region in KL is older than the one observed in Europe. Additional features of the KL sample support this view. First, we observed new variation at six sites in KL. Two of these SNPs are doubletons. Furthermore, KL has retained ancestral variation as shared polymorphic sites between KL and ZBM indicate. In addition, haplotypes in KL are recombinants since some variants are located on different alleles in the African sample (*i.e.* at fragments 4 and 8). In contrast, only a single allele went to fixation in the European population.

Considering both derived populations, this pattern is rather unlikely under the assumption of neutral drift. First, the Kuala Lumpur sample shows on average much lower levels of nucleotide diversity across 27 putatively neutrally evolving loci than Europe. We would therefore expect that nucleotide variation in the *wapl* region is approx. the same in both derived populations or even lower in KL. However, the opposite pattern was observed. Second, we detected high LD in loci flanking the *wapl* fragments in the European sample. In KL no such pattern was found. We therefore propose that the *wapl* sweep region in the KL sample is older than in the European one. Nucleotide variation is best explained by an additional selective sweep that affected the European population exclusively. KL, on the other hand, displays the signature of an old sweep where nucleotide variation is slowly being restored. Additional sequencing of the entire *ph-d* – *Pgd* region (as in chapter 2) in the KL sample might help to further uncover the processes that shaped nucleotide variability in this region of the genome.

3.4 SUMMARY

We obtained population genetic parameter estimates from six SE Asian samples and compared them to those obtained from a European and an African sample. We observed substantially lower levels of nucleotide diversity in SE Asia than in either Africa or Europe. In particular, samples taken from peripheral populations (*e.g.* Manila and Cebu, located on the Philippines) show a paucity of haplotypes. Common summary statistics indicate that genetic drift had a substantial impact on these populations, which also led to considerable population substructure. One sample, *i.e.* Kuala Lumpur, shows rather high levels of heterozygosity among all SE Asian samples and is on average least differentiated from these. This indicates that the Kuala Lumpur population is ancestral to the other SE Asian populations, which is supported by a high number of shared polymorphic sites. Finally, we find the selective sweep observed in the *wapl* region to be older than in Europe.

Conclusions

D. melanogaster likely originated in sub-Saharan Africa (DAVID and CAPY 1988, LACHAISE and SILVAIN 2004) and only recently expanded its range (LACHAISE *et al.* 1988). Therefore, it is assumed that Drosophila adapted to new habitats, as environmental conditions in (*e.g.*) Europe or Asia are different from those in the ancestral range of this species. It is therefore worthwhile to find the signatures of recent selection in the genome and to determine the mutations that have led to a phenotype. This is particularly interesting, as it has been shown that selection can leave a footprint in the genome, known as a selective sweep. Neutral variants can rise in frequency due to linkage to a positively selected site ("genetic hitchhiking", MAYNARD SMITH and HAIGH 1974). Therefore, the region close to the selected site typically shows a severe reduction in polymorphism (KIM and STEPHAN 2002, PRZEWORSKI 2002). However, the out-of-Africa expansion of *D. melanogaster* was accompanied by a severe reduction in effective population size ("bottleneck", see GLINKA *et al.* 2003, HADDRILL *et al.* 2005, OMETTO *et al.*, 2005). Thus, a local reduction in heterozygosity can be due to either demography or selection. Comparing regional levels of polymorphism between populations might help to distinguish between these two forces. In addition, local levels of nucleotide variability can be tested against the background of various demographic scenarios (see OMETTO *et al.*, 2005, LI and STEPHAN 2006).

This thesis addressed several questions regarding the selective and demographic history of *D. melanogaster*:

1. Is it possible to attribute a local reduction in polymorphism to selection?

In chapter 1 we analyzed a particular genomic region in detail. We found a severe reduction of nucleotide variability in a European sample of *D. melanogaster*. Standard neutrality tests and recently proposed likelihood ratio tests (KIM and

STEPHAN 2002, KIM and NIELSEN 2004, JENSEN *et al.* 2005) rejected the standard neutral model. Our results were also robust against simple bottleneck scenarios. This argues in favor of positive selection having shaped the region under investigation. However, we also observed a significant reduction in heterozygosity in the corresponding region of a putative ancestral population (Zimbabwe). This complicates the localization of the target of selection in the derived population, since both demography and selection probably have shaped the region under consideration. In chapter 2 we therefore concentrated on the African sample, as the valley of reduced variation was much narrower. We almost completely sequenced this valley and found further support for a selective sweep. In chapter 3 we revisited the sweep region in another derived sample (from SE Asia) and also detected the sweep signature. However, the pattern of variation suggests that the sweep is older that in the European sample. This argues in favor of an additional selective sweep that only affected the European population. The joint analysis of several derived populations might therefore faciliate the detection of positive selection.

2. Can we determine the time since the fixation of a positively selected site? (Chapter 2)

It has recently been shown that the signature of a selective sweep can be used to determine its age (PRZEWORSKI 2003). New mutations arising after the fixation of a beneficial mutation leave a characteristic pattern of nucleotide variation (*e.g.* an excess of low frequency variants or a surplus of haplotypes). However, the detection of positive selection is limited to the timeframe of ~$0.1N_e$ generations (*e.g.* KIM and STEPHAN 2002, PRZEWORKSI 2002), since new variation is slowly building up and equilibrium conditions will be reached eventually. This timeframe corresponds to 100,000 generations in Drosophila (assuming $N_e = 10^6$). Assuming 10 or 5 generations per year this corresponds to 10,000 or 20,000 years, respectively. In our African sample we dated the fixation of the beneficial mutation to >10,000 years ago. However, a deviation from equilibrium conditions, *e.g.* a population expansion, can complicate this analysis, since the signature of an expansion may resemble that produced by a selective sweep (*e.g.* an excess of low frequency variants). The

incorporation of non-equilibrium conditions into tests of selection might help to determine the time since the fixation of a positively selected site more precisely.

3. *Is it possible to detect the mutation that was positively selected? (Chapter 2)*

Mutations typically result in changes in either protein structure or gene expression. The former is harder to detect on the molecular level, as only single bp changes might be involved (FFRENCH-CONSTANT *et al.* 1993, WEILL *et al.* 2004) and the consequences of these changes are not easy to evaluate. In chapter 2 we studied the region around the estimated position of the target of selection. We detected several fixed differences between *D. melanogaster* and *D. simulans* / *D. yakuba*, which might result in different transcription factor binding sites. We therefore suggest that the gene under investigation is specializing on different functions in the *D. melanogaster* lineage. A closer examination of the molecular evolution of this gene and its paralogue might help to confirm our assumptions.

Gene expression differences, in contrast, are easier to detect, as modern molecular techniques, such as microarrays or qRT-PCR are available. We found significant differences in gene expression between our population samples. For one gene these differences correlate with the presence of mutations in the promotor region. Our prelimiary analyses therefore suggest, that selection may have acted on a variant that alters gene expression (in the European sample). However, since gene expression is affected by multiple factors, a more detailed analysis might be necessary.

4. *What about the population structure of SE Asian* D. melanogaster*? (Chapter 3)*

The analysis of an additional population sample might help to distinguish between the effects of demography and selection. However, the demographic *status quo* of these samples has to be determined first. We therefore analyzed the genetic structure of six Southeast Asian *D. melanogaster* samples. Population genetic parameter estimates revealed that these samples have a small effective population size, probably due to severe population bottlenecks. Yet, one sample (Kuala Lumpur) seems to

deviate least from neutral equilibrium expectations. Furthermore, we found all SE Asian samples to be highly differentiated from one another and from the European and the African samples as well. Our demographic analysis suggests that the Kuala Lumpur population and the European one had a common ancestor that originated from Africa about 20,000 ya. After a phase of shared expansion the ancestor split into to separate populations ~12,000 ya. The analysis of a larger number of loci in the Kuala Lumpur sample might help to study the demographic history of SE Asian *D. melanogaster* more precisely. Finally, this sample could be used to study differentiating selection.

Literature cited

AGUADÉ, M., N. MIYASHITA and C. H. LANGLEY, 1989 Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607-615.

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

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

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

AYALA, F. J., E. S. BALAKIREV and A. G. SAEZ, 2002 Genetic polymorphism at two linked loci, *Sod* and *Est-6*, in *Drosophila melanogaster*. Gene **300:** 19-29.

BANTIGNIES, F., C. GRIMAUD, S. LAVROV, M. GABUT and G. CAVALLI, 2003 Inheritance of Polycomb-dependent chromosomal interactions in Drosophila. Genes Dev. **17:** 2406-2420.

BAUDRY, E., N. DEROME, M. HUET and M. VEUILLE, 2006 Contrasted polymorphism patterns in a large sample of populations from the evolutioanry genetics model *D. simulans*. Genetics **173:** 759-767.

BAUDRY, E., B. VIGINIER and M. VEUILLE, 2004 Non-African populations of *Drosophila melanogaster* have a unique origin. Mol. Biol. Evol. **21:** 1482-1491.

BAUER DUMONT, V. and C. F. AQUADRO, 2005 Multiple signatures of positive selection downstream of *Notch* on the X chromosome of *Drosophila melanogaster*. Genetics **171:** 639-653.

BEGUN, D. J. and C. F. AQUADRO, 1992 Levels of naturally occuring 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, 1994 Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of Drosophila: Selection and geographic differentiation. Genetics **136:** 155-171.

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.

BETRÁN, E. and M. LONG, 2003 *Dntf-2r*, a young Drosophila retroposed gene with specific male expression under positive darwinian selection. Genetics **164:** 977-988.

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., A. FLEDEL-ALON, S. WILLIAMSON, R. NIELSEN, M. T. HUBISZ *et al.*, 2005 Natural selection on protein-coding genes in the human genome. Nature **437:** 1153–1157.

CARTHARIUS, K., K. FRECH, K. GROTE, B. KLOCKE, M. HALTMEIER *et al.*, 2005 MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics **21:** 2933-2942.

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., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. **15:** 538-543.

CHARLESWORTH, D., B. CHARLESWORTH and M. T. MORGAN, 1995 The pattern of

neutral molecular variation under the background selection model. Genetics **141:** 1619-1632.

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

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.

CRICK, F. H. C., 1958 On protein synthesis. Symp. Soc. Exp. Biol. **12:** 138-63.

DABORN, P. J., J. L. YEN, M. R. BOGWITZ, G. LE GOFF, E. FEIL *et al.* 2002 A single P450 alleleassociated with insecticide resistance in *Drosophila.* Science **297:** 2253-2256.

DARWIN, C., 1839 Journal of researches into the geology and natural history of the surveying voyages of H.M.S. Adventure and Beagle, Colburn, London. In: DARWIN, C., 1989, Voyage of the Beagle. Penguin, London.

DARWIN, C., 1859 The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. Murray, London. In: DARWIN, C., 1985 The origin of species, Penguin classics, London.

DAS, A., S. MOHANTY and W. STEPHAN, 2004 Inferring the population structure and demography of *Drosophila ananassae* from multilocus data. Genetics **168:** 1975-1985.

DAVID, J. R., C. BOCQUET and E. PLA, 1976 New results on the genetic characteristics of the Far East race of *Drosophila melanogaster.* Genet. Res. **28:** 253-260.

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

DEATRICK, J., M. DALY, N. B. RANDSHOLT and H. W. BROCK, 1991 The complex genetic locus *polyhomeotic* in *Drosophila melanogaster* potentially encodes two homologous zin-finger proteins. Gene **105:** 185-195.

DEPAULIS, F., S. MOUSSET and M. VEUILLE, 2003 Power of neutrality tests to detect bottlenecks and hitchhiking. J. Mol. Evol. **57:** S190-S200.

DE VRIES, H., 1901 Die Mutationstheorie. Versuche und Beobachtungen über die Entstehung von Arten im Pflanzenreich. Veit, Leipzig.

DIERINGER, D., V. NOLTE and C. SCHLÖTTERER, 2005 Population structure in African *Drosophila melanogaster* revealed by microsatellite analysis. Mol. Ecol. **14:** 563-573.

DURA, J.-M., N. B. RANDSHOLT, J. DEATRICK, I. ERK, P. SANTAMARIA *et al.*, 1987 A complex genetic locus, polyhomeotic, is required for segmental specification and epidermal development in *D. melanogaster*. Cell **51:** 829-839.

EANES, W. F., 1999 Analysis of selection on enzyme polymorphisms. Annu. Rev. Ecol. Syst. **30:** 301-326.

ENARD, W., P. KHAITOVICH, J. KLOSE, S. ZÖLLNER, F. HEISSIG *et al.*, 2002 Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340-343.

EXCOFFIER, L., G. LAVAL and S. SCHNEIDER, 2005 Arlequin ver. 3.0: an integrated software package for population genetic data analysis. Evol. Bioinf. Onl. **1:** 47-50.

EXCOFFIER, L., P. E. SMOUSE and J. M. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondiral DNA restricion data. Genetics **131:** 479-491.

FAIRBANKS, R. G., 1989 A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on the Younger Dryas event and deep ocean circulation. Nature **342:** 637-642.

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

FLYBASE CONSORTIUM, 2003 The Flybase database of the Drosophila genome projects and community literature. Nucleic Acids Res. **31:** 172-175.

FFRENCH-CONSTANT, R. H., T.A. ROCHELAU, J. C. STEICHEN and A. E. CHALMERS, 1993 A point mutation in a Drosophila GABA receptor confers insecticide resistance. Nature **363:** 449-451.

FU, Y.-X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915-925.

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

GALTIER, N., M. GOUY and C. GAUTIER, 1996 SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phyolgeny. Comput. Appl. Biosci. **12:** 543-548.

GALTIER, N., F. DEPAULIS and N. H. BARTON, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics **155:** 981–987.

GILLESPIE, J. H., 2004 Population genetics, a concise guide, 2nd edition. The John Hopkins University Press, 2004.

GLINKA, S., D. DE LORENZO and W. STEPHAN, 2006 Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster*. Mol. Biol. Evol. **23:** 1869-1878.

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.

GLINKA, S., W. STEPHAN and A. DAS, 2005 Homogeneity of common cosmopolitan inversion frequencies in Southeast Asian *Drosophila melanogaster*. J. Genet. **84:** 173-178.

GORDO, I. and B. CHARLESWORTH, 2001 Genetic linkage and molecular evolution. Curr. Biol. **11:** R684-R686.

GOUDET, J., M. RAYMOND, T. DE MEEÜS and F. ROUSSET, 1996 Testing differentation in diploid populations. Genetics **144:** 1933-1940.

GREGORY, S. L., R. D. KORTSCHAK, B. KALIONIS and R. SAINT, 1996

Characterization of the *dead ringer* gene identifies a novel, highly conserved family of sequence-specific DNA-binding proteins. Mol. Cell. Biol. **16:** 792-799.

GRUMBLING, G., V. STRELETS and THE FLYBASE CONSORTIUM, 2006 FlyBase: anatomical data, images and queries. Nucleic Acids Res. **34:** D484-D488.

HADDRILL, P. R., K. R. THORNTON, B. CHARLESWORTH and P. ANDOLFATTO, 2005 Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. **15:** 790-799.

HAMBLIN, M. T., A. M. CASA, H. SUN, S. C. MURRAYA. H. PATERSON *et al.*, 2006 Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. Genetics 1**73:** 953-964.

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.

HARR, B., C. VOOLSTRA, T. J. HEINEN, J. F. BAINES, R. ROTTSCHEIDT *et al.*, 2006 A change of expression in the conserved signaling gene MKK7 is associated with a selective sweep in the western house mouse *Mus musculus domesticus*. J. Evol. Biol. **19:** 1486-1496.

HARTL, D. L. and A. G. CLARK, 1997 Principles of population genetics, 3rd edition. Sinauer Associates, Sunderland, USA.

HILL, W. G. and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226-231.

HODGSON, J. W., N. N. CHENG, D. A. R. SINCLAIR, M. KYBA, N. B. RANDSHOLT and H. W. BROCK, 1997 The *polyhomeotic* locus of *Drosophila melanogaster* is transcriptionally and post-transcriptionally regulated during embryogenesis. Mech. Dev. **66:** 69-81.

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

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

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

HUDSON, R. R., K. BAILEY, D. SKARECKY, J. KWIATOWSKI and F. J. AYALA, 1994 Evidence for positive selection in the *superoxide dismutase (sod)* region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

INGER, R. F. and H. K. VORIS, 2001 The biogeographical relations of the frogs and snakes of Sundaland. J. Biogeogr. **28:** 863-891.

INGHAM, P. W. and A. MARTINEZ-ARIAS, 1986 The correct activation of *Antennapedia* and bithorax complex genes requires the *fushi tarazu* gene. Nature **324:** 592-597.

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.

KALARI, K. R., M. CASAVANT, T. B. BAIR, H. L. KEEN, J. M. COMERON *et al.*, 2006 First exons and introns – a survey of GC content and gene structure in the human genome. In Silico Biol. **6:** 0022.

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

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 A microsatellite variability screen for positive selection associated with the "out of Africa" habitat expansion of *Drosophila melanogaster*. Genetics **165:**

1137-1148.

KAYSER, M., S. BRAUER and M. STONEKING, 2003 A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol. Biol. Evol. **20:** 893-900.

KEALHOFER, L., 2002 Changing perceptions of risk: the development of agro-ecosystems in Southeast Asia. Am. Anthr. **104:** 178-194.

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

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

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

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

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

KIMURA, M., 1971 Theoredical foundation of population genetics at the molecular level. Theor. Pop. Biol. **2:** 174-208.

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

KIMURA, M. and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725-738.

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

KLEIN, R. G., 1999 The human career: human biological and cultural origins, 2nd edition. The University of Chicago Press, Chicago, USA.

KREITMAN, M., 2000 Methods to detect selection in populations with applications to the human. Annu. Rev. Genomics Hum. Genet. **1:** 539-559.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings in

Bioinf. **5:** 150-163.

LACHAISE, D., M. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup, pp. 159–225 in *Evolutionary Biology*, edited by M. K. HECHT, B. WALLACE and G. T. PRANCE. Plenum, New York.

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

LAMARCK, J.-B., 1809 Philosophie Zoologique. 1909: Zoologische Philosophie (Deutsch von SCHMIDT, H.). Kröner, Leipzig.

LANGLEY, C. H., B. P. LAZZARO, W. PHILLIPS, E. HEIKKINEN and J. BRAVERMAN, 2000 Linkage disequilibrium and the site frequency spectra in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster* X chromosome. Genetics **156:** 1837-1852.

LEHMANN, M., 2004 Anything else but GAGA: a nonhistone protein complex reshapes chromatin structure. Trends Genet. **20:** 15-22.

LEMEUNIER, F., J. R. DAVID, L. TSACAS and M. ASHBURNER, 1986 The *melanogaster* species group, pp. 147-256 in: The genetics and biology of *Drosophila*, Vol. 3e, edited by M. Ashburner and E. Novitski, Cademic Press, New York.

LI, Y.-J., Y. SATTA and N. TAKAHATA, 1999. Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes Get. Syst. **74:** 117-127.

LI, H. and W. STEPHAN, 2005 Maximum likelihood methods for detecting recent positive selection and localizing the selected site in the genome. Genetics **171:** 377-384.

LI, H. and W. STEPHAN, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet., in press.

LUND, A. H. and M. VAN LOHUIZEN, 2004 Polycomb complexes and silencing mechanisms, Curr. Opin. Cell Biol. **16:** 239-246.

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

MAYR, E., 2001 Das ist Evolution. Goldmann, München.

MCVEAN, G., P. AWADALLA and P. FEARNHEAD, 2002 A coalescent-based method for detecting and estimating recombination rates form gene sequences. Genetics **160:** 1231-1241.

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.

MUTERO, A., M. PRALAVORIO, J.-M. BRIDE and D. FOURNIER, 1994 Resistance-associated point mutations in insecticide-insensitive acetylcholinesterase. Proc. Natl. Acad. Sci, USA **91:** 5922-5926.

NEI, M., 1987 *Molecular evolutionary genetics*. Columbia University press, New York.

NURMINSKY, D. I., M. V. NURMINSKAYA, D. DE AGUIAR and D. L. HARTL, 1998 Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature **396:** 572-575.

OHTA, T., 1972 Population size and rate of evolution. J. Mol. Evol. **1:** 305-314.

OMETTO, L., 2006 The selective and demographic history of *Drosophila melanogaster*. PhD thesis.

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

OSADA, N., M. H. KOHN and C.-I WU, 2006 Genomic inferences of the *cis*-regulatory nucleotide polymorphisms underlying gene expression differences between *Drosophila melanogaster* mating races. Mol. Biol. Evol. **23:** 1585-1591.

PAGE, R. D. M., 1996 TREEVIEW: An application to display phylogenetic trees on personal computers. Comp. Appl. Biosci. **12:** 357-358.

PAYSEUR, B. A. and M. W. NACHMAN, 2002 Natural selection at linked sites

in humans. Gene **300:** 31-42.

PINHASI, R., J. FORT and A. J. AMMERMAN, 2005 Tracing the origin and spread of agriculture in Europe. PloS Biol. **3:** e410.

POOL, J. E., V. BAUER DUMONT, J. L. MUELLER and C. F. AQUADRO, 2006 A scan of molecular variation leads to the narrow localization of a selective sweep affecting both afrotropical and cosmopolitan populations of *Drosophila melanogaster*. Genetics **172:** 1093-1105.

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945-959.

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

PRZEWORSKI, M., 2003 Estimating the time since the fixation of a beneficial allele. Genetics **164:** 1667-1676.

PRZEWORSKI, M., J. D. WALL and P. ANDOLFATTO, 2001 Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 291-298.

QUESADA, H., U. E. M. RAMIREZ, J. ROZAS and M. AGUADÉ, 2003 Large-scale adaptive hitchhiking upon high recombination in *Drosophila melanogaster*. Genetics **165:** 895-900.

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.

RANDI, E., M. PIERPAOLI, M. BEAUMONT, B. RAGNI and A. SFORZI, 2001 Genetic identification of wild and domestic cats (*Felis silvestris*) and their hybrids using bayesian clustering methods. Mol. Biol. Evol. **18:** 1679-1693.

RIEDL, R., 2003 Kulturgeschichte der Evolutionstheorie. Springer Verlag, Berlin.

RONALD, J., R. B. BREM, J. WHITTLE and L. KRUGLYAK, 2005 Local regulatory variation in *Saccharomyces cerevisiae*. PLoS Genet. **1:** e25.

ROSENBERG, N. A. and M. NORDBORG, 2002 Genealogical trees, coalescent theory

and the analysis of genetic polymorphisms. Nat. Rev. Genet. **3:** 380-390.

ROZAS, J., J. C. SÁNCHEZ-DEL BARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

SAEZ, A. G., A. TATARENKOV, E. BARRIO, N. H. BECERRA and F. J. AYALA, 2003 Patterns of DNA sequence polymorphism at *Sod* vicinities in *Drosophila melanogaster*: Unraveling the footprint of a recent selecctiv sweep. Proc. Natl. Acad. USA **100:** 1793-1798.

SAITOU, N and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.

SANTIAGO, E. and A. CABALLERO, 2005 Variation after a selective sweep in a subdivided population. Genetics **169:** 475-483.

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.

SCHLÖTTERER, C., 2002 A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics **160**: 753-763.

SCHLÖTTERER, C., H. NEUMEIER, C. SOUSA and V. NOLTE, 2006 Highly structured Asian *Drosophila melanogaster* populations: a new tool for hitchhiking mapping? Genetics **172:** 287-292.

SCHLÖTTERER, C., C. VOGL and D. TAUTZ, 1997 Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. Genetics **146:** 309-320.

SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169:** 1601-1615.

SCHMITZ, S., 1983 Charles Darwin. Leben, Werk, Wirkung. Econ, Düsseldorf.

SCOTT, M. J. and J. C. LUCCHESI, 1991 Structure and expression of the *Drosophila*

melanogaster gene encoding 6-phosphogluconate dehydrogenase. Gene **109:** 177-183.

SERRANO, N., H. W. BROCK, C. DEMERET, J.-M. DURA, N. B. RANDSHOLT *et al.*, 1995 *Polyhomeotic* appears to be a target of Engrailed regulation in *Drosophila*. Development **121:** 1691-1703.

SHANDALA, T., R. D. KORTSCHAK, S. GREGORY and R. SAINT, 1999 The *Drosophila dead ringer* gene is required for early embrynoci patterning through regulation of *argos* and *buttonhead* expression. Development **126:** 4341-4349.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555-562.

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

STEPHAN, W., and C. H. LANGLEY, 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila melanogaster* populations. I. Contrasts between the *vermilion* and *forked* loci. Genetics **121:** 89-99.

STEPHAN, W., T. H. E. WIEHE and M. W. LENZ, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237-254.

STEPHAN, W., Y. S. SONG and C. H. LANGLEY, 2006 The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics **172:** 2647-2663.

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

STRANGER, B. E., M. S. FORREST, A. G. CLARK, M. J. MINICHIELLO, S. DEUTSCH *et al.*, Genome-wide associations of gene expression variation in humans. PloS Genet. **1:** e78.

STUPAR, R. M. and N. M. SPRINGER, 2006 *Cis*-transcriptional variation in maize

inbred lines B73 and Mo17 leads to additive expression patterns in the F_1 hybrid. Genetics **173:** 2199-2210.

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.

THE FLYBASE CONSORTIUM, 2003 The FlyBase database of the Drosophila genome Projects and community literature. Nucleic Acids Res. **31:** 172-175.

THOMSON, G., 1977 The effect of a selected locus on linked neutral loci. Genetics **85:** 753-788.

THORNTON, K. and P. ANDOLFATTO, 2006 Approximate bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics **172:** 1607-1619.

VERNI, F., R. GANDHI, M. L. GOLDBERG and M. GATTI, 2000 Genetic and molecular analysis of *wings apart-like (wapl)*, a gene controlling heterochromatin organization in *Drosophila melanogaster*. Genetics **154:** 1693-1710.

VOGL, C., A. DAS, M. BEAUMONT, S. MOHANTY and W. STEPHAN, 2003 Population subdivision and molecular sequence variation: theory and analysis of *Drosophila ananassae* data. Genetics **165:** 1385-1395.

VORIS, H. K., 2000 Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. J. Biogeogr. 27: 1153-1167.

WAHLUND, S., 1928 Zussammensetzung von Populationen und Korrelations-erscheinungen vom Standpunkt der Vererbungslehre aus betrachet. Hereditas **11:** 65-105.

WAKELEY, J., 2006 Coalescent theory, an introduction. Roberts and Company, Greenwood Village.

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

WEBB, T. 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.

WEILL, M., C. MALCOLM, F. CHANDRES, M. MOGENSEN, A. BERTHOMIEU *et al.*, 2004 The unique mutation in ace-1 giving high insecticide resistance is easily detectable in mosquito vectors. Ins. Mol. Biol. **13:** 1-7.

WEISSMANN, A., 1885 Die Continuität des Keimplasmas als Grundlage einer Theorie der Vererbung. Jena.

WITTKOPP, P. J., 2004. Evolution of *cis*-regulatory sequence and function in Diptera. Heredity **97:** 139-147.

ZHANG, Z., P. J. BRADBURY, D. E. KROON, T. M. CASSTEVENS and E. D. BUCKLER, 2006 TASSEL 2.0, a software package for association and diversity analyses in plants and animals. Available at www.maizegenetics.net.

Appendix A. Tables and Figures

TABLE A1. Polymorphism and summary statistics for the individual fragmens in the *ph-d – Pgd* region.

[a] fragment identification number

[b] identifier for merged fragmentds

[c] position relative to the first bp sequenced in the region

[d] number of bp used for analyses

[e] observed number of polymorphic sites

[f] number of haplotypes

[g] haplotype diversity (NEI 1987)

[h] Watterson's (1975) estimator, based on the number of segregating sites

[i] Tajima's (1983) estimator, based on the average number of paiwise differences

[k] Tajima's (1989) *D* statistic to test the standard neutral model

[l] linkage disequilibrium (KELLY 1997). Only informative sites were used

[m] number of bp used for analyses (with outgroup sequence)

[n] number of fixed differences between species

[o] divergence between species

NA denotes a parameter estimate that is not available, due to limited data.
Only estimates for noncoding loci are presented.

The table is presented on the following page.

TABLE A1.

FIN [a]	Locus [b]	Position [c]	L [d]	S [e]	h [f]	Hd [g]	θ [h]	π [i]	D [k]	Z_{nS} [l]	L [m]	D_s [n]	K [o]
1	**1**	1 - 515	276	3	4	0.682	0.0038	0.0031	-0.579	NA	264	16	0.06
2		623 - 1109											
3		1,920 - 2,521											
4		2,717 - 3,464											
5		3,562 - 4,051											
6		4,112 - 4,886											
7		4,583 - 5,198											
8	**2**	5,246 - 5,884	639	3	4	0.727	0.0016	0.0017	0.322	0.180	NA	NA	NA
9		5,724 - 6,351	469	2	3	0.530	0.0014	0.0012	-0.382	NA	470	31	0.05
10		6,492 - 7,078	585	3	4	0.455	0.0017	0.0009	-1.629	NA	577	16	0.03
11	**3**	6,712 - 7,335	259	4	5	0.576	0.0051	0.0031	-1.385	NA	259	18	0.05
12		7,290 - 7,858	497	2	3	0.530	0.0013	0.0012	-0.382	NA	450	30	0.07
13		7,793 - 8,493	637	5	6	0.682	0.0026	0.0019	-0.988	0.067	351	52	0.15
14		8,585 - 9,195	613	8	5	0.788	0.0043	0.0033	-0.933	0.211	602	41	0.07
15	**4**	9,038 - 9,637	437	4	5	0.667	0.0030	0.0031	0.104	0.556	549	44	0.11
16		9,409 - 10,122	488	2	3	0.439	0.0014	0.0010	-0.850	NA	446	31	0.05
17		10,124 - 10,578											
18		10,625 - 11,327											
19		11,683 - 12,171											
20		12,118 - 12,846											
21		12,680 - 13,402											
22		14,006 - 14,614											
23		14,424 - 15,094											
24		14,548 - 15,254											
25		14,694 - 15,274	136	0	1	0.000	0.0000	0.0000	NA	NA	116	3	0.03
26		17,086 - 17,557	472	2	3	0.439	0.0014	0.0010	-0.850	NA	NA	NA	NA
27		17,349 - 17,936	380	1	2	0.167	0.0009	0.0004	-1.141	NA	379	18	0.05
28	**5**	17,739 - 18,402	468	2	3	0.318	0.0014	0.0007	-1.451	NA	476	7	0.02
29		18,134 - 18,861	461	5	5	0.576	0.0036	0.0018	-1.831	NA	456	32	0.07
30		18,706 - 19,437	578	2	3	0.318	0.0012	0.0006	-1.451	NA	451	15	0.03
31		19,020 - 19,699											
32	**6**	19,762 - 20,412	649	5	5	0.576	0.0026	0.0013	-1.831	NA	631	25	0.04
33	**7**	20,506 - 21,096	590	6	7	0.879	0.0034	0.0031	-0.324	0.246	NA	NA	NA
34		21,794 - 22,542	730	9	9	0.909	0.0041	0.0021	-2.016	NA	639	50	0.09
35		22,023 - 22,796	258	1	2	0.167	0.0013	0.0007	-1.141	NA	NA	NA	NA
36	**8**	22,727 - 23,481	686	4	3	0.318	0.0019	0.0016	-0.661	1.000	649	48	0.07
37		23,138 - 23,782	313	3	4	0.455	0.0032	0.0016	-1.629	NA	314	43	0.15
38		23,550 - 24,062	282	1	2	0.167	0.0012	0.0006	-1.141	NA	294	23	0.08
39		23,962 - 24,725	651	9	7	0.833	0.0046	0.0029	-1.459	0.040	636	56	0.09
40	**9**	24,783 - 25,403	615	15	9	0.909	0.0081	0.0082	0.081	0.272	590	36	0.07
41	**10**	25,477 - 26,121	645	14	9	0.955	0.0072	0.0046	-1.547	0.226	569	22	0.04
42		25,986 - 26,720	347	8	6	0.758	0.0076	0.0045	-1.618	NA	344	27	0.08
43		26,457 - 26,759											
44		26,612 - 26,957											
45		26,809 - 27,633											
46		27,702 - 28,259											
47	**11**	28,600 - 29,149	291	2	2	0.303	0.0023	0.0021	-0.248	1.000	275	20	0.07
48		28,921 - 29,344	219	2	3	0.621	0.0030	0.0032	0.153	NA	NA	NA	NA
49	**12**	29,341 - 29,658	318	11	9	0.909	0.0115	0.0095	-0.717	0.309	273	12	0.05
50		29,505 - 29,981	323	2	3	0.318	0.0021	0.0010	-1.451	NA	NA	NA	NA
51		29,896 - 30,413	434	1	2	0.182	0.0008	0.0004	-1.129	NA	NA	NA	NA
52		30,437 - 31,070											
53		31,042 - 31,811											

FIGURE A1. Segregating sites in the *ph-d* – *Pgd* region.

[a] SNP identifier.

[b] Location of SNP: 3', 3'-flanking region; 5', 5'-flanking region; in, intron; n, noncoding site; r, replacement site; s, synonymous site.

[c] Line identifier: *D. melanogaster* lines mel #1 – mel #19 correspond to the European sample, mel #82 – mel #398 correspond to the African one.

[d] *D. simulans.*

[e] *D. yakuba.*

[f] consensus: The variant that shows the highest frequency in *D. melanogaster.*

NA denotes missing sequence information.

Grey fields indicate SNPs with (almost) fixed differences between *D. melanogaster* samples.

The corresponding position of a SNP is shown in the lower part of the figure. Black boxes indicate exons, lines connecting boxes indicate introns.

The figure is shown on the following pages.

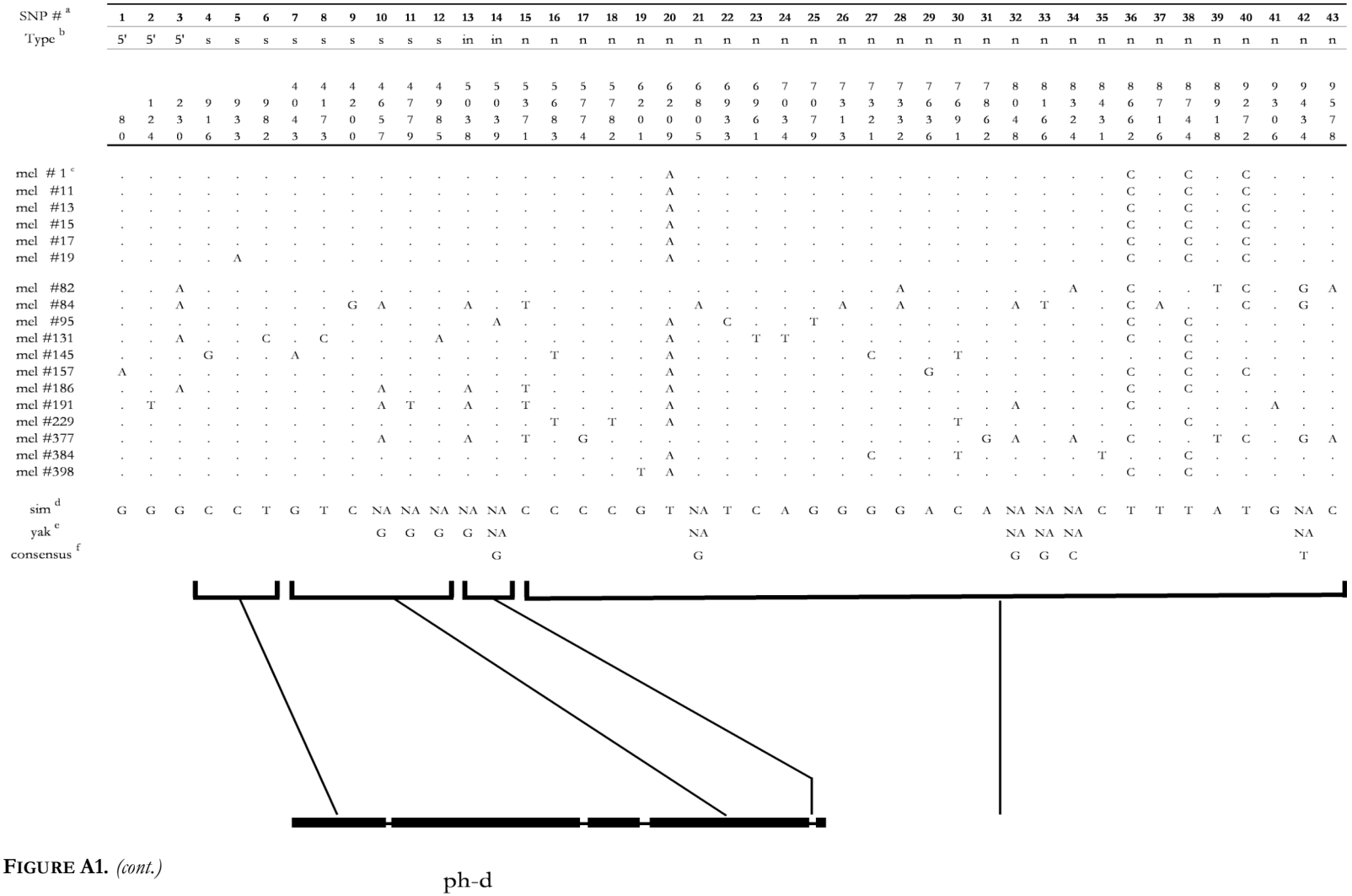


FIGURE A1. *(cont.)*

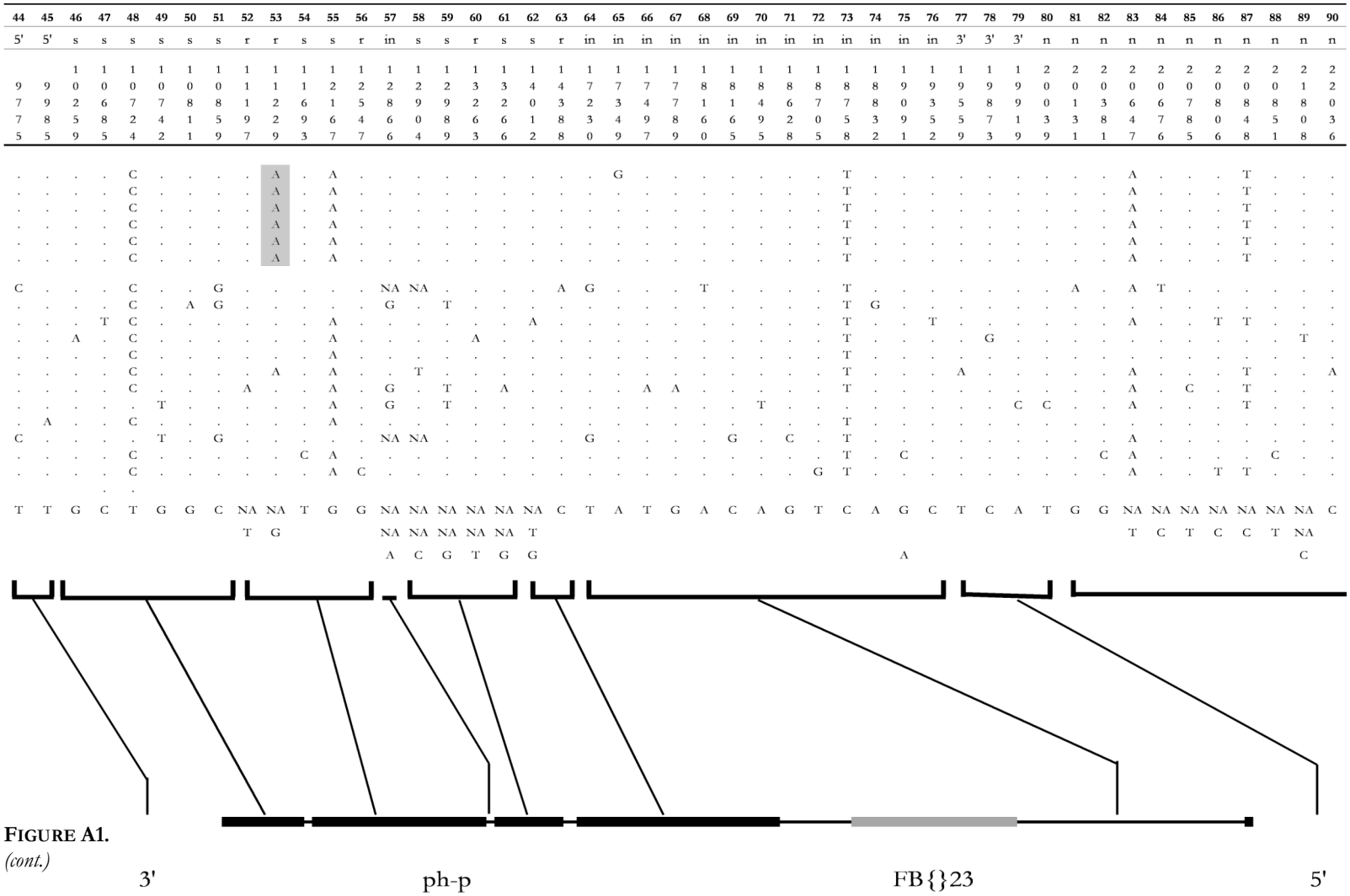


FIGURE A1.
(cont.)

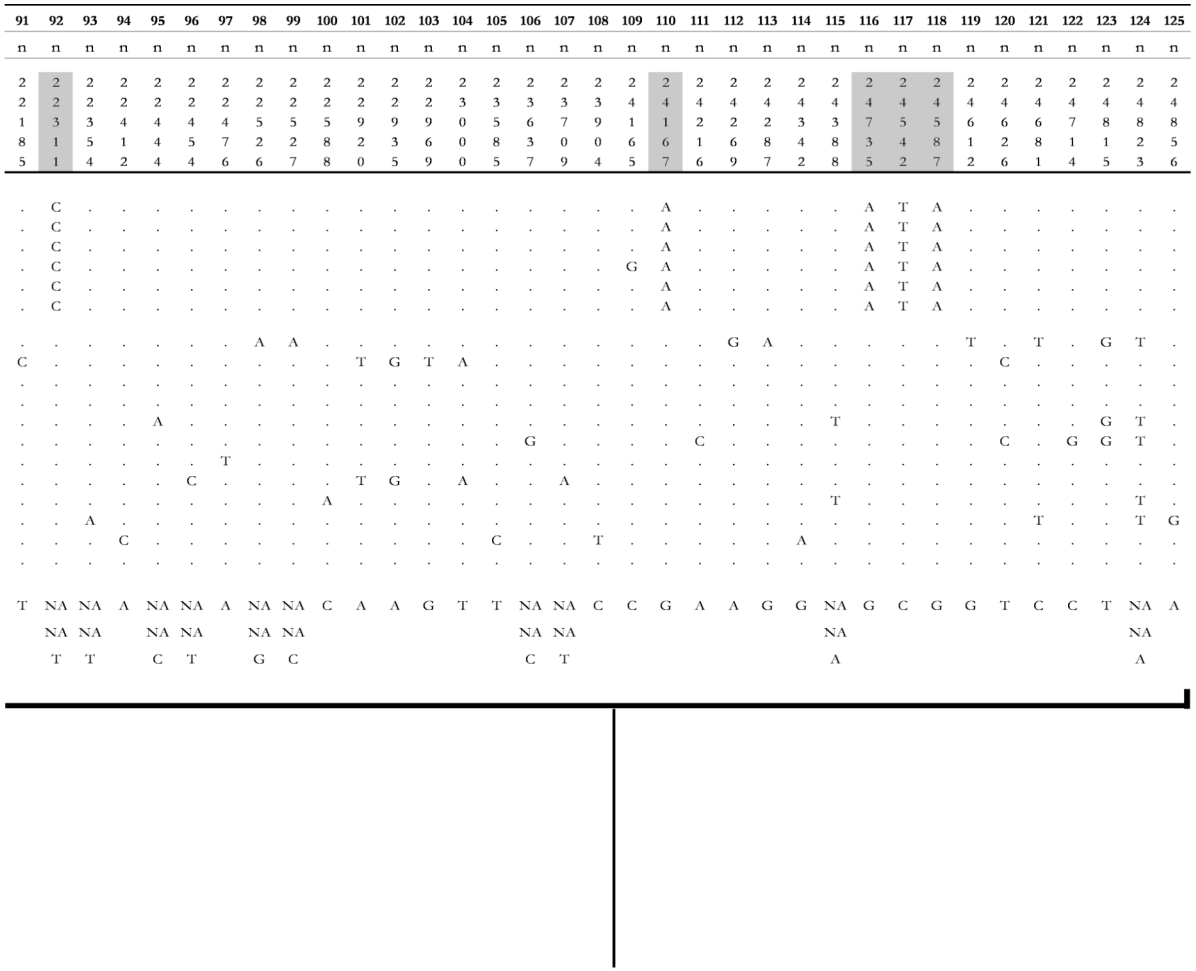


FIGURE A1. *(cont.)*

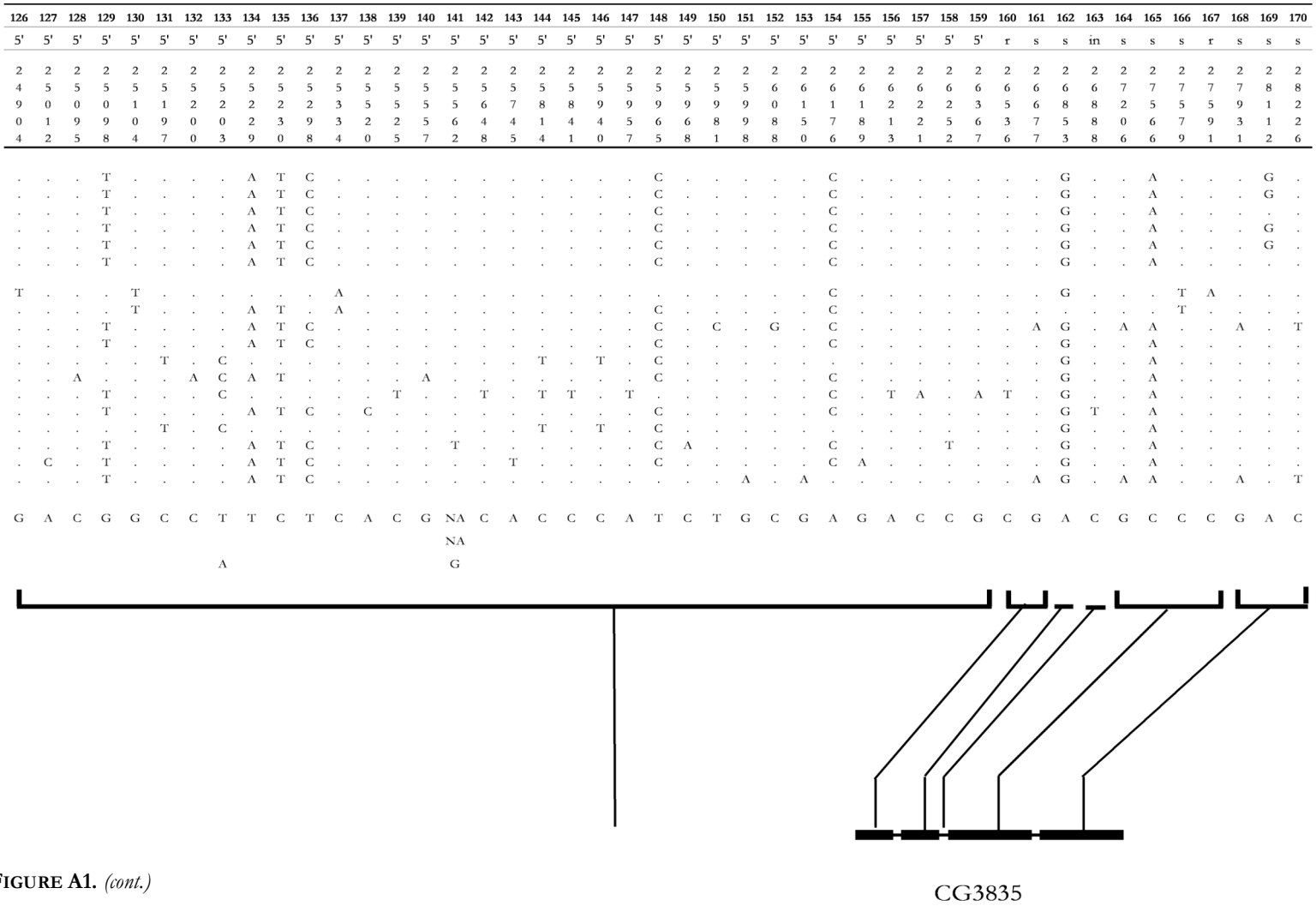


FIGURE A1. *(cont.)*

171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	SNP #
5'	5'	r	in	in	in	in	in	in	in	in	in	in	in	in	in	in	in	in	in	in	in	in	r	s	s	s	3'	3'	3'	n	Type
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
8	8	8	9	9	9	9	9	9	9	9	9	9	9	9	9	9	0	0	0	0	0	0	1	1	1	1	1	1	1	2	
7	7	9	1	2	3	3	3	4	4	5	5	5	5	5	7	7	1	4	4	4	5	5	1	4	5	7	8	0	9	1	
7	9	9	5	7	5	7	8	0	2	1	5	6	7	8	1	9	5	6	6	8	4	6	2	3	2	8	0	4	3	0	
1	2	0	4	1	3	6	3	2	2	6	3	3	2	5	5	2	7	6	8	2	6	9	3	6	3	7	3	8	0	1	
A	G	.	.	A	.	.	.	.	.	.	G	.	C	G	.	T	.	A	.	.	G	T	.	.	.	.	.	.	.	.	mel #1
A	G	.	.	A	.	.	.	.	.	.	G	.	C	G	.	T	.	A	.	.	G	T	.	.	.	.	.	.	.	.	mel #11
.	G	G	.	A	.	.	.	.	.	.	G	.	C	.	.	T	.	A	.	.	G	T	.	.	.	.	.	.	.	.	mel #13
A	G	.	.	A	.	.	.	.	.	.	G	.	C	G	.	T	.	A	.	.	G	T	.	.	.	.	.	.	.	.	mel #15
A	G	.	.	A	.	.	.	.	.	.	G	.	C	G	.	T	.	A	.	.	G	T	.	.	.	.	.	.	.	.	mel #17
A	G	.	.	A	.	.	.	.	.	.	G	.	C	G	.	T	.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	mel #19
.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	T	T	T	T	.	G	.	mel #82
.	.	.	.	.	T	.	.	.	.	.	G	A	C	.	.	T	.	.	T	.	.	.	.	T	T	T	T	.	G	.	mel #84
.	.	.	.	.	T	.	.	.	.	A	.	.	.	.	.	T	.	.	.	.	.	.	.	T	T	T	.	A	.	.	mel #95
.	.	.	.	.	.	.	.	T	.	.	G	.	C	.	.	.	.	.	.	A	.	.	.	T	T	T	.	.	.	.	mel #131
.	.	.	T	.	T	.	.	.	C	.	G	.	C	.	.	T	.	.	.	.	.	.	.	T	T	T	.	.	.	.	mel #145
A	G	.	.	A	.	.	.	.	.	.	G	.	C	G	.	T	.	A	.	.	G	T	.	.	.	.	.	.	.	.	mel #157
A	G	.	.	.	.	G	.	.	.	.	G	.	C	.	T	T	.	.	.	.	G	.	T	.	.	T	.	.	.	.	mel #186
.	.	.	T	.	T	.	.	.	C	.	G	.	C	.	.	T	.	.	.	.	.	.	.	T	T	T	.	.	.	G	mel #191
.	.	.	T	.	T	.	.	.	C	.	G	.	C	.	.	T	.	.	.	.	.	.	.	T	NA	NA	NA	NA	NA	NA	mel #229
.	.	.	T	.	T	.	T	.	.	.	G	.	C	.	.	T	A	.	.	.	.	.	.	T	T	T	.	.	.	.	mel #377
.	.	.	T	.	T	.	.	.	C	.	G	.	C	.	.	T	.	.	.	.	.	.	.	T	T	T	.	.	.	.	mel #384
.	.	.	.	.	T	.	.	.	.	A	.	.	.	.	.	T	.	.	.	.	.	.	.	T	T	T	.	A	.	.	mel #398
G	A	A	G	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	C	C	C	C	C	A	A	sim
				C	G	A	G	C	T	G	A	T	A	C	C	G	G	C	C	G	A	A	G								yak
												C																			consensus

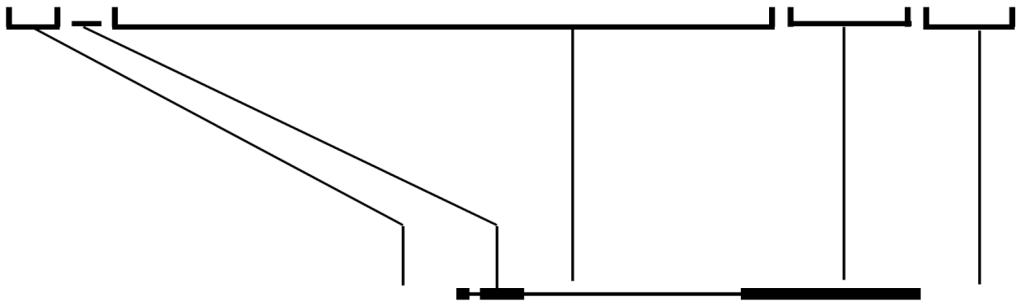
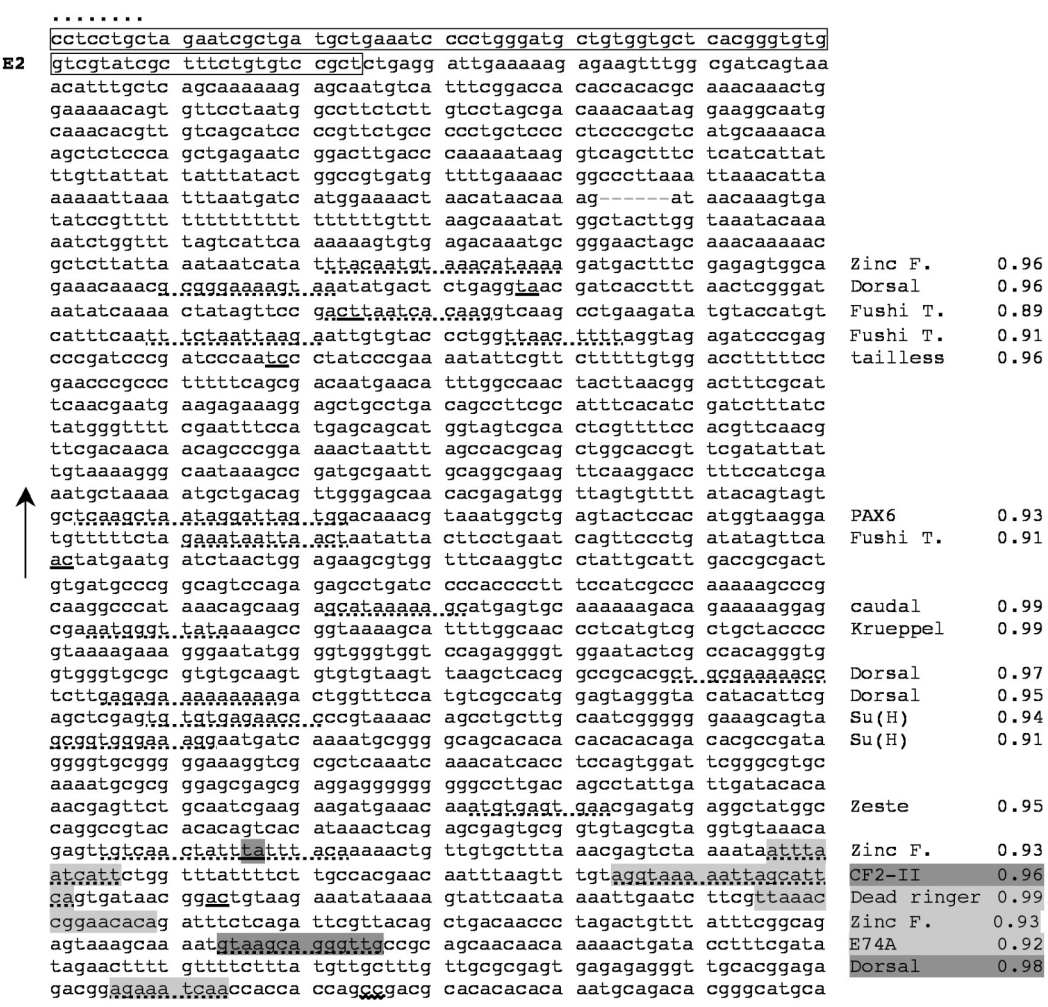


FIGURE A1. *(cont.)*

FIGURE A2. Transcription factor binding sites at *ph-p*.

Direction of transcription is bottom to top (indicated by vertical arrows). The position of the transposon (not present in our samples) is indiacted by dashes. Putative transcription factor binding sites (TFB) are denoted by dashed lines. Solid lines indicate indel positions. Exons are shown in boxes. E1, first exon; E2, second exon. The known transcription start point is indicated by a horizontal arrow. Dark grey backgrounds indicate putative TFBs private to *D. simulans*, those private to *D. melanogaster* are highlighted in light grey. The numbers represent sequence similarity scores to known TFB sequences.



(cont.)

FIGURE A2. *(cont.)*

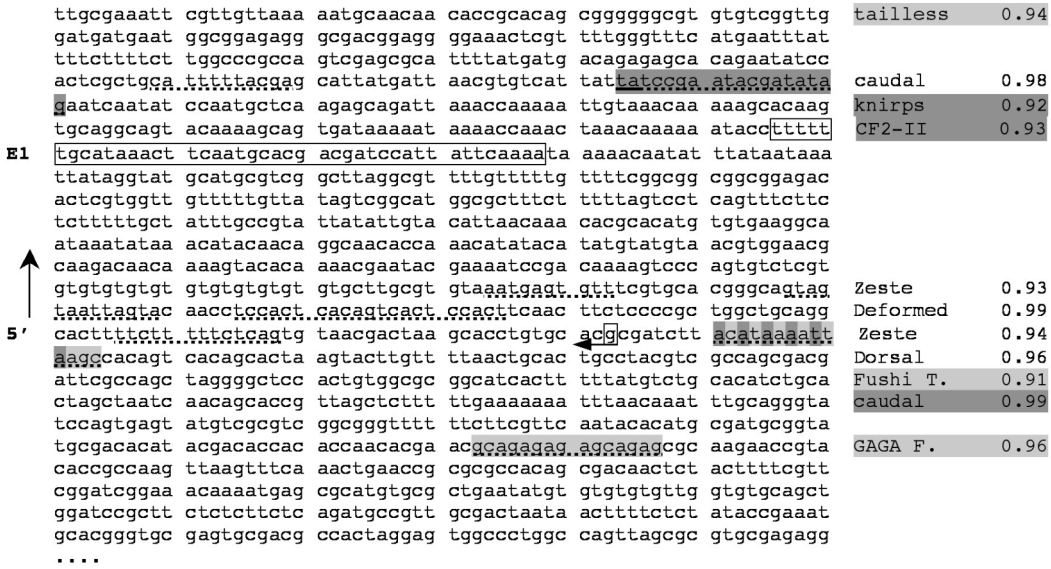


TABLE A2. Sampling locations of *D. melanogaster*.

Sample	Location	Country	Position
EUR	Leiden	The Netherlands	52°N 004°E
ZBM	Lake Kariba	Zimbabwe	NA
KL	Kuala Lumpur	Malaysia	03°N 102°E
KK	Kota Kinabalu	Malaysia	06°N 116°E
BKK	Bangkok	Thailand	14°N 101°E
CNX	Chiang Mai	Thailand	18°N 099°E
CEB	Cebu	Philippines	10°N 123°E
MNL	Manila	Philippines	14°N 121°E

TABLE A3. Parameter estimates for individual loci in SE Asia.

[a] Location of fragment: ig, intergentic; in, intron.

[b] Recombination rate (x10^{-8} rec/bp/gen), following COMERON *et al.* (1999).

[c] Length in bp.

[d] Parameters: *n*, sample size; *S*, number of observed segregating sites; *h*, number of haplotypes; *Hd*, haplotype diversity; and *D*, Tajima's *D*.

NA, not available.

Locus	loc.[a]	*r* [b]	*L* [c]	*Par.* [d]	KL	KK	BKK	CEB	MNL	CNX	ZBM	EUR
78	ig	3.55	520	*n*	14	12	15	12	7	11	12	12
				S	8	3	0	7	7	7	22	7
				h	4	2	1	6	2	4	12*	5
				Hd	0.780	0.545	NA	0.828	0.571	0.491	1.000*	0.788
				π	0.0074	0.0032	NA	0.004	0.0077	0.0037	0.0111	0.0068
				θ	0.0048	0.0019	NA	0.0045	0.0055	0.0046	0.0141	0.0045
				D	1.962	2.123*	NA	-0.435	2.071*	-0.786	-0.927	2.0211*
206	ig	3.00	504	*n*	12	11	16	12	7	13	12	12
				S	3	3	2	4	6	3	30	5
				h	3	5*	3	5	2	3	11*	4
				Hd	0.591	0.855*	0.608	0.833	0.476	0.564	0.985*	0.561
				π	0.0017	0.0028	0.0014	0.0033	0.0057	0.0015	0.015	0.0022
				θ	0.002	0.002	0.0012	0.0026	0.0005	0.0019	0.0203	0.0034
				D	-0.429	1.316	0.378	0.908	0.847	-0.645	-1.048	-1.291
237	ig	4.14	499	*n*	12	12	16	12	7	12	12	12
				S	22	4	5	16	1	15	42	23
				h	4	3	3	3	2	4	12*	8
				Hd	0.773	0.667	0.708	0.621	0.286	0.742	1.000*	0.9390
				π	0.0166	0.0032	0.0048	0.0171	0.0006	0.0107	0.0273	0.0200
				θ	0.0146	0.0027	0.003	0.0106	0.0008	0.01	0.0309	0.0153
				D	0.599	0.667	1.898	2.615*	-1.006	0.304	-0.534	1.369
259	in	4.81	295	*n*	12	12	15	11	7	13	12	12
				S	11	6	2	2	0	1	17	15
				h	4	3	3	3	1	2	7	3
				Hd	0.712	0.682	0.562	0.618	NA	0.538	0.773	0.530
				π	0.0143	0.0083	0.0021	0.0025	NA	0.0018	0.0153	0.0184
				θ	0.0124	0.0067	0.0021	0.0023	NA	0.0011	0.0207	0.0173
				D	0.668	0.897	-0.024	0.199	NA	1.475	-1.135	0.290
359	ig	3.06	526	*n*	14	12	16	12	7	13	12	12
				S	2	3	2	6	5	3	22	10
				h	3	3	2	4	3	3	12*	5
				Hd	0.385	0.621	0.125	0.561	0.714	0.667	1.000*	0.788
				π	0.0008	0.0026	0.0004	0.0036	0.0055	0.0025	0.0124	0.0059
				θ	0.0012	0.0019	0.0012	0.0038	0.0039	0.0018	0.0139	0.0063
				D	-0.959	0.993	-1.498*	-0.149	1.982*	1.115	-0.486	-0.236

(cont.)

TABLE A3. *(continued)*

Locus	loc.	*r*	*L*	*Par.*	KL	KK	BKK	CEB	MNL	CNX	ZBM	EUR
392	in	4.62	447	*n*	11	12	14	12	7	8	12	12
				S	4	2	2	3	3	2	13	4
				h	5	2	3	3	2	3	7	4
				Hd	0.782	0.409	0.582	0.682	0.571	0.607	0.773	0.697
				π	0.0025	0.0018	0.0015	0.0025	0.0038	0.0022	0.0059	0.0035
				θ	0.0031	0.0015	0.0014	0.0022	0.0027	0.0017	0.0104	0.0030
				D	-0.639	0.688	0.179	0.472	1.811*	0.932	-1.836*	0.627
422	in	2.78	578	*n*	12	12	16	12	7	13	11	12
				S	17	0	1	1	1	1	28	18
				h	5	1	2	2	2	2	11*	7
				Hd	0.780	NA	0.525	0.167	0.571	0.385	1.000*	0.909
				π	0.0067	NA	0.0009	0.0003	0.0009	0.0006	0.0150	0.0122
				θ	0.0103	NA	0.0005	0.0005	0.0006	0.0005	0.0165	0.0103
				D	-1.511*	NA	1.4737*	-1.141	1.342	0.426	-0.420	0.793
431	ig	2.49	468	*n*	13	12	13	12	7	13	12	12
				S	2	2	2	4	0	0	6	5
				h	3	2	2	3	1	1	4	6
				Hd	0.410	0.303	0.462	0.530	NA	NA	0.455	0.758
				π	0.0014	0.0013	0.0020	0.0029	NA	NA	0.0022	0.0022
				θ	0.0014	0.0014	0.0014	0.0029	NA	NA	0.0044	0.0035
				D	0.097	-0.248	1.214	-0.057	NA	NA	-1.894*	-1.224
721	ig	4.43	378	*n*	14	12	14	12	7	12	12	11
				S	1	1	1	1	0	1	31	24
				h	2	2	2	2	1	2	10	3
				Hd	0.527	0.303	0.495	0.167	NA	0.485	0.955	0.473
				π	0.0014	0.0008	0.0013	0.0004	NA	0.0013	0.0304	0.0205
				θ	0.0008	0.0008	0.0008	0.0009	NA	0.0009	0.0281	0.0217
				D	1.434	-0.195	1.212	-1.141	NA	1.066	0.372	-0.271
727	ig	2.68	413	*n*	12	12	16	12	7	13	12	10
				S	8	3	6	3	1	2	16	6
				h	5	3	4	3	2	3	11*	4
				Hd	0.727	0.682	0.758	0.439	0.571	0.564	0.985*	0.822
				π	0.0059	0.0033	0.0070	0.0024	0.0014	0.0023	0.0128	0.0067
				θ	0.0064	0.0024	0.0044	0.0024	0.0010	0.0015	0.0128	0.0051
				D	-0.293	0.993	2.067*	-0.028	1.342	1.438	-0.021	1.281

FIGURE A3. *wapl* – region SNP data of Kuala Lumpur, Europe and Africa.

EUR, Europe; KL, Kuala Lumpur; ZBM, Africa.

D. sim, D. simulans (outgroup).

KL04 is set as reference on top of the figure.

Grey shaded backgrounds indicate segregating sites private to the KL sample.

NA, no data available.

- indicates indel

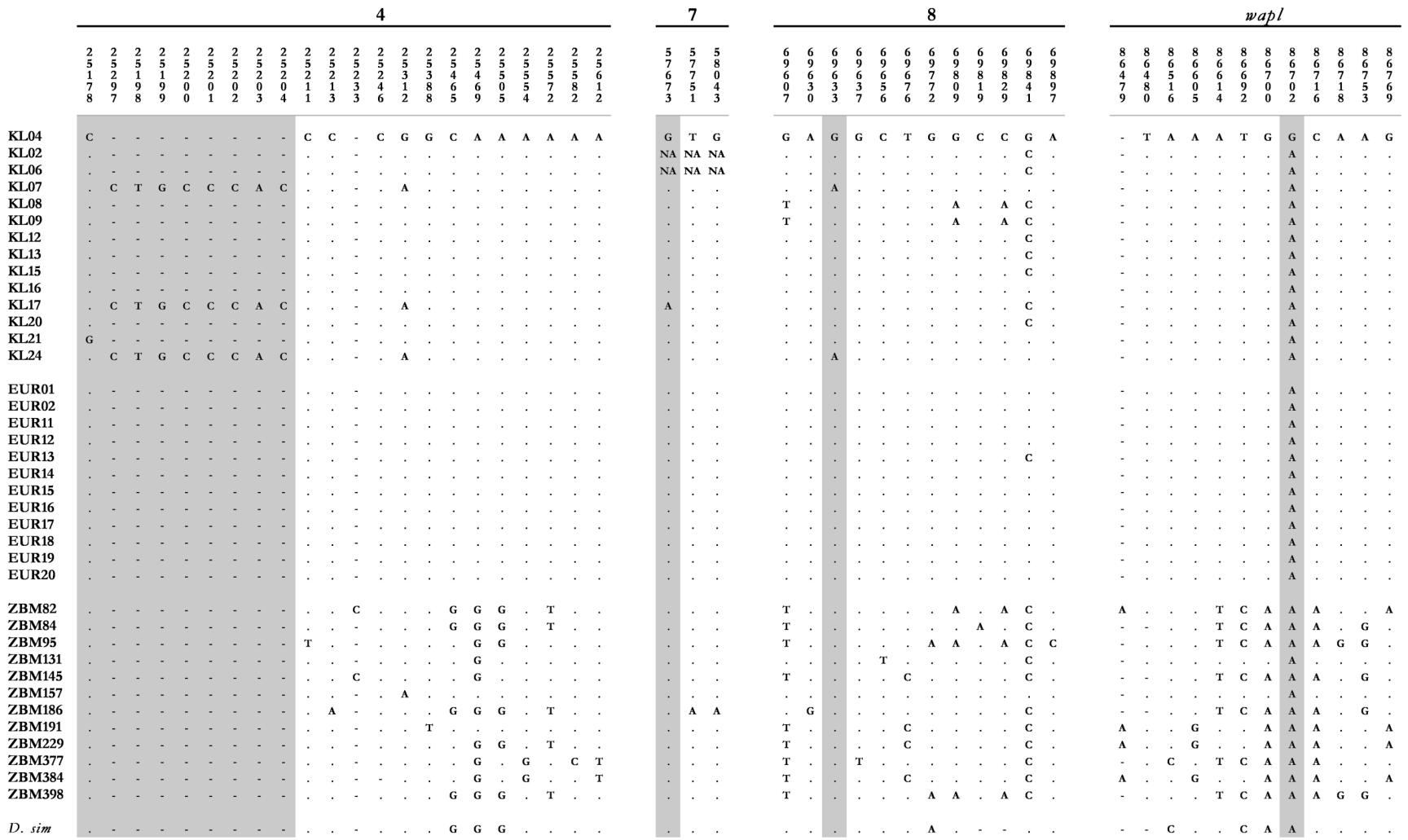

4
7
8
wapl
KL04
KL02
KL06
KL07
KL08
KL09
KL12
KL13
KL15
KL16
KL17
KL20
KL21
KL24
EUR01
EUR02
EUR11
EUR12
EUR13
EUR14
EUR15
EUR16
EUR17
EUR18
EUR19
EUR20
ZBM82
ZBM84
ZBM95
ZBM131
ZBM145
ZBM157
ZBM186
ZBM191
ZBM229
ZBM377
ZBM384
ZBM398
D. sim

FIGURE A3 *(cont.)*

										9									
9567	**9570**	**9598**	**9607**	**9637**	**9658**	**9623**	**9692**	**9632**	**9643**	**9658**	**9663**	**9629**	**9629**	**9633**	**9633**	**9633**	**9633**	**9633**	**9638**
C	C	G	G	A	T	C	A	G	G	G	C	C	C	G	A	A	C	A	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	-	T	T	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	
.	.	.	.	.	.	.	C	.	.	C	.	.	T	A	.	.	.	.	
T	.	.	C	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	
.	T	C	.	.	.	.	.	.	T	.	.	.	T	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
T	.	.	C	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
T	.	.	C	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
T	.	.	C	G	.	A	C	.	.	.	T	.	T	.	.	.	.	.	
.	.	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	
.	.	.	.	.	G	.	.	.	.	.	.	T	T	.	.	.	.	.	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	

Appendix B. Protocols

PROTOCOL B1. DNA isolation

DNA was isolated from 10-15 flies using the Puregene DNA isolation kit (Gentra Systems, Minneapolis, USA):

Cell Lysis

1) Chill a 1.5 ml tube containing 300 µl of Cell Lysis Solution (put 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

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

Protein precipitation

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

DNA precipitation

1) 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).
2) Mix the sample by inverting gently 50 times.
3) Centrifuge at 13,000–16,000 rpm for 1 minute.
4) Pour off supernatant and drain tube on clean absorbent paper.
5) Add 300 µl of 70% ethanol and invert tube several times to wash the DNA pellet.
6) 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.
7) Invert and drain the tube on clean absorbent paper and allow to dry at room temperature for 15 minutes.

DNA hydration

1) Add 50 µl of DNA Hydration Solution. (Note: sometimes H_2O was used)
2) 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.
3) If particles 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.
4) Store DNA at 2–8 °C (or at –20 °C for long term storage).

PROTOCOL B2. PCR clean-up

PCR reactions were cleaned using EXOSAP-IT (USB, Cleveland, USA)

1) Add 1 µl of ExoSAP-IT to 10 µl of PCR product.
2) Mix and incubate at 37 °C for 30 minutes.
3) Heat to 80 °C for 15 minutes.

PROTOCOL B3. Sequening reactions (for MegaBace)

Sequening was performed separately for forward and reverse primers (both strands were sequenced), following the DYEnamic ET Terminator Cycle Sequencing Kit protocol (Amersham Biosciences, Buckinghamshire, UK):

1) Mix 2 μl of primer (forward OR reverse) with 3 μl distilled water and 4 μl sequencing mix.
2) Add 1 μl purified PCR product.
3) Centrifuge at low speed.

Sequencing reaction:

1) 26 amplification cycles:
   Denaturation 95 °C, 20 seconds
   Annealing 50 °C, 15 seconds
   Extension 60 °C, 60 seconds
2) Hold at 4 °C.
3) Store at –20°C.

PROTOCOL B4. Sequencing clean-up (for MegaBace)

Prior to the run on a MegaBace sequencer the sequencing product has to be cleaned. The protocol refers to a 96-well plate. Volumes to add are for a single well, unless stated otherwise.

1) In a 1.5 ml tube, prepare a solution containing 1 ml of distilled water and 200 μl of Sodium-acetate/EDTA (1/10 Vol.).
2) Add 12 μl of the solution to 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) Discard the supernatant.
7) Short inverted spin for ~30 seconds at 300 rpm.
8) Rinse with 150 µl of 70% ethanol.
9) Centrifuge 10 minutes at 3000 rpm.
10) Discard the supernatant.
11) Let the ethanol evaporate by leaving the plate at room temperature for 5–15 minutes.
12) The plate can be stored at –20 °C before being run on the sequencer.
13) Before the add 15 µl of distilled water.
14) Vortex to elute the DNA pellet.
15) Centrifuge briefly.

PROTOCOL B5. Sequening reactions (for ABI)

Sequening was performed separately for forward and reverse primers (both strands were sequenced), using the ABI BigDye Terminator v1.1 sequencing kit (Applied Biosystems, Buckinghamshire, UK):

1) Mix 1 µl 5x buffer with 2 µl of sequencing mix and 3 µl of destilled water.
2) Add 2 µl of forward OR reverse primer.
3) Add 2 µl of PCR product.

Sequencing reaction:

1) Denaturation 95°C, 60 seconds
2) 40 amplification cycles:
   Denaturation 95 °C, 10 seconds
   Annealing 50 °C, 15 seconds
   Extension 60 °C, 4 minutes
3) Hold at 4 °C.

4) Add 15 μl of destilled water prior to sequening or store at –20°C.

PROTOCOL B6. RNA isolation

This protocol is a modification of the Invitrogen Life Technologies Trizol protocol. (Invitrogen, Carlsbad, USA). It is available at the Drosophila Genomics Resource Center website (http://dgrc.cgb.indiana.edu/microarrays/downloads.html).

1) To 50 mg snap 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.
2) Incubate at room temperature for 5 minutes.
3) Centrifuge at 12,000 rpm for 10 minutes at 4 °C to pellet insoluble debris such as exoskeleton.
4) Transfer the supernatant to a new microcentrifuge tube, taking great care not to take pellet or fat layer.
5) Add 200 μl of Chloroform to each tube.
6) Shake vigorously by hand (do not vortex).
7) Incubate tubes at room temperature for 3 minutes.
8) Centrifuge at 10,000 rpm for 15 minutes at 4 °C.
9) Transfer upper aqueous phase (~0.6 ml) to a fresh RNAse-free microcentrifuge tube.
10) Add 0.5 ml isopropanol
11) Incubate at room temperature for 10 minutes.
12) Centrifuge at 12,000 rpm for 10 minutes at 4 °C.
13) Remove the supernatant and wash the pellet with 1 ml 75% ethanol.
14) Centrifuge at 7,500 rpm for 5 minutes at 4 °C.
15) Remove the supernatant.
16) Centrifuge briefly and carefully remove the last of the supernatant with a micropipette.

17) Air dry for 10 minutes.
18) Resuspend the pellet in 100 µl RNase-free water.
19) Quantify a 1/100 dilution of the RNA on spectrophotometer.
20) Store at -20 °C.

PROTOCOL B7. cDNA synthesis

First-strand cDNA synthesis was done following the ThermoScript RT protocol (Invitrogen, Carlsbad, USA).

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

Random primers	1 µl
10 pg to 5 µg of total RNA	x µl,
10 mM dNTP mix	2 µl
Sterile, 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
0.1 M DTT	1 µl
Sterile, distilled water	1 µl
ThermoScript RT (15 U/ml)	1 µl

3) Incubate tube at 25 °C for 10 minutes.
4) Mix contents of the tube 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.

PROTOCOL B8. qRT-PCR

For realtive quantification of gene expression the protocol TaqMan gene expression assay protocol for quantification on the Applied Biosystems 7500 Fast Real-Time PCR system was used (Applied Biosystems, Foster City, USA):

Mix 1 µl of TaqMan gene expression assay (20X) with 9 µl of cDNA template (~50 µg) and 10 µl of TaqMan Universal Master Mix.

qRT-PCR reaction:

40 amplification cycles:

Denaturation	95°C	15 seconds
Annealing/Extension	60 °C	60 seconds

Data were collected at the end of each amplification cycle

Curriculum vitae

Name:	Oliver Steffen Beißwanger	
Date and place of birth:	20. August 1975, Stuttgart	
Nationality:	German	
Family status:	unmarried	
Address:	Fraunhoferstr. 9 RGB 80469 München	
Education:	2003-2006	PhD student, LMU München
	1996-2003	LMU München, Diploma in Biology
	1986-1995	Karl-Ritter-von-Frisch-Gymnasium Moosburg, Abitur
Professional life:	2003-2006	Research assistant, LMU München
	1996-2003	Associate in the department 'inpatient service', Klinikum Freising GmbH
	1995	Associate in the department 'motor liability insurance', R+V Allgemeine Versicherung AG, München

Publications

BEISSWANGER, S., W. STEPHAN and D. DE LORENZO, 2006 Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. Genetics **172**: 265- 274.

WERNER S. H., U. I. WALTHER, K. VOGL, S. BEISSWANGER, C. MAYR und S. C. WALTHER, 2002 Vergleichende Bestimmung des zellulären Zinkgehaltes mittels ICP nach Druckaufschluß mit einer luminometrischen Methode mittels Zincon ohne Druckaufschluß. In: Anke, M., R. Müller, U. Schäfer, und M. Stoepler (eds.): *Macro and Trace Elements*, Schubert-Verlag Leipzig, pp. 13-18.

Conferences and Posters

BEISSWANGER, S. Population structure and selection in *Drosophila melanogaster* samples from Africa, Europe and SE Asia. Annual meeting for the Society for Molecular Biology and Evolution: Genomes, Evolution & Bioinformatics 2006, 24 – 28 May 2006, Tempe, USA. *(oral 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, Krakow, Poland.

BEISSWANGER, S., W. STEPHAN and D. DE LORENZO. Molecular evidence for a selective sweep in the *wapl*-region of *Drosophila melanogaster*. Symposium on Mechanisms of adaptation, genetic differentiation and speciation, March 2004, Tutzing. *(poster presentation)*.

Acknowledgements

Ich möchte mich zunächst bei Herrn Prof. Stephan dafür bedanken, dass er es mir ermöglicht hat meine bereits in der Diplomarbeit begonnene wissenschaftliche Arbeit fortzusetzten. Mit seiner Hilfe habe ich tiefere Einblicke in die Populationsgenetik, insbesondere zu 'selective sweeps' erhalten. Es war für mich auch sehr interessant an der Populationsstruktur von Drosophila melanogaster zu arbeiten.

Ein herzliches Dankeschön geht an David De Lorenzo, den ich schon seit meiner Zeit als HiWi kenne. Er hatte immer nützliche Ratschläge und war insbesondere beim Anfertigen meines ersten Manuskriptes von großer Hilfe. Natürlich danke ich auch der gesamten Drosophila Gruppe, besonders Sascha Glinka und Lino Ometto. Wir hatten viele ergiebige Gespräche und Diskussionen über populationsgenetische Aspekte bei Drosophila.

Vielen Dank an Herrn Prof. John Parsch und seine Gruppe für all die Hilfe in den letzten Jahren. Besonders die beiden Johns hatten immer gute Ratschläge zu experimentellen Arbeiten. Ich danke auch Laura Rose für ihre Hilfe bzgl. Phylogenetik und ihre Bereitschaft meine Arbeit zu korrigieren. Besonderer Dank gilt der Gruppe von Joachim Hermisson. Mit Pleuni Pennigs hatte ich viele Diskussionen über 'selective sweeps'. Mit Nina Stoletzki und Katrin Kümpfbeck teilte ich viele Kaffeepausen, bei denen wir interessante Gespräche über Evolution und das Leben an sich hatten.

Ein herzliches Dankeschön geht auch an Aparup Das für seine Hilfe bei der Arbeit mit Drosophila und an Haipeng Li und Claus Vogl für ihre Hilfe bei der Datenanalyse. Mein besonderer Danke gilt der technischen Assistenz, insbesondere Traudl Feldmaier-Fuchs und Anne Wilken. Vielen Dank für das Fliegenfutter und die unermüdliche Hilfe im Labor! Ohne sie wäre ich wohl noch im Labor und würde pipettieren.

Nicht zuletzt möchte ich mich bei Stefan Donhauser und meiner Familie für jegliche Unterstützung in all den Monaten bedanken!