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

Natural variation in *Drosophila melanogaster*:

A survey of genome-wide DNA sequence polymorphism and gene expression diversity, and the development of new bioinformatic tools

Stephan Hutter

aus
Salzburg, Österreich

2007

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 nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen habe.

EHRENWÖRTLICHE VERSICHERUNG

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

München, 31.05.2007

Stephan Hutter

1. Gutachter: Prof. Dr. Wolfgang Stephan

2. Gutachter: Prof. Dr. John Parsch

Dissertation eingereicht am: 31.05.2007

Mündliche Prüfung am: 11.09.2007

Note

In this dissertation I present the work of my doctoral research from June 2003 until May 2007. It is organized in three chapters. All of them are the result of collaboration with numerous other scientists.

For the work presented in CHAPTER 1 Steffen Beisswanger and I generated the data. Haipeng Li and I analyzed the data and Haipeng Li developed the theoretical methods. The writing was done by Haipeng Li and myself and revised by Wolfgang Stephan. The study was carried out under the supervision of David de Lorenzo and Wolfgang Stephan. This chapter has been published under the following title:

HUTTER, S., H. LI, S. BEISSWANGER, D. DE LORENZO and W. STEPHAN, 2007 Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosome-wide SNP data. Genetics doi: 10.1534/genetics.107.074922.

For CHAPTER 2 the data was generated by Sarah Saminadin-Peter and me. I did the analysis and the writing. The manuscript was revised by John Parsch and Wolfgang Stephan. The study was supervised by John Parsch. A paper based on the findings described in this chapter has been submitted for publication under the title:

HUTTER, S., S. S. SAMINADIN-PETER, W. STEPHAN and J. PARSCH, 2007 Gene expression variation in African and European populations of *Drosophila melanogaster*. Genome Biology (submitted)

For CHAPTER 3 Albert Vilella and I developed the bioinformatic tools and wrote the online documentation. Albert Vilella and Julio Rozas analyzed the data. Julio Rozas wrote the manuscript and supervised the study. This chapter has been published under the title:

HUTTER, S., A. J. VILELLA and J. ROZAS, 2006 Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics **7:** 409.

Contents

Summary

This work is dedicated to studying natural variation in *D. melanogaster* at the DNA sequence and gene expression level. In addition I present a new version of the DNA polymorphism analysis program VariScan, which includes significant improvements.

In CHAPTER 1 I describe a genome scan of single nucleotide polymorphism in two natural *D. melanogaster* populations (from Africa and Europe) on the third chromosome. Together with polymorphism data previously published for the X chromosome of the same populations, this allows a comparative study of the polymorphism patterns of the X chromosome and an autosome. The frequency spectrum of mutations and the patterns of linkage disequilibrium are investigated. The observed patterns indicate that there is a significant difference in the behavior of the two chromosomes, as has already been suggested by previous studies. To uncover the reasons for this a coalescent based maximum likelihood method is applied that incorporates the effects of demographic history and unequal sex ratios. For the African population the differential behavior of the chromosomes can be explained by its demographic history and an excess of females. In Europe, a population bottleneck and an excess of males alone cannot explain the patterns we observe. The additional action of positive selection in this population is proposed as a possible explanation.

In CHAPTER 2 I investigate the variation in gene expression of the two aforementioned populations. Whole-genome microarrays are used to study levels of expression for 88% of all known genes in eight adult males from both populations. The observed levels of expression variation are equal in Africa and Europe, despite the fact that DNA sequence variation is much higher in Africa. This is evidence for the action of stabilizing selection governing levels of expression polymorphism. Supporting this view, genes involved in many different functions, and are therefore on strong selective constraint, show less variation than do genes with only few functions. The experimental design allows the search for genes which differ in their expression patterns between Europe and Africa and might therefore have undergone adaptive evolution. Detected candidates include genes putatively involved in insecticide resistance and food choice. Surprisingly, many genes over-expressed in Africa are involved in the formation and function of the flying apparatus.

In CHAPTER 3 I present version 2 of the program VariScan. This program was designed to analyse patterns of DNA sequence polymorphism on a chromosomal scale. The functionality of the core analysis tool, the wavelet decomposition, is described. In addition, multiple improvements to the previous version are presented. The program now supports the "pairwise deletion" option. This is essential for analysing data at the chromosome scale, since such data often contains incomplete information. It is now possible to add outgroup information, which allows the calculation of additional statistics. Furthermore, the separate analysis of different predefined chromosomal regions is added as an option. To increase the user friendliness, a graphical user interface is now included as part of the software package. Finally, VariScan is applied to published and computer-generated data and the ability of the wavelet-based analysis to uncover chromosomal regions with interesting DNA polymorphism patterns is demonstrated.

General Introduction

"The many slight differences which appear in the offspring from the same parents, or which it may be presumed have thus arisen, from being observed in the individuals of the same species inhabiting the same confined locality, may be called individual differences. No one supposes that all the individuals of the same species are cast in the same actual mould. These individual differences are of the highest importance for us, for they are often inherited, as must be familiar to every one; and they thus afford materials for natural selection to act on and accumulate, in the same manner as man accumulates in any given direction individual differences in his domesticated productions. These individual differences generally affect what naturalists consider unimportant parts; but I could show, by a long catalogue of facts, that parts which must be called important, whether viewed under a physiological or classificatory point of view, sometimes vary in the individuals of the same species. I am convinced that the most experienced naturalist would be surprised at the number of the cases of variability, even in important parts of structure, which he could collect on good authority, as I have collected, during a course of years."

CHARLES DARWIN (1872)

In the aforementioned citation from the 6th edition of "*The origin of the species: by means of natural selection*" Darwin describes one of the basic requirements for evolution to occur in the first place: Natural variation. For Darwin, this variation was the substrate on which natural selection acted. Those small differences between individuals were chosen by natural selection that provided the highest fitness in the local environment. This process would then lead to adaptive evolution of populations, which ultimately led to the formation of new species.

The desire to understand the exact dynamics which take place in such variable populations lead to the creation of a whole new scientific field within evolutionary biology: population genetics. One of the main focuses of this science was to investigate how new variants (or mutations) arose in a population and subsequently changed in frequency. Mathematical models were developed that described how these mutations were passed on from generation to generation. In the 1960s a theory was proposed that had a significant impact on the field of population genetics: the neutral theory of molecular evolution (KIMURA 1968). This theory suggests that natural selection is not necessary to explain a vast amount of differences observable between populations or species. The underlying assumption was that the majority of arising mutations is deleterious and will be quickly purged from the population. Variation that remains in the population is thought to have no effect on the fitness of individuals and therefore behave neutrally. These mutations will then be passed on from generation to generation. This is a random process, and in populations with finite population sizes this inevitably leads to the random loss or fixation of mutations. This effect is called genetic drift. Since this random process will fix different mutations in different populations, these populations will differentiate without the aid of natural selection. Furthermore, the variation observable within a population is only governed by its effective population size (N_e) and the rate at which neutral mutations occur (μ).

In contrast to widespread belief, the neutral theory does not reject the possibility of adaptation through positive Darwinian selection. It simply states that the role of genetic drift is a dominant one and that positive selection contributes only little to the differentiation of populations. Up to this day there is a heated debate on this topic which revolves around one central question: What are the frequencies of neutral and non-neutral mutations and what are their relative contributions to the differentiation of populations and the formation of new species? Since the neutral theory of evolution depends on only few parameters it can be modeled relatively easily. These mathematical models can make predictions on how natural variation should behave under neutral conditions. Nowadays, the neutral model of evolution is used as a null hypothesis in countless numbers of studies. In particular, surveys of variation at the DNA level studying single nucleotide polymorphisms (SNPs) rely on coalescent-based neutral models (KINGMAN 1982). With the help of coalescent theory, neutral expectations for SNP-based statistics describing the frequency spectrum of mutations or the association of alleles can be obtained. Deviations from these expectations are often considered as evidence for the action of natural selection.

In the early 1990s a study of BEGUN and AQUADRO (1992) added new insight to the question of validity of the neutral model. The authors found that levels of DNA polymorphism were positively correlated with the recombination rate in the fruit fly *Drosophila melanogaster*. This is only expected under the neutral model if the mutation rate, μ, is larger in regions of higher recombination rate. This possibility was rejected, since this would also imply that regions with higher recombination rates would show a higher level of divergence to the sister species *D. simulans*. Such a pattern was not observed. The authors therefore suggest that genetic hitchhiking caused by positive selection (MAYNARD SMITH and HAIGH 1974) led to the observed correlation. In this scenario, a positively selected mutation arises in the population, rapidly increases in frequency and eventually becomes fixed. Since neutral variants on the same chromosome as the selected mutation are physically linked, they "hitchhike" along to fixation. This results in a chromosomal region depleted of variation, also known as a "selective sweep". The size of this region depends on the local recombination rate. Recombination breaks up the association of neutral variants with the selected mutation and therefore allows polymorphism to be retained. As a result, regions of high recombination show higher levels of variation than do regions of low recombination after being affected by a selective sweep. Since the study of BEGUN and AQUADRO (1992) shows a significant correlation between DNA polymorphism and the recombination rate, the authors conclude that positive selection occurs frequently.

An alternative explanation was put forward by CHARLESWORTH *et al.* (1993). In their model, a mutation that is deleterious arises in the population. Since this mutation decreases the fitness of the individuals carrying it, it will be removed by natural selection. However, since neutral variants residing on the same chromosome as the negatively selected allele are also removed from the population, this form of selection, called "background selection", has the ability to reduce neutral variation. This effect will also depend on the recombination rate. If the recombination rate is high, the neutral polymorphism has an increased opportunity to recombine away from the chromosome carrying the deleterious allele.

Finding ways to distinguish between the effects of hitchhiking and background selection has been the focus of many theoretical studies (*e.g.*, KIM and STEPHAN 2000, INNAN and STEPHAN 2003). A pattern that is unique to genetic hitchhiking is the creation of distinct valleys of reduced neutral polymorphism in regions of medium and high recombination. Since variants close to the selected site have less opportunity to break up the physical linkage by recombination than those with larger distances, the reduction of variation cased by hitchhiking will be more extreme in regions close to the selected locus. This negative correlation between the physical distance and the reduction of variation will create a distinct valley, with the lowest amount of variation just around the selected mutation. Theoretical studies have shown that such valleys caused by hitchhiking should be detectable at the DNA level in regions of medium to high recombination (KIM and STEPHAN 2002).

This finding led to the idea of so called genome scans. In such a scan, neutral polymorphism of multiple loci distributed along a chromosome is measured within a population. Those loci that show a reduction in variation relative to the chromosomal average and do not have reduced mutation rate (as inferred from divergence to an outgroup species) are then putative candidates for being affected by a hitchhiking event. The polymorphism pattern of such candidate regions is then inspected in more detail, and tests are applied to confirm that positive selection truly acted in this region (KIM and STEPHAN 2002).

One of the first studies applying such a genome scan approach was carried out by GLINKA *et al.* (2003). Here, the authors sequenced 125 X chromosomal loci in two populations of *D. melanogaster*. This species has long been a model organism in population genetics. Additionally, its genome sequence is known and well annotated, which facilitates the choice of neutral markers. The studied loci were located in introns and intergenic regions, since these sequences are thought to evolve neutrally. The two populations studied came from Zimbabwe and the Netherlands. They were chosen because of the biogeographic history of *D. melanogatser*. The species originated in sub-Saharan Africa and expanded its range into Europe after the last glaciation ~15,000 years ago (DAVID and CAPY 1988, LACHAISE *et al.* 1988). Since the conquest of the temperate European continent by a tropical species was most likely only possible by adaptation to the new environment, the search for positive selection was focused on the Dutch population. The putatively ancestral African population served as a control to which the derived population could be compared.

This study yielded some interesting results. The chromosome-wide patterns of polymorphism for the African and the European population were not in accordance with the standard neutral model. The deviations from neutrality in both populations were also highly unlikely to be the result of positive selection alone. The reason for this observed disparity is the demographic history of both populations. Events such as population size expansions or population bottlenecks in the recent history of a population can create polymorphism patterns that deviate drastically from the standard neutral model, which assumes a constant population size. The standard neutral model is therefore an inappropriate null hypothesis for populations with fluctuating population size. The solution to this problem is the inclusion of demographic history into the null hypothesis of tests searching for positive selection. Such a test has been developed recently and applied to an expanded dataset of the aforementioned study, resulting in the detection of multiple regions where positive selection may have occurred (OMETTO *et al.* 2005, LI and STEPHAN 2006).

Thus far, genome scans in *D. melanogaster* populations have focused on the X chromosome, because natural selection is thought to be easier to detect than on autosomes (CHARLESWORTH *et al.* 1987). This leaves us virtually ignorant about the patterns of neutral SNP polymorphism on non-sex chromosomes in natural populations. ANDOLFATTO (2001) compiled variation estimates of X-linked and autosomal genes from multiple different studies in order to compare the relative amount of polymorphism in African and non-African *D. melanogaster*. The drawback of this approach is that these estimates do not come from true population samples. That caveat aside, the main finding was that X-linked and autosomal variation do not behave as expected under the standard neutral model. Since males carry only one copy of the X-chromosome, the population size of these sex chromosomes should be 3/4 of that of autosomes. Since variation under the standard neutral model depends only on μ and N_e, neutral variation on the X chromosome should also be 3/4 of the autosomal variation. This of course only holds if mutation rates are equal on both chromosomes, but this seems to be the case (BETANCOURT *et al.* 2002). However, ANDOLFATTO (2001) finds that the ratio of X-linked to autosomal variation is much larger than 3/4 in African, and much lower than 3/4 in non-African flies. A study of microsatellite variation using true population samples of *D.*

melanogaster confirms these patterns (KAUER *et al.* 2002). However, the handicap of microsatellite studies is that microsatellites have vastly different locus-specific mutation rates, and comparisons may therefore be biased. Nevertheless, biologists have tried to explain the observed deviations from neutral expectations. Possible explanations include the differential impact of positive selection on the two chromosome types and unequal sex ratios leading to deviations from the expected ratio of 3/4 for the chromosomal population sizes (CHARLESWORTH 2001).

In CHAPTER 1 I present the first SNP based genome scan of *D. melanogaster* on an autosome. Since we use the same populations studied by GLINKA *et al.* (2003), we can obtain estimates of the ratio of X-linked to autosomal variation using true appropriate population samples. The deviation from the neutral expectation reported in earlier studies is confirmed. In order to find the reason for these disparities, sophisticated coalescent-based models are applied, which include the effects of demography and unequal sex ratios. This allows us to test if these two forces are enough to explain the patterns of polymorphism we see in the African and European populations. In addition, this autosomal genome scan is the starting point in the search for positive selection other than on the X chromosome, and tests such as the one developed by LI and STEPHAN (2006, see above) are currently being applied to the new data. The future results of this search will help us understand the effect of selection on sex chromosomes and autosomes and should be useful in gaining further knowledge about the differences in the evolution of these two types of chromosomes.

As soon as candidate regions for adaptive evolution are found by the aforementioned genome scan, the question of the nature of the positively selected mutation arises. In general, two types of mutations are thought to have the potential to create positively selected alleles. First, mutations that modify the amino acid sequence of a protein may be targets of selection. Such changes can alter the structure of the protein, which might then influence the efficiency of the protein function or even lead to the acquisition of totally new functions. Second, regulatory mutations that change the expression level or pattern of a gene have long been suspected to significantly contribute to evolutionary change. The quantitative levels of a gene product can have a major influence on the phenotype of an organism, and it has been hypothesized that such regulatory changes may even play a bigger role in evolution than do changes at the amino acid level (KING and WILSON 1975).

Finding these positively selected mutations in candidate regions is a challenging task. Such regions contain often dozens of genes which all might have been the target of selection. The putatively selected amino acid changing variants can be found by sequencing the coding regions of these genes in multiple populations and looking for fixed non-synonymous mutations in the population containing the sweep. In the case of regulatory changes, mutations altering the expression level of a gene usually occur in *cis*-regulatory elements located in the flanking regions of the gene or the first intron. In addition to comparing the sequences of these elements in different populations, an alternative approach is to compare levels of gene expression directly. Genes that show distinctly different expression levels between populations and that are monomorphic in the population where selection is thought to have taken place are good candidates for being targets of adaptive evolution.

Microarray technology allows us to search for such differentially expressed genes on a genome wide scale. Two-channel microarrays can detect differences in the levels of mRNA abundance between individuals and are therefore a valuable tool for detecting differentially expressed genes. In CHAPTER 2, a large-scale microarray study is presented in which the genome-wide expression patterns of the European and African populations studied in CHAPTER 1 are analyzed. Our approach allows us to detect genes that have distinctly different expression patterns between the two populations and are therefore candidates for adaptation. Furthermore, it should be able to find positively selected regulatory changes which might have been found by genome scans at the DNA polymorphism level. If the selective event occurred too far in the past, the signatures created by hitchhiking may be no longer detectable. In addition, such signatures can often be obscured by demographic events. These problems should not occur in our microarray study, since we are not relying on indirect signatures of selection. We are directly investigating the trait under selection, the gene expression phenotype. Another interesting possibility of this study is the investigation of variation at the gene expression level within populations. Adding such data to what we already know about DNA polymorphism (see CHAPTER 1) should help us expand our understanding of how natural variation is shaped in *D. melanogaster* populations.

As mentioned above, patterns of DNA nucleotide polymorphism can be powerful tools in describing phenomena such as positive selection or the demographic history of populations. Over the last two decades many statistics and tests have been developed that make use of SNP data. Well known examples are Tajima's *D* (TAJIMA 1989), which investigates the frequency spectrum of mutations or statistics describing the association of alleles like r^2 (HILL and ROBERTSON 1968). For over a decade, DnaSP (ROZAS and ROZAS 1995) has been the *de facto* standard program for calculating such statistics from DNA sequence alignment data. It provides an easy to use interface, and is also able to perform sophisticated types of analyses, such as sliding window methods.

Today, the widespread availability of technology such as high throughput capillary sequencers or SNP detection microarrays allows the generation of nucleotide polymorphism data at very large scales. Genotyping projects have created SNP datasets for whole genomes, typing dozens or even hundreds of individuals. Examples include the International HapMap Project in humans (http://www.hapmap.org), the Mouse Genome Resequencing Project (http://mouse.perlegen.com/mouse) or the 50 Genomes project of *D. melanogaster* (http://www.dpgp.org). Such large datasets provide the opportunity for new and innovative approaches of analysing the data, but also require new and sophisticated bioinformatic tools.

With the program VariScan (VILELLA *et al.* 2005) we introduced a new form of analysis of SNP data on a chromosomal scale. The core idea is the decomposition of the chromosome wide signal of statistics describing the polymorphism pattern by means of wavelet transformation. In CHAPTER 3 a description of the functionality of this approach is provided. Version 2 of VariScan, which contains many new features, is also presented. The program now supports the inclusion of outgroup sequences, which allows the calculation of powerful statistics such as Fay and Wu's *H* (FAY and WU 2000) and others. In the aforementioned large-scale datasets, information about the genotype is not always available for all studied individuals at each position of the chromosome. Under the conservative "complete deletion" option, such sites would be removed from the analysis, resulting in the loss of large amounts of data. VariScan 2 is able to overcome these limitations by introducing the "pairwise deletion" option. Furthermore, the user is now able to predefine specific regions of the chromosome to be analyzed, hence allowing the separate analysis of genomic regions with different functional properties (*e.g.*, coding *versus* non-coding regions, *etc.*). The user friendliness of the program is increased by the inclusion of a graphical user interface. Finally, we apply VariScan to a published and computer-simulated dataset and demonstrate the power of the wavelet-based analysis to detect local reductions of polymorphism, similar to those created by a selective sweep.

LITERATURE CITED

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

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

BETANCOURT, A. J., D. C. PRESGRAVES and W. J. SWANSON, 2002 A test for faster X evolution in *Drosophila*. Mol Biol Evol **19:** 1816-1819.

CHARLESWORTH, B., 2001 The effect of life-history and mode of inheritance on neutral genetic variability. Genet Res **77:** 153-166.

CHARLESWORTH, B., J. A. COYNE and N. H. BARTON, 1987 The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. **130:** 113-146.

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

DARWIN, C., 1872 *The origin of species: by means of natural selection*. 6th Edition, Murray, London, UK.

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

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

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.

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

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

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.

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

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

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

KING, M. C., and A. C. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science **188:** 107-116.

KINGMAN, J.F.C., 1982 On the genealogy of large populations. J Appl Prob **19A:** 27-43.

LACHAISE, D., M. L. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS and M. ASHBURNER, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol Biol **22:** 159-225.

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

ROZAS, J., and R. ROZAS, 1995 DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. Comput Appl Biosci **11:** 621-625.

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

VILELLA, A. J., A. BLANCO-GARCIA, S. HUTTER and J. ROZAS, 2005 VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics **21:** 2791-2793.

1. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosome-wide SNP data

ABSTRACT

It has been hypothesized that the ratio of X-linked to autosomal sequence diversity is influenced by unequal sex ratios in *D. melanogaster* populations. We conducted a genome scan of single nucleotide polymorphism (SNP) diversity of 378 autosomal loci in a derived European population and of a subset of 53 loci in an ancestral African population. Based on these data and our already available X-linked data, we used a coalescent-based maximum likelihood method to estimate sex ratios and demographic histories simultaneously for both populations. We confirm our previous findings that the African population experienced a population size expansion while the European population suffered a population size bottleneck. Our analysis also indicates that the female population size in Africa is larger than or equal to the male population size. In contrast, the European population shows a huge excess of males. This unequal sex ratio and the bottleneck alone, however, cannot account for the overly strong decrease of X-linked diversity in the European population (compared to the reduction on the autosome). The patterns of the frequency spectrum and the levels of linkage disequilibrium observed in Europe suggest that, in addition, positive selection must have acted in the derived population.

INTRODUCTION

In recent years genomic scans of DNA sequence variation have been performed for multiple species. These studies became possible by the availability of full genome sequences, and data is now available from a variety of organisms such as humans (AKEY *et al.* 2004), *Arabidopsis* (SCHMID *et al.* 2005) and *Drosophila* (GLINKA *et al.* 2003, ORENGO and AGUADÉ 2004). These data sets provide useful tools to address questions such as estimating population sizes and demographic histories of a species. Another major focus of these investigations was to find footprints of positive selection (selective sweeps). Positive (directional) selection is predicted to locally reduce DNA polymorphism around a selected site (MAYNARD-SMITH and HAIGH 1974). This effect is expected to be stronger and more abundant on the X chromosome than on autosomes (CHARLESWORTH *et al.* 1987).

Most studies, especially in *Drosophila melanogaster*, have concentrated on the sex chromosome (GLINKA *et al.* 2003, ORENGO and AGUADÉ 2004, HARR *et al.* 2002). A conclusion of these and other works was that demographic events such as population size expansions and bottlenecks are major factors in shaping the polymorphism pattern of *D. melanogaster* (*e.g.*, GLINKA *et al.* 2003, ANDOLFATTO 2001, HADDRILL *et al.* 2005b). This makes it difficult to detect signatures of positive selection and quantify their contribution to the pattern of genome-wide diversity. Recent studies have tried to overcome this difficulty by using models that take demographic history into account (OMETTO *et al.* 2005, LI and STEPHAN 2006).

Another approach for estimating effects of selection was proposed by AQUADRO *et al.* (1994). Since the diversity-reducing effect of positive selection is larger on the X chromosome than the autosome, populations that are subject to positive selection should exhibit lower ratios of X-linked to autosomal diversity than the expected 3/4 under standard neutral conditions. Theoretical work shows that this expectation holds if a substantial amount of positive mutations is recessive (BETANCOURT *et al.* 2004). On the other hand, if there is background selection acting in the population the ratio is predicted to be higher than 3/4 (CHARLESWORTH *et al.* 1993). A comparison of X-linked and autosomal polymorphism might therefore be helpful in identifying the prevalent mode of selection within a population.

So far, data quantifying the amount of autosomal variation based on SNPs is rather scarce for *D. melanogaster*. ANDOLFATTO (2001) used a data set compiled from different previous studies to compare levels of X-linked and autosomal diversity in African and non-African flies. *D. melanogaster* is thought to have originated in sub-Saharan Africa and have only relatively recently (10,000-15,000 years ago) colonized the rest of the world (DAVID and CAPY 1988, LACHAISE *et al.* 1988). It could therefore be informative to compare patterns between putatively ancestral and derived populations. The drawback of this study, however, is that it combines samples from many different publications. The data set is therefore not a representative population sample. In addition, the loci are mostly genes that were chosen in the original studies because of their unusual patterns of polymorphism. Thus, they may not represent an unbiased set of loci as those used in genome scans (see above). That being said, the comparison of X-linked and autosomal diversity showed a clear pattern: the X chromosome exhibited more variation than expected under standard neutrality in Africa and too few polymorphisms outside Africa. While it is tempting to attribute this to different modes of selection acting in these environments, other forces can also lead to such an observation. There is evidence that effective population sizes for males and females are not equal in natural populations of *D. melanogaster* (reviewed in CHARLESWORTH 2001). If this is the case then also the ratio of X-linked to autosomal variation will not correspond to the standard expectation. Since females carry 2/3 of the X chromosomes in a population but only 1/2 of the autosomes changes in female population size will affect X chromosomes more strongly than autosomes. In other words, if the female population size is larger than the male one, the X-chromosomal to autosomal ratio of diversity is expected to be larger than 3/4, while lower female population size leads to ratios below 3/4. Only recently a study surveying SNP polymorphism in an unbiased set of coding regions was published (SHAPIRO *et al.* 2007). Yet this study does not address the relationship between autosomal and X-linked diversity.

KAUER *et al.* (2002) performed a genome scan of variability of 133 microsatellite loci in African and non-African populations on both the X chromosome and autosomes. Comparing heterozygosities of microsatellites can be problematic, because each locus has a specific mutation rate. So there is a risk of having mutational biases. The data was again from multiple populations in and outside Africa, but the authors corrected for possible biases by forming across-population averages. The results of this study confirm the findings of ANDOLFATTO (2001). The authors conclude that background selection shaped polymorphism in the ancestral African populations while positive selection was prevalent outside Africa. However, unequal sex ratios could also have contributed to the patterns (KAUER *et al.* 2002). Both the SNP and the microsatellite study did not account for the demographic history of Drosophila. Events such as population size expansions and bottlenecks might have different effects on X chromosomes and autosomes further complicating a direct comparison of both chromosomes.

We present here the first genome scan of an autosome in *D. melanogaster* using non-coding SNP markers. We surveyed a total of 378 loci located on chromosome 3 in a European population. In addition, we analyzed a random subset of 53 loci in an African population to get an estimate of autosomal diversity for an ancestral population. The individuals analyzed come from the same populations that have already been used in previous scans of X-linked diversity (GLINKA *et al.* 2003, OMETTO *et al.* 2005). By combining the data sets we now have the opportunity to study X chromosomes and autosomes within a single ancestral and a derived population. To make use of the additional information that SNP data can provide, we also analyzed statistics describing the frequency spectrum of mutations as well as linkage disequilibrium (LD). To address the possibility of unequal sex ratios and demography shaping polymorphism, we extended the likelihood method described in LI and STEPHAN (2006). This approach allows us to estimate the most likely sex ratios in both populations while taking their demographic history into account.

MATERIALS AND METHODS

Biological material: The individuals come from a total of 24 highly inbred lines sampled from an African population from Zimbabwe and a European population from The Netherlands (GLINKA *et al.* 2003).

Chromosomal inversions: For inversion analysis we individually crossed male flies to virgin Canton-S females, homozygous for the standard chromosome arrangement. We prepared salivary glands from late F_1 third-instar larvae, maintained at 18°C. Several larvae per *D. melanogaster* line were dissected to account for low-frequency inversions. Polytene chromosomes were prepared using the lacto-acetic orcein method and viewed under an inverted phase contrast microscope. Banding patterns and inversion breakpoints were identified according to standard chromosome maps (LEFEVRE 1976).

Data collection: Primers for 378 loci (125 on chromosome 3L, 253 on chromosome 3R) were designed based on the *D. melanogaster* genome Release 3.2 (http://www.flybase.org). Loci were chosen evenly spaced along the chromosome and located in intronic or intergenic regions. All 378 genomic regions were sequenced in the European sample, while only a random subset of 53 fragments (17 on 3L, 36 on 3R) was sequenced in the African one.

Sequence data was obtained by means of capillary sequencing on both strands. As the *D. melanogaster* genome Release 4.2.1 became available during the course of our study, we re-checked all our loci for overlap with coding regions. One locus (namely 3-480) overlapped partially with a coding exon. The overlapping part was removed from the alignment and only the non-coding sequence was used for analysis. The new sequences were deposited into the European Molecular Biology Laboratory (EMBL) database under accession numbers AM701830 to AM706343.

To obtain an outgroup sequence for each of our fragments homologous sequences were searched in the *D. simulans* genome, Mosaic Assembly Release 1.0 (Genome Sequencing Center, WUSTL School of Medicine) via BLAST. Only for four fragments a homolog could not be found, resulting in a total of 374 alignments in the European and 51 alignments in the African population that contain outgroup information.

Based on the annotation of the *D. melanogaster* genome Release 4.2.1, we updated our published data on the X-linked non-coding loci (OMETTO *et al.* 2005). Loci that showed complete or almost complete overlap with coding regions were removed from the data set. Where only a partial overlap existed, the coding regions were removed from the alignments. The final X chromosomal data set consisted of 259 loci for the European and 249 loci for the African population.

Statistical analysis: We estimated basic population genetic statistics, such as levels of nucleotide diversity per site measured as average pair wise distance π (TAJIMA 1983) and Watterson's estimator θ (WATTERSON 1975) as well as Tajima's *D* statistic (TAJIMA 1989). We calculated the *P* value of Tajima's *D* based on coalescent simulations assuming the standard neutral model (10,000 iterations). SCHAEFFER (2002) noted that it is difficult to compare values of Tajima's *D* between different loci, if they show differences in sample size and/or number of segregating sites. The author therefore suggests using the ratio of Tajima's D to D_{min} to overcome this limitation, where D_{min} is the absolute value of the theoretical minimum of *D*. Thus, the D/D_{min} statistic and *D* have the same sign. Since the loci we are dealing with here do indeed vary greatly in sample size and number of segregating sites, we used the D/D_{min} statistic (SCHAEFFER 2002) to summarize the frequency spectrum when building averages over multiple loci or comparing loci among each other. The LD statistic Z_{nS} (KELLY 1997) and divergence (*K*) corrected for multiple hits (JUKES and CANTOR 1969) to *D. simulans* were estimated by the program VariScan (HUTTER *et al.* 2006). Since Z_{nS} values are biased by the sample size of the locus we cannot compare loci with different allele numbers directly (similar to Tajima's *D*). We therefore used the following approach: Only loci with eight alleles or more were considered. If a locus contained more than eight alleles all $\binom{n}{8}$ combinations of alignments with exactly eight alleles where generated and the average Z_{nS} value was calculated. This statistic which we will call Z_{nS8} was used to describe the locus-specific LD.

To estimate recombination rates for each locus we used the program Recomb-Rate (COMERON *et al.* 1999) which follows an approach by KLIMAN and HEY (1993). Levels of recombination are given as recombination rate per base pair per generation × 10^{-8}.

Demographic modeling and estimation of the sex ratio in the African population: For a diploid sexual population, we denote the number of females as N_f, the number of males as N_m, the effective population size of the X chromosome as N_x, the effective population size of the autosome as N_a, and the sex ratio as $\beta = N_f / N_m$. We assume that the sex ratio is constant (over time) within a population, but may vary from one population to another one. Then we have $N_x = \frac{9N_mN_f}{4N_m + 2N_f} = \frac{9\beta}{4+2\beta}N_m$, and $N_a = \frac{4N_mN_f}{N_m + N_f} = \frac{4\beta}{1+\beta}N_m$ (HEDRICK 2000), where N_x and N_a are large relative to the sample size. Their ratio is

$$N_x/N_a = 9(1+\beta)/8(2+\beta). \qquad (1)$$

N_x/N_a has the lower bound $\lim_{\beta\to 0}(N_x / N_a) = 0.5625$ and the upper bound $\lim_{\beta\to\infty}(N_x / N_a) = 1.125$.

Following LI and STEPHAN (2006), we assume that the demographic history of the African population is characterized by an instantaneous expansion model. That is, the effective population size of the X chromosome increased instantaneously from N_{x1} to N_{x0} at $2N_{x0}t_x$ $(= 2N_{a0}t_a)$ generations ago, where N_{x1} and N_{x0} are the ancestral and current effective population sizes of the X chromosome in the African population, respectively, and N_{a0} is the current effective population size of the autosome in the African population. Then we have $t_a = \frac{9(1+\beta)}{8(2+\beta)} t_x$. Thus the time back to the expansion for the autosome (in units of $2N_{a0}$) may be different from that for the X chromosome (in units of $2N_{x0}$), while the strength of the expansion is the same (*i.e.*, $N_{x0} / N_{x1} = N_{a0} / N_{a1}$), where N_{a1} is the ancestral effective population size of the autosome in the African population.

The average mutation rates of the X and the autosome are denoted by $\overline{\mu}_x$ and $\overline{\mu}_a$ (per base pair), respectively. They are estimated from divergence between *D. melanogaster* and *D. simulans*. We then have $\theta_x = 4N_{x0}\overline{\mu}_x$ and $\theta_a = 4N_{a0}\overline{\mu}_a = \frac{8(2+\beta)\overline{\mu}_a}{9(1+\beta)\overline{\mu}_x}\theta_x$. Thus, the unknown parameters in the model are θ_x, t_x, N_{x0} / N_{x1} and β.

Following LI and STEPHAN (2006), we summarize the SNP data in terms of the mutation frequency spectrum (MFS). The likelihood for the *k*-th locus on the X chromosome is then given as $L_{xk} = P(MFS \mid \overline{G}_{xk}) = \prod_{i=1}^{n_k-1} P(\xi_{ik} \mid E(l_{ik}))$, where $\overline{G}_{xk}$ is a set of $(n_k - 1)$ expected branch lengths under the demographic scenario for the X chromosome. The branch length is scaled so that one unit represents $2N_{x0}$ generations; n_k is the sample size of the *k*-th locus, ξ_{ik} the number of derived mutations carried by *i* sampled chromosomes for the *k*-th locus, and $E(l_{ik})$ the expected length of branches with *i* descendants for the *k*-th locus under the demographic scenario. $P(\xi_{ik} \mid E(l_{ik}))$ is given by the Poisson probability,

i.e., $P(\xi_{ik} \mid E(l_{ik})) = \frac{\lambda_{ik}{}^{\xi_{ik}} e^{-\lambda_{ik}}}{\xi_{ik}!}$, with $\lambda_{ik} = E(l_{ik})\theta_{xk} / 2$, where $\theta_{xk} = 4N_{x0}\mu_k$, and μ_k is the mutation rate of the *k*-th locus (per locus).

The likelihood for the *k*-th locus on the autosome is given as

$L_{ak} = P(MFS \mid \overline{G}_{ak}) = \prod_{i=1}^{n_k-1} P(\xi_{ik} \mid E(l_{ik}))$, and $P(\xi_{ik} \mid E(l_{ik})) = \frac{\lambda_{ik}{}^{\xi_{ik}} e^{-\lambda_{ik}}}{\xi_{ik}!}$, with

$\lambda_{ik} = E(l_{ik})\theta_{ak} / 2$, where $\theta_{ak} = 4N_{a0}\mu_k$, and the branch length is scaled so that one unit represents $2N_{a0}$ generations. $\overline{G}_{xk}$ is defined in analogy to $\overline{G}_{ak}$ for the X chromosome.

Then $L = \prod_{k=1}^{m_a} L_{ak} \times \prod_{k=1}^{m_x} L_{xk}$, where m_a and m_x are the numbers of loci on the autosome and X chromosome, respectively. A grid search is performed to maximize the likelihood.

Demographic modeling and estimation of the sex ratio in the European population: Since there is convincing evidence that the European population is derived from an ancestral African population (GLINKA *et al.* 2003, BAUDRY *et al.* 2004), we use a two-population model (Figure 2 of LI and STEPHAN 2006) to infer the demographic history and the sex ratio of the European population. In the following, the indices *A* and *E* distinguish the model parameters for the African and the European populations, respectively.

Similar to the definitions for the African population, we have $\beta_A = N_{fA} / N_{mA}$,

$N_{xA0}/N_{aA0} = 9(1+\beta_A)/8(2+\beta_A)$, $\theta_{xA} = 4N_{xA0}\overline{\mu}_x$, $\theta_{aA} = 4N_{aA0}\overline{\mu}_a = \frac{8(2+\beta_A)\overline{\mu}_a}{9(1+\beta_A)\overline{\mu}_x}\theta_{xA}$, and

$t_{aA0} = \frac{9(1+\beta_A)}{8(2+\beta_A)} t_{xA0}$. For the European population, we have $\beta_E = N_{fE} / N_{mE}$,

$N_{xE0}/N_{aE0} = 9(1+\beta_E)/8(2+\beta_E)$, $\theta_{xE} = 4N_{xE0}\overline{\mu}_x$, and $\theta_{aE} = 4N_{aE0}\overline{\mu}_a = \frac{8(2+\beta_E)\overline{\mu}_a}{9(1+\beta_E)\overline{\mu}_x}\theta_{xE}$. We assume that the sex ratio of the European population is constant and may be different from that of the ancestral African population.

Following LI and STEPHAN (2006), we assume that the demographic history of the European population is characterized by an instantaneous bottleneck model. The demographic history of the European population for the X chromosome is parameterized by t_{xE0}, t_{xE1}, $strength_{xE}$ $(= N_{xE0} / N_{xE1})$ and $ratio_{xE}$ $(= N_{xA0} / N_{xE0})$, where time is given in units of $2N_{xA0}$ generations. Similarly, the demographic history of the European population for the autosome is parameterized by t_{aE0} $(= \frac{9(1+\beta_A)}{8(2+\beta_A)}t_{xE0})$, t_{aE1} $(= \frac{9(1+\beta_A)}{8(2+\beta_A)}t_{xE1})$, $strength_{aE}$ $(= N_{aE0} / N_{aE1} = strength_{xE})$, and $ratio_{aE}$ $(= N_{aA0} / N_{aE0} = \frac{(2+\beta_A)(2+\beta_E)}{(1+\beta_A)(1+\beta_E)}ratio_{xE})$, where time is measured in units of $2N_{aA0}$ generations. Thus, the unknown parameters in the two-population model are t_{xE0}, t_{xE1}, $strength_{xE}$, $ratio_{xE}$ and β_E because the parameters for the African (ancestral) population are estimated according to the procedure described above.

We summarize the SNP data in the two populations in terms of the joint mutation frequency spectrum. The maximum likelihood method outlined previously (LI and STEPHAN 2006) is used to estimate the demographic scenario and the sex ratio in the derived European population. In this analysis, we only used the fragments that are sequenced in both populations.

Likelihood ratio test and likelihood-based confidence intervals: The likelihood ratio test (LRT) is a statistical test of the goodness-of-fit. If the null model and the alternative model are hierarchically nested, and the former model has one parameter less than the latter, then we define $\zeta = \log(\max L_1 / \max L_0)$, where L_1 and L_0 are the likelihoods for the alternative and null models, respectively. Then we have $\zeta \geq 0$ because of $\max L_1 \geq \max L_0$. Since ζ may not be approximated by a χ^2 distribution with 1 *df*, we obtain the empirical distribution of ζ from 1000 simulated datasets under the null model. The LRT (a one-tail test) is conducted as follows: we reject the null model at the 5% significance level if $\zeta_{95\%} < \zeta_{obs}$, where $\zeta_{95\%}$ is the critical value.

For the African population, polymorphism data sets of the X and autosome are simulated conditional on the constant population size model and the local recombination rate (COMERON *et al.* 1999). The coalescent process was described previously (LI and STEPHAN 2006). We assume that there is no recombination within loci since the average fragment length is ~500bp. The sex ratio is 1. An empirical distribution of ζ is shown in Figure 1.1.

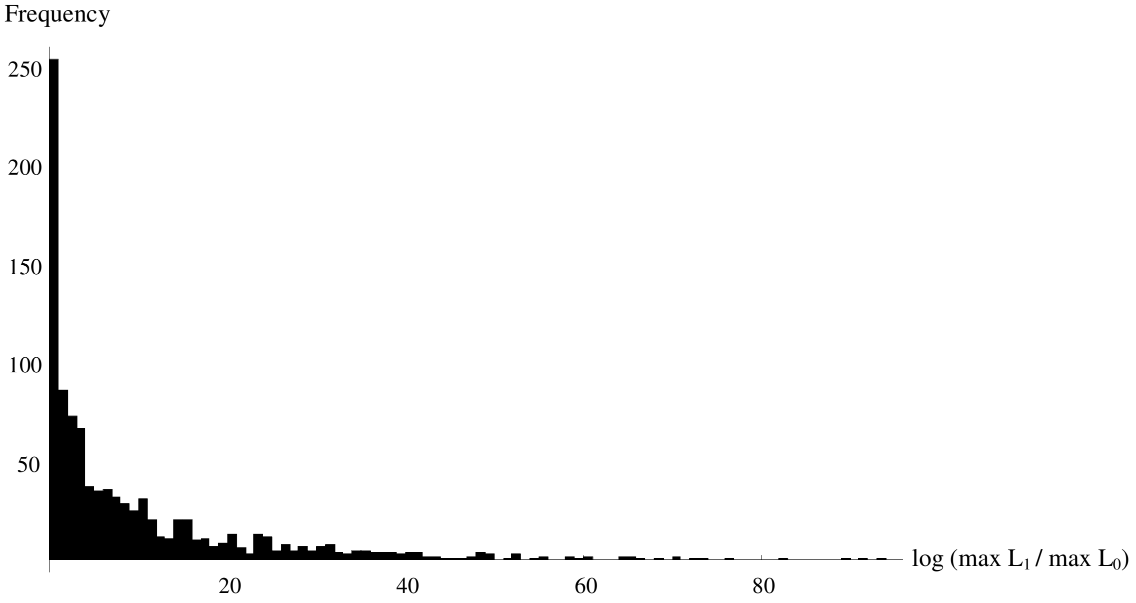


Figure 1.1: Empirical distribution of $\zeta = \log(\max L_1 / \max L_0)$ which is obtained by analyzing 1000 simulated African datasets conditional on the null hypothesis (see text). The analysis is based on the profile likelihood for sex ratio. The critical value ($\zeta_{95\%} = 44.46$) is obtained from the empirical distribution, and is further used in the LRT.

For the European population, a two-population model is used to simulate the joint MFS (LI and STEPHAN 2006). The estimated African expansion scenario is used, and we assume that no bottleneck occurred when the European population is derived from the ancestral African population. The current population size is equal between the two populations. The joint MFS of the X chromosome is simulated conditional on the local recombination rate (COMERON *et al.* 1999). To simulate the joint MFS of the autosome, we assume that the autosomal loci are independent. The sex ratio is 1 for both populations.

In case of a multi-parameter model ($\vartheta_1, \vartheta_2, \ldots, \vartheta_k$), we may be interested in the confidence interval (CI) of one parameter at a time, say in ϑ_1. Let $L_p(\vartheta_1) = \max_{\vartheta_2,\ldots,\vartheta_k} L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k)$ be the profile likelihood. Then the likelihood-based approximative 95% CI is $\left\{ \vartheta_1, \log \frac{L_p(\hat{\vartheta}_1)}{L_p(\vartheta_1)} \le \zeta_{95\%} \right\}$, where $\hat{\vartheta}_1$ is the maximum likelihood estimate of ϑ_1 (PAWITAN 2001).

RESULTS

Inversion patterns: In the African sample we detected inversions on chromosome 3L in line 145. Inversions on 3R were found in lines 131, 157 and 229. In the European sample we detected an inversion on 3R in line 13. We did not observe any 3L inversions in Europe. Genes included in an inversion do not recombine with genes on the standard chromosome, with the exception of rare double crossovers (WESLEY and EANES 1994, AULARD *et al.* 2002). Therefore, we excluded lines showing an inversion from subsequent analyses.

Autosomal polymorphism patterns of the European population: We analyzed a total of 378 fragments located on both arms of chromosome 3 (125 fragments on 3L and 253 fragments on 3R) in 11 inversion-free lines of the European. The length of fragments (excluding insertions and deletions) ranges between 162 and 672 bp with a mean of 536 bp. On average data could be obtained from 10.8 lines. Fragments located on 3L have an average distance of 63 kb encompassing a total of 7.2 Mb and show recombination rates of 3.7 to 5.0 $\times 10^{-8}$. Fragments located on 3R are on average 46 kb apart, spanning a total region of 11.7 Mb. Their recombination rates lie between 1.2 and 3.9×10^{-8}. Data from both chromosomal arms were pooled and analyzed jointly to cover a broad range of recombination rates.

Of the 378 loci surveyed, only a single fragment has no polymorphism. This lack of polymorphism does not result from an overly short alignment (639 bp) or reduced mutation rate (divergence to *D. simulans* is 0.084). The θ value averaged over all fragments is 0.0068 with a standard error (SE) of 0.0002. This is only approximately half as high as the diversity levels reported for synonymous sites on autosomes in non-African populations (0.0155, ANDOLFATTO 2001). The mean (SE) divergence to *D. simulans* is 0.050 (0.0016). For the D/D_{min} statistic describing the frequency distribution we observed a mean (SE) of -0.08 (0.027) which is close to the standard neutral expectation. When looking at Tajima's *D* values of the loci individually we found that a total of 47 of 377 fragments differ from standard neutral expectations; 15 are significantly positive and 32 significantly negative. Z_{nS8} values show an average (SE) of 0.40 (0.010). The summary statistics are shown in Table 1.1.

Autosomal polymorphism patterns of the African population: A random subset of 53 fragments was sequenced in the African population. 36 fragments are located on 3R and 17 on 3L. Again only lines harboring no inversions were used. On average data could be obtained for 7.96 lines.

Mean (SE) level of diversity is 0.0114 (0.0011). No invariant loci were observed. Comparing this to previous results from synonymous sites in African populations we can see that our diversity estimates are again lower (0.0161 on autosomes, ANDOLFATTO 2001).

Table 1.1: Summary statistics of the four data sets analyzed

	Autosome Europe	Autosome Africa	X chromosome Europe[b]	X chromosome Africa[b]
Number of loci	378	53	259	249
Avg. sample size	10.96	7.96	11.86	11.72
Avg. length of alignment [a]	536.1	510.5	527.7	501.3
Avg. (SE) recombination rates × 10^{-8}	3.3 (0.06)	3.4 (0.16)	3.5 (0.07)	3.5 (0.07)
Avg. (SE) GC-content in %	41.4 (0.003)	41.6 (0.007)	39.1 (0.004)	39.0 (0.004)
Number of invariant loci	1	0	21	0
Avg. (SE) θ in %	0.68 (0.02)	1.14 (0.11)	0.47 (0.03)	1.34 (0.05)
Avg. (SE) π in %	0.66 (0.03)	1.04 (0.10)	0.48 (0.03)	1.17 (0.04)
Avg. (SE) divergence (*K*) in %	5.0 (0.16)	5.2 (0.46)	6.8 (0.22)	6.4 (0.19)
Avg. (SE) θ/K	0.16 (0.005)	0.25 (0.025)	0.08 (0.004)	0.22 (0.006)
Avg. (SE) D/D_{min}	-0.08 (0.027)	-0.25 (0.052)	-0.09 (0.045)	-0.32 (0.018)
Significant Tajima's *D* values (+/-)	15/32	0/3	25/50	0/21
Avg. (SE) Z_{nS8}	0.40 (0.010)	0.29 (0.019)	0.51 (0.018)	0.20 (0.005)

[a] Excluding indels
[b] The updated data set (Ometto *et al.* 2005) based on the annotation of the *D. melanogaster* genome Release 4.2.1

D/*D*$_{min}$ has a mean of -0.25 with a SE of 0.052. A locus-by-locus inspection of the frequency spectrum revealed that three loci deviate significantly from the standard neutral model; all of them have negative values. Values of Z_{nS8} show a mean (SE) of 0.29 (0.019). The numbers are summarized in Table 1.1.

Contrasting autosomal patterns between populations: At first we investigated if our African subset of loci deviates from the European set in terms of average recombination or mutation rate (estimated by divergence to *D. simulans*). Neither recombination rate (Mann-Whitney *U* test, $P = 0.77$) nor divergence (Mann-Whitney *U* test, $P = 0.91$) is statistically different. These factors should therefore not influence our comparisons between the populations. Average levels of diversity are significantly lower in Europe (Mann-Whitney *U* test, $P < 0.001$). The mean *D*/*D*$_{min}$ is significantly higher in Europe (Mann-Whitney *U*-test, $P = 0.02$). Both populations are known to have undergone different demographic events (LI and STEPHAN 2006). Such events affect not only means of summary statistics describing patterns of polymorphism but also their variance. Therefore contrasting variances of statistics such as *D*/*D*$_{min}$ and Z_{nS8} can provide useful information on the demographic history (*e.g.*, HADDRILL *et al.* 2005b). Tables 1.2 and 1.3 summarize the variances of *D*/*D*$_{min}$ and Z_{nS8} for our data sets. To find out if empirical variances differ between chromosomes and populations we conducted Levene tests. Comparing variances of *D*/*D*$_{min}$ between European and African autosomes we find that the empirical variance is significantly higher in Europe (Levene test, $P = 0.01$). To check if this larger variance also results in an increase of significant Tajima's *D* values we conducted a Fisher's Exact test. Neither the number of positive values ($P = 0.14$) nor the number of negative values ($P = 0.37$) is significantly increased in the European population. LD behaves similar to the frequency spectrum: The mean value of Z_{nS8} (Mann-Whitney *U* test, $P < 0.001$) and its variance (Levene test, $P < 0.001$) are significantly elevated in Europe.

Table 1.2: Pair wise tests for differences in variance of D/D_{min}

	D/D_{min}		
	X chromosome	Autosome	Levene test
Europe	0.492	0.269	$P < 0.001$
Africa	0.085	0.144	$P = 0.012$
Levene test	$P < 0.001$	$P = 0.009$	

X-linked data: We calculated the same statistics from the updated X-linked data set of OMETTO *et al.* (2005) for comparison although the updating produces little differences in the summary statistics (Table 1.1).

We compared the X-linked data between populations in the same way as the autosomal ones. Mutation and recombination rates do not differ between Africa and Europe (Mann-Whitney *U*-test, $P = 0.30$ and $P = 0.84$, respectively) as found for the autosomal loci. Average levels of diversity and the variance are significantly reduced in Europe (Mann-Whitney *U* test, $P < 0.027$, and Levene test, $P < 0.001$). The average D/D_{min} value is higher in Europe than in Africa, but this difference is not statistically significant (Mann-Whitney *U* test, $P = 0.27$). The failure to detect a difference here might be due to the high variance in the European population. It is significantly higher than in Africa (Levene test, $P < 0.001$) and in fact it is the highest in our data set (Table 1.2). This high variance in the frequency spectrum also leads to an increased number of significant Tajima's *D* values in Europe. Both the number of significantly positive and negative values is higher than expected when compared to the African population (Fisher's Exact test, $P < 0.001$ in both cases). Comparison of LD showed that the average value of Z_{nS8} is higher in Europe (Mann-Whitney *U* test, $P < 0.001$), and so is its variance (Levene test, $P < 0.001$).

Table 1.3: Pair wise tests for differences in variance of Z_{nS8}

	Z_{nS8}		
	X chromosome	Autosome	Levene test
Europe	0.068	0.038	$P < 0.001$
Africa	0.007	0.019	$P = 0.060$
Levene test	$P < 0.001$	$P < 0.001$	

Comparison of X chromosomes and autosomes within populations: It is important to note that levels of divergence are different when comparing the chromosomes. Divergence is elevated on the X chromosome in the European (Mann-Whitney *U* test, $P < 0.001$) and the African data sets (Mann-Whitney *U* test, $P = 0.001$). This cannot be due to systematic differences in mutation rates since studies have shown that evolutionary rates do not differ between the chromosomes in *D. melanogaster* (BETANCOURT *et al.* 2002). HADDRILL *et al.* (2005a) found a negative correlation of divergence and GC-content in introns and concluded that base composition influences the local mutation rate. Confirming this hypothesis, we observed a significantly elevated GC-content in our autosomal loci for both populations (Mann-Whitney *U* test, $P < 0.001$ in Europe and $P = 0.008$ in Africa). To account for the resulting differences in mutation rate we corrected the levels of diversity by dividing individual θ values by the local divergence. These θ/K values were then used for estimating the ratios of X-chromosomal to autosomal diversity. Expanding the results of HADDRILL *et al.* (2005a) we find that the effect of base composition is not only confined to introns. We took the combined 232 purely intergenic loci from both chromosomes of the European data set and correlated the GC-content with divergence. A significant negative correlation can be observed (Spearman correlation coefficient $R = -0.477$, $P < 0.001$).

We also tested if levels of recombination are comparable between our X-linked and autosomal data sets. A Mann-Whitney *U* test shows that levels of recombination are neither different for the African data sets ($P = 0.47$) nor the European ones ($P = 0.057$). It should, however, be noted that the *P* value of the European population is close to significance.

The ratios of X-linked to autosomal polymorphism are 0.49 for the European and 0.90 for the African population. Previous studies report much higher numbers. ANDOLFATTO (2001) finds a ratio of 0.66 in non-African and 1.60 in African flies using synonymous sites. KAUER *et al.* (2002) used microsatellite heterozygosity and observed ratios of 0.78 outside of Africa and 1.20 in Africa. It should be noted that these studies do not correct for possible differences in mutation rate so these ratios might be subject to mutational biases. When leaving θ values uncorrected for our data set we observe ratios of 0.66 and 1.17 in Europe and Africa, respectively. These numbers are in good agreement with the uncorrected numbers reported in previous studies. BEGUN and WHITLEY (2000) looked at ratios of diversity in non-African *D. simulans* populations. They could correct for different mutation rates since *D. melanogaster* was available as outgroup. Using these corrected diversity levels they found a X-chromosomal to autosomal ratio of polymorphism of 0.49 which is exactly the value we obtained for our European data set.

We wanted to know if the observed diversity ratios were significantly different from the expected 0.75 assuming a sex ratio of 1:1 in an equilibrium population. To do this we multiplied X-linked data by 4/3 and performed Mann-Whitney *U* tests. We find that the X chromosome lacks diversity in Europe ($P < 0.001$) but is too diverse in Africa ($P < 0.001$). We also compared the patterns of the frequency spectrum and LD. In Europe the average values of D/D_{min} do not differ between the X chromosome and autosome (Mann-Whitney *U* test, $P = 0.48$) while the variance is increased for the X chromosome (Levene test, $P < 0.001$). This larger variance leads to significantly more loci deviating from standard neutral expectations on the X chromosome. There are more loci with significantly positive (Fischer's

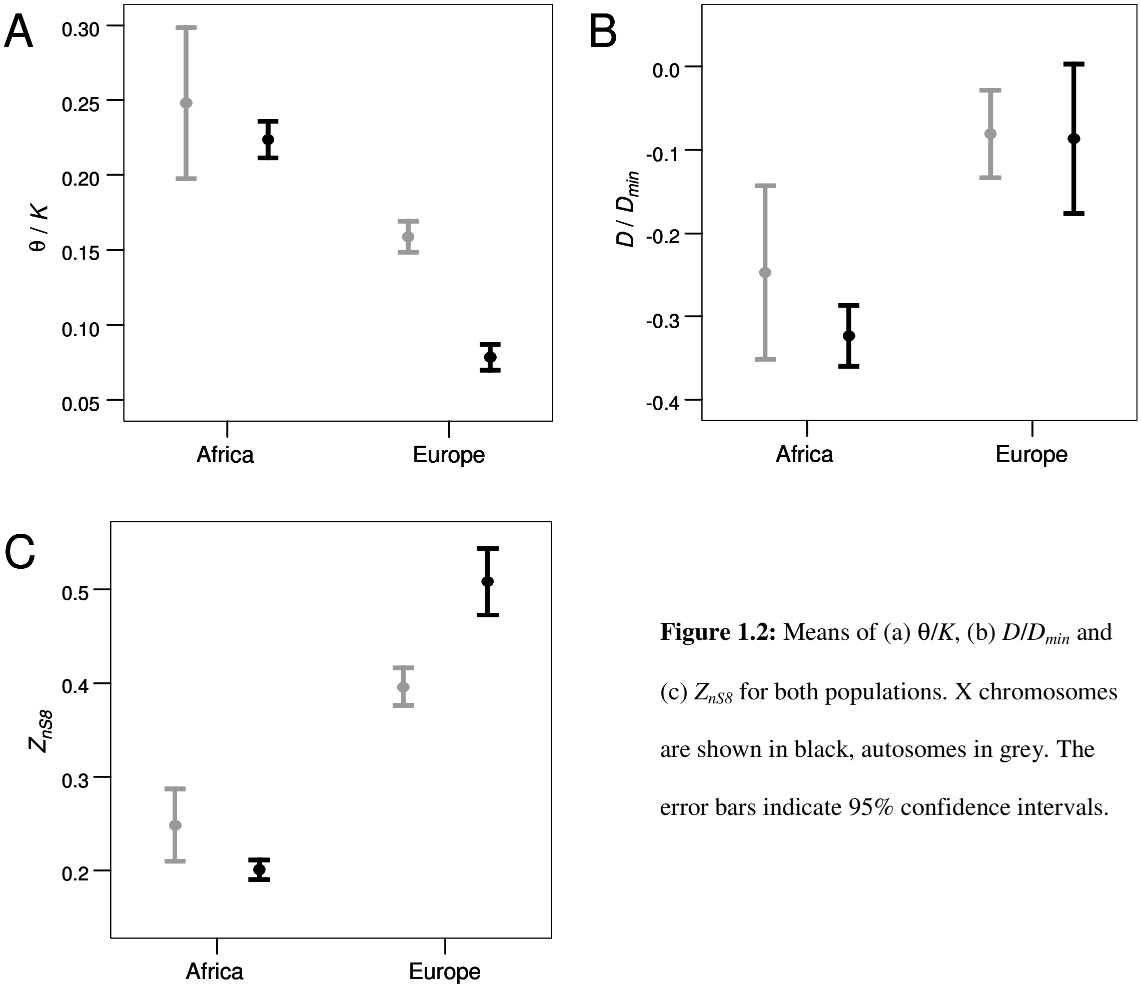


Figure 1.2: Means of (a) θ/*K*, (b) D/D_{min} and (c) Z_{nS8} for both populations. X chromosomes are shown in black, autosomes in grey. The error bars indicate 95% confidence intervals.

Exact test, $P = 0.003$) and significantly negative values ($P < 0.001$) than expected in comparison with the autosome. The Z_{nS8} statistic shows an elevated level of LD on the X chromosome (Mann-Whitney *U* test, $P < 0.001$) along with an elevated variance (Levene test, $P < 0.001$). A study examining the pattern of LD in non-African *D. simulans* found a very similar pattern (WALL *et al.* 2002). In the African population mean D/D_{min} values do not differ statistically either (Mann-Whitney *U* test, $P = 0.13$) while the variance is higher on the autosome (Levene test, $P = 0.012$). The mean LD is higher on the autosome (Mann-Whitney *U* test, $P = 0.001$) while the variances do not differ significantly (Levene test, $P = 0.06$). The means and 95% confidence intervals of θ/*K*, *D*/*D*$_{min}$ and *Z*$_{nS8}$ for all four data sets are pictured in Figure 1.2 for better comparison.

Inferring the demographic history and the sex ratio in the African and European populations: Based on X-linked data it has been proposed that the African population has expanded in recent time (OMETTO *et al.* 2005, LI and STEPHAN 2006). To further examine this hypothesis, we compare the instantaneous population expansion model with the standard neutral model, using both our X-linked and autosomal data sets. That is, the null hypothesis is $N_{xA0} = N_{xA1}$ and $N_{aA0} = N_{aA1}$ (the simple model), and the alternative hypothesis is $N_{xA0} \geq N_{xA1}$ and $N_{aA0} \geq N_{aA1}$ (the complex model). The LRT (*P* < 0.01, the critical value $\zeta_{99\%} = 41.21$) suggests that the expansion model ($\max \log(L) = -4754.2$) explains the features of the polymorphism data in the African sample significantly better than the standard neutral model ($\max \log(L) = -4896.3$).

Since the *D. melanogaster* lineage split from *D. simulans* approximately 2.3 million years ago (LI *et al.* 1999) and the average divergence over loci on the X and third chromosome is 0.0667 and 0.0522, respectively, $\overline{\mu}$ is 1.450×10^{-9} and 1.135×10^{-9} per site per generation (assuming 10 generations per year). Thus, in the expansion model, $\hat{\theta}_{xA0} (= 4N_{xA0}\overline{\mu}_x)$ is 0.050 (*i.e.*, $\hat{\theta}_{aA0} = 4N_{aA0}\overline{\mu}_a = 0.047$). Then $\hat{N}_{xA0}$ and $\hat{N}_{aA0}$ are 8.621×10^6 and 10.40×10^6, respectively, and the ratio of population sizes (N_{xA0} / N_{aA0}) is 0.829 with a 95% confidence interval of (0.636, 1.08). The ratio N_{xA0} / N_{aA0} is slightly less than the ratio of X-linked to autosomal polymorphism obtained from Watterson's θ (0.90). When the ratio of X-linked to autosomal polymorphism is inferred as the ratio N_{xA0} / N_{aA0}, it is assumed that the African population is under equilibrium. However, after the population size expands, genetic diversity on the X chromosome is expected to reach equilibrium before that of autosomes because of the smaller effective population size of the X chromosome, resulting in an excess of diversity on the X (see also DISCUSSION).

The estimated time to the expansion in the past is 60,300 years with a 95% confidence interval of (13.8, 172.0) ky, which is very similar with our previous estimate (LI and STEPHAN 2006). The strength of the expansion, measured by the ratio of the current size to the size before the expansion, is 5.0 (2.0, 12.0). The sex ratio ($N_f \,/\, N_m$) is 1.8, but the LRT suggests that the number of females in the African population is not significantly larger than the number of males ($P > 0.05;\ \max\log(L) = -4756.6$ *versus* -4754.0; the critical value $\zeta_{95\%} = 44.46$).

For the European population, the time of the out-of-Africa migration is 17,500 years, which is similar to our previous estimate from the X-linked data set (LI and STEPHAN 2006), and $\hat{N}_{xE0} = 0.958 \times 10^6$. We find $\hat{\beta} = 0$ (0, 0.082), and the ratio of population sizes ($N_{xE0} \,/\, N_{aE0}$) is 0.5625. The LRT suggests that there is a vanishingly small percentage of females in the European population ($P < 0.01;\ \max\log(L) = -12452.9$ *versus* -12548.8; the critical value $\zeta_{99\%} = 8.45$). The unrealistic estimate of the sex ratio suggests that the low diversity on the X chromosome in the European population cannot be explained by the bottleneck and the smaller effective population size for the X chromosome alone.

Compared to the autosomal diversity in the European population, the reduced diversity on the X chromosome could be due to purifying selection alone. For this reason, following FU (1997) and SMITH and EYRE-WALKER (2002), we repeated the analyses disregarding the singletons (*i.e.*, the mutations carried by a single chromosome in the sample). In this case, we also have $\hat{\beta} = 0$ (0, 0.31), which may suggest that purifying selection does not play a major role.

DISCUSSION

Different levels of nucleotide diversity between the X and third chromosome and between ancestral and derived populations: Our scan of sequence diversity on the third chromosome in an ancestral African and a derived European population of *D. melanogaster* shows a pattern similar to that already found by GLINKA *et al.* (2003) on the X chromosome in the same populations: the European population exhibits reduced levels of polymorphism. However, while the level of diversity drops to 35% on the X chromosome (relative to the ancestral one), the European autosome retains 62% of the ancestral θ value. Thus, the question arises what forces lead to this differences in reduction of polymorphism. In addition, the ancestral population shows a ratio of X-chromosomal to autosomal diversity of 0.90, which is significantly higher than the expectation of 0.75 under standard neutral conditions. Following the more severe drop-off of diversity on the X chromosome, this leads to a ratio of 0.49 in Europe that is significantly lower than the standard neutral expectation. Such disparities have already been reported in previous studies (ANDOLFATTO 2001, KAUER *et al.* 2002).

It is now well established that the reduction of variation in the derived population can be attributed mainly to a population size bottleneck during range expansion (HADDRILL *et al.* 2005b, OMETTO *et al.* 2005, LI and STEPHAN 2006). However, it appears that autosomes and sex chromosomes were affected differently during this colonization process. CHARLESWORTH (2001) suggested that an unequal sex ratio might explain these observations. If any of both sexes experiences a large variance in reproductive success (NUNNEY 1993) or if a substantial fraction of individuals fail to reproduce during their lifetime (CHARLESWORTH 2001), this will lead to a reduction of effective population size. Reproduction in natural populations of Drosophila is highly unlikely to occur by random mating, and in fact there is abundant evidence (CROW and MORTON 1955, BOULÉTREAU 1978, SOLLER *et al.* 1999) that sexual selection and environmental factors might lead to different population sizes for males and females. If these effects cause an unequal sex ratio, this will be reflected in the ratio of X-chromosomal to autosomal effective population sizes. Since males carry only one X chromosome as opposed to two in females, differences in male population size have a smaller effect on X-linked than autosomal diversity. Therefore, unequal sex ratios will lead to a deviation from the standard expectation of 0.75 for ratios of diversity. A goal of this study was to obtain accurate estimates of the sex ratios in both populations, taking their demography into account.

Demography and sex ratio in the African population: Our estimation of the demographic history of the African population using the combined X-chromosomal and autosomal data set confirms the results of LI and STEPHAN (2006) where only X-linked data was used, although we find that the empirical distribution of ζ does not follow a χ^2 distribution (Figure 1.1), which was used in our previous analysis (LI and STEPHAN 2006). The most likely scenario for the African population is a population size expansion approximately 60,000 years ago. The estimate of the sex ratio suggests that the female population size is 1.8 times larger than the male one in Africa. Although this difference is not significant, it may suggest that an unequal sex ratio played an important role in shaping X-chromosomal and autosomal diversity in Africa. CHARLESWORTH (2001) tried to explain the excess of female effective population size in the ancestral population: females are in good breeding condition in Africa, so there is little variance in reproductive success. Males, on the other hand, are subject to strong sexual selection that reduces the male effective population size.

In addition to our maximum likelihood approach, an estimation of the sex ratio can be obtained directly from levels of diversity. If we take the ratio of θ/K values as a proxy for the ratio of population sizes of X chromosomes and autosomes and plug them into equation (1), we obtain a female/male ratio of 3.0 for the African population. This ratio is larger than the estimate (1.8) from our analysis. The difference between both values may be understood as follows. The populations are not in equilibrium. In such a case estimating population sizes from diversity (using Watterson's θ) will lead to an underestimation of effective population size. Since the X chromosome has a smaller population size than autosomes it will reach equilibrium faster after the expansion. Therefore the underestimation of population size will not be as extreme on the X as on the autosome. This leads to a bias towards higher female/male ratios. Even though we estimate an elevated population size for females, the population size of the X chromosome is still considerably smaller than that of autosomes. Following equation (1) one would need a seven-fold excess of females to achieve equal population sizes for X chromosomes and autosomes, but our estimate is well below that value. Therefore the argument of a smaller X-chromosomal N_e still holds.

KAUER *et al.* (2002) used equation (1) to test if unequal sex ratios could explain their data. But even when they assumed a 50-fold excess of female population size the X chromosome seemed too variable. This is because their ratio of microsatellite heterozygosity was 1.21, which is larger than the limiting value of 9/8 that can be achieved when male population size approaches zero. A reason for this high ratio could be a bias in mutation rate. If we leave θ uncorrected for our SNP data set we find that the ratio equals 1.18 for the African population. This ratio corresponds well with the microsatellite data; so a biased mutation rate might indeed explain these findings.

Not only do we see differences in overall levels of diversity between X chromosomes and autosomes but there are also differences in the frequency spectrum. D/D_{min} values are slightly (although not significantly) more negative on the X chromosome and the variance is reduced. Population size expansions are known to lead to an increase of low frequency variants (*i.e.*, creating negative Tajima's *D* values), and reduce the variance of Tajima's *D* across the loci. The magnitude of this effect depends on the parameter values of the expansion. Although the population size expansion affects both the X chromosome and the autosome, the effect may not be the same, and the expansion scenario is different for the X chromosome and the autosome from a coalescent point of view. Since N_{xA0} is less than N_{aA0}, we have $t_x = 0.035$ and $t_a = 0.029$. Coalescent simulations show that the expansion scenario for the X chromosome results in more negative values of Tajima's *D* and a lower variance than that for the autosome (results not shown). Therefore, our observation on D/D_{min} supports the hypothesis that the African population may have undergone a population size expansion.

D/D_{min} is not the only summary statistic that shows differences between chromosomes. The average Z_{nS8} is significantly lower on the X chromosome implying that there is less LD. The variance is also lower on the X chromosome, and this difference is marginally significant ($P = 0.06$). The effect of population size expansions on LD is very similar to that on the frequency spectrum. Average LD tends to get lower and the variance becomes smaller if the population underwent a size increase. Coalescent simulations show that expansions with X-chromosomal parameter values produce lower averages and smaller variances, in accordance with the trend we find in our data.

Thus, our analyses suggest that the population size expansion (about 60,000 years ago) together with an unequal sex ratio can account for the genome-wide patterns of polymorphism that we observe on the X chromosome and the autosome in the African population.

Demography and sex ratio in the European population: Our inference of the demography and sex ratio for the European population produced a surprising result. The most likely demographic scenario is a population bottleneck with subsequent expansion, but the estimated female/male ratio is zero implying that there are an extremely small percentage of females present in Europe. In such a case, we have $N_f = 4/9N_x$. We estimate that the current effective population for the X chromosome in the European population is 0.958×10^6, which suggests that the number of females is 0.426×10^6, whereas the number of males in the

European population is much larger. This suggests that the lower X-linked diversity cannot be explained by a biased sex ratio alone.

Previous work on non-African *D. melanogaster* (ANDOLFATTO 2001, KAUER *et al.* 2002) found higher ratios of X-chromosomal to autosomal diversity than we observed. However, since these analyses do not control for mutation rate, they may be biased (see above). Another study comparing X-chromosomal and autosomal polymorphism in non-African *D. simulans* (BEGUN and WHITLEY 2000) found exactly the same estimate we obtained in our European data set. This work provides a good comparison to our analysis, because levels of diversity were also corrected (by dividing by divergence). Mutational biases should therefore not affect the results.

Demography and sex ratio alone cannot account for the genomic polymorphism patterns observed, while they may have played important roles in shaping polymorphism in Europe. There is independent ecological evidence that males have a higher population size than females in Europe (BOULÉTREAU 1978), so that the sex ratio seems to be inverted relative to Africa. It has been proposed that this is because of poor breeding conditions in Europe (CHARLESWORTH 2001). The inverted sex ratio in the derived population suggests that the X chromosome underwent a more severe bottleneck than the autosome.

Population size reductions, much like expansions, have distinct effects on the frequency spectrum of polymorphisms and LD. In the case of a population size reduction the frequency spectrum tends to show an excess of intermediate frequency polymorphisms. This leads to elevated Tajima's *D* values. In addition, the variance of Tajima's *D* tends to get larger. A similar pattern is created for the Z_{nS8} statistic. Average LD and its variance increase. It is important to note that these predictions are only sufficiently understood for simple models of population size reduction. The demographic history of the European population, however, seems to be rather complex (Figure 1b of LI and STEPHAN 2006). First, it was derived from an ancestral African population that was not in equilibrium (see above). Of course one could question if the Zimbabwe population actually reflects the true ancestral population. But recent studies have shown that there is a signal for expansion in the vast majority of African populations that have been surveyed (POOL and AQUADRO 2006). It is therefore very likely that the true ancestral population also showed this pattern. Second, the derived population experienced a population size expansion of its own, subsequent to the colonization of Europe. It is difficult to assess how these different events will contribute to the overall patterns of polymorphism.

For average D/D_{min}, we find that it is nearly identical on the X chromosome and autosome in Europe. Interestingly both values are very close to zero as in the case of a standard neutral population. However, since our estimation of the European demography rejects this model, this is most likely the result of a much more complex process. The variance of D/D_{min} shows a large difference between the chromosomes. The larger variance on the X chromosome may again be explained by the different times back to the bottleneck event (in the coalescent view). In the case of very recent bottlenecks slightly "older" events tend to cause more variance in D/D_{min}. A comparison of means and variances of Z_{nS8} between the chromosomes also shows a similar pattern. LD is higher on the X chromosome and has a larger variance. This is also expected under a simple bottleneck model. In a similar case, WALL *et al.* (2002) explained the patterns of LD they found in non-African *D. simulans* (which mimic our findings) with a simple bottleneck. Even though the European population has a complex demographic past the patterns of the frequency spectrum (D/D_{min}) and LD agree well with the predictions of a simple population size reduction. In conclusion, the bottleneck seams to have been the dominant demographic force shaping the patterns of polymorphism we observe in Europe today.

However, the main question remains: What led to the extreme differences in average diversity between the X chromosome and the autosome? We have shown that demography and unequal sex ratio alone cannot account for these differences. BEGUN and WHITLEY (2000) propose that genetic hitchhiking caused by positive selection might have played a major role in reducing X-linked diversity in their *D. simulans* data, since it is thought that positive selection should affect X chromosomes stronger than autosomes (AQUADRO *et al.* 1994). Theoretical work has shown that this claim holds if recombination occurs in both sexes (BETANCOURT *et al.* 2004). In the case of *Drosophila* where recombination only occurs in females, on the other hand, this effect is only visible if mutations are partially recessive (BETANCOURT *et al.* 2004). Thus, if a large fraction of advantageous mutations are indeed recessive (for instance, as suggested by ZEYL *et al.* 2003) the effect should also be visible in *Drosophila*.

Positive selection might also help explain the pattern we see for the frequency spectrum in the European population: A population size bottleneck creates an excess of intermediate frequency variants and this effect is larger for the X chromosome. Genetic hitchhiking, on the other hand, tends to create low-frequency polymorphisms (BRAVERMAN *et al.* 1995) and also is supposed to influence the X chromosome to a larger extent (see above). If both forces act simultaneously their effects on the average levels of Tajima's *D* might cancel out. At the same time this will result in a large variance of Tajima's *D*. Since we expect the effect of both forces to be more pronounced on the X chromosome, this might explain why the X chromosome harbors more loci that deviate from standard neutral expectations (32% of all loci containing polymorphism) than the autosome (only 12%) and, at the same time, average levels of D/D_{min} are approximately equal. The effects on average levels of D/D_{min} cancel out while the increases in variance add up.

The effect of hitchhiking on the genome-wide averages of LD is complex and not well understood. PRZEWORSKI (2002) shows that recent selective sweeps can substantially increase levels of LD. However, the created signal disappears rapidly after fixation of the selected mutation. Furthermore, if there is recurrent positive selection in the population the overall effect of multiple sweeps might even lead to a slight decrease in LD. A more recent study

(STEPHAN *et al.* 2006) shows that hitchhiking can destroy preexisting LD. Depending on the haplotype on which a beneficial mutation appears it will either lead to an increase or a decrease of already present LD. The average over all possibilities leads to a level of LD that will then be slightly lower after the sweep than it was before the emergence of the positively selected mutation. In summary, both studies suggest that the overall effect of hitchhiking is not one of an increase of LD as has been previously thought. On a genome-wide level, hitchhiking rather tends to slightly decrease LD. We conclude that the pattern of LD we observe in our data was therefore mainly shaped by demographic events as these leave a more pronounced signature. The effects of selection on the variance of LD are clearer. Hitchhiking increases the variance (STEPHAN *et al.* 2006). This increase in variance will be more pronounced on the X chromosome since we expect positive selection to be more prevalent there. Therefore, positive selection could have reinforced the differences in variance of LD between the two chromosomes created by the bottleneck.

In summary, the observed patterns of polymorphism provide evidence that population size bottlenecks and positive selection have acted simultaneously in the European population.

ACKNOWLEDGEMENTS

We thank K. Bhuiyan, A. Fabry and A. Wilken for excellent technical assistance. We also thank J. Baines, J. Hermisson, P. Pennings and P. Pfaffelhuber for constructive and helpful advice and discussions. We are grateful to the Volkswagen-Foundation for financial support (grant number I/78 815).

LITERATURE CITED

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

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

AQUADRO, C. F., D. J. BEGUN and E. C. KINDAHL, 1994 Selection, recombination, and DNA polymorphism in Drosophila, pp. 46-56 in *Non-neutral evolution: theories and molecular data*, edited by B. Golding, Chapman and Hall, New York.

AULARD, S., J. R. DAVID and F. LEMEUNIER, 2002 Chromosomal inversion polymorphism in Afrotropical populations of *Drosophila melanogaster*. Genet Res **79:** 49-63.

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

BEGUN, D. J., and P. WHITLEY, 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc Natl Acad Sci U S A **97:** 5960-5965.

BETANCOURT, A. J., Y. KIM and H. A. ORR, 2004 A pseudohitchhiking model of X vs. autosomal diversity. Genetics **168:** 2261-2269.

BETANCOURT, A. J., D. C. PRESGRAVES and W. J. SWANSON, 2002 A test for faster X evolution in *Drosophila*. Mol Biol Evol **19:** 1816-1819.

BOULÉTREAU, J., 1978 Ovarian activity and reproductive potential in a natural population of *Drosophila melanogaster*. Oecologia **35:** 319-342.

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.

CHARLESWORTH, B., 2001 The effect of life-history and mode of inheritance on neutral genetic variability. Genet Res **77:** 153-166.

CHARLESWORTH, B., J. A. COYNE and N. H. BARTON, 1987 The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. **130:** 113-146.

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. AGUADE, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics **151:** 239-249.

CROW, J. F. and N. E. MORTON, 1955 Measurement of gene-frequency drift in small populations. Evolution **5:** 202-214.

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

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

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.

HADDRILL, P. R., B. CHARLESWORTH, D. L. HALLIGAN and P. ANDOLFATTO, 2005a Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. Genome Biol **6:** R67.

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

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 U S A **99:** 12949-12954.

HEDRICK, P. W., 2000 *Genetics of populations*. Jones and Bartlett Publishers, Sudbury, Mass.

HUTTER, S., A. J. VILELLA and J. ROZAS, 2006 Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics **7:** 409.

JUKES, T. H. and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21-132 in *Mammalian protein metabolism*, edited by H. N. Munro, Academic Press, New York.

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.

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

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

LACHAISE, D., M. L. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS and M. ASHBURNER, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol Biol **22:** 159-225.

LEFEVRE, G., 1976 A photographic representation and interpretation of the polytene chromosomes of *Drosophila melanogaster* salivary glands, pp. 32-66 in *The genetics and biology of Drosophila*, vol. 1a, edited by M. Ashburner and E. Novitski, Academic Press, New York.

LI, H., and W. STEPHAN, 2006 Inferring the Demographic History and Rate of Adaptive Substitution in *Drosophila*. PLoS Genet **2**: e166.

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

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

NUNNEY, L., 1993 The influence of mating system and overlapping generations on effective population size. Evolution **47:** 1329-1341.

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

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

PAWITAN, Y., 2001 *In all likelihood: Statistical modeling and inference using likelihood.* Oxford University Press, New York.

POOL, J. E., and C. F. AQUADRO, 2006 History and structure of sub-Saharan populations of *Drosophila melanogaster*. Genetics **174:** 915-929.

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

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

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.

SHAPIRO, J. A., W. HUANG, C. ZHANG, M. J. HUBISZ, J. LU *et al.*, 2007 Adaptive genic evolution in the *Drosophila* genomes. Proc Natl Acad Sci U S A **104:** 2271-2276.

SMITH, N. G., and A. EYRE-WALKER, 2002 Adaptive protein evolution in *Drosophila*. Nature **415:** 1022-1024.

SOLLER, M., M. BOWNES and E. KUBLI, 1999 Control of oocyte maturation in sexually mature *Drosophila* females. Dev Biol **208:** 337-351.

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

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.

WALL, J. D., P. ANDOLFATTO and M. PRZEWORSKI, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203-216.

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

WESLEY, C. S., and W. F. EANES, 1994 Isolation and analysis of the breakpoint sequences of chromosome inversion In(3L)Payne in *Drosophila melanogaster*. Proc Natl Acad Sci U S A **91:** 3132-3136.

ZEYL, C., T. VANDERFORD and M. CARTER, 2003 An evolutionary advantage of haploidy in large yeast populations. Science **299:** 555-558.

2. A survey of variation in gene expression in two natural populations of *Drosophila melanogaster*

ABSTRACT

Changes in levels of gene expression can have large consequences for the phenotype of an organism. Variation of gene expression in natural populations might therefore serve as a substrate for positive Darwinian selection, and hence play a role in the adaptation of populations to their local habitats. We profiled the genome-wide expression of adult males of 16 *Drosophila melanogaster* lines. The flies come from two natural populations with vastly different environments, Zimbabwe and the Netherlands. Our approach allowed us to estimate the levels of gene expression variation within each population and detect genes that show expression patterns which differ significantly between the two populations and hence are candidates for local adaptation. We find that variation is equal in both populations. This argues for stabilizing selection as the major force shaping expression polymorphism. This was also previously suggested by mutation accumulation studies. Supporting this view, genes that are under increased selective constraint, because they are involved in many different biological processes, tend to be less variable. We observe that there is substantial population differentiation at the gene expression level. Candidates to have undergone adaptive evolution include genes that are putatively involved in ecological traits such as insecticide resistance or food choice. Surprisingly, many other candidate genes play a role in the formation and function of the flying apparatus.

INTRODUCTION

Uncovering the underlying basis of phenotypic variation in natural populations and phenotypic differences between species has been the focus of many studies over the past decades. Two main types of genetic changes are believed to contribute to phenotypic differences between individuals: (a) mutations that alter the amino acid sequence of a protein and (b) mutations that alter levels of gene expression (KING and WILSON 1975). From the very beginning there has been debate about the relative contributions of these two sources of variation. Already in the 1970s the intriguing observation that humans and chimpanzees do not show much differentiation in the molecular sequence of serum albumins (WILSON *et al.* 1974) led the authors to believe that changes in gene regulation might have a larger impact on the phenotype than do mutations at the protein level. This view has been reinforced over time by the growing availability of DNA sequence data of closely related species which show only few differences in coding regions. Comparative analysis of the complete genomes of humans and chimpanzees, for example, revealed a strikingly small amount of differentiation at the protein level (~0.2%) despite the considerable phenotypic differences between species (THE CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM 2005). Due to this and the increasing number of studies which provide evidence that regulatory changes can have drastic effects on various phenotypic traits, it is nowadays undisputed that gene expression polymorphism is an important source of adaptive variation (WHITEHEAD and CRAWFORD 2006a).

The advent of microarray technology allows us to investigate the differences in transcript abundance between individuals for a large number of genes. To date, numerous studies have investigated the variation of gene expression in natural populations for a broad range of species, including yeast (CAVALIERI *et al.* 2000, TOWNSEND *et al.* 2003, FAY *et al.* 2004), fish (OLEKSIAK *et al.* 2002, AUBIN-HORTH *et al.* 2005, WHITEHEAD and CRAWFORD 2006b) and hominids (ENARD *et al.* 2002, STRANGER *et al.* 2005).

The fruit fly *Drosophila melanogaster* is of particular interest, because it has long served as model organism for population genetics. Variation at the DNA level in natural populations has been surveyed extensively in microsatellite (*e.g.* KAUER *et al.* 2002) and single nucleotide polymorphism (SNP) studies (*e.g.* OMETTO *et al.* 2005, SHAPIRO *et al.* 2007). These studies have confirmed that *D. melanogaster* originated from an ancestral population in sub-Saharan Africa and only relatively recently expanded to the rest of the world, a scenario that had already been suggested by earlier studies (LACHAISE *et al.* 1988, DAVID and CAPY 1988). Populations residing in the ancestral species range nowadays show the signal of a population size expansion (GLINKA *et al.* 2003, POOL and AQUADRO 2006) while derived populations show the signature of a population bottleneck (ORENAGO and AGUADÉ 2004, OMETTO *et al.* 2005) and extensive theoretical studies (HADDRILL *et al.* 2005, LI and STEPHAN 2006) have estimated parameters for these demographic events.

Most surveys studying gene expression variation in *D. melanogaster* have focused on a small number of laboratory strains derived from non-African populations (JIN *et al.* 2001, RIFKIN *et al.* 2003, GIBSON *et al.* 2004). Thus they do not offer a complete view of expression variation within the species. They are also only of limited value if one wants to estimate the effects of historic demographic events, such as bottlenecks, on levels of gene expression variation within natural populations. An exception is the study of MEIKLEJOHN *et al.* (2003). Here the authors investigated gene expression polymorphism in adult males of eight strains of *D. melanogaster*, including four strains from an ancestral population from Zimbabwe and four non-African (cosmopolitan) lab strains. This study uncovered greater levels of variation than previous studies due to its inclusion of ancestral, African strains. There were, however, some limitations to this work. For example, the sample size was relatively small, with only four African and four non-African strains. Furthermore, the cosmopolitan sample was not from a single population, but instead was a mixture of North American and Asian laboratory stocks. Finally, the MEIKLEJOHN *et al.* (2003) study used microarrays made from the Drosophila

Gene Collection (DGC) 1.0 which contained probes for ~6000 genes identified from an initial EST screen of the *D. melanogaster* genome (RUBIN *et al.* 2000); therefore the resulting arrays represented only 42% of the genes predicted in the genome.

Here we present a survey of gene expression variation in adult males of 16 strains from two natural populations of *D. melanogaster*. Polymorphism of the African and European populations used in this study has already been well characterized at the DNA level (GLINKA *et al.* 2003, OMETTO *et al.* 2005, see also CHAPTER 1). The amplicon-based microarrays we used cover almost the complete genome (88% of all predicted genes) and therefore provide a comprehensive and unbiased platform to study variation in gene expression. We show that our experiment has high statistical power to detect expression differences between strains. We contrast the levels of expression polymorphism between the two populations and compare our findings with previous results (MEIKLEJOHN *et al.* 2003) and the expectations derived from DNA data. Our extensive analyses include the effects of chromosomal location (X-linked *versus* autosomal) and functional diversity on the level of gene expression variation. Finally, our experimental design allows us to detect genes differing significantly in expression on a population level and thus reveals candidates for genes that have undergone adaptive regulatory evolution accompanying the out-of-Africa range expansion of the species. We present a list of interesting candidates and correlate our findings with candidate regions for positive selection in the European and African populations detected by studies using SNP data (OMETTO *et al.* 2005, LI and STEPHAN 2006).

MATERIALS AND METHODS

Drosophila lines: Flies are from the Dutch and Zimbabwean populations described in GLINKA *et al.* (2003). The eight highly-inbred lines per population used for the study were randomly chosen. The flies were kept on standard fly food at 22 °C and a 15h-9h light-dark cycle.

Microarray platform: The platform used was a genome-wide *D. melanogaster* microarray obtained from the Drosophila Genomics Resource Center (Bloomington, Indiana, USA) called DGRC-1. This microarray consists of 13,921 exonic PCR amplicons (100-600 bp in length) representing 11,895 unique genes, which is equivalent to 88% of the genome (based on genome annotation 4.1). Since the amplicons were based on an earlier annotation of the genome (namely 3.1), some genes are not present on the array according to updated annotations, while other genes are represented by more than one amplicon.

Experimental design: To asses the amount of expression differentiation between any given pair of lines, we developed a hybridization scheme that allowed us to compare all of these

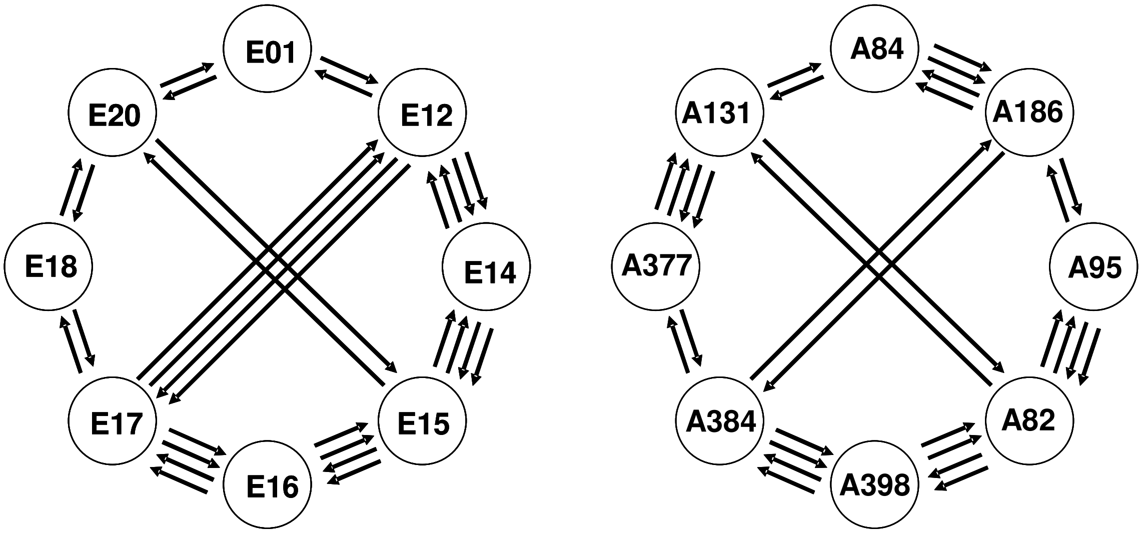


Figure 2.1: Comparison structure of the European (left) and the African (right) population. Arrowheads indicate red, arrow bases green labelled samples.

lines among each other while keeping the total number of hybridizations at an experimentally feasible level. The starting point was two circular designs with cross connections that connected only individuals within the two populations (Figure 2.1). Each pairwise comparison included a dye-swap to account for dye effects. Dye-swaps are indicated in Figure 2.1 as arrows; each arrowhead symbolizes a red labelled sample, each arrow base a green labelled sample. Additionally, inter-population comparisons were performed (Figure 2.2). The inclusion of these comparisons allowed us to investigate gene expression differences between lines of different populations and therefore obtain an estimate of population differentiation.

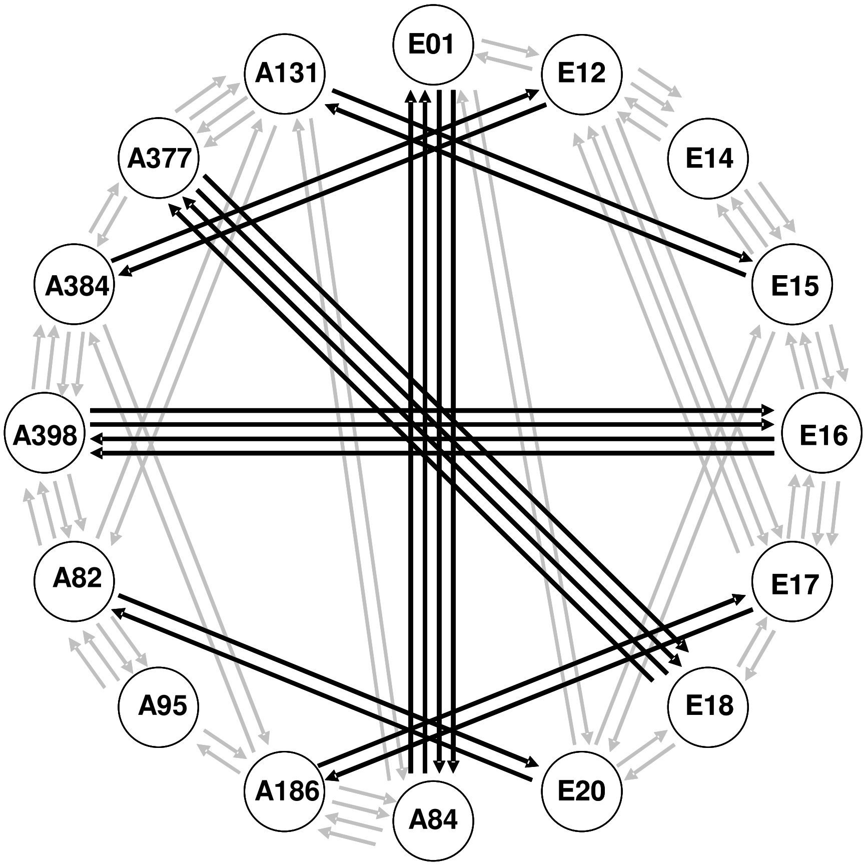


Figure 2.2: Comparison structure of the total experiment. Grey arrows are comparisons within populations (see also Figure 2.1); black arrows are comparisons between populations.

RNA extraction and hybridization procedure: We extracted RNA from 70-75 adult males that were 4-6 days of age using the recommended DGRC protocol (https://dgrc.cgb.indiana.edu/microarrays/support/protocols.html). Labelling and hybridization was performed using Invitrogen (Carlsbad, California, USA) Alexa Fluor chemistry following protocols provided by the DGRC. For each pairwise comparison dye-swaps were performed. In this case the same RNA was used for both hybridizations. Otherwise RNA was extracted independently for each pairwise comparison. Hybridized slides were scanned using an aQuire microarray scanner (Genetix, New Milton, UK) directly after hybridization.

Normalization of raw data: To obtain a normalized ratio of the red/green signal for each spot on our arrays we applied a three-step procedure that is implemented in CARMAweb (RAINER *et al.* 2006). This web service provides different methods that correct for (i) local background effects, (ii) within array variation, and (iii) between array variation. To make use of between array normalization, dye-swap arrays were normalized as pairs. The optimal methods for each of the three steps were obtained by extensive testing of our data set, resulting in the "minimum" method for background correction, the "printtiploess" method for within array correction and the "quantile" method for between array correction.

Relative expression levels per line and quality control: The normalized red/green ratios for each slide were used as input for BAGEL (TOWNSEND and HARTL 2002). This program uses a Markov Chain Monte Carlo based algorithm to estimate the relative expression levels of all lines for any given gene. The levels of gene expression are given relative to the line with the lowest level (meaning that the lowest value is always 1). Furthermore a test for differential expression for any pairwise comparison is performed.

In addition to a first analysis with BAGEL, we performed a second run where we removed spots which did not show a significant signal of expression and therefore did not present meaningful red/green ratios. A significant signal was defined on a per slide basis using the negative controls. We defined as negative controls those 182 spots on the array which consist of exogenic DNA (for example, genes amplified from yeast or *Escherichia coli*). For each array the distribution of the signals above background for these negative controls was determined separately for each channel. Subsequently, the signal intensity in each channel for each spot representing a gene was compared to the negative distribution. If the signal of the spot fell within the upper 5% of the negative distribution in each channel, the gene was considered to be "expressed". If a spot presented a signal that is lower than this threshold in either of the two channels, then we assumed that the red/green ratios for these spots do not provide meaningful information about the relative gene expression levels and they were excluded from further analysis. This approach automatically removes non-expressed genes from the BAGEL input and selectively eliminates spots of low quality that occur randomly on some arrays and may influence the analysis for genes that are expressed.

Quantification of expression diversity, false discovery rate and statistical power: As a measurement of expression diversity within a gene we used two statistics: the standard deviation (SD) of relative expression levels and the number of significant pairwise differences. In order to find a reasonable significance threshold for the tests for pairwise differences, we created a randomized input file for BAGEL. Randomization was performed by sampling with replacement within each hybridization (*i.e.*, randomizing within a column), therefore keeping the proportion of missing data per hybridization constant. The results of this randomized input were then used to set a reasonable significance threshold for pairwise differences by looking at the false discovery rate (FDR).

As a measurement of statistical power to detect expression differences, we calculated the GEL_{50} statistic (TOWNSEND 2004) for our BAGEL runs. The GEL_{50} statistic is defined as the expression difference at which there is a 50% chance of detecting this difference as being significant at the 5% level. To obtain this statistic, the following approach is applied (TOWNSEND 2004): All pairwise comparisons of differential gene expression performed by BAGEL are assigned either a 1 if they are significant or a 0 if they are non-significant at the 5% level. These zeros and ones are then plotted on a graph as a function of the expression difference (*i.e.*, the fold-change) which was tested (on a $\log_2$ scale). Afterwards a logistic function is fitted through these data points. The GEL_{50} statistic is then defined as the value (*i.e.*, the fold-change) at which the logistic function reaches 0.5.

Average expression levels and recombination rates: In order to quantify a gene's overall level of expression, the following approach was applied: For each slide the raw mean signal above background for the 48 replicates of *Act5c* (acting as positive controls) was calculated for each channel. This value was used as a slide specific reference. Subsequently, for every spot on the array the signal above background was divided by the reference signal. The result was used as a quantifier of gene expression, with values larger than 1 denoting expression levels higher than those of *Act5c*, and values smaller than 1 indicating expression levels lower than those of *Act5c.* Since this was done on all slides and for both channels separately, the mean of all 160 resulting estimates of gene expression was used for further analysis.

Recombination rates were assigned to each probe based on the cytological location of their corresponding gene. The local recombination rate of each cytological band was estimated by the program "recomb-rate" (COMERON *et al.* 1999). Levels of recombination are given as recombination rate per base pair per generation $\times 10^{-8}$.

Detection of differentially expressed genes between populations: To find genes that differ in expression between the African and the European populations, we reformatted our BAGEL input. We used only hybridizations where an African line was directly compared to a European line. This resulted in a total of 20 hybridizations (black arrows in Figure 2.2). In addition all African lines were combined to one single node named “Africa” and all European lines where combined to a node named “Europe”. Thus, with this approach the different lines used within each population can be considered as biological replicates. The result is a comparison structure that has a much higher power to detect differences, since only two nodes are compared to each other with a relatively high number of hybridizations. We performed two separate BAGEL runs: one including all red/green ratios and one where uninformative (non-expressed) spots were excluded. The procedure for removing uninformative spots was the same as described above.

As for the first BAGEL analysis, a randomized data set was created to estimate the FDR and choose an appropriate significance threshold. In contrast to the previous analysis, we did not randomize within each hybridization (*i.e.*, column), but within each gene (*i.e.*, row). By doing this, we ensured that the proportion of missing data (*i.e.*, data points removed by the quality control step) remained constant for each gene.

Gene ontology: A list of all known gene ontology (GO) terms describing molecular functions and biological processes associated with the probes on the microarrays was downloaded directly from the DGRC. Of the 13,921 probes representing a gene, at least one biological process was known for 8,251 and at least one molecular function was annotated for 8,523. We calculated the number of unique GO terms describing molecular functions and biological processes associated with each probe to get an estimate of its “functional diversity”.

Candidate regions for positive selection on the X chromosome: We investigated the expression patterns of genes lying in candidate regions of positive selection on the X chromosome as defined by two different studies. For the European population we looked at the six regions found to be candidates by the method II^{all} of OMETTO *et al.* (2005). Additionally, the data of LI and STEPHAN (2006) were investigated. For this study, the regions defined as candidates were those where the selection model performed better than the neutral model for at least three estimates of selection strength (Table S3 of LI and STEPHAN 2006). The same approach was applied for finding putative candidate regions for selection in the African population (Table S4 of LI and STEPHAN 2006).

RESULTS

Statistical power: We calculated the GEL_{50} statistic (TOWNSEND 2004) to compare the statistical power of our experimental design to earlier works (*e.g.* MEIKLEJOHN *et al.* 2003). The result of the logistic regression is plotted in Figure 2.3a. It can be seen that small expression differences are often found to be non-significant (visible as zeroes for small fold-changes) while large fold-changes are often significant at the 5% level (visible as ones for large fold-changes). The logistic regression reaches a value of 0.5 at a $\log_2$ fold-change of 0.945 (dashed line in Figure 2.3a). This corresponds to a GEL_{50} of 1.93 ($2^{0.945} = 1.93$), which means that in our experimental design we have a 50% chance of detecting 1.93-fold expression differences as significant at the 5% level. This value is larger than the GEL_{50} of the MEIKLEJOHN *et al.* (2003) data set ($GEL_{50} = 1.64$, CLARK and TOWNSEND 2007). Therefore our experiment seems to have a lower resolution, which is surprising since the ratio of hybridizations to nodes (80/16) in our experiment should provide a better resolution than the MEIKLEJOHN *et al.* (2003) experiment (hybridizations/nodes = 23/8). The obvious explanation for this is the lack of quality control in our data. All spots on the microarrays were included in

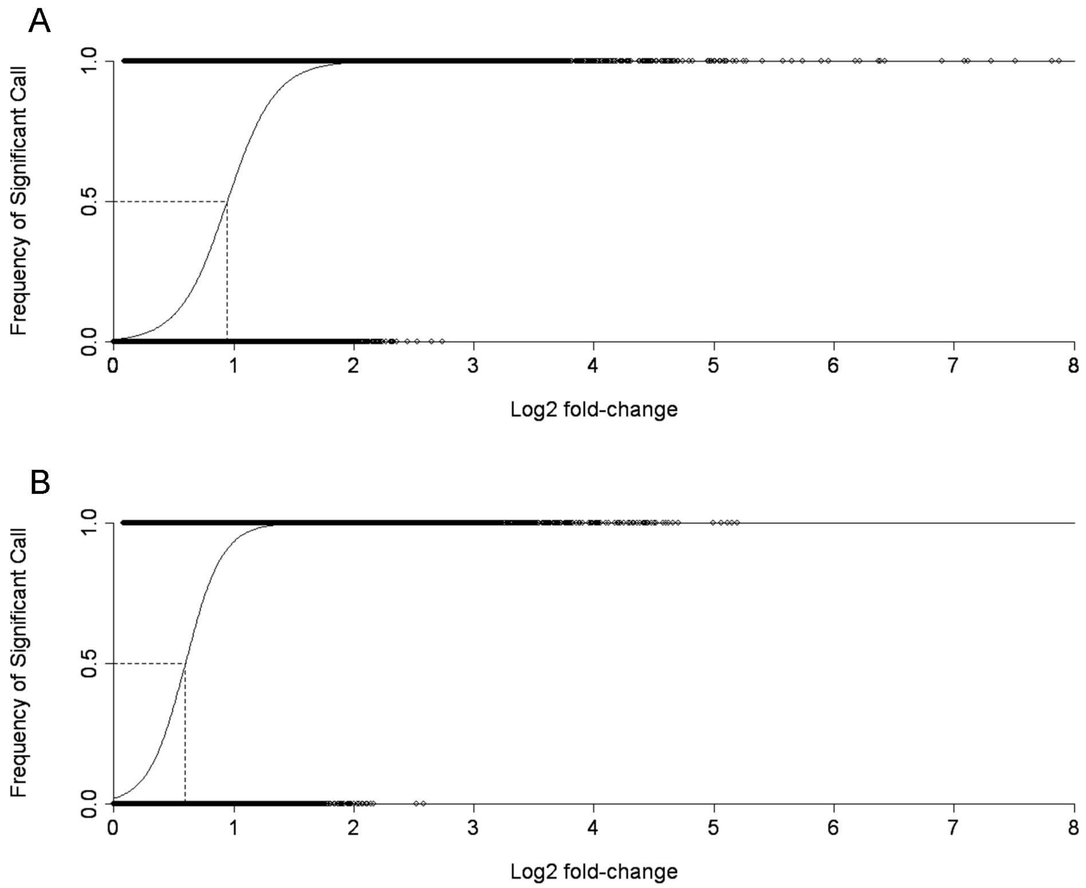


Figure 2.3: Logistic regression of the probability of detecting gene expression differences at the $P < 0.05$ level. (a) the regression of the BAGEL run with all data included, (b) the regression of the BAGEL run after quality control. The dashed line defines the GEL_{50} value on a $\log_2$ scale.

our analysis, regardless of their signal quality. This resulted in an increased variance of the estimation of gene expression levels which consequently led to a reduction of power to detect differences between lines.

In order to eliminate the effect of bad quality spots, we removed all spots showing poor hybridization intensity from the analysis (see MATERIALS AND METHODS). This step also automatically removes genes with very low or no expression from the data, since these genes will not present enough high quality signal to obtain expression estimates for all the lines studied. As a result, the data set was reduced from 13,921 probes representing 11,895 unique genes to 5,048 probes representing 4,512 unique genes. Therefore, 37.9% of all genes on the arrays can be detected as significantly expressed in all 16 of our adult male lines. We recalculated the GEL_{50} value for our "high-quality" data set (Figure 2.3b). The GEL_{50} statistic dropped to 1.51 indicating that our quality control improved the power to detect differences of expression. This high-quality data set is used for all further analyses. We also calculated the GEL_{50} values separately for detecting differences within or between populations for the high-quality data set. The GEL_{50} was 1.512 for within Europe, 1.508 within Africa and 1.513 between populations. So the resolution to detect differences in any of these three comparisons is approximately equal. This confirms that our experimental design is well balanced and does not have any biases in detecting differential expression within or between populations.

Gene expression levels: In order to estimate the overall expression level of each gene, we estimated the average signal intensity for each probe relative to the control gene *Act5c*. It should be noted that the signal intensities on the microarray do not necessarily correlate perfectly with the level of mRNA abundance, since the efficiency of hybridization might differ between cDNAs. Nevertheless it can be used as a crude estimator. The median signal for all genes in the data set is 0.105 relative to *Act5c* with a standard deviation (SD) of 0.708. Figure 2.4 shows the distribution of the logarithm of relative signal intensities. It suggests that there is a skew towards relatively low expression levels with a distinct tail of genes with higher expression. This is expected, as *Act5c* is known to be one the genes with the highest expression levels in *D. melanogaster*. We used the relative signal intensity to check if expression levels differed for genes located on the autosomes and the X chromosome, but no significant difference was found (Mann-Whitney *U* test, $P = 0.215$).

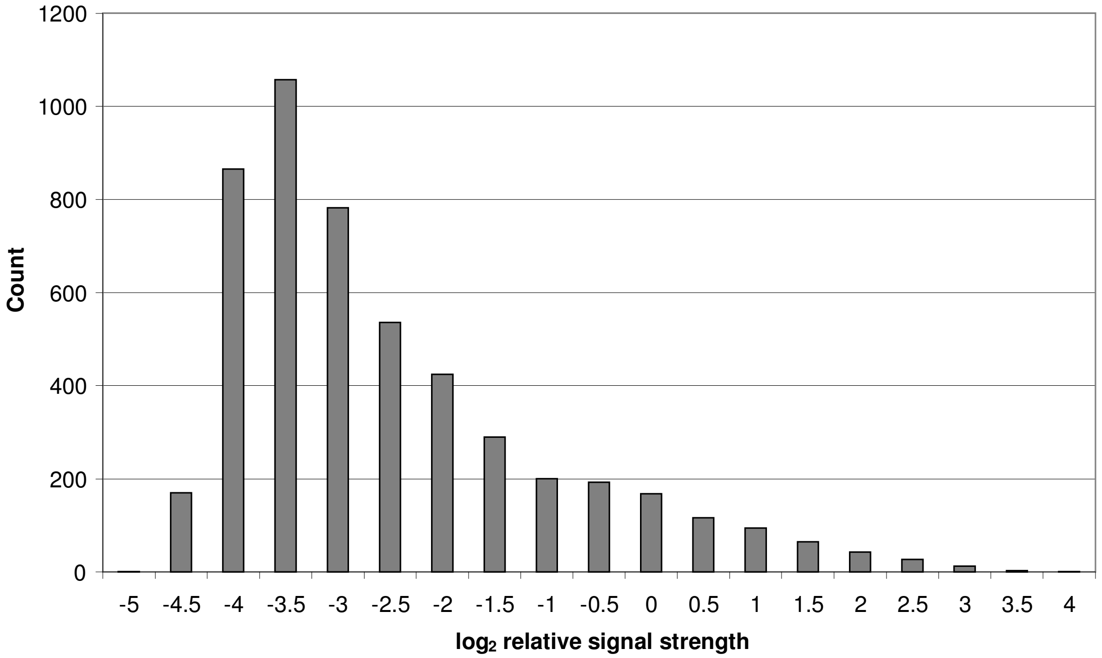


Figure 2.4: Histogram of the average expression levels of all probes relative to *Act5c* on a $\log_2$ scale.

Total number of differentially expressed genes: Since the number of tests for pairwise differences was extremely high (5,048 probes × 120 pairwise comparisons = 605,760 tests), we could not operate with the standard 5% significance level due to the problem of multiple testing. We therefore created randomized data sets to estimate the FDR at any given significance level. Table 2.1 shows the number of tests which produce a significant result for differences in expression along with the numbers expected by chance for a range of significance levels. We can see that the percentages of significant tests found in the randomized data set correspond relatively well to the preset *P* values. This is an indication that our randomization procedure worked properly. As expected, using a *P* value of 0.05 would lead to a very large number of false positives in our original data. We therefore decided to use a *P* value of 0.001 which corresponds to a FDR of 6.9%. This is close to the FDR of

Table 2.1: False discovery rates for different levels of significance

P value	$\mathrm{Sig}_{\mathrm{data}}{}^{a}$ (%)	$\mathrm{Sig}_{\mathrm{rnd}}{}^{b}$ (%)	FDR
0.05	110,285 (18.21%)	54,105 (8.93%)	49.06%
0.01	44,081 (7.28%)	10,301 (1.70%)	23.37%
0.005	31,670 (5.29%)	5,249 (0.87%)	16.57%
0.001	16,564 (2.73%)	1,147 (0.19%)	6.92%
0.0005	13,219 (2.18%)	622 (0.10%)	4.71%

[a] Number of tests significant in the original data set
[b] Number of tests significant in the randomized data set

5.2% used in the study of MEIKLEJOHN *et al.* (2003), and allows us to make unbiased comparisons.

Using this cut-off we find that 1,894 probes show significant differences for at least one pairwise comparison (Table 2.2). This estimate of 37.5% of all probes showing expression polymorphism is close to that reported previously (MEIKLEJOHN *et al.* 2003). Since 413 genes are represented by multiple probes in our data set, we checked if the percentage of polymorphic genes corresponded to the number we find when comparing the probes. If we define a gene as polymorphic if at least one of its probes shows a pairwise difference, then we find that 38.9% of all expressed genes are polymorphic. If we apply a stricter criterion and only consider those genes as polymorphic where all probes show significant differences this number drops to 35.1%. Since the overall effect of including multiple probes per gene is rather small, we will present the results on a "per-probe" basis throughout this paper.

A total of 964 probes (19.1%) showed differences within the European population, 1,039 probes (20.6%) showed differences within the African population and 1,600 probes (31.7%) showed differences when comparing European to African lines (inter-population comparison). The elevated number of probes in the inter-population comparison is somewhat

Table 2.2: Expression polymorphism by population

	Polymorphic genes		
	Total number (%)	Mean per PW (SD)[a]	Mean Δ per probe (%)
Overall	1894 (37.5%)	138.0 (53.0)	3.28 (2.73%)
Europe	964 (19.1%)	126.5 (43.7)	0.702 (2.51%)
Africa	1039 (20.6%)	125.9 (47.8)	0.698 (2.49%)
Between	1600 (31.7%)	148.4 (57.3)	1.88 (2.94%)

[a] Average number of probes found to be differentially expressed for each pairwise (PW) comparison between all lines within the corresponding data set

expected, since it has a higher number of pairwise tests than within population comparisons (64 as opposed to 28).

Expression differences between individual lines: We investigated the number of differentially expressed probes individually for each pairwise comparison. Table 2.3 shows the results of these comparisons. The number of differentially expressed probes is given below the diagonal, while the numbers expected at random are given above the diagonal. On average 138 probes showed differential expression for each individual pairwise comparison. Given the overall number of 1,894 probes that show differences, this number is surprisingly small, even more so when taking into account that the MEIKLEJOHN *et al.* (2003) data set presented an average of 498 genes differentially expressed in each pairwise comparison with a total number of 2,289 differentially expressed genes. It shows that in our data set there is not much overlap in the lists of differentially expressed genes for the 120 pairwise comparisons. This effect is also visible when comparing the number of pairwise differences detected for each probe. The histogram (Figure 2.5) shows that a large fraction of probes only show significant differences for one or two out of the 120 pairwise comparisons.

Table 2.3: Number of probes detected as differentially expressed between lines

	E01	*E12*	*E14*	*E15*	*E16*	*E17*	*E18*	*E20*	*A82*	*A84*	*A95*	*A131*	*A186*	*A377*	*A384*	*A398*
E01		9	2	7	6	11	7	6	10	9	13	8	17	12	14	5
E12	168		5	10	7	7	4	8	18	12	14	13	7	10	16	11
E14	74	151		8	7	3	4	2	3	10	9	9	6	8	3	5
E15	93	145	137		8	6	7	9	9	6	14	19	11	8	7	11
E16	99	111	92	76		4	3	7	9	9	13	7	5	5	10	16
E17	80	255	114	151	221		5	4	12	9	15	6	6	6	9	10
E18	91	99	92	96	98	94		5	9	7	12	4	6	8	5	5
E20	139	156	106	174	145	117	168		9	19	25	9	12	10	8	16
A82	131	164	109	92	142	148	104	280		16	25	20	11	11	9	7
A84	180	132	108	79	97	110	79	154	72		23	10	15	10	14	9
A95	216	220	153	112	168	299	165	322	180	127		19	23	13	13	13
A131	109	121	95	42	98	98	129	150	133	80	188		8	6	12	16
A186	118	147	93	83	110	52	105	165	89	167	192	105		6	6	8
A377	126	180	131	105	139	120	229	188	97	78	178	116	88		7	4
A384	128	228	123	135	197	160	148	187	148	112	240	105	157	102		4
A398	180	222	161	145	275	157	110	245	54	66	200	93	109	84	164	

Numbers below the diagonal come from the original data set, numbers above the diagonal are form the randomization.

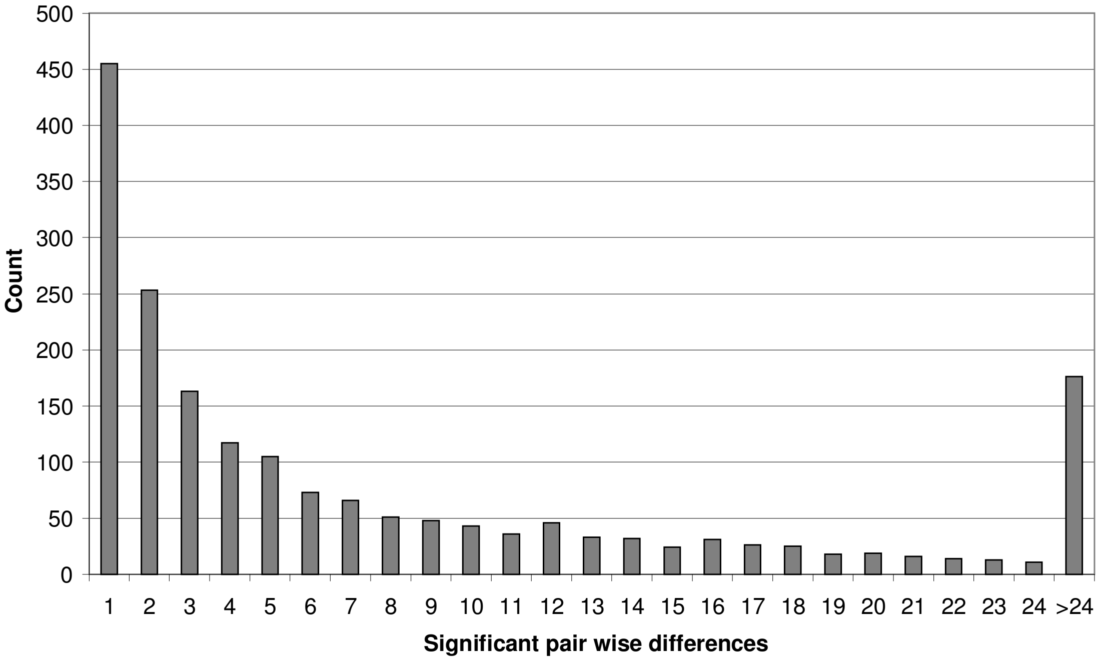


Figure 2.5: Histogram of the number of significant pairwise differences for all expressed probes.

Expanding this approach to investigate differences within and between populations we see a pattern that mimics our findings for the total number of differentially expressed probes. Comparisons between two European lines show on average differences in 126.5 probes, when two African lines are compared differences can be found in 125.9 probes and when a European line is compared to an African line on average 148.4 probes are significantly different (Table 2.2). Since these numbers are independent from the number of pairwise comparisons, it can be said that there is an excess of significantly different probes in the inter-population comparisons. This approach also produced some surprising results. The two lines which showed the most similarity in terms of their expression profile did not come from within a single population. The European line *E15* and the African line *A131* showed only 42 differentially expressed probes (Table 2.3). On the other hand, the two most divergent lines do come from different populations, with the European line *E20* and the African line *A95* differing in expression for 322 probes (Table 2.3). This wide range in the probe number difference is also visible in the distribution of the number of differentially expressed probes. For the inter-population comparisons the SD of differentially expressed probes is 57.3, for comparisons within the African population it is 47.8 and only 43.7 for the within Europe comparisons (Table 2.2).

Measurements of expression polymorphism: To get an estimate of the overall level of expression polymorphism within a gene we looked at two statistics: (1) the number of significant pairwise comparisons per probe (Figure 2.5) and (2) the SD of the relative expression level per line. Genes that present a high level of expression polymorphism should have a large number of significant pairwise differences as well as a large SD. When we correlate these two measurements we indeed find that they are positively correlated (Spearman correlation coefficient $R = 0.233$, $P < 0.001$). Even though this correlation is significant, it is surprisingly weak. One explanation may be the effect of the gene expression level (estimated by signal strength, see above) on these two statistics. We find that there is a strong negative correlation between the SD and the mean expression level (Spearman correlation coefficient $R = -0.529$, $P < 0.001$). This means that genes that are lowly expressed generally show a large SD of relative gene expression between lines. This is understandable, since lowly expressed genes have a low signal on the microarrays and are therefore more prone to experimental variance (*e.g.* background noise on the arrays). On the other hand we find a strong positive correlation between mean expression levels and the number of significant pairwise differences per line (Spearman correlation coefficient $R = 0.439$, $P < 0.001$). This is again an expected result. Since experimental variance such as background noise has only little influence on the signals of highly expressed genes, the confidence intervals of expression estimates will be relatively small. As a consequence, it will be easier to find significant differences between two lines.

Of the two above statistics, the number of significance differences (which we will call Δ) seems to be a more meaningful indicator of the level of gene expression polymorphism than SD. Experimental noise is bound to generate many false positives when using SD as a measurement, especially in the class of lowly expressed genes. Since the vast majority of probes have expression levels (or signal intensities) which are relatively low (Figure 2.4), this effect may have a big influence on our analysis. The statistic Δ on the other hand can be viewed as conservative. While it may fail to detect differences within lowly expressed genes, the differences that are being detected most likely reflect "true" differences in expression which are also biologically meaningful. We therefore use Δ as a measurement of expression polymorphism within a gene for further analysis.

We see that Δ follows the pattern we see for the number of differentially expressed genes with the European and African population (Table 2.2). Probes have a similar level of polymorphism in African (0.698) and Europe (0.702) and a Mann-Whitney *U* test of the two populations is not significant ($P = 0.086$). The between population comparisons show a larger number of significant tests (1.88), but this high number is partly due to the increased number of tests in this data set (64 pairwise comparisons between populations, 28 comparisons within populations). However, when we normalize the numbers by dividing by the number of tests (visible as percentages in Table 2.2), the between population comparisons still present the highest number of significant pairwise tests (Mann-Whitney *U* test, $P < 0.001$).

The magnitude of expression differences: Not only were we interested in the number of probes that showed differential expression, but also in the magnitude of these differences. Of all 605,760 pairwise tests for expression differences a total of 16,564 were significant at the 0.001 level. Figure 2.6 shows a histogram of the relative fold-changes of these differences. The median fold-change of significant differences is 1.74. The smallest change that was detected as being significant was a 1.11-fold change, the largest was a change of over 36-fold.

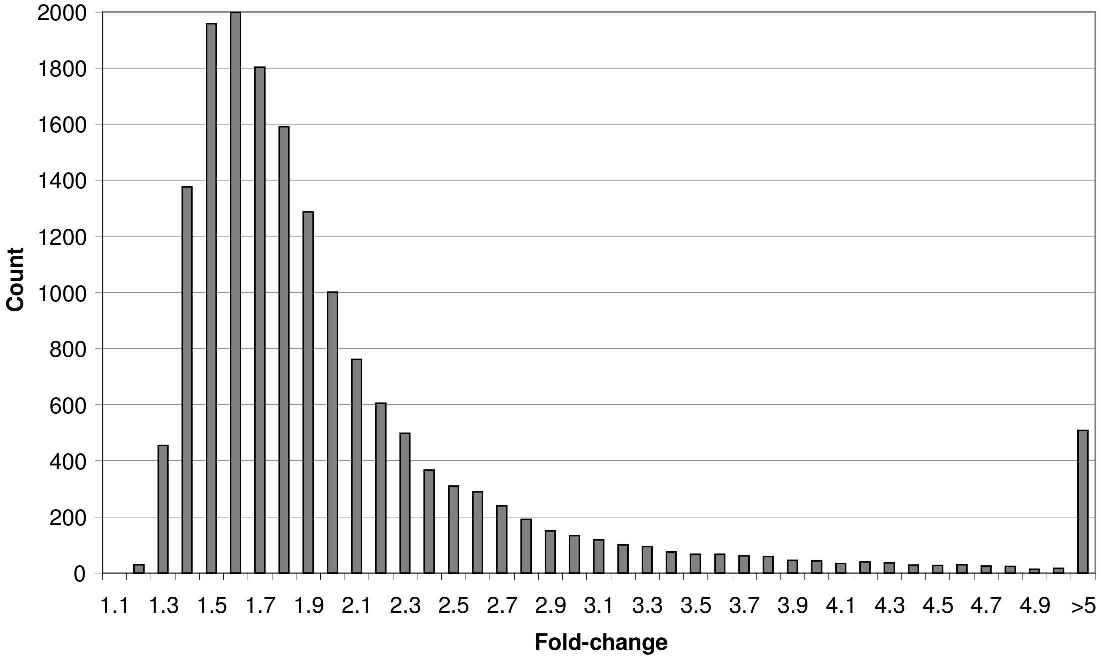


Figure 2.6: Histogram of the fold-changes in expression for comparisons significant at the *P* < 0.001 level.

As can be seen in the figure, the majority of changes are relatively small being in between 1.2 and 2-fold.

The effect of recombination rate and chromosomal location on expression differences: We tested if there was an influence of recombination rate on expression polymorphism similar to the effect seen for DNA polymorphism (*e.g.* BEGUN and AQUADRO 1992, GLINKA *et al.* 2003). We grouped our probes in bins of high (2,331 probes in total) and low (2,717 probes in total) recombination rate. The boundary value between these bins was a recombination rate of 0.0025 which is the median value of our data set. We then used a Fisher Exact test to see if the proportion of genes that showed expression polymorphism differed between these two bins. The result turned out to be non-significant (two-tailed test, *P* = 0.60).

Table 2.4: Expression polymorphism on the X chromosome and autosomes

	X chromosome	Autosomes	X/A ratio[a]
Number and percentage of polymorphic genes			
Overall	335 (35.8%)	1559 (37.9%)	0.945 (*P* = 0.22)
Europe	155 (16.5%)	809 (19.7%)	0.838 (*P* = 0.027)
Africa	168 (17.9%)	871 (21.2%)	0.844 (*P* = 0.025)
Between	277 (29.6%)	1323 (32.2%)	0.919 (*P* = 0.12)
Average number of pairwise differences per probe			
Overall	2.425	3.477	0.697 (*P* = 0.040)
Europe	0.495	0.749	0.661 (*P* = 0.014)
Africa	0.521	0.739	0.705 (*P* = 0.017)
Between	1.409	1.989	0.708 (*P* = 0.035)

[a] Deviations from 1:1 expectations for the X/A ratios were tested with two-tailed Fisher's Exact tests for the percentage of polymorphic genes and with Mann-Whitney *U* tests for the average number of pairwise differences

We examined if the proportion of polymorphic genes differed between the autosomes and the sex chromosome. We did this analysis with both populations combined, but also separately for the European and the African population. In addition, we examined the average number of significant pairwise differences per gene to include information about the magnitude of diversity within a polymorphic gene. The results are summarized in Table 2.4. If we examine the complete data set we find that the percentage of polymorphic genes is slighter higher on the autosomes (37.9%) than on the X chromosome (35.8%). This difference is, however, not significant (Fisher’s Exact test, *P* = 0.22). The picture changes if we consider the two populations separately. The X chromosome harbours proportionally fewer polymorphic genes than the autosomes in both the European and the African population (X/A ratio of 0.838 in Europe and 0.844 in Africa) and both ratios deviate significantly from the 1:1

expectation (Table 2.4). When considering the number of probes that differ between populations in at least one pairwise comparison we see that the X chromosome also carries fewer of these probes compared to the autosomes, but this difference is not significant (Two-tailed Fisher's Exact test, $P = 0.12$).

The average number of pairwise differences (Δ) per probe behaved differently. We found that genes located on autosomes show on average more pairwise differences (3.477) than X-linked genes (2.425) when the data set is considered as a whole. This means that if a gene is found to be polymorphic, it tends to exhibit a larger amount of variation if it is located on an autosome. This pattern was also observed when considering the European and the African population separately, as well as comparing differences between populations. The ratio of the average pairwise differences between X chromosomes and autosomes was less than one (0.661 in Europe, 0.705 in Africa, 0.708 between populations) and was always significant (confirmed by means of Mann-Whitney *U* tests, Table 2.4).

The effect of gene function on expression differences: For a sizable fraction of our data set, the biological processes and/or molecular functions of the genes were known. Of the 5,048 probes we found to be expressed, biological processes were known for 3,217 probes, and 3,275 probes had at least one known molecular function. Some of the probes were associated with more than one GO term. Regarding biological processes, the most extreme case was the gene *Egrf,* which was involved in 62 different biological processes. For the number of different molecular functions associated with a gene, *ninaC* presented the most extreme case. Eleven different molecular functions were associated with this gene. We wanted to know if the number of different processes and functions had an influence on the gene expression diversity of the genes. Our hypothesis was, that if a gene was involved in many different biological processes (or had many different molecular functions) it would be under more constraint than genes associated with only few biological processes (or molecular functions)

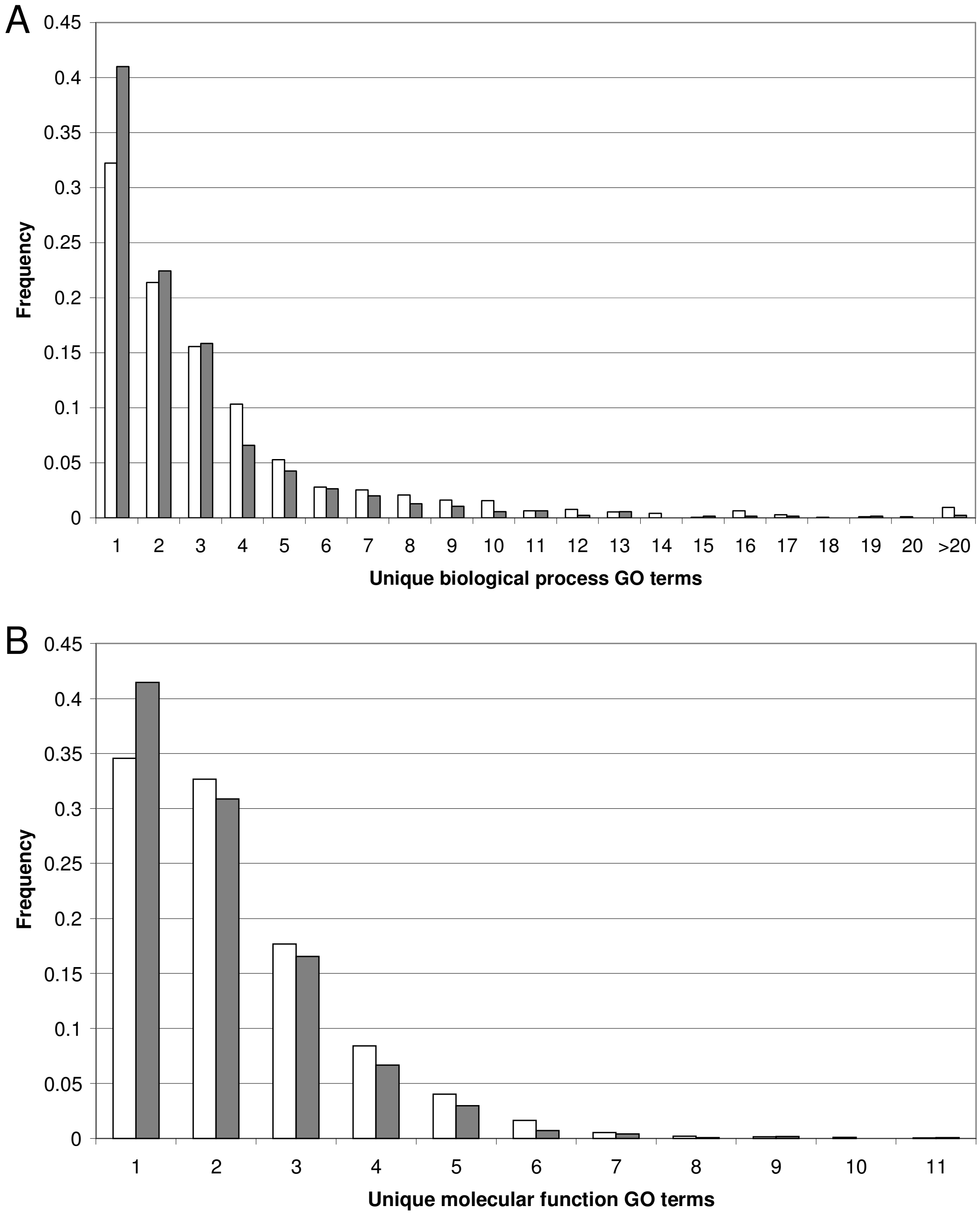


Figure 2.7: Histogram of the number of unique GO terms associated with monomorphic probes (white) and polymorphic probes (grey). (a) GO terms related to biological processes, and (b) GO terms related to molecular functions.

and therefore be less polymorphic. To test this hypothesis we examined the number of GO terms for probes we found to be polymorphic in expression and compared them to probes which were monomorphic. A comparative histogram is shown for biological processes (Figure 2.7a) and molecular functions (Figure 2.7b). In Figure 2.7a it can be seen that in the polymorphic data set probes with a low number of processes (three or less) are overrepresented, while in the monomorphic data set probes associated with four or more processes are comparatively more prevalent. A Mann-Whitney *U* test confirms that polymorphic probes are associated with fewer GO terms than monomorphic probes (*P* < 0.001). A histogram of the number of different molecular functions show the same trend (Figure 2.7b). Here the Mann-Whitney *U* test also finds that polymorphic probes have fewer molecular functions than monomorphic probes (*P* < 0.001).

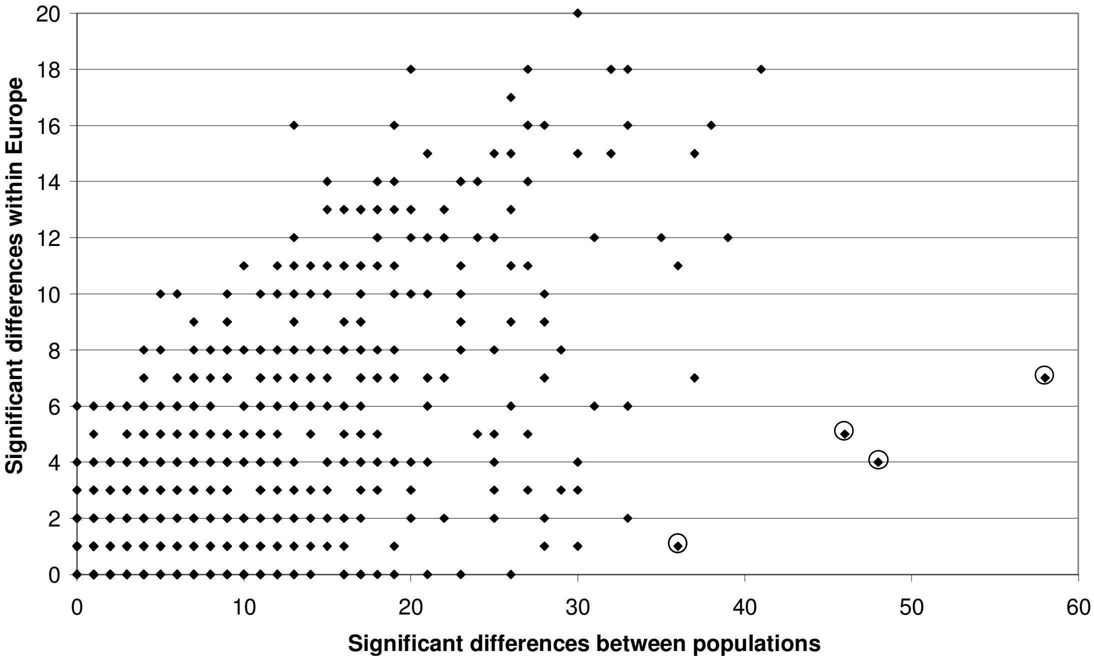


Figure 2.8: Number of significant pairwise differences between populations (X axis) plotted against the number of significant pairwise differences within Europe (Y axis) for each probe. Probes that are putative candidates for adaptation are marked with circles.

Expression differences between populations: In order to find probes that show differences in expression on a population scale (and are therefore candidates for adaptation), we examined the significant pairwise differences between European and African lines. If a probe shows a distinctly different expression pattern between the two populations, then all 64 inter-population pairwise comparisons (eight European lines × eight African lines) should turn out to be significantly different. Figure 2.8 shows the number of significant inter-population comparisons plotted against the number of significant pairwise comparisons within the European population. It can be seen that the maximum number of significant inter-population comparisons is 60 and only few probes show a large number (> 32) of differences between both populations. We find that the number of pairwise differences between populations and within Europe are positively correlated (Spearman correlation coefficient $R = 0.669$, $P <$

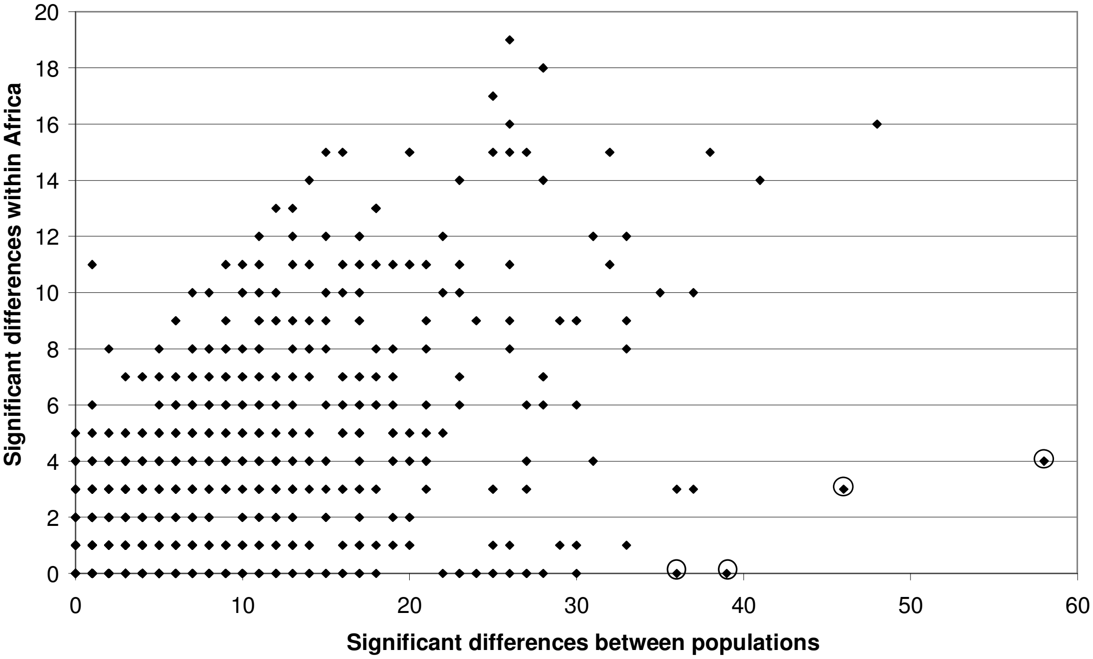


Figure 2.8: Number of significant pairwise differences between populations (X axis) plotted against the number of significant pairwise differences within Africa (Y axis) for each probe. Probes that are putative candidates for adaptation are marked with circles.

0.001), so probes that are highly differentiated between populations also show large variation within the European population. The same trend can be observed when plotting inter-population differences against differences within the African population (Figure 2.9). Here both statistics are also positively correlated (Spearman correlation coefficient $R = 0.668$, $P < 0.001$). Genes that have undergone adaptive changes to their specific environment through directional selection should not only show distinct differences between populations, but also show reduced variation within the population in which selection operated. This pattern can be observed for some of the probes in our data set. These data points are marked by circles in Figures 8 and 9. The genes represented by these probes should make good candidates for those that have undergone adaptive regulatory evolution.

The above definition of candidate genes for adaptation is unsatisfying since it requires thresholds to be set for the minimum number of differences between populations and the maximum number of differences within a population, and the assignment of such thresholds is somewhat arbitrary. We applied a different approach to find distinctly differentially expressed probes between populations by pooling all lines of each population into one node and then using BAGEL to find differences between the African and the European node (see MATERIALS AND METHODS for details). With this approach, BAGEL will estimate the average expression level for each population and test for significant differences. Since the polymorphism within a population will affect the variance of this estimate, only those differences will be detected as significant where the within population variation is small compared to the between population difference. This new comparison scheme should also be much more powerful to detect differences since it has only two nodes to compare with 20 hybridizations. Figure 2.10a shows a plot of the logistic regression performed to obtain the GEL_{50} value for this experiment. The GEL_{50} is 1.33, which, as expected, is lower (*i.e.* better) than in our original 16-node experiment without quality control. When we removed non-detectable signals, we obtained a surprising result: The GEL_{50} statistic increased to 1.41

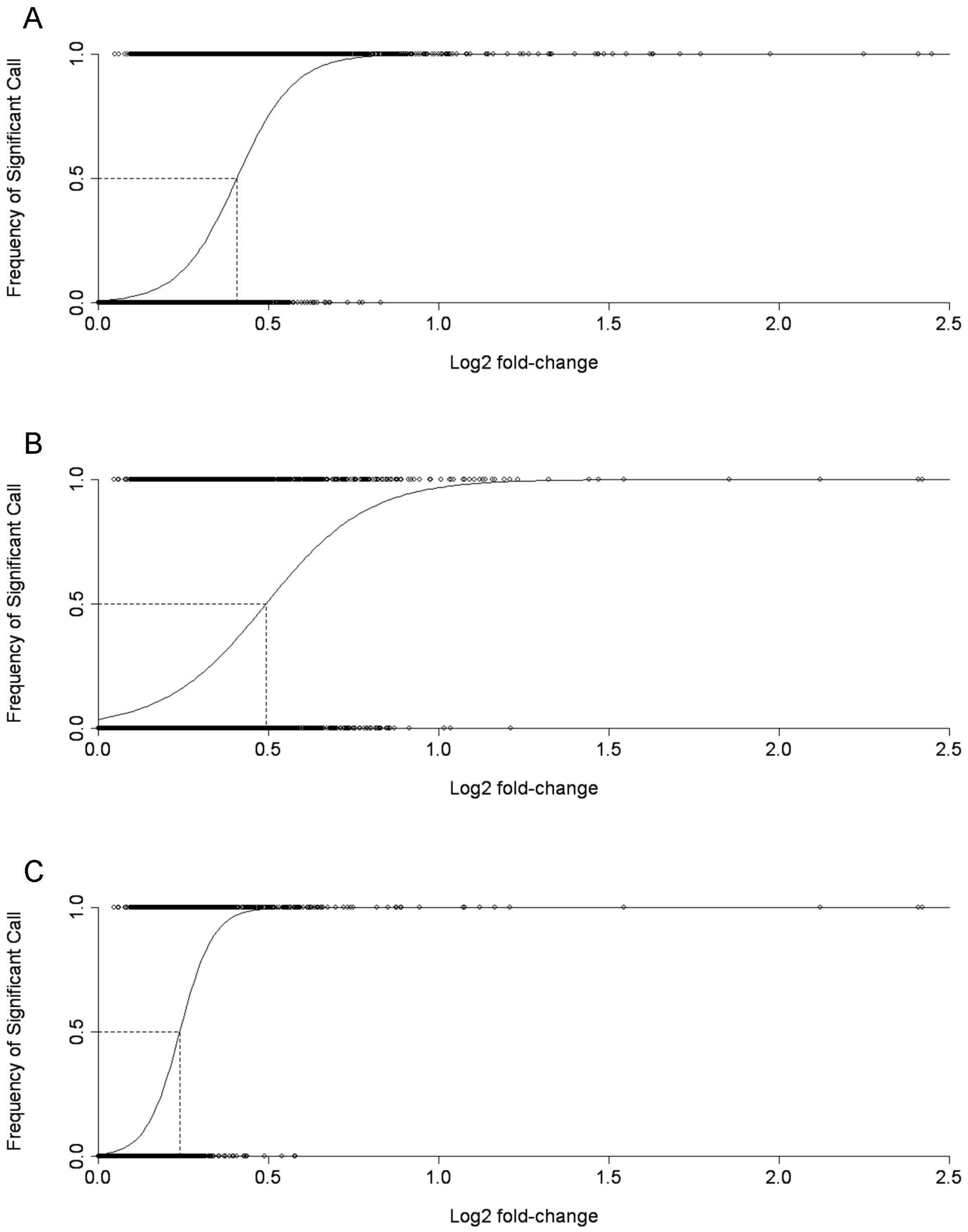


Figure 2.10: Logistic regression of the probability of detecting gene expression differences at the $P < 0.05$ level for the two-node experiment. (a) the regression of the BAGEL run with all data included, (b) the regression of the BAGEL run after quality control, and (c) the quality controlled data set with at least nine data points. The dashed line defines the GEL_{50} value on a $\log_2$ scale.

(Figure 2.10b). This might be due to the lack of data for some probes. BAGEL is able to perform calculations with as few as three hybridizations. If our quality control step removed many non-detectable signals from the input of a probe, then the few remaining signals could lead to a biased estimate of large expression differences for some probes. This is simply the effect of an increased variance due to a low number of data points. In addition these relatively large differences will be not detected as significant, since the low number of data points reduces the power to do so. These probes are visible in Figure 2.10b as the large number of points on the X-axis (meaning not detected as significantly different) with relatively high fold-changes. We confirmed that these probes are actually the ones with very few data points (data not shown). To eliminate this effect we defined a threshold of nine data points as the minimum number of data used as input for BAGEL. The result was that the number of probes analyzed by BAGEL dropped from 9,396 (probes with three or more data points) to 5,087 (probes with nine or more data points). The threshold of nine was chosen so that the resulting number of probes was approximately that of our previous experiment (5,048 probes for the 16 node experiment). The logistic regression of the new "high quality" data set (Figure 2.10c) showed the desired effect: The GEL_{50} statistic dropped to 1.18. This data set was used for further analysis.

As with the first experiment, we used a randomized data set to calculate the FDR and adjust our *P* value for differential expression. We chose a *P* value cut-off of 0.002, which leads to an FDR of 8.6% in our new experiment and corresponds well to the FDR of our 16 node experiment. At this significance level, 161 probes (3.2%) were found to be different between Europe and Africa. Again the magnitude of expression differences was relatively low, with the median value of fold-change being 1.32 and the maximum being 5.36. Interestingly, over-expression seems to be more common in the African population. Of the 161 differentially expressed probes, 85 (52.8%) are expressed at a higher level in Africa while only 76 probes (47.2%) are over-expressed in Europe. In addition the magnitude of the expression difference is larger for probes over-expressed in Africa (median fold-change: 1.35) than for probes over-expressed in Europe (median fold-change: 1.27) and this difference is significant (Mann-Whitney *U* test, $P = 0.044$). We investigated the chromosomal distribution of the differentially expressed probes and found that neither the X chromosome nor the autosomes were enriched for these probes (Fisher's Exact Test, $P = 0.83$).

Candidate genes of adaptation: We compiled a list of the ten probes which had the largest over-expression in the European population (Table 2.5). Also shown is the number of significant pairwise differences within the European and African population as well as the differences between populations, as detected by the original 16 node experiment. Table 2.6

Table 2.5: Top 10 candidate probes with over-expression in Europe

Probe ID	Gene name	Chromosome	$\mathrm{Exp_E} / \mathrm{Exp_A}$[a]	*P* value	Δ_E[b]	Δ_A[b]	Δ_B[b]
INC118A02	*Cyp6g1*	2R	4.35	$P < 0.0001$	4	16	48
INC067C10	*CG9509*	X	2.31	$P < 0.0001$	3	1	29
INC071G07	*CG32919*	3R	1.85	$P = 0.0002$	2	0	25
INC022G08	*dpr15*	3R	1.80	$P = 0.0001$	4	5	17
INC075G06	*Men*	3R	1.76	$P = 0.0001$	5	3	46
INC111G04	*Obp56d*	2R	1.68	$P < 0.0001$	12	10	35
INC149B06	*CG15036*	X	1.59	$P = 0.0001$	1	2	6
INC042C06	*α-Man-I*	X	1.56	$P = 0.0001$	3	1	25
INC031G12	*CG18135*	3L	1.52	$P = 0.0003$	4	11	20
INC157E04	*CG13183*	2R	1.51	$P = 0.0019$	11	15	26

[a] $\mathrm{Exp_E} / \mathrm{Exp_A}$ quantifies the fold-change of expression

[b] Δ_E, Δ_A, and Δ_B are the number of pairwise differences within Europe, within Africa and between populations detected in the 16 node experiment.

shows the top ten probes which are significantly over-expressed in Africa. It is interesting to note that all probes chosen to be candidates from the original 16 node experiment (circled data points in Figures 8 and 9) also show up in these two tables. The list of genes over-expressed in Europe contains some genes of unknown function, but also genes which have been well characterized such as the cytochrome P450 gene *Cyp6g1*, the odorant-binding protein *Obp56d* or the malic enzyme gene *Men* (see DISCUSSION for details). The top-ten list of genes over-expressed in Africa shows an interesting pattern. Three of these genes (*CG7214*, *Act88F* and *TpnC41C*) are involved in the morphogenesis of the wings or the function and formation of the indirect flight musculature (see DISCUSSION).

Table 2.6: Top 10 candidate probes with over-expression in Africa

Probe ID	Gene name	Chromosome	Exp_A / Exp_E[a]	*P* value	Δ_E[b]	Δ_A[b]	Δ_B[b]
INC010H12	*CG7214*	2L	5.36	$P < 0.0001$	1	0	36
INC012F04	*CG7203*	2L	5.31	$P < 0.0001$	7	4	58
INC115H09	*Act88F*	3R	2.92	$P < 0.0001$	7	3	37
INC016E09	*CG3301*	3R	2.24	$P < 0.0001$	12	0	39
INC125A09	*TpnC41C*	2R	2.11	$P = 0.0001$	2	0	16
INC107F03	*fit*	3R	1,92	$P < 0.0001$	7	7	28
INC026C04	*Nplp3*	3L	1.85	$P = 0.0017$	2	0	28
INC155G06	*CG8661*	X	1.84	$P = 0.0007$	12	9	24
INC057A05	*Mipp1*	3L	1.83	$P = 0.0002$	0	7	23
INC044H01	*CG8997*	2L	1.67	$P < 0.0001$	14	0	24

[a] Exp_A / Exp_E quantifies the fold-change of expression

[b] Δ_E, Δ_A, and Δ_B are the number of pairwise differences within Europe, within Africa and between populations detected in the 16 node experiment.

Two studies (OMETTO *et al.* 2005, LI and STEPHAN 2006) have used DNA polymorphism data to find regions on the European X chromosome that were target of recent positive selection (candidates for selective sweeps). In addition the LI and STEPHAN (2006) study characterized such regions of putative positive selection in Africa. We looked if any of our differentially expressed genes fall within the predicted regions of both studies. We find that six probes (out of a total of 31 X-linked probes with differential expression) fall within putative selective sweep regions. They are summarized in Table 2.7. One of the probes (representing the gene *CG9509*) was also found to be part of the top-ten probes over-expressed in Europe (Table 2.5). One finding is that the direction of the expression change is not correlated with the population where the putative selective sweep occurred. Out of the five genes lying in putative European sweep regions two show over- and three show under-

Table 2.7: Differentially expressed genes lying in putative selective sweep regions

Probe ID	Gene name	Exp_E[a]	Exp_A[a]	*P* value	Sweep regions?	
					OMETTO[b]	LI[c]
Signature of positive selection in Europe						
INC067C10	*CG9509*	2.31	1.00	$P < 0.0001$	Yes	-
INC089F12	*sesB*	1.00	1.37	$P = 0.0007$	Yes	Yes
INC080B11	*CG14419*	1.00	1.36	$P < 0.0001$	Yes	-
INC029G05	*caz*	1.00	1.31	$P = 0.0006$	-	Yes
INC044B07	*CG5877*	1.30	1.00	$P = 0.0012$	Yes	-
Signature of positive selection in Africa						
INC037H12	*Pgd*	1.00	1.50	$P = 0.0002$	-	Yes

[a] Exp_E and Exp_A are the relative expression levels in Europe and Africa respectively
[b] Putative selective sweep region according to OMETTO *et al.* (2005), Method $\mathrm{II}^{\mathrm{all}}$
[c] Putative selective sweep region according to LI and STEPHAN (2006)

expression in Europe. Genes of particular interest in Europe are the transcription factor *caz* and the mitochondrial membrane protein *sesB* (see DISCUSSION). The candidate gene for the African population *Pgd* lies in a region which has been extensively characterized at the DNA level and is thought to have undergone selection not only in the African but also in the European population (BEISSWANGER *et al.* 2006).

DISCUSSION

Patterns of gene expression polymorphism: The data presented here represents the first study of gene expression variation in a truly natural population of derived *D. melanogaster*. The inclusion of the African lines in our experiment allows us to conduct a comprehensive survey of expression variability of the species. The large amount of replication in our study allows us to detect even relatively small differences in gene expression, as suggested by the GEL_{50} statistic and the application of strict quality control criteria increases the resolution of our experimental design even more (Figure 2.3). In total we found that 4,512 genes (represented by 5,048 probes) were expressed at detectable levels. This number is slightly lower than the 4,905 clones passing the quality control step of the MEIKLEJOHN *et al.* (2003) study. Still the numbers agree reasonably well, given the differences in experimental design and quality control approaches. In total we found that 37.5% of the probes we studied showed at least one significant pairwise comparison and were therefore labelled as polymorphic. This is similar to the findings of MEIKLEJOHN *et al.* (2003). Nevertheless, it might not be valid to compare these two numbers directly. The differences in experimental design, investigated populations, sample size and chosen *P* value cut-off have a large influence on the percentage of genes (or probes) found to be differentially expressed (CLARK and TOWNSEND 2007).

A detailed inspection of expression polymorphism revealed that the amount of expression variation does not differ between the European and the African population. This might seem somewhat surprising, since large scale genome scans have shown that the African population harbours much more variation at the DNA level than the European population (*e.g.* GLINKA *et al.* 2003). This observation is attributed mainly to the fact that the African population has a larger effective population size than the European one (see CHAPTER 1). On the other hand, the DNA polymorphism studied in such genome scans consists mainly of SNPs which are thought to evolve neutrally. While some authors suggest that gene expression also evolves in a neutral manner (KHAITOVICH *et al.* 2004), more recent studies provide evidence that this is not the case (*e.g.* LEMOS *et al.* 2005). Regulatory changes have a direct impact on the phenotype and might affect the fitness of the organism. Most of these changes will have a deleterious effect and the levels of gene expression should therefore be under stabilizing selection. The patterns of expression polymorphism we observe might therefore be explained by a mutation-selection balance model, where arising mutations are mostly deleterious and quickly get purged from the population. In such a case the observable variation is solely dependent on the mutation rate and the selection coefficient against deleterious mutations (which should be equal in both of our studied populations), and independent of the population size (GILLESPIE 1998). The differences in population size between the African and European populations should therefore have no effect on their relative levels of gene expression variation. Evidence that stabilizing selection is a key factor in governing expression variation has already been found by previous studies. Mutation accumulation experiments in *Caenorhabditis elegans* (DENVER *et al.* 2005) and *D. melanogaster* (RIFKIN *et al.* 2005) have shown that spontaneous mutations are able to create abundant variation in gene expression. However, when comparing the levels of expression variation in mutation accumulation lines to the levels found in natural isolates, it can be seen that variation in natural populations is significantly lower (DENVER *et al.* 2005). Additionally, expression divergence between closely related species was much lower than expected under a neutral model (RIFKIN *et al.* 2005). These results suggest that stabilizing selection plays a dominant role in shaping gene expression variation in natural populations.

The amount of expression differences between populations was higher than within populations. This is easily understandable, as both populations have undergone population differentiation since the colonization of Europe approximately 17,000 years ago (LI and STEPHAN 2006, see also CHAPTER 1). This pattern is consistent independently of which type of statistic is used to describe expression polymorphism (Table 2.2). If expression levels were found to be significantly different between lines, the magnitude of the expression difference was found to be relatively small (~1.5-fold) for the majority of changes (Figure 2.6). This has already been reported in many other studies (*e.g.* OLEKSIAK *et al.* 2002, TOWNSEND *et al.* 2003) and seems to be a general biological pattern, observable in a broad range of species.

Contrasting the amount of expression polymorphism between the X chromosome and autosomes we find a significant paucity of polymorphic genes on the X chromosome. This might be the effect of the chromosomal location of male-biased genes. Male-biased genes are genes which are expressed at higher levels in males than in females and several studies have shown that these genes are preferably located on the autosomes of *D. melanogaster* (PARISI *et al.* 2003, RANZ *et al.* 2003). The study of MEIKLEJOHN *et al.* (2003) showed that male-biased genes generally show higher levels of expression variation than do unbiased or female-biased genes. The overrepresentation of these highly diverse male-biased genes on the autosomes could therefore have led to an increased level of expression diversity on a chromosomal level.

Effects of gene function: The question arises, if there are any qualitative differences between genes that show polymorphism in expression and those which are monomorphic. In other words: What is the reason that some genes vary in their expression levels while others have equal expression in all studied lines. We argue that one of the main forces governing expression polymorphism is stabilizing selection (see above). Under this view, each gene possesses an expression state which can be viewed as optimal. Any regulatory mutations that result in an expression state which departs from this optimum would lead to a decreased fitness of the organism and are selected against by stabilizing selection. Of course the fitness effect of a departure from the optimum will not be the same for different genes. Some genes are of larger importance than others and changes in these particular genes should have a more drastic effect on the fitness of the individual than changes in genes of less importance. Therefore stabilizing selection should act much more strongly on important genes and these genes should be monomorphic (or have reduced polymorphism).

Since *D. melanogaster* is well studied model organism, there is extensive knowledge about the functions of many genes in the genome. Unfortunately, even though this species is widely used in genetic and physiological research, little is known about the ecology of *D. melanogaster*. It is therefore not easy to categorize genes in to classes of major or minor importance in terms of their contribution to fitness. Furthermore the interaction between the over 10,000 genes in the genome is far from being resolved. So instead of trying to divide our genes into groups of important and non-important genes we applied a different approach. We looked at the number of different biological processes that a gene is involved in. While some genes are specialised to fulfil one specific biological function, others are involved in a multitude of different biological processes. A change of expression level in one of these multi-functional genes might therefore disrupt many different biological pathways and hence lead to a large decrease in fitness. Our expectation is that these genes should be under more stabilizing selection than genes with only few functions and consequently show lower levels of expression diversity. Of course there are some caveats to this approach. Just because a gene is specialised in only one function, does not mean it is of less importance. This one function could be absolutely essential to the viability of the organism and changes in expression levels could have drastic fitness effects. Additionally, the characterization of the gene functions for all genes in the *D. melanogaster* genome is far from being complete. For some genes the function is completely unknown, while for others we have only partial knowledge of the processes they are involved in.

Even when keeping these caveats in mind, the pattern we observe is striking (Figure 2.7). Genes that show differences in gene expression between individual strains tend to have fewer functions than those that are monomorphic. This pattern is consistent whether we investigate the number of different biological processes or the number of different molecular functions a gene is associated with. This reinforces our view that stabilizing selection is the dominant force when it comes to shaping the patterns of expression polymorphism in natural populations.

Finding candidate genes for adaptation: Our main approach for finding genes that are differentially expressed between Europe and Africa was the so-called two-node experiment (see MATERIALS AND METHODS). In this experiment we grouped all lines of a single population into one node (one European and one African). BAGEL was then used to estimate the average relative expression levels for the European and African populations and test if the expression differences between the two populations were significant. Whether or not there is statistical support for differential expression depends on the confidence intervals of the estimated average expression levels. Genes that show a high level of variation within populations will also have large confidence intervals for the estimate of expression for the given population. Good statistical support for differential expression is only obtained if the confidence intervals for the expression estimates of the European and African population show minimal overlap. Hence only those genes will be identified as candidates that have small confidence intervals for the population specific expression estimate (meaning there is little variation in expression level among the lines within a population), and at the same time the estimates of the average expression of the European and African populations are relatively far apart. It is therefore not surprising, that the probes assigned to be candidates following the

16-node experiment (circled data points in Figures 2.8 and 2.9), even though chosen somewhat arbitrarily, also show up in the lists of the best candidates derived from the two-node experiment (Tables 2.5 and 2.6). These genes fulfill the requirements of good candidates: a large amount of differentiation between populations and only little within population variation.

The pattern that genes show a large difference between populations but only little variation within a population will only be visible if the two population-specific expression levels represent optimal expression phenotypes that are adapted to both local environments. Of course other scenarios of adaptation are also possible. A gene could be under relaxed selective constraint in one population (*i.e.*, gene expression of this gene does not have to follow a specific pattern), but have undergone adaptive directional selection in the other. In such a case the first population should have high within-population polymorphism, while the other should be (nearly) monomorphic. Our approach should still be able to detect these kinds of genes, if the difference in average expression level between populations is large enough to overcome the within-population variation of the population with relaxed selective constraint. The gene *CG3301* seems to be such a case (Table 2.6). Looking at the Δ values, we see that there is a large amount of diversity within Europe, while the African population is almost monomorphic. The gene *Cyp6g1* shows a similar pattern (Table 2.5). Here the African population is relatively variable, while polymorphism in the Europe is reduced. In the following we will discuss the most promising candidate genes found by our approach.

Genes over-expressed in Europe: The gene with the highest amount of over-expression in Europe was *Cyp6g1*. This cytochrome P450 gene is very well studied, and it has been found, that over-expression of this gene leads to an increased resistance to insecticides such as DDT (DABORN *et al.* 2002). The exposure to insecticides represents a strong example of natural selection, and being able to overcome the effects of such substances would be associated with a large increase in fitness. In Europe large areas are used for agriculture, and the use of DDT was widespread until late in the 20th century. In general, the resistance to insecticides is viewed as a classical example of man-made natural selection (reviewed in FFRENCH-CONSTANT *et al.* 2004). PEDRA *et al.* (2004) studied the transcription profile of DDT-resistant *Drosophila* strains and found, that genes related to lipid metabolism showed different expression levels when compared to DDT-sensitive strains. Interestingly, we also find genes involved in lipid metabolism in our Top 10 list of genes over-expressed in Africa. The malic enzyme (*Men*) oxidizes malate to pyruvate and concurrently reduces NADP to NADPH, which is a major reductant in lipid biosynthesis (WISE and BALL 1964). Studies of DNA polymorphism and enzymatic activity of naturally occurring alleles of *Men* have already suggested, that this gene might be a target of positive selection (MERRITT *et al.* 2005). The exact function of the gene *CG32919* is not known, but its gene product shows homologies with fatty acyl-CoA reductases in other organisms (HUBBARD *et al.* 2007). These enzymes are involved in the synthesis of ether lipids and are conserved in a broad range of species (CHENG and RUSSELL 2004). The presence of the P450 gene and the two lipid-associated genes in our candidate list suggests that insecticide resistance might indeed drive natural selection in the European Drosophila population.

Another gene over-expressed in Europe is the odorant-binding protein *Obp56d*. This gene is a member of a gene family which is responsible for the olfactory sense, and it has been shown that olfactory genes are involved in mating behavior of *D. melanogaster* (MACKAY *et al.* 2005). A recent study suggests that odorant-binding proteins also play a significant role in taste perception and food preference of fruit flies (MATSUO *et al.* 2007). Interestingly, another gene on our list is associated with taste perception. The gene *dpr15* is member of the defective proboscis extension response (DPR) gene family, and NAKAMURA *et al.* (2002) have shown that these genes are involved in salt perception. We can therefore speculate that *Obp56d* and *dpr15* might be involved in the food choice of *D. melanogaster*.

The gene α Mannosidase I (*α-Man-I*) is involved in the metabolism of Asn-linked oligosaccharides in the Golgi apparatus (DEWALD and TOUSTER 1973). This gene is also part of a larger gene family, and a study investigating the phylogenetic relationship of these genes in multiple species suggests that these genes have been target of positive Darwinian selection (GONZALEZ and JORDAN 2000).

The gene *CG9509*, which encodes for a choline dehydrogenase and plays a role in the mesoderm development of *Drosophila* (FURLONG *et al.* 2001), was also found to be over-expressed in cosmopolitan lab strains when compared to African flies (MEIKLEJOHN *et al.* 2003). So over-expression of this gene seems to be common in derived populations of *D. melanogaster*. For the other genes in the Top 10 list, functions are either unknown or too vague to draw any reasonable conclusions on their possible contribution to the phenotype.

Genes over-expressed in Africa: The investigation of the genes over-expressed in Africa reveals an interesting pattern. The gene with the highest level of over-expression (5.36-fold) is *CG7214*. Even though the exact function of his gene is unknown, it has been shown that it is involved in wing morphogenesis (REN *et al.* 2005). *Act88F* encodes for an actin that is found predominantly in the indirect flight muscle of *Drosophila* and is responsible for the correct myofibril formation (KARLIK *et al.* 1984). The gene *TpnC41C* encodes for a subunit of troponin that is only found in the indirect flight muscle, and which is essential for the muscle contraction (QIU *et al.* 2003). Expanding our list beyond the Top 10 genes presented in Table 2.6 we find even more genes associated with the flying apparatus. The genes *Mlc1*, *Mlc2* and *fln* also show a significant over-expression in Africa that is more than 1.5-fold. The myosin light chain proteins *Mlc1* and *Mlc2* are part of the thick filament in *Drosophila* flight muscles (VIGOREAUX 2001) and the gene *flightin* (*fln*) is essential for the assembly and stability of the flight muscles (REEDY *et al.* 2000). This might be related to differences in the ratio of wing-size/body-size between African and European flies. It is known that *D. melanogaster* populations living close to the equator have smaller wings relative to their body-size than flies inhabiting higher latitudes (AZEVEDO *et al.* 1998). It has also been shown that flies that have a small wing area relative to their body size have higher frequencies of wing-beat to overcome the small lift provided by their wings (REED *et al.* 1942). We therefore hypothesize that the higher expression levels of muscle genes enables African flies to maintain a high-frequency wing-beat. Direct measurements of relative wing sizes and wing-beat frequencies in our surveyed populations will allow a test of this hypothesis.

Other genes found to be over-expressed in Africa also have interesting functions. *CG3301* encodes for an oxidoreductase. This gene was also found to be significantly over-expressed in African flies when compared to cosmopolitan lab strains in the study of MEIKLEJOHN *et al.* (2003). Additionally, it is one of the genes which are differentially expressed between DDT-resistant and DDT-sensitive strains of *D. melanogaster* (PEDRA *et al.* 2004). However, the expression patterns of this gene reported for the two resistant stains in the DDT study are inconclusive. The laboratory selected DDT-resistant strain *Rst(2)DDT*$^{91\text{-}R}$ showed an over-expression of *CG3301* when compared to the DDT-sensitive strain *Canton-S*. The wild-caught DDT-resistant strain *Rst(2)DDT*^{Wisconsin}, on the other hand, showed an under-expression. Furthermore, we see that within-population variation for this gene is high in Europe and low in Africa (Table 2.6). This suggests that if positive selection acted, it must have happened in Africa. Therefore, the adaptive acquisition of insecticide resistance by the European population does not seem to be the driving force behind the gene expression differentiation of *CG3301*.

The gene *Nplp3* encodes a neuropeptide-like precursor (BAGGERMAN *et al.* 2002), and differences in expression might have an influence on behavioral phenotypes. The gene *Mipp1* is an inositol polyphosphate phosphatase which hydrolizes inositol phosphates. These are abundant metabolites found in a variety of plant tissues as a storage form of phosphate, but are also used as messengers to translate extra-cellular signals to the cytoplasm (STREB *et al.*

1983). Since inositol phosphates are found in many plats, differences in expression of *Mipp1* between Europe and Africa might reflect different food choices.

An interesting case is the gene *fit*, which stands for *female-specific independent of transformer*. This gene is expressed mainly in fat cells and is highly female-biased (FUJII and AMREIN 2002). Another study has shown that this gene is involved in oocyte maturation, a process which only occurs in females (NAKAHARA *et al.* 2005). The question arises if the differences we observe in the males studied here simply reflects differences which are present (and were maybe selected for) in females, or if this gene also has a function in males which could have been the target of selection. A survey of the expression pattern in European and Africa females might give us some more insight into the behavior of this gene.

Unfortunately, not much is known about the gene with second largest expression difference (5.31-fold) between Europe and Africa, *CG7203*. Its function is unknown, but a search for homologous proteins in other species revealed a homology with a putative cuticle protein in the yellow fever mosquito *Aedes aegypti* (HUBBARD *et al.* 2007). Interestingly, this gene is in close proximity to *CG7214* (less than 10kb away). Since these two genes show a similar expression pattern and they are orientated in the same direction on the chromosome, it might be possible that they are controlled by the same *cis*-regulatory elements.

Candidates lying in putative selective sweep regions: The studies of OMETTO *et al.* (2005) and LI and STEPHAN (2006) investigated the patterns of DNA polymorphism of the X chromosome of the two populations studied here. They defined regions where demography alone could not explain the observed patterns and which therefore were candidate regions of recent positive selection. Finding the gene that was the target of selection, however, is a difficult task. The six candidate regions defined by the method II^{all} of OMETTO *et al.* (2005) span on average 400kb of the chromosome. The approach of LI and STEPHAN (2006) is more precise, as the surveyed part of the X chromosome was clustered into windows of 100kb, and eleven of these windows were found to be putative candidate regions for positive selection in the European population. Still these regions are quite large and contain many genes that could have been the target of selection. In order to find this target, some of these regions have been characterized in greater detail at the DNA level, and patterns of SNP polymorphism were used to narrow down the region which was affected by the selective sweep (BEISSWANGER *et al.* 2006).

If selection acted on a mutation in a *cis*-regulatory sequence that altered the expression phenotype of the gene under control, we can use our expression data to find such a gene. We therefore specifically looked at genes that showed differences in expression between the European and African population and lie in regions that are candidates of positive selection as suggested by the aforementioned studies. If positively selected mutations in *cis*-elements are the main source of expression differentiation between the two populations, then we should see an excess of differentially expressed genes lying in candidate regions of positive selection. We do not observe such a pattern; the proportion of differentially expressed genes lying in selective sweep regions does not deviate from expectations under a random distribution (data not shown). In total, five genes were found to be candidates of positive selection in the European population (Table 2.7). One of these genes, *CG9509* is among the genes with the largest amount of over-expression in the European population and was also found to be over-expressed in other derived strains (see above). Another interesting gene is *stress sensitive B* (*sesB*). It lies in a region that was identified as a potential target of a selective sweep in both the OMETTO *et al.* (2005) and the LI and STEPHAN (2006) study. The gene encodes for a mitochondrial membrane protein, and ZHANG *et al.* (1999) have identified this protein as an adenine nucleotide translocase, a class of proteins which is required for the exchange of ADP and ATP across the mitochondrial membrane. The name of the gene is derived from the fact that *sesB* mutants are extremely sensitive to stress and mechanical shock (HOMYK 1977). Interestingly, mutations at this gene also decrease the flight ability of the individual (HOMYK and SHEPPARD 1977). This expands the list of candidates which possibly contribute to the differentiation of flight behavior of European and African *D. melanogaster* (see above).

We find that the gene *Pgd* lies in a candidate region for selection in the African population. This region has been extensively characterized at the DNA polymorphism level for both the European and the Africa population (BEISSWANGER *et al.* 2006). The main finding of this study was that selection not only occurred in the African population, but also in Europe. The DNA polymorphism pattern suggests that *Pgd* actually is not the target of selection in the African population, since it lies outside of the valley of reduced DNA diversity associated with a selective sweep. It could, however, be a potential target in Europe. Sequencing of the 5' region of *Pgd* showed that some SNPs which were fixed or at high frequency in Europe were in low frequency in Africa. The effect of these mutations on the expression phenotype is up to now not well understood and will require further research (S. BEISSWANGER, personal communication).

ACKNOWLEDGMENTS

We would like to thank Jeff Townsend for helpful comments on how to conceive the experiment and interpret the data and John Baines for sharing his expertise on microarrays and the lively discussion throughout the course of the experiment. We also thank Bea Stiening for her excellent technical assistance. Special thanks go out to Anna Bauer-Mehren for her work on normalizing our microarray data. This work was supported by grants from the Deutsche Forschungsgemeinschaft and the Volkswagen-Foundation.

LITERATURE CITED

AZEVEDO, R. B. R., A. C. JAMES, J. MCCABE and L. PARTRIDGE, 1998 Latitudinal variation of wing:thorax size ratio and wing-aspect ratio in *Drosophila melanogaster*. Evolution **52:** 1353-1362.

BAGGERMAN, G., A. CERSTIAENS, A. DE LOOF and L. SCHOOFS, 2002 Peptidomics of the larval *Drosophila melanogaster* central nervous system. J Biol Chem **277:** 40368-40374.

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

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.

CAVALIERI, D., J. P. TOWNSEND and D. L. HARTL, 2000 Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. Proc Natl Acad Sci U S A **97:** 12369-12374.

CHENG, J. B., and D. W. RUSSELL, 2004 Mammalian wax biosynthesis. I. Identification of two fatty acyl-Coenzyme A reductases with different substrate specificities and tissue distributions. J Biol Chem **279:** 37789-37797.

CLARK, T.A., and J. P. TOWNSEND, 2007 Quantifying variation in gene expression. Mol Ecol **16:** 2613-2616.

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.

DABORN, P. J., J. L. YEN, M. R. BOGWITZ, G. LE GOFF, E. FEIL *et al.*, 2002 A single p450 allele associated with insecticide resistance in *Drosophila*. Science **297:** 2253-2256.

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

DENVER, D. R., K. MORRIS, J. T. STREELMAN, S. K. KIM, M. LYNCH *et al.*, 2005 The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nat Genet **37:** 544-548.

DEWALD, B., and O. TOUSTER, 1973 A new alpha-D-mannosidase occurring in Golgi membranes. J Biol Chem **248:** 7223-7233.

FAY, J. C., H. L. MCCULLOUGH, P. D. SNIEGOWSKI and M. B. EISEN, 2004 Population genetic variation in gene expression is associated with phenotypic variation in *Saccharomyces cerevisiae*. Genome Biol **5:** R26.

FFRENCH-CONSTANT, R. H., P. J. DABORN and G. LE GOFF, 2004 The genetics and genomics of insecticide resistance. Trends Genet **20:** 163-170.

FUJII, S., and H. AMREIN, 2002 Genes expressed in the *Drosophila* head reveal a role for fat cells in sex-specific physiology. Embo J **21:** 5353-5363.

FURLONG, E. E., E. C. ANDERSEN, B. NULL, K. P. WHITE and M. P. SCOTT, 2001 Patterns of gene expression during *Drosophila* mesoderm development. Science **293:** 1629-1633.

GIBSON, G., R. RILEY-BERGER, L. HARSHMAN, A. KOPP, S. VACHA *et al.*, 2004 Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. Genetics **167:** 1791-1799.

GILLESPIE, J. H., 1998 *Population genetics : a concise guide*. The Johns Hopkins University Press, Baltimore, Md.

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.

GONZALEZ, D. S., and I. K. JORDAN, 2000 The alpha-mannosidases: phylogeny and adaptive diversification. Mol Biol Evol **17:** 292-300.

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.

HOMYK, T., 1977 Behavioral mutants of *Drosophila melanogaster*. II. Behavioral analysis and focus mapping. Genetics **87:** 105-128.

HOMYK, T., and D. E. SHEPPARD, 1977 Behavioral mutants of *Drosophila melanogaster*. I. Isolation and mapping of mutations which decrease flight ability. Genetics **87:** 95-104.

HUBBARD, T. J., B. L. AKEN, K. BEAL, B. BALLESTER, M. CACCAMO *et al.*, 2007 Ensembl 2007. Nucleic Acids Res **35:** D610-617.

JIN, W., R. M. RILEY, R. D. WOLFINGER, K. P. WHITE, G. PASSADOR-GURGEL *et al.*, 2001 The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet **29:** 389-395.

KHAITOVICH, P., G. WEISS, M. LACHMANN, I. HELLMANN, W. ENARD *et al.*, 2004 A neutral model of transcriptome evolution. PLoS Biol **2:** E132.

KING, M. C., and A. C. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science **188:** 107-116.

LACHAISE, D., M. L. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS and M. ASHBURNER, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol Biol **22:** 159-225.

LEMOS, B., C. D. MEIKLEJOHN, M. CACERES and D. L. HARTL, 2005 Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution **59:** 126-137.

LI, H., and W. STEPHAN, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet **2:** e166.

MACKAY, T. F., S. L. HEINSOHN, R. F. LYMAN, A. J. MOEHRING, T. J. MORGAN *et al.*, 2005 Genetics and genomics of *Drosophila* mating behavior. Proc Natl Acad Sci U S A **102 Suppl 1:** 6622-6629.

MATSUO, T., S. SUGAYA, J. YASUKAWA, T. AIGAKI and Y. FUYAMA, 2007 Odorant-binding proteins *OBP57d* and *OBP57e* affect taste perception and host-plant preference in *Drosophila sechellia*. PLoS Biol **5:** e118.

MEIKLEJOHN, C. D., J. PARSCH, J. M. RANZ and D. L. HARTL, 2003 Rapid evolution of male-biased gene expression in *Drosophila*. Proc Natl Acad Sci U S A **100:** 9894-9899.

MERRITT, T. J., D. DUVERNELL and W. F. EANES, 2005 Natural and synthetic alleles provide complementary insights into the nature of selection acting on the *Men* polymorphism of *Drosophila melanogaster*. Genetics **171:** 1707-1718.

NAKAHARA, K., K. KIM, C. SCIULLI, S. R. DOWD, J. S. MINDEN *et al.*, 2005 Targets of microRNA regulation in the *Drosophila* oocyte proteome. Proc Natl Acad Sci U S A **102:** 12023-12028.

NAKAMURA, M., D. BALDWIN, S. HANNAFORD, J. PALKA and C. MONTELL, 2002 Defective proboscis extension response (DPR), a member of the *Ig* superfamily required for the gustatory response to salt. J Neurosci **22:** 3463-3472.

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

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

PARISI, M., R. NUTTALL, D. NAIMAN, G. BOUFFARD, J. MALLEY *et al.*, 2003 Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science **299:** 697-700.

PEDRA, J. H., L. M. MCINTYRE, M. E. SCHARF and B. R. PITTENDRIGH, 2004 Genome-wide transcription profile of field- and laboratory-selected dichlorodiphenyltrichloroethane (DDT)-resistant *Drosophila*. Proc Natl Acad Sci U S A **101:** 7034-7039.

RANZ, J. M., C. I. CASTILLO-DAVIS, C. D. MEIKLEJOHN AND D. L. HARTL, 2003 Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science **300:** 1742-1745.

REED, S. C., C. M. WILLIAMS and L. E. CHADWICK, 1942 Frequency of wing-beat as a character for separating species races and geographic varieties of *Drosophila*. Genetics **27:** 349-361.

REN, N., C. ZHU, H. LEE and P. N. ADLER, 2005 Gene expression during *Drosophila* wing morphogenesis and differentiation. Genetics **171:** 625-638.

RIFKIN, S. A., D. HOULE, J. KIM and K. P. WHITE, 2005 A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438:** 220-223.

RIFKIN, S. A., J. KIM and K. P. WHITE, 2003 Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nat Genet **33:** 138-144.

RUBIN, G. M., L. HONG, P. BROKSTEIN, M. EVANS-HOLM, E. FRISE *et al.*, 2000 A *Drosophila* complementary DNA resource. Science **287:** 2222-2224.

STRANGER, B. E., M. S. FORREST, A. G. CLARK, M. J. MINICHIELLO, S. DEUTSCH *et al.*, 2005 Genome-wide associations of gene expression variation in humans. PLoS Genet **1:** e78.

STREB, H., R. F. IRVINE, M. J. BERRIDGE and I. SCHULZ, 1983 Release of Ca2+ from a nonmitochondrial intracellular store in pancreatic acinar cells by inositol-1,4,5-trisphosphate. Nature **306:** 67-69.

THE CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437:** 69-87.

TOWNSEND, J. P., 2004 Resolution of large and small differences in gene expression using models for the Bayesian analysis of gene expression levels and spotted DNA microarrays. BMC Bioinformatics **5:** 54.

TOWNSEND, J. P., D. CAVALIERI and D. L. HARTL, 2003 Population genetic variation in genome-wide gene expression. Mol Biol Evol **20:** 955-963.

TOWNSEND, J. P., and D. L. HARTL, 2002 Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. Genome Biol **3:** RESEARCH0071.

WHITEHEAD, A., and D. L. CRAWFORD, 2006a Variation within and among species in gene expression: raw material for evolution. Mol Ecol **15:** 1197-1211.

WHITEHEAD, A., and D. L. CRAWFORD, 2006b Neutral and adaptive variation in gene expression. Proc Natl Acad Sci U S A **103:** 5425-5430.

WILSON, A. C., L. R. MAXSON and V. M. SARICH, 1974 Two types of molecular evolution. Evidence from studies of interspecific hybridization. Proc Natl Acad Sci U S A **71:** 2843-2847.

WISE, E. M., JR., and E. G. BALL, 1964 Malic Enzyme and Lipogenesis. Proc Natl Acad Sci U S A **52:** 1255-1263.

ZHANG, Y. Q., J. ROOTE, S. BROGNA, A. W. DAVIS, D. A. BARBASH *et al.*, 1999 *stress sensitive B* encodes an adenine nucleotide translocase in *Drosophila melanogaster*. Genetics **153:** 891-903.

3. Genome-wide DNA polymorphism analyses using VariScan

ABSTRACT

DNA sequence polymorphisms analysis can provide valuable information on the evolutionary forces shaping nucleotide variation, and provides an insight into the functional significance of genomic regions. The recent ongoing genome projects will radically improve our capabilities to detect specific genomic regions shaped by natural selection. Current available methods and software, however, are unsatisfactory for such genome-wide analysis. We have developed methods for the analysis of DNA sequence polymorphisms at the genome-wide scale. These methods, which have been tested on a coalescent-simulated and actual data files from mouse and human, have been implemented in the VariScan software package version 2.0. Additionally, we have also incorporated a graphical-user interface. The main features of this software are: i) exhaustive population-genetic analyses including those based on the coalescent theory; ii) analysis adapted to the shallow data generated by the high-throughput genome projects; iii) use of genome annotations to conduct a comprehensive analyses separately for different functional regions; iv) identification of relevant genomic regions by the sliding-window and wavelet-multiresolution approaches; v) visualization of the results integrated with current genome annotations in commonly available genome browsers. VariScan is a powerful and flexible suite of software for the analysis of DNA polymorphisms. The current version implements new algorithms, methods, and capabilities, providing an important tool for an exhaustive exploratory analysis of genome-wide DNA polymorphism data.

INTRODUCTION

The comparative analysis of DNA sequence variation within species (polymorphism) and between species (divergence) is a powerful approach to understand the evolutionary process (*e.g.* HUDSON *et al.* 1987, MCDONALD and KREITMAN 1991), and represents an insight into the functional significance of genomic regions (for instance, see HUGHES and YEAGER 1998). Particularly, the detection of both positive and negative purifying selection at the molecular level is of major interest. Since positive Darwinian selection is ultimately responsible for evolutionary adaptations, the detection of genomic regions driven by positive selection has profound implications in evolutionary biology as well as in understanding the gene function. The identification of regions evolving by negative selection is also very important as conserved regions are most likely to be functionally significant. The inference of such evolutionary process requires knowing how within-species DNA sequences change under neutrality (KIMURA 1983). In this context, the coalescent theory (KINGMAN 1982, HUDSON 1990) has become the primary framework for the analysis of DNA polymorphism data.

Currently, there are few convincing studies on the action of recent -or ongoing- positive selection at the intraspecific level (*e.g.* SABETI *et al.* 2002, QUESADA *et al.* 2003, MEKEL-BOBROV *et al.* 2005). Apparently, the most important difficulty is that demographic events such as migration, population expansions or bottlenecks can mimic the signature of selective processes; therefore, it is not easy to detect the specific imprint of positive selection on individual genes or on short stretches of DNA. The distinction between natural selection and other demographic events requires the surveys of large genome regions (for instance, see QUESADA *et al.* 2003, ORENGO and AQUADÉ 2004, HADDRILL *et al.* 2005, NORDBORG *et al.* 2005). The detection of negative purifying selection on DNA sequences, on the contrary, has been much easier (KREITMAN 1983); in fact negative selection is acting continuously while positive selection is much more episodic. Indeed, there are many surveys where the action of negative purifying selection has been detected even at non-coding DNA regions (*e.g.* ANDOLFATTO 2005, MACDONALD and LONG 2005). Undoubtedly, such studies will provide fundamental insights into the functional significance of non-coding DNA. Even so, there are very few studies analysing the within and between-species patterns of nucleotide variation at the genome-wide scale.

Recent genome projects efforts, as the HapMap (http://www.genome.gov/10001688), ENCODE (http://www.genome.gov/10005107), SimYak (http://www.dpgp.org/sim_yak/), DPGP (http://www.dpgp.org/about_dpgp/) and the Mouse Genome Resequencing Project (http://www.niehs.nih.gov/crg/cprc.htm) will change radically our capabilities to detect specific genomic regions shaped by natural selection. Although with different goals, these projects will generate SNPs (single nucleotide polymorphisms) data from many whole-genome copies. A limiting critical point has been the absence of adequate bioinformatic tools for such analysis. Although there are powerful programs for molecular population genetic analyses [for instance, ProSeq (FILATOV 2002), DnaSP (ROZAS *et al.* 2003) and Arlequin (EXCOFFIER *et al.* 2005)], they are not completely satisfactory for the high-throughput kind of data released by these projects.

Here, we describe version 2 of the VariScan software (VILELLA *et al.* 2005). In this new version we implemented new methods and features for an exhaustive analysis of DNA sequence polymorphisms at the genome-wide scale, using a graphical user-friendly interface. In particular, the current version of the software allows i) reading several informative-rich genome-wide data files; ii) estimating many population genetic parameters including coalescent-based statistics; iii) a separate analysis for different genomic regions, functional categories, chromosome locations, etc; iv) adapted analysis for shallow data generated by high-throughput genome projects; v) the identification of relevant genomic regions by using the sliding-window (*e.g.* ROZAS and ROZAS 1995) and wavelet-multiresolution approaches

(ARNEODO *et al.* 1995, MALLAT 1989, LIÒ 2003); vi) the visualization of the results integrated with current genome annotations in the most commonly available genome browsers.

IMPLEMENTATION

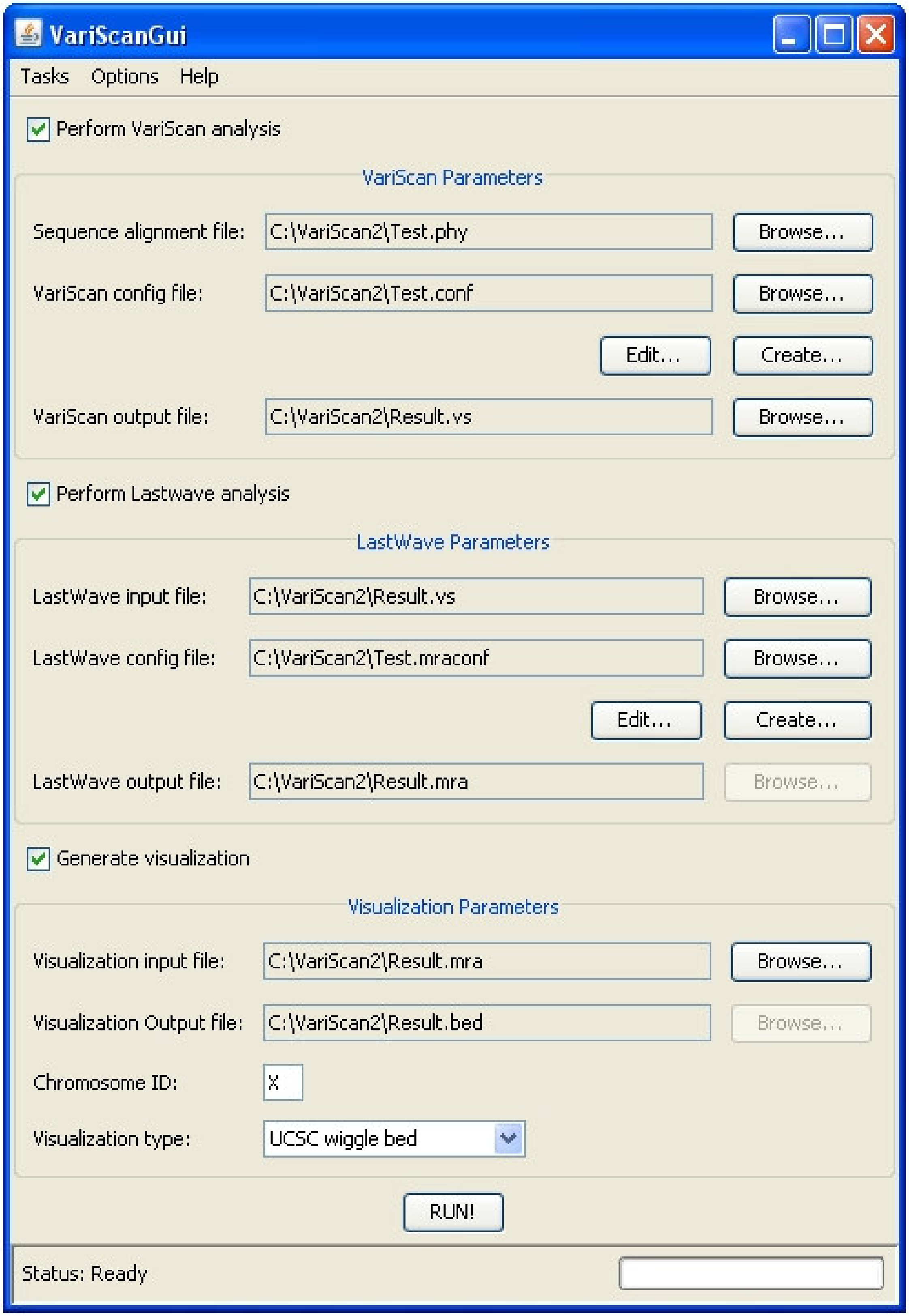
VariScan main algorithms are written in ANSI C. The software also includes a number of scripts written in Perl, and a GUI front-end developed in Java. VariScan currently runs on a wide variety of platforms, such as Linux, MacOS X and Win32. VariScan also uses the LastWave version 2.0 software (http://www.cmap.polytechnique.fr/~bacry/LastWave/) that is invoked from the Java front-end.

RESULTS

New features: VariScan version 2 incorporates substantial improvements over version 1: it implements many new methods and features and also includes a graphical user-friendly interface. Specifically, VariScan 2 allows handling input data files with DNA sequence information from (one or more) outgroup species. This feature allows the current version of VariScan conducting divergence estimates, neutrality tests and other parameters requiring such information. The second major improvement is the possibility to conduct separate analysis of different genomic regions (in exonic, intronic, etc), functional categories (such those defined in the Gene Ontology) and chromosome locations. In addition, VariScan version 2 implements new features to visualize the results of the sliding-window, as well of the wavelet-multiresolution approaches, integrated with current genome annotations in the most commonly available genome browsers. Since the data analysis by using such methods is complicated, we have incorporated an easy-to-use graphical user interface which allows conducting all needed computing steps, including those of the wavelet-multiresolution methods.

Overview: VariScan can read multiple alignment formats as MAF, MGA, PHYLIP, XMGA as those used in the HapMap project (http://www.genome.gov/10001688), with DNA sequence polymorphism data (within-species variation), and also with interspecific nucleotide variation (outgroup information). The software allows conducting exhaustive population-genetic analyses using genome annotations, and permits the visualization of the results integrated in the most commonly available genome browsers. The analysis can be performed using the available GUI (Graphical User Interface) (Figure 3.1) or under a command-line mode.

Molecular population genetics analysis: VariScan computes state-of-the-art population genetic parameters and coalescent-based statistics including those requiring outgroup nucleotide information (KINGMAN 1982, HUDSON 1990, NEI 1987, ROSENBERG and NORDBORG 2002). In particular, VariScan calculates (1) the standard summary statistics of nucleotide polymorphism and divergence levels (NEI 1987, DEPAULIS and VEUILLE 1998), such as the population mutational parameter (θ), nucleotide diversity (π), haplotype diversity or the number of nucleotide substitutions per site (*K*); (2) linkage disequilibrium based-statistics: *D*' (LEWONTIN 1964), r^2 (HILL and ROBERTSON 1968) and Z_{nS} (KELLY 1997); (3) neutrality-based tests: Tajima's *D* (TAJIMA 1989), Fu and Li's *D**, *F**, *D* and *F* (FU and LI 1993), Fu's F_S (FU 1997), and Fay and Wu's *H* (FAY and WU 2000). All parameters and statistics can be conducted by means of the sliding window (SW) (ROZAS and ROZAS 1995), or the multiresolution analysis (MRA) approaches (ARNEODO *et al.* 1995, MALLAT 1989, LIÒ 2003).



☑ Perform VariScan analysis

Sequence alignment file: C:\VariScan2\Test.phy

VariScan config file: C:\VariScan2\Test.conf

VariScan output file: C:\VariScan2\Result.vs

☑ Perform Lastwave analysis

LastWave input file: C:\VariScan2\Result.vs

LastWave config file: C:\VariScan2\Test.mraconf

LastWave output file: C:\VariScan2\Result.mra

☑ Generate visualization

Visualization input file: C:\VariScan2\Result.mra

Visualization Output file: C:\VariScan2\Result.bed

Chromosome ID: X

Visualization type: UCSC wiggle bed

Status: Ready

Figure 3.1: Graphical User Interface of VariScan showing the major options of analysis.

Missing data: Previous statistics are commonly estimated after excluding all sites with alignment gaps or missing data (*i.e.* the standard *Complete Deletion* option). However, current genome sequencing projects are generating high-throughput data with a large number of sites with missing information. For example, only ~10% of the polymorphic sites identified in the PATIL *et al.* (2001) study were typed in all 20 chromosomes. Therefore, it is clearly convenient to develop and implement statistics that could capture relevant information included from sites with missing data (about 90% in PATIL *et al.*'s data). Here, we have implemented a version of π (π_m) dealing with missing data. We define π_m (per site) as

$$\pi_m = \frac{k}{l} \qquad (1)$$

where l is the net number of positions surveyed (see below), and k is the average number of nucleotide differences that is given by

$$k = \sum_{i=1}^{m} h_i \qquad (2)$$

where m is the total number of positions (including sites with missing information, but excluding all positions with alignment gaps), and h_i is heterozygosity at site i, that is defined as

$$h_i = \begin{cases} \dfrac{n_i}{n_i - 1}\left(1 - \sum_{j=1}^{4} {x_{ij}}^2\right), n_i > 1 \\ 0, otherwise \end{cases} \qquad (3)$$

where n_i is the total number of chromosomes (sequences) excluding those with missing data at site i (*i.e.*, the net sample size), and x_{ij} is the relative frequency of nucleotide variant j (j = 1, 2, 3, and 4 correspond to A, C, G, and T) at site i. We denote as l (the net number of positions) the total number of positions excluding those sites with $n_i \leq 1$. In estimating π_m, all sites with alignment gaps should be excluded from the analysis. The rationale for this criterion is that while missing data are likely accumulated at random, alignment gaps are not; indeed, two (or more) sequences with gaps in a given position likely correspond to a single insertion/deletion event occurred in a common ancestor.

Analysis of different functional regions: VariScan allows a fine and detailed analysis of the pattern and levels of nucleotide variation at different functional regions. More precisely, it allows a separate analysis of different genomic regions (*e.g.* intergenic, non-coding, exonic, intronic, etc.), functional categories (a particular Gene Ontology category), or chromosome locations (specific chromosomal bands or arms, etc.). For the analysis VariScan uses current genome annotations available in public databases. This task is accomplished by a Perl script (*gff2bdf.pl*) that parses the appropriate genome information contained in a GFF (General Feature Format) file (http://www.sanger.ac.uk/Software/formats/GFF/) and returns a BDF (Block Data File) file directly used by VariScan. The BDF format, which is very similar to that used in VISTA server (http://genome.lbl.gov/vista/index.shtml), consists of a tab-delimited list of the relevant positions (the chromosome positions of the genome feature on the reference sequence) to be analysed. *gff2bdf.pl* incorporates several pre-defined filter options; the script, nevertheless, can be easily adapted to accommodate specific or more complex analyses.

Wavelet transform and multiresolution analysis: VariScan incorporates both the standard SW and the wavelet-based methods to identify particular genome features along the DNA sequence. The wavelet transform (WT), like Fourier transform, is a mathematical transformation widely used to extract information from signals. A signal can be resolved simultaneously in time (or space) and frequency domain by WT. The Fourier transform, on the contrary, only contains frequency information and, therefore, fails to detect spectral components localized in the time (or space) domain. Therefore, wavelet-based analysis provides a method to decompose the signal into high and low frequencies and therefore it is useful in extracting feature information at different scales. For the present analysis, time/space and frequency should be regarded as the position of the nucleotide sequence (a multiple alignment of nucleotide sequence data) and the relevant parameter intensity (levels of nucleotide diversity, linkage disequilibrium, etc), respectively. In this context, the signal is the profile of the relevant statistic along the DNA sequence. Here, we used the WT to decompose the signal into high and low frequencies for detecting global and local relevant features from genome-scale DNA polymorphism data. There are two basic kinds of WT, continuous (CWT) and discrete (DWT). The CWT of a signal *x*(*t*) is defined as

$$CWT_x^{\psi}(\tau,s)=\Psi_x^{\psi}(\tau,s)=\frac{1}{\sqrt{s}}\int_{-\infty}^{\infty}x(t)\psi^*\left(\frac{t-\tau}{s}\right)dt$$

where τ represents translation (time/space shift), s represents scale (or dilation; the inverse of the frequency), ψ(*t*) is the transforming function or mother wavelet, and the asterisk denotes a complex conjugate. There are a number of suitable mother wavelet functions; the choice of the particular mother wavelet to be used, nevertheless, should be adapted to the actual information to be extracted from the signal. Signals are analysed by CWT, which is obtained by scaling and translating (shifting) the mother wavelet along the signal. This process generates the wavelet coefficients (which represent the fit between the function and a particular scale-time of signal) that capture relevant information from a signal.

Here, we used the DWT (discrete wavelet transform), which is just the discrete version of CWT, because of the discrete nature of the signal to be analysed (DNA polymorphism data). The signal, which can be envisaged as a one-dimensional vector (of length *L*), is analysed by the *wtrans1d* module of LastWave v2.0 software (http://www.cmap.polytechnique.fr/~bacry/LastWave/) using Daubechies' D4 (DAUBECHIES 1992) as the default wavelet filter since it is adequate for locating features, such as peaks and valleys, from a signal (LIÒ and VANNUCCI 2000). The DWT analysis requires a signal to have a number of points equal to some power of two. For this purpose, and to avoid the boundary effect problem, we used the mirror padding method. With this approach the signal is extended by mirroring both ends at the boundaries, to achieve a total length (*L*') as a power of two. After the WT analysis, the padding tags are discarded and the original signal (of length *L*) is recovered. DWT can be conducted by means of the MRA (MALLAT 1989). This method uses a fast algorithm based on orthogonal wavelets, leading to the decomposition of a signal into different resolution levels; consequently, it enables the extraction of valuable information at different scales. Under this method, the original signal is decomposed by two complementary filters (half-band filters). As a result, the signal is split into two equal parts: one including the high-frequency components (detail coefficients), and the other with low frequency components (approximation coefficients) (Figure 3.2). While details are not further analysed, the approximation component is successively decomposed, split into two new high and low frequency components. The decomposition process can continue hierarchically until the detail component consists of a single coefficient. Orthogonal wavelets allow for the further reconstruction of the signal, which can be used for an easy location of features along the DNA sequence.

In the context of DNA polymorphism analysis, the signal is the raw profile of the statistic (for instance, nucleotide diversity or linkage disequilibrium levels) obtained along the DNA sequence. The signal is further decomposed to all analysed levels (MRA analysis) using

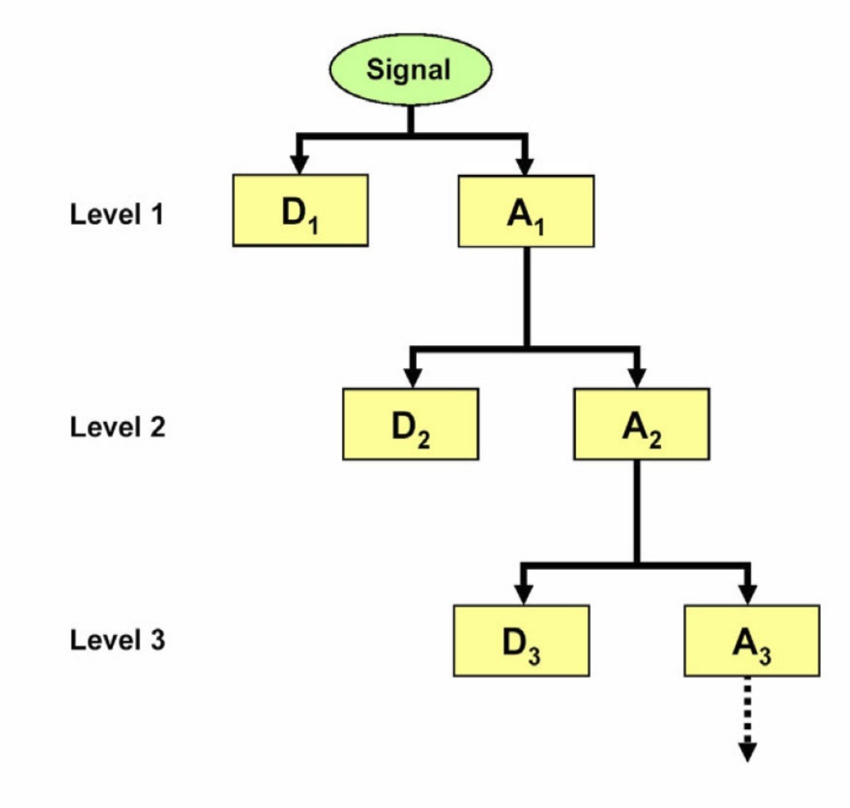


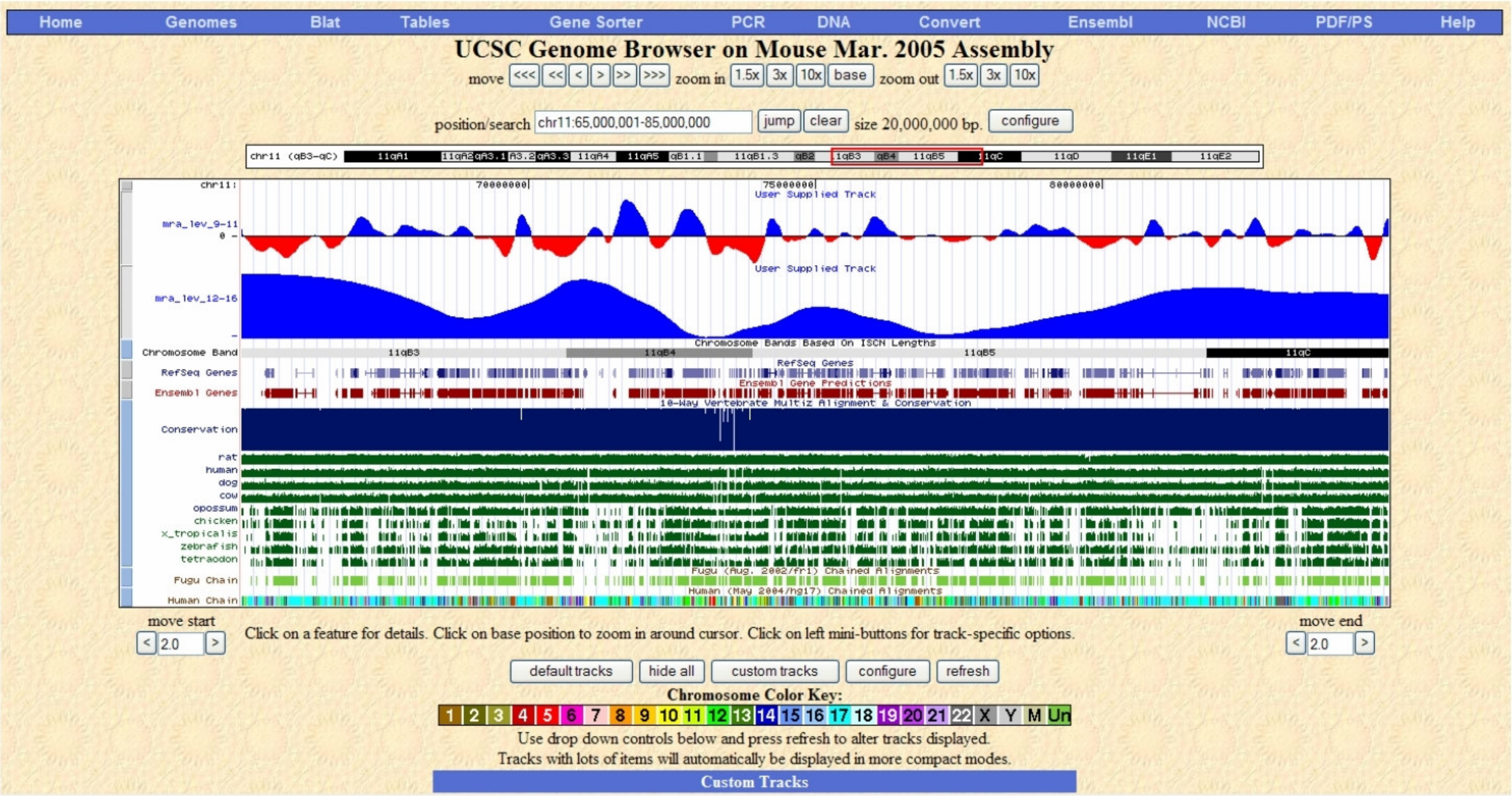
Figure 3.2: Wavelet decomposition tree. MRA allows for the decomposition of a signal into several resolution levels. First, the original signal (with a power of two points) is decomposed by two complementary half-band filters (high-pass and low-pass filters) that divide a spectrum into high-frequency (detail coefficients; D1) and low-frequency (approximation coefficients; A1) components (bands). For example, the low-pass filter will remove all half-band highest frequencies. Information from only the low frequency band (A1), with a half number of points, will be filtered in the second decomposition level. The A2 outcome will be filtered again for further decomposition.

the orthogonal wavelet decomposition method. The orthogonal property of Daubechies' wavelets allows for reconstruction of the signal. The outcome is the reconstructed wavelet-transform profiles of the population genetic parameter along the sequence, which can be used for detecting global and local relevant features (*i.e.* at different resolution scales) on genome-wide DNA polymorphism data.

Output visualization: The SW and MRA results can easily be visualized in available genome browsers (see Figure 3.3), such as the Human Genome Web Browser at UCSC (KENT *et al.* 2002) and any Web browser using Gbrowse (STEIN *et al.* 2002). This is accomplished by writing the relevant outcome in the so-called custom annotation track formats. In this way, the relevant results (profile of the haplotype or nucleotide diversity along the DNA sequence) can be visualized integrating available genome features (genes, repetitive or intergenic regions, *etc*.).

Data analysis: We tested the performance of the methods implemented in the VariScan software by analysing two qualitatively different data sets: i) a computer-simulated data set generated by applying coalescent methods, and ii) SNP data from the Mouse Genome Resequencing Project (http://www.niehs.nih.gov/crg/cprc.htm) and from the PATIL *et al.* (2001) study in human. MRA analysis conducted using windows of 1 bp captures all information of the data. Small windows, however, increase the computational RAM-time requirements, and in fact are not strictly necessary. However, we can use larger windows without losing interesting features. Even so, unlike the SW analyses, the MRA results are nearly independent of the chosen window length. Moreover, the SW would likely fail in detecting small-size features at the whole genome scale. For the MRA analysis, the optimal window size to detect most of the interesting features will depend on the current nucleotide diversity values and on of the sample size of the study. These values will be the input (the signal) for the MRA. From a practical standpoint, analysis of 10–30 sequences may be conducted by using non-overlapping windows of 50–500 bp for per-site θ values of 0.01, up to 500–5000 bp for $\theta = 0.001$, as in Drosophila and humans, respectively.

Computer-simulated data set: We generated random data sets based on the simplest non-recombining coalescent model (HUDSON 1990) as follows: i) generation of evolutionary times



UCSC Genome Browser on Mouse Mar. 2005 Assembly

position/search chr11:65,000,001-85,000,000 size 20,000,000 bp.

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options.

move start 2.0

move end 2.0

Chromosome Color Key:

Use drop down controls below and press refresh to alter tracks displayed.

Tracks with lots of items will automatically be displayed in more compact modes.

Figure 3.3: Visualization on the UCSC browser of the MRA analysis based on θ values from the mouse genome resequencing project data. The USCS browser shows a 20 Mb-region (within positions 65,000,001–85,000,000). The first two tracks (customer tracks) represent the signal reconstruction of low-frequency bands with information from 9 to 11 MRA levels (first track), and from 12 to 16 MRA levels (second track).

and the gene genealogy (fixing the number of sequences); ii) incorporation of Poisson-randomly distributed mutations (fixing the population mutation parameter θ). Subsequently, we modify this data set by changing (at specific locations) the applied θ value. In particular, we reduced nucleotide diversity values continuously and symmetrically. We made changes at two different levels: iii) one or more chromosome-wide nucleotide diversity reductions; iv) additional reductions at narrow regions. These changes were conducted by using different intensity (parameter α; the degree of nucleotide diversity reduction) and stretch lengths (parameter β; β specifies the half-length of the affected region) values. Therefore, the simulated data set mimics the effects caused by partial selective sweeps upon different nucleotide diversity levels. The analysis of one of these simulated data files is given in Figure 3.4. It can be seen that the MRA technique recovers the two different intensity types of distorted regions included in the data: nucleotide diversity reductions affecting small DNA stretches are detected at lower MRA levels while more genome-wide reductions are identified at higher levels.

DNA polymorphism data from the Mouse Genome Resequencing Project: The Mouse Genome Resequencing Project (http://mouse.perlegen.com/mouse/index.html) is conducting a genome-wide DNA resequencing survey in 15 inbred strains of mice using an array-based resequencing technology. In spite that the project in not finished yet, some chromosomes are quite well covered. Here, we use VariScan to analyse the levels of nucleotide diversity along the chromosome 11 (121,803,636 bp; NCBI build 34 which corresponds to the UCSC release of March 2005). Since the polymorphism data were determined in inbred strains (and therefore homozygous) we will consider one sequence per strain (*i.e.* the sample size is 15). The mouse chromosome 11 data set contains 262,988 SNPs; not all of these SNPs, nevertheless, were typed in all 15 strains because of experimental errors (the average number of missing chromosomes per site was 2.30; and only 91,119 SNPs were typed in all 15

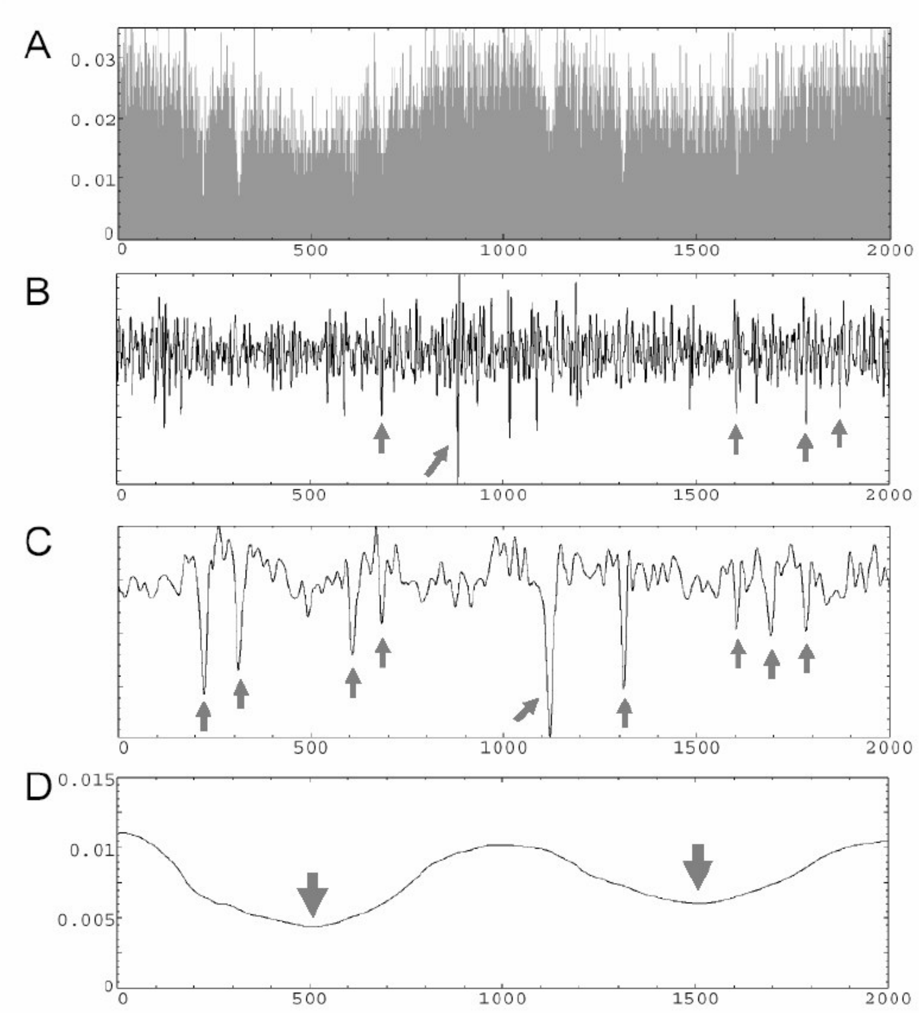


Figure 3.4: Application of the MRA analysis to the coalescent-simulated data set. The data contains 10 sequences of 2,000,000 bp each, and it was generated applying a per-site value of θ = 0.01. Upon this raw data set, we made two different levels of changes: i) two wide reductions in nucleotide diversity levels (g_1: α = 1/3, β = 500,000; g_2: α = 1/2, β = 500,000); and ii) 11 local valleys of reduced variability (v_1: α = 1/4, β = 20,000; v_2: α = 1/4, β = 15,000; v_3: α = 1/4, β = 10,000; v_4: α = 1/4, β = 5,000; v_5: α = 1/4, β = 2,000; v_6: α = 1/3, β = 20,000; v_7: α = 1/3, β = 10,000; v_8: α = 1/3, β = 5,000; v_9: α = 1/2, β = 10,000; v_{10}: α = 1/2, β = 5,000; v_{11}: α = 1/2, β = 2,000). (a) nucleotide diversity profile obtained by SW using non-overlapping windows of 50 bp; (b) Signal reconstruction of low-frequency bands with information from 7 to 8 MRA levels, showing the location (in arrows) of 5 depleted-variation regions ($v_{4–5}$, v_8, $v_{10–11}$; β ≤ 5,000). c) Signal reconstruction from 9 to 12 MRA levels, showing the location (in arrows) of 9 depleted-variation regions ($v_{1–4}$, $v_{6–10}$; 5,000 ≤ β ≤ 20,000). d) Signal reconstruction from 13 to 15 MRA levels, showing the location (in arrows) of the two broad areas with reduced levels of variation ($g_{1–2}$; β = 500,000). The nucleotide sequence positions (X axis) are given in kb.

strains). Estimates of nucleotide diversity (π_m) were $\pi_m = 0.00072$. Nonetheless, since many repetitive regions of the chromosome were not completely resequenced, current nucleotide diversity values likely are underestimated. Nucleotide diversity values along the chromosome, nevertheless, contain much more information than the global π values. For instance, the SW method allows identifying constrained regions, and it could facilitate the detection of the distinctive fingerprint of positive selection. The MRA analysis is clearly a much more useful method for detecting specific genomic features at different scales. Additionally, the results of these analyses can be visualized integrated with current genome annotations using available genome browsers (Figure 3.3). The MRA analysis revealed a strong heterogeneous nucleotide diversity profile along the DNA region, including a number of peaks and valleys. Although it is premature to determine the evolutionary meaning of these regions, the joint visualization of the MRA results with current annotated genomic features (genes, haplotype information, etc) is a comprehensive tool for their characterization and further understanding.

Human chromosome 21 data set: This data set (PATIL *et al.* 2001) contains the 35,989 SNPs identified in the survey of 32.4 Mb (21.7 Mb after excluding repetitive-masked positions; nearly all human chromosome 21) in 20 ethnically diverse individuals using high-density oligonucleotide arrays. However, for an easy and comprehensible interpretation of the results we do not use this raw data. First, we excluded all singletons variants because the used array-based technology had little power in their identification. Second, we only analysed SNPs confirmed in the NCBI build 34 of the human genome (PATIL *et al.*'s data were based on an older NCBI build). Third, we focused the analysis on SNPs located in the longest contig (NT_002836; named NT_011512 in NCBI build 34 release) of PATIL *et al.*'s data, since there were missing regions between contigs. In total, we analyzed 21,218 SNPs (there were 21,840 in PATIL *et al.*'s data) in a region of 28.6 Mb long (the net number of sites *l* was 19.1 Mb after excluding repetitive-masked positions) in 20 individuals.

For the total NT_011512 contig data, only 2097 SNPs (10%) were typed in all 20 chromosomes, resulting on 3.87 missing chromosomes per site. Estimates of nucleotide diversity (π_m) was $\pi_m = 0.00044$. This value is lower than that reported in PATIL *et al.*'s study ($\pi = 0.00072$) (see also INNAN *et al.* 2003); these estimates, however, are not completely comparable because we are using only a subset of PATIL *et al.*'s data. Particularly, we have not taken into account singleton information, while the expected frequency of singletons (mutations occurring on the external branches of the genealogy) for a neutrally evolving region in a sample of 20 sequences is 0.297 (0.321 if we consider that the net number of chromosomes is 16). Thus, roughly 30% of the SNPs should be singletons, although the actual value is likely higher since many human regions have negative Tajima's *D* values. Considering this 30% as the true percentage of singletons in the sample, the π_m estimates for the total contig would be 0.00050.

DISCUSSION

Detecting the action of positive natural selection is critical to understand and identify the evolutionary forces that have shaped organismal traits and genomes. Despite the profound implications in evolutionary biology and in medicine currently there are few convincing evidences of the action of positive selection. Since purifying selection weeding out deleterious mutations operates continuously, their detection had been much easier. Indeed, the detection of evolutionarily conserved regions has been proven to be a very effective method for the identification of functionally important regions, such as regulatory elements. The detection of the distinctive signature of natural selection can, nevertheless, be detected by analysing the spatial distribution of polymorphisms across the genome; essentially, positive natural selection causes a distinctive fingerprint on the pattern of nucleotide variation both in the target of selection but also in their surrounding regions. For instance, the selective sweep (or hitchhiking effect) produced when selection drives an advantageous mutation to fixation, will affect variation at relatively short DNA sequence stretches (of some kb; the magnitude of the effect is determined by the relative strength of selection and recombination) (KAPLAN *et al.* 1989, KIM and STEPHAN 2002). On the other hand, demographic effects will have a genome-wide signature. The identification of the specific regions evolving under natural selection at the genome scale requires, however, new analytical methods and bioinformatics tools. In spite of the impressive recent development of such methods (HUDSON 2002), nevertheless, they are not fully adequate for a genome-wide analysis.

In this context, VariScan software overcomes many limitations of current software and methods, and it is useful as an exploratory tool in the analysis of DNA polymorphism at the genome scale. VariScan can handle the vast amount of DNA polymorphism data generated by large genome-based projects, and implements efficient methods, such as SW and MRA, to determine the common patterns of nucleotide variation and to identify specific features, along large (chromosome-wide) DNA fragments. The SW has been extensively used in DNA polymorphism studies for exploratory data analysis (ROZAS et al 2003). This method allows obtaining a relevant parameter profile (*e.g.* nucleotide or haplotype diversity, linkage disequilibrium) along a DNA region and, therefore, is instrumental in detecting the distinctive footprint of natural selection, mainly in genome wide-based analysis. Unfortunately, the determination of the appropriate window size represents an important limitation of the method. This is a critical point because the accuracy of extracting features from DNA sequence data (*i.e.*, the signature of natural selection) strongly depends on the window size. Although there have been some statistical attempts to determine the window size (TAJIMA 1991, FARES *et al.* 2002), the usual approach is by trial-and-error. The MRA-based analysis, on the contrary, can be used to detect genomic features even at different resolution scales; for example, features in various nucleotide diversity backgrounds. Therefore, the method can be helpful in detecting relevant features from DNA polymorphism data at a genome-wide scale, such as conserved regions, peaks and valleys of nucleotide diversity, linkage disequilibrium clusters, *etc*. that, in turn, might reveal the distinctive footprint left by the action of natural selection.

CONCLUSION

In summary, the version 2 of the VariScan software implements new methods and features for an exhaustive DNA sequence polymorphism analysis at the genome-wide scale. We have tested the performance of the methods implemented in the software by analysing computer-simulated and real data sets.

AVAILABILITY AND REQUIREMENTS

Project name: VariScan

Project home page: http://www.ub.es/softevol/variscan.

Source code, executables and documentation are available from this site.

Operating system(s): Linux, Mac OSX, Windows

Programming languages: ANSI C, Java, Perl

Other requirements: Java 1.4 or higher, Perl 5.6 or higher

License: GNU GPL

ACKNOWLEDGEMENTS

We are very grateful to M. Aguadé, S. O. Kolokotronis, S. Ramos-Onsins, M. Schuster and C. Segarra for their valuable comments and suggestions. This work was supported by grants BMC2001–2906 and BFU2004–02253 from the Dirección General de Investigación Científica y Técnica (DGICYT) of Spain.

LITERATURE CITED

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

ARNEODO, A., E. BACRY, P. V. GRAVES and J. F. MUZY, 1995 Characterizing long-range correlations in DNA sequences from wavelet analysis. Physical Review Letters **74:** 3293-3296.

DAUBECHIES, I., 1992 Ten lectures on wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia.

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

EXCOFFIER, L., G. LAVAL, S. SCHNEIDER, 2005 Arlequin (version 3): An integrated software package for population genetics data analysis. Evol Bioinformatics Online **1:** 47-50.

FARES, M. A., S. F. ELENA, J. ORTIZ, A. MOYA and E. BARRIO, 2002 A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J Mol Evol **55:** 509-521.

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

FILATOV, D.A., 2002 ProSeq: A software for preparation and evolutionary analysis of DNA sequence data sets. Mol Ecol Notes **2:**621-624.

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.

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.

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

HUDSON, R.R., 1990 Gene genealogies and the coalescent process. Oxf Surv Evol Biol **7:**1-44.

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

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

HUGHES, A. L., and M. YEAGER, 1998 Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev Genet **32:** 415-435.

INNAN, H., B. PADHUKASAHASRAM and M. NORDBORG, 2003 The pattern of polymorphism on human chromosome 21. Genome Res **13:** 1158-1168.

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

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

KENT, W. J., C. W. SUGNET, T. S. FUREY, K. M. ROSKIN, T. H. Pringle *et al.*, 2002 The human genome browser at UCSC. Genome Res **12:** 996-1006.

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

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

KINGMAN, J.F.C., 1982 On the genealogy of large populations. J Appl Prob **19A:** 27-43.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature **304:** 412-417.

LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49-67.

LIÒ, P., 2003 Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics **19:** 2-9.

LIÒ, P., and M. VANNUCCI, 2000 Finding pathogenicity islands and gene transfer events in genome data. Bioinformatics **16:** 932-940.

MACDONALD, S. J., and A. D. LONG, 2005 Identifying signatures of selection at the enhancer of split neurogenic gene complex in *Drosophila*. Mol Biol Evol **22:** 607-619.

MALLAT, S.G., 1989 A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell **11:** 674-693.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351:** 652-654.

MEKEL-BOBROV, N., S. L. GILBERT, P. D. EVANS, E. J. VALLENDER, J. R. ANDERSON *et al.*, 2005 Ongoing adaptive evolution of ASPM, a brain size determinant in *Homo sapiens*. Science **309:** 1720-1722.

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

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol **3:** e196.

ORENGO, D. J., AND M. AGUADÉ, 2004 Detecting the footprint of positive selection in a european population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. Genetics **167:** 1759-1766.

PATIL, N., A. J. BERNO, D. A. HINDS, W. A. BARRETT, J. M. DOSHI *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294:** 1719-1723.

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

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

ROZAS, J., and R. ROZAS, 1995 DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. Comput Appl Biosci **11:** 621-625.

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

SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. LEVINE, D. J. RICHTER *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832-837.

STEIN, L. D., C. MUNGALL, S. SHU, M. CAUDY, M. MANGONE *et al.*, 2002 The generic genome browser: a building block for a model organism system database. Genome Res **12:** 1599-1610.

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

TAJIMA, F., 1991 Determination of window size for analyzing DNA sequences. J Mol Evol **33:** 470-473.

VILELLA, A. J., A. BLANCO-GARCIA, S. HUTTER and J. ROZAS, 2005 VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics **21:** 2791-2793.

Curriculum Vitae

Personal Details

Name:	Stephan Hutter
Date of birth:	November 26th 1977
Place of birth:	Salzburg, Austria
Nationality:	Austrian
Marital status:	Single
Current work address:	LMU Biozentrum Grosshaderner Strasse 2 82152 Planegg-Martinsried, Germany
Telephone:	++49 - 89 - 2180 74 101
Fax:	++49 - 89 - 2180 74 104
E-Mail:	hutter@zi.biologie.uni-muenchen.de

Education

Jun 2003 – May 2007	University of Munich (LMU), Munich, Germany PhD student in biology
Sep 1997 – Apr 2003	University of Munich (LMU), Munich, Germany Diploma in biology
Oct 1996 – Aug 1997	University of Technology (TU), Munich, Germany Studies of computer science
Sep 1987 – Jun 1996	Erasmus-Grasser-Gymnasium, Munich, Germany High school graduation

Additional Research and teaching Experience

Oct 2002 – Dec 2002	University of Munich (LMU), Munich, Germany Teaching assistant, Mathematics for Biologists lecture
Jul 2001 – Oct 2001	University of Munich (LMU), Munich, Germany Research assistant, Section of Evolutionary Biology

Fellowships

Jan 2004 – Jun 2004	University of Barcelona (UB), Barcelona, Spain Marie-Curie-Fellowship, awarded by the European Commission

List of Publications

HUTTER, S., H. LI, S. BEISSWANGER, D. DE LORENZO and W. STEPHAN, 2007 Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosome-wide SNP data. Genetics doi: 10.1534/genetics.107.074922.

HUTTER, S., A. J. VILELLA and J. ROZAS, 2006 Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics **7:** 409.

VILELLA, A. J., A. BLANCO-GARCIA, S. HUTTER and J. ROZAS, 2005 VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics **21:** 2791-2793.

General Acknowledgements

First, I would like to thank Prof. Wolfgang Stephan for the opportunity to conduct my PhD research in the field of population genetics under his supervision. I would also like to thank Prof. John Parsch for collaborating with me on the microarray project and sharing his rich insight.

I owe many thanks to my co-authors Steffen Beisswanger, Haipeng Li and Sarah Saminadin-Peter. Without their contributions the exiting results presented in my thesis would not have been possible. Muchas Gracias go out to my Spanish colleagues! To David de Lorenzo for creating a wonderful atmosphere in the lab and to Julio Rozas and Albert Vilella for the more than pleasant time I had in Barcelona.

I thank John Baines, Sascha Glinka, Lino Ometto, Pleuni Pennings and Nicolas Svetec not only for their helpful scientific discussion, but also for the many cheerful beer garden visits. Of course I must also thank all other members of the Department of Evolutionary Biology: assistants, postdocs and students. Everybody has been very kind and helpful during the four years of my PhD project.

I would like to express my gratitude towards all the technicians that have helped me collect the enormous amounts of data and keep the flies happy: Kawsar Bhuiyan, Alexandra Fabry, Traudl Feldmaier-Fuchs, Bea Stiening and Anne Wilken. Thank you! I would also like to thank Katrin Kümpfbeck for helping me find my way through the bureaucratic system of the university.

Last, but certainly not least, I would like to thank my parents for their ongoing support throughout all these years.