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

Evolution of gene expression and gene-regulatory sequences in *Drosophila melanogaster*

Sarah Sylvie Saminadin-Peter
aus
Basse-Terre, Guadeloupe
2008

Erklärung

Diese Dissertation wurde im Sinne von §12 der Promotionsordnung von Prof. Dr. John Parsch 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 Dissetation von mir selbständig, ohne unerlaubte Hilfe angefertigt wurde.

München, 04.11.2008

Sarah Saminadin-Peter

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

Dissertation eingereicht am: 04.11.2008
Mündliche Prüfung am: 05.12.2008

Table of contents

List of tables and figures

Note

In this thesis I present my doctoral research, all of which has been done by me except for the following: In chapter 1, Stephan Hutter performed half of the microarray experiments (those within the European population) and also the BAGEL and GO statistical analyses. Both Stephan Hutter and John Parsch contributed to writing the manuscript that served as the basis for chapter 1. In chapter 3, the polymorphism and divergence analysis of non-coding regions was performed with the help of a computer program developed by Sergej Nowoshilow.
The simulation of demographic and neutral scenarios described in chapter 4 was performed by Pavlos Pavlidis.

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

Hutter, S.*, **S. Saminadin-Peter***, W. Stephan and J. Parsch (2008). "Gene expression variation in African and European populations of *Drosophila melanogaster*." Genome Biology **9**(1): R12
* The author contributed equally to this work.

List of Abbreviations

ADH	Alcohol dehydrogenase
α	Strength of selection
BAGEL	Bayesian analysis of gene expression levels
bp	Base pair
CI	Confidence interval
CLR	Composite likelihood ratio
Ct	Threshold cycle value
D	Divergent sites between species
DDT	Dichloro-diphenyl-trichloroethane
DGRC	Drosophila genomics resource center
f	Female data
FAD	Flavine adenine dinucleotide
GMC	Glucose-methanol-choline
GO	Gene Ontology
GOF	Goodness-of-fit
H	Fay and Wu's *H*
HKA	Hudson-Kreitman and Aguadé
I	Intron
K	Divergence between species
kb	Kilobases
L	Sequence length (in bp)
m	Male data
MK test	McDonald-Kreitman
n	Sample size
NA	Not available or not applicable
NADP	Nicotinamide adenine dinucleotide phosphate
Ne	Effective population size
NI	Neutrality index
ns	Non-synonymous sites
NSG	Non-significant Genes, those with no expression difference between the European and African populations
P	Polymorphic sites

π	Nucleotide diversity based on the average pairwise differences
qPCR	Quantitative real-time PCR
r	Recombination rate
R	Correlation coefficient
s	Synonymous sites
S	Number of segregating sites
SD	Standard deviation
SF	SweepFinder
SG	Significant Genes, those with significant expression differences between the European and African populations
θ	Nucleotide diversity based on S
TD	Tajima’s *D*
UTR	Untranslated region
5’	Region upstream of the start codon ATG

Zusammenfassung

Anhand dieser Arbeit, untersuchte ich die Rolle von genregulatorischen Veränderungen in der Evolution von *Drosophila melanogaster*. Der erste Schritt beinhaltete eine Studie der Variation der Genexpression mittels „whole-genome“ Microarrays. Ich untersuchte acht Stämme aus einer Urpopulation aus Afrika und acht Stämme aus einer davon abstammenden Population aus Europa. Der experimentelle Aufbau erlaubte es mir, sowohl Expressionsunterschiede innerhalb einer Population als auch zwischen den Populationen zu erfassen. Die Höhe der Variation der Genexpression innerhalb der beiden Populationen war nahezu gleich, dagegen eine höhere Variation wurde zwischen den beiden Populationen gemessen. Der überwiegende Anteil der Variation der Genexpression innerhalb der beiden Populationen wird durch „stabilizing selection“ limitiert. Jedoch einige Gene, welche zwischen den Populationen differentiell exprimiert sind, könnten Ziele von positiver Selektion sein. Diese kodierenden Proteine sind in Prozesse wie Insektizidresistenz, Wahl der Nahrungsquelle, Lipidmetabolismus und Flug involviert. Diese Gene sind gute Kandidaten um adaptive regulatorische Evolution, welche mit der aus-Afrika Migration von *D. melanogaster* verbunden ist, zu untersuchen.

Um die Genauigkeit des Microarray-Experimentes zu verifizieren, untersuchte ich die Variation der Genexpression mittels quantitativer Real-Time PCR (qPCR) in einer Teilmenge von Genen, welche für das Microarray-Experiment verwendet wurde. Die qPCR ist eine weitere Methode zur Messung der Genexpression. Ich habe die „fold-changes“ der Genexpression zwischen den Paaren von Stämmen mittels beider Methoden verglichen. Zusätzlich habe ich das Muster der Variation der Genexpression zwischen männlichen und weiblichen Fliegen verglichen. Der qPCR-Ansatz hat die generelle Genauigkeit des Microarray-Experimentes bestätigt und die gemessenen „fold-changes“ der beiden Methoden waren sehr stark übereinstimmend. Die Expressionsunterschiede zwischen den Stämmen tendierten dazu relativ gleich zwischen männlichen und weiblichen Fliegen zu sein. Jedoch konnten Ausnahmen dieses generellen Muster beim paarweisen Vergleich der „fold-changes“ einzelner Gene gefunden werden, welche Unterschiede in der Expression zwischen männlichen und weiblichen Fliegen zeigten.

Ich untersuchte auch die molekulare Evolution und Populationsgenetik von protein-kodierenden und stromaufwärts gelegenen regulatorischen Regionen von Genen, welche Anzeichen von adaptiver Evolution auf der Ebene der Genregulation zeigten. Diese Gene repräsentieren eine Teilmenge der Gene, welche signifikante Unterschiede in der Genexpression zwischen der afrikanischen und europäischen Population zeigten. Eine Anzahl von Kontrollgenen, welche keine signifikanten Unterschiede in der Genexpression zwischen den beiden Populationen zeigten, wurde auch in die Analyse integriert. Zusammenfassend habe ich Anzeichen für positive Selektion als auch „purifying selection" in kodierenden und nicht-kodierenden Regionen gefunden. Jedoch, die Muster der Polymorphismen und Divergenzen zeigten keine signifikanten Unterschiede zwischen Kandidatengenen und Kontrollgenen.

Eines der Gene, welches ein interessantes Expressionsmuster im Microarray und qPCR Experiment zeigte, war Bestandteil von weiteren populationsgenetischen Untersuchungen. Dieses Gen *CG9509* hatte eine zweifach höhere Expression im europäischen Stamm als im afrikanischen Stamm. Die kodierende und stromaufwärts gelegene Region dieses Gen zeigt Anzeichen von wiederkehrender positiver Selektion, seit der Spaltung von *D. melanogaster* und ihrer Schwesterspezies *D. sechellia*. Eine Untersuchung der Polymorphismen der *CG9509* Region enthüllte ein 1,2 kb großes Segment, welches die mutmaßliche Promotorregion des Genes beinhaltet und keine Polymorphismen in der europäischen Population zeigte. Die europäische Population hat mehrere fixierte und nahezu fixierte abgeleitete Mutationen in dieser Region. Diese Beobachtungen verbunden mit der statistischen Analyse unterstützt das Anzeichen eines „selective sweep" in der europäischen Population. Der „selective sweep" wurde wahrscheinliche durch lokale Adaption auf dem Level der Genexpression hervor gerufen.

General Introduction

Understanding the evolutionary forces that shape the diversity within and among species is an important aim in biology and population genetics. While the environment makes an important contribution to differences among individuals, it is the heritable, genetic differences that are of greatest interest to population geneticists and evolutionary biologists. This genetic diversity is influenced both by processes that affect the entire genome, such as demographic events, and by processes that act on particular regions of the genome to modify the fitness of the organism, such as natural selection. Charles Darwin, in his famous 1859 treatise “On the Origin of Species”, introduced the notion of evolution by means of natural selection. However, the mechanism of heritability of variable characters that explains their maintenance in the population was first described by Gregor Johann Mendel in his 1865 paper entitled “Versuche über Pflanzenhybriden” (Experiments in Plant Hybridization).

Darwin’s studies of diversity were limited to the morphological phenotypic variants that could be distinguished easily among individuals. As technology progressed, it became possible to study variation at the molecular level. For example, Efemov and Braend (1964) used starch gel electrophoresis to demonstrate protein polymorphism within humans. This technique was widely used in the 1960’s and 1970’s to reveal levels of protein polymorphism within a large number of species. Given the nature of the genetic code, it was assumed that polymorphism seen at the level of protein sequence was the result of underlying DNA sequence differences, although the specific details of the DNA variation were not known until DNA sequencing methods were developed. For example, DNA sequencing of the gene encoding the *Drosophila melanogaster* alcohol dehydrogenase (ADH) protein showed that amino-acid differences result from single nucleotide differences at nonsynonymous sites in the *Adh* gene and that many additional variants were present at synonymous (silent) sites in the gene, which did not alter the protein’s amino acid sequence (Kreitman 1983). The rapid improvement of DNA sequencing technologies resulted in the availability of entire genome sequences for several eukaryotic organisms and also led to the development of high throughput methods to analyze gene expression on a genomic scale. The standard method for this consists of the use of competitive hybridizations of cDNA to high density microarrays in order to the relative expression level of a large number of genes between two samples or individuals (Lockhart *et al.* 1996) (Figure 1).

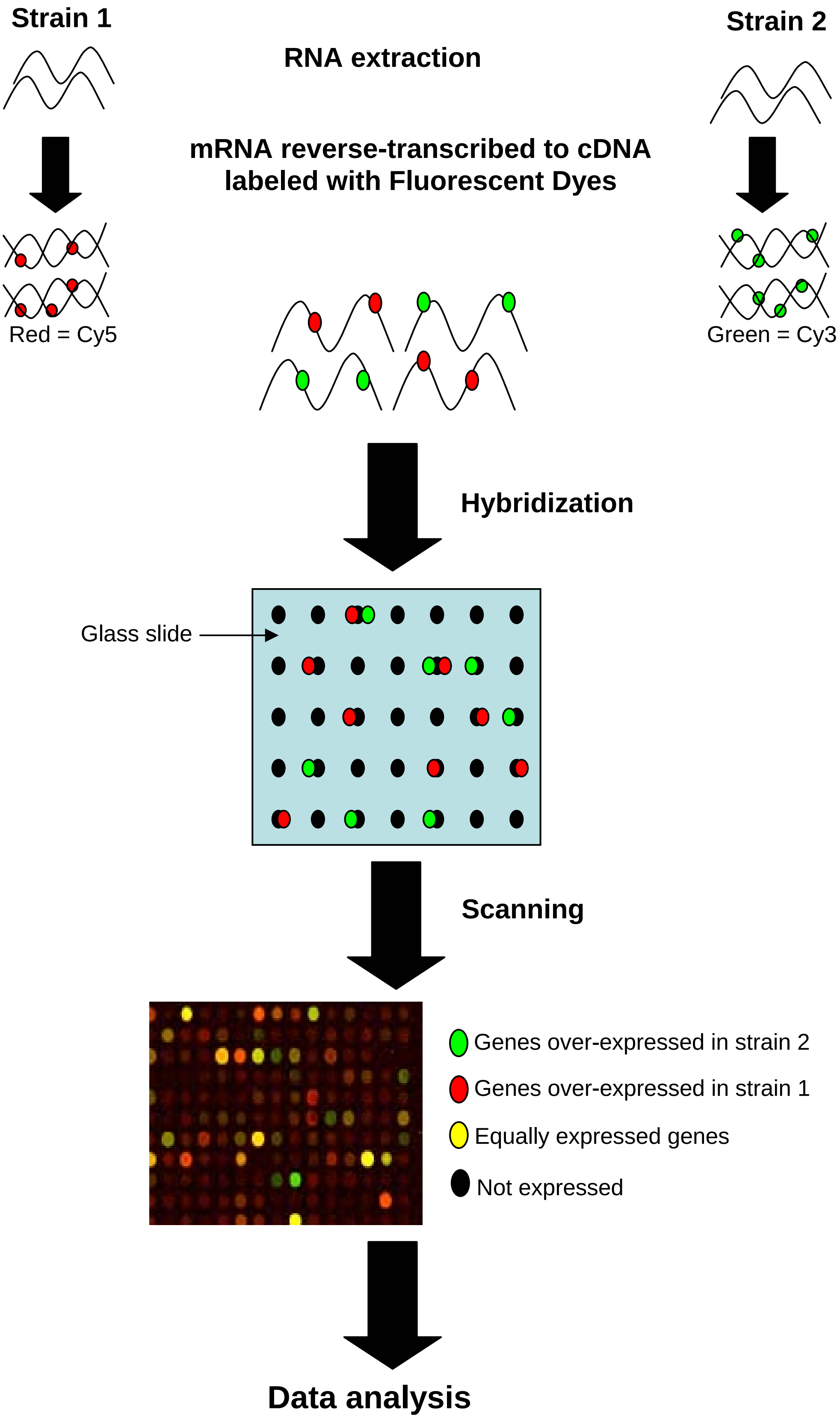

Strain 1
RNA extraction
Strain 2
mRNA reverse-transcribed to cDNA
labeled with Fluorescent Dyes
Red = Cy5
Green = Cy3
Hybridization
Glass slide
Scanning
🟢 Genes over-expressed in strain 2
🔴 Genes over-expressed in strain 1
🟡 Equally expressed genes
● Not expressed
Data analysis

Figure 1 : Principle of cDNA microarray assay.

As microarray data became available, it became clear that variation in levels of gene expression was also abundant within and between species, and highlighted the potential for gene regulatory changes to be targets of natural selection.

Beginning in 1969, Britten and Davidson pointed out the relative importance of regulatory changes in adaptation and species differentiation. In fact, DNA sequence analysis revealed a nearly 99% identity between human and chimpanzee. However, the great morphological and the cognitive differences between these two species are undeniable. In the 1970's, King and Wilson (1975) suggested that the phenotypic differences between the two species were primarily due to changes in gene regulation. An initial study based on microarray technologies suggested that the main difference between humans and chimpanzees was in the genes expressed in the brain (Enard *et al.* 2002; Khaitovich *et al.* 2005). However, this finding did not hold up and later microarray experiments suggested that the largest difference was in genes expressed in testis, while the smallest difference was in genes expressed in the brain. Differences in gene expression are also the major cause of sexual dimorphism, the phenotypic differences between males and females (Parisi *et al.* 2004). The importance of gene regulatory changes in environmental adaptation has also been documented. For example, resistance to the insecticide DDT was found to be linked to an allele of the cytochrome P450 gene, *Cyp6g1*. Microarray analysis indicated that *Cyp6g1* expression is three times higher in resistant flies than in susceptible flies (Daborn *et al.* 2002). More generally, extensive gene expression variation has been found between and within several species (Whitehead and Crawford 2006). In this thesis, I mainly concentrate on variation at the level of gene expression and its underlying genetic and evolutionary causes.

Over fifty years ago, Jacob and Monod (1961) first speculated about the evolutionary importance of non-coding regions. The non-coding DNA occupies the vast majority of most eukaryotic genomes and previous surveys have shown that a relatively large fraction of non-coding DNA is conserved among species (Waterston *et al.* 2002; Siepel *et al.* 2005; Stark *et al.* 2007). In 2005, Andolfatto, used two closely-related species of Drosophila to show that many types of non-coding DNA evolve more slowly (*i.e.*, are under greater selective constraint) than synonymous sites. Andolfatto (2005) also used polymorphism data to estimate that 40–70% of the interspecific differences in non-coding DNA were driven by positive selection. This study confirmed previous results from Kohn *et al.* (2004) who used the McDonald-Kreitman (1991) approach to estimate that ~50% of all substitutions in the 700 bp upstream of genes had been fixed by positive selection. These findings argue for the functional importance of non-coding DNA. These non-coding regions may contain regulatory elements, such as transcription factor binding sites, transposable element insertions, or small RNAs. In general, regulatory elements can be classified into two types: *cis*- and *trans*-elements. The *cis*-elements are located near the gene they regulate. They can be found near the transcription start site or in an enhancer located in the non-coding sequences surrounding the transcribed region (Figure 2). They can alter the transcription rate and/or the half-life of the transcript (RNA stability) (Wray *et al.* 2003). The *trans*-elements are mainly transcription factors that are unlinked to the genes they regulate (Figure 2). Changes in *trans*-factors often have a pleiotropic effect, as they affect the expression of many genes.

The relationship between the two types of elements is complex and determining how changes at the DNA level change the phenotype remains an exciting challenge. Several studies argue in favor of the predominant role of *cis*-regulatory changes to explain phenotypic diversity (Wittkopp *et al.* 2004; Stranger *et al.* 2005; Osada *et al.* 2006). In the review of Wray (2007), several examples highlight the phenotypic consequences of *cis*-regulatory mutations (see Table 1 from Wray 2007). One main reason for this is that *cis*-mutations are more readily studied at molecular level. In fact, they are less difficult to pursue than *trans*-mutations because they are physically linked to the gene that they regulate. In addition, the pleiotropic effects of mutations in *trans*-acting factors often makes them difficult to study.

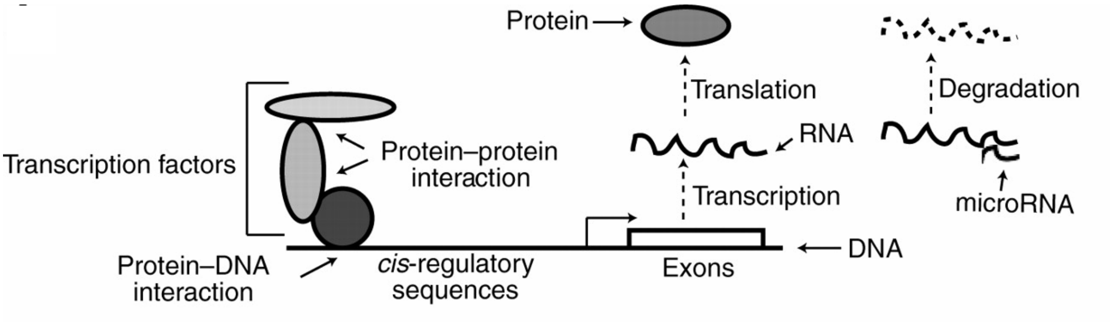


Figure 2: The interactions between transcription factor proteins (*trans* elements) and *cis*-regulatory DNA sequences, from (Wittkopp 2007).

To attribute functional significance to non-coding DNA, it is important to demonstrate the non-neutral evolution (Kimura 1983) of such regions. The signature of selective constraint is manifested by a reduction of polymorphism and divergence, as well as an excess of rare variants (Figure 3). As a neutral control for these measures, synonymous sites are typically used. The McDonald and Kreitman (1991) test, which compares divergence between species to polymorphism within species, can also be applied to detect differences in the evolution of neutral sites (synonymous sites) and putatively selected sites (non-synonymous or non-coding regions). In this case, a relative excess of interspecific divergence at the tested sites is taken as support for their adaptive evolution. Other neutrality tests, such as Tajima's *D* (1989) and the HKA test (Hudson *et al.* 1987), can also be applied.

Although statistical analyses of sequence variation within and between species can provide evidence for the contribution of non-coding DNA to phenotypic variation and adaptation, the ultimate test is to get direct experimental evidence for the functional significance of a putative regulatory region. One such approach is to perform reporter assays that specifically test the effects of a putative regulatory sequence on gene expression. A limitation of this technique, which has been used successfully in some cases, is mainly that the reporter assays are not sensitive enough to detect small differences in expression (Bird *et al.* 2006). An alternative approach to detect genetic variants that affect gene expression is association mapping to find eQTLs (expression quantitative trait loci) (Cowles *et al.* 2002). In the future, we can assume that technologies will be developed that are more sensitive to small expression changes and a wide range of experimental methods will be available to functionally test for allele- or haplotype-specific effects on gene expression.

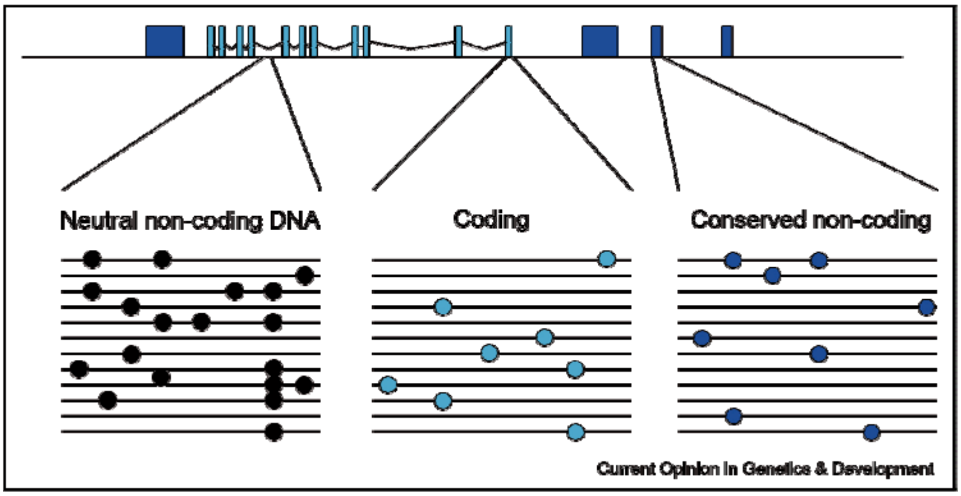


Figure 3 : Distribution of new variants in haplotypes in three different classes of sequences (neutral non-coding, coding and conserved non-coding). This figure illustrates the predominant signal of purifying selection in functional sequences, from (Bird *et al.* 2006).

In this thesis, I used the fruitfly, *Drososphila melanogaster*, as a model system to study gene expression evolution. *D. melanogaster* is used in many areas of biological research, mainly because it is easy to maintain in the laboratory and has a short generation time of approximately 10 to 14 days. In addition, this species is a human commensal and easy to find worldwide. This enables researchers to study the behavior, population structure, demography, and adaptive history of *D. melanogaster*. In 1988, David and Capy suggested a sub-Saharan origin of this species and its recent migration to temperate regions after the last glaciation (10,000 to 15,000 years ago). The demographic history of *D. melanogaster* has been studied extensively (Glinka *et al.* 2003; Haddrill *et al.* 2005; Ometto *et al.* 2005; Li and Stephan 2006; Pool and Aquadro 2006), which provides background information on the non-selective forces that affect levels of nucleotide diversity. A genome scan performed on the X chromosome revealed that the nucleotide variation is higher in the African population, as would be expected for an ancestral population (Glinka *et al.* 2003; Ometto *et al.* 2005). Migration to new habitats, such as the temperate regions of Europe, was likely accompanied by adaptation to new biotic and abiotic factors, including differences in temperature, humidity, food, and pathogens. By comparing a putatively ancestral population from Africa (Zimbabwe) with a derived population from Europe (the Netherlands), we have a powerful and unique opportunity to look for both traits and genes that have been involved in the process of adaptation to new environmental conditions.

In this thesis, I address several basic questions about gene regulatory evolution, which can be summarized as follows:

1) *How much gene expression variation is present within and between populations? Are there fixed expression differences between derived and ancestral populations? (Chapter 1)*

We performed a gene expression variation survey of *Drosophila melanogaster* using whole-genome microarrays. We surveyed eight strains from an ancestral African population and eight strains from a derived European population following an experimental design that allowed us to detect significant expression differences within and between populations. We find nearly the same level of gene expression variation within the two populations and a higher amount between the two populations. Most gene expression variation within populations seems to be limited by the action of stabilizing selection. However, some genes that are differentially expressed between the two populations might be targets positive selection. These genes are good candidates for studying adaptive regulatory evolution that accompanied the out-of-Africa migration of *D. melanogaster*.

2) *Can the microarray results be confirmed using another method to measure gene expression? Does the pattern of expression observed for adult males also hold for adult females? (Chapter 2)*

Quantitative Real-Time PCR (qPCR) was used to validate the microarray experiments for a subset of genes analyzed in Chapter 1. I compared the fold-changes in gene expression between pairs of strains determined by the two methods. I also compared the pattern of expression variation in male and females flies. The qPCR approach supported the general accuracy of the microarray experiments. Expression differences among the strains tended to be similar for male and females. However, exceptions to this general pattern could be found by looking at the pairwise fold-changes for individual genes, some of which differed in expression pattern between males and females.

3) *Is there evidence for selection on the coding and non-coding regulatory regions of the candidate genes for regulatory adaptation? (Chapter 3)*

From a subset of genes that showed a significant difference in gene expression between the African and European populations, I investigated the molecular evolution of the protein-encoding and upstream regulatory regions. I also performed the same analysis on a set of control genes with no significant difference in expression between the two populations. I found that the coding and the regulatory regions showed evidence of both positive and purifying selection. However, the selective pressures seem to be the same for the differentially expressed genes and the control genes.

4) *Is there evidence of recent local adaptation (i.e., a selective sweep) associated with gene expression changes in the European population? (Chapter 4)*

The microarray and qPCR analyses uncovered a gene with an interesting pattern of gene expression. This gene, *CG9509,* has twofold higher expression in the European strains than in the African strains. The coding and the upstream regions of this gene show evidence of recurrent positive selection since the split of *D. melanogaster* and *D. sechellia*. A polymorphism survey of the *CG9509* region uncovered 1.2 kb segment, which included the putative *CG9509* promoter that showed no polymorphism in the European population. This region also contains several fixed or nearly-fixed derived mutations. This observation, coupled with statistical analysis, provides evidence for a selective sweep in the European population. The selective sweep was likely driven by local adaptation at the level of gene expression.

Chapter 1 Gene expression variation in African and European populations of *Drosophila melanogaster*

1.1 Introduction

Phenotypic diversity is abundant within and between species and can be generated through two major mechanisms: variation in protein structure (*i.e.* amino acid sequence) or variation in protein abundance (*i.e.* gene expression). Over the past several decades, most molecular evolutionary and population genetic studies have focused on the former. This was mainly for practical purposes, as technologies and statistical methods for analyzing structural variation were widely available. However, the importance of gene expression in the generation of phenotypic diversity has long been suspected (King and Wilson 1975). Recent advances in microarray technologies now permit large-scale investigation of differences in transcript abundance among individuals and gene expression surveys have shown that natural variation in transcript abundance is widespread in many different species, ranging from yeast to human (Oleksiak *et al.* 2002; Townsend *et al.* 2003; Morley *et al.* 2004; Stupar and Springer 2006).

The fruit fly *Drosophila melanogaster* has long served as an important model for genetic studies, and is also an important model system for population genetics. Variation at the DNA level in natural populations has been surveyed extensively in microsatellite (Kauer *et al.* 2002) and single nucleotide polymorphism (SNP) studies (Ometto *et al.* 2005; Shapiro *et al.* 2007). The origin and the demographic history of this species are also of interest. Previous surveys pointed out the putative sub-Saharan origin of *D. melanogaster* and its recent migration to the rest of the world (David and Capy 1988; Lachaise *et al.* 1988). Current populations in the ancestral range show a signal of population size expansion (Glinka *et al.* 2003; Pool and Aquadro 2006), while derived populations show the signature of a population bottleneck (Orengo and Aguade 2004; Ometto *et al.* 2005). Extensive theoretical studies have estimated the population genetic parameters associated with these demographic events (Haddrill *et al.* 2005; Li and Stephan 2006).

Most surveys of 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 detect the effects of demographic events, such as bottlenecks or range expansion, on levels of gene expression variation within natural populations. An exception is the study of Meiklejohn *et al.* (2003), which 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, presumably due to its inclusion of the 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.* study used microarrays designed from an early expressed sequence tag screen of the *D. melanogaster* genome (Rubin *et al.* 2000) that covered only 42% of the predicted genes.

In the present study, we measure gene expression variation in adult males of 16 strains from two natural populations of *D. melanogaster*, including eight strains from Africa (Zimbabwe) and eight strains from Europe (the Netherlands) by using whole-genome microarrays. DNA sequence polymorphism has already been thoroughly characterized in these two populations (Glinka *et al.* 2003; Ometto *et al.* 2005; Hutter *et al.* 2007). At the level of gene expression, we find nearly equal amounts of variation within the two populations, but higher amounts in between-population comparisons. We find that genes with male-biased expression exhibit higher levels of variation than those with female-biased or unbiased expression, which has implications for the chromosomal distribution of expression-variable genes. Finally, our experimental design allows us to detect genes that differ significantly in expression between the European and African populations, and thus reveals candidates for genes that have undergone adaptive regulatory evolution accompanying the out-of-Africa range expansion of the species.

1.2 Materials and Methods

Experimental design

Flies were from the European (the Netherlands) and African (Zimbabwe) populations described in Glinka *et al.* (2003). The eight highly-inbred strains per population used for the study were randomly chosen. The flies were reared on standard cornmeal-molasses medium at 22° C and a 15h-9h light-dark cycle.

The platform used was a genome-wide *D. melanogaster* microarray obtained from the Drosophila Genomics Resource Center (DGRC; Bloomington, Indiana, USA) known as 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 these probes were designed to an earlier annotation of the genome (namely 3.1), some genes are not represented on the array according to updated annotations, while others are represented by more than one probe.

In order to compare all of the strains while keeping the number of hybridizations practical we used a “loop design” (Churchill 2002) (Figure 4). We probed each slide with labeled cDNA from two strains. Cross connections were performed to join strains within each of the two populations (solid arrows in Figure 4). To connect the two loops and allow for comparisons between populations, inter-population hybridizations were performed (dashed arrows in Figure 4). Each pairwise comparison included a dye swap. In a total, 30 hybridizations within each population and 20 hybridizations between populations were performed.

RNA extraction and hybridization

RNA was extracted from 70-75 adult males that were 4-6 days of age using the DGRC protocol (*https://dgrc.cgb.indiana.edu/microarrays*). Reverse transcription and labeling were performed with the SuperScript Plus Indirect cDNA Labeling System and Alexa Fluor 555 and 647 dyes (Invitrogen, Carlsbad, California, USA). RNA from the same extraction was used for the dye-swap replicates. Otherwise, RNA was extracted from a new cohort of flies for each pairwise comparison of strains. Hybridizations were performed following DGRC protocols and arrays were scanned using an aQuire microarray scanner (Genetix, New Milton, UK). All array data have been submitted to the GEO database (*http://www.ncbi.nlm.nih.gov/geo*) under accession numbers GSM219761-GSM219840 (platform GPL3830, series GSE8843). All protocol details are available in Appendix A.

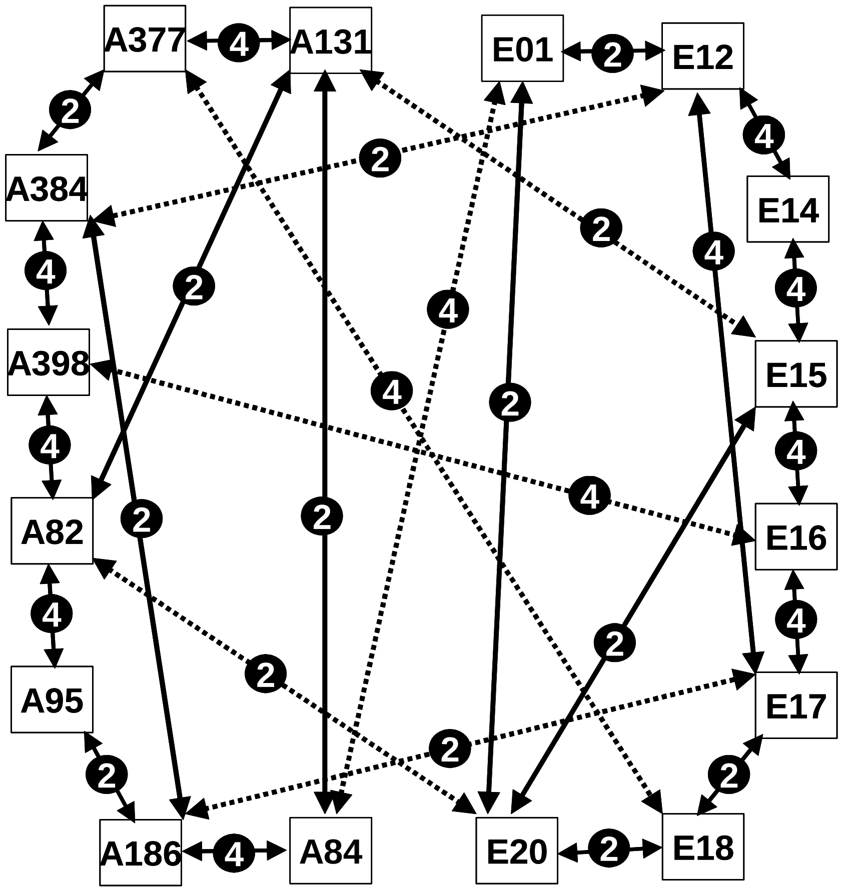


Figure 4: Microarray experimental design.
Each rectangle represents a different *D. melanogaster* strain with 'A' indicating African strains and 'E' indicating European strains. The numbers in black circles represent the total number of replicate hybridizations between the two samples. Arrows in opposite directions represent the dye-swap replicates. All of the pairwise hybridizations include dye-swap replicates. Solid arrows represent hybridizations within each population and dashed arrows represent hybridizations between populations.

Normalization of raw data

To normalize the signal intensity of the two dye channels for each spot on our arrays, we applied a three-step procedure that is implemented in CARMAweb (Rainer *et al.* 2006). This is a web-based interface of the Bioconductor package (Gentleman *et al.* 2004) that provides algorithms to correct for local background effects, within-array variation, and between-array variation. For these corrections, we used the "minimum", "printtiploess", and "quantile" options, respectively. Between-array normalization was performed using the dye-swap replicates for each pairwise comparison of strains.

Data analysis, quality control and statistical power

The normalized expression ratios for each slide were used as input for BAGEL (Townsend and Hartl 2002). This program uses a Markov Chain Monte Carlo algorithm to estimate the relative expression levels of all strains for any given gene. Furthermore, the probability of a gene being differentially expressed between any two strains in the data set is computed.

As a means of quality control, we removed spots that did not show a significant signal of expression, which was determined on a per-slide basis using negative controls probes included on the DGRC-1 arrays. Negative controls were defined as the 182 spots on the array consisting of exogenic DNA (*e.g.*, genes amplified from yeast, *Escherichia coli*, Arabidopsis). 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 a spot fell within the upper 5% of the negative distribution in each channel, the gene was considered "expressed". If a spot presented a signal that was lower than this threshold in either of the two channels, then it was considered "non-expressed" and was excluded from further analysis.

To determine the experiment-wide false discovery rate (FDR), we repeated the BAGEL analysis on a randomized version of our final data set. Randomization was performed on the input file by sampling with replacement within each hybridization (*i.e.*, randomizing within a column), thereby maintaining the underlying data structure (*e.g.*, missing data) within each hybridization. Random sampling was carried out until a total of 5,048 randomized probes were generated, which corresponds to the total number of expressed probes in the original data set. This allowed for an easy and direct comparison of observed and randomized data.

To estimate the power of our experiment to detect expression differences between strains, we calculated the GEL_{50} statistic, which has been proposed as a standard measure to compare studies of expression variation across different experiments and platforms (Townsend 2004). The GEL_{50} is defined as the expression difference at which there is a 50% chance of detecting significance at the 5% level. To obtain this statistic, all pairwise comparisons of differential gene expression are assigned a value of one if they are significant or zero if they are non-significant. These zeros and ones are then plotted on a graph as a function of the expression difference (*i.e.*, the fold-change) between the two samples (on a $\log_2$ scale). Afterwards, a logistic function is fitted through the data points and the GEL_{50} is defined as the fold-change at which the logistic function reaches 0.5.

Detection of differentially expressed genes between populations

To identify genes that differ in expression between the African and the European populations, we repeated the BAGEL analysis using only hybridizations in which an African strain was compared directly to a European strain. This resulted in a total of 20 hybridizations (dashed arrows in Figure 4). All African strains were combined into a single node named "Africa" and all European strains where combined into a node named "Europe". With this approach, the different strains used within each population can be considered as biological replicates. To determine the FDR, a randomized data set was created by permuting the expression ratios of the replicate hybridizations within each gene (*i.e.*, randomizing within a row). This has the effect of randomly assigning the ratio of each hybridization as either Europe/Africa or Africa/Europe. It ensures that the proportion of missing data remains constant in the overall data set as well as within each gene, leading to equal distributions of missing data per gene in the observed and the randomized data sets. Furthermore, the randomized data set automatically contained 5,089 randomized probes that could be directly compared to the observed data. Additionally, we created a randomized data set using the approach of the 16-node analysis (see above) for comparison. Both methods produced very similar results (data not shown) and the first approach was used for our analysis.

Gene ontology analysis

Analysis of the gene function was done by Gene Ontology process available on web-based tool g:Profiler (Reimand *et al.* 2007). This tool introduces a new correction for multiple testing (called g:SCS) that takes the hierarchical nature of GO terms into account. The GO terms identified molecular functions and biological processes associated with the 153 genes significantly differentially expressed between European and African lines.

Analysis of chromosomal and gene expression location

The chromosomal locations of the 153 genes significantly differently expressed between the two populations were provided by the web-tool FlyMine (Lyne *et al.* 2007). The web-tool FlyAtlas (Chintapalli *et al.* 2007) was used to identify tissue-enriched gene expression for the above genes . We consider only tissue expression information for the following: brain, head (including brain), crop, Malphigian tubule, ovaries (excluding spermatheca, uterus), testes (excluding accessory glands), male accessory glands, and larval fat body. All of the information concerning the dataset is available on the web site (*http://www.flyatlas.org*).

1.3 Results

Data quality and detection of gene expression

We performed a total of eighty microarray hybridizations, each of which was a head-to-head comparison of two *D. melanogaster* strains (Figure 4). After quality control, 5,048 probes representing 4,512 unique genes had sufficient signal quality to estimate their relative expression level in all 16 strains. This corresponds to ~40% of all genes on the array. The relative expression level of each gene in each strain was estimated using BAGEL (Bayesian Analysis of Gene Expression Levels) (Townsend and Hartl 2002) and the statistical power of our experiment to detect expression differences between strains was determined by calculating the GEL_{50} statistic (Townsend 2004) (see Materials and Methods). This measurement allows the comparison of diverse microarray studies and the power of their experiments to detect gene expression differences. For the complete 16-strain analysis, the GEL_{50} value is 1.51. In other words, given our experimental design and data quality, there is a 50% chance of detecting a 1.51-fold expression difference as significant at the 5% level (Table 1). This value compares well with those of similar experiments in fish, yeast, flies, and plants (Clark and Townsend 2007), and is slightly better that that of the study of Meiklejohn *et al.* (2003) (GEL_{50} = 1.64), which also examined four African and four non-African *D. melanogaster* strains. GEL_{50} values were also calculated within or between populations separately. The GEL_{50} was 1.512 within Europe, 1.508 within Africa, and 1.513 between populations, indicating that the power to detect differences in any of these three comparison schemes 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 (Table 1).

Table 1: Statistical power of the experiments to detect expression differences.

	GEL_{50}
Overall*	1.510
Within Europe*	1.512
Within Africa*	1.508
Between *	1.513
Between §	1.180

*using 16 nodes
§using 2 nodes

Total number of differentially expressed genes

Since the number of tests for pairwise differences in expression was extremely high (5,048 probes × 120 pairwise comparisons = 605,760 tests), we could not operate with the conventional 5% significance level due to the problem of multiple testing. We therefore created randomized data sets to estimate the false discovery rate (FDR) at any given significance level (Table 2, 16-node experiment). For all further analyses, we use a *P*-value cut-off of 0.001, which corresponds to a FDR of 6.9% and is similar to the FDR of 5.2% used in the study of Meiklejohn *et al* (2003).

Table 2: Number of significant tests and false discovery rates (FDR) for different *P*-value cut-offs.

	16-nodes analysis		Two-nodes analysis	
P-value	Significant tests	FDR	Significant tests	FDR
0.05	110,285 (18.21%)	0.4906	991 (19.47%)	0.4834
0.02	63,636 (10.51%)	0.3285	562 (11.04%)	0.3292
0.01	44,081 (7.28%)	0.2337	380 (7.47%)	0.2237
0.005	31,670(5.23%)	0.1657	269 (5.29%)	0.1710
0.002	21,480 (3.55%)	0.1024	161 (3.16%)	0.0870
0.001	16,564 (2.73%)	0.0692	109 (2.14%)	0.0550

Using this cut-off, we found that 1,894 (37.5%) of the probes showed significant differences for at least one pairwise comparison (Table 3), which was slightly lower than the proportion (46.7%) reported by Meiklejohn *et al.* Since 413 genes were represented by multiple probes in our data set, we checked how well the percentage of polymorphic genes corresponded to the number of polymorphic probes. If a gene was considered polymorphic when at least one of its probes showed a significant pairwise difference between strains, then 38.9% of all expressed genes were polymorphic. If a stricter criterion was applied and only genes for which all probes showed a significant difference were considered polymorphic, this dropped to 35.1%. The overall effect of including multiple probes per gene was rather small. Unless noted otherwise, we 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 (20.6%) showed differences within the African population, and 1,600 (31.7%) showed differences when comparing European to African strains (inter-population comparisons), (Table 3). The higher number of differences for the inter-population comparisons was somewhat expected, since there were more pairwise tests than for the within-population comparisons (64 as opposed to 28).

Table 3: Polymorphism in expression.

	Polymorphic probes		Mean pairwise differences per probes in %§
	Total number (%)	Mean per PW (SD)#	
Overall	1,894 (37.5%)	138.0 (53.0)	2.73
Europe	964 (19.1%)	126.5 (43.7)	2.51
Africa	1,039 (20.6%)	125.9 (47.8)	2.49
Between	1,600 (31.7%)	148.4 (57.3)*	2.94*

# Average number and standard deviation (SD) of probes differentially expressed for each pairwise (PW) comparison between all strains within the corresponding data set.

§ Average percentage of pairwise comparisons showing differential expression for a probe

* Significant ($P < 0.001$) based on Mann-Whitney U test

Expression differences between individual strains

We also investigated the number of differentially expressed probes for each pairwise comparison. The complete pairwise comparison matrix is provided in figure 5. On average, 138 probes showed differential expression for each individual pairwise comparison (Table 3). Given the overall number of 1,894 probes that showed differences, this number was surprisingly small, even more so when taking into account that the Meiklejohn *et al.* (2003) study detected an average of 498 differentially expressed genes per pairwise comparison with a total number of 2,289 differentially expressed genes. This reveals 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.

	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	

Figure 5: Complete pairwise comparison matrix.
Numbers below the diagonal are the numbers of differentially expressed genes for each pairwise comparison of strains at $P < 0.001$; numbers above the diagonal are from a randomization of the data (*i.e.*, the expected number of false positives)

Expanding this approach to investigate differences within and between populations, we see a pattern resembling that for the total number of differentially expressed probes. On average, comparisons between two European strains showed differences in 126.5 probes, comparisons between two African strains showed differences in 125.9 probes, and comparisons between a European and an African strain showed differences in 148.4 probes (Table 3). Since these numbers are independent from the number of pairwise comparisons, we conclude that there is an excess of differentially expressed probes in the inter-population comparisons (Mann-Whitney *U* test, $P = 0.019$).

To examine expression variation on a gene-by-gene basis, we determined the percentage of significant pairwise differences per probe. In general, this measure of variation followed the pattern seen for the number of differentially expressed genes within the European and African populations presented above (Table 3). The level of expression polymorphism was similar within Africa (2.49%) and Europe (2.51%) and a Mann-Whitney *U* test of the two populations was not significant ($P = 0.086$). The between-population comparisons showed a larger proportion of significant tests (2.94%) and this was significantly larger than the within-population polymorphism (Mann-Whitney *U* test, $P < 0.001$). However a neighbor-joining tree based on a gene expression pairwise comparisons (Figure 6) showed no distinct separation of the two populations (Figure 6b) as it is the case with the neutral DNA data (Figure 6a). This result could be due to the relatively small number of gene expression differences between the two populations in comparison to the expression variation within populations (Table 3).

Expression polymorphism of X-linked and autosomal genes

We compared the levels of polymorphism for genes residing on the X chromosome to those located on the autosomes and found a systematic difference between these two classes. Levels of expression polymorphism were consistently lower for X-linked genes, irrespective of whether they were measured within or between populations or in the complete data set.

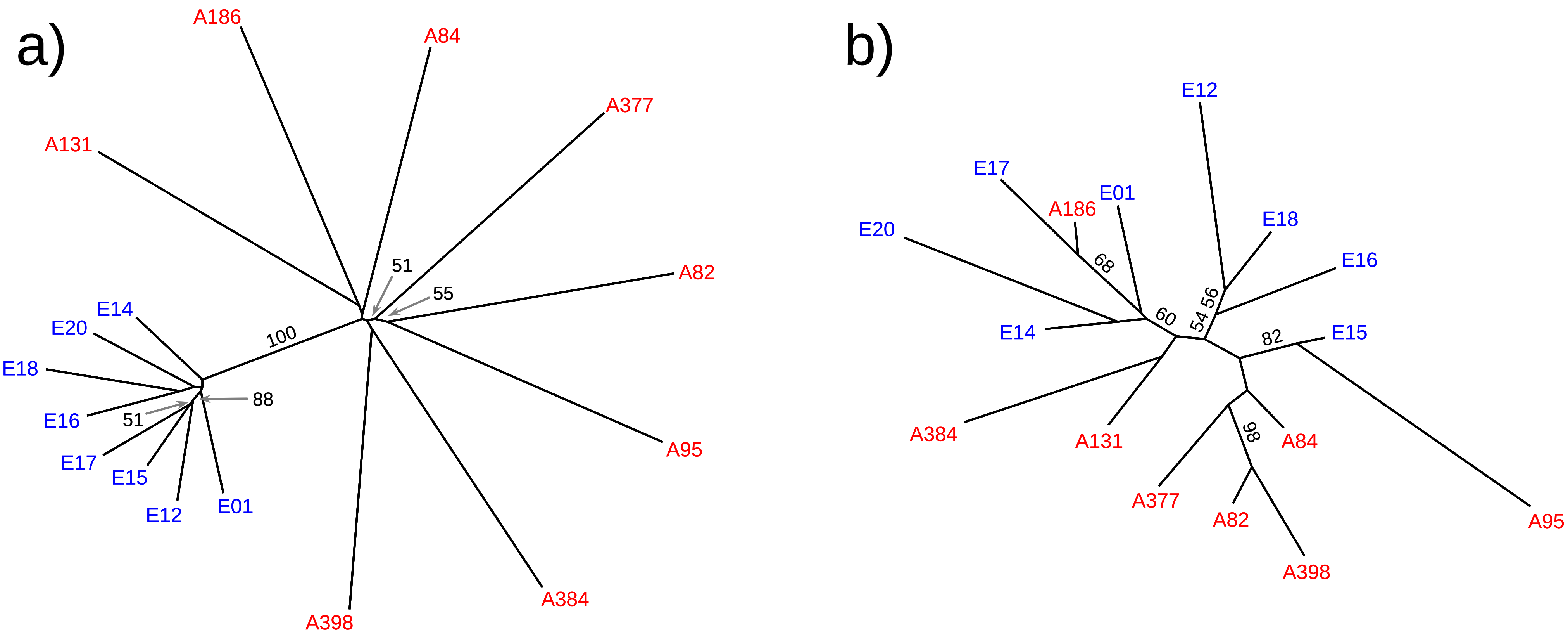


Figure 6: Neighbor-joining trees of the 16 *D. melanogaster* strains based on (a) DNA polymorphism data from the study of Ometto *et al.* (2005) and (b) the gene expression distance matrix in figure 5. Trees were created using the PHYLIP software package (Felsentein, 2005). Bootstrap values for the gene expression tree were generated by sampling with replacement of the gene list and are given as percentages for nodes with a support of >50%. As expected for neutral DNA data from two semi-isolated, random-mating populations, the tree based on non-coding SNPs forms two distinct star-like clades separating the African and the European strains (in red and blue, respectively). Note the increased length of the African external branches indicating elevated polymorphism in Africa. For the tree based on expression differences, strains from the same population do not form monophyletic clades. This is due to the small number of genes showing distinct population specific expression relative to overall levels of polymorphism. Nevertheless, strains originating from the same population tend to cluster together, even though bootstrap support for internal nodes is low.

Variability on the X chromosome was only about 70% of that on the autosomes when measured as percentage of pairwise differences per probe, and this dearth of polymorphism was statistically significant for all four comparison schemes (Table 4). The same trend was found when using the percentage of polymorphic probes as a statistic, although the differences between chromosomal classes were not as pronounced (Table 4).

Table 4: Expression polymorphism on the X chromosome and autosomes.

	X chromosome	Autosomes	X/A ratio *
Number and percentage of polymorphic probes			
Overall	335 (35.8%)	1,559 (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%)	1,323 (32.2%)	0.919 (*P* = 0.12)
Average percentage of pairwise differences			
Overall	2.02	2.90	0.697 (*P* = 0.040)
Europe	1.77	2.68	0.661 (*P* = 0.014)
Africa	1.86	2.94	0.705 (*P* = 0.017)
Between	2.20	3.11	0.708 (*P* = 0.035)

*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.

Expression polymorphism of sex-biased genes

To investigate the contribution of genes with sex-biased expression to overall levels of gene expression variation, we used the consensus results of three independent experiments that directly compared male *versus* female gene expression in *D. melanogaster* (Parisi *et al.* 2003; Ranz *et al.* 2003; Gibson *et al.* 2004) and two different criteria for the classification of sex-biased genes, one based on fold-change and one based on statistical significance (Gnad and Parsch 2006). We detected the highest fraction of expressed genes within the male-biased class and the lowest fraction within the female-biased class (Table 5). This is expected, since adult male flies were used as the RNA source for all of our experiments. Meiklejohn *et al.* (2003) reported that, when assayed in adult males, genes with male-biased expression were significantly more variable than genes with female-biased or unbiased expression. We observed the same pattern for the genes in our data set: male-biased genes were consistently more variable than genes of the other two classes, and this pattern held for both the European and African populations (Table 5). Female-biased genes tended to have the least expression variation (Table 5). This low variation cannot be explained simply by the lack of expression of the female-biased genes in adult males, because only genes with detectable expression were used in the analysis.

Table 5: Expression variation in sex-biased genes.

	Two-fold			FDR 10%		
Sex-bias classification*	Male	Female	Unbiased	Male	Female	Unbiased
Number of genes on array	669	768	3,891	1,228	857	1,534
Percentage of genes detected as expressed	61†	22	41	67†	33	41
Percentage of expressed genes						
Variable in Europe	20‡	12	16	22†	13	15
Variable in Africa	28†	15	16	27†	16	17
Variable overall	42†	32	31	45†	31	32
Differentially expressed between populations	1.21§	2.86	3.54	2.46	1.75	3.10
Average percentage of pairwise differences						
Within Europe	2.50†	1.14	2.07	2.93†	1.50	1.82
Within Africa	3.96†	1.08	1.75	3.57†	1.32	1.75
Overall	3.16†	1.09	2.21	3.35†	1.50	2.00

*Sex-biased gene sets are defined using Sebida (Gnad and Parsch 2006)

† Significantly different from both female and unbiased ($P < 0.05$) by Fisher's exact test (percentages) or Mann-Whitney *U* test (pairwise differences).

‡ Significantly different from female ($P < 0.05$) by Fisher's exact test.

§ Significantly different from unbiased ($P < 0.05$) by Fisher's exact test

Expression differences between populations

In order to find genes that differ in expression on a population scale (and are therefore candidates for local adaptation), we pooled all strains of each population into a single node. We came out with two nodes, one for Africa and one for Europe, with which we implement BAGEL software to find differences (see Materials and Methods). With this approach, BAGEL estimates the average expression level for each population and tests 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 be much more powerful to detect differences since it has only two nodes to compare with 20 hybridizations. As an additional quality control step, we required that each probe be detected as "expressed" (see Materials and Methods) in at least nine of the 20 hybridizations. A total of 5,089 probes representing 4,528 genes passed the quality control. The GEL_{50} for this design was 1.18 (Table 1), which, as expected, was lower (*i.e.,* better) than in the original 16-node analysis.

As with the first analysis, we used a randomized data set to calculate the FDR and adjust our *P* value for differential expression (Table 2, two-node experiment). We chose a *P* value cut-off of 0.002, which leads to an FDR of 8.7% and corresponds well to the FDR of the 16-node experiment (6.9%). At this significance level, 161 probes representing 153 genes were differentially expressed between Europe and Africa (Figure 7). A complete list of these probes is provided as Appendix C.

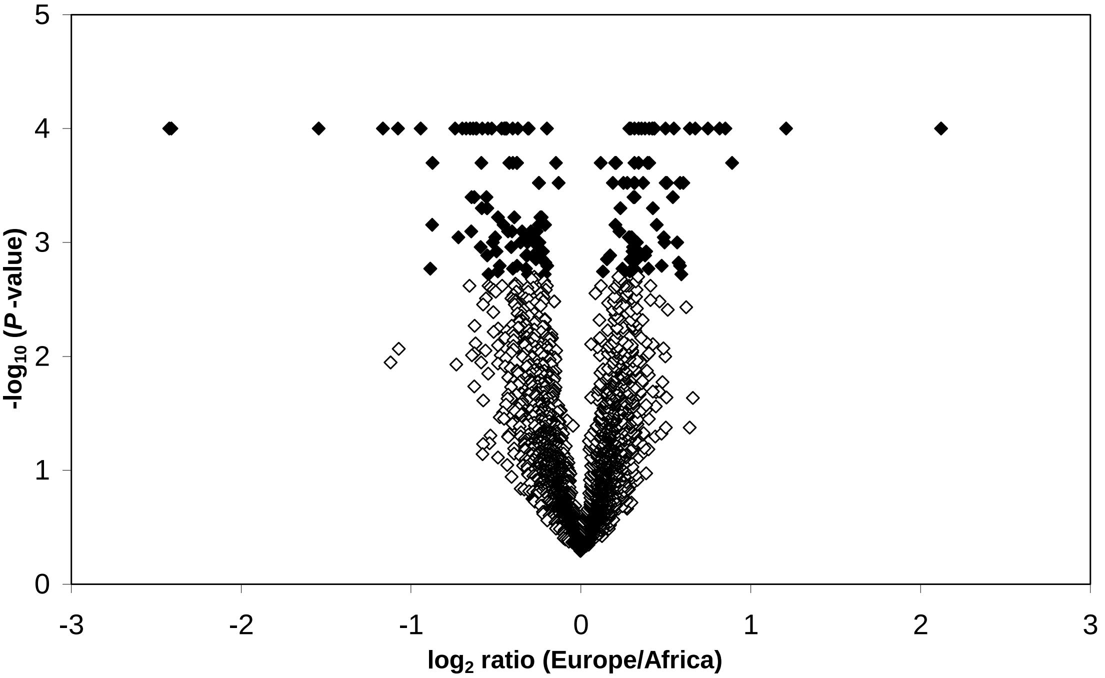


Figure 7: Volcano plot of gene expression differences between European and African populations.
The open squares indicate non-significant expression ratios. The black squares indicate significant differentially expressed genes between Africa and Europe (*P* < 0.002).

The magnitude of expression differences was relatively low, with the median fold-change difference being 1.32 and the maximum being 5.36. Of the 161 differentially expressed probes, 85 (52.8%) were expressed at a higher level in Africa and 76 (47.2%) were expressed at a higher level in Europe, but this difference was not significant (Fisher’s exact test, *P* = 0.26). A comparison on a per-gene basis showed a similar pattern: 80 genes were over-expressed in Africa and 73 in Europe (Fisher’s exact test, *P* = 0.25). The magnitude of the expression difference was 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 was significant (Mann-Whitney *U* test, *P* = 0.044). Neither the X chromosome nor the autosomes were enriched for these genes (Fisher’s exact test, *P* = 0.83). The differently expressed genes between the two populations are mainly located on the autosomal chromosomes with the maximum on the third chromosome (N=67) and twenty seven on the X-chromosome (Figure 8).

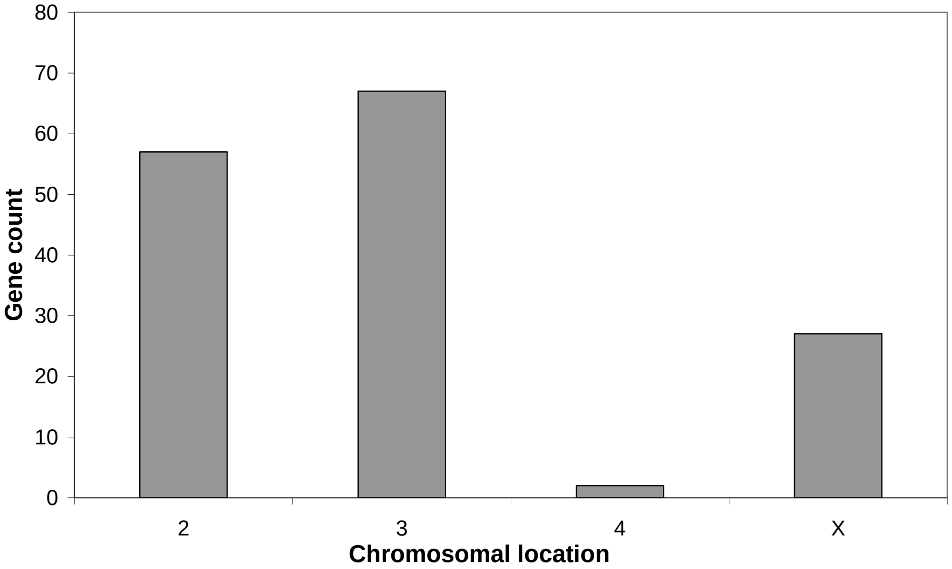


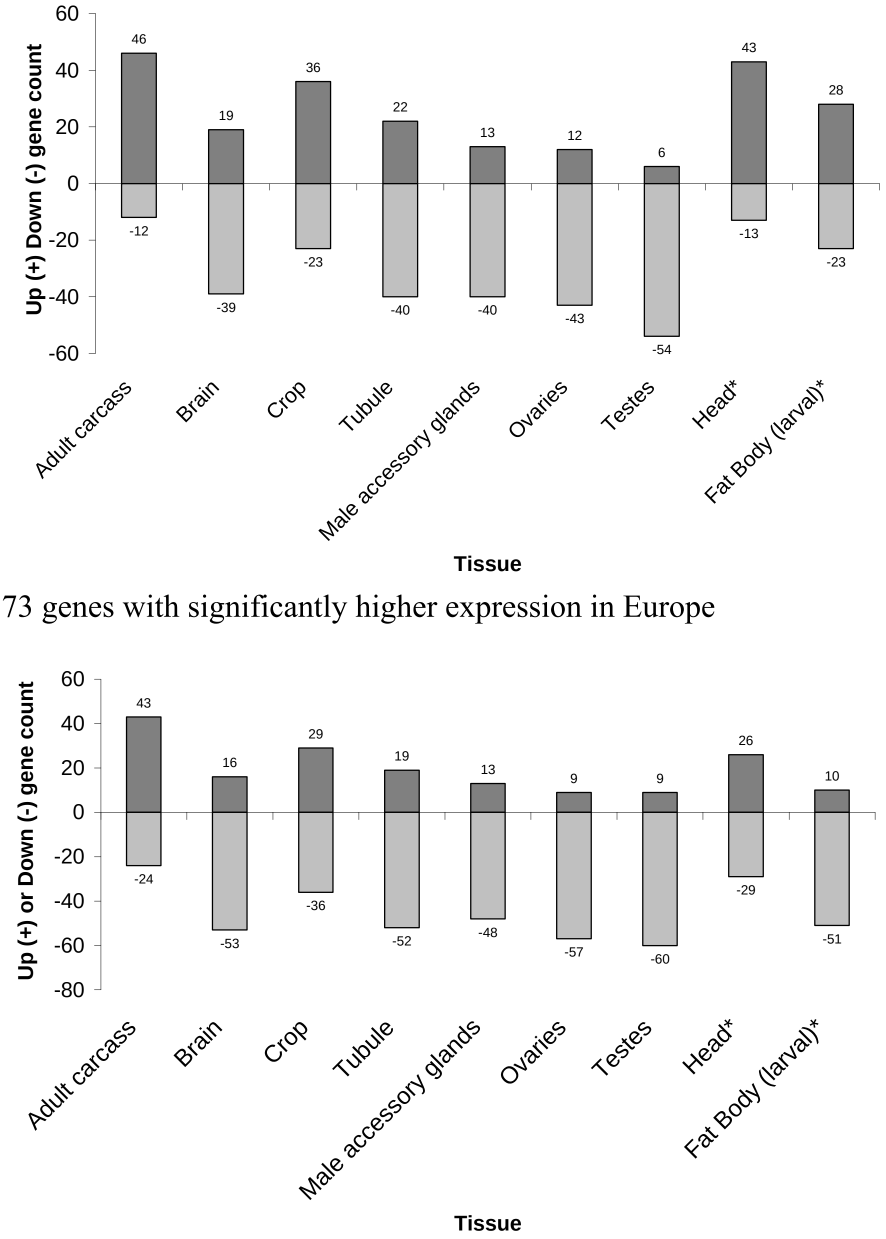
Figure 8: Chromosomal location of significant differentially expressed genes between Europe and Africa.

There was also no enrichment of sex-biased genes. If anything, sex-biased genes were under-represented among those showing expression differences between the populations (Table 5).

Functional analysis and tissue-enriched expression of candidate genes

Using the FlyAtlas database (Chintapalli *et al.* 2007) we identified the different tissues in which the significantly overexpressed genes of each population were expressed (Figure 9a and b). A high gene expression signal was present for both populations in crop and adult carcass. However there was no expression enrichment in brain, tubule, male accessory glands, ovaries or testes. Highly expressed genes in Europe show tissue expression enrichment in head (almost twice that of highly expressed genes in Africa), which was significant by Fisher's exact test ($P < 0.001$), and in larval fat body (almost three times that of highly exprssed African genes; Fisher's exact test, $P < 0.001$) (Figure 9).

Some GO categories were significantly over-represented among the 153 genes with expression differences between populations (Table 6). Furthermore, for some categories the expression differences were biased towards a certain direction. For example, the genes associated with the actin cytoskeleton were all over-expressed in the African population. The GO categories "actin filament" and "structural constituent of cytoskeleton" were also exclusively composed of these genes.



a) Based on 73 genes with significantly higher expression in Europe

b) Based on 80 genes with significantly higher expression in Africa

Figure 9: Specific tissue enrichment for genes significantly differentially expressed between Europe and Africa. * Tissue specific enrichment significantly different between the two populations according to Fisher's exact test ($P < 0.001$).

Table 6: GO terms over-represented in the list of differentially expressed genes between European and African populations.

Go number	GO term	Genes in genome	Genes in list	*P*-value
Biological process				
GO:0005975	Carbohydrate metabolic process	347	14	7.34E-05
GO:0032787	Monocarboxylic acid metabolic process	48	6	2.24E-05
GO: 0006631	Fatty acid metabolic process	38	5	8.67E-05
Molecular function				
GO:0016491	oxidoreductase activity	523	20	4.07E-06
GO:0004448	isocitrate dehydrogenase activity	4	3	6.83E-06
GO:005200	Structural constituent of cytoskelton	12	4	9.34E-06
Cellular Component				
GO:0015629	actin cytoskeleton	47	8	7.68E-08
GO:0005884	actin filament	10	5	5.71E-08

Interestingly, some genes highly over-expressed in African lines were involved in wing morphogenesis, such as *CG7214* (Figure 10). Some other genes like *Act88F* and *TpnC41C* are muscle components that might also play a role in flight. In the top-ten gene list of over-expressed genes in the African population, most are associated with musculature and flight and are mainly expressed in adult carcass (Figure 9b). Other interesting genes are present in this list such as *CG8997* and *Nplp3,* which are an odorant binding protein and a neuropeptide, respectively (Figure 10). In contrast, genes involved in fatty acid metabolism had a higher level of expression in the European population. These genes also form the GO category “monocarboxylic acid metabolic process”. The gene with the highest over-expression in the European population is *Cyp6g1*. This gene is well known because it has been found to be overexpressed in flies resistant to DDT (Daborn *et al.* 2002). Other genes with putative functions in lipid biosynthesis are also present in the list, such as *CG9509*, Malic enzyme (*Men*), and *CG18135* (Figure 10). The gene *Dpr15* shows some similarity with immunoglobulin. Another protein important in odorant perception is in top-ten list of the genes highly expressed in Europe, *Opb56d*.

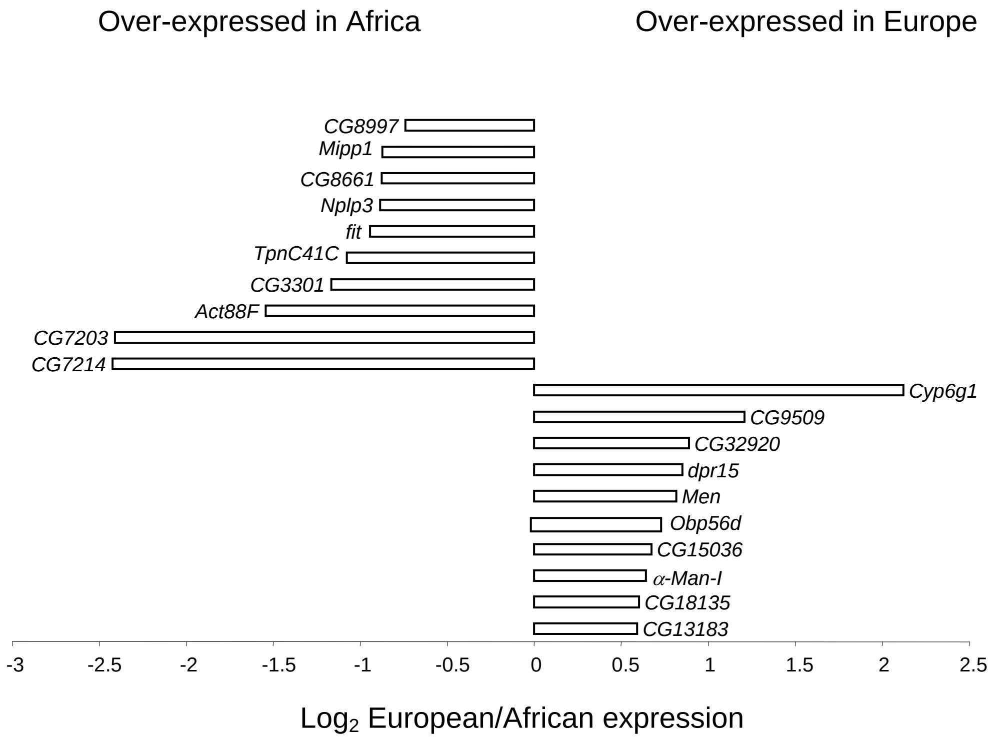


Figure 10: Top-ten list of the genes over-expressed in each population with $P < 0.002$.

1.4 Discussion

Patterns of gene expression polymorphism

Our survey of gene expression variation is the largest performed to date in *D. melanogaster* and the first to include a truly natural, derived population. In combination with the ancestral African population, this provides a comprehensive picture of expression variability in the species. However, it should be noted that the amount of expression variation detected among inbred strains may differ from that in natural populations for several reasons. First, inbred strains are expected to be homozygous over a large proportion of the genome and thus the effects of dominance on gene expression will not be detected (Gibson *et al.* 2004). Second, the process of inbreeding itself may act like an environmental stress and lead to changes in the expression of genes involved in metabolism and stress resistance (Kristensen *et al.* 2005). Third, mutations that alter levels of gene expression may accumulate in inbred strains during the time that they are maintained in the laboratory (Rifkin *et al.* 2003). Finally, since all strains were reared in a common laboratory environment, it is not possible to detect genotype-by-environment interactions that affect gene expression. While the above limitations are inherent to this type of microarray study, we expect the general patterns of gene expression polymorphism observed among inbred strains to be robust to these factors and to reflect the patterns present in natural populations.

One pattern we observed was that the amount of expression variation did not differ between the European and the African populations (Table 3). This might seem somewhat surprising, since large-scale genome scans have shown that the African population harbors much more variation (over twice as much) at the DNA level than the European population (*e.g.* Glinka *et al.* 2003), an observation that is consistent with the inferred demographic history of these populations and with the African population having a larger effective size (Li and Stephan 2006; Hutter *et al.* 2007). However, the DNA polymorphism studied in such genome scans consists mainly of non-coding SNPs, which are thought to evolve (nearly) neutrally. While some authors suggest that differences in gene expression also reflect changes that are selectively neutral (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. Thus, the patterns of expression polymorphism that we observe could be explained by a mutation-selection balance model, where mutations affecting expression level are mostly deleterious and are quickly purged from the population. In such a case, the observable variation depends on the mutation rate and the selection coefficient against deleterious mutations (which should be equal in both of our studied populations), and is independent of the population size (Gillespie 1998). Evidence that stabilizing selection is a key factor governing expression variation has already been found in several studies. For example, 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 within species, as well as expression divergence between species.

We observed a higher number of expression differences between populations than within populations, and this result was consistent regardless of the statistic used to quantify expression polymorphism (Table 3). This increased inter-population expression divergence is likely a consequence of population differentiation since the colonization of Europe approximately 16,000 years ago (Li and Stephan 2006; Hutter *et al.* 2007). Some of this expression divergence may reflect adaptation to the temperate environment, which would result in genes that show relatively low expression polymorphism within populations, but high expression divergence between populations (discussed below). Nevertheless, the number of genes showing population specific expression patterns is relatively low compared to overall levels of polymorphism. The results of the two-node analysis show that only 161 probes have expression levels that are population specific (~3% of all expressed probes) while 37.5% of all expressed probes show expression differences between at least two strains in the 16-node analysis. In our case, differences at the gene expression level therefore only have little power to group individual strains by population when used, for example, as an input for tree building methods compared to trees derived from neutral DNA data (Figure 6).

In both populations, X-linked genes showed consistently less expression polymorphism than autosomal genes (Table 4). This appears to be a result of the unequal genomic distribution of sex-biased genes. Previous studies have shown that male-biased genes are significantly under-represented on the X chromosome (Parisi *et al.* 2003; Ranz *et al.* 2003) and also show the highest levels of expression polymorphism (Meiklejohn *et al.* 2003). These results are confirmed in our data. Only 9% of the male-biased genes detected as expressed are X-linked; the corresponding proportions for female-biased and unbiased genes are 23% and 17%, respectively. Additionally, we find that male-biased genes show the highest levels of gene expression polymorphism (Table 5). Thus, the reduced expression polymorphism on the X chromosome could be explained by its paucity of male-biased genes. The slight over-abundance of female-biased genes, which show the least expression polymorphism, on the X chromosome, may also contribute to this pattern.

Candidate genes for adaptation

To identify genes that are differentially expressed between Europe and Africa, we employed a two-node analysis (see Materials and Methods), in which all strains from each population were grouped into a single node. An interesting finding was that genes encoding proteins involved in muscle formation were consistently over-expressed in the African population. Two of these genes (*Act88F* and *TpnC41C*) encode proteins that are predominately found in the indirect flight musculature (Karlik *et al.* 1984; Qiu *et al.* 2003).

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 (David and Bocquet 1975; 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. This over-expression of muscle-related genes could be the result of direct selection on their expression, but could also be a downstream effect of selection for increased number or size of muscle cells in African flies. In this context, it is noteworthy that the gene *CG7214*, which has the largest magnitude of over-expression in the African population (5.36-fold), is expressed during wing morphogenesis (Ren *et al.* 2005), although its exact function remains unknown. Direct measurements of relative wing sizes, wing-beat frequencies, and number and size of muscle cells in our surveyed populations will provide insight into the phenotypes associated with these gene expression differences. The abundance of highly differentially expressed muscle-related genes in our top ten-list might also be the reason why a stronger signal of expression was detected in adult carcass (Figure 9b and 10). The over-expresed genes in Africa also included a neuropeptide and an odorant binding protein, which might play a role in the fly’s perception of its environment.

Genes associated with fatty acid metabolism showed consistent over-expression in the European population. The fat body of Drosophila plays an important role in the detoxification of xenobiotics and the defense response to microbial infections and can be viewed as the functional equivalent of the mammalian liver (Lemaitre and Hoffmann 2007; Yang *et al.* 2007). This observation might explain why we detect more tissue expression enrichment in larval fat body of the European population (Figure 9a). Most of the study comparing the expression profile of dichloro-diphenyl-trichloroethane (DDT)-resistant and DDT-sensitive strains of *D. melanogaster* revealed differences in the expression levels of lipid metabolism genes between these strains (Pedra *et al.* 2004).

The malic enzyme gene (*Men*), which shows 1.76-fold over-expression in the European population, is of particular interest in this context. This enzyme oxidizes malate to pyruvate and concurrently reduces nicotinamide adenine dinucleotide phosphate (NADP) to NADPH, which is a major reductant in lipid biosynthesis (Wise and Ball 1964). A study of DNA polymorphism and enzymatic activity of naturally occurring alleles of *Men* revealed clear differences between African and non-African populations (Merritt *et al.* 2005). The allelic state of this gene influences not only the abundance of triglycerides in flies, but also the activity of isocitrate dehydrogenase (*Idh*). We find that expression levels of *Idh* also differ between European and African flies (represented by two probes, showing 1.24-fold and 1.18-fold over-expression in Europe), indicating that not only DNA polymorphism, but also variation in expression plays a role in the interaction of these two genes.

The *CG9509* gene, which has 2.31-fold over-expression in Europe, is a protein with unknown function that contains a FAD (Flavine Adenine Dinucleotide) oxydoreductase domain and a GMC (Glucose-Methanol-Choline) oxydoreductase domain, which might indicate a putative metabolic function. The gene is mainly expressed in tubules, which have an exocrine function. The top-ten list of the genes over-expressed in Europe includes a gene (*dpr15*) with immunoglobulin function that is mainly expressed in the thoracicoabdominal ganglion and might be associated with immune function. A classic example of expression differences leading to adaptive phenotypes is the cytochrome P450 gene *Cyp6g1*. It has been shown that over-expression of this gene leads to increased DDT resistance (Daborn *et al.* 2002). In our microarray data set, this gene shows the largest magnitude of over-expression in the European population (4.35-fold). The consistent pattern of higher expression levels in European flies for the above genes provides evidence that the acquisition of resistance against insecticides, such as DDT, is an important adaptive trait for flies living in the European habitat.

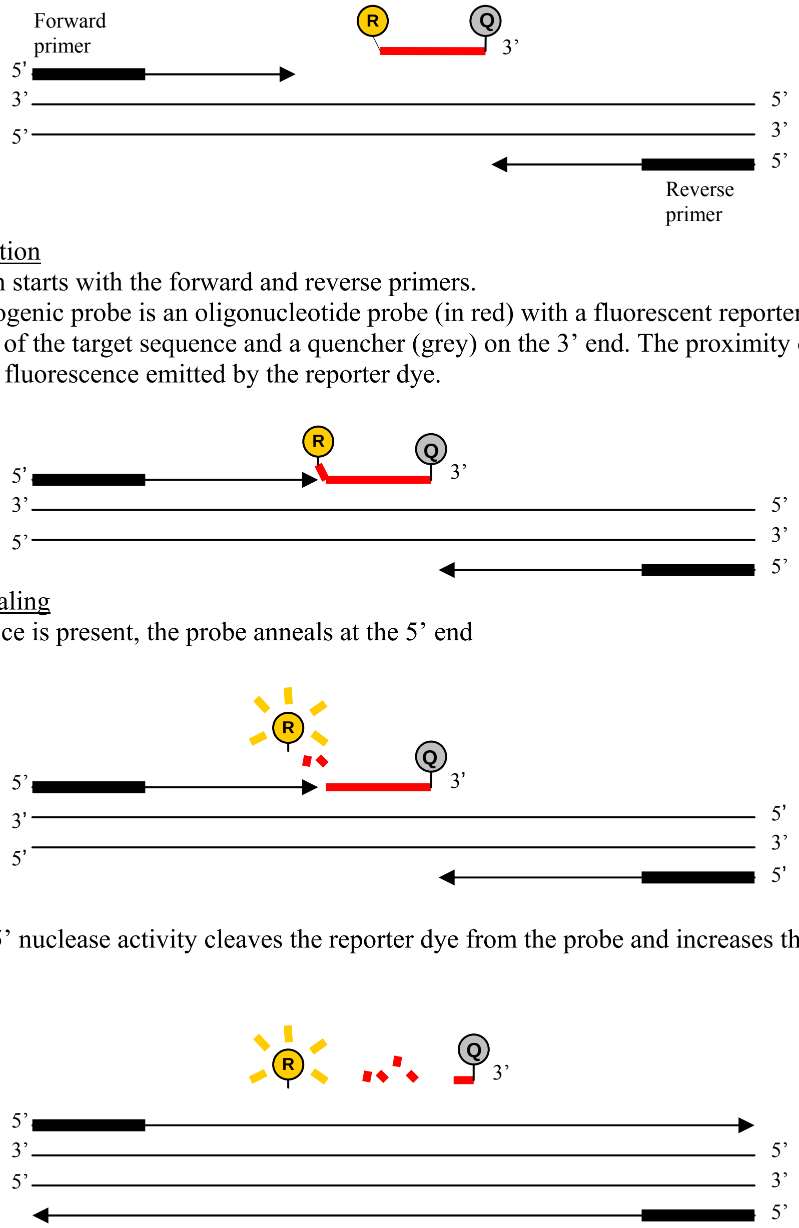
It is noteworthy that a protein involved in odorant perception is over-expressed in each population, *Opb56d* in Europe and *CG8997* in Africa. This may indicate the importance of environmental perception within each habitat (Vieira *et al.* 2007). In general, however, it seems that selection operates differently on different functional classes of genes in each population. In Africa, selection acts on components of the flight musculature, while in Europe it acts on genes involved in detoxification and immune response.

Chapter 2 Validation of microarray results and comparison of patterns of expression variation in males and females of *Drosophila melanogaster*

2.1 Introduction

The microarray experiments described in chapter 1 provide a global picture of gene expression variation among 16 strains of *Drosophila melanogaster* and between two populations (Africa and Europe). However, because microarrays query the expression level of many thousands of genes simultaneously, it is often difficult to extract accurate expression information pertaining to specific genes. There are several reasons for this (Draghici *et al.* 2006). First, because microarrays survey so many genes, there is a problem of multiple testing and one usually needs to set a very conservative standard for detecting statistically significant differences in expression, or else accept a relatively high rate of false positives. Second, because microarrays contain thousands of probes, it is difficult to ensure the quality of each individual probe on the array. Even for high-quality arrays, such as those produced by the DGRC, the quality may vary from probe to probe. Third, for exon-based microarrays, such as those used in chapter 1, the probes are relatively long (from 100-600 bp: average = 410 bp) and the amount of non-specific or background hybridization may vary from probe to probe depending on length or other sequence properties. Finally, the large-scale nature of the microarrays limits the number of different samples that can be analyzed. For example, the experiments of chapter 1 focused only on adult males. We have no information about the expression of these genes in adult females.

For the above reasons, it is important to verify and extend the microarray results using an independent method that is highly gene-specific. This is most commonly done by quantitative real-time PCR (qPCR) using Taqman technology (Livak *et al.* 1995). Generally, this method uses specific PCR primers to amplify cDNA generated from mRNA of the sample of interest and the level of gene expression is quantified by the incorporation of an intercalating fluorescent dye into the amplified products. The TaqMan method (Livak *et al.* 1995) is a modification of this approach that uses a specific, fluorescently-labeled probe to detect PCR amplification (Figure 11).



Step 1: Polymerization
The polymerization starts with the forward and reverse primers.
The TaqMan fluorogenic probe is an oligonucleotide probe (in red) with a fluorescent reporter dye (yellow) bound to the 5'end of the target sequence and a quencher (grey) on the 3' end. The proximity of the quencher greatly reduces the fluorescence emitted by the reporter dye.

Step 2: Probe annealing
If the target sequence is present, the probe anneals at the 5' end

Step 3: Cleavage
DNA polymerase 5' nuclease activity cleaves the reporter dye from the probe and increases the reporter dye signal

Step 4: Polymerization completed
The probe is removed from the target strand, allowing primer extension to continue to the end of the template strand.

Figure 11: Quantitative Real-Time PCR using the on TaqMan method.

In many species males and females show extremely different phenotypes such as sexual size dimorphism (size differences), presence of horns, antlers, tusks or color patterns. Males and females also differ in behavior. Examples include differential investment in offspring and mate choice (*e.g.*, females favoring males with ornaments) (Andersson and Simmons 2006; Clutton-Brock 2007). All of these differences are grouped under the term of sexual dimorphism. Drosophila sexes show relatively modest phenotypic sexual dimorphism. These differences concern mainly the abdominal pigmentation and the presence of combs on the male's two first legs. At the genomic level, males and females are identical, except for a degenerate Y chromosome (with about 20 genes) that is present only in males (Malone and Oliver 2008). This implies that most of the sexual dimorphism that we observe is due to gene regulation differences (Ellegren and Parsch 2007). Previous surveys of gene expression in the two sexes indicate that more than 50% of the genome shows sex-biased expression in *Drosophila melanogaster* (Ranz *et al.* 2003; Parisi *et al.* 2004; McIntyre *et al.* 2006). This means that most genes are expressed at a higher level in one sex than in the other.

The results of chapter 1 indicate that, in adult males, more than 2,000 genes vary in expression across the 16 strains analyzed, and over 120 genes show significant expression differences between the European and African populations. It was already shown that differences exist between these two populations in traits like cold tolerance, starvation resistance and fecundity (Greenberg *et al.* 2003). Other characteristics, such as female mating choice, also differ between European and African strains. The African females mate preferentially with African males, while cosmopolitan (*i.e.*, non-African) females mate indiscriminately with either African or cosmopolitan males (Osada *et al.* 2006; Michalak *et al.* 2007). These characteristics might also be attributable to differences in gene expression between the two populations and their environmental adaptation (Greenberg *et al.* 2003). These observations suggest that patterns of gene expression variation within and between populations may differ between males and females.

In this chapter, qPCR is used to validate the microarray expression results for a subset of genes described in chapter 1. This includes pairwise comparisons of gene expression levels in adult males among strains, comparisons of differences between populations, and comparisons of expression variation among adult females. This approach reveals a strong correlation between male and female qPCR and microarray measurements. This result suggests that the microarrays provide an accurate representation of gene expression variation among strains and that, in general, these differences are the same in males and females. However, the comparison of the pairwise fold changes in expression for each gene revealed cases in which the expression variation observed in males differed from that observed in females.

2.2 Material and Methods

Gene expression experiments

The quantitative real-time PCR (qPCR) experiments were performed on adult female and male flies between 4 and 6 days old. Total RNA was extracted from the same lines used in Chapter 1 using Trizol reagent (Invitrogen Carlsbad, California, USA) and either 30 female or 45 male flies per extraction. This included genes that showed a high number of significant expression differences within Europe (*CG18180* and *CG8997*), within Africa (*CG15281* and *CG5791*), or in the combined sample (*Cyp6a2* and *CG18179*). In addition, we included genes showing significant expression differences between the two populations, including two with higher expression in Europe (*Cyp6g1* and *CG9509*) and two with higher expression in Africa (*CG7214* and *CG7203*). Finally, we included two control genes that did not show any significant expression differences within or between populations (*Nap1* and *CG15295*).

Prior to qPCR, 5 µg of total RNA was reverse transcribed using Superscript II reverse transcriptase (Invitrogen, Carlsbad, California, USA) and random hexamer primers. The resulting cDNA was used at 1:40 dilution for qPCR using TaqMan probes and a 7500 Fast Real-Time PCR System (Applied Biosciences, Foster City, CA, USA). The probe IDs for the target genes (in the order listed above) were as follows: Dm01801887_s1, Dm01791303_g1, Dm01791414_s1, Dm02147133_g1, Dm01817955_g1, Dm01801878_s1, Dm01819889_g1, Dm01838873_g1, Dm02365366_s1, Dm01809356_g1, Dm01842610_g1, and Dm02539051_s1. All protocol details are available in Appendix A.

Analysis of gene expression

Three replicate assays were performed for each sample and the threshold cycle value (Ct) was averaged across these replicates. Expression levels of the target genes were standardized using the ribosomal protein gene, *RpL32* (Dm02151827_g1) as an endogenous control. For this, a ΔCt value was calculated by subtracting the control Ct value (*RpL32*) from the target Ct value (Figure 12). The fold-change, which represents the difference in expression between two samples (ΔCt_1 and ΔCt_2), was calculated as $2^{-(\Delta Ct_1 - \Delta Ct_2)}$. For comparisons between the European and African populations, ΔCt values were averaged within each population and the African value was used as ΔCt_2 for fold-change calculation

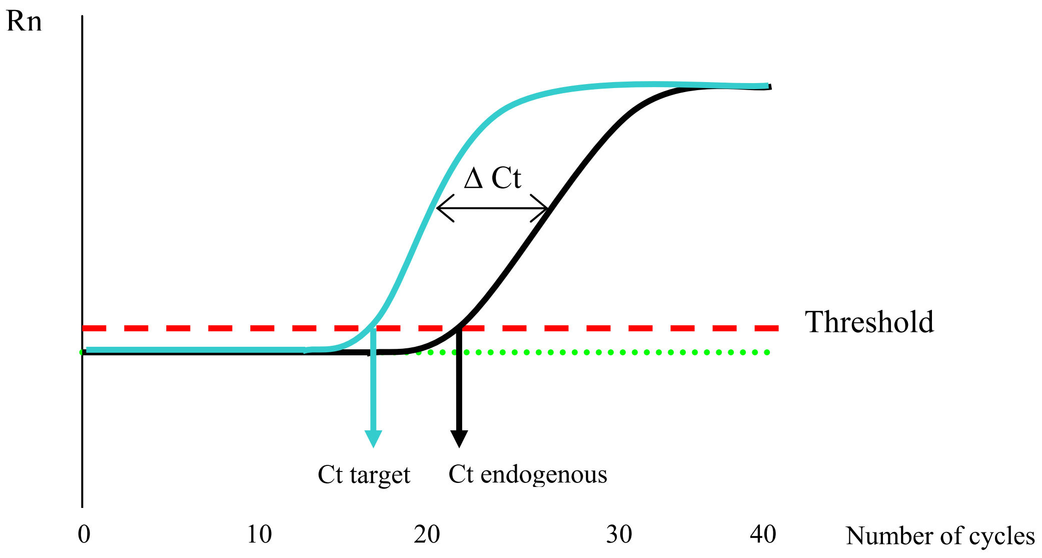


Figure 12: Principle of qPCR analysis for a single sample model.
Rn is emission intensity of the dye.
Ct for the threshold cycle is the fractional cycle number at which the fluorescence passes the threshold. The threshold (red line), automatically determined by the detection software, is the red line whose intersection with the amplification plot defines the Ct.
ΔCt is the normalize value for the sample: the Ct value of the target gene (blue line) is subtracted from that of the endogenous gene (black line).
The green dotted line is the no template control, which does not contain template and it is used to verify amplification quality.

2.3 Results and Discussion

We used qPCR to verify our microarray gene expression results. We also performed qPCR on female samples to examine potential differences in the pattern of expression variation between male and female flies.

Comparison between microarray and male qPCR

We performed qPCR on 12 candidate genes and investigated the fold change in expression between strains for 1154 pairwise comparisons in both microarray and qPCR experiments. Indeed, by measuring fold-changes, we could verify the individual genes' correlation for both experiments. The complete list of the samples used is given in the appendix (see Appendix C). There is a highly significant correlation between the qPCR fold-changes using male samples and the microarray experiment fold-changes with Pearson‘s R = 0.6 (*P*-value < 0.0001) (Figure 13).

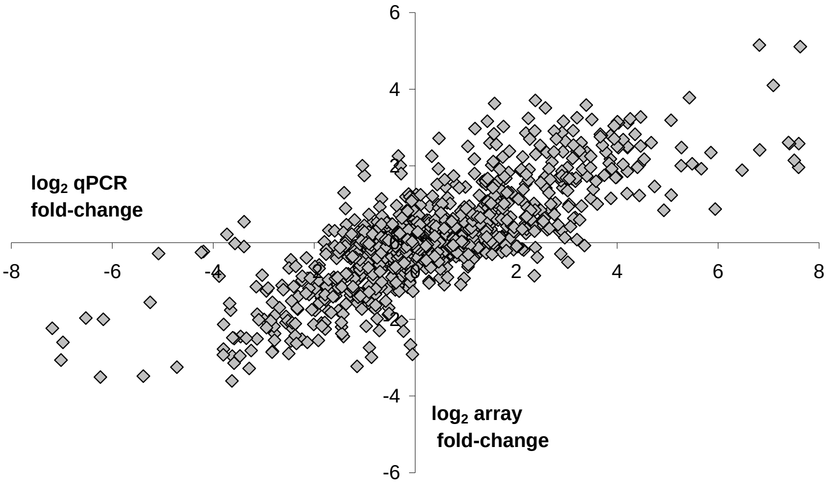


Figure 13: Correlation between fold-changes in expression measured by microarray and qPCR for male samples. Data are from 1154 pairwise comparison of strains across 12 different genes (Pearson's R = 0.6, *P* < 0.0001). Several outlying points are not shown on the graph (see Appendix D)

To verify gene expression differences between populations, as determined by the two-node analysis (see Chapter 1), we used six genes, including two that were significantly over-expressed in Europe, two that were significantly over-expressed in Africa and two with no significant difference between the populations. These last genes were considered as control genes (see Material and Methods). The qPCR experiments in adult males confirmed the microarray results for the differently expressed genes between the two populations (Table 7).

Table 7: Comparison of quantitative PCR and microarray measurements of adult male and female gene expression differences between populations.

Gene	qPCR male E/A (*P*-value)	qPCR female E/A (*P*-value)	microarray E/A (*P*-value)
Cyp6g1	3.26 (0.0008)	3.45 (0.0008)	2.12 ($P < 0.0001$)
CG9509	0.99 (0.003)	1.687 (0.02)	1.21 ($P < 0.0001$)
CG7214	-2.61 (0.002)	-3.96 (0.0008)	-2.42 ($P < 0.0001$)
CG7203	-1.57 (0.005)	-1.33 (0.09)	-2.41 ($P < 0.0001$)
Nap1	-0.46 (0.02)	0.88 (0.09)	0.03 (0.35)
CG15295	0.26 (0.60)	-0.76 (0.30)	0.14 (0.17)

Values represent the $\log_2$ of the mean fold-change difference in expression between the European and African populations as determined by qPCR or microarray. *P*-values were determined by Mann-Whitney *U* test.

For all of these genes a significant difference in expression between the European and African populations was detected (*P*-value < 0.01). Concerning the control genes, only the gene *Nap1* qPCR result showed a significant difference between the populations (*P*-value < 0.05) with higher expression detected in Africa (Table 7). The control genes were chosen because they showed no significant differences between the two populations in the microarray experiments. It is well known that the qPCR technique is more sensitive than the microarray technique and this could explain the above results (Draghici *et al.* 2006). Differences in primer/probe size and location between qPCR and microarray might also contribute to the difference between the two techniques (Canales *et al.* 2006). It should also be noted that no multiple-test correction was applied in the qPCR analysis and that the *Nap1* gene is no longer significant after correction for multiple tests.

Patterns of gene expression variation in male and female flies

Previous microarray surveys showed that many genes are sex-biased in expression (Ranz *et al.* 2003). To investigate, the sex-biased expression of our gene list we used the Sebida database (Gnad and Parsch 2006), which summarizes results of published microarray experiments (Table 8). The results shown are based on three independent surveys directly comparing male versus female gene expression in *D. melanogaster* (Gibson *et al.* 2004; Parisi *et al.* 2004; Wayne *et al.* 2007). The candidate genes are often classified as sex-biased (male or female) or unbiased. However, it is clear from table 8 that there are many inconsistencies in the classification of sex bias from experiment to experiment.

Table 8: Sex-biased expression of genes according to the Sebida database.

Gene	M/F (Gibson *et al.* 2004)§	M/F (Parisis *et al.* 2004) *[whole]**	Expression bias (Wayne *et al.* 2007)
Cyp6g1	3.0239 (*P*=0)	2.6751 (*P* = 0.0082)	male
CG7214	NA	1.3342 (*P* = 0.0688)	female
CG7203	NA	1.2283 (*P* = 0.1746)	unbiased
CG9509	0.9908 (*P* = 0.318)	1.9915 (*P* = 0.0059)	male
Cyp6a2	2.6632 (*P* = 0)	1.6493 (*P* = 0.0511)	male
CG18179	NA	0.3188 (*P* = 0.0083)	male
CG5791	0.6712 (*P* = 0.497)	1.2425 (*P* = 0.2128)	male
CG15281	NA	3.0435 (*P* = 0.0096)	NA
CG8997	1.2702 (*P* = 0.047)	0.8306 (*P* = 0.4134)	male
CG18180	0.8229 (*P* = 0)	0.3086 (*P* = 0.0042)	female
CG15295	1.3634 (*P* = 0)	0.8637 (*P* = 0.0412)	male
Nap1	1.3634 (*P* = 0)	0.2130 (*P* = 0.0009)	female

M/F: Male/Female gene expression ratio
NA indicates that data were not available
P-values are indicated in parentheses
§ from the Gibson *et al.* (2004) survey only the value for two strains combined is shown
* from the Parisi *et al.* (2004) survey only the value from whole flies is shown

We performed qPCR on female samples using 12 candidate genes and the same strains used for the male experiments (see Appendix D). The qPCR experiments in adult females confirm are largely consistent with the result from the male gene expression analysis (Table 7). The results are similar for male and female qPCR measurements, except for the gene *CG7203* which shows no significant difference between the European and African populations by female qPCR. The control gene *Nap 1* has a non-significant *P*-value ($P > 0.05$) between the populations for female qPCR and is closer to the microarray result. We compared the fold-change for 1154 gene and strain pairs between female qPCR and microarray, and male and female qPCR (Figures 14 and 15, respectively). The comparisons show a highly significant correlation between the female samples by qPCR and the microarray experiments with Pearson's R = 0.6 ($P < 0.0001$). This correlation coefficient is similar to that measured by microarray and qPCR for the male samples. The correlation between male and female qPCR samples is also strong, with Pearson's R = 0.8 ($P < 0.0001$) (Figure 15).

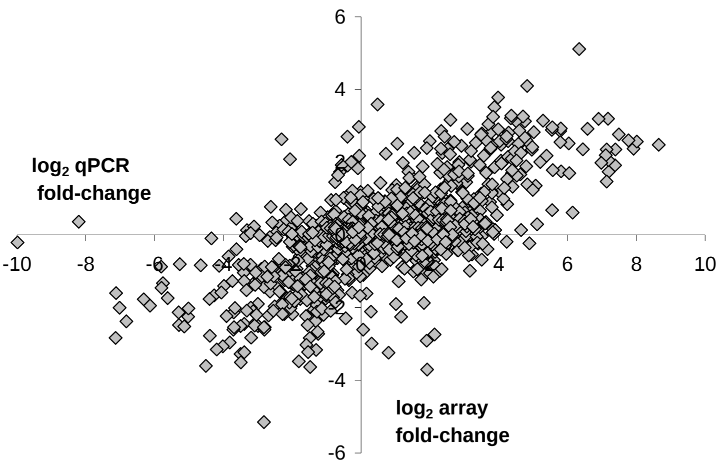


Figure 14: Correlation between fold-changes in expression measured by microarray and qPCR for male samples. Data are from 1154 pairwise comparison of strains across 12 different genes (Pearson's R = 0.6, $P <$ 0.0001). Several outlying points are not shown on the graph (see Appendix D).

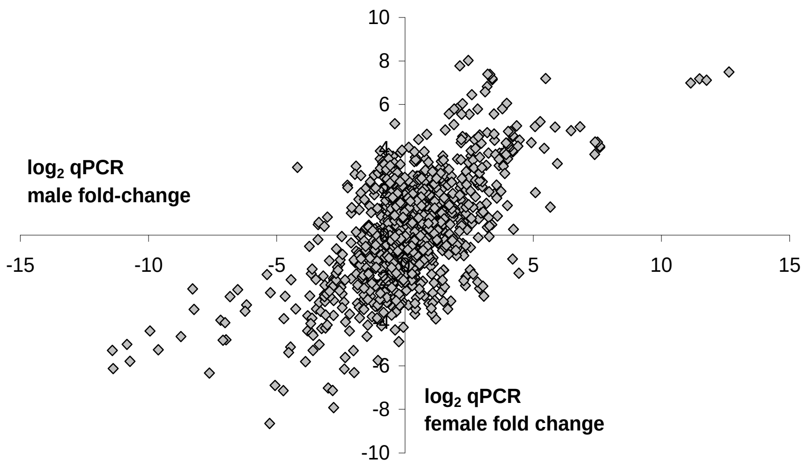


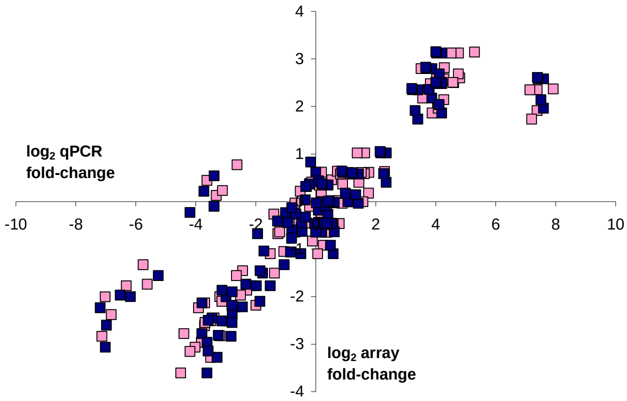
Figure 15: Correlation between fold-changes in expression measured by microarray and qPCR for female samples. Data are from 1154 pairwise comparison of strains across 12 different genes (Pearson's R=0.8, $P <$ 0.0001). Several outlying points are not shown on the graph (see Appendix D).

Overall, the genes analyzed by qPCR are more male-biased than female-biased (Table 8) and since the microarrays were performed on males only, one might expect to see a stronger correlation between male qPCR versus microarray than female qPCR versus microarray. However, our results indicate a similar correlation coefficient for both comparisons. This might indicate that, within strains, these genes show similar levels of expression in males and females. This result is even stronger, when we compare qPCR gene expression fold-changes between males and females (Figure 15). However, we cannot exclude excluded that there are more subtle differences between expression in males and females that cannot be detected even with qPCR. This is supported by the male/female expression ratios in Table 8, which indicate that most of the gene expression differences between males and females are less than two-fold, regardless of the studies considered.

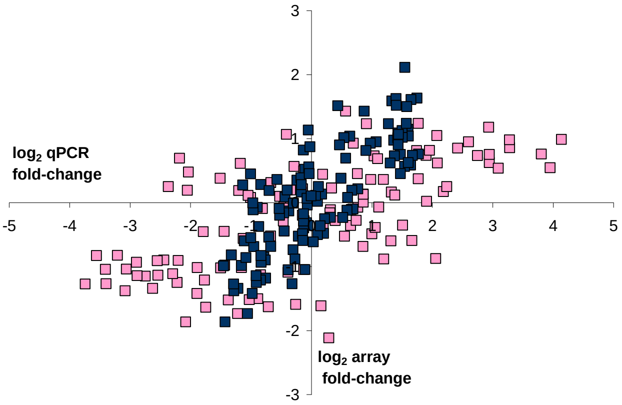
If the pattern of expression differs between males and females, looking at the pattern for each gene separately might be a more sensitive way to detect these differences. Because the endogenous control gene (*RpL32*) used for qPCR differs in expression between males and females, it is not possible to directly compare male/female expression ratios of each gene within each strain. However, we can compare the relative expression levels among strains for each sex. This pattern of expression is similar for the majority of the genes: *Cyp6g1*, *CG7214*, *CG7203*, *CG18179*, *CG15281*, *Cyp6a2* and *CG5791* (Figure 16a, 16c, 16e, 16g, 16f, 16i, and 16l). For these genes, there is a strong correlation between qPCR and microarray expression with a highly significant *P*-value ($P < 0.0001$). The gene *CG15295* shows a negative correlation ($R_F = -0.14$, $R_M = -0.27$) for qPCR versus microarray expression for both sexes (Figure 16h). This gene was a negative control that showed no significant differences in expression among strains in the microarray analysis. This may be a case where the microarray is less sensitive than the qPCR or where the microarray probe is subject to cross-hybridization and does not provide a high-quality gene-specific signal. For the genes *CG9509*, *CG8997* and *CG18180*, the correlation is stronger between male qPCR fold-changes and microarray fold-changes than for the female comparisons (respectively, $R_F = 0.64$, $R_M = 0.80$; $R_F = 0.42$, $R_M = 0.77$; $R_F = 0.59$, $R_M = 0.93$) (Figure 16b, 16d and 16j). This follows the expectation, since the microarray experiments were performed on males only. The *Nap1* gene shows a positive correlation between male microarray expression and female qPCR expression ($R_F = 0.49$), but a negative correlation between male microarray expression and male qPCR expression ($R_M = -0.16$) (Figure 16k). For the male samples, there is almost no correlation between microarray and qPCR. This gene belongs to the control set and this result was already shown in Table 7, which compares European and African populations. Since *Nap1* expression did not differ significantly among strains in the male microarray experiments, one would not expect to see a strong correlation in the male qPCR experiment. For this gene, the strains A377, A186, and A82 are outliers on the graph because they show very low *Nap1* expression in the female qPCR experiments (see Appendix D). Interestingly, the *Nap1* gene is the one that shows the strongest and most consistent female-biased expression (Table 8).

Figure 16: Correlation between fold-changes in expression measured by qPCR (male and female) and microarray (male) for each gene.
N indicates the sample size for each gene
R_F and R_M are the correlation coefficients for female and male, respectively
Blue squares indicate male data
Pink squares indicate female data

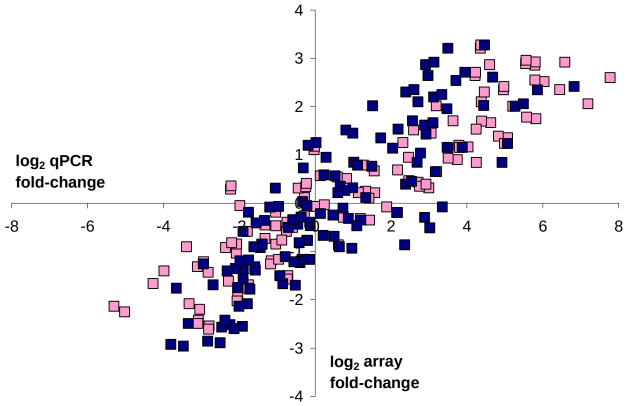
a) *Cyp6g1*
N=120
Pearson's R_F=0.89, $P < 0.0001$
Pearson's R_M=0.88, $P < 0.0001$



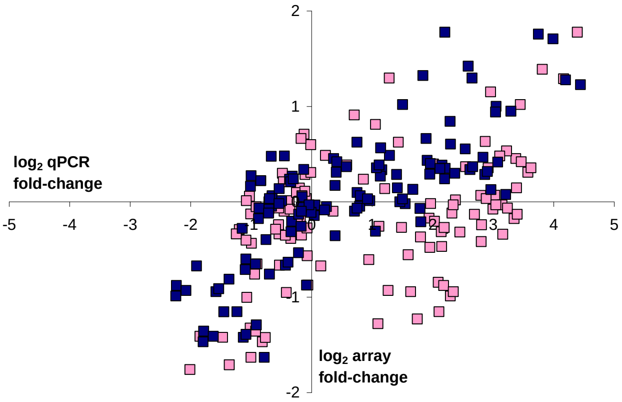
b) *CG9509*
N=120
Pearson's R_F=0.64, $P < 0.0001$
Pearson's R_M=0.80, $P < 0.0001$



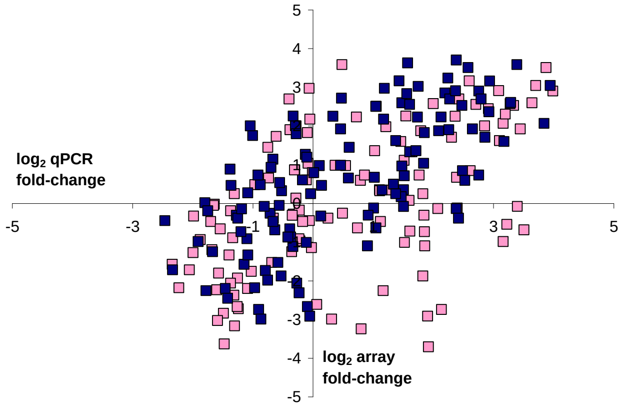
c) *CG7214*
N=120
Pearson's R_F=0.90, $P < 0.0001$
Pearson's R_M=0.85, $P < 0.0001$



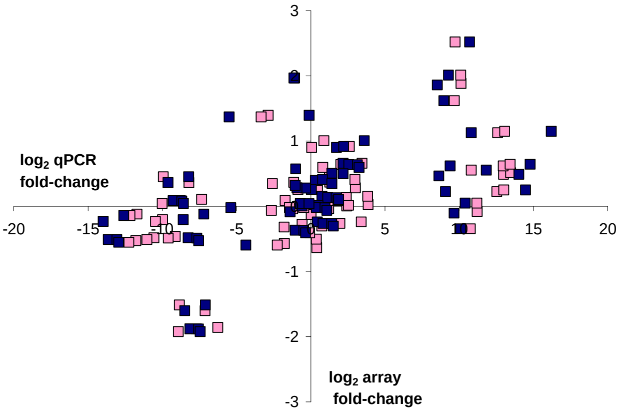
d) *CG8997*
N=120
Pearson's R_F=0.42, $P < 0.0001$
Pearson's R_M=0.77, $P < 0.0001$



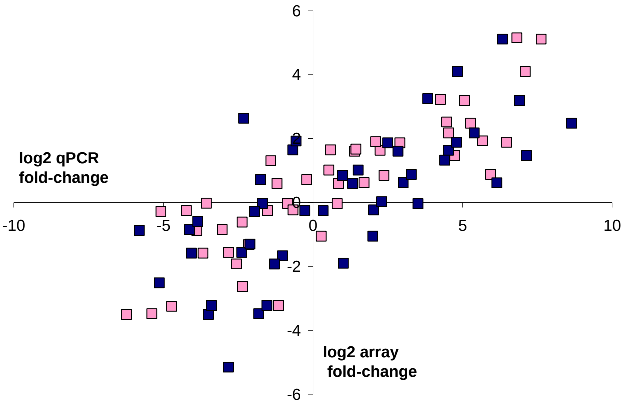
e) *CG7203*
N=120
Pearson's R_F=0.51, $P < 0.0001$
Pearson's R_M=0.67, $P < 0.0001$



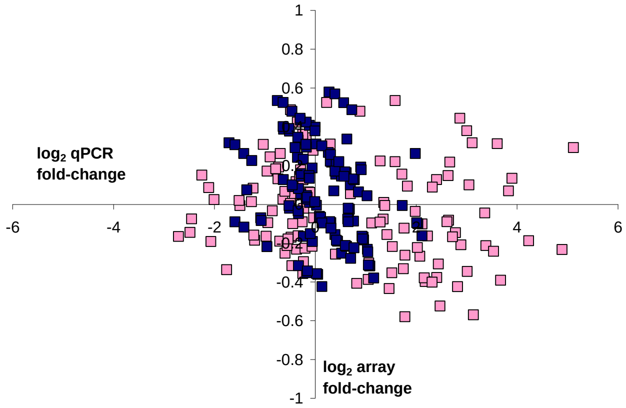
f) *CG15281*
N=78
Pearson's R_F=0.52, $P < 0.0001$
Pearson's R_M=0.54, $P < 0.0001$



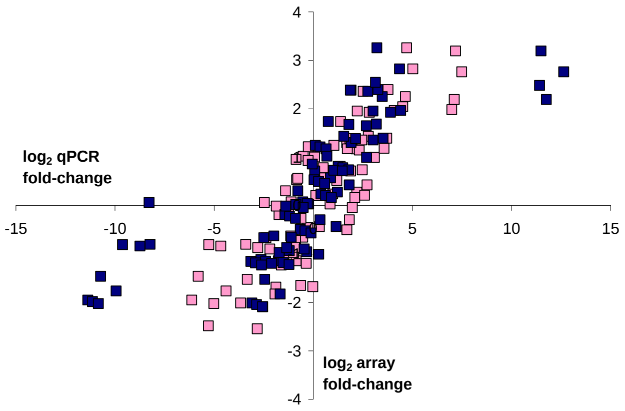
g) *CG18179*
N=45
Pearson's R_F=0.86, $P < 0.0001$
Pearson's R_M=0.70, $P < 0.0001$



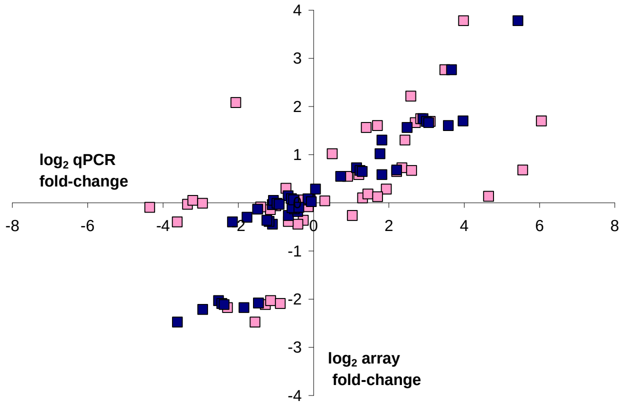
h) *CG15295*
N=120
Pearson's R_F=-0.14, $P < 0.0001$
Pearson's R_M=-0.27, $P < 0.002$



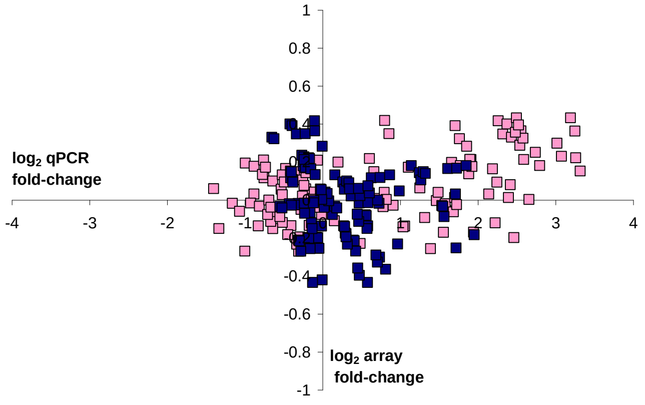
i) *Cyp6a2*
N=91
Pearson's R_F=0.84, $P < 0.0001$
Pearson's R_M=0.78, $P < 0.0001$



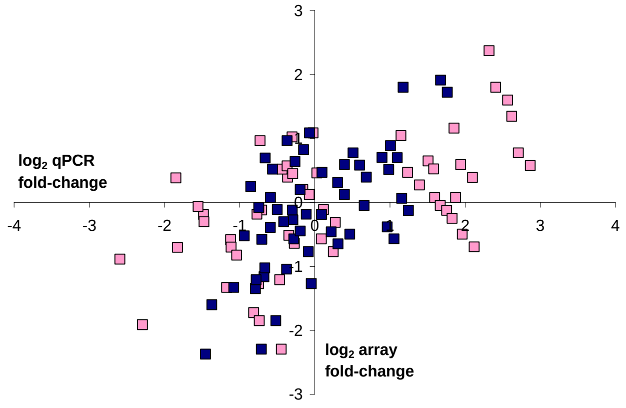
j) *CG18180*
N=45
Pearson's R_F=0.59, $P < 0.0001$
Pearson's R_M=0.93, $P < 0.0001$



k) *Nap1*
N=120
Pearson's R_F=0.49, $P < 0.0001$
Pearson's R_M=-0.16, $P < 0.0001$



l) *CG5791*
N=55
Pearson's R_F=0.57, $P < 0.0001$
Pearson's R_M=0.64, $P < 0.0001$



Interestingly, the highly expressed genes in the African population (*CG7203* and *CG7214*) show the same pattern of expression in both sexes (Table 7). The same is true for *Cyp6g1,* which shows a high level of expression in the European population in both sexes (Table 7). This suggests that the pressure of selection due to the environment is similar for males and females. The gene *CG9509* has a pattern of expression that differs slightly between the two sexes (Table 7). This gene shows significantly higher expression in Europe than in Africa for both sexes, but the expression difference between populations appears to be greater for females. Unfortunately, nothing is known about the function of this gene. A detailed analysis of DNA sequence polymorphism and divergence in the coding and 5' upstream regions of *CG9509* is presented in chapter 4.

For our data, qPCR confirmed the microarray results both within and between populations for male samples. Similar results were obtained for the female samples. However, for some genes, the pattern of expression variation differed between male and female qPCR measurements and microarray measurements. These observations suggest that by increasing the number of genes surveyed by qPCR, we might detect more differences between males and females. Alternatively, this could be examined on a larger scale by repeating the microarray experiments with female-derived cDNA. We also expect that some of the candidate genes for adaptation (those expressed differently between the European and African populations) will differ between males and females, revealing sex-specific differences in the adaptation process. Meiklejohn *et al*. (2003) used intra- and interspecific expression data to show that the selective pressure acting on gene expression differs between male-biased and female-biased genes. Several studies have shown extensive gene expression differences between males and females, as well as differences in expression between African and cosmopolitan strains (Jin *et al.* 2001; Gibson *et al.* 2004; Michalak *et al.* 2007). For example, Jin *et al*. (2001) found that sex-by-genotype interactions are responsible for 10% of the overall gene expression variation in *D. melanogaster*. A comparison of a French strain and a Zimbabwe strain showed that 10% of the genes showed more than a two-fold difference in expression (Osada *et al.* 2006). These genes are involved in different pathways, such as: cell communication, signal transduction and phototransduction and may play a major role in environmental adaptation. Sex-biased expression is a an important factor in evolution and may also play a role in speciation (Zhang *et al.* 2007)

Chapter 3 The relationship between gene expression polymorphism and DNA sequence polymorphism in coding and non-coding regions of genes in natural populations of *Drosophila melanogaster*

3.1 Introduction

DNA sequence polymorphisms are a major source of heritable phenotypic differences among individuals. However, in most cases, the relationship between particular sequence variants and an organism's phenotype is unknown. Understanding the complex relationship between genotype and phenotype would provide insight into the mechanism of adaptation. One phenotype that may be especially amenable for investigating this relationship is gene expression. This is because, in many cases, DNA sequence variants may be tightly linked to the genes whose expression they affect. Genome-wide microarray analysis is a recently developed technology that can be used to measure gene expression on a global scale and can be combined with population-level DNA sequencing to investigate the relationship between genotype and phenotype. Previous studies in this area have shown that genes with significant expression variation tend to be more divergent between species at the amino acid level (Meiklejohn *et al.* 2003; Holloway *et al.* 2007). Additionally, highly expressed genes tend to show lower levels of polymorphism and divergence in their coding regions (Nuzhdin *et al.* 2004).

Transcription is regulated by interactions between *cis*-regulatory elements, which are physically linked to the gene they regulate (often in the proximal promoter), and *trans*-regulatory elements that are not linked to the gene and located further away on the chromosome or on a different chromosome. These *trans*-elements typically encode regulatory proteins such as transcription factors (Rockman and Kruglyak 2006). The relative contribution of changes in *cis*- and *trans*-regulatory elements to gene expression variation remains controversial. However, several studies suggest that changes in *cis* may be the main contributor to mRNA expression differences observed between individuals (Cowles *et al.* 2002; Rockman and Wray 2002; Morley *et al.* 2004) or between species (Jeong *et al.* 2008; Wittkopp *et al.* 2008). The c*is*-regulatory region of a gene is often located within a region of ~1 kb immediately upstream of the start codon (Wray *et al.* 2003) and it has already been demonstrated that nucleotide changes in this regions can affect the phenotype and fitness of the organism (Daborn *et al.* 2002; De Gobbi *et al.* 2006; Wray 2007).

Transcriptional profiling with whole-genome microarrays has been used to survey gene expression variation between two natural populations of *Drosophila melanogaster*, including eight strains from Africa (Zimbabwe) and eight from Europe (the Netherlands; see Chapter 1 and Hutter *et al.* (2008)). This study revealed that 153 genes differed significantly in expression between the two populations at a false discovery rate (FDR) of 9%. This represents ~3% of all genes detected as expressed in the adult males, and these genes are good candidates for genes that have undergone adaptive regulatory evolution accompanying the out-of-Africa expansion of *D. melanogaster*, which occurred around 10,000-15,000 years ago (David and Capy 1988; Lachaise *et al.* 1988). The underlying genetic basis for the difference in gene expression between the two populations is unknown, although it is likely that a substantial fraction is caused by polymorphism in *cis*-regulatory sequences. In this chapter, I investigate this possibility by sequencing the coding and upstream regulatory regions of genes that differ in expression between the populations and comparing their molecular evolution to that of a set of control genes that do not differ in expression between populations. In general, I find evidence for both purifying and positive selection in coding and regulatory regions, although the amount and type of DNA sequence variation do not differ greatly between the genes that vary in expression and the control genes.

3.2 Materials and Methods

PCR amplification and DNA sequencing

Based on *Drosophila melanogaster* DNA sequences available from Ensembl 4.4 (www.ensembl.org), primers were designed to amplify the coding region and approximately 1 kb upstream of the candidate and control genes using the Primer3 software (*http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi*) (Rozen and Skaletsky 2000). For the region of interest, the primers were designed to amplify overlapping fragments of around 800 bp. Genomic DNA extraction was performed on one male fly of each strain (see appendix B). Fragments were PCR-amplified using PeqLab specific buffer (Erlangen, Germany) and the PCR fragments were purified using ExoSAP-IT (USB, Cleveland, USA). Sequencing reactions using ABI BigDye ddNTPs were read by an ABI 3730 automated sequencer (Applied Biosystems, Foster City, CA, USA). Sequence reads were aligned with Seqman (DNAstar, Madison, WI, USA). The PCR products were sequenced in both forward and reverse directions. All protocol details are available in Appendix A.

Analysis

Several annotation changes occurred between Ensembl 4.4 and the current genome release, version 5.0. The gene *CG9973* could only be aligned for exons 1 and 2 after a change in its annotation. The gene *CG12683* was no longer annotated as coding, however it was detected as expressed in the microarray; only its upstream regions were used for the analysis. For *CG876*8 and *CG15314,* the outgroup (*D. simulans*) did not have ATG as the start codon, this codon was not included in the analysis.

The consensus regions (intergenic and coding) of the *D. melanogaster* strains and of the outgroup were aligned with the ClustalW algorithm using the BioEdit application (BioEdit v7.0.9(Thompson *et al.* 1994)) and manually corrected. For the outgroup, sequence data available on the University of California Santa Cruz (UCSC) Genome Browser (Hinrichs *et al.* 2006) was used. Whenever possible the *Drosophila simulans* DNA sequence was used as the outgroup, otherwise *D. sechellia* was used. The local recombination rate was estimated for each gene with the computer program “Recomb-rate” developed by Comeron *et al*. (1999).

All of the basic polymorphism and divergence analyses were performed using the program DnaSP 4.50.3 (Rozas *et al.* 2003). For each locus, different sites were considered: upstream sites (5’) containing 5’ UTR and intergenic regions, synonymous sites (S), non-synonymous sites (NS), and intron sites (I). We assumed that synonymous sites evolve neutrally and could be used as a control to detect selection at the other types of sites. We separated the sites into those that are polymorphic within *D. melanogaster* (P) and those that show a fixed differences between *D. melanogaster* and the outgroup (D). The MK test (McDonald and Kreitman 1991) was performed to compare fixed differences between *D. melanogaster* and the outgroup and the polymorphisms within *D. melanogaster*.

The neutrality index (NI) was calculated using synonymous sites (indicated by *s*) and non-synonymous sites or upstream sites or intron sites (indicated by *x*) (Rand and Kann 1996): $NI = \frac{(Px/Dx)}{(Ps/Ds)}$

For all site categories, nucleotide diversity was estimated by the both θ (Watterson 1975) and π (Tajima 1989) and divergence was estimated by *K* (Nei 1987). The neutral equilibrium model was tested using Tajima's *D* (Tajima 1989) and the mutli-locus HKA test (Hudson *et al.* 1987). Both tests were performed using the program, HKA (kindly provided by J. Hey) in which the test statistics were compared with distributions generated by 10,000 neutral coalescent simulations (Kliman *et al.* 2000).

3.3 Results and Discussion

Description of the data

Based on the two-node analysis of Chapter 1, 153 genes showing significant gene expression differences between African and European strains were identified (see Chapter 1). From this list we chose 23 genes showing significant expression differences between the two populations for further population genetic analysis. We also included nine control genes with no significant difference in expression between the two populations (Table 9). The genes are from diverse chromosomal locations (eight genes are located on the X chromosome and 24 on the autosomes) and a diverse range of functional categories (or unknown function) to avoid bias towards genes of a particular genomic location or functional class (Table 9). We excluded genes if they contained coding sequences of another gene within 1 kb upstream of their start codon, because constraints on the coding region may affect the evolution of the intergenic region. For each gene sequenced the entire coding region as well as the putative 5' regulatory region (or proximal promoter), which we defined as the region ~1 kb upstream of the start codon. Thus, the 5' upstream region is composed of both 5' UTR and intergenic sequences (Table 10). In all cases, sequencing was performed in both an ancestral African population and a derived European population.

Polymorphism and divergence

The African and European populations were chosen because they are expected to have different demographic and adaptive histories. The African population of *D. melanogaster* is ancestral and appears to be near equilibrium, although it may have undergone a mild population size expansion within the past 60 million years (Glinka *et al.* 2003; Ometto *et al.* 2005; Li and Stephan 2006). In contrast, the European population has experienced a strong bottleneck upon leaving Africa about 12,000 years ago (Li and Stephan 2006). The consequence of this is that DNA sequence polymorphism is considerably reduced in the European population (Begun and Aquadro 1993; Baudry *et al.* 2004; Haddrill *et al.* 2005). Because the bottleneck occurred relatively recently, the majority of the polymorphisms present in Europe are expected to be present also in Africa. The sequences analyzed here also indicate that there is more polymorphism in the African population than in the European population. For example, at synonymous sites π is almost twice as high in Africa than in Europe (Table 10).

Table 9: List of the genes used for this study.

CG number	Gene name	Two-node	Expression		Function	Loc.	Cytogenetic map	Rec/bp§
			Afro	Euro				
CG9509		s	1	2.311	dehydrogenase	X	13A	4.83E-08
CG9511		s	1	1.461	unknown	2L	26D	4.4E-08
CG1468		s	1	1.35	unknown	X	9A	3.54E-08
CG5386		s	1	1.151	unknown	3R	94A	3.85E-08
CG5154	Idgf5	s	1	1.407	chitinase	2R	55C	1.37E-08
CG14629		s	1	1.258	unknown	X	1E	7E-10
CG7203		s	5.317	1	unknown	2L	28C	4.36E-08
CG7214		s	5.361	1	unknown	2L	28C	4.36E-08
CG8997		s	1.671	1	unknown	2L	34D	2.65E-08
CG7953		s	1.411	1	unknown	2L	34D	2.65E-08
CG7916		s	1.562	1	unknown	2L	34D	2.65E-08
CG33306		s	1.354	1	unknown	2L	34D	2.65E-08
CG5178	Act88F	s	2.917	1	actin filament	3R	88F	2.294E-08
CG5402		s	1.53	1	unknown	3R	98A	3.55E-08
CG5144		s	1.439	1	kinase	3L	66F	4.53E-08
CG10912		s	1.622	1	unknown	2R	55B	1.44E-08
CG10597		s	1.32	1	unknown	X	15F	2.78E-08
CG4734		s	1.47	1	unknown	2R	50A	2.25E-08
CG8661		s	1.836	1	unknown	X	15F	2.78E-08
CG9973		s	1.329	1	unknown	3L	63A	3.73E-08
CG5623		s	1.17	1	unknown	3R	89A	2.37E-08
CG13061	Nplp3	s	1.849	1	neuropeptide	3L	72E	1.47E-08
CG5210	Chit	s	1	1.321	chitinase	2R	53D	1.89E-08
CG8768		ns	1	1.009	unknown	2R	49C	2.15E-08
CG13675		ns	1	1.114	unknown	3L	66B	4.81E-08
CG9602		ns	1.071	1	ligase activity	3R	87F	1.85E-08
CG16916	Rpt3	ns	1	1.071	protease	X	10A	2.55E-08
CG5832	Hmx	ns	1	1.071	protein binding	3R	90B	2.85E-08
CG12912		ns	1	1.049	unknown	2R	46F	1.29E-08
CG12683		ns	1	1.1	unknown	X	4D	4.54E-08
CG14503	Tango8	ns	1	1.068	unknown	2R	55C	1.37E-08
CG15314		ns	1	1.018	unknown	X	9A	3.54E-08

(CG numbers are from Flybase release 5.4; the two-node test indicates whether gene expression is significantly different (s) or not (ns) between the two populations. Expression indicates the relative gene expression in the African and European populations (see to Chapter 1). Loc, is the gene's chromosomal location.

§based on Comeron *et al*. (1999).

Table 10: Polymorphism and divergence in coding and non-coding regions.

Region	number of loci	Mean size [a]	Mean π [b]	Mean K [c]	D [d]	P [e]	p-value [f]	P§	*P*-value [g]
Europe									
Synonymous	32	197.43	1.37	11.41	582	262	-	169	-
Non-synonymous	32	628.73	0.13***	2.12***	356	81	$1x10^{-6}$	52	$4.4x10^{-5}$
Non-coding	32	605.97	0.96	7.97	1982	750	0.044	516	0.284
Upstream region	32	1024.72	0.74**	5.89***	1653	633	0.068	438	0.378
• 5'UTR	21	54.76	0.27	5.03	48	11	0.056	7	0.094
• Intergenic region	32	990.66	0.75	5.92	1605	622	0.099	432	0.468
Intron	24	187.21	1.19	10.04	329	117	0.072	78	0.202
Africa									
Synonymous	32	197.65	2.16	11.79	543	454	-	271	-
Non-synonymous	32	630.25	0.27***	2.13***	343	146	$1x10^{-6}$	67	$1x10^{-7}$
Non-coding	32	605.97	1.40	7.96	1857	1235	0.001	611	$1x10^{-6}$
Upstream region	32	1024.72	1.09***	5.84***	1548	1042	0.004	510	$6x10^{-6}$
• 5'UTR	21	54.76	0.67	5.20	46	24	0.081	10	0.018
• Intergenic region	32	988.78	0.75	5.85	1502	1018	0.006	500	$9x10^{-6}$
Intron	24	187.21	1.71*	10.08	309	193	0.009	101	0.002

[a] Mean number of sites sequenced per locus

[b] Weighted average within *D. melanogaster* populations of pairwise diversity per 100 sites

[c] Weighted average of between species pairwise divergence per 100 sites (between *D. melanogaster* populations and the outgroup)

[d] Number of fixed differences between species

[e] Number of polymorphisms within the population

[f] *P*-value from two-tailed Fisher exact test using all polymorphisms

§ Number of polymorphisms, excluding singletons

[g] *P*-value from two-tailed Fisher exact test after excluding all singleton polymorphisms

Asterisks indicate a significant Wilcoxon test comparing synonymous sites to non-synonymous sites, upstream sites, or to intron sites: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

In Africa, we observed a highly significant difference for divergence K and π in the synonymous sites relative to the upstream regions and the non-synonymous sites (Wilcoxon test, $P < 0.0001$) (Table 10). For the introns, the difference in polymorphism with synonymous sites was significant ($P > 0.05$). However, there was no significant difference in divergence between intronic and synonymous (Table 10). In Europe there were significant differences between synonymous sites, non-synonymous and upstream sites for π and K (Table 10). This was not the case for the introns, in which the π and K values were not different from those of the synonymous sites (Table 10). The overall pattern is that levels of polymorphism and divergence are reduced at non-synonymous and upstream sites, which suggests that these sites are under functional and selective constraints.

To better determine the type of selection acting on coding and non-coding regions, MK tests were performed (McDonald and Kreitman 1991). This test allows us determine if positive selection has played a role in shaping the evolution of the different classes sites. We compared the levels of polymorphism within a population to the divergence between *D. melanogaster* and its sisters species *D. simulans* and *D. sechellia*. The comparison was made between a putatively selected class of sites (*e.g.*, non-synonymous, upstream, intron) and sites that are assumed to evolve neutrally (synonymous sites). If there has been recurrent positive selection, the relative ratio of divergence to polymorphism will be higher at the selected sites than the control sites. To perform the MK test, we summed the polymorphism and divergence counts within each class of sites over all genes. This pooling of data increases the power to detect departures from neutrality (Table 10). Within Europe, we observed a significant signal of positive selection when comparing non-synonymous sites to synonymous sites and when comparing the entire non-coding region to synonymous sites (Table 10). In Africa, there was a significant excess of divergence for non-synonymous sites ($P = 10^{-6}$) and for all non-coding region in general ($P = 0.001$), with the exception of the 5' UTR (Table 10). This suggests that a significant proportion of the divergence between species is driven by positive selection. The presence of slightly deleterious mutations, which are kept at low frequency by purifying selection, can bias the MK-test away from the detection of positive selection (Fay *et al.* 2001; Charlesworth and Eyre-Walker 2008). This is because slightly-deleterious mutations are subject to weak negative selection and contribute more to polymorphism than to divergence. To increase the power of the MK test, we removed all low frequency polymorphisms (singletons) and repeated the analysis (Andolfatto 2005; Proeschel *et al.* 2006). When applying this method, all of the putatively selected classes showed a significant ($P < 0.05$)

signal of positive selection in Africa, while only non-synonymous sites were significant ($P < 4.4\text{x}10^{-5}$) in Europe (Table 10).

To quantify the contribution of individual genes to the above patterns, we also performed the MK test separately for each gene (Table 11). For each gene we also calculated the neutrality index (NI) (Rand and Kann 1996). An NI < 1 is indicative of the positive selection, while NI > 1 can be explained by balancing selection or weak purifying selection. To distinguish between these two possibilities (balancing selection and weak purifying selection), we also examined the Tajima'*s D* value. A negative Tajima'*s D* value indicates an excess of rare variants and would support the hypothesis of purifying selection, while a positive value is evidence for balancing selection, since it indicates an excess of variants in intermediate frequency. For the coding regions, six genes gave a significant MK test (Table 11A). The genes *CG5623*, *CG7916*, *CG8997* and *CG9505* showed evidence of recurrent positive selection (NI < 1) (Table 11A). In Africa, the gene *CG14629* has a NI > 1 and a negative Tajima's *D* value for synonymous and non-synonymous sites (indicating an excess of rare variants). This suggests that the gene is subject to weak purifying selection. This gene is located in a part of the genome with a low recombination rate ($7\text{x}10^{-10}$ recombination events per bp) (Table 9). In such regions, both positive and negative selection are expected to be less effective due to Hill-Robertson interference (Hill and Robertson 1966; Zhang and Parsch 2005). For this reason, this gene was excluded from further analyses.

The upstream regions of seven genes were detected as being under positive selection (NI < 1) (Table 11B), with *CG9602* significant only in Europe and *CG12912* significant only in Africa. The gene *CG5178*, which encodes an Actin protein shows evidence of balancing selection in the European population (NI > 1 and Tajima's *D* > 0). In both populations, *CG5402* shows the same pattern, with NI > 1 and Tajima's *D* > 0. In the African population, the gene *CG10912* has NI > 1 and Tajima's *D* < 0, which suggests weak purifying selection. Two genes located on the X chromosome, *CG9509* and *CG16916,* showed very low levels of polymorphism in Europe (Table 11B), which could be a remnant of the European bottleneck, or could result from a recent selective sweep in or near these genes. In the intronic regions, three genes (gene *CG5210*, gene *CG7916*, gene *CG16916*) gave evidence of positive selection (Table 11C). Interestingly, the gene *CG7916* shows evidence of positive selection at all sites analyzed.

A) Coding region

In Europe

Gene	A/E	Chr	Dn	Pn	Ds	Ps	NI	TDs	TDn
CG5623	1.170	3R	16	1	16	10	0.1*	-0.42	-1.12
CG7916	1.562	2L	6	1	14	20	0.12*	0.83	-1.15

In Africa

Gene	A/E	Chr	Dn	Pn	Ds	Ps	NI	TDs	TDn
CG14629	0.795	X	7	7	23	3	7.67*	-1.63	-0.91
CG7916	1.562	2L	6	1	13	24	0.09*	-0.35	-1.15
CG8997	1.670	2L	10	2	13	16	0.16*	-0.40	1.26
CG9509	0.432	X	49	8	40	40	0.16***	-0.36	-1.58
CG5832	ns	3R	0	3	21	8	NA*	0.56	-1.58

B) Upstream region

In Europe

Gene	A/E	Chr	$D_{5'}$	$P_{5'}$	Ds	Ps	NI	$TD_{5'}$
CG5402	1.530	3R	42	37	20	1	17.6***	0.23
CG5178	2.617	3R	39	50	14	6	2.99*	0.90
CG9511	0.684	2L	76	21	40	24	0.46*	1.01
CG7916	1.562	2L	69	35	14	20	0.36*	-0.59
CG9509	0.432	X	84	1	44	7	0.07**	1.49
CG9602	ns	3R	34	9	11	12	0.24*	-0.22
CG16916	ns	X	56	1	20	5	0.07**	-1.14

In Africa

Gene	A/E	Chr	$D_{5'}$	$P_{5'}$	Ds	Ps	NI	$TD_{5'}$
CG5402	1.530	3R	39	41	20	1	21.03***	0.75
CG10912	1.622	2R	82	46	33	6	3.09*	-1.11
CG9511	0.684	2L	75	18	38	43	0.21***	-0.57
CG7916 §	1.562	2L	71	34	13	24	0.26***	0.11
CG8997 §	1.670	2L	71	34	13	16	0.39*	
CG9509	0.432	X	76	33	40	40	0.43**	-1.14
CG12912	ns	2R	78	39	3	7	0.21*	-1.39
CG16916	ns	X	48	29	17	26	0.40*	0.08

C) Intronic region

In Europe

Gene	A/E	Chr	D_I	P_I	Ds	Ps	NI	TD_I
CG5210	0.757	2R	43	8	23	16	0.27**	1.44

In Africa

Gene	A/E	Chr	D_I	P_I	Ds	Ps	NI	TD_I
CG5210	0.757	2R	34	18	19	32	0.31**	-1.07
CG7916	1.562	2L	8	3	13	24	0.20*	1.54
CG16916	ns	X	16	4	17	26	0.16**	-0.94

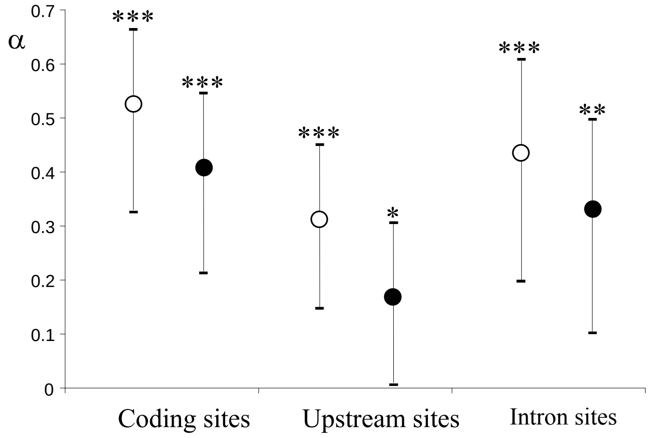
Table 11: Genes with significant McDonald-Kreitman tests.
(A/E, indicates the Africa/Europe gene expression ratio for the genes showing a significant gene expression differences between the populations and *ns* indicates that the genes did not differ significantly in expression between the populations.
Chr is the chromosomal location.
Ps, Pn, $P_{5'}$, P_I, Ds, Dn, $D_{5'}$and D_I are respectively polymorphism and divergence for synonymous, non-synonymous, upstream and intron sites.
NI is the neutrality index of Rand and Kann (1996).
TDs, TDn, TD_5 and TD_I are Tajima's *D* values for synonymous, non-synonymous, upstream and intron sites, respectively.
Asterisks indicate a significant Fisher's exact test comparing divergence and polymorphism of synonymous to non-synonymous sites, upstream sites and to intron sites: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, NA: not applicable
P-value is from two-tailed Fisher exact test
§ CG8997 and CG7916 shared the same upstream region.

In the dataset of genes with significant expression differences between populations (SG), 39% are significant for the MK –test. In the control (NSG) dataset, 33% of the genes gave a significant MK. Thus, there is very little difference between the two groups of genes. Overall, positive selection seems to be a major force in the evolution of non-synonymous and non-coding sites in many genes.

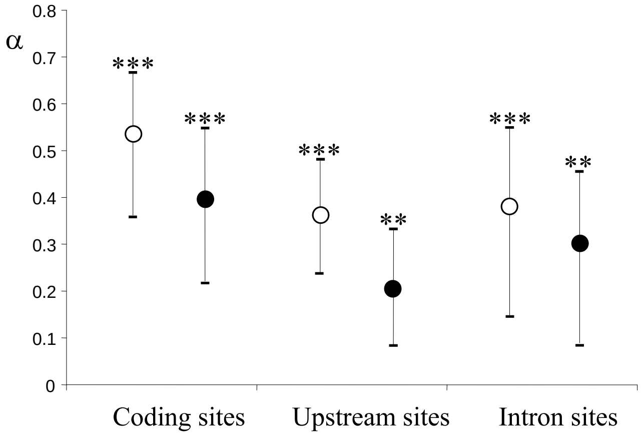
To quantify the fraction of divergence attributable to positive selection, we calculated the parameter α using a maximum-likelihood method (Bierne and Eyre-Walker 2004). In the coding regions of the African population, α values are 40% for the SG and NSG datasets (Figure 17A). The percentage of nucleotide replacements driven by positive selection in the upstream regions of all genes is 20%. It is and 17% when considering only genes differentially expressed between populations (Figure 17A). In the intron regions we estimate that 30% of the nucleotide replacements were driven by positive selection in all genes. For SG datatset, the corresponding estimate is 33%. In Africa, the α value for all classes of sites is significantly greater than zero based on a likelihood ratio test (Figure 17A). In the European population, the test was applied to coding and upstream sites only, because the introns did not contain enough polymorphic sites to run the test (Figure 17B). In the coding region, 36% of the amino acid replacements in all genes are estimated to be driven by positive selection. The corresponding estimate for differentially expressed genes is 40% (Figure 17B). In the upstream regions, α is 16% for all genes and 7% for SG dataset, but it is not significantly different from zero according to the likelihood ratio test. In order to reduce the effect of weak purifying selection, all of the analyses were performed after removing the singleton polymorphisms. In Africa, α values for coding sites were 53% for both datasets. For upstream regions, the values of α were 36% and 31% for all genes and for the differentially expressed genes, respectively. For the intronic sites, α values were 38% and 44% for all genes and SG dataset, respectively. All of the tests were significant (Figure 17A).

A) In Africa

- Differentially expressed genes

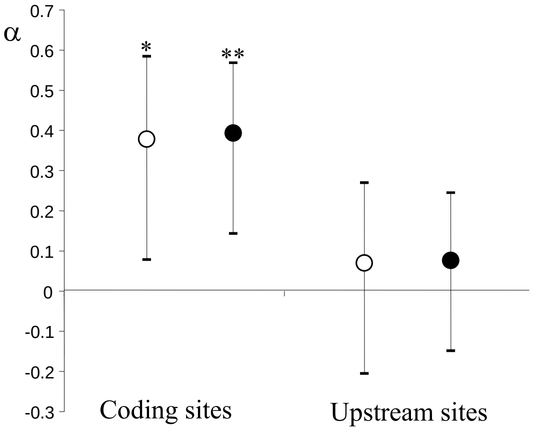


- All expressed genes



B) In Europe

- Differentially expressed genes



- All expressed genes

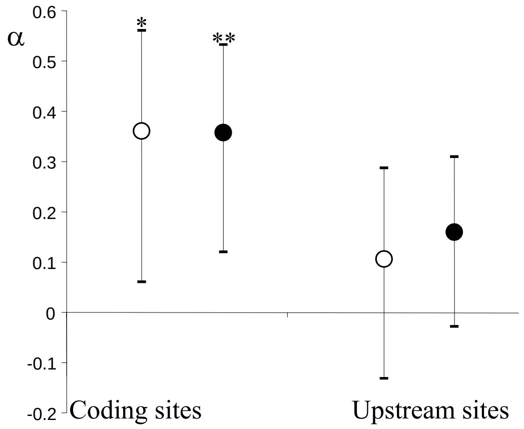


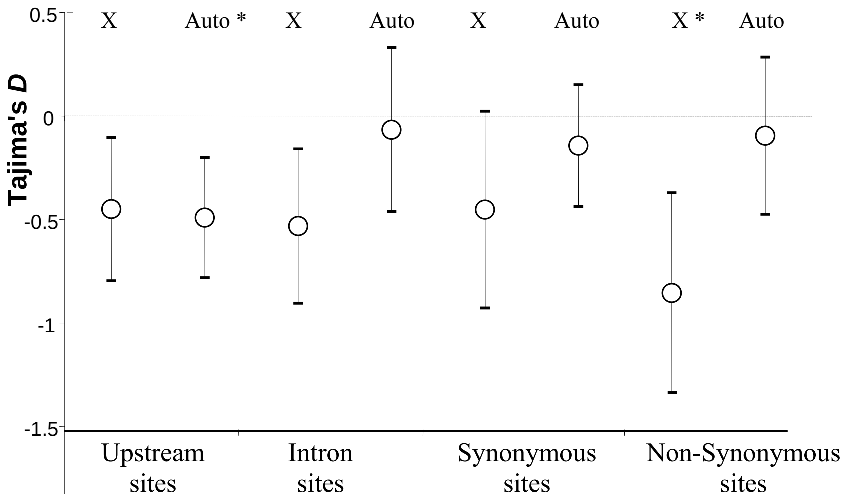
Figure 17: Proportion of amino acid substitutions driven by adaptive evolution.
Estimation of α the fraction of nucleotide divergence driven by positive selection calculation using maximum likelihood method (Bierne and Eyre-Walker 2004). Error bars represent 95% confidence intervals. Solid circles indicate data including singletons and open circles indicate data excluding the singletons. Asterisks indicate a significant likelihood-ratio test: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$

Within the coding regions of the European population, α values for all of the genes and for the differentially expressed genes were 36% and 38%, respectively, and both were significant for positive selection by the likelihood ratio test (Figure 17B). For the upstream regions, α values were 11% and 7% for all genes and for the differently expressed genes, respectively, and were significantly greater than zero (Figure 17B). The α values are lower in Europe than in Africa, which suggests the presence of more segregating deleterious mutations in Europe.

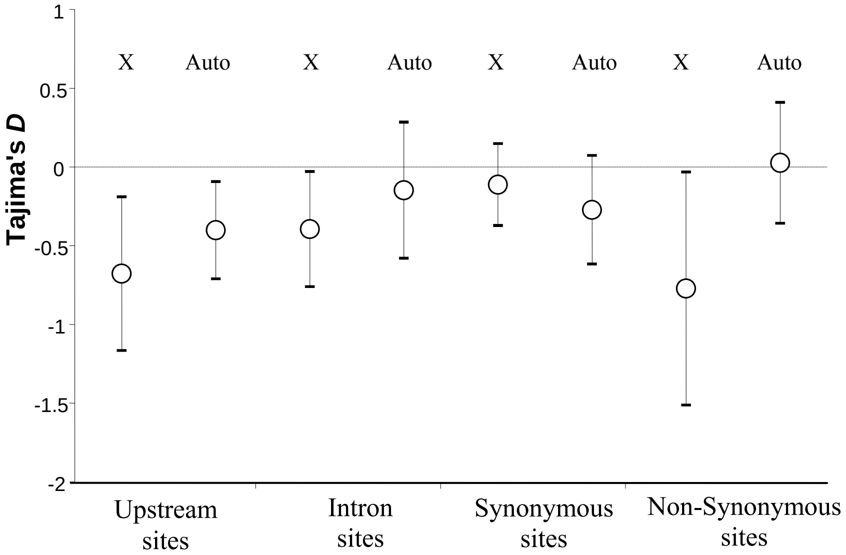
The reduced levels of polymorphism and divergence at non-synonymous and upstream sites (Table 11) relative to synonymous sites suggest that the former are under increased selective constraint. If so, one would expect Tajima's *D* to be skewed towards a negative value at these sites if purifying selection acts to keep deleterious variants at lower frequency than those are that are neutral (Tajima 1989; Andolfatto 2005; Haddrill *et al.* 2008). In Africa, considering all of the genes, Tajima's *D* is significantly skewed towards negative values for the autosomal genes in upstream sites (Figure 18A). The genes located on the X chromosome are significantly skewed towards negative values for non-synonymous sites only. These observations support the assumption that at least some of the non-synonymous and non-coding sites evolve under purifying selection. On the autosomes, the upstream regions show the strongest effect of weak purifying selection, while on the X chromosome the non-synonymous sites show the strongest effect. For the autosomes, this pattern is similar in both data sets (SG and NSG) (Figure 18A).

A) In Africa

- All expressed genes

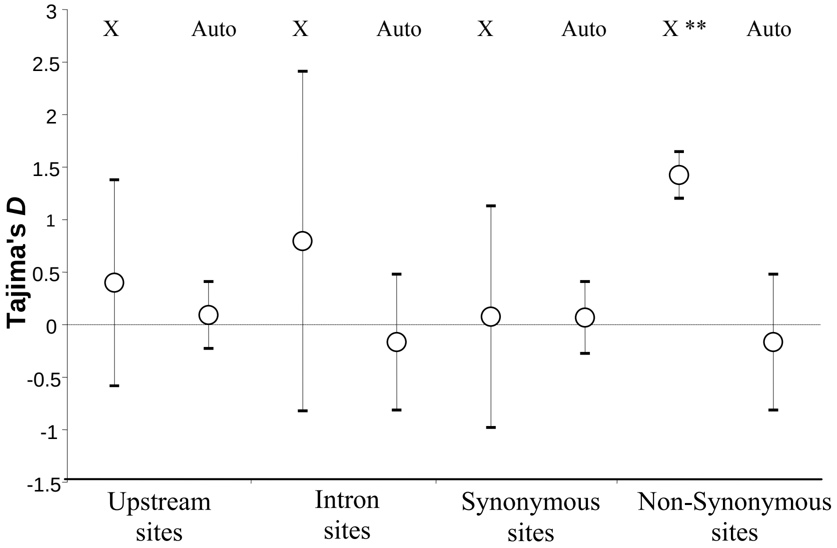


- Differentially expressed genes



B) In Europe

- All expressed genes



- Differentially expressed genes

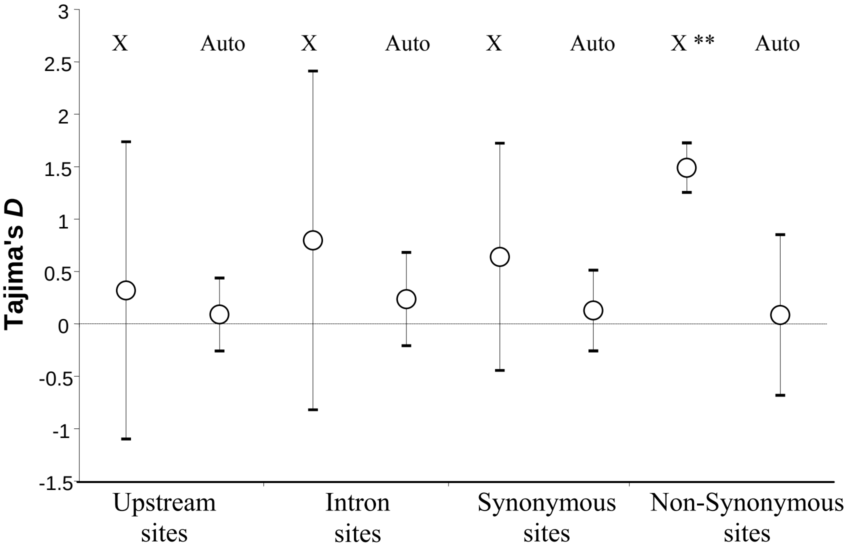


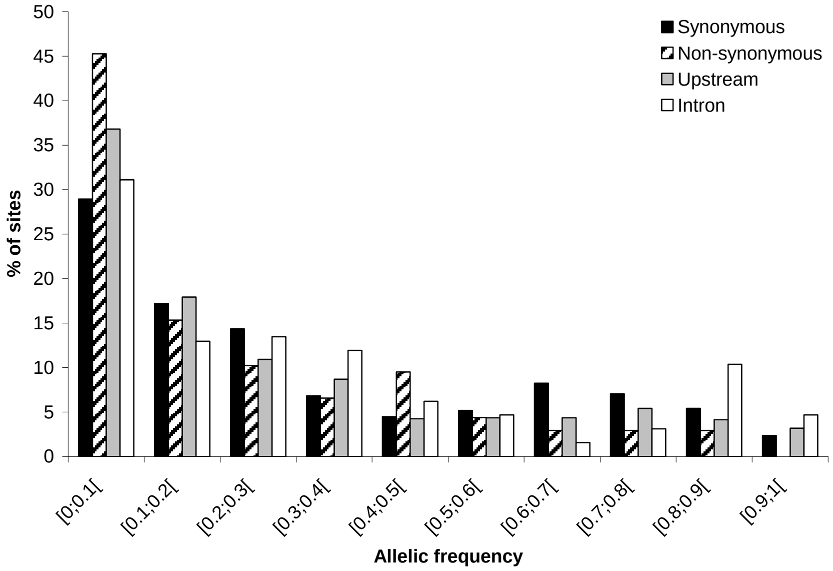
Figure 18: Mean values of Tajima's *D*.
The expectation of Tajima's *D* under the neutral equilibrium model is indicated by the dotted line. The Tajima's *D* values for genes located on the X chromosome (X) and on the autosomes (Auto) is shown for each class of sites. Error bars indicate the 95% confidential intervals of the mean. Asterisks indicate a significant difference from the neutral equilibrium model: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$

For the X chromosome, both selected classes skewed toward more negative values than the synonymous sites. This is mainly due to differences at synonymous sites between the two datasets. When considering all the genes, the value of *D* was more negative (-0.45) than when considering only the differentially expressed genes (-0.11) (Figure 18A). However, the Tajima's *D* values are not significantly different between any of the different classes examined and they overlap in their 95% confidence interval (CI). Tajima's *D* is sensitive to demographic events, and the European population is known to have undergone a recent bottleneck. Because of this, we expect that Tajima's *D* will skew toward positive values (Glinka *et al.* 2003; Ometto *et al.* 2005). This pattern is observed in our data in Europe, where most of the Tajima's *D* values are positive (Figure 18B). The Tajima's *D* values for the autosomal genes indicate the prevalence of purifying selection at non-synonymous sites (Tajima's $D < 0$, considering all the data) (Figure 18B). However, on the X chromosome the upstream and the non-synonymous sites had higher values of Tajima's *D* than synonymous sites (Figure 18B).

We also examined the complete frequency distribution of the polymorphisms at the different classes of sites. This distribution is pretty similar when all genes or when only differentially expressed genes are included. The same is true when all genes are included or when only autosomal genes are included. In the African population, the frequency spectrum combining all the loci did not show any difference between synonymous sites, introns, and upstream sites (Figure 19A). Significant differences were found only between synonymous and non-synonymous sites for all genes or just the autosomal genes, with an excess of low frequency polymorphisms at the nonsynonymous sites (respectively $P = 0.027$ and $P = 0.016$, Figure 19A). This result confirms the prevalence of purifying selection at non-synonymous sites from the autosomal genes already found by the Tajima's *D* analysis. We find that, for any given class of sites sites, there is an excess of low frequency variants (more than 20% of the sites) (Figure 19A). It is likely that this pattern is due to a recent population size expansion in the African population (Glinka *et al.* 2003; Li and Stephan 2006). In Europe, the frequency spectrum showed a significant excess of low polymorphism in the non-synonymous genes for all autosomal genes (Figure 19B). This result suggests the importance of purifying selection at non-synonymous sites. However, according to the distribution that we observe, it seems that intermediate polymorphism is more common in synonymous, upstream and intron sites. This pattern could be explained by the recent bottleneck that occurred in the European population (Ometto *et al.* 2005).

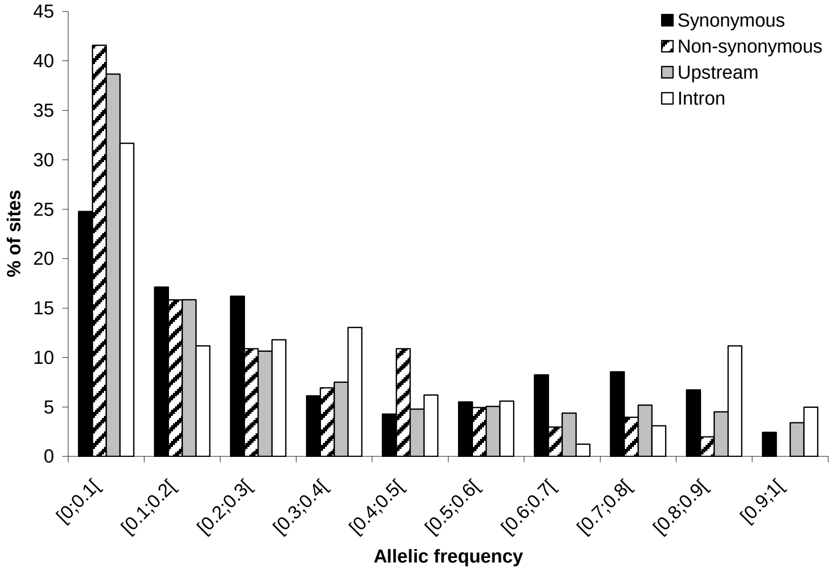
A) In the African population

- All genes included



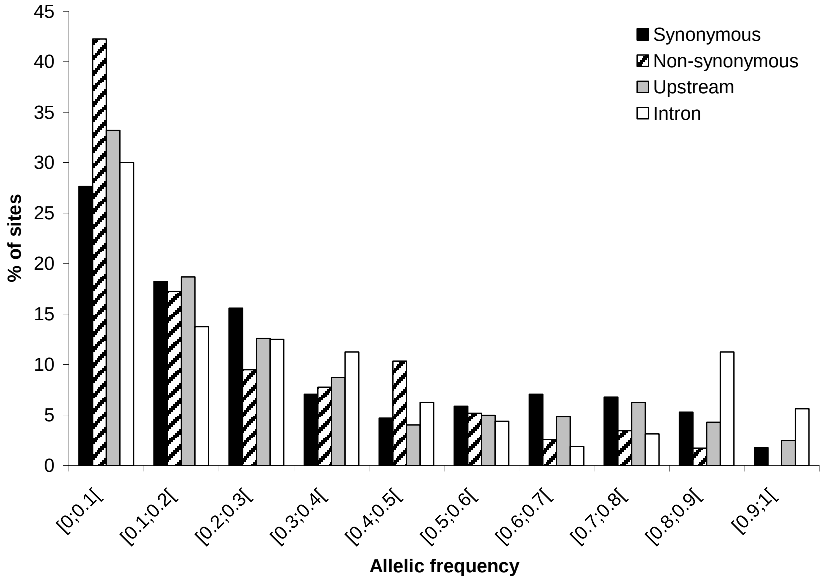
‡ Fisher's exact test for non-synonymous, upstream, intron sites: *P* = 0.027, *P* = 0.292, *P* = 0.877

- All autosomal genes included



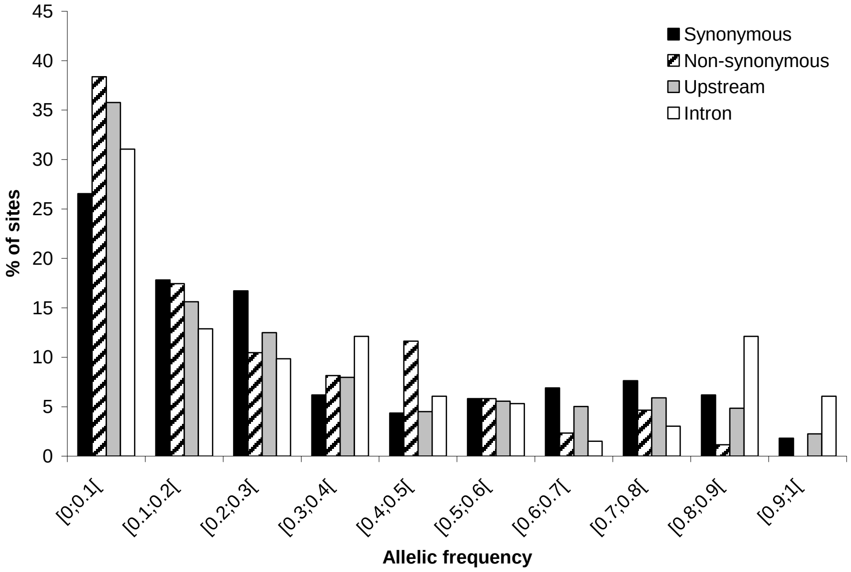
‡ Fisher's exact test for non-synonymous, upstream, intron sites: *P* = 0.016, *P* = 0.066, *P* = 0.347

- All significantly differentially expressed genes between the two populations



‡ Fisher’s exact test for respectively non-synonymous, upstream, intron sites: *P* = 0.0535, *P* = 0.539, *P* = 0.876

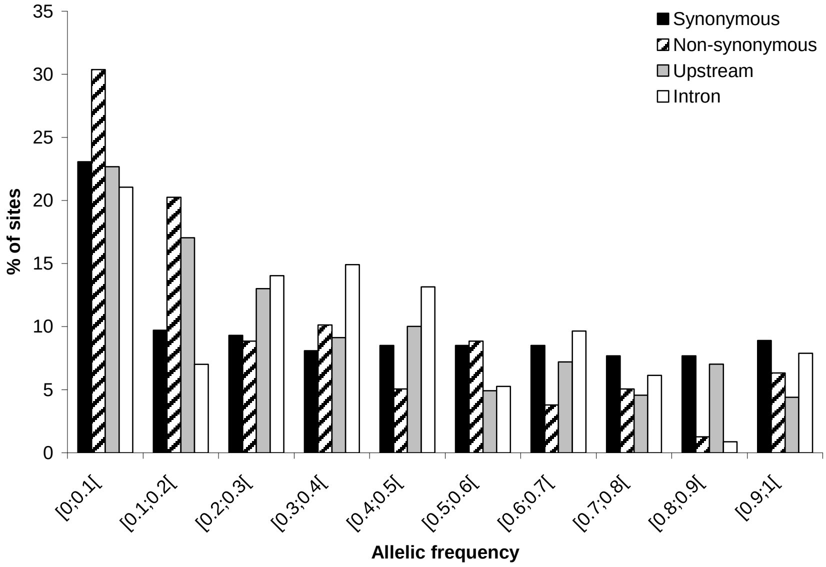
- All autosomal significantly differentially expressed genes between the two populations



‡ Fisher’s exact test for non-synonymous, upstream, intron sites: *P* = 0.095, *P* = 0.169, *P* = 0.531

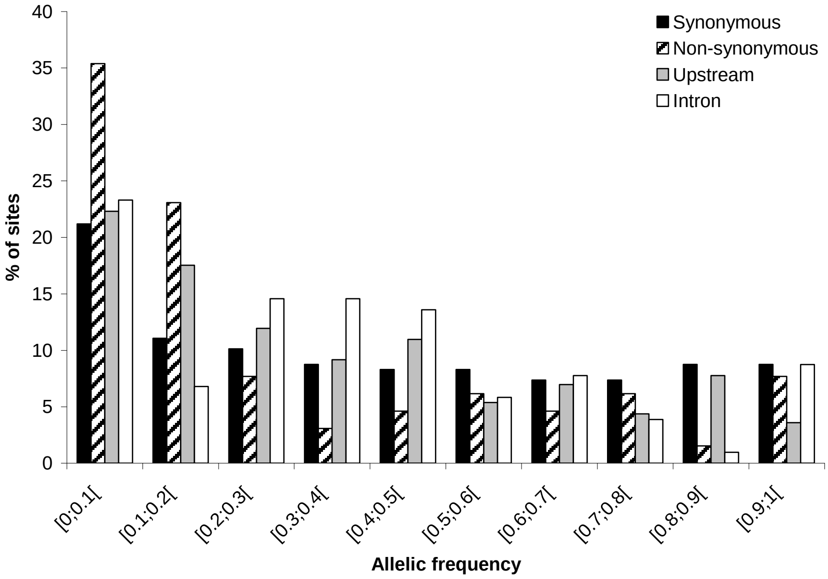
B) In the European population

- All genes included



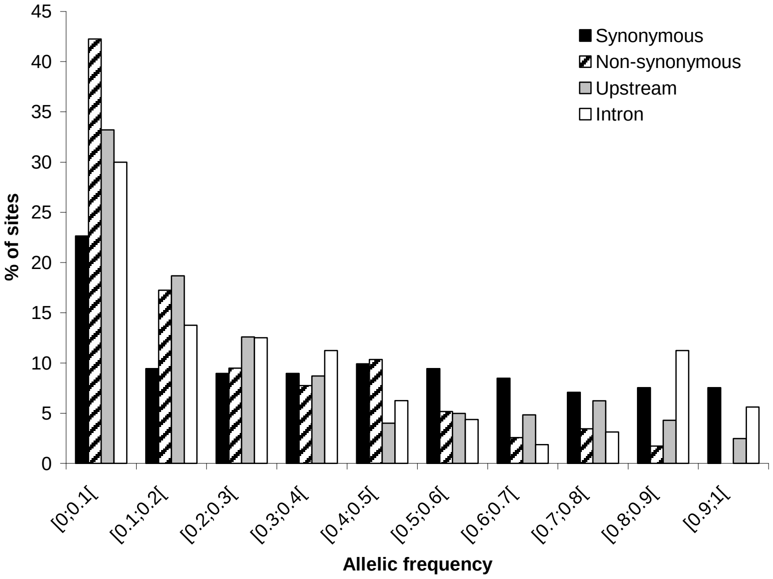
‡ Fisher's exact test for non-synonymous, upstream, intron sites: *P* = 0.266, *P* = 1, *P* = 0.864

- All autosomal genes included



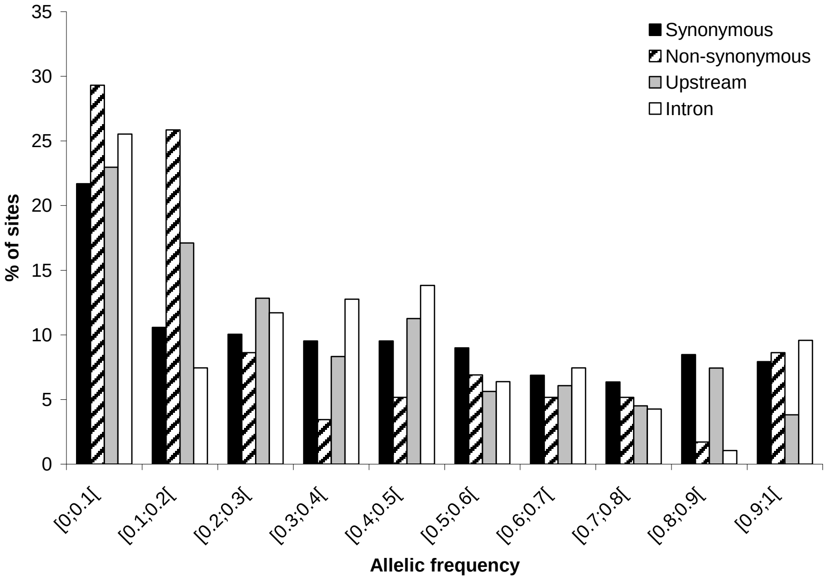
‡ Fisher's exact test for non-synonymous, upstream, intron sites: *P* = 0.040, *P* = 1, *P* = 0.864

- All significantly differentially expressed genes



‡ Fisher’s exact test for non-synonymous, upstream, intron sites: *P* = 0.743, *P* = 1, *P* = 0.864

- All autosomal significantly differentially expressed genes between the two populations



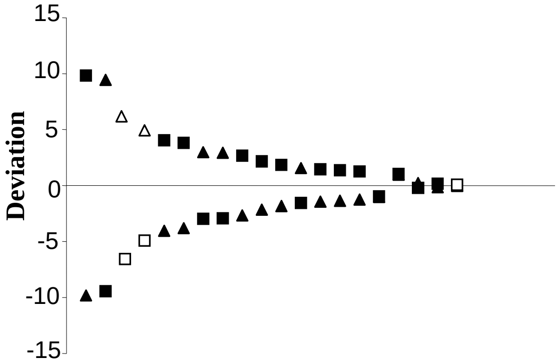
‡ Fisher’s exact test for non-synonymous, upstream, intron sites: *P* = 0.330, *P* = 1, *P* = 0.620

Figure 19: Frequency Spectra.
‡ Fisher ‘s exact test comparing the number of low frequency polymorphisms (between 0 and 0.1 allelic frequency) to the rest of the polymorphisms for the synonymous sites to the non-synonymous, upstream and intron sites.
P-value is from two-tailed Fisher’s exact test

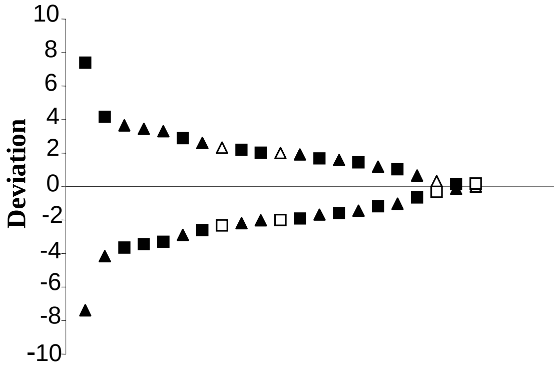
However the interpretation of the Tajima's *D* analysis is difficult because of its sensitivity to demographic events (Tajima 1989). There were slightly more negative Tajima's *D* values in Africa than in Europe. The African population is thought to have experienced an expansion that resulted in an excess of rare variants, while the European population has experienced a bottleneck, which resulted in more intermediate frequency variants. From the Tajima's *D* values, the differences observed between the autosomes and the X chromosome seem to be due to the reduction of the population size in Europe after the bottleneck. The reduction of population size affected the X chromosome more than the autosomes, which have the a more similar *Ne* in the two populations (Hutter *et al.* 2007).

To further determine whether the observed polymorphism and divergence deviates from the neutral equilibrium model, we used a multi-locus version of the HKA test (Hudson *et al.* 1987). This test compares the ratio of polymorphism to divergence at several loci. Under neutrality, these ratios are expected to be equal across loci. For all the autosomal genes from the different classes and from both populations, no significant departure from neutrality was observed (Figure 20). The small sample size could explain the lack of power to reject the neutral model with the HKA test.

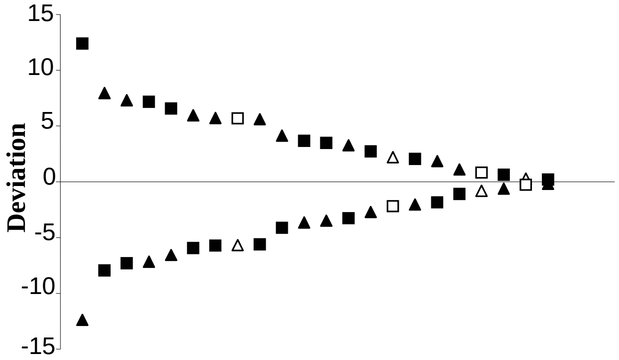
- Intron sites in the African population (ns for HKA test)



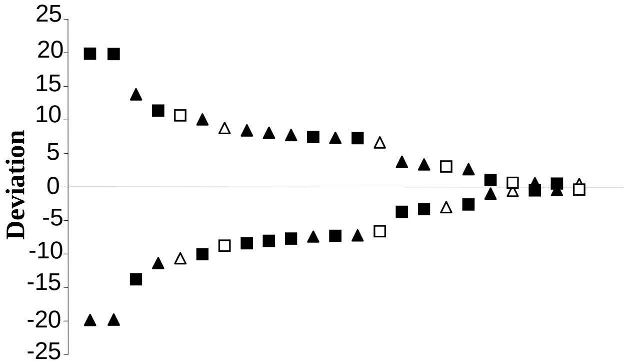
- Non-Synonymous sites in the African population (ns for HKA test)



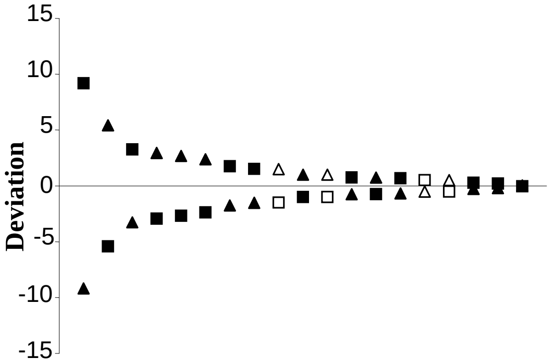
- Synonymous sites in the African population (ns for HKA test)



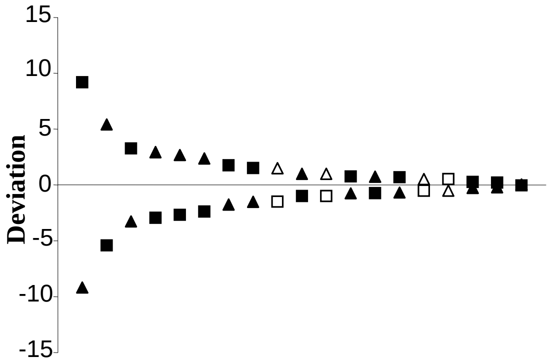
- Upstream sites in the African population (ns for HKA test)



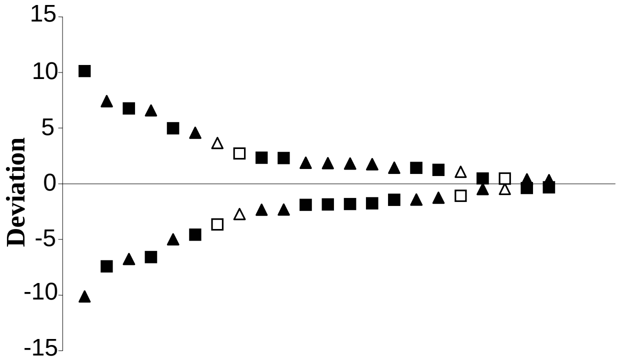
- Intron sites in the European population (ns for HKA test)



- Non-Synonymous sites in the European population (ns for HKA test)



- Synonymous sites in the European population (ns for HKA test)



- Upstream sites in the European population (ns for HKA test)

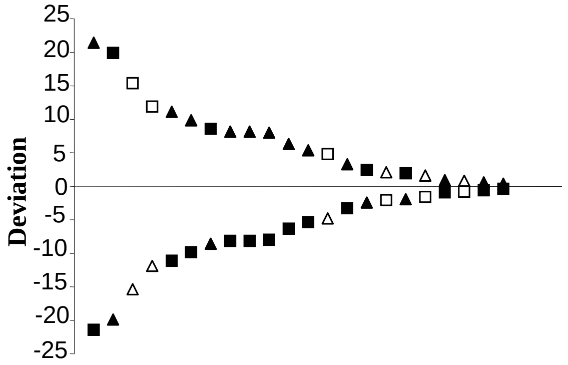
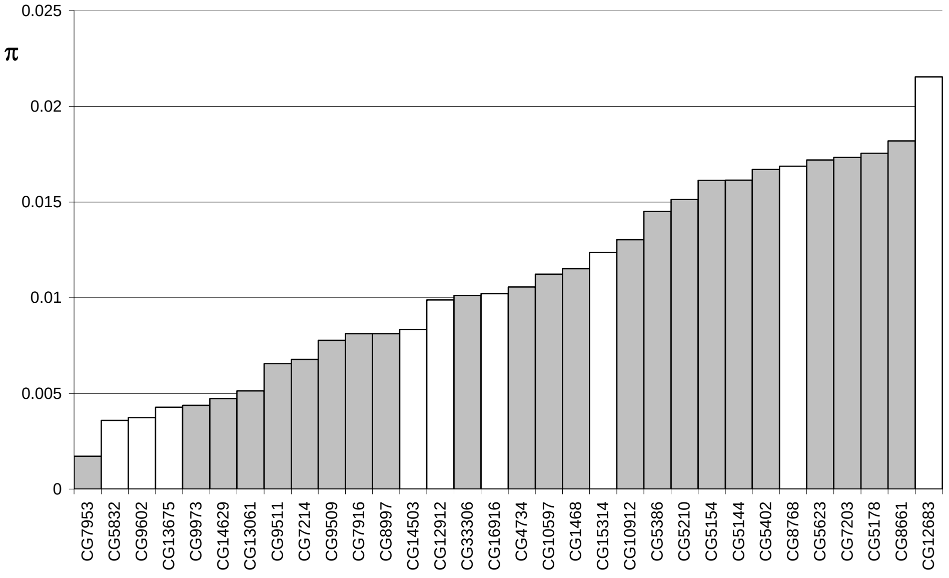


Figure 20: Multi-locus HKA test for autosomal genes.
Genes with no polymorphism or low recombination rate (< 10^{-9} Rec/bp) were excluded. Triangles indicate polymorphism and squares indicate divergence. Solid points represent genes that show a significant difference in gene expression between the populations, while and open points represent genes that do not differ in expression. Value above (below) the x-axis indicate a positive (negative) deviation from the expected value under neutraility.

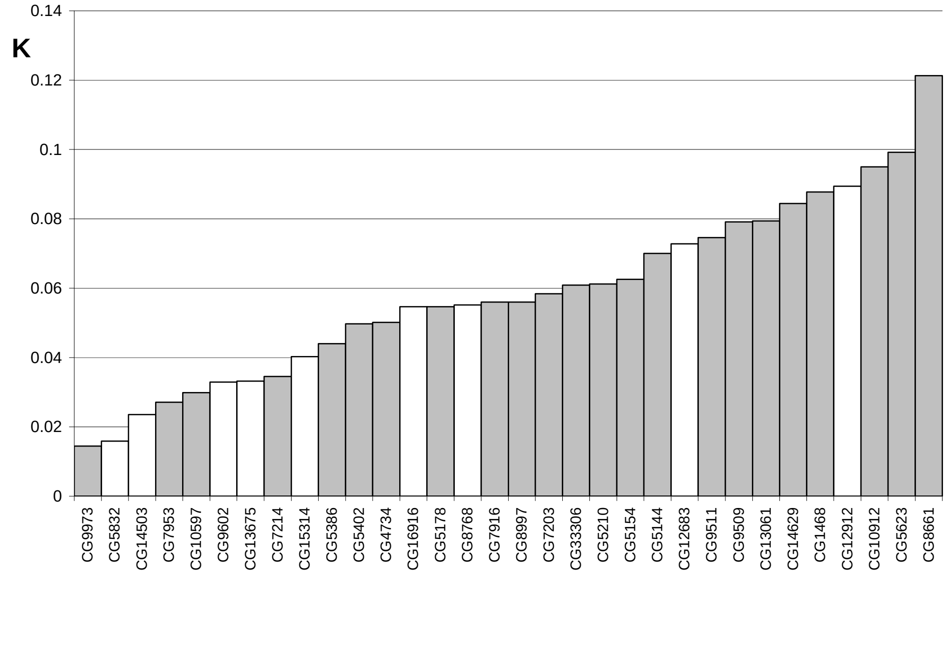
Proximal promoter regions

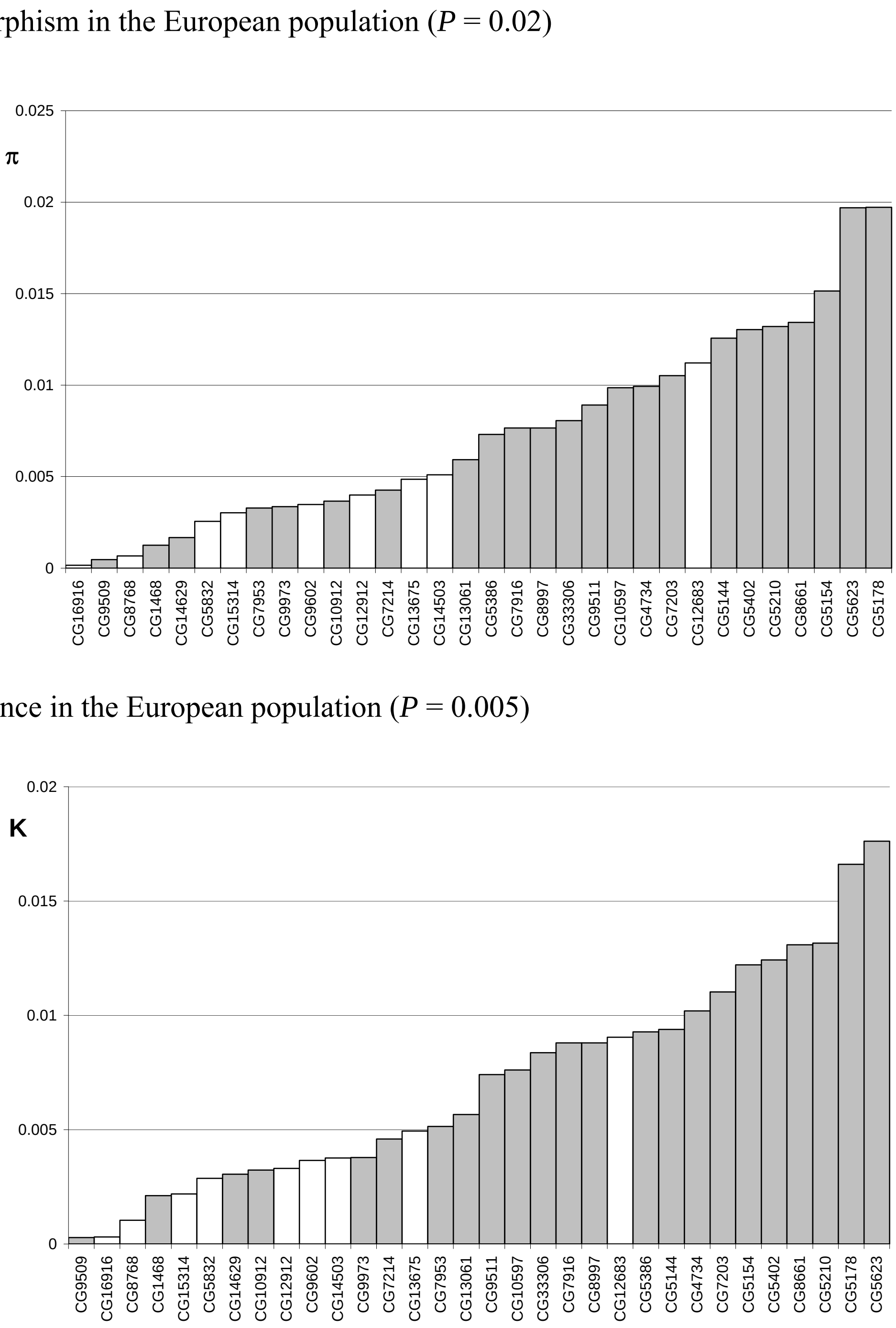
To further investigate the causes of differences in expression between Europe and Africa, I focused on polymorphism and divergence between two groups of genes: genes that are significantly differentially expressed (SG) between the populations and those with no difference in expression (NSG) between the populations (Figure 21). If the expression differences are due to DNA sequence variation in *cis*-regulatory regions, one might expect there to be differences in the type or amount of upstream polymorphism between SG and NSG. In Africa, there is no significant difference between SG and NSG for the amount of polymorphism or divergence (Figure 21A and B). In Europe, both divergence and polymorphism are lower for NSG than for SG (respectively P = 0.02, Figure 21C and P = 0.005, Figure 21 D).

A) Polymorphism in the African population (*P* = 0.463)



B) Divergence in the African population (*P* = 0.082)





C) Polymorphism in the European population ($P = 0.02$)

D) Divergence in the European population ($P = 0.005$)

Figure 21: Rank of polymorphism, π, and divergence, *K* in the upstream regions of all genes.
Solid bars indicate genes that differ significantly in expression between populations (SG) and open bars indicate genes that do not differ significantly in expression between populations (NSG). The Mann-Whitney test was used to compare the rank distribution of SG and NSG.

I also searched for fixed derived mutations in the upstream sites in each population, which may play a causative role in the expression difference between the two populations. To do this, I investigated the polymorphic derived mutations in both the SG and NSG datasets (respectively, Table 12A, B and C). In both datasets, most of the derived mutations are present at low frequencies (less than 10%) in both populations. In other words, these are all singleton mutations. The mutations that are of particular interest are those that are in high frequency in one population, but at low frequency in the other. This is because they are the best candidates to play a role in the difference in expression between Europe and Africa. In Africa, six sites have fixed or almost-fixed derived mutations in the sample (frequencies between 90% and 100%) (Table 12A). Most of these sites are in a gene that is differentially expressed between populations, *CG9509*. One other such site is in the gene C*G7203* (Table 12C), and two additional sites are in *CG14503* and *CG13675* (Table 12B). In total, there were 12 derived polymorphisms that were fixed or almost fixed in Europe, but at low frequency (between 0% and 10%) in Africa. Within the genes differentially expressed between populations (SG), a total of eight sites showed this pattern, five in *CG9509*, two in *CG1468* and one in *CG10912* (Table 12C). In the NSG dataset, there were two sites in *CG8768*, one site in *CG14503,* and one site in *CG16916* (Table 12B).

Table 12: Frequency of derived mutations in the African and European populations.

A) All the genes

Frequency in Africa (%)

	10	20	30	40	50	60	70	80	90	100
100	6	0	0	3	0	1	4	1	3	19
90	3	0	0	0	2	5	3	4	3	24
80	4	6	4	11	2	2	0	0	6	19
70	6	1	6	0	8	2	3	5	7	12
60	8	3	20	4	1	4	4	2	1	7
50	12	4	3	6	3	1	2	7	2	4
40	43	3	8	5	5	6	7	2	3	7
30	63	8	12	3	7	3	15	6	1	4
20	143	11	8	0	11	4	2	3	0	4
10	403	60	25	20	21	4	12	8	3	12
0										

Frequency in Europe (%)

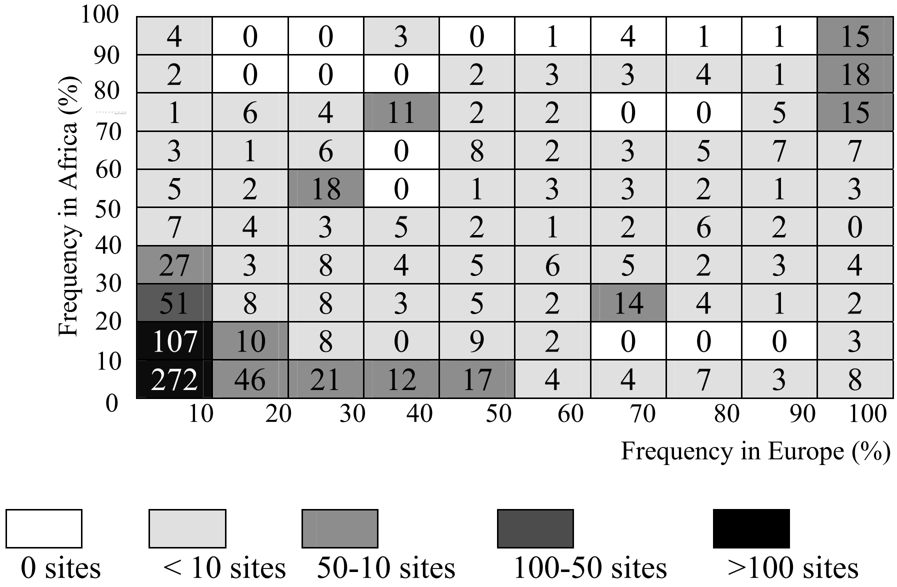
B) All genes with no difference in expression

Frequency in Africa (%)

	10	20	30	40	50	60	70	80	90	100
100	2	0	0	0	0	0	0	0	2	4
90	1	0	0	0	0	2	0	0	2	6
80	3	0	0	0	0	0	0	0	1	4
70	3	0	0	0	0	0	0	0	0	5
60	3	1	2	4	0	1	1	0	0	4
50	5	0	0	1	1	0	0	1	0	4
40	16	0	0	1	0	0	2	0	0	3
30	12	0	4	0	2	1	1	2	0	2
20	36	1	0	0	2	2	2	3	0	1
10	131	14	4	8	4	0	8	1	0	4
0										

Frequency in Europe (%)

C) All genes with difference in expression



10	20	30	40	50	60	70	80	90	100
4	0	0	3	0	1	4	1	1	15
2	0	0	0	2	3	3	4	1	18
1	6	4	11	2	2	0	0	5	15
3	1	6	0	8	2	3	5	7	7
5	2	18	0	1	3	3	2	1	3
7	4	3	5	2	1	2	6	2	0
27	3	8	4	5	6	5	2	3	4
51	8	8	3	5	2	14	4	1	2
107	10	8	0	9	2	0	0	0	3
272	46	21	12	17	4	4	7	3	8

All derived SNPs and INDELs are included.

A cluster of differentially expressed genes on Chromosome arm 2L

Four of the genes that differ in expression between Europe and Africa are clustered together on chromosome arm 2L. These genes are located in the intron of the *CenG1A* gene (Figure 22). In this region, the genes *CG33306*, *CG7953*, *CG8997*, and *CG7916* all have significantly higher expression in Africa than in Europe, with *CG8997* and *CG7916* having the highest relative expression difference (Table 13). Nothing is known about the function of these genes, however they all show an enrichment of expression in the midgut (Chintapalli *et al.* 2007). The organization of these genes suggests that they may share a common upstream regulatory region located between *CG8997* and *CG7916* (Figure 22). The gene *CG7916* showed evidence of recurrent positive selection in its coding, upstream, and intronic regions. The gene *CG8997* also showed evidence of recurrent positive selection in its coding and upstream regions (Table 11A, B and C). To further investigate the expression of these genes, I performed qPCR (quantitative reverse-transcription real-time PCR as in Chapter 2) of *CG7916* and *CG8997*, which confirmed the microarray results; both showed a significant difference in expression between Europe and Africa (Table 14).

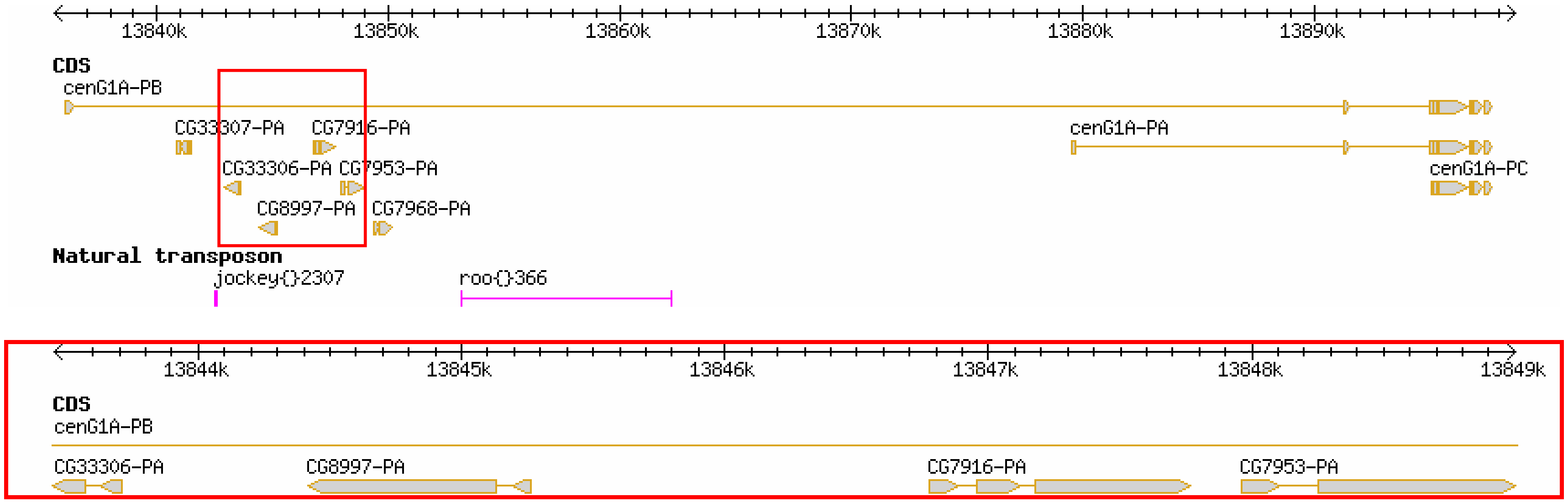


Figure 22: The *cenG1A* region on chromosome arm 2L.
The genes of interest are located between the first and the second exon of the *Centaurin gamma 1A* (*cenG1A*) gene. For each gene, the transcriptional unit is shown. Annotated transposable elements in the sequenced *D. melanogaster* strain are shown in pink. This figure was obtained using the FlyBase Genome Browser 5.0 *(http://flybase.bio.indiana.edu/cgi-bin/gbrowse/dmel/).*

Table 13: Relative levels of gene expression for four genes in the *cenG1A* region.

	Gene name			
Population	CG3306	CG8997	CG7916	CG7953
Africa	1.35	1.67	1.56	1.41
Europe	1	1	1	1

Relative expression levels are from microarray analysis (see Chapter 1). All of the genes are significantly differently expressed between the European and African populations with a *P*-value < 0.002 (two-node analysis).

Table 14: Relative levels of gene expression for *CG7916* and *CG8997* in the European and African populations.

	Europe	Africa	*P*-value*
CG7916	1	1.91	0.02
CG8997	1	2.58	0.01

* Based on Mann-Whitney test comparing European and African strains.

An investigation of their expression in the individual strains revealed that the two genes are expressed at significantly different levels from each other (Wilcoxon test, *P* < 0.0001, Figure 23). On a strain-by-strain basis within populations, there does not appear to be tightly co-regulated expression of the two genes (Figure 23). Landry *et al.* (2005) showed that the expression variation of *CG8997* within species is caused by both *cis*- and *trans*-acting factors. The observation that there is not a tight co-regulation of the genes *CG8997* and *CG7916* may be explained by the influence of some external factors, such as the *trans* factors, that act independently on one of the genes. This hypothesis may be supported by the presence of transposable elements (pink in Figure 22) that can also play a role in gene regulation.

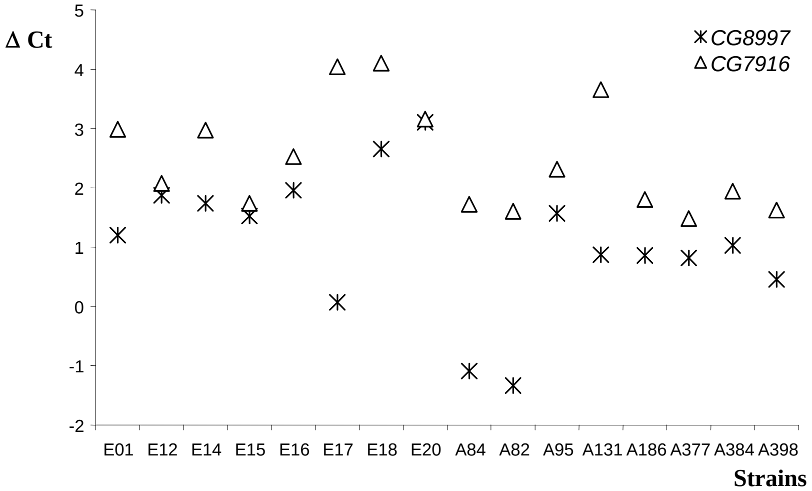


Figure 23: Gene expression levels estimated by qPCR for *CG8997* and *CG7916.* Δ Ct values werecalculated for each strain. The expression level of the two genes is significantly different by the Wilcoxon test (*P* < 0.0001).

The goal of this chapter was to determine the genetic variation associated with gene expression differences between two natural populations of *D. melanogaster*. An original aspect of the study is the use of two groups of genes for which high-quality, population level gene expression data are available. One group of genes differed significantly in expression between two populations (SG), while the other group did not differ in expression and served as a control dataset (NSG). A similar approach was used by Brown and Feder (2005), who compared DNA sequence variation of the proximal promoter (~1 kb upstream of the start codon) of genes that either differed or did not differ in expression between two laboratory strains of *D. melanogaster* . Those authors found no clear differences in the amount or type of polymorphism in the two groups of genes.

In the present study, the candidate genes were differentially expressed between two populations, one from Europe and the other from Africa, composed of eight strains for each population. The analysis of DNA variation did not reveal great differences between the two groups of genes. Thus, it is difficult to relate DNA variation to gene expression variation. Overall, we detected no differences in polymorphism or divergence between the regulatory regions of the differentially-expressed genes and the control genes. A possible explanation for this is that only a small fraction of segregating mutations may affect the level gene. However, previous studies suggests the functional importance of many sites in the non-coding upstream regions of genes (Halligan *et al.* 2004; Andolfatto 2005; Kern and Begun 2005; Haddrill *et al.* 2008). The sequences studied here appear to have been subject to both positive and negative selection. Evidence for negative (purifying) selection was mainly detected in the African population as a skew in the frequency spectrum towards rare variants.

It appears that the evolutionary constraints that we observe are a feature of upstream regions and are not specific to the SG or NSG dataset. Only some of the genes appear to have undergone recurrent positive selection in their upstream regulatory regions. Thus, there is some indication of the importance of *cis*-regulatory elements in the adaptation of *D. melanogaster*, but it is likely that *trans*-regulatory elements are also involved (Wittkopp 2005). The genes with a history of positive selection between *D. melanogaster* and its sibling species may be good candidates to explain local adaptation within *D. melanogaster* populations. Little is known about the function of these genes, however for some of them we may get an idea about their function from their expression pattern or annotation. *CG7916* and *CG7953* are differentially expressed between virgin and mated males (McGraw *et al.* 2004) and may play a role in male reproduction. The gene *CG5178*, also known as *Act88*F, belongs to the Actin gene family. This gene is expressed in the muscles and may also play a role in male reproduction (McGraw *et al.* 2004). The gene *CG5178* shows expression enrichment in adult carcass according to FlyAtlas (Chintapalli *et al.* 2007), but little else is known of its function. Some of the genes show a major enrichment of expression in one or two tissues. *CG9511* is highly expressed in the head and carcass, *CG9602* in testis, *CG10912* in midgut, *CG5402* in male accessory gland, and *CG9509* in Malpighian tubules.

Chapter 4 Evidence for a selective sweep in the regulatory region of a gene that differs in expression between European and African populations of *Drosophila melanogaster*

4.1 Introduction

Understanding the genetic changes that underlie phenotypic adaptation is a major goal of evolutionary genetics. One of the most important factors that can lead to adaptation is environmental changes. An excellent model system for studying such adaptation is *Drosophila melanogaster*, a species that originated in sub-Saharan Africa about 2.5 million years ago and spread around the world by following human migration within the past 10,000–15,000 years (David and Capy 1988). This expansion to new non-African habitats is expected to have been accompanied by adaptation to the new environmental conditions.

Population genetic and genomics methods can be applied to map regions of the genome that are associated with adaptation to the non-African environment (Pavlidis *et al.* 2008). At the DNA level, positive selection leads to the fixation of a beneficial allele, thereby reducing polymorphism in the surrounding, linked genomic region (Maynard Smith and Haigh 1974). This process is known as genetic hitchhiking or a selective sweep and can also lead to an excess of high-frequency derived mutations at sites near the target of selection (Fay and Wu 2000). Within the selected region, there will be a valley of reduced polymorphism. This is often considered to be the signature of a selective sweep, and can be used to identify genomic fragments that have experienced recent positive selection. However, expansion to new environments is typically accompanied by population size bottlenecks, which can also lead to a decrease in polymorphism in derived populations. For *D. melanogaster*, it is clear that the out-of-Africa expansion was accompanied by such a bottleneck (Begun and Aquadro 1993; Glinka *et al.* 2003; Baudry *et al.* 2004; Haddrill *et al.* 2005; Li and Stephan 2006). Because a bottleneck may eliminate variation at particular loci or alter the frequency spectrum of polymorphisms (Tajima 1989; Depaulis *et al.* 2003), it can potentially mimic the predictions of the selective sweep model and lead to the false inference of positive selection. Thus, it is important to account for the demographic history of a population when searching for the targets of selective sweeps.

DNA sequence polymorphism in gene regulatory regions may be a major source of phenotypic variation and a target of natural selection. However, most of the gene expression variation present in natural populations appears to be deleterious and kept at low frequency in the population by negative selection. In Chapter 1 and Hutter *et al.* (2008), a relatively low fraction of genes (3%) were found to show high expression divergence between European and African populations, but low gene expression variation within each population. Within populations, there appears to be strong stabilizing selection on gene expression levels. The prevalence of stabilizing selection on gene expression variation has also been shown in mutation accumulation experiments (Denver *et al.* 2005; Rifkin *et al.* 2005). Since there appears to be strong stabilizing selection on gene expression levels, genes that show large and consistent differences between populations are likely to represent cases where adaptive regulatory evolution has occurred within one or both populations.

One gene that is of particular interest in the above context is *CG9509*. This gene is expressed over twice as high in European strains as in African strains (see Chapter 1, 2 and Hutter *et al.* 2008). A previous microarray survey from Meiklejohn *et al.* (2003) also found a similar 2-fold difference in the expression of *CG9509* between Cosmopolitan (Asian and North American) strains and African strains. Thus, this gene is a good candidate for one that has undergone adaptive regulatory evolution to the non-African environment. It is thought that DNA sequence variation in the *cis*-regulatory regions of genes (primarily in the 5’ upstream region) plays an important role in gene regulatory evolution (Wray *et al.* 2003; Hughes *et al.* 2006). To investigate a possible adaptive role of sequence variation in the *cis*-regulatory region of *CG9509,* we sequenced the coding and upstream regions of this gene, as well as two neighboring genes, *CG14406* and *CG12398*. Of these genes, only *CG9509* shows evidence of expression in adult male flies. *CG9509* is also the only one of these genes that shows evidence of recurrent positive selection in its coding and upstream region since the split of *D. melanogaster* and *D. sechellia*.

Our polymorphism survey uncovered a region of around 1.2 kb that shows no polymorphism in the European population sample, but has several fixed (or nearly fixed) differences between the European and the African populations. The region that is monomorphic in Europe includes the 5’ upstream region of *CG9509,* which may be responsible for the observed gene expression difference between the two populations. Several statistical tests suggest that a selective sweep has occurred in this region in the European population.

4.2 Materials and Methods

PCR amplification and DNA sequencing

We amplified a region between the nucleotides 14,800,704 and 14,806,152 (release 5.0; http://flybase.org) on the X chromosome, which included the second exon of the gene *CG12398*, the intergenic region between *CG12398* and *CG14406*, the entire coding region of *CG14406*, the intergenic region between *CG14406* and *CG9509*, and the entire coding region of *CG9509* (Figure 24). DNA sequencing was performed using twenty-four highly-inbred *D. melanogaster* strains from two populations: 12 from a European population (Leiden, The Netherlands) and 12 from an African population (Lake Kariba, Zimbabwe). The European strains were kindly provided by A. J. Davis, and the African strains by C. F. Aquadro. The primers for sequencing and PCR amplification were designed with the software Pimer3 (Rozen and Skaletsky 2000) to give fragments with a maximum of ~1000 overlapping base pairs to cover the whole region. The list of primers is available in Appendix B.

PCR was performed using PeqLab reagents (Erlangen, Germany). Afterwards, the PCR products were purified using ExoSAP-IT (USB, Cleveland, USA). Sequencing reactions using ABI BigDye ddNTPs were read by an ABI 3730 sequencer (Applied Biosystems, Foster City, CA, USA). Sequence reads were aligned with Seqman (DNAstar, Madison, WI, USA). Chromatograms were visually inspected and all polymorphic sites were visually confirmed. Final alignments were made using the CLUSTALW algorithm as implemented in BioEdit v7.0.9 (Thompson *et al.* 1994) and manually corrected. For the interspecific comparison of *Drosophila sechellia,* the published genomic DNA sequence was used (*Drosophila* 12 Genomes Consortium 2007). The sequence was downloaded from the UCSC Genome Browser (*http://www.genome.ucsc.edu/*) (Hinrichs *et al.* 2006). All protocol details are available in Appendix A.

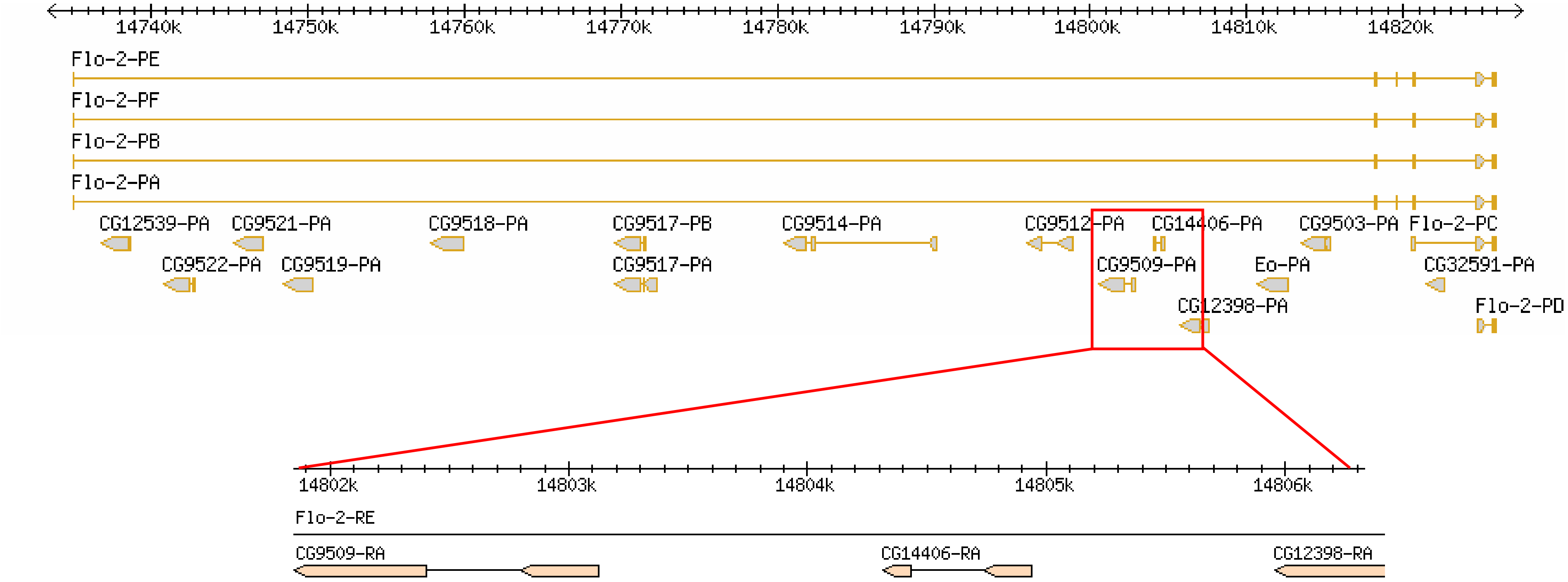


Figure 24: Map of the region of interest. The genes are located on the X chromosome within the second intron of the *Flo-2* gene.
The location of the genes (nucleotide coordinates) according to Genome release 5.0 are shown at the top. An enlargement of the sequenced region and the transcriptional units contained within are shown below. This figure was modified from the FlyBase Genome Browser (http://flybase.org/cgi-bin/gbrowse/dmel/).

Data analysis

Standard polymorphism and divergence analyses were performed using DnaSP 4.50.3 (Rozas *et al.* 2003). For each locus, several different sets of sites were considered: upstream sites (5'), which include 5' UTR and intergenic regions, synonymous sites (S), non-synonymous sites (NS), and intron sites (I). We assumed that synonymous sites evolve neutrally and used them as the reference for comparisons of other sites. Within each set, sites were separated into those which are polymorphic within *D. melanogaster* (P) and those that show a fixed difference between *D. melanogaster* and the outgroup (D). The MK test (McDonald and Kreitman 1991) was performed to compare fixed differences between *D. melanogaster* and *D. sechellia* to polymorphisms within *D. melanogaster*. To determine the relative contribution of the theses two parameters to the MK test, we calculated the proportion of divergence driven by positive selection (α) (Smith and Eyre-Walker 2002).

$$\alpha = 1 - \frac{DsPx}{DxPs}$$

Here the synonymous sites are denoted as (s) and non-synonymous sites, upstream sites and intron sites as (x).

The nucleotide diversity was estimated by the statistics θ (Watterson 1975) and π (Tajima 1989), which allow for a test of neutral equilibrium using Tajima's *D* (Tajima 1989).

Likelihood analysis and sweep localization

We computed the likelihood of a selective sweep model versus the neutral model for the polymorphism data using a composite maximum likelihood ratio (CLR) test (Kim and Stephan 2002). This method allows for the detection and localization of the selective sweep. In this test, the maximum likelihood of observing derived variants at a polymorphic site under the selective sweep model (H_A) is compared to what is expected under the standard neutral model (H_0) and a statistic Λ_{CLR} is calculated (Pavlidis *et al.* 2008). Two neutral models were considered in order to estimate the null distribution of $\Lambda_{CLR.}$ The first one is derived from the standard neutral model with the population at equilibrium, whereas the second model includes population size changes consistent with the inferred demographic history of *D.*

melanogaster (Li and Stephan 2006). Because the use of the CLR test with a demographic model can give high false-positive rates, we also applied the corrective methods proposed by Thornton and Jensen (2007).

If a selective sweep is inferred, the CLR method estimates the location of the advantageous mutation (*X*) and the strength of selection ($\alpha = 1.5\ N_e\, s$, where N_e is the effective population size and *s* the selection coefficient). The recombination rate (*r*) for the region considered was assumed to be $0.48\text{x}10^{-8}$ (per site per generation) according to the computer program "Recomb-rate" (Comeron *et al.* 1999). The null hypothesis of the test is based on the standard neutral model. In case the null hypothesis was rejected by the CLR test, we analyzed the same data using a goodness-of-fit (GOF) test (Jensen *et al.* 2005). This test compares the observed data to a selective sweep model.

Another CLR test, *SweepFinder* (SF), proposed by Nielsen *et al.* (2005) was also applied to the polymorphism data. As with the test of Kim and Stephan (2002), *SweepFinder* uses the CLR approach to distinguish between a neutral model and a selective model. If a selective sweep is inferred, the SF method estimates the location of the advantageous mutation (*X*) and the strength of selection ($\alpha = r/s\ln(2N_e)$, where N_e is the effective population size, *s* the selection coefficient and *r* is the recombination rate mentioned above) (Nielsen *et al.* 2005).

The likelihood ratios were compared to a cumulative frequency distribution of likelihood ratios obtained from 10,000 simulations using the program *ms* (Hudson 2002).

4.3 Results

Candidate region

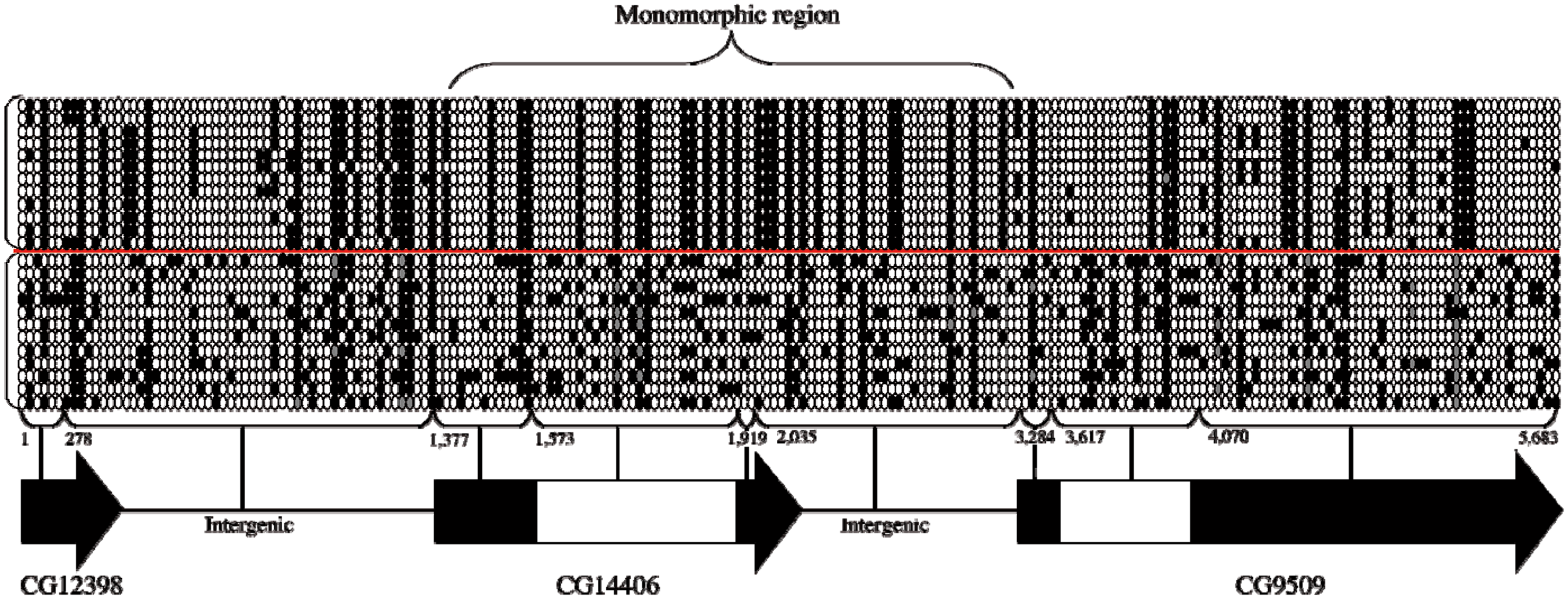
Our previous microarray survey of the European and African *D. melanogaster* populations (Chapter 1; Hutter *et al.* 2008) revealed that the gene *CG9505* has significantly higher expression in European flies than in African flies (Table 15). For males, the difference in expression between the two populations was greater than two-fold in both microarray and qPCR (quantitative real-time PCR) experiments (2.31-fold and 2.02-fold, respectively; Table 15). For females assayed by qPCR, the difference was 1.69-fold. (Table15). The previous microarray survey of Meiklejohn *et al.* (2003) showed a similar difference in expression between Cosmopolitan strains (from Asia and North America) and African strains (more than 2-fold, Table 15).

Table 15: Gene expression data for the *CG9509* gene.

	Populations		*P-value*
	Cosmopolitan	African	
Meiklejohn *et al.* 2003	2.97	1.11	0.001
	European	African	
Microarray (Chapter 1)	2.31	1	0.0001
qPCR (Chapter 2) male	2.02	1	0.003
qPCR (Chapter 2) female	1.69	1	0.02

CG14406 and *CG12398* were not detected as expressed in the microarray analyses

All of the above differences in expression were highly significant ($P \leq 0.02$). *CG9509* is located on the X chromosome within the second intron of the *Flo-2* gene (Figure 24) and is predicted to have several different functions, including mesoderm development and alcohol metabolism. However, the exact function of *CG9509* is unknown. The gene encodes a protein possibly involved in binding flavin-adenine dinucleotide (FAD) in choline deshydrogenase activity (inferred from structural similarity). According to the FlyAtlas database, *CG9509* is expressed mainly in the Malpighian tubules (Chintapalli *et al.* 2007). In the upstream region of this gene, a previous genome scan from Lino Ometto's thesis (Ometto *et al.* 2005) identified a valley of reduced nucleotide variation in the European population.



○ Ancestral SNP

◍ Derived SNP with low frequency

● Derived SNP with high frequency

Figure 25: Graphical representation of nucleotide polymorphism in the *CG9509* region. Each column represents a polymorphic site and each row a different strain. The twelve rows above the red line represent the European strains, and the twelve rows below represent the African strains. For each polymorphic site, the ancestral state is indicated by an open circle and the derived state by a solid circle. The location of the three transcriptional units contained within the region are shown below, with solid boxes indicating exons and the open boxes indicating introns. The arrowheads indicate the direction of transcription. Numbers indicate the relative nucleotide coordinates of the first position of each region.

To determine the boundaries of this low polymorphism region in Europe, we sequenced a region encompassing *CG9509* and two of its neighboring upstream genes, *CG14406* and *CG12398* (Figure 24 and 25) in both European and African population samples. *CG14406* and *CG12398* showed no evidence of expression in the microarray analysis of Hutter *et al*. (2008).The gene *CG14406* was predicted computationally, but it has no annotated function. It also shows no enrichment of expression in any of the tissues surveyed by Chintapalli *et al*. (2007). *CG12398* shows a weak signal of expression in the ovary (Chintapalli *et al.* 2007) and appears to be involved in chorion formation (Fakhouri *et al.* 2006).

Nucleotide variation in the region of interest

The entire sequenced region encompasses a total of 5.6 kb. Within this, there is a monomorphic region (a region without polymorphism) of about 1.2 kb in the European population (Figure 25). This monomorphic region includes the entire *CG14406* gene, the intergenic region between *CG14406* and *CG9509*, and the first exon of *CG9509* (Figure 25). Polymorphism in the African population is consistently high across the entire region (Figure 26). Furthermore, the divergence between *D. melanogaster* and *D. sechellia*, a closely related species, is also relatively high (7%) for the entire region. In the monomorphic region, the divergence value is also relatively high (> 6%, Figure 26). These observations indicate that the reduced polymorphism seen in the European population is not due to the region having an overall low mutation rate.

The McDonald-Kreitman (MK) test (McDonald and Kreitman 1991), coupled with the calculation of α (Smith and Eyre-Walker 2002), was performed to test for selection in the coding and non-coding regions of the above genes. Under neutrality, the ratio of polymorphisms to fixed differences should be identical for synonymous sites and the other classes of sites being tested (non-synonymous, upstream, and intron sites). An excess of polymorphism relative to divergence at the tested sites can be interpreted as a signal of balancing selection (or weak purifying selection), while the opposite pattern (a relative excess of divergence at the tested sites) is indicative of positive selection. Positive selection will produce positive α values, with greater values indicating a greater signal of positive selection. Negative α values provide evidence for balancing (or weak purifying) selection. We detected evidence of positive selection for the *CG9509* coding region in the African population (α = 0.84 and highly significant MK test (P < 0.001), Table 16). In Europe, the test is non-significant for both *CG9509* and *CG14406*.

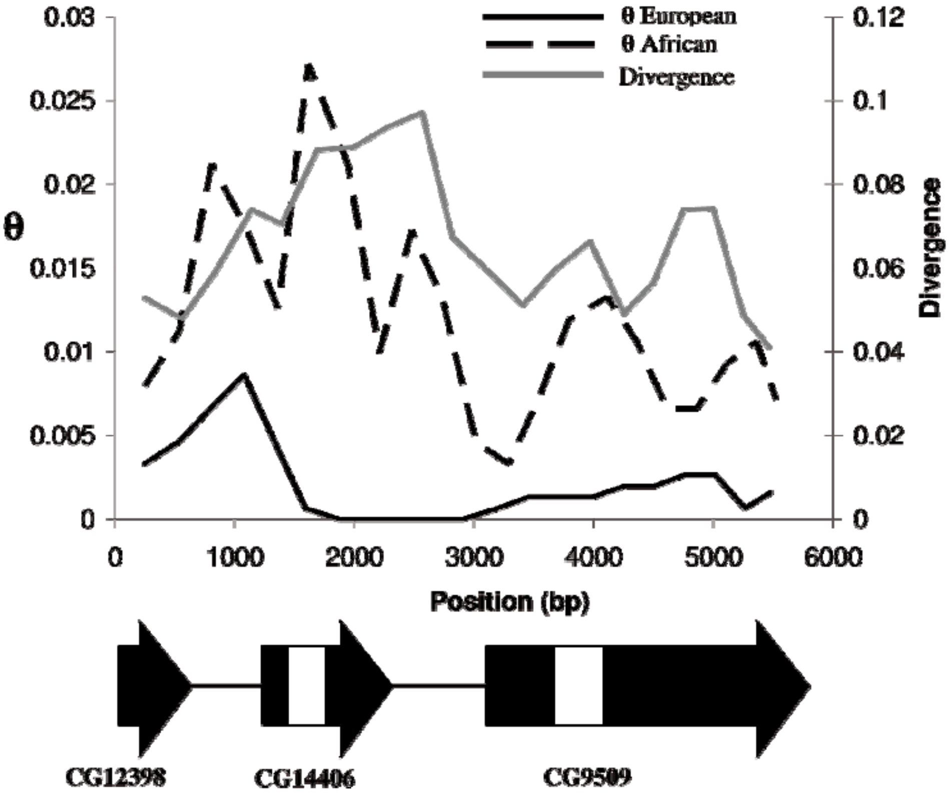


Figure 26: Polymorphism and Divergence in the *CG9509* region.
Nucleotide variation θ in the European (solid line) and African (dashed lines) populations. The average divergence to *D. sechellia* is indicated by the shaded line. The location of the three transcriptional units contained within the region are shown below, with solid boxes indicating exons and open boxes indicating introns. The arrowhead indicates the direction of transcription.

Table 16: McDonald-Kreitman test for the *CG9509* and *CG14406* genes.

Gene Name	D_S	P_S	D_N	P_N	α_{coding}	$D_{5'}$	$P_{5'}$	α_{interG}	D_I	P_I	α_{intron}
European population											
CG9509	44	7	51	3	0.63	84	1	0.93**	28	1	0.78
CG14406	8	1	7	0	1.00	53	20	-2.02*	42	0	1.00
African population											
CG9509	40	40	49	8	0.84***	76	33	0.57**	25	16	0.36
CG14406	5	8	7	3	0.73	45	43	0.40	21	31	0.08

Ps, Pn, $P_{5'}$, P_I, Ds, Dn, $D_{5'}$, D_I are polymorphism and divergence for synonymous, non-synonymous, intergenic and intron sites. α, divergence driven by positive selection (Smith and Eyre-Walker, 2002); Asterisks indicate significant p-value for Fisher exact test comparing divergence and polymorphism of synonymous sites to non-synonymous sites, integenic (interG) sites and intron sites: "*"=p<0.05, "**"=p<0.01, "***"=p<0.001

The intergenic region between *CG9509* and *CG14406* shows significant evidence of positive selection in both the European ($\alpha = 0.93$) and African samples ($\alpha = 0.57$) (Table 16).
In Europe, the intergenic region between *CG14406* and *CG12398* shows significantly more polymorphism than divergence in comparison to the synonymous sites (Table 16).

We further investigated the above departures from neutrality by examining the frequency of segregating polymorphisms with Tajima's *D* statistic (Tajima 1989). The Tajima's *D* (TD) values tend to be positive in the European population for *CG9509* and *CG14406* for all classes of sites, except for the *CG9509* intron (Table 17A). A positive TD value indicates the presence of intermediate frequency variants, while a negative TD indicates an excess of rare variants. In Africa, TD tends to be negative for *CG9509* at the different classes of sites and also for the non-coding region of *CG14406*. For the entire region, Fay and Wu's *H* (2000) indicates the presence of high frequency derived variants in the European population that are located on both sides of the monomorphic region (Figure 27). These results in negative *H* values are not significantly different from zero.

Table 17: Tajima's D values for synonymous, non-synonymous, intergenic and intron sites: TDs, TDn, $TD_{5'}$, TD_{I}.

A) European

Gene Name	TD_S	TD_N	$TD_{5'}$	TD_I
CG9509	0.65	1.85	1.49	-0.85
CG14406	1.41	1.65	0.68	NA

B) African

Gene Name	TD_S	TD_N	$TD_{5'}$	TD_I
CG9509	-0.36	-1.57	-1.14	-0.11
CG14406	0.39	0.78	-0.37	-0.24

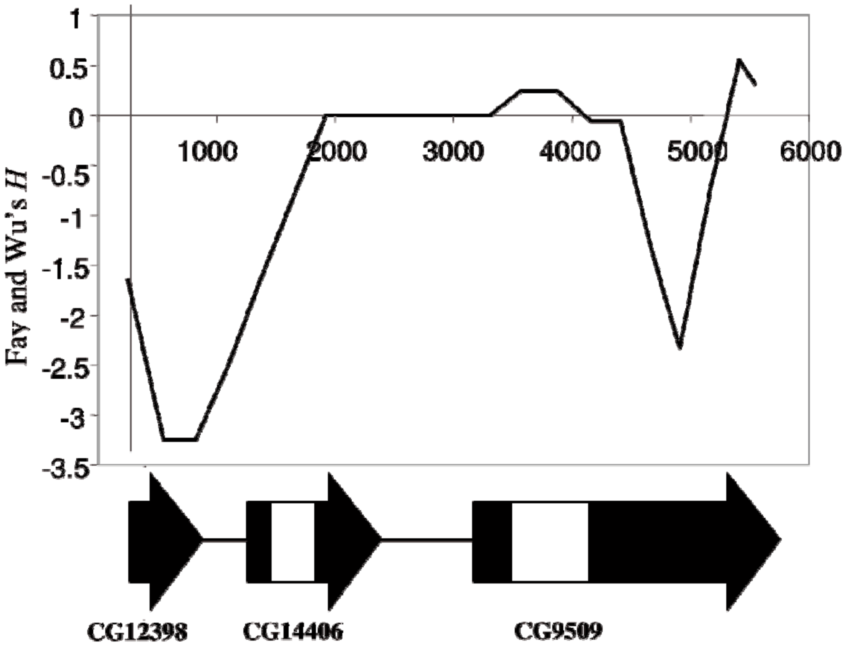


Figure 27: Fay and Wu's *H* in the European population.
The test was non-significant with the conservative assumption of no recombination. The location of the three transcriptional units contained within the region are shown below, with solid boxes indicating exons and open boxes indicating introns. The arrowhead indicates the direction of transcription.

Likelihood of a selective sweep

The low polymorphism observed in the European population may result from the bottleneck that accompanied the colonization of Europe, or it may result from a selective sweep driven by adaptation to the new environment. To try to distinguish between these two possibilities, we performed several sequence-based tests for selective sweeps.

First, we determined if the reduction in variation observed in the European population departed the expectations of a neutral equilibrium model. To do this, we applied the composite likelihood ratio (CLR) test of Kim and Stephan (2002). We performed the test on the European and African sequences independently using a standard neutral model as the null hypothesis and using sequences from non-coding regions only. For the test we used θ values of 0.0046 for Europe and 0.0131 for Africa, which correspond to the estimates of Glinka *et al.* (2003) for the entire non-coding region of the X chromosome (Table 18). We compared the likelihood ratio of the observed data to that obtained from 10,000 neutral coalescence simulations. The probability of neutral evolution producing the reduction in polymorphism seen in Europe is very low ($P = 0.0001$), with $\Lambda_{\mathrm{CLR}} = 11.225$ (Table 18). We also performed another test, *SweepFinder*, that was proposed by Nielsen *et al.* (2005). This test also uses a composite likelihood approach and is similar to the Kim and Stephan test (2002). However, SF differs from the Kim and Stephan (2002) approach in that the null hypothesis is derived from the background pattern of variation observed in the data and not based on a specific evolutionary model as the previous test, which assumed a neutral equilibrium model. Despite this difference, the two tests give similar results, with *SweepFinder* also providing strong evidence for a departure from neutral evolution (Λ_{SF} = 16.34, *P* = 0.0002, Table 18). We conclude that the European polymorphism pattern cannot be explained by a simple model of neutral evolution. In contrast, the African population shows no significant departure from neutrality by either the CLR or the *SweepFinder* test (respectively, *P* = 0.111 and *P* = 0.176, Table 18).

Table 18: Maximum-Likelihood analysis based on Kim and Stephan (2002), Jensen *et al.* (2005) and Nielsen *et al.* (2005).

A) Europe

θ =0.0046, n=12, L=3,746 bp, S = 23			
		Standard neutral model	Demographic scenario
	Parameters	*P-value*	*P-value*
CLR/GOF			
Λ_{CLR}	11.23	0.0001	0.123
α	726.47		
X	2,362		
Λ_{GOF}	245.97	0.521	
SweepFinder			
Λ_{SF}	16.34	0.0002	0.068
α	$4.44\text{x}10^{-4}$		
X	2,194		

B) Africa

Under the standard neutral model

θ = 0.0131, n=12, L=3,776 bp, S = 125		
	Parameters	*P-value*
CLR		
Λ_{CLR}	5.13	0.111
SweepFinder		
Λ_{SF}	3.35	0.176

The tests were performed using both the standard neutral model and demographic scenario (bottleneck) proposed by Li and Stephan (2006). The demographic model was integrated into the CLR distribution following the approach of Thornton and Jensen (2007).
θ is the estimated nucleotide diversity parameter (Watterson 1975).
n indicates the number of sequences
L is the length (in bp) of the non-coding sequence
S indicates the number of segregating sites

Because the CLR test rejected the neutral equilibrium model in the European population, we also applied the goodness-of-fit (GOF) test developed by Jensen *et al.* (2005). With this test, we can determine if the CLR rejection can be explained by a specific evolutionary process. In this case, we test whether or not the observed data can be explained by a selective sweep. The GOF test p-value is 0.521, which indicates that the selective sweep scenario cannot be rejected. We then tested a second scenario, which includes a bottleneck event. For this, the demographic parameters inferred from Li and Stephan (2006) were used to estimate the null distribution of Λ_{CLR}. The *P*-value under the bottleneck scenario was 0.123, which means that we cannot rule out a bottleneck as the cause of the observed reduction in polymorphism (Table 18). A similar non-significant p-value was obtained using *SweepFinder* and including a bottleneck scenario (*P* = 0.068). Thus, in both cases, the statistical tests cannot distinguish between a selective sweep and a population size bottleneck as the cause for the reduced European variation. This is illustrated in Figure 28 in which the scenarios described above are combined to show that we cannot distinguish between a selective sweep selection and a bottleneck.

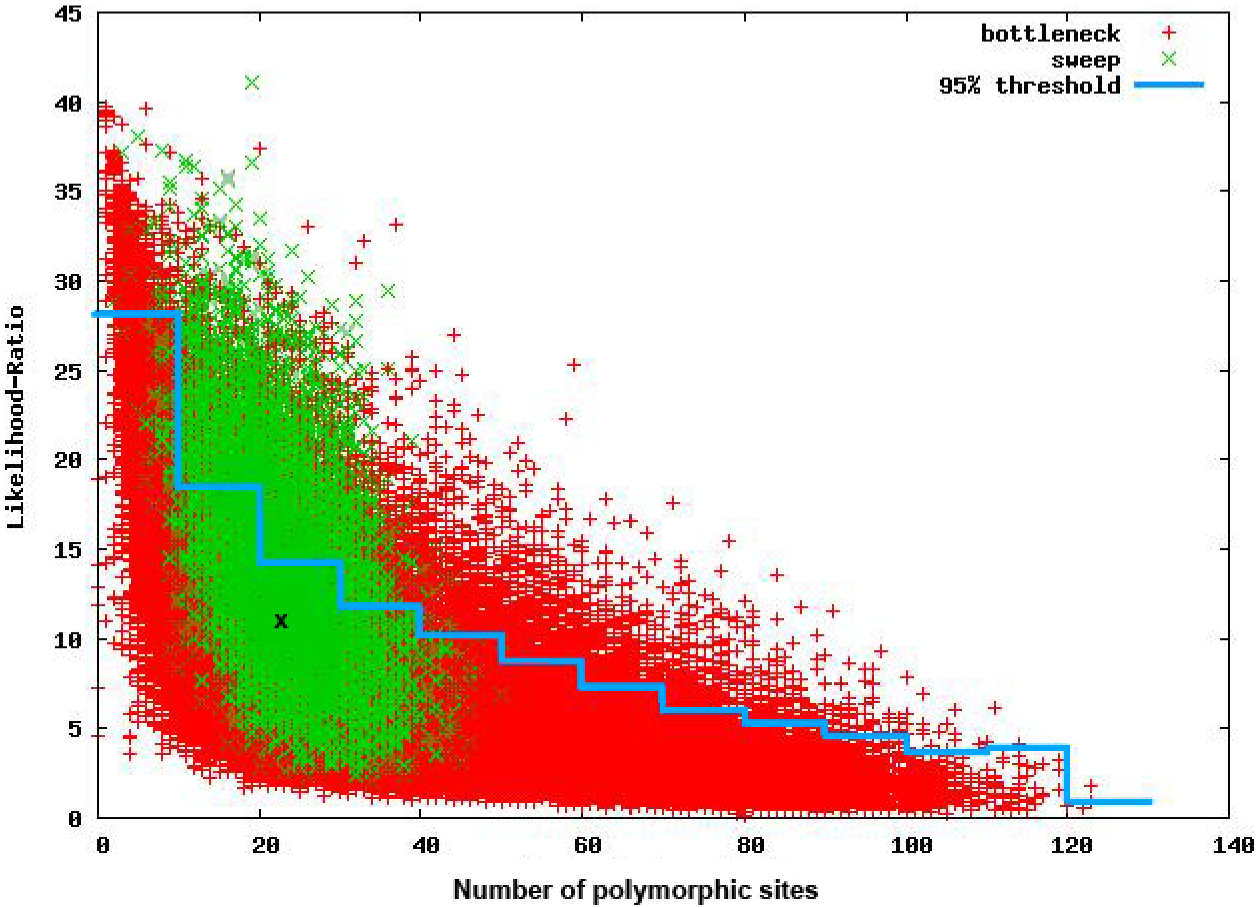


Figure 28: Results of simulations to determine the relative likelihood of selection or demographic scenarios to explain the observed polymorphism in the European population. The black cross indicates the observed value for the European population sample. For each scenario, 10,000 coalescence simulations were performed. Bottleneck parameters are based on Li and Stephan (2006).

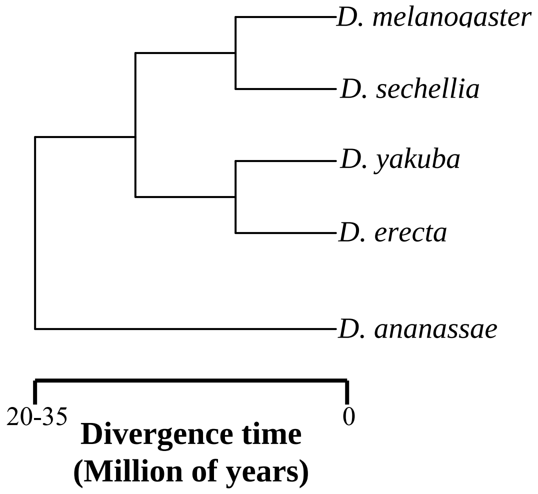
Putative target of selection

In the case that positive selection has played a role in shaping the observed polymorphism in Europe, both *SweepFinder* and the Kim and Stephan test (2002) allow us to estimate the approximate position of the putative selected site. In the European sequences, this site is predicted to be approximately at position 2,362 (CLR) or 2,194 (SF) of the sequenced fragment (Table 18). This corresponds to the intergenic region between *CG14406* and *CG9509*. Since this is the upstream region of *CG9509*, it is likely to contain the regulatory elements controlling *CG9509's* expression. Thus, sequence differences between the European and African populations within this region might explain the observed difference in expression of *CG9509* between the European and African populations (Table 16). To identify such putative regulatory changes, we compared the *D. melanogaster* sequences with those of other *Drosophila* species: *D. sechellia D. yakuba*, *D. erecta* and *D. ananassae* (see the phylogeny in Figure 29). Mutations that occur in well-conserved regions are good candidates for those that function in gene regulation. We identified several derived mutations that were fixed in the European population, but conserved across the other *Drosophila* species (Table 20). These mutations may occur in *cis*-regulatory elements and are a good starting point for further functional analysis.

Table 19: Fixed (or nearly fixed) differences between European and African strains in the upstream region of *CG9509*. Strains starting with “E” are from the European population and those starting with “A”are from the African population. At the bottom are four outgroup species, *D. sech, D. yak, D. ere, D. ana* which correspond to *Drosophila sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*. Bold writing indicates derived mutations. The analysis was performed with Evo Printer (*http://evoprinter.ninds.nih.gov/index.html*) *melanogaster* subgroup.

	Position on the fragment					
	2,112	2,131	2,461-2,465	2,517	2,533	2,563
	Relative position to start codon of CG9509					
Strains	-1,173	-1,153	-824 to -820	-768	-752	-722
E01	**A**	A	**Gap**	**C**	**A**	**G**
E02	**A**	A	**Gap**	**C**	**A**	**G**
E11	**A**	A	**Gap**	**C**	**A**	**G**
E12	**A**	A	**Gap**	**C**	**A**	**G**
E13	**A**	A	**Gap**	**C**	**A**	**G**
E14	**A**	A	**Gap**	**C**	**A**	**G**
E15	**A**	A	**Gap**	**C**	**A**	**G**
E16	**A**	A	**Gap**	**C**	**A**	**G**
E17	**A**	A	**Gap**	**C**	**A**	**G**
E18	**A**	A	**Gap**	**C**	**A**	**G**
E19	**A**	A	**Gap**	**C**	**A**	**G**
E20	**A**	A	**Gap**	**C**	**A**	**G**
A82	C	T	ATATA	G	T	A
A84	C	T	ATATA	G	T	A
A95	C	T	ATATA	G	T	A
A131	**A**	T	ATATA	G	T	A
A186	C	T	ATATA	G	T	A
A145	C	T	ATATA	G	T	A
A157	C	T	ATATA	G	T	A
A186	C	T	ATATA	G	T	A
A191	C	T	ATATA	G	T	A
A229	C	T	ATATA	G	**A**	**G**
A377	C	T	ATATA	G	T	A
A384	C	T	ATATA	G	T	A
A398	C	T	ATATA	**C**	**A**	A
D. sech	C	A	ATATA	G	T	A
D. yak	C	T	ATATA	G	T	A
D. ere	C	T	ATATA	G	T	A
D. ana	-	-	ATATA	G	T	A

Figure 29: Phylogenetic relationship and estimated divergence times of species in the melanogaster subgroup. Data from Da Lage *et al*. (1998)



4.4 Discussion

This chapter focused on a candidate sweep region that encompasses 5.6 kb of the X chromosome and includes the entire *CG14406* and *CG9509* genes, as well as part of the *CG12398* gene. The gene *CG9509* was of particular interest, because it showed a significant difference in expression between African and non-African populations of *D. melanogaster* (Meiklejohn *et al.* 2003; Hutter *et al.* 2008). These populations have different demographic histories, with the derived European population having experienced a bottleneck that accompanied the out-of-Africa migration (David and Capy 1988; Lachaise *et al.* 1988). This migration was presumably accompanied by adaptation to the new, non-African environment. One way that such adaptation might occur is through changes in gene expression, which may be caused by the fixation of beneficial mutations in either *cis* or *trans* regulatory elements (Wray *et al.* 2003; Rockman and Kruglyak 2006). Since the former will be physically linked to the gene that they regulate, it should be possible to identify them using a combination of population-level gene expression and DNA sequence polymorphism data. Furthermore, since gene expression levels appear to be under strong stabilizing selection (see Chapter 1 and Hutter *et al.* 2008), it is unlikely that large, between-population differences in expression will occur due to neutral drift. Thus, the identification of such genes from microarray studies can reveal promising candidate regions for fine-scale selective sweep mapping.

At the DNA level, the MK test revealed evidence of recurrent positive selection on the coding and upstream region of *CG9509* since the split of *D. melanogaster* and *D. sechellia*. In addition, the European population showed unusually low polymorphism in this region, with a 1.2-kb region being monomorphic. A region of the genome that has recently experienced an episode of positive selection is expected to show a local reduction of variability (Schlotterer 2003; Thornton *et al.* 2007), which is often considered to be the signature of a selective sweep (Pool *et al.* 2005; Beisswanger *et al.* 2006; Harr *et al.* 2006). This is because a positively-selected mutation will rapidly go to fixation within a population and, thereby remove variation at the selected site as well as its flanking regions. This process is also known as genetic hitchhiking (Smith and Haigh 1974). It is possible, however, that demographic processes, such as population size bottlenecks, may reduce variation in a way that appears very similar to the signature of a selective sweep. This is especially relevant to *D. melanogaster*, because the European population is known to have experienced a bottleneck as it migrated from its ancestral African home range (Ometto *et al.* 2005; Li and Stephan 2006).

Two statistical tests applied to the *CG9509* region, the CLR test of Kim and Stephan (2002) and *SweepFinder* (Nielsen *et al.* 2005) indicated that the reduced polymorphism observed in the European population cannot be explained by the standard neutral model. Furthermore, the GOF test of Jensen *et al. (2005)* indicated that the reduction in polymorphism was consistent with the action of a selective sweep. There was also an excess of high-frequency derived variants flanking both sides of the monomorphic region, which is expected in the case of a selective sweep (Fay and Wu 2000). Although all of the above patterns are consistent with a selective sweep, our simulations show that they can also be caused by a bottleneck combined with recombination, which may create a pattern that resembles that of a selective sweep (Barton 1998; Thornton and Jensen 2007). The small size of the monomorphic region (1.2 kb) might reduce the statistical power to detect a selective sweep. Previous investigations of selective sweeps have focused on regions with lengths greater than 10 kb (DuMont and Aquadro 2005; Pool *et al.* 2005; Jensen *et al.* 2007).

In contrast to the European population, there is no evidence for a selective sweep in Africa (CLR and SF are both non-significant). Furthermore, we found no shared haplotypes between the two populations. This suggests that, if a sweep has occurred, it is limited to the European population. This is consistent with a sweep that is caused by adaptation to a new environment. Because the high level of *CG9509* expression observed in the European population was also observed in other non-Africa strains, it is likely that the sweep occurred soon after the out-of-Africa migration. It has already been demonstrated that a large fraction of non-coding DNA is under selective constraint (Bergman and Kreitman 2001; Andolfatto 2005; Bachtrog and Andolfatto 2006). This includes both positive and negative selection in putative gene-regulatory regions. Thus, regulatory polymorphisms may be an important source of adaptive variation to new environmental conditions.

To explain the significant difference in expression of *CG9509* between Europe and Africa, we searched for the presence of derived mutations that are fixed in Europe, but absent or present at low frequency (<10%) in the African population. Such mutations were present mainly in the intergenic region between *CG9509* and *CG14406*, which suggests their importance in gene regulation. Many of the sites where these mutations occurred were well conserved between *D. melanogaster* and *D. ananassae*, which diverged at least 35 million years ago (Da Lage *et al.* 1998). Mutations should accumulate randomly in nonfunctional regions of DNA, therefore only functionally important regions, such as gene-regulatory elements, will be conserved between distantly-related species (Doniger *et al.* 2005; Gompel *et al.* 2005; Macdonald and Long 2005).

The detection of a putative selective sweep in a gene-regulatory region by statistical methods should eventually be verified by functional experiments. In this particular case, no information is available on the function of the *CG9509* gene or the effect of mutations in this gene on the organism's phenotype or fitness. However, the functional role of naturally occurring variants in the upstream region of *CG9509* could be investigated using transgenic reporter gene constructs (Parsch 2004). Such experiments could demonstrate that the region in question is responsible for the observed gene expression difference between populations.

Summary

In this work, I investigate the role of gene regulatory changes in the evolution of *Drosophila melanogaster*. As a first step, I performed a survey of gene expression variation in the species using whole-genome microarrays. I surveyed eight strains from an ancestral African population and eight strains from a derived European population using an experimental design that allowed for the detection of expression differences within and between populations. Levels of gene expression variation were nearly equal within the two populations, but a higher amount of variation was detected in comparisons between the two populations. Most gene expression variation within populations appears to be limited by stabilizing selection. However, some genes that are differentially expressed between the two populations might be targets of positive selection. Some of these encode proteins associated with insecticide resistance, food choice, lipid metabolism, and flight. These genes are good candidates for studying adaptive regulatory evolution that accompanied the out-of-Africa migration of *D. melanogaster*.

To verify the accuracy of the microarray experiments, I performed quantitative Real-Time PCR (qPCR), which is another method to measure gene expression, on a subset of genes. I compared the fold-changes in gene expression between pairs of strains determined by the two methods. I also compared the pattern of expression variation in male and female flies. The qPCR approach supported the general accuracy of the microarray experiments, as the fold-changes measured by the two techniques were highly correlated. Expression differences among the strains tended to be similar for male and females. However, exceptions to this general pattern could be found by looking at the pairwise fold-changes for individual genes, some of which differed in expression pattern between males and females.

I also investigated the molecular evolution and population genetics of the protein-encoding and upstream regulatory regions of genes that have potentially undergone adaptive evolution at the gene-regulatory level. These genes represent a subset of the genes that showed a significant difference in gene expression between the African and European populations. A set of control genes, which showed no significant difference in expression between the two populations, was also included in the analysis. Overall, I found evidence for both positive and purifying selection in the coding and non-coding regions. However, patterns of polymorphism and divergence did not differ significantly between the candidate genes and the control genes.

One of the genes that showed an interesting pattern of expression in the microarray and qPCR experiments was subjected to further, more detailed population genetic analysis. This gene, *CG9509,* has twofold higher expression in the European strains than in the African strains. The coding and the upstream regions of this gene show evidence of recurrent positive selection since the split of *D. melanogaster* and its close relative, *D. sechellia*. A polymorphism survey of the *CG9509* region uncovered a 1.2-kb segment, which included the putative *CG9509* promoter that showed no polymorphism in the European population. The European population also has several fixed or nearly-fixed derived mutations in this region. These observations, coupled with statistical analysis, provide evidence for a selective sweep in the European population. The selective sweep was likely driven by local adaptation at the level of gene expression.

Literature cited

Andersson, M. and L. W. Simmons (2006). "Sexual selection and mate choice." Trends Ecol Evol **21**(6): 296-302.

Andolfatto, P. (2005). "Adaptive evolution of non-coding DNA in Drosophila." Nature **437**(7062): 1149-52.

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.

Bachtrog, D. and P. Andolfatto (2006). "Selection, Recombination and Demographic History in Drosophila miranda." Genetics **174**(4): 2045-2059.

Barton, N. H. (1998). "The effect of hitch-hiking on neutral genealogies." Genetical Research **72**: 123-133.

Baudry, E., B. Viginier and M. Veuille (2004). "Non-African populations of Drosophila melanogaster have a unique origin." Mol Biol Evol **21**(8): 1482-91.

Begun, D. J. and C. F. Aquadro (1993). "African and North American populations of Drosophila melanogaster are very different at the DNA level." Nature **365**(6446): 548-50.

Beisswanger, S., W. Stephan and D. De Lorenzo (2006). "Evidence for a Selective Sweep in the wapl Region of Drosophila melanogaster." Genetics **172**(1): 265-274.

Bergman, C. M. and M. Kreitman (2001). "Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences." Genome Res **11**(8): 1335-45.

Bierne, N. and A. Eyre-Walker (2004). "The Genomic Rate of Adaptive Amino Acid Substitution in Drosophila." Mol Biol Evol **21**(7): 1350-1360.

Bird, C. P., B. E. Stranger and E. T. Dermitzakis (2006). "Functional variation and evolution of non-coding DNA." Curr Opin Genet Dev **16**(6): 559-64.

Britten, R. J. and E. H. Davidson (1969). "Gene regulation for higher cells: a theory." Science **165**(891): 349-57.

Brown, R. P. and M. E. Feder (2005). "Reverse transcriptional profiling: non-correspondence of transcript level variation and proximal promoter polymorphism." BMC Genomics **6**: 110.

Canales, R. D., Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, et al. (2006). "Evaluation of DNA microarray results with quantitative gene expression platforms." Nat Biotechnol **24**(9): 1115-22.

Charlesworth, J. and A. Eyre-Walker (2008). "The McDonald-Kreitman test and slightly deleterious mutations." Mol Biol Evol **25**(6): 1007-15.

Chintapalli, V. R., J. Wang and J. A. Dow (2007). "Using FlyAtlas to identify better Drosophila melanogaster models of human disease." Nat Genet **39**(6): 715-20.

Churchill, G. A. (2002). "Fundamentals of experimental design for cDNA microarrays." Nat Genet **32 Suppl**: 490-5.

Clark, T. A. and J. P. Townsend (2007). "Quantifying variation in gene expression." Mol Ecol **16**: 2613 - 2616.

Clutton-Brock, T. (2007). "Sexual selection in males and females." Science **318**(5858): 1882-1885.

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**(1): 239-49.

Cowles, C. R., J. N. Hirschhorn, D. Altshuler and E. S. Lander (2002). "Detection of regulatory variation in mouse genes." Nat Genet **32**(3): 432-7.

Da Lage, J. L., E. Renard, F. Chartois, F. Lemeunier and M. L. Cariou (1998). "Amyrel, a paralogous gene of the amylase gene family in Drosophila melanogaster and the Sophophora subgenus." Proc Natl Acad Sci U S A **95**(12): 6848-53.

Daborn, P. J., J. L. Yen, M. R. Bogwitz, G. Le Goff, E. Feil, S. Jeffers, N. Tijet, T. Perry, D. Heckel, P. Batterham, et al. (2002). "A single p450 allele associated with insecticide resistance in Drosophila." Science **297**(5590): 2253-6.

David, J. R. and C. Bocquet (1975). "Evolution in a cosmopolitan species: genetic latitudinal clines in Drosophila melanogaster wild populations." Experientia **31**(2): 164-6.

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

De Gobbi, M., V. Viprakasit, J. R. Hughes, C. Fisher, V. J. Buckle, H. Ayyub, R. J. Gibbons, D. Vernimmen, Y. Yoshinaga, P. de Jong, et al. (2006). "A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter." Science **312**(5777): 1215-7.

Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch and W. K. Thomas (2005). "The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans." Nat Genet **37**(5): 544-8.

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

Doniger, S. W., J. Huh and J. C. Fay (2005). "Identification of functional transcription factor binding sites using closely related Saccharomyces species." Genome Res **15**(5): 701-9.

Draghici, S., P. Khatri, A. C. Eklund and Z. Szallasi (2006). "Reliability and reproducibility issues in DNA microarray measurements." Trends in Genetics **22**(2): 101-109.

Draghici, S., P. Khatri, A. C. Eklund and Z. Szallasi (2006). "Reliability and reproducibility issues in DNA microarray measurements." Trends Genet **22**(2): 101-9.

DuMont, V. B. and C. F. Aquadro (2005). "Multiple signatures of positive selection downstream of notch on the X chromosome in Drosophila melanogaster." Genetics **171**(2): 639-53.

Efremov, G. and M. Braend (1964). "Serum Albumin: Polymorphism in Man." Science **146**: 1679-80.

Ellegren, H. and J. Parsch (2007). "The evolution of sex-biased genes and sex-biased gene expression." Nat Rev Genet **8**(9): 689-698.

Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, et al. (2002). "Intra- and interspecific variation in primate gene expression patterns." Science **296**: 340 - 343.

Fakhouri, M., M. Elalayli, D. Sherling, J. D. Hall, E. Miller, X. Sun, L. Wells and E. K. LeMosy (2006). "Minor proteins and enzymes of the Drosophila eggshell matrix." Dev Biol **293**(1): 127-41.

Fay, J. C. and C.-I. Wu (2000). "Hitchhiking Under Positive Darwinian Selection." Genetics **155**(3): 1405-1413.

Fay, J. C., G. J. Wyckoff and C. I. Wu (2001). "Positive and negative selection on the human genome." Genetics **158**(3): 1227-34.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**: R80.

Gibson, G., R. Riley-Berger, L. Harshman, A. Kopp, S. Vacha, S. Nuzhdin and M. Wayne (2004). "Extensive Sex-Specific Nonadditivity of Gene Expression in Drosophila melanogaster." Genetics **167**(4): 1791-1799.

Gillespie, J. H. (1998). Baltimore, MD: The Johns Hopkins University Press.

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**(3): 1269-78.

Gnad, F. and J. Parsch (2006). "Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression." Bioinformatics **22**(20): 2577-9.

Gnad, F. and J. Parsch (2006). "Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression." Bioinformatics **22**: 2577 - 2579.

Gompel, N., B. Prud'homme, P. J. Wittkopp, V. A. Kassner and S. B. Carroll (2005). "Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila." Nature **433**(7025): 481-7.

Greenberg, A. J., J. R. Moran, J. A. Coyne and C. I. Wu (2003). "Ecological adaptation during incipient speciation revealed by precise gene replacement." Science **302**(5651): 1754-1757.

Haddrill, P. R., D. Bachtrog and P. Andolfatto (2008). "Positive and negative selection on noncoding DNA in Drosophila simulans." Mol Biol Evol **25**(9): 1825-34.

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.

Halligan, D. L., A. Eyre-Walker, P. Andolfatto and P. D. Keightley (2004). "Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila." Genome Res **14**(2): 273-9.

Harr, B., C. Voolstra, T. J. Heinen, J. F. Baines, R. Rottscheidt, S. Ihle, W. Muller, F. Bonhomme and D. Tautz (2006). "A change of expression in the conserved signaling gene MKK7 is associated with a selective sweep in the western house mouse Mus musculus domesticus." J Evol Biol **19**(5): 1486-96.

Hill, W. G. and A. Robertson (1966). "The effect of linkage on limits to artificial selection." Genet Res **8**(3): 269-94.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, et al. (2006). "The UCSC Genome Browser Database: update 2006." Nucl. Acids Res. **34**(suppl_1): D590-598.

Holloway, A. K., M. K. Lawniczak, J. G. Mezey, D. J. Begun and C. D. Jones (2007). "Adaptive gene expression divergence inferred from population genomics." PLoS Genet **3**(10): 2007-13.

Hudson, R. R. (2002). "Generating samples under a Wright-Fisher neutral model of genetic variation." Bioinformatics **18**(2): 337-8.

Hudson, R. R., M. Kreitman and M. Aguade (1987). "A test of neutral molecular evolution based on nucleotide data." Genetics **116**(1): 153-9.

Hughes, K. A., J. F. Ayroles, M. M. Reedy, J. M. Drnevich, K. C. Rowe, E. A. Ruedi, C. E. Caceres and K. N. Paige (2006). "Segregating variation in the transcriptome: cis regulation and additivity of effects." Genetics **173**(3): 1347-55.

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 chromosomewide single nucleotide polymorphism data." Genetics **177**: 469 - 480.

Hutter, S., S. Saminadin-Peter, W. Stephan and J. Parsch (2008). "Gene expression variation in African and European populations of Drosophila melanogaster." Genome Biology **9**(1): R12.

Jacob, F. and J. Monod (1961). "Genetic regulatory mechanisms in the synthesis of proteins." J Mol Biol **3**: 318-56.

Jensen, J. D., V. L. Bauer DuMont, A. B. Ashmore, A. Gutierrez and C. F. Aquadro (2007). "Patterns of Sequence Variability and Divergence at the diminutive Gene Region of Drosophila melanogaster: Complex Patterns Suggest an Ancestral Selective Sweep." Genetics **177**(2): 1071-1085.

Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro and C. D. Bustamante (2005). "Distinguishing between selective sweeps and demography using DNA polymorphism data." Genetics **170**(3): 1401-10.

Jeong, S., M. Rebeiz, P. Andolfatto, T. Werner, J. True and S. B. Carroll (2008). "The evolution of gene regulation underlies a morphological difference between two Drosophila sister species." Cell **132**(5): 783-93.

Jin, W., R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel and G. Gibson (2001). "The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster." Nat Genet **29**(4): 389-95.

Karlik, C. C., M. D. Coutu and E. A. Fyrberg (1984). "A nonsense mutation within the act88F actin gene disrupts myofibril formation in Drosophila indirect flight muscles." Cell **38**: 711 - 719.

Kauer, M., B. Zangerl, D. Dieringer and C. Schlotterer (2002). "Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of Drosophila melanogaster." Genetics **160**: 247 - 256.

Kern, A. D. and D. J. Begun (2005). "Patterns of polymorphism and divergence from noncoding sequences of Drosophila melanogaster and D. simulans: evidence for nonequilibrium processes." Mol Biol Evol **22**(1): 51-62.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann and S. Paabo (2005). "Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees." Science **309**(5742): 1850-4.

Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge and S. Paabo (2004). "A neutral model of transcriptome evolution." PLoS Biol **2**: e132.

Kim, Y. and W. Stephan (2002). "Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome." Genetics **160**(2): 765-777.

Kimura, M. (1983). The neutral theory of molecular evolution. Cambridge, Cambrige University Press.

King, M. C. and A. C. Wilson (1975). "Evolution at two levels in humans and chimpanzees." Science **188**: 107 - 116.

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, J. Wakeley and J. Hey (2000). "The population genetics of the origin and divergence of the Drosophila simulans complex species." Genetics **156**(4): 1913-31.

Kohn, M. H., S. Fang and C. I. Wu (2004). "Inference of positive and negative selection on the 5' regulatory regions of Drosophila genes." Mol Biol Evol **21**(2): 374-83.

Kreitman, M. (1983). "Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster." Nature **304**(5925): 412-7.

Kristensen, T. N., P. Sorensen, M. Kruhoffer, K. S. Pedersen and V. Loeschcke (2005). "Genome-wide analysis on inbreeding effects on gene expression in Drosophila melanogaster." Genetics **171**: 157 - 167.

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.

Landry, C. R., P. J. Wittkopp, C. H. Taubes, J. M. Ranz, A. G. Clark and D. L. Hartl (2005). "Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila." Genetics **171**(4): 1813-22.

Lemaitre, B. and J. Hoffmann (2007). "The host defense of Drosophila melanogaster." Annu Rev Immunol **25**: 697 - 743.

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 Int J Org Evolution **59**: 126 - 137.

Li, H. and W. Stephan (2006). "Inferring the Demographic History and Rate of Adaptive Substitution in Drosophila." PLoS Genetics **2**(10): e166.

Livak, K. J., S. J. Flood, J. Marmaro, W. Giusti and K. Deetz (1995). "Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization." PCR Methods Appl **4**(6): 357-62.

Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-80.

Lyne, R., R. Smith, K. Rutherford, M. Wakeling, A. Varley, F. Guillier, H. Janssens, W. Ji, P. McLaren, P. North, et al. (2007). "FlyMine: an integrated database for Drosophila and Anopheles genomics." Genome Biol **8**(7): R129.

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**(3): 607-19.

Malone, J. H. and B. Oliver (2008). "The sex chromosome that refused to die." Bioessays **30**(5): 409-11.

Maynard Smith, J. and J. Haigh (1974). "The hitch-hiking effect of a favourable gene." Genet Res **23**(1): 23-35.

McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature **351**(6328): 652-4.

McGraw, L. A., G. Gibson, A. G. Clark and M. F. Wolfner (2004). "Genes regulated by mating, sperm, or seminal proteins in mated female Drosophila melanogaster." Curr Biol **14**(16): 1509-14.

McIntyre, L., L. Bono, A. Genissel, R. Westerman, D. Junk, M. Telonis-Scott, L. Harshman, M. Wayne, A. Kopp and S. Nuzhdin (2006). "Sex-specific expression of alternative transcripts in Drosophila." Genome Biology **7**(8): R79.

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**(17): 9894-9.

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.

Michalak, P., J. H. Malone, I. T. Lee, D. Hoshino and D. Ma (2007). "Gene expression polymorphism in Drosophila populations." Mol Ecol **16**(6): 1179-89.

Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman and V. G. Cheung (2004). "Genetic analysis of genome-wide variation in human gene expression." Nature **430**(7001): 743-7.

Nei, M. (1987). Molecular Evolutionary Genetics. New York, Columbia Univ. Press.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark and C. Bustamante (2005). "Genomic scans for selective sweeps using SNP data." Genome Res **15**(11): 1566-75.

Nuzhdin, S. V., M. L. Wayne, K. L. Harmon and L. M. McIntyre (2004). "Common pattern of evolution of gene expression level and protein sequence in Drosophila." Mol Biol Evol **21**(7): 1308-17.

Oleksiak, M. F., G. A. Churchill and D. L. Crawford (2002). "Variation in gene expression within and among natural populations." Nat Genet **32**(2): 261-6.

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**(10): 2119-30.

Ometto, L., W. Stephan and D. De Lorenzo (2005). "Insertion/deletion and nucleotide polymorphism data reveal constraints in Drosophila melanogaster introns and intergenic regions." Genetics **169**(3): 1521-7.

Orengo, D. J. and M. Aguade (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.

Osada, N., M. H. Kohn and C. I. Wu (2006). "Genomic Inferences of the cis-Regulatory Nucleotide Polymorphisms Underlying Gene Expression Differences between Drosophila melanogaster Mating Races." Mol Biol Evol **23**(8): 1585-1591.

Parisi, M., R. Nuttall, P. Edwards, J. Minor, D. Naiman, J. Lu, M. Doctolero, M. Vainer, C. Chan, J. Malley, et al. (2004). "A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults." Genome Biol **5**(6): R40.

Parisi, M., R. Nuttall, D. Naiman, G. Bouffard, J. Malley, J. Andrews, S. Eastman and B. Oliver (2003). "Paucity of genes on the Drosophila X chromosome showing male-biased expression." Science **299**(5607): 697-700.

Parsch, J. (2004). "Functional analysis of Drosophila melanogaster gene regulatory sequences by transgene coplacement." Genetics **168**(1): 559-61.

Pavlidis, P., S. Hutter and W. Stephan (2008). "A population genomic approach to map recent positive selection in model species." Mol Ecol.

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 USA **101**: 7034 - 7039.

Pool, J. E. and C. F. Aquadro (2006). "History and Structure of Sub-Saharan Populations of Drosophila melanogaster." Genetics: genetics.106.058693.

Pool, J. E., V. Bauer DuMont, J. L. Mueller and C. F. Aquadro (2005). "A scan of molecular variation leads to the narrow localization of a selective sweep affecting both Afrotropical and cosmopolitan populations of Drosophila melanogaster." Genetics: genetics.105.049973.

Proeschel, M., Z. Zhang and J. Parsch (2006). "Widespread adaptive evolution of Drosophila genes with sex-biased expression." Genetics: genetics.106.058008.

Qiu, F., A. Lakey, B. Agianian, A. Hutchings, G. W. Butcher, S. Labeit, K. Leonard and B. Bullard (2003). "Troponin C in different insect muscle types: identification of two isoforms in Lethocerus, Drosophila and Anopheles that are specific to asynchronous flight muscle in the adult insect." Biochem J **371**: 811 - 821.

Rainer, J., F. Sanchez-Cabo, G. Stocker, A. Sturn and Z. Trajanoski (2006). "NCBI Gene Expression Omnibus (GEO)
CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis." Nucleic Acids Res **34(Web Server issue)**: W498 - W503.

Rand, D. M. and L. M. Kann (1996). "Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans." Mol Biol Evol **13**(6): 735-48.

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.

Reimand, J., M. Kull, H. Peterson, J. Hansen and J. Vilo (2007). "g:Profiler - a web-based toolset for functional profiling of gene lists from large-scale experiments." Nucleic Acids Res **35** (Web Server issue): W193 - W200.

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." **438**(7065): 220-223.

Rifkin, S. A., J. Kim and K. P. White (2003). "Evolution of gene expression in the Drosophila melanogaster subgroup." Nat Genet **33**(2): 138-44.

Rockman, M. V. and L. Kruglyak (2006). "Genetics of global gene expression." Nat Rev Genet **7**(11): 862-72.

Rockman, M. V. and G. A. Wray (2002). "Abundant raw material for cis-regulatory evolution in humans." Mol Biol Evol **19**(11): 1991-2004.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer and R. Rozas (2003). "DnaSP, DNA polymorphism analyses by the coalescent and other methods." Bioinformatics **19**(18): 2496-7.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-86.

Rubin, G. M., L. Hong, P. Brokstein, M. Evans-Holm, E. Frise, M. Stapleton and D. A. Harvey (2000). "A Drosophila complementary DNA resource." Science **287**: 2222 - 2224.

Schlotterer, C. (2003). "Hitchhiking mapping--functional genomics from the population genetics perspective." Trends Genet **19**(1): 32-8.

Shapiro, J. A., W. Huang, C. Zhang, M. J. Hubisz, J. Lu, D. A. Turissini, S. Fang, H. Y. Wang, R. R. Hudson, R. Nielsen, et al. (2007). "Adaptive genic evolution in the Drosophila genomes." Proc Natl Acad Sci USA **104**: 2271 - 2276.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res. **15**(8): 1034-1050.

Smith, J. M. and J. Haigh (1974). "The hitch-hiking effect of a favourable gene." Genet Res **23**(1): 23-35.

Smith, N. G. and A. Eyre-Walker (2002). "Adaptive protein evolution in Drosophila." Nature **415**(6875): 1022-4.

Stark, A., M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, et al. (2007). "Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures." Nature **450**(7167): 219-32.

Stranger, B. E., M. S. Forrest, A. G. Clark, M. J. Minichiello, S. Deutsch, R. Lyle, S. Hunt, B. Kahl, S. E. Antonarakis, S. Tavare, et al. (2005). "Genome-Wide Associations of Gene Expression Variation in Humans." PLoS Genet **1**(6): e78.

Stupar, R. M. and N. M. Springer (2006). "Cis-transcriptional variation in maize inbred lines B73 and Mo17 lead to additive expression patterns in the F1 hybrid." Genetics.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-95.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Thornton, K. R. and J. D. Jensen (2007). "Controlling the false-positive rate in multilocus genome scans for selection." Genetics **175**(2): 737-50.

Thornton, K. R., J. D. Jensen, C. Becquet and P. Andolfatto (2007). "Progress and prospects in mapping recent selection in the genome." Heredity **98**(6): 340-8.

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**(6): 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.

Vieira, F. G., A. Sanchez-Gracia and J. Rozas (2007). "Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution." Genome Biol **8**(11): R235.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Watterson, G. A. (1975). "On the number of segregating sites in genetical models without recombination." Theor Popul Biol **7**(2): 256-76.

Wayne, M. L., M. Telonis-Scott, L. M. Bono, L. Harshman, A. Kopp, S. V. Nuzhdin and L. M. McIntyre (2007). "Simpler mode of inheritance of transcriptional variation in male Drosophila melanogaster." Proc Natl Acad Sci U S A **104**(47): 18577-82.

Whitehead, A. and D. L. Crawford (2006). "Variation within and among species in gene expression: raw material for evolution." Mol Ecol **15**(5): 1197-211.

Wise, E. M. and E. G. Ball (1964). "Malic enzyme and lipogenesis." Proc Natl Acad Sci USA **52**: 1255 - 1263.

Wittkopp, P. J. (2005). "Genomic sources of regulatory variation in cis and in trans." Cell Mol Life Sci **62**(16): 1779-83.

Wittkopp, P. J. (2007). "Variable gene expression in eukaryotes: a network perspective." J Exp Biol **210**(Pt 9): 1567-75.

Wittkopp, P. J., B. K. Haerum and A. G. Clark (2004). "Evolutionary changes in cis and trans gene regulation." Nature **430**(6995): 85-8.

Wittkopp, P. J., B. K. Haerum and A. G. Clark (2008). "Regulatory changes underlying expression differences within and between Drosophila species." Nat Genet **40**(3): 346-50.

Wray, G. A. (2007). "The evolutionary significance of cis-regulatory mutations." Nat Rev Genet **8**(3): 206-216.

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman and L. A. Romano (2003). "The evolution of transcriptional regulation in eukaryotes." Mol Biol Evol **20**(9): 1377-419.

Yang, J., C. McCart, D. J. Woods, S. Terhzaz, K. G. Greenwood, R. H. ffrench-Constant and J. A. Dow (2007). "A Drosophila systems approach to xenobiotic metabolism." Physiol Genomics **30**: 223 - 231.

Zhang, Y., D. Sturgill, M. Parisi, S. Kumar and B. Oliver (2007). "Constraint and turnover in sex-biased gene expression in the genus Drosophila." Nature **450**(7167): 233-7.

Zhang, Z. and J. Parsch (2005). "Positive correlation between evolutionary rate and recombination rate in Drosophila genes with male-biased expression." Mol Biol Evol **22**(10): 1945-7.

Appendix A

DNA EXTRACTION

Protocols of Purgene DNA Isolation kit for one fly (Gentra Systems, Minneapolis, MN, USA).

1. Add 1 fly to the 50 µL chilled Cell-lysis-Solution place on ice in 1.5 mL tube and grind the fly completely.
2. Add an additional 49.5 µL Cell-lysis-Solution and 0.5 µL Proteinase K (20 mg/mL).
3. Homogenize thoroughly.
4. Incubate 1 hour at 55°C then 10 minutes at 65°C.
5. Add 1.5 µL RNase solution A (2 mg/mL) to the tube
6. Mix the sample by inverting the tube and incubate 15 minutes at 37°C.
7. Cool sample at room temperature.
8. Add 33µL Protein-Precipitation-Solution and vortex.
9. Keep the tube 5 minutes on ice.
10. Centrifuge at 13200 rpm for 5 minutes.
11. Transfer the supernatant to a clean 1.5 mL tube.
12. Add 150 µL of 100% Isopropanol and mix the sample by inverting gently.
13. Incubate 5 minutes at room temperature.
14. Centrifuge at 13200 rpm for 5 minutes, remove supernatant.
15. Wash the pellet with 150 µL of 70% ethanol.
16. Centrifuge 2 min at 12000 rpm, carefully remove the ethanol.
17. Dry the pellet at room temperature, resuspend the pellet in 50 µL of distilled water.

POLYMERASE-CHAIN-REACTION (PCR)

Protocols of PeqLab PCR kit (PEQLAB Biotechnologie GMBH, Erlangen, Germany)

Indicated are volumes for a volume final of 25 µL and in parentheses, the concentration

Distilled water	19.75 µL
Buffer specific bufferS(10X)	2.5 µL
dNTP's (12.5mM)	0.5 µL
Forward Primer (10µM)	0.5 µL
Reverse Primer (10µM)	0.5 µL
Taq polymerase (	0.25 µL
DNA template	1 µL

PCR standard run program:

Indicated are the temperature and the duration of each step

1. Initial denaturation	94°C	2 min
2. 30 amplification cycles		
Denaturation	94°C	45 sec
Annealing	X°C	45 sec

	Elongation	72°C	50 sec (for 800 pb)
		X is specific to each pair of primers (between 55°C and 62°C)	
3.	Final elongation	72°C	7 min
4.	Hold	4°C	until storage at -20°C

PCR products are verified on 1%agarose gel

PCR products purification

PCR product	20 µL
ExoSAP-IT (USB, Cleavland, OH, USA)	1 µL

Thermocycler programs

37°C	30 min
80°C	15 min
4°C	until the storage at -20°C

SEQUENCING REACTION

Protocol according to the ABI sequencer 3730 using DYEnamic ET terminator cycle sequencing kit (Applied Biosystems, Foster City, CA, USA)

General condition for a 10 µL reaction
Indicated are volumes and in parentheses the concentration

Big Dye v1.1 (2.5X)	1.5 µL
Sequencing buffer (5X)	1.25µL
Primer (10µM)	2 µL
DNA template	3µL
Distilled water	2.25µL

Thermocycler programs

1.	denaturation	96°C	1 min
2.	40 amplification cycles		
	Denaturation	96°C	10 sec
	Annealing	50°C	15 sec
	Elongation	60°C	4 min
3.	Hold	4°C	until storage at -20°C
4.	Add to sample 10 µL of distilled water		

QUANTITATIVE REAL-TIME PCR

RNA extraction for quantitative Real-Time PCR

1. Collect three sets of flies (15 male flies or 10 female flies per set) and transfer each set to individual 1.5 mL tubes on ice.
2. Add 200 µL Trizol (Invitrogen, Carlsbad, CA, USA) to each tube and grind the flies completely. Combine the three tubes in one 1.5 mL tube. Maintain the tubes on ice
3. Add an additional 400 µL Trizol for a total of 1 mL. Mix by inverting the tube and incubate at room temperature for 5 min.
4. Centrifuge 12,000g at 4°C for 10 min. Transfer supernatant to a clean tube.
5. Add 200 µL chloroform, mix well by shaking the tubes vigorously and incubate at room temperature for 3 min.
6. Centrifuge 12,000g at 4°C for 10 min and transfer the upper phase to a clean Rnase-free.
7. Add 500 µL isopropanol and incubate at room temperature for 10 min.
8. Centrifuge 12,000g at 4 °C for 10min and remove the supernatant.
9. Wash the pellet with 1 mL 75% ethanol prepared with Rnase-free water
10. Remove the ethanol completely and air dry for 10 min.
11. Resuspend the pellet in 50µL of Rnase-free water.

RNA Quantification

RNA was quantifying by spectrophotometer (Bio Rad SmartSpec 3000, Hercules, CA, USA) in 2µL of sample and 98 µL of Rnase-free water.
Get concentration by A_{260} (1abs=40 µg/mL=40ng/µL)

First-Strand cDNA Synthesis Using SuperSript II RT
(Invitrogen, Carlsbad, CA, USA)

Indicated are volumes for a volume final of 20 µL and in parentheses, the concentration

1. Add the following components to a Rnase-free 200µL tube:

Random primers (50-250ng)	1 µL
Total RNA (5µg)	x µL
dNTP's (10 mM each)	1 µL
Sterile, distilled water	to 12 µL

2. Heat mixture to 65°C for 5 min and quick chill on ice. Collect the contents of the tube by brief centrifugation and add:

First-Strand Buffer (5X)	4 µL
DTT (0.1M)	2 µL

3. Mix contents of the tube gently and incubate at 42°C for 2 min.
4. Add 1µL of SuperScipt II RT and mix by pipetting gently up and down and add sterile, distilled water to a 20 µL final volume
5. Incubate tube at 25°C for 10 min
6. Incubate 50 min at 42 °C

7. Inactivate the reaction by heating at 70 °C for 15 min

Real-time PCR reaction

Reactions are prepared for a volume of 20 μL according to TaqMan Gene Expression Assays protocol (Applied Biosystems, Foster City, CA, USA)

TaqMan Universal PCR Master Mix, No AmpErase UNG (2X)	10 μL
TaqMan Gene expression Assay (specific primer) (20 X)	1 μL
cDNA (1/40 diluted in Rnase-free water)	9 μL

Thermocycler programs

For 7500 Fast Real-Time PCR System (Applied Biosciences, Foster City, CA, USA)

1. Initial Setup	95°C	10min
2. 40 amplification cycles		
Denaturation	95°C	15 sec
Annealing	60°C	1 min

MICROARRAY

RNA extraction for Microarray

1. Collect three sets of 23-25 males aged 4-6 days, reared several weeks in advances under standard conditions (22°C and 15h-9h light-dark cycle).
2. Knock flies out and transfer each set to individual 1.5 mL tubes on ice.
3. Add 200 μL Trizol (Invitrogen, Carlsbad, CA, USA) to each tube and grind the flies completely. Combine the three tubes in one 1.5 mL tube. Maintain the tubes on ice
4. Add an additional 400 μL Trizol for a total of 1 mL. Mix by inverting the tube and incubate at room temperature for 5 min.
5. Centrifuge 12,000g at 4°C for 10 min. Transfer supernatant to a clean tube.
6. Add 200 μL chloroform, mix well by shaking the tubes vigorously and incubate at room temperature for 3 min.
7. Centrifuge 12,000g at 4°C for 10 min and transfer the upper phase to a clean Rnase-free.
8. Add 500 μL isopropanol and incubate at room temperature for 10 min.
9. Centrifuge 12,000g at 4 °C for 10min and remove the supernatant.
10. Wash the pellet with 1 mL 75% ethanol prepared with Rnase-free water
11. Remove the ethanol completely and air dry for 10 min.
12. Resuspend the pellet in 30μL of Rnase-free water.
13. Quantify the RNA

cDNA Synthesis and Amino Allyl Labeling

The following protocol is making use Invitrogen packages: Super Script Indirect cDNA Labeling System, Alexa Fluor 555 and Alexa Fluor 647 Reactive Dye DecaPacks and cDNA Labeling Purification Module (Invitrogen, Carlsbad, CA, USA)

First Strand cDNA Synthesis

1. Mix and centrifuge each component in 200 µL RNase-free tubes

Components	Volume
30µg total RNA	X µL
Anchored Oligo (dT) Primer (2.5µg/µl)	2 µL
DEPC treated dH2O	To 18 µL

2. Incubate 70°C for 5 min, and then quick on ice for 1 min.
3. A following to each tube:

Component	Volume
5X First-Strand Buffer	6 µL
0.1 DTT	1.5 µL
dNTP mix	1 µL
RNaseOUT (40/µL)	1 µL
SuperScript III RT	2 µL
Final Volume	30 µL

4. Mix gently and collect the contents of each tube by briefly centrifugation. Incubate at 46°C for 3 hours.

Hydrolysis and Neutralization

1. Add 15 µL of 1N NaOH to reaction tube from the First Strand cDNA synthesis reaction. Mix thoroughly.
2. Incubate tube at 70°C for 10 min.
3. Add 15 µL of 1N HCl immediately after the 10 min incubation to neutralize the pH and mix gently.

Purifying First-Strand cDNA

1. Add 24 µL 3M Sodium Acetate, pH 5.2 and 2 µL glycogen to the combined neutralized reactions and mix.
2. Add 360 µL ice-cold 100% ethanol and mix by vortexing.
3. Place at -20°C for at least 1 hr.
4. Centrifuge at 14,00g at 4°C for 20 min. Carefully remove and discard the supernatant.
5. Wash the pellet with 1 mL 75% ethanol and centrifuge at 14,000g at 4°C for 2 min. Carefully remove and discard the supernatant.

6. Air dry the samples to evaporate any ethanol that may still be on the sample. Sample will turn from white to clear and viscous (glass-like) when ready. Avoid over-drying, as it will be harder to resuspend the samples.
7. Resuspend each sample in 5 µL of 2X Coupling Buffer.

Labeling with Fluorescent Dye

1. Add 2 µL of DMSO directly to each dye vial and mix thoroughly
2. Centrifuge vials briefly.
3. Add the DMSO/dye solution to the tube from the ethanol precipitation step above. Add 3 µL of DEPC-treated H2O to bring the final volume of the sample to 10 µL.
4. Mix samples by vortexing, centrifuge briefly and incubate at room temperature in the dark for 1 hour.

Purifying Labeled cDNA

1. Add 700 µL of Binding Buffer to the reaction tube containing the labeled cDNA and vortex briefly.
2. Transfer each labeled solution in a collection tube with preinserted Spin Cartridge and then load the cDNA/Binding Buffer solution directly into the Spin Cartridge.
3. Centrifuge at 3,300g in a microcentrifuge for 1 min. remove the collection tube and discard the flow-through.
4. Place the Spin Cartridge in the same collection tube and add 600 µL of Wash Buffer to the column.
5. Centrifuge at maximum speed for 30 sec. Rove the collection tube and discard the flow-through.
6. Place the Spin cartridge in the same collection tube and centrifuge at maximum speed for 30 sec to remove any residual Wash Buffer. Remove the collection tube and discard.
7. Place the Spin Cartridge onto a new amber collection tube
8. Add 20 µL of DEPC-treated water to the center of the Spin Cartridge and incubate at room temperature for 1 min.
9. Centrifuge at the maximum speed for 1 min to collect the purified cDNA.
10. Dry down the sample in the speed vac for around 30 min.
11. Resuspend the sample in 12-24 µL in 40 µL of Pronto Long Oligo

Prehybridization and Hybridization

The following protocols are using in accordance with Pronto Microarray Hybridization Kit with small modification (Corning, Akasaka, Japan)

Preparation of Wash solution

Wash Solution 1	
Deionized water	1,118.75 mL
Universal Wash Reagent A	125 mL
Universal Wash Reagent B	6.25 mL

Wash Solution 2

Deionized water	3,562.5 mL
Universal Wash Reagent A	187.5 mL

Wash Solution 3

Deionized water	3,000 mL
Wash Solution 2	750 mL

Presoak and Prehybritation

1. Heat required volumes of both Pronto! Universal Pre-Soak Solution and Pronto! Universal Pre-Hybridization Solution to 42°C for at least 30 min.
2. Add 250 µL Sodium Borohydride Solution to 24.75 mL of 42°C Universal Pre-Soak solution
3. Immerse arrays in solution and incubate at 42°C for 20 min.
4. Transfer arrays to Wash Solution 2 and incubate at ambient temperature for 30 sec.
5. Transfer to a fresh container of Wash Solution 2 for 30 sec.
6. Transfer arrays to 42°C Universal Pre-Hybridization Solution (from step 1)
7. Transfer arrays to a fresh arrays to Wash Solution 2 and incubate at ambient temperature for 1 min.
8. Transfer arrays to Wash Solution 3 and incubate at ambient temperature for 30 sec. Repeat this sep one time
9. Dip arrays in nuclease-free water at ambient temperature (22-25°C), and dry by centrifuging at 1,600g for 2 min.

Hybridization

1. Combine the two samples into a 200 µL tube and mix well.
2. Incubate the labeled cDNA solution at 95°C for 5 min.
3. Centrifuge the cDNA at 13,500g for 2 min.
4. Place the array in chamber. Transfer the cDNA onto the surface of the printed side of the array slide and cover with cover glass.
5. Close the chamber and incubate it at 42°C for 20 hours in water bath.

Post-Hybridization Washes

1. Heat Wash Solution in a container to 42°C for at least 30 min.
2. Disassemble the hybridization chamber
3. Immerge arrays in Wash Solution 1 at 42°C for 1-2 min until the cover glass falls from the slide
4. Transfer arrays to a fresh container of Wash Solution 1 at 42°C and incubate for 5 min
5. Transfer arrays to Wash Solution 2 at ambient temperature and incubate 10 min.
6. Transfer to Wash Solution 3 at ambient temperature and incubate for 2 min. Repeat this step twice.
7. Dry arrays by centrifugation at 1,600g for 2 min
8. The arrays is ready to scan.

Appendix B

Gene	Forward primer (5’-3’)	Reverse (5’-3’)	Ta [§]
CG9509part1	GCCCCTGTTCAATTTATTCG	TTCTGAATCGGCATCATCAC	60°C
CG9509part2	GGCTGCAGCTCTTAAATGGC	ACGAGGACGTTGACTTAGCC	56°C
CG9509part3	CCAATGGCTAAGTCAACGTCC	CAAAGAATAGTGCCGGCAAC	58°C
CG9509part4	CCCACACCAACACCATACC	CTCCACATATGGCTGTCCCAAC	56°C
CG9509part5	GATGGTCGCTGCTATTGGC	CTTGAATGGATAGACCCTTGG	58°C
CG9509part6	ACGCAATCTCCAGGATCATGTC	CGTGGGCTAAACTTGTTGCTAAG	60°C
CG14406	CAACGATCCATCGGGTATG	CGCCAAATTTAAACCAGCAC	60°C
CG12398	GTGATGCAATCCCTGAATG	GCAGTGGCTGCATTCGTTG	58°C
CG9511part1	TTCCTCCAGCCATGACTCCTTGG	GGCAAAGAGCGTAAAAATGGGG	58°C
CG9511part2	CCATTTTTCCCCATTTTTACGC	GACCGAACTCAAACTGAGCG	58°C
CG9511part3	GTGGATTCATGCCACATCCC	GATCCAGCTGGGGAGCAAC	58°C
CG9511part4	GCTTTCGCACTGGTTAATCG	TCTTGTCCAGCTTGGCACTG	58°C
CG9511part5	CACCTGGACACGGATATGGAGTAC	CCAAAAGCACGCAACTCTCATC	58°C
CG1468part1	GATGTCACCGACCACAGATCAAAG	CCAAACCAATATTCAGTACGTTTC	55°C
CG1468part2	GTTTTCTTCCAGCCGGTCTC	GCTTTCGGTAGTAGACGTCC	55°C
CG1468part3	GAGAAGCTCCTCCAAAACAGTC	GGTCAACCTCCTGATACTCCTG	55°C
CG1468part4	GCCCAAGAAGTTGGTTGTCC	CTGTCAACCCTCGAAAAGAC	55°C
CG5386part1	AAATTCAGGCCGCCCTTTG	GCTTTCAACAAAAAGGGCTC	58°C
CG5386part2	GGATTTGAGTTTAAGAGCCC	GAGAACTAACTCCGGCTAAC	58°C
CG5386part3	TGTTGCACGAGTTACTGGGC	CAAGGTTATTCAAAATCCTCG	58°C
CG5386part4	CAACAGCAACCAAAATGGC	ACACACATGTGCACGGCAAC	58°C
CG5154part1	GGCTCTCACCATTTGGCTTTATG	GTGTCACTTCCACAACTTC	60°C
CG5154part2	CCATGTATGAGTGGTGGTC	CTCCAGTTCAGCCAGCGAC	60°C
CG5154part3	GTTCATCAAAACAAAGCGTC	AGCTCTTCGAGAAGTGTGGC	60°C
CG5154part4	CGGAACTGAACAACAACGG	CAGGATTCCACTTGGCCGG	60°C
CG14629part1	TTTCTTGCTCGCTGCTCC	TTTAGTAGGGTGCTACCC	60°C
CG14629part2	CTCGATCTCACGCTCTTTC	AAAAGCGGACAAAGCGACC	60°C
CG14629part3	CTCAGACATTGACCCCTCC	GACTCCACGATATTGCTGG	60°C
CG14629part4	TCTTTATCGCCGAAGAGGG	CAGCTTAACACTTTGGGC	60°C
CG7203part1	GGGATCGTATGCATTGTGATC	CCCACACACATTCAGCACCA	59°C
CG7203part2	GCGAAAGTGGTTTGATTTAC	CGTTGGCCAAGTGCTTTTAT	59°C
CG7203part3	TTGATTCGGGAAGGTCAAGG	GGTCTTGGCCACATAGCTG	59°C
CG7203part4	GACCCGTGCTCATCCAGTC	GTGAGTCGATCCCCTGGTAA	59°C
CG7214part1	CGGCTGACTAACTTTTTGGC	CTAAGCCACACCCCTCTCAC	58°C
CG7214part2	CCCTCGGCTGTAAATTCTTG	GAACACTCCTGCCAATGTCC	58°C
CG7214part3	CAGAATCAACTCGTTCGCTG	TCATCCGGTTGGAGTAGTCG	58°C
CG8997part1	GCCTGAGTGGCTTGTTTTG	GGTCTTACGAAAACTTTCCA	56°C
CG8997part2	TTAGATCCGCAGAAGTGCATG	GCCACAAAAGCCAAAATAGC	56°C
CG8997part3	GGTGGCTAATCAAATCGTTC	GGACCTCCAATGACTTCAGC	56°C
CG8997part4	GACCTACAAGTTCACTTTCCGC	GGACACAACTTGAGATTAGACC	56°C

Gene	Forward primer (5'-3')	Reverse (5'-3')	Ta §
CG7953part1	TTAGCTTGCTGTCCGCCACC	CACGTTCAGCTTGAACTTCTG	59°C
CG7953part2	AAGCAGATGGAGTGCGGATG	GAGATAATCGATGTGTTCCC	59°C
CG7916part1	TTGTGGAGGAGTGTCAAGATC	CGCCTGAGTGGCTTGTTTTG	58°C
CG7916part2	ATCCAGACGATTAACACCTG	ACCGATCCGTTCAAGATCAG	58°C
CG7916part3	AATCAGGTGGCTGGCTATAC	GCGGACAGCAAGCTAAAGAG	58°C
CG7916part4	TCTACAAGTTCAGTTGCGCC	CGCATGCTTGATTGTGACTT	58°C
CG33306part1	CGCATGCTTGATTGTGACTT	GAGCCAGGTGACGAATTCTTG	58°C
CG33306part2	GAGTGGTAAACAAATCATCGG	GCTACCATTTTGCAGGACG	58°C
CG33306part3	AGTTGCCGAACCTTTGTCAG	CCGCACGTTTTTGACCACCG	58°C
CG5178part1	GCCACCAATTCACTTCCGAG	GAGTGCCAGGAGAGAGAAAG	58°C
CG5178part2	CCATTGGCAGCAGGATTGG	CCTTGGATCCTTCGACATTG	58°C
CG5178part3	GATATAAAGGCAGGACAGACCG	ATGGGGTACTTCAGCGTCAG	58°C
CG5178part4	GATGATGCGGGTGCATTAG	TGTAGACGGTCTCGTGGATG	58°C
CG5178part5	CCGATTACCTGATGAAGATCC	GAAGGATGAGCACCGACAAC	58°C
CG5402part1	CCCTTCCTTCATTTGACAGC	CTGCTCTCGAGCTGAATATTCTC	58°C
CG5402part2	CAGTCACACCTGAGGCAAATG	TGATTCGCCTCCATCATCACC	58°C
CG5402part3	GGAGTTTTAGGCCTTATGGG	CAGCCTCCTATGAGCGTAGAAC	58°C
CG5144part1	CTCGCTGCTGGCTTCTTTTTG	TAGCACGGCTCCAAACCATTG	59°C
CG5144part2	ATTGTGCGAACGAGAACAGG	GCACAGCGGCCTAGTAATTG	59°C
CG5144part3	CTAGCCATTTGGTTTAATGTCAC	CCATGCGATCGTAGACCTTG	59°C
CG5144part4	GCTTCAAGAAGACGGACAAACAG	GCAGCAGCAACAACGGATAAC	59°C
CG10912part1	CTATTTGCAGAAATGAGGC	CCATCAAAAAGCACCTTG	58°C
CG10912part2	ACCCGTTGAGATAAAGAATGC	GGCTTTTCTCCTATCTTGATGC	58°C
CG10912part3	GCATCAAGATAGGAGAAAAGCC	GCCACTGCTGTTTTCAACG	58°C
CG10912part4	CCTGCTACTCCACTGAGGTAAG	GCCATTCAGACGAAGATAAG	58°C
CG10957part1	GTGGGAGCATTTATTTGTTG	CCTCGAATGACCTAGTGGCT	57°C
CG10957part2	CTCGACGTTCATACAAACG	GTCTTAAAAGCCCGAAAAG	57°C
CG10957part3	GCATTTTCGCTTGCATAAC	AATGTGGACCTTAGTGTTGCCC	57°C
CG10957part4	GAAGATCCATAGGACTTGGAC	TATTGGTCGGATTCAGATTCG	57°C
CG4734part1	TCCTCGGTTTCACTCAGGTTC	TCCAGAAACTTGCAATGGTG	58°C
CG4734part2	CGATAAATGCAGTTCCAGC	TCGACCTCGAAGTAATCACG	58°C
CG4734part3	GAAGAACTGCATTGCAAGCGC	GAAGTAGGTGGCGTGCTTGTG	58°C
CG4734part4	AAGGATGGCTTCTTTGTTGC	CCCAAGCGATTGTACCTGCC	58°C
Cg8661part1	CATGGCGCTGCACATTAATAG	CTAAATGTTAGGATTGCACGG	58°C
Cg8661part2	CATTTAATCTGCAAATCGCCC	GCGTTGATCAATTTGGTCAC	58°C
Cg8661part3	GAACTGGACGATGTGACCG	CGCTTCCTTTTACTTTACGAG	58°C
CG9973part1	TTGGCCGCTCATTATATAGTC	GAACTCACTGGTTTTGCGTG	58°C
CG9973part2	TTGCCTAGCTGCCGTTTTAC	TGGAACCGTGGTATTGCTG	57°C
CG9973part3	ATGGCGTCTTCAATTTTAGC	CTGCTACTGAATTGGGTCC	57°C
CG9973part4	CTCCAAGCTGCCGTCGTTTC	CATCGTTGCTATCCCCCATG	58°C
CG9973part5	GGTCAAGGTCGAGAAGTTGC	CCGTGAATGAATCTGTGGG	58°C
CG9973part6	CAGTGGATGCGACAAGCAGG	CAGTTTGCACCGATTCCCAG	58°C

Gene	Forward primer (5’-3’)	Reverse (5’-3’)	Ta §
CG14503part1	TGCTACGGTGTTAACATGTGC	GGATCTGAGAAGATTTGTTGGG	58°C
CG14503part2	CTACCTTTCTGGCGATAAGC	CCATTTCACTGGGTTCTGCTTC	58°C
CG5623part1	CTGCCACTGTATCTGTAACCTG	GCGAAAAGTTAATGGAACCAAG	58°C
CG5623part2	TGAGCAAGTTTCTTATCGCGC	CGGCTCTTATGAAATTCGG	58°C
CG5623part3	GCGCGATTTATGTGCGTTTG	CGAAGAACAGACGGAGGAAG	58°C
CG13061part1	CATCCTTTGAAAAATGTTTCTTC	GAATGCAAACCCTTCCGG	58°C
CG13061part2	TGCACTGGGAGAAAAGTC	CAGATTCCAGGTGCCACAAT	58°C
CG13061part3	TCAGTTTGGTTCTATGTGGC	GCATTATAAACAGGGCACAG	58°C
CG5210part1	CAAATAAACCGAAGGGAGATG	AAACAAAGCTAAGCTCGGACG	54°C
CG5210part2	GGATTGAACAGCGATTAGGGC	ATTATCGCTTCCACACTCGACCC	60°C
CG5210part3	CGATTCCGCCAGTTTCGTC	GACATCCAAGCCATCGAATC	56°C
CG5210part4	ATCGATGAACTGGAACCGGC	TGAGCAGCCACATTGAACTC	54°C
CG5210part5	AATTGGCAGCCTGTGGAAG	GTAACCAAGAAGATGCGATAGCG	58°C
CG8768part1	GAAGCGCAATAATTGAGGAGC	CGCTTAAAGTGGACTGCTGCC	54°C
CG8768part2	CAATTTAGGTTCAGTGTCTGC	CATTTACCACGGCGTTTACGG	60°C
CG8768part3	GTGTCGTAATTCGGGAAAAG	GATGTACTGTATCAAGCTGC	58°C
CG8768part4	GAACTCGCGAATCAACTCCTC	ATTTTCGGAGGGCGTGAT	58°C
CG13675part1	CCTGGAAAGTTATTAGCGTG	CCATGAAATGCCGAAAAGAG	58°C
CG13675part2	ATTGCCAATCACTTTGCCC	GTGGGACTCTTCCCACCTTT	58°C
CG13675part3	TGCCTGCGCTGAGATAATTC	GATTGCCGTGTTGCTGTTG	58°C
CG13675part4	GTCGAATGTGAACTGCGAGG	CAACGGGTACATTTTGGGTA	58°C
CG9602part1	ACCGTCTACGACATGTGGC	TTGCGCACAAATGTTGGCCC	60°C
CG9602part2	CATGTCGTGGTCGAAGTTGC	AGTCGCTGACCAGACCGG	60°C
CG9602part3	CAGCCTCCGTAGAAGTAAACC	CCAGGAGGTAATACCCGTAGC	60°C
CG16916part1	GTATTCCGGGATTTCACGG	CCATTTTGACGCGATTTCG	61°C
CG16916part2	AGGACAACGACGACAACGAC	CCCGATATATACCAGCTCTG	58°C
CG16916part3	GGAAATCGAAATCGCGTC	GTTGTACGGCATTCTTTTCG	58°C
CG16916part4	CCGATTGAAAGAGCAGCAC	CATCGAAATGGAGCTATCCG	60°C
CG16916part5	AACTGGACATGGAGGATCTC	CAATGCTCCTTGTTTGCATC	58°C
CG5832part1	GGAATTCCGATGCGTTCGAC	GGGTTGCTACTACTCCAAG	58°C
CG5832part2	GTGTGAGGTTAAGTGTTGCAGC	GTTCGGGCTCATGGAAGTG	58°C
CG5832part3	CAACACCACTGCCAGCTCGAAC	GGCATCCTCGTCGTCCATC	58°C
CG12912part1	GGGTCATTGTGGTCATAGC	GGGCTTGACTTTCGACATTTAC	58°C
CG12912part2	ATGTCGAAAGTCAAGCCCAG	ATTTTGCGAAGCAGCTGAC	58°C
CG12912part3	CACACACACTCGGATTTTCC	CTGCTTCTGTATCTGTGTCC	58°C
CG12912part4	CACACAGCTGGTGGTTGTTG	CACCTGTGCGTTTCGTTTC	58°C
CG12683part1	GAGAGGACAATGGGAACGTTC	CTCTCTTCGCGTTTTCTCTTAG	58°C
CG12683part2	AAAGGGAAAGCGACAGGTC	CACTTTCGCCTGATGTCTCTTG	58°C
CG15314part1	TTTGCCCTTGAACAGCGG	GCGCTAGTTTTGCTTTGTCC	58°C
CG15314part2	TTATCTCGAAGGTGTCTGCG	CGCGGTACATGGAATTATATTG	58°C
CG15314part3	CATGTAGCGCGCTTTTGAATTG	CGACAAGTTTCTTCGTTGC	58°C

Note: PCR primers were used for both PCR and sequencing and the primers were designed for overlapping fragments ~ 800 bp from 5’ (part1) to 3’.

§ Ta indicates the annealing temperature used for the PCR program

Appendix C

CG Number	Gene Name	Europe	Africa	Chr
CG8453	*Cyp6g1*	4.350	1	2R
CG9509	*CG9509*	2.311	1	X
CG10097	*CG10097*	1.853	1	3R
CG10095	*dpr15*	1.803	1	3R
CG10120	*Men*	1.762	1	3R
CG11218	*Obp56d*	1.679	1	2R
CG15036	*CG15036*	1.595	1	X
CG32684	*a-Man-I*	1.560	1	X
CG18135	*CG18135*	1.518	1	3L
CG13183	*CG13183*	1.507	1	2R
CG18135	*CG18135*	1.499	1	3L
CG33272	*CG33272*	1.498	1	3L
CG32372	*CG32372*	1.491	1	3L
CG18345	*trpl*	1.481	1	2R
CG9511	*CG9511*	1.461	1	2L
CG9280	*Glt*	1.456	1	2L
CG11804	*ced-6*	1.421	1	2R
CG3943	*kraken*	1.413	1	2L
CG1865	*Spn43Ab*	1.412	1	2R
CG5154	*Idgf5*	1.407	1	2R
CG31358	*CG31358*	1.403	1	3R
CG33271	*CG33271*	1.390	1	3L
CG10345	*CG10345*	1.363	1	3R
CG7052	*TepII*	1.351	1	2L
CG1468	*CG1468*	1.350	1	X
CG16953	*CG16953*	1.342	1	3R
CG12703	*CG12703*	1.339	1	X
CG3523	*CG3523*	1.338	1	2L
CG5210	*Chit*	1.321	1	2R
CG3050	*Cyp6d5*	1.321	1	3R
CG5315	*CG5315*	1.318	1	3R
CG12304	*CG12304*	1.313	1	3L
CG5877	*CG5877*	1.305	1	X
CG31764	*vir-1*	1.300	1	2L
CG17836	*CG17836*	1.300	1	3R
CG12262	*CG12262*	1.290	1	3L
CG4264	*Hsc70-4*	1.281	1	3R
CG14648	*CG14648*	1.266	1	3R
CG10146	*AttA*	1.266	1	2R
CG4389	*CG4389*	1.264	1	2L

CG Number	Gene Name	Europe	Africa	Chr
CG14629	*CG14629*	1.258	1	X
CG13091	*CG13091*	1.257	1	2L
CG6718	*CG6718*	1.255	1	3L
CG1106	*Gel*	1.246	1	3R
CG6953	*fat-spondin*	1.246	1	2R
CG3308	*CG3308*	1.245	1	3R
CG8913	*Irc*	1.245	1	3R
CG14224	*CG14224*	1.242	1	X
CG5119	*pAbp*	1.242	1	2R
CG4475	*Idgf2*	1.242	1	2L
CG8639	*Cirl*	1.240	1	2R
CG1106	*Gel*	1.240	1	3R
CG10033	*for*	1.237	1	2L
CG5393	*apt*	1.237	1	2R
CG7176	*Idh*	1.235	1	3L
CG11763	*micr*	1.230	1	2R
CG12070	*Sap-r*	1.225	1	3R
CG1554	*RpII215*	1.225	1	X
CG1483	*Map205*	1.221	1	3R
CG9429	*Crc*	1.221	1	3R
CG1146	*CG1146*	1.219	1	3L
CG11567	*Cpr*	1.216	1	2L
CG8983	*ERp60*	1.209	1	2R
CG7470	*CG7470*	1.197	1	3L
CG17246	*Scs-fp*	1.189	1	2R
CG7176	*Idh*	1.185	1	3L
CG11395	*CG11395*	1.175	1	2R
CG2727	*emp*	1.171	1	2R
CG8430	*Got1*	1.155	1	2R
CG5386	*CG5386*	1.151	1	3R
CG18815	*CG18815*	1.150	1	3L
CG17292	*CG17292*	1.139	1	2L
CG31893	*Peritrophin-15b*	1.127	1	2L
CG1522	*cac*	1.112	1	X
CG32245	*CG32245*	1.095	1	3L
CG5580	*sbb*	1.084	1	2R
CG3805	*CG3805*	1	1.095	2L
CG5683	*Aef1*	1	1.107	3L
CG4027	*Act5C*	1	1.147	X
CG13908	*CG13908*	1	1.149	3L
CG5028	*CG5028*	1	1.155	3R
CG15316	*CG15316*	1	1.157	X
CG4027	*Act5C*	1	1.160	X

CG Number	Gene Name	Europe	Africa	Chr
CG4027	*Act5C*	1	1.168	X
CG5623	*CG5623*	1	1.170	3R
CG4000	*CG4000*	1	1.174	3R
CG6803	*Zeelin1*	1	1.180	3R
CG9214	*Tob*	1	1.184	X
CG4027	*Act5C*	1	1.186	X
CG4027	*Act5C*	1	1.187	X
CG32130	*stv*	1	1.187	3L
CG10039	*CG10039*	1	1.194	2L
CG4626	*fz4*	1	1.198	X
CG12233	*l(1)G0156*	1	1.202	X
CG18290	*Act87E*	1	1.211	3R
CG12278	*CG12278*	1	1.227	3R
CG2258	*CG2258*	1	1.237	X
CG31072	*Lerp*	1	1.242	3R
CG8210	*Vha14*	1	1.242	2R
CG1793	*MED26*	1	1.244	4
CG11303	*TM4SF*	1	1.245	2R
CG1572	*CG1572*	1	1.245	X
CG4412	*ATPsyn-Cf6*	1	1.249	3R
CG3127	*Pgk*	1	1.252	2L
CG7390	*smp-30*	1	1.270	3R
CG13047	*CG13047*	1	1.279	3L
CG6163	*CG6163*	1	1.294	3L
CG3140	*Adk2*	1	1.297	2R
CG6544	*fau*	1	1.298	3R
CG4533	*l(2)efl*	1	1.299	2R
CG3606	*caz*	1	1.312	X
CG11765	*Prx2540-2*	1	1.320	2R
CG10597	*CG10597*	1	1.320	X
CG33254	*CG33254*	1	1.321	X
CG9032	*sun*	1	1.325	X
CG9973	*CG9973*	1	1.329	3L
CG6579	*CG6579*	1	1.339	2L
CG1674	*CG1674*	1	1.346	4
CG33306	*CG33306*	1	1.354	2L
CG2017	*CG2017*	1	1.358	3R
CG14419	*CG14419*	1	1.362	X
CG4843	*Tm2*	1	1.368	3R
CG16944	*sesB*	1	1.371	X
CG13375	*CG13375*	1	1.382	X
CG4795	*Cpn*	1	1.392	3R
CG7409	*CG7409*	1	1.402	3L

CG Number	Gene Name	Europe	Africa	Chr
CG6457	*yip7*	1	1.404	3L
CG7953	*CG7953*	1	1.411	2L
CG6803	*Zeelin1*	1	1.419	3R
CG32030	*CG32030*	1	1.432	3L
CG5144	*CG5144*	1	1.439	3L
CG7266	*Eip71CD*	1	1.458	3L
CG5177	*CG5177*	1	1.462	2L
CG12408	*TpnC4*	1	1.465	2R
CG4898	*Tm1*	1	1.466	3R
CG33138	*CG33138*	1	1.468	2R
CG4734	*CG4734*	1	1.470	2R
CG7478	*Act79B*	1	1.495	3L
CG32031	*Argk*	1	1.498	3L
CG3724	*Pgd*	1	1.501	X
CG13057	*retinin*	1	1.506	3L
CG5402	*CG5402*	1	1.530	3R
CG8137	*Spn2*	1	1.533	2L
CG10842	*Cyp4p1*	1	1.546	2R
CG7445	*fln*	1	1.553	3L
CG7916	*CG7916*	1	1.562	2L
CG9676	*CG9676*	1	1.564	X
CG2184	*Mlc2*	1	1.572	3R
CG5596	*Mlc1*	1	1.597	3R
CG10912	*CG10912*	1	1.622	2R
CG9441	*Pu*	1	1.649	2R
CG8997	*CG8997*	1	1.671	2L
CG4123	*Mipp1*	1	1.832	3L
CG8661	*CG8661*	1	1.836	X
CG13061	*Nplp3*	1	1.849	3L
CG17820	*fit*	1	1.923	3R
CG2981	*TpnC41C*	1	2.110	2R
CG3301	*CG3301*	1	2.244	3R
CG5178	*Act88F*	1	2.917	3R
CG7203	*CG7203*	1	5.317	2L
CG7214	*CG7214*	1	5.361	2L

List of the genes significantly differentially expressed between European and African populations (P < 0.002)

Appendix D

ΔCt values obtained by qPCR of male samples for the 12 genes surveyed.

Expression ratios for several comparisons involving these strains (which have very low expression levels for *CG15281* and *CG9438*) fell outside of the axis boundaries of Figure X and are, therefore, not shown (bold number). However, all data points were included in the regression analysis.

Empty cells indicate that expression levels of the gene were not measured for the corresponding strain.

	European samples								African samples							
Gene	E01	E12	E14	E15	E16	E17	E18	E20	A82	A84	A95	A131	A186	A377	A384	A398
CG8453	-2.394	-2.773	-2.564	-2.973	-2.886	-2.355	-2.794	-2.976	0.429	4.622	1.229	-0.625	-1.900	-1.554	1.233	0.894
CG7214	10.616	10.247	9.685	7.927	8.635	9.829	10.028	11.583	4.758	7.641	8.239	7.142	7.782	6.907	8.111	7.119
CG7203	3.694	2.883	3.582	3.059	1.783	3.163	2.496	4.248	1.912	1.312	1.453	1.324	0.408	2.789	0.305	2.828
CG9509	2.983	2.295	3.063	2.464	2.394	3.768	2.495	2.611	4.041	3.367	3.826	3.936	2.930	4.066	3.894	3.899
CG9438	5.893	3.447	**14.171**	3.029	2.796	2.414	1.539	2.758	5.942	4.226	5.433	—	3.331	2.687	—	4.553
CG18179	6.808	2.548	1.739	—	—	4.647	—	9.366	7.014	3.132	8.215	3.982	—	—	2.272	—
CG5791	3.927	4.323	3.338	—	4.585	3.733	3.428	—	4.392	5.102	3.648	4.437	—	4.024	—	—
CG15281	**16.096**	**17.114**	5.286	—	8.427	7.019	8.893	7.465	**21.482**	**15.985**	7.848	—	8.544	6.709	—	7.505
CG8997	1.202	1.871	1.739	1.524	1.957	0.067	2.654	3.105	-1.336	-1.092	1.568	0.870	0.860	0.817	1.028	0.454
CG18180	2.75	—	-0.306	—	2.180	1.458	—	5.125	3.664	1.510	3.276	—	—	2.601	—	2.690
CG15295	10.775	10.643	11.401	10.726	11.499	11.166	11.344	11.801	10.749	12.758	11.164	11.685	11.048	11.676	11.034	11.338
CG5330	3.304	3.167	3.407	3.182	3.021	2.752	2.703	3.174	1.458	2.598	3.003	2.722	2.443	2.991	3.068	2.779

ΔCt values obtained by qPCR of female samples for the 12 genes surveyed.

Expression ratios for several comparisons involving these strains (which have very low expression levels for *CG15281*) fell outside of the axis boundaries of Figure X and are, therefore, not shown (bold number). However, all data points were included in the regression analysis.

Empty cells indicate that expression levels of the gene were not measured for the corresponding strain.

	European samples								African samples							
Gene	E01	E12	E14	E15	E16	E17	E18	E20	A82	A84	A95	A131	A186	A377	A384	A398
CG8453	0.569	0.879	0.680	0.504	0.319	-0.216	0.313	0.561	7.700	4.589	4.383	1.943	2.081	1.385	5.075	4.073
CG7214	14.907	17.135	14.190	13.150	13.746	14.934	15.153	14.921	9.944	10.691	11.831	12.867	9.118	12.112	9.361	10.554
CG7203	4.515	7.911	6.035	8.022	7.676	6.339	6.367	7.738	5.680	6.433	5.314	6.421	4.576	6.650	4.036	4.824
CG9509	4.244	3.719	3.572	4.578	5.625	6.663	5.754	3.386	5.466	6.319	6.014	5.339	6.782	7.522	6.467	7.128
CG9438	9.987	6.656	12.455	5.471	6.323	5.344	4.967	7.162	9.047	8.054	7.795	—	7.439	5.279	—	7.187
CG18179	8.909	5.519	2.007	—	—	4.506	—	8.344	10.657	4.849	6.805	6.534	—	—	3.518	—
CG5791	6.186	8.127	7.782	—	7.855	6.028	6.302	—	8.149	8.594	6.274	8.895	—	7.421	—	—
CG15281	**18.883**	**17.107**	6.298	—	5.896	6.355	7.178	9.758	**19.356**	**16.021**	8.324	—	7.106	5.922	—	8.902
CG8997	0.962	2.915	-0.244	2.447	2.977	3.153	3.919	0.811	2.557	1.906	0.108	2.992	3.103	2.902	2.074	-0.471
CG18180	4.119	—	1.417	—	2.818	1.913	—	5.401	7.467	3.847	3.116	—	—	4.261	—	4.519
CG15295	15.013	14.657	16.771	14.765	14.521	14.058	14.701	11.654	15.260	14.319	11.879	15.552	13.237	14.286	12.940	12.540
CG5330	2.079	1.436	-0.473	0.381	1.769	1.957	2.105	-0.416	2.321	2.773	0.185	2.064	2.600	2.841	1.844	1.287

Microarray values (Chapter 1) of the 12 genes surveyed in qPCR experiment

	European samples								African samples							
Gene	E01	E12	E14	E15	E16	E17	E18	E20	A82	A84	A95	A131	A186	A377	A384	A398
CG8453	12.1755	8.8902	6.85066	5.6806	6.4347	8.8291	8.7144	8.7096	1.7079	1.4564	1.5631	4.2879	5.6975	5.8307	1	1.255
CG7214	1.04871	1.28159	2.97314	3.39	2.3945	2.2621	1.3276	1	5.3347	6.545	5.6253	4.0665	7.5992	6.0995	7.772	9.6937
CG7203	1	1.05315	2.00973	1.603	1.9809	3.6677	1.6603	1.455	4.7715	13.051	9.4506	7.526	6.102	6.6789	12.019	7.9421
CG9509	3.32308	3.10393	2.67437	2.8285	3.0584	1.841	4.3269	3.6374	1	1.6063	1.4373	2.0647	1.5122	1.8325	1.4005	1.5081
CG9438	1	2.87245	1.04437	4.1236	4.0326	4.7663	7.0844	5.8321	1.8152	3.5472	1.862	—	4.2476	9.5639	—	1.8282
CG18179	3.78905	35.5351	34.5771	—	—	9.5038	—	1	6.1987	11.368	9.3448	11.158	—	—	17.132	—
CG5791	3.48013	2.31113	3.30245	—	3.6003	3.0319	3.7623	—	4.9062	1	5.177	2.0343	—	2.5115	—	—
CG15281	7.22478	10.7353	15.7578	—	10.148	8.4359	7.8532	9.9714	7.1053	2.7487	10.124	—	10.427	11.09	—	8.3506
CG8997	2.22339	3.09695	2.44681	2.678	1.6905	2.6789	1	1.3947	3.2669	3.3826	2.621	2.5608	2.7649	2.5026	2.6555	3.4286
CG18180	4.32819	—	13.7305	—	4.6433	6.7809	—	1	4.2309	5.5713	4.5204	—	—	4.092	—	4.2589
CG15295	1.49375	1.11377	1.18266	1.1627	1.0653	1.3262	1.3509	1.4488	1.1665	1.2458	1.0065	1.3202	1	1.3074	1.2409	1.0386
CG5330	1	1.10959	1.28713	1.0097	1.2051	1.023	1.0264	1.3505	1.1345	1.0002	1.1602	1.054	1.0969	1.1573	1.011	1.029

Appendix E

Gene	Pop	Outgroup	Loc	Region	lines	Total sites	S sites	NS sites	π_S	π_{NS}	θ_S	θ_{NS}	K_S	K_{NS}
CG7203	E	sim	2L	coding	12	435	126.128	308.872	0.02182	0.00231	0.01838	0.00214	0.08454	0.0084
CG7214	E	sim	2L	coding	12	426	114.167	311.833	0.00359	0	0.0029	0	0.02602	0
CG5623	E	sim	3R	coding	12	837	190.59	646.41	0.01558	0.00026	0.01737	0.00051	0.11917	0.0253
CG5402	E	sim	3R	coding	12	459	114.154	344.846	0.00426	0.00242	0.0029	0.0048	0.19811	0.0584
CG5178	E	sim	3R	coding	11	1128	274.625	853.376	0.00866	0	0.00746	0	0.06822	0
CG5144	E	sim	3L	coding	11	1158	265.528	892.473	0.02139	0.00077	0.02314	0.00115	0.18945	0.0149
CG8661	E	sim	X	coding	11	594	143.556	450.444	0.01046	0.00218	0.01189	0.00152	0.14255	0.0198
CG5386	E	sim	3R	coding	12	750	179.846	570.154	0.00907	0.00189	0.01289	0.00232	0.05055	0.0169
CG10912	E	sim	2R	coding	10	813	199.651	613.349	0.00435	0.00192	0.00531	0.00173	0.19511	0.0847
CG5154	E	sim	2R	coding	11	1329	316.486	1012.32	0.02144	0.00468	0.02265	0.00472	0.15987	0.0241
CG9511	E	sim	2L	coding	9	1203	285.767	917.235	0.03136	0.00176	0.0309	0.00201	0.19696	0.0082
CG5210	E	sim	2R	coding	11	1356	332.695	1023.31	0.02162	0.00103	0.01642	0.00067	0.09174	0.008
CG9973	E	sim	3L	coding	11	1914	442.64	1474.36	0.01401	0.00207	0.01388	0.00208	0.06385	0.0143
CG14629	E	sim	X	coding	12	951	231.654	719.347	0.00144	0.00057	0.00286	0.00046	0.10733	0.0102
CG1468	E	sech	X	coding	11	525	119.961	405.639	0.00612	0.00468	0.00572	0.00337	0.31186	0.1243
CG10597	E	sech	X	coding	12	681	185.564	495.436	0.01935	0.00196	0.01249	0.00134	0.08453	0.0256
CG4734	E	sech	2R	coding	10	975	236.591	738.41	0.03411	0.00317	0.02839	0.00335	0.10753	0.0163
CG7953	E	sech	2L	coding	12	297	210.038	680.962	0.00645	0.00085	0.00788	0.00097	0.12819	0.0183
CG7916	E	sech	2L	coding	12	861	198.679	662.321	0.03962	0.00025	0.03333	0.0005	0.11247	0.0092
CG8997	E	sech	2L	coding	12	774	174.974	599.025	0.01726	0.00157	0.02082	0.00055	0.11841	0.0173
CG33306	E	sech	2L	coding	11	242	169.514	556.486	0.03154	0.00196	0.02417	0.00123	0.19906	0.0119
CG13061	E	sech	3L	coding	12	247	77.885	189.112	0.00214	0	0.00425	0	0.11171	0.0663
CG9509	E	sech	X	coding	12	1938	466.796	1471.2	0.0058	0.00106	0.00497	0.00068	0.10771	0.0369
CG13675	E	sim	3L	coding	12	840	169.141	670.859	0.01703	0.0005	0.0137	0.00099	0.07127	0.0032
CG9602	E	sim	3R	coding	11	504	116.681	387.319	0.02762	0	0.03511	0	0.16588	0.0052
CG15314	E	sim	X	coding	11	573	154.708	418.291	0.00401	0.00122	0.00441	0.00083	0.02875	0.0178
CG14503	E	sim	2R	coding	11	129	27.542	101.458	0	0.00179	0	0.00337	0	0.0108
CG12912	E	sim	2R	coding	9	306	68.75	247	0.00324	0	0.00535	0	0.01709	0.0081
CG5832	E	sim	3R	coding	12	768	180.205	587.795	0.01357	0.00098	0.0147	0.00113	0.1534	0.0014
CG8768	E	sim	2R	coding	12	906	215.743	690.257	0	0.00048	0	0.00096	0.1425	0.0119
CG16916	E	sech	X	coding	12	1239	290.013	948.989	0.00288	0	0.00571	0	0.08029	0.0021

						MK test						Singleton		Tajima's D	
Gene	Pop	Outgroup	Loc	Region	lines	Dn	Pn	Ds	Ps	NI	p-value	S	N	S	N
CG7203	E	sim	2L	coding	12	2	2	8	7	1.143	1	1	1	0.73	0.23
CG7214	E	sim	2L	coding	12	0	0	3	1	NA	NA	0	0	0.55	NA
CG5623	E	sim	3R	coding	12	16	1	16	10	0.1	0.029*	3	1	-0.42	-1.12
CG5402	E	sim	3R	coding	12	16	5	20	1	6.25	0.184	0	5	1.08	-1.83
CG5178	E	sim	3R	coding	11	0	0	14	6	NA	NA	1	0	0.64	NA
CG5144	E	sim	3L	coding	11	12	3	38	18	0.528	0.527	10	2	-0.34	-1.13
CG8661	E	sim	X	coding	11	8	2	17	5	0.85	1	3	0	-0.46	1.33
CG5386	E	sim	3R	coding	12	9	4	7	7	0.444	0.440	5	1	-1.16	-0.65
CG10912	E	sim	2R	coding	10	48	3	34	3	0.708	0.693	2	0	-0.65	0.40
CG5154	E	sim	2R	coding	11	21	14	36	21	1.143	0.827	7	5	-0.24	-0.04
CG9511	E	sim	2L	coding	9	5	5	40	24	1.667	0.500	6	2	0.07	-0.53
CG5210	E	sim	2R	coding	11	7	2	23	16	0.411	0.451	2	0	1.41	1.65
CG9973	E	sim	3L	coding	11	18	9	19	18	0.528	0.306	8	3	0.04	-0.02
CG14629	E	sim	X	coding	12	7	1	23	2	1.643	1	2	0	-1.45	0.55
CG1468	E	sech	X	coding	11	44	4	30	2	1.364	1	1	1	0.21	1.42
CG10597	E	sech	X	coding	12	12	2	11	7	0.262	0.235	0	0	2.15	1.36
CG4734	E	sech	2R	coding	10	10	7	16	19	0.589	0.555	3	2	0.94	-0.23
CG7953	E	sech	2L	coding	12	12	2	22	5	0.733	1	2	1	-0.67	-0.36
CG7916	E	sech	2L	coding	12	6	1	14	20	0.117	0.045*	6	1	0.83	-1.15
CG8997	E	sech	2L	coding	12	10	1	14	11	0.127	0.059	5	0	-0.71	4.23
CG33306	E	sech	2L	coding	11	6	2	25	12	0.694	1	3	0	1.32	1.82
CG13061	E	sech	3L	coding	12	12	0	8	1	0	0.429	1	0	-1.14	NA
CG9509	E	sech	X	coding	12	51	3	44	7	0.37	0.193	2	0	0.65	1.85
CG13675	E	sim	3L	coding	12	2	2	9	7	1.286	1	1	0	0.95	-1.45
CG9602	E	sim	3R	coding	11	2	0	11	12	0	0.48	6	0	-0.93	NA
CG15314	E	sim	X	coding	11	7	1	4	2	0.286	0.538	1	0	-0.28	1.17
CG14503	E	sim	2R	coding	11	1	1	0	0	NA	NA	0	1	NA	-1.13
CG12912	E	sim	2R	coding	9	2	0	7	3	0	1	2	0	-0.50	NA
CG5832	E	sim	3R	coding	12	0	2	21	8	NA	0.097	4	1	-0.31	-0.39
CG8768	E	sim	2R	coding	12	8	2	28	0	NA	0.064	0	2	NA	-1.46
CG16916	E	sech	X	coding	12	2	0	20	5	0	1	0	5	-1.83	NA

Gene	Pop	Outgroup	Loc	Region	lines	Total sites	S sites	NS sites	π_S	π_{NS}	θ_S	θ_{NS}	K_S	K_{NS}
CG7203	A	sim	2L	coding	12	435	126.154	308.846	0.02665	0.00418	0.03675	0.00429	0.08452	0.0095
CG7214	A	sim	2L	coding	12	426	114.013	311.987	0.01179	0.00053	0.01452	0.00106	0.03593	0.00027
CG5623	A	sim	3R	coding	11	837	190.722	646.278	0.01095	0.00113	0.01611	0.00106	0.1202	0.02605
CG5402	A	sim	3R	coding	10	456	113.242	342.758	0.00315	0.00867	0.00315	0.00722	0.19793	0.06011
CG5178	A	sim	3R	coding	11	1128	274.569	853.431	0.01054	0	0.01119	0	0.06138	0
CG5144	A	sim	3L	coding	10	1158	265.53	892.471	0.02918	0.00102	0.033894	0.001188	0.19162	0.01482
CG8661	A	sim	X	coding	9	594	143.533	450.466	0.0294	0.00414	0.030761	0.004084	0.12983	0.02
CG5386	A	sim	3R	coding	10	762	183.455	578.545	0.00853	0.0047	0.009634	0.00611	0.04899	0.01873
CG10912	A	sim	2R	coding	10	801	196.318	604.682	0.01026	0.00731	0.0108	0.00643	0.20357	0.08316
CG5154	A	sim	2R	coding	9	1329	316.683	1012.318	0.02824	0.00468	0.032532	0.004362	0.16156	0.02406
CG9511	A	sim	2L	coding	9	1203	285.267	917.735	0.0541	0.00334	0.05546	0.00321	0.20091	0.0092
CG5210	A	sim	2R	coding	10	1356	332.924	1023.077	0.03673	0.00126	0.03398	0.00104	0.09225	0.00707
CG9973	A	sim	3L	coding	10	1914	440.183	1473.817	0.02229	0.0032	0.02088	0.00312	0.06177	0.01328
CG14629	A	sim	X	coding	12	951	231.833	719.167	0.00216	0.00247	0.00429	0.00322	0.10766	0.01156
CG1468	A	sech	X	coding	9	531	121.433	409.567	0.01759	0.00245	0.01818	0.0027	0.3389	0.13195
CG10597	A	sech	X	coding	10	714	194.515	519.485	0.01734	0.0039	0.01636	0.00544	0.08137	0.02446
CG4734	A	sech	2R	coding	11	975	236.166	738.834	0.04367	0.0039	0.039033	0.003235	0.10876	0.0168
CG7953	A	sech	2L	coding	12	297	210.115	680.885	0.0117	0.00049	0.00946	0.00097	0.12252	0.01809
CG7916	A	sech	2L	coding	12	861	198.372	662.6628	0.03695	0.00025	0.04006	0.0005	0.11217	0.00924
CG8997	A	sech	2L	coding	12	774	174.795	599.205	0.0275	0.00158	0.030311	0.001105	0.11105	0.01723
CG33306	A	sech	2L	coding	12	242	169.385	556.615	0.02659	0.00204	0.0215	0.00178	0.19435	0.01223
CG13061	A	sech	3L	coding	10	252	72.652	179.349	0.00522	0.0041	0.009731	0.003942	0.12232	0.05433
CG9509	A	sech	X	coding	12	1938	466.078	1471.942	0.02617	0.00109	0.02842	0.0018	0.11234	0.03479
CG13675	A	sim	3L	coding	12	831	168.192	662.808	0.01538	0	0.01378	0	0.0668	0.00302
CG9602	A	sim	3R	coding	12	504	116.705	387.295	0.03088	0	0.03121	0	0.1673	0.00518
CG15314	A	sim	X	coding	11	567	153.181	413.819	0.01172	0.00317	0.01783	0.00413	0.02843	0.01801
CG14503	A	sim	2R	coding	11	141	29.5	111.5	0	0.00392	0	0.00306	0	0.0115
CG12912	A	sim	2R	coding	11	315	69.806	245.195	0.03629	0.00372	0.03424	0.00696	0.10765	0.01008
CG5832	A	sim	3R	coding	10	783	181.939	601.06	0.01755	0.001	0.01554	0.00176	0.14909	0.00183
CG8768	A	sim	2R	coding	10	906	215.894	690.106	0.03974	0.00326	0.03766	0.00307	0.13904	0.01007
CG16916	A	sech	X	coding	11	1239	289.611	949.39	0.02502	0.00019	0.03065	0.00036	0.08254	0.00221

						MK test						Singleton		Tajima’s D	
Gene	Pop	Outgroup	Loc	Region	lines	Dn	Pn	Ds	Ps	NI	p-value	S	N	N	S
CG7203	A	sim	2L	coding	12	2	4	8	14	1.143	1	8	3	-1.17	-0.09
CG7214	A	sim	2L	coding	12	0	1	3	5	NA	1	3	1	-0.69	-1.15
CG5623	A	sim	3R	coding	11	16	2	17	9	0.236	0.093	6	1	-1.35	0.20
CG5402	A	sim	3R	coding	10	17	7	20	1	8.235	0.051	0	2	0.00	0.85
CG5178	A	sim	3R	coding	11	0	0	14	9	NA	NA	4	0	-0.24	NA
CG5144	A	sim	3L	coding	10	12	3	38	27	0.352	0.148	15	1	-0.63	-0.51
CG8661	A	sim	X	coding	9	7	5	14	12	0.833	1	6	2	-0.21	0.06
CG5386	A	sim	3R	coding	10	9	10	6	5	1.33	1	2	8	-0.46	-1.02
CG10912	A	sim	2R	coding	10	44	11	33	6	1.375	0.600	3	2	-0.21	0.62
CG5154	A	sim	2R	coding	9	20	12	34	28	0.729	0.516	15	6	-0.66	0.35
CG9511	A	sim	2L	coding	9	6	8	38	43	1.178	1	16	3	-0.12	0.18
CG5210	A	sim	2R	coding	10	6	3	19	32	0.297	0.145	8	1	0.39	0.78
CG9973	A	sim	3L	coding	10	16	13	17	26	0.531	0.232	8	5	0.32	0.12
CG14629	A	sim	X	coding	12	7	7	23	3	7.667	0.018*	3	5	-1.63	-0.91
CG1468	A	sech	X	coding	9	48	3	32	6	0.333	0.163	2	2	-0.14	-0.35
CG10597	A	sech	X	coding	10	11	8	11	9	0.889	1	3	6	0.26	-1.22
CG4734	A	sech	2R	coding	11	10	7	14	27	0.363	0.142	7	3	0.55	0.83
CG7953	A	sech	2L	coding	12	12	2	21	6	0.583	0.692	1	2	0.90	-1.44
CG7916	A	sech	2L	coding	12	6	1	13	24	0.09	0.031*	11	1	-0.35	-1.15
CG8997	A	sech	2L	coding	12	10	2	13	16	0.163	0.038*	5	2	-0.40	1.26
CG33306	A	sech	2L	coding	12	6	3	25	11	1.136	1	2	1	0.98	0.48
CG13061	A	sech	3L	coding	10	9	2	8	2	0.889	1	1	1	-1.50	0.13
CG9509	A	sech	X	coding	12	49	8	40	40	0.163	1.00E-05***	17	5	-0.36	-1.58
CG13675	A	sim	3L	coding	12	2	0	9	7	0	0.497	2	0	0.45	NA
CG9602	A	sim	3R	coding	12	2	0	11	11	0	0.482	2	0	-0.04	NA
CG15314	A	sim	X	coding	11	6	5	3	8	0.313	0.387	7	3	-1.42	-0.89
CG14503	A	sim	2R	coding	11	1	1	0	0	NA	NA	0	0	NA	0.68
CG12912	A	sim	2R	coding	11	2	5	3	7	1.071	1	1	5	0.24	-1.78
CG5832	A	sim	3R	coding	10	0	3	21	8	NA	0.033*	2	3	0.56	-1.58
CG8768	A	sim	2R	coding	10	5	6	17	23	0.887	1	6	2	0.26	0.26
CG16916	A	sech	X	coding	11	2	1	17	26	0.327	0.561	15	1	-0.85	-1.14

										MK test				
Gene	Pop	Outgroup	Loc	Region	lines	sites	π'	θ'	K'	D5'	P5'	p-value	Singletons	Tajima's. D
CG7203	E	sim	2L	Intergenic; 5UTR	12	1084	0.010517	0.01102	0.061828	49	36	0.784	13	-0.209
CG7214	E	sim	2L	Intergenic; 5UTR	12	1005	0.00426	0.00459	0.034618	30	11	1	5	-0.308
CG5623	E	sim	3R	Intergenic; 5UTR	12	1187	0.019691	0.017619	0.098676	90	56	1	14	0.550
CG5402	E	sim	3R	Intergenic; 5UTR	12	1130	0.013036	0.01242	0.051271	42	37	0.0003	9	0.228
CG5178	E	sim	3R	Intergenic; 5UTR	11	1114	0.019718	0.0166	0.053346	39	50	0.047*	3	0.899
CG5144	E	sim	3L	Intergenic; 5UTR	11	1097	0.012568	0.009384	0.069775	61	26	0.853	7	1.585
CG8661	E	sim	X	Intergenic; 5UTR	11	1004	0.013433	0.013079	0.120564	97	37	0.797	12	0.128
CG5386	E	sim	3R	Intergenic	12	992	0.00731	0.009274	0.040826	32	25	0.768	11	-0.955
CG10912	E	sim	2R	Intergenic; 5UTR	10	1070	0.003655	0.003229	0.097576	94	8	1	0	0.586
CG5154	E	sim	2R	Intergenic; 5UTR	11	971	0.015147	0.012206	0.064793	47	34	0.599	7	1.133
CG9511	E	sim	2L	Intergenic; 5UTR	9	1130	0.00891	0.007407	0.07752	76	21	0.032*	3	1.012
CG5210	E	sim	2R	Intergenic; 5UTR	11	862	0.013205	0.013155	0.059913	37	32	0.688	4	0.018
CG9973	E	sim	3L	Intergenic	11	1084	0.003353	0.003779	0.014651	13	11	1	7	-0.491
CG14629	E	sim	X	Intergenic; 5UTR	12	947	0.001677	0.003057	0.084191	73	8	1	8	-1.836*
CG1468	E	sech	X	Intergenic; 5UTR	11	1080	0.001252	0.00211	0.086824	87	7	1	6	-1.654*
CG10597	E	sech	X	Intergenic	12	591	0.009854	0.007609	0.02772	13	9	1	1	1.315
CG4734	E	sech	2R	Intergenic	10	1039	0.009935	0.010198	0.048914	42	27	0.151	12	-0.124
CG7953	E	sech	2L	Intergenic; 5UTR	12	194	0.003287	0.005138	0.028008	5	3	0.346	2	-1.186
CG7916	E	sech	2L	Intergenic; 5UTR	12	1465	0.007661	0.008795	0.057189	69	35	0.015*	19	-0.590
CG8997	E	sech	2L	Intergenic; 5UTR	12	1465	0.007661	0.008795	0.056882	69	35	0.359	19	-0.590
CG33306	E	sech	2L	Intergenic; 5UTR	11	675	0.008066	0.008362	0.05925	32	16	1	8	-0.159
CG13061	E	sech	3L	Intergenic; 5UTR	12	954	0.005929	0.005662	0.079739	67	15	1	4	0.206
CG9509	E	sech	X	Intergenic	12	1115	0.000465	0.000282	0.079893	84	1	0.004**	1	1.487
CG13675	E	sim	3L	Intergenic	12	1141	0.004863	0.004937	0.031853	28	16	0.765	2	-0.065
CG9602	E	sim	3R	Intergenic; 5UTR	11	1082	0.00347	0.003654	0.035264	34	9	0.013*	4	-0.219
CG12683	E	sim	X	Intergenic	12	1090	0.011217	0.009045	0.075461	68	27	NA	4	1.087
CG15314	E	sim	X	Intergenic	11	1048	0.003031	0.002188	0.04523	43	6	0.206	0	1.566
CG14503	E	sim	2R	Intergenic	11	1163	0.005093	0.003764	0.023018	21	13	NA	2	1.549
CG12912	E	sim	2R	Intergenic	9	1083	0.003993	0.003304	0.090968	88	10	0.010*	1	0.973
CG5832	E	sim	3R	Intergenic	12	1016	0.002554	0.002871	0.017509	16	9	0.566	4	-0.449
CG8768	E	sim	2R	Intergenic; 5UTR	12	913	0.000662	0.001033	0.058148	51	2	0.542	2	-1.180
CG16916	E	sech	X	Intergenic; 5UTR	12	1060	0.000157	0.000311	0.054783	56	1	0.009**	1	-1.141

										MK test				
Gene	Pop	Outgroup	Loc	Region	lines	sites	π	θ'	K'	D5'	P5'	p-value	Singletons	Tajima's. D
CG7203	A	sim	2L	Intergenic; 5UTR	12	1061	0.017325	0.018953	0.058371	43	57	0.638	30	-0.401
CG7214	A	sim	2L	Intergenic; 5UTR	12	995	0.006783	0.008967	0.034535	29	23	0.454	19	-1.096
CG5623	A	sim	3R	Intergenic; 5UTR	11	1185	0.017195	0.016765	0.099182	88	52	0.830	17	0.123
CG5402	A	sim	3R	Intergenic; 5UTR	10	1131	0.016702	0.014499	0.049711	39	41	5.00E-5**	12	0.749
CG5178	A	sim	3R	Intergenic; 5UTR	11	1136	0.017544	0.014451	0.054683	42	45	0.349	12	1.019
CG5144	A	sim	3L	Intergenic; 5UTR	10	1092	0.016144	0.017773	0.070019	54	51	0.429	26	-0.455
CG8661	A	sim	X	Intergenic; 5UTR	9	987	0.018194	0.018139	0.121315	95	46	0.262	15	0.016
CG5386	A	sim	3R	Intergenic	10	985	0.014509	0.016094	0.044007	32	40	0.747	23	-0.485
CG10912	A	sim	2R	Intergenic; 5UTR	10	1057	0.013028	0.016818	0.094983	82	46	0.017*	33	-1.114
CG5154	A	sim	2R	Intergenic; 5UTR	9	974	0.016138	0.018055	0.062541	45	45	0.621	25	-0.547
CG9511	A	sim	2L	Intergenic; 5UTR	9	1132	0.006556	0.007395	0.074551	75	18	4.00E-6**	10	-0.566
CG5210	A	sim	2R	Intergenic; 5UTR	10	866	0.015129	0.014806	0.061207	40	33	0.0682	12	0.106
CG9973	A	sim	3L	Intergenic	10	1122	0.004382	0.005049	0.014488	13	15	0.628	5	-0.613
CG14629	A	sim	X	Intergenic; 5UTR	12	932	0.004728	0.005551	0.084424	67	15	0.553	6	-0.643
CG1468	A	sech	X	Intergenic; 5UTR	9	1068	0.011517	0.013951	0.087752	80	36	0.092	25	-0.894
CG10597	A	sech	X	Intergenic	10	587	0.011228	0.013076	0.029889	14	17	0.572	12	-0.690
CG4734	A	sech	2R	Intergenic	11	1039	0.010563	0.012495	0.050111	40	36	0.080	20	-0.730
CG7953	A	sech	2L	Intergenic; 5UTR	12	194	0.00172	0.003422	0.027117	5	2	1	2	-1.456
CG7916	A	sech	2L	Intergenic; 5UTR	12	1502	0.008115	0.007921	0.056025	71	34	8.00E-4**	15	0.112
CG8997	A	sech	2L	Intergenic; 5UTR	12	1502	0.008115	0.007921	0.056025	71	34	0.031*	15	0.112
CG33306	A	sech	2L	Intergenic; 5UTR	12	684	0.010119	0.012754	0.060885	34	25	0.282	15	-0.932
CG13061	A	sech	3L	Intergenic; 5UTR	10	941	0.00513	0.005805	0.079404	66	13	1	6	-0.541
CG9509	A	sech	X	Intergenic	12	1110	0.007773	0.010366	0.079104	76	33	0.007**	23	-1.142
CG13675	A	sim	3L	Intergenic	12	1140	0.004284	0.006114	0.033179	29	21	1	13	-1.324
CG9602	A	sim	3R	Intergenic; 5UTR	12	1072	0.003739	0.006274	0.032894	32	19	0.437	16	-1.788
CG12683	A	sim	X	Intergenic	11	1090	0.021546	0.022562	0.072814	60	65	NA	30	-0.218
CG15314	A	sim	X	Intergenic	11	1006	0.012362	0.013197	0.040282	31	35	0.328702	18	-0.299
CG14503	A	sim	2R	Intergenic	11	1160	0.008342	0.009133	0.023564	20	30	NA	15	-0.404
CG12912	A	sim	2R	Intergenic	11	1083	0.009884	0.013958	0.0894	78	39	0.035*	31	-1.387
CG5832	A	sim	3R	Intergenic	10	1025	0.003589	0.003377	0.015874	13	10	0.257	3	0.278
CG8768	A	sim	2R	Intergenic; 5UTR	10	876	0.016874	0.016297	0.055177	36	37	0.556	16	0.174
CG16916	A	sech	X	Intergenic; 5UTR	11	1057	0.010214	0.010042	0.054667	48	29	0.022*	12	0.080

											MK				
Gene	Pop	Loc	Outgroup	Loc	Region	lines	sites	π	θ	K	D5'	P5'	p-value	Singletons	Tajima's. D
CG7203	E	2L	sim	2L	Intergenic	12	1079	0.010565	0.01107	0.062127	49	36	0.783	13	-0.209
CG7214	E	2L	sim	2L	Intergenic	12	917	0.004666	0.005029	0.036879	29	11	1	5	-0.308
CG5623	E	3R	sim	3R	Intergenic	12	1132	0.0201	0.01788	0.097419	84	54	1	14	0.580
CG5402	E	3R	sim	3R	Intergenic	12	1102	0.013362	0.012731	0.049706	39	37	1.00E-04***	9	0.229
CG5178	E	3R	sim	3R	Intergenic	11	1039	0.021149	0.017802	0.055273	37	50	0.045*	3	0.900
CG5144	E	3L	sim	3L	Intergenic	11	1042	0.013213	0.009865	0.071531	59	26	0.855	7	1.586
CG8661	E	X	sim	X	Intergenic	11	994	0.013566	0.013208	0.120699	96	37	0.797	12	0.128
CG5386	E	3R	sim	3R	Intergenic	12	992	0.00731	0.009274	0.040826	32	25	0.768	11	-0.955
CG10912	E	2R	sim	2R	Intergenic	10	1039	0.003762	0.003323	0.100692	94	8	1	0	0.586
CG5154	E	2R	sim	2R	Intergenic	11	922	0.014601	0.011739	0.06366	44	31	0.719	6	1.142
CG9511	E	2L	sim	2L	Intergenic	9	980	0.009281	0.007452	0.081481	69	18	0.028*	2	1.216
CG5210	E	2R	sim	2R	Intergenic	11	790	0.013996	0.013923	0.063989	36	31	0.686	4	0.024
CG9973	E	3L	sim	3L	Intergenic	11	1084	0.003353	0.003779	0.014651	13	11	1	7	-0.491
CG14629	E	X	sim	X	Intergenic	12	944	0.001682	0.003066	0.084474	73	8	1	8	-1.836
CG1468	E	X	sech	X	Intergenic	11	1058	0.001276	0.002152	0.084497	83	7	1	6	-1.654
CG10597	E	X	sech	X	Intergenic	12	591	0.009854	0.007609	0.02772	13	9	1	1	1.315
CG4734	E	2R	sech	2R	Intergenic	10	1039	0.009935	0.010198	0.048914	42	27	0.151	12	-0.124
CG7953	E	2L	sech	2L	Intergenic	12	142	0.003315	0.004678	0.016015	2	2	0.212	1	-0.854
CG7916	E	2L	sech	2L	Intergenic	12	1401	0.00801	0.009196	0.059135	68	35	0.015*	19	-0.591
CG8997	E	2L	sech	2L	Intergenic	12	1401	0.00801	0.009196	0.058814	68	35	0.362	19	-0.591
CG33306	E	2L	sech	2L	Intergenic	11	639	0.00823	0.008298	0.05949	31	15	1	7	-0.037
CG13061	E	3L	sech	3L	Intergenic	12	921	0.006133	0.005856	0.08276	67	15	1	4	0.206
CG9509	E	X	sech	X	Intergenic	12	1115	0.000465	0.000282	0.079893	84	1	0.004**	0	1.487
CG13675	E	3L	sim	3L	Intergenic	12	1141	0.004863	0.004937	0.031853	28	16	0.765	2	-0.065
CG9602	E	3R	sim	3R	Intergenic	11	1018	0.00368	0.003875	0.036506	33	9	0.025*	4	-0.220
CG12683	E	X	sim	X	Intergenic	12	1090	0.011217	0.009045	0.075461	68	27	NA	4	1.087
CG15314	E	X	sim	X	Intergenic	11	1048	0.003031	0.002188	0.04523	43	6	0.206	0	1.566
CG14503	E	2R	sim	2R	Intergenic	11	1163	0.005093	0.003764	0.023018	21	13	NA	2	1.549
CG12912	E	2R	sim	2R	Intergenic	9	1083	0.003993	0.003304	0.090968	88	10	0.010	1	0.973
CG5832	E	3R	sim	3R	Intergenic	12	1016	0.002554	0.002871	0.017509	16	9	0.566	4	-0.449
CG8768	E	2R	sim	2R	Intergenic	12	844	0.000713	0.001113	0.060534	49	2	0.536	2	-1.180
CG16916	E	X	sech	X	Intergenic	12	935	0.000177	0.000352	0.052031	47	1	0.016*	1	-1.141

										MK test				
Gene	Pop	Loc	Outgroup	Region	lines	sites	π	θ	K	D5'	P5'	p-value	Singletons	Tajima's D.
CG7203	A	2L	sim	Intergenic	12	1056	0.017405	0.019041	0.058659	43	57	0.638	30	-0.402
CG7214	A	2L	sim	Intergenic	12	907	0.006887	0.008736	0.036524	28	20	0.445	16	-0.946
CG5623	A	3R	sim	Intergenic	11	1130	0.017346	0.016967	0.097367	82	50	0.827	17	0.1072
CG5402	A	3R	sim	Intergenic	10	1103	0.01694	0.014543	0.048014	36	40	4.00E-5	11	0.8117
CG5178	A	3R	sim	Intergenic	11	1062	0.018771	0.01546	0.057628	41	45	0.349	12	1.0208
CG5144	A	3L	sim	Intergenic	10	1036	0.016234	0.017362	0.071656	53	47	0.525	22	-0.322
CG8661	A	X	sim	Intergenic	9	977	0.018379	0.018323	0.121459	94	46	0.262	15	0.0158
CG5386	A	3R	sim	Intergenic	10	985	0.014509	0.016094	0.044007	32	40	0.747	23	-0.485
CG10912	A	2R	sim	Intergenic	10	1026	0.013415	0.017319	0.098049	82	46	0.017*	33	-1.115
CG5154	A	2R	sim	Intergenic	9	925	0.015298	0.017805	0.060645	42	42	0.617	25	-0.724
CG9511	A	2L	sim	Intergenic	9	982	0.007326	0.00815	0.078537	68	17	1.00E-5**	9	-0.504
CG5210	A	2R	sim	Intergenic	10	794	0.0154	0.0148	0.064949	39	30	0.043*	10	0.1975
CG9973	A	3L	sim	Intergenic	10	1122	0.004382	0.005049	0.014488	13	15	0.628	5	-0.613
CG14629	A	X	sim	Intergenic	12	929	0.004743	0.005569	0.084713	67	15	0.553	6	-0.643
CG1468	A	X	sech	Intergenic	9	1046	0.011184	0.01355	0.085534	77	34	0.091	24	-0.893
CG10597	A	X	sech	Intergenic	10	587	0.011228	0.013076	0.029889	14	17	0.572	12	-0.69
CG4734	A	2R	sech	Intergenic	11	1039	0.010563	0.012495	0.050111	40	36	0.080	20	-0.73
CG7953	A	2L	sech	Intergenic	12	142	0.002351	0.004678	0.015416	2	2	0.268	2	-1.458
CG7916	A	2L	sech	Intergenic	12	1438	0.008476	0.008273	0.057867	70	34	9.00E-4**	15	0.112
CG8997	A	2L	sech	Intergenic	12	1438	0.008476	0.008273	0.057867	70	34	0.032*	15	0.112
CG33306	A	2L	sech	Intergenic	12	648	0.010666	0.013444	0.061062	32	25	0.275	15	-0.933
CG13061	A	3L	sech	Intergenic	10	908	0.004813	0.005632	0.082086	66	12	0.657	6	-0.673
CG9509	A	X	sech	Intergenic	12	1110	0.007773	0.010366	0.079104	76	33	0.007**	23	-1.142
CG13675	A	3L	sim	Intergenic	12	1140	0.004284	0.006114	0.033179	29	21	1	13	-1.324
CG9602	A	3R	sim	Intergenic	12	1008	0.003968	0.006659	0.033993	31	19	0.438	16	-1.789
CG12683	A	X	sim	Intergenic	11	1090	0.021546	0.022562	0.072814	60	65	NA	30	-0.218
CG15314	A	X	sim	Intergenic	11	1006	0.012362	0.013197	0.040282	31	35	0.329	18	-0.299
CG14503	A	2R	sim	Intergenic	11	1160	0.008342	0.009133	0.023564	20	30	NA	15	-0.404
CG12912	A	2R	sim	Intergenic	11	1083	0.009884	0.013958	0.0894	78	39	0.035*	31	-1.387
CG5832	A	3R	sim	Intergenic	10	1025	0.003589	0.003377	0.015874	13	10	0.256	3	0.278
CG8768	A	2R	sim	Intergenic	10	807	0.018047	0.017232	0.057275	34	36	0.558	15	0.2322
CG16916	A	X	sech	Intergenic	11	932	0.010485	0.010284	0.050846	39	26	0.049*	11	0.091

										MK test				
Gene	Pop	Loc	Outgroup	Region	lines	sites	π	θ	K	D5'	P5'	p-value	Singletons	Tajima's D
CG7203	E	2L	sim	5'UTR	12	5	0	0	0	0	0	NA	0	NA
CG7214	E	2L	sim	5'UTR	12	88	0	0	0.011451	1	0	1	0	NA
CG5623	E	3R	sim	5'UTR	12	55	0.011101	0.012139	0.125015	6	2	0.681	0	-0.252
CG5402	E	3R	sim	5'UTR	12	28	0	0	0.115613	3	0	1	0	NA
CG5178	E	3R	sim	5'UTR	11	75	0	0	0.027152	2	0	1	0	NA
CG5144	E	3L	sim	5'UTR	11	55	0	0	0.037275	2	0	0.564	0	NA
CG8661	E	X	sim	5'UTR	11	10	0	0	0.107326	1	0	1	0	NA
CG10912	E	2R	sim	5'UTR	10	31	0	0	0	0	0	NA	0	NA
CG5154	E	2R	sim	5'UTR	11	49	0.025666	0.0212	0.086426	3	3	0.666	1	0.731
CG9511	E	2L	sim	5'UTR	9	150	0.006438	0.007109	0.052142	7	3	0.738	1	-0.363
CG5210	E	2R	sim	5'UTR	11	72	0.004559	0.004757	0.016596	1	1	1	0	-0.101
CG14629	E	X	sim	5'UTR	12	3	0	0	0	0	0	NA	0	NA
CG1468	E	X	sech	5'UTR	11	22	0	0	0.208224	4	0	1	0	NA
CG7953	E	2L	sech	5'UTR	12	52	0.003212	0.006395	0.06177	3	1	1	1	-1.148
CG7916	E	2L	sech	5'UTR	12	64	0	0	0.01579	1	0	0.428	0	NA
CG8997	E	2L	sech	5'UTR	12	64	0	0	0.01579	1	0	1	0	NA
CG33306	E	2L	sech	5'UTR	11	36	0.005068	0.009544	0.054998	1	1	1	1	-1.140
CG13061	E	3L	sech	5'UTR	12	33	0	0	0	0	0	NA	0	NA
CG9602	E	3R	sim	5'UTR	11	64	0	0	0.01579	1	0	1	0	NA
CG8768	E	2R	sim	5'UTR	12	69	0	0	0.02956	2	0	NA	0	NA
CG16916	E	X	sech	5'UTR	12	125	0	0	0.075694	9	0	NA	0	NA

										MK test				
Gene	Pop	Loc	Outgroup	Region	lines	sites	π	θ	K	D5'	P5'	Singletons	p-value	Tajima's D
CG7203	A	2L	sim	5'UTR	12	5	0	0	0	0	0	0	NA	NA
CG7214	A	2L	sim	5'UTR	12	88	0.005703	0.011375	0.014341	1	3	3	1	-1.648
CG5623	A	3R	sim	5'UTR	11	55	0.014014	0.012519	0.137485	6	2	0	0.694	0.368
CG5402	A	3R	sim	5'UTR	10	28	0.007177	0.012732	0.119791	3	1	1	0.3	-1.127
CG5178	A	3R	sim	5'UTR	11	74	0	0	0.013637	1	0	0	1	NA
CG5144	A	3L	sim	5'UTR	10	56	0.014424	0.025684	0.040352	1	4	4	0.163	-1.712
CG8661	A	X	sim	5'UTR	9	10	0	0	0.107326	1	0	0	1	NA
CG10912	A	2R	sim	5'UTR	10	31	0	0	0	0	0	0	NA	NA
CG5154	A	2R	sim	5'UTR	9	49	0.032437	0.022872	0.099258	3	3	0	1	1.622
CG9511	A	2L	sim	5'UTR	9	150	0.001426	0.002362	0.048972	7	1	1	0.058	-1.091
CG5210	A	2R	sim	5'UTR	10	72	0.012135	0.014875	0.021128	1	3	2	1	-0.669
CG14629	A	X	sim	5'UTR	12	3	0	0	0	0	0	0	NA	NA
CG1468	A	X	sech	5'UTR	9	22	0.028305	0.034218	0.201587	3	2	1	0.228	-0.608
CG7953	A	2L	sech	5'UTR	12	52	0	0	0.060032	3	0	0	0.592	NA
CG7916	A	2L	sech	5'UTR	12	64	0	0	0.01579	1	0	0	0.368	NA
CG8997	A	2L	sech	5'UTR	12	64	0	0	0.01579	1	0	0	0.466	NA
CG33306	A	2L	sech	5'UTR	12	36	0	0	0.057721	2	0	0	0.578	NA
CG13061	A	3L	sech	5'UTR	10	33	0.014276	0.010789	0.009146	0	1	0	0.273	0.834
CG9602	A	3R	sim	5'UTR	12	64	0	0	0.01579	1	0	0	1	NA
CG8768	A	2R	sim	5'UTR	10	69	0.002904	0.005141	0.03107	2	1	1	0.575	-1.118
CG16916	A	X	sech	5'UTR	11	125	0.00819	0.008239	0.083783	9	3	1	0.048*	-0.021

										MK test				
Gene	Pop	Loc	Outgroup	Region	lines	sites	π	θ	K	DI	PI	p-value	Singletons	Tajima's D
CG7203	E	2L	sim	intron	12	263	0.032714	0.02822	0.109669	21	20	1	6	0.719
CG7214	E	2L	sim	intron	12	117	0.003505	0.002836	0.069349	7	1	1	0	0.543
CG5623	E	3R	sim	intron	12	66	0.006224	0.005034	0.164522	9	1	0.127	0	0.545
CG5402	E	3R	sim	intron	12	217	0.006373	0.009972	0.199929	36	7	0.255	6	-1.422
CG5178	E	3R	sim	intron	11	538	0.010611	0.009099	0.082813	36	14	1	4	0.745
CG8661	E	X	sim	intron	11	62	0.017205	0.011095	0.163154	8	2	1	0	1.697
CG5386	E	3R	sim	intron	12	160	0.004083	0.004151	0.014196	1	2	1	1	-0.048
CG10912	E	2R	sim	intron	10	75	0.003875	0.003852	0.150059	10	1	1	0	0.015
CG5154	E	2R	sim	intron	10	112	0.016271	0.014125	0.102471	8	4	1	2	0.588
CG5210	E	2R	sim	intron	11	494	0.007236	0.005375	0.099216	43	8	0.009**	1	1.438
CG9973	E	3L	sim	intron	11	523	0.007872	0.007066	0.042371	18	10	0.324	5	0.493
CG1468	E	X	sech	intron	11	109	0.018557	0.014009	0.153056	12	5	0.080	0	1.256
CG4734	E	2R	sech	intron	10	232	0.029088	0.024356	0.106332	18	14	0.466	4	0.914
CG10597	E	X	sech	intron	12	136	0.010772	0.00734	0.048077	5	3	1	0	1.541
CG7953	E	2L	sech	intron	12	94	0.010205	0.018496	0.114727	8	5	0.246	5	-1.812
CG7916	E	2L	sech	intron	12	121	0.013453	0.010672	0.077568	8	3	0.091	0	0.923
CG8997	E	2L	sech	intron	12	64	0.01121	0.01042	0.065259	4	0	0.143	0	0.223
CG33306	E	2L	sech	intron	11	55	0.018067	0.025251	0.10447	4	4	0.427	3	-1.059
CG13061	E	3L	sech	intron	12	109	0.01025	0.01042	0.130801	12	3	1	1	-0.058
CG9509	E	X	sech	intron	12	361	0.00119	0.001679	0.082388	28	1	0.247	1	-0.851
CG13675	E	3L	sim	intron	12	221	0.021224	0.016465	0.040676	7	6	0.165	0	1.228
CG9602	E	3R	sim	intron	11	61	0.025047	0.016705	0.131617	6	3	0.444	0	1.728
CG8768	E	2R	sim	intron	12	171	0	0	0.023764	4	0	NA	0	NA
CG16916	E	X	sech	intron	12	132	0	0	0.132209	16	0	NA	0	NA

										MK test				
Gene	Pop	Loc	Outgroup	Region	lines	sites	π	θ	K	DI	PI	p-value	Singletons	Tajima's D
CG7203	A	2L	sim	intron	12	256	0.033647	0.040619	0.111771	18	28	1	18	-0.796
CG7214	A	2L	sim	intron	12	117	0.013327	0.020078	0.06623	5	6	1	5	-1.334
CG5623	A	3R	sim	intron	12	66	0.016714	0.015682	0.170982	9	3	0.714	0	0.228
CG5402	A	3R	sim	intron	10	232	0.016224	0.01416	0.197593	38	8	0.254	4	0.652
CG5178	A	3R	sim	intron	9	538	0.010427	0.010818	0.083568	37	16	0.596	7	-0.164
CG8661	A	X	sim	intron	11	62	0.025055	0.030275	0.165791	8	4	0.504	3	-0.745
CG5386	A	3R	sim	intron	10	160	0.001251	0.002213	0.013241	2	1	1	1	-1.114
CG10912	A	2R	sim	intron	10	68	0.017667	0.012732	0.149197	8	3	0.392	0	1.406
CG5154	A	2R	sim	intron	9	97	0.024809	0.024465	0.077145	6	4	1	0	0.064
CG5210	A	2R	sim	intron	10	494	0.014293	0.01835	0.097118	34	18	0.006**	14	-1.068
CG9973	A	3L	sim	intron	10	517	0.011152	0.01011	0.039922	16	15	0.349	4	0.478
CG1468	A	9	sech	intron	9	109	0.012295	0.012062	0.13232	12	3	0.701	1	0.079
CG4734	A	2R	sech	intron	11	232	0.025119	0.026389	0.11268	17	16	0.159	5	-0.220
CG10597	A	X	sech	intron	10	136	0.014519	0.015759	0.044688	5	5	1	3	-0.329
CG7953	A	2L	sech	intron	12	94	0.014241	0.016159	0.112664	8	3	1	1	-0.469
CG7916	A	2L	sech	intron	11	121	0.015307	0.010672	0.078332	8	3	0.0398*	0	1.538
CG8997	A	2L	sech	intron	12	64	0.018697	0.015685	0.06668	4	1	0.335	1	0.637
CG33306	A	2L	sech	intron	12	55	0.022369	0.030724	0.114351	5	5	0.283	4	-1.023
CG13061	A	3L	sech	intron	10	109	0.003132	0.005544	0.140881	13	2	1	2	-1.409
CG9509	A	X	sech	intron	12	359	0.013506	0.013853	0.089906	25	16	0.336	5	-0.109
CG13675	A	3L	sim	intron	12	220	0.030737	0.026349	0.04689	7	13	0.313	4	0.742
CG9602	A	3R	sim	intron	12	62	0.027681	0.021325	0.130317	6	4	0.712	1	1.063
CG8768	A	2R	sim	intron	10	171	0.021632	0.024929	0.039625	2	12	0.102	7	-0.607
CG16916	A	X	sech	intron	11	132	0.007753	0.010418	0.137155	16	4	0.003**	2	-0.943

Appendix F

ΔCt values obtained by qPCR of male samples

Strains	CG8997	CG7916
E01	1.202	2.987
E12	1.871	2.075
E14	1.739	2.971
E15	1.524	1.737
E16	1.957	2.525
E17	0.067	4.042
E18	2.654	4.101
E20	3.105	3.157
A82	-1.092	1.722
A84	-1.336	1.603
A95	1.568	2.312
A131	0.87	3.655
A186	0.86	1.805
A377	0.817	1.48
A384	1.028	1.942
A398	0.454	1.624

Curriculum Vitae

Sarah Sylvie SAMINADIN-PETER
January 14th 1979
French

CONTACT

Ludwig-Maximilians-Universität München
Department Biologie II
Großhaderner Str. 2
82152 Planegg-Martinsried Germany
00-49-89-2180-74-109
peter@zi.biologie.uni-muenchen.de

RESEARCH INTERESTS

Genomics, molecular evolution, gene expression and regulation, statistics, genetics, ecology

EDUCATION

Ludwig-Maximilians- University, Munich, Germany
July 2005-current
PhD. Evolutionary Biology, supervisor: Prof. John Parsch

University of Tours, France
2003
Master's degree in Evolution and Control of populations

University of Marseille, France
2001
Bachelor's degree in Biology of Marine Organisms

RESEARCH EXPERIENCE

Research assistant **January - June 2004**
Institute of Genetics and Molecular and Cellular Biology (IGBMC)
Strasbourg, France
"Isolation and functional studies of proteins mutated in Bardet-Biedl syndromes"
Supervisors: Dr. Didier Devys and Prof. Jean-Louis Mandel

Master Thesis **January - August 2003**
IFREMER, Department of Genetics and Pathology
La Tremblade, France
"Inter and Intraspecific molecular evolution in the oyster, *Crassostrea gigas* protease inhibitor Cg-TIMP"
Supervisor: Dr. Timothy Sharbel

Master Fellow **March - May 2002**
Marine Biology Station of Concarneau
Concarneau, France
"Populations genetics of Clam species *Ruditapes philippinarum* and *R. decussatus* from south Brittany coast"
Supervisor: Prof. Alain Van Wormhoundt

Publications and Presentations

Hutter, S., **S. S. Saminadin-Peter**, W. Stephan and J. Parsch (2008) "Gene expression variation in African and European populations of *Drosophila melanogaster*" Genome Biology 9:R12, 2008

"Gene expression variation in natural populations of *Drosophila melanogaster*" talk at 20th European Drosophila Research Conference, Vienna, Austria, September 2007

"Gene expression variation in natural populations of *Drosophila melanogaster*" talk at « Petit Pois Déridé », Poitier, France, August 2007

"Evolutionary and functional genomics of *Drosophila* gene expression" talk at Theoretical Genetics Molecular Evolution and Diversity Summer School, Edinburgh, UK, July, 2006

"Evolution of duplicated Cg-TIMP gene in the oyster *Crassostrea gigas*: can we use the signature of natural selection to identify loci involved with parasite resistance?" talk at the 7th Evolutionary Biology Meeting, Marseille, France, June, 2003

Acknowledgements

I would like to thank Prof. John Parsch for giving me the opportunity to work under his supervision. With his help I was able to expand my understanding of evolutionary genomics and improve my scientific writing. I will always have fond memories of our visits to Octoberfest. Also, to John Baines who was available to answer my questions and who was there to give me advice on my experimental setup and procedure. I would like to extend my gratitude to Prof. Wolfgang Stephan and the members of my committee for their insightful comments that helped to advance my project. Thank you to my colleagues in the Parsch group: Matthias Proeschel, Winfried Hense, Zhi Zhang who showed me how to work with microarray experiments and how to cook the Mapu Tofu, Xiao Liu for our lively chat and Claus Kemkemer for his always pleasant mood and his great advice.

I shared many memories with the other Drosophilist people, past and present members: Lino Ometto, Sacha Glinka, Steffen Beisswenger, Stephan Hutter, Annegret Werzner, Nicolas Svetec. They took the time to explain to me the genome scan and the sweep and how to work with our dear *Drosophila melanogaster* and its two famous populations.

I have to give a special thanks to Stefan Laurent, Pavlos Pavlidis and Sergej Nowoshilow for their great help with the simulations, programming problems and sequence analysis. I don't forget the two bioinformatitions students who collaborated on the microarray analysis Annahita Oswald and Anna Bauer-Mehren.

The work atmosphere would not be the same without the Tomato's group: Laura Rose, Anja Hörger, Iris Fischer, Carlos Merino-Mendès, Lukas Grzeskowiak, Hilde Lainer, Simone Lange, Letizia Camus-Kulandaivelu, Aurelien Tellier and Robert Piskol the RNA's specialist. In addition, this job would never be finished on time without the technical assistance from Beate Stiening for the microarrays and Yvonne Cämmerer and Hedwig Gebhart for the PCR and sequencing. You did an amazing job on this project. And finally, I would like to thank Simone Lange for her helpful technical advice.

Thanks to Anne Wilken and Anica Vrljic for all the washing, cleaning and preparation of food. To Katrin Kümpfbeck who helped to deal with the German administration with her contagious enthusiasm.

Critical comments and brain storming are always necessary to come out with new ideas or to forget about the bad ones. This was done in French, of course, with Letizia, Aurélien and Stefan. But when things are not going well, then you need a shoulder to lean on,

thank you to Letizia and Carlos who always listened to my complaints. Thank you also Constanze Schmit for your good job in the lab. A special thanks also to Anna Gabrenya and Rebecca Meredith who correct my English mistakes.

They are not members of the Evolution group anymore, but they brought me a lot: Traudl, for our great exchange, particularly about culture and literature; Ann who showed me around Bayern and helped me to appreciate this region; Thomas who shared with me his interest in Andechs; David for the Latin nightlife; people from the former Joachim Hermisson group, Nina for the hiking fun and Pleuni for the creative plans and the invitation to be a part of Volvox creation.

I would like to give a special thanks to the members of Volvox who help us by their enthusiasm to organize things and the organization members with particular thanks for Rita Verma. You did a good job and keep going.

I would like to express my appreciation for the Casodom Organization that awarded me with a grant for my success in biological studies. Their grants support students from the outside regions of France (DOM) and try to promote success in the school systems and our integration in the French society. I cannot forget to acknowledge the Guadeloupe Region for its support by way of the president Mr. Victorin Lurel.

Thank to my Munich friends and particularly Mite and Poupou who improved my PhD life with barbecues, biergartens and parties. I would like to thank my friends and relatives, who despite the distance, are always there for me, particularly: Thierry, Mimi and Christel.

Lots of love to my mother and my brother Remy who sent me their support and a great deal of positive encouragement. Mèci dè toujou ba mwen fòs la.

Finally, to Kiki, who shares my successes, my hopes and my doubts. Thanks to you for always caring for me and following my dreams.