A genome-wide scan for genes under balancing selection in *Drosophila melanogaster*

Dissertation

an der Fakultät für Biologie

der Ludwig-Maximilians-Universität

München

vorgelegt von

Myriam Florence Croze

aus Montélimar

München, Januar 2017

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

Tag der Einreichung: 17.01.2017

Tag der mündlichen Prüfung: 13.03.2017

Erklärung:

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

Eidesstattliche Erklärung:

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt ist.

München, den 17.01.2017

Myriam Croze

Table of contents

Note 7

List of abbreviations 9

List of figures 10

List of tables 11

CHAPTER 1: INTRODUCTION 14

1.1 History of the main theories 14

1.2 Characteristics of balancing selection 15

1.2.1 Definition of balancing selection 15

1.2.2 Methods to detect balancing selection 18

1.3 Evidence of balancing selection 21

1.3.1 Balancing selection in immunity 21

1.3.2 Examples of balancing selection 22

1.4 Balancing selection in *Drosophila melanogaster* 25

1.4.1 *Drosophila melanogaster* as model organism 25

1.4.2 Evidence of balancing selection in *Drosophila* 27

1.5 Aim of the project 29

CHAPTER 2: MATERIALS AND METHODS 30

2.1 Sequence data 30

2.2 Window-size analysis 31

2.3 Coalescent simulations 32

2.3.1 Joint effects of selection and demography 32

2.3.2 Genome-scan analysis 33

2.3.3 GO enrichment analysis 35

2.3.4 Linkage disequilibrium analysis 36
2.3.5 Trans-species polymorphisms 37
2.4 Analysis of candidate genes 37
CHAPTER 3: RESULTS 39
3.1 Genome-scan for balancing selection 39
3.1.1 Choice of the method to detect balancing selection 39
3.1.2 Choice of the window-size for the genome scan 41
3.1.3 Genome-scan analysis 44
3.1.4 Candidate genes 46
3.1.5 GO terms 49
3.2 Analysis of the candidate genes *chm and CG15818* 55
CHAPTER 4: DISCUSSION 65
4.1 Detection of footprints of balancing selection 65
4.2 Evidence of balancing selection in *D. melanogaster* 67
4.2.1 Candidate genes 67
4.2.2 Function of the candidate genes 69
4.3 Balancing selection in immunity 70
4.4 The candidate genes *chm* and *CG15818* 73
4.5 Conclusion and perspectives 75
APPENDIX A: MATERIALS AND METHODS 78
APPENDIX B: RESULTS 81
BIBLIOGRAPHY 99
ACKNOWLEDGEMENTS 116

Note

In this thesis, I present my doctoral research about balancing selection in *Drosophila melanogaster*. All the analytical work has been done by me except for the following: Daniel Živković and Andreas Wollstein determined the demographic model for the European and African populations. Vedran Božičević and I did the Gene Ontology analysis. Pablo Duchén provided the PERL script to extract SNP information from the *D. simulans* data and to do the Linkage Disequilibrium analysis.

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

Croze M., D. Živković, W. Stephan and S. Hutter (2016). "Balancing selection on immunity genes: review of the current literature and new analysis in *Drosophila melanogaster"*. *Zoology* 119: 322-329.

Croze M., A. Wollstein, V. Božičević, D. Živković, W. Stephan and S. Hutter (2017). "A genome-wide scan for genes under balancing selection in *Drosophila melanogaster*". (Accepted in *BMC Evolutionary Biology*).

List of abbreviations

bp	base pair
CLECT	C-type lectin / C-type lectin-like
DNA	deoxyribonucleic acid
DPGP	*Drosophila* Genomics Project
F_{ST}	genetic differentiation
GO	gene ontology
HAT	Histone acetyltransferase
HKA	Hudson Kreitman Aguade test
JNK	c-Jun N-terminal kinases
LD	linkage disequilibrium
MHC	major histocompatibility complex
MK	McDonald Kreitman test
Ne	current population size
NGS	next generation sequencing
NS	nonsynonymous SNP
SFS	site frequency spectrum
SNP	single nucleotide polymorphism
TSP	trans-species polymorphism

List of figures

Figure 1: Schematic representation of the signatures of balancing selection……p.16

Figure 2: Demographic models for the autosomal chromosomes of the European (A) and African (B) populations……..p.33

Figure 3: The site frequency spectrum (SFS) under balancing selection……..p.39

Figure 4: Proportion of candidate windows as a function of window size (in bp) for each chromosome (2L, 2R, 3L, 3R and X) in the African and the European populations……p.41

Figure 5: Power analysis for different window sizes (in bp)……p.43

Figure 6: Map of the genes *CG15818* and c*hm*..……p.55

Figure 7: Schematic representations of the domains present on the proteins *CG15818* and *chm*……p.56

Figure 8: Polymorphism table of the candidate region in Africa……...p.58

Figure 9: Polymorphism table of the candidate region in Europe……..p.59

Figure 10: Representation of LD (r^2) for the two candidate regions……p.61-62

List of tables

Table 1: Statistical values of the mean of θ_w and Tajima's *D* for each chromosome and population……...p.44

Table 2: List of candidate genes shared by the African and European populations……...p.47

Table 3: List of enriched GO terms for the European population……..p.50-51

Table 4: List of the best candidate genes for the European and African populations with a *p*-value < 10^{-4} for θ_w and Tajima's *D*……...p.53

Summary

In the history of population genetics balancing selection has been considered as an important evolutionary force, yet until today little is known about its abundance and its effect on patterns of genetic diversity. Several well-known examples of balancing selection have been reported from humans, plants, and parasites. However, only very few systematic studies have been carried out to detect genes under balancing selection. We performed a genome scan in *Drosophila melanogaster* to find signatures of balancing selection in a derived (European) and an ancestral (African) population. We screened a total of 34 genomes searching for regions of high genetic diversity and an excess of SNPs with intermediate frequency. In total, we found 183 candidate genes: 141 in the European population and 45 in the African one, with only three genes shared between both populations. Most differences between both populations were observed on the X chromosome, though this might be partly due to false positives. Functionally, we find an overrepresentation of genes involved in neuronal development and circadian rhythm. Furthermore, some of the top genes we identified are involved in innate immunity. Finally, we decided to study in more details two of our best genes (*chm* and *CG15818*) in order to see if we observe other patterns of balancing selection in our candidate genes. At the protein level, we found evidence of polymorphisms (including non-synonymous polymorphisms) at intermediate frequency in linkage disequilibrium (LD). In addition, we also found haplotype structure in the European and African populations. These results confirm that these genes are effectively under balancing selection and that our method allowed us to detect genes under balancing selection.

CHAPTER 1

INTRODUCTION

1.1 History of the main theories

At both phenotypic and genetic levels, large diversity is observed in natural populations. Although a high level of genetic polymorphism has been observed in most species, the reason for the maintenance of this genetic variation is still unclear. Based on this observation, Dobzhansky (1955) proposed a model called “balanced hypothesis” which suggests that many genes are polymorphic and that these polymorphisms are maintained by heterozygote advantage (also called overdominant selection). This model was opposed to the classical view of Muller (1958), who believed that individuals in a population are homozygous at most loci. For him, deleterious alleles are removed by natural selection and the main force acting on the genome is purifying selection. These two models were debated during several years until new methods and technologies such as protein electrophoresis showed a high variability in natural populations of humans and *Drosophila pseudoobscura* (Lewontin and Hubby 1966). Lewontin and Hubby proposed heterozygote advantage as a possible explanation, although balancing selection alone cannot explain all the high variability observed in the genome (Lewontin 1974), particularly due

to the segregating load (i.e. deleterious mutation on an individual will induce a reduction of its fitness).

To explain this high variability observed in the genome, Kimura (1968) proposed the neutral theory of molecular evolution. This theory states that the majority of polymorphisms are neutral or nearly neutral and that they are maintained through the joint action of mutation and random genetic drift rather than selection. In the 1980s, the study of genetic diversity and molecular evolution moved to the DNA level. Thanks to the advent of these new technologies and molecular population genetics, the neutral theory was rigorously tested (Kreitman 1983; Hudson *et al.* 1987). This led to the conviction that the neutral theory alone cannot explain the observed patterns of DNA polymorphism within populations and the divergence between species. Natural selection has to be invoked to explain the patterns observed.

Balancing selection, which maintains genetic diversity within populations, is one of these selective forces. However, only few studies have identified this type of selection at the DNA level. Therefore, balancing selection is thought to be rare and specific only to some classes of genes like those related to immunity (Asthana *et al*. 2005; Andrés *et al*. 2009; Quintana-Murci and Clark 2013).

1.2 Characteristics of balancing selection

1.2.1 Definition of balancing selection

Balancing selection is described as a selective force which maintains genetic variation within a population (Stahl *et al.* 1999). Three different mechanisms are known for balancing selection: (i) heterozygote advantage (or overdominant selection), (ii) negative frequency-dependent selection, and (iii) spatio-temporally fluctuating selection. In the case of overdominant selection, heterozygote alleles are maintained in the genome because heterozygous individuals

have a selective advantage over both homozygotes. Concerning negative frequency-dependent selection, a rare allele will be favored and its frequency will increase until it reaches equilibrium or it starts to be selected against. Finally, for the last mechanism the frequency of an allele will depend on the environment and time. An allele will be favored in one habitat or under certain environmental conditions but disfavored in another. For example the fluctuation of pathogens in an environment or the climate can change the frequency of alleles (Asthana *et al.* 2005; Charlesworth 2006; Charlesworth and Charlesworth 2010, Bergland *et al.* 2014). However, these different kinds of mechanisms will produce similar polymorphism patterns and consequently, it will be difficult to differentiate them.

However, balancing selection can act on different timescales from very long-term (in different species) to short-term (only within a population) selection, and following how long the selection has been acting, the signatures observed in the genome will be different (Woolhouse *et al.* 2002; Brown and Tellier 2011; Fijarczyk and Babik 2015). Recent balancing selection is difficult to detect because it will generate signatures similar to positive selection. It will increase linkage disequilibrium around the selected locus and reduce differentiation among populations which is similar of what is observed in the case of incomplete sweeps (Hermisson and Pennings 2005). Consequently, it is hard to differentiate recent balancing selection from positive selection. Older balancing selection will change the genealogies of the gene under selection compared to genealogies under neutrality. In this case, balancing selection will produce longer internal branches (Hein *et al.* 2005) and increase the diversity around the target of selection by hitchhiking such that an excess of alleles at intermediate frequency compared to neutrality will be observed (Charlesworth 2006; Charlesworth and Charlesworth 2010). Concerning ancient balancing selection, one of its main characteristics is the presence of trans-species polymorphism (TSP). In this case, a polymorphism, which appeared before the time of species divergence, will be maintained in two or more species (Klein 1987). Recently, several studies have looked for evidence of balancing selection in different organisms and different methods have been developed to look for features of balancing selection.

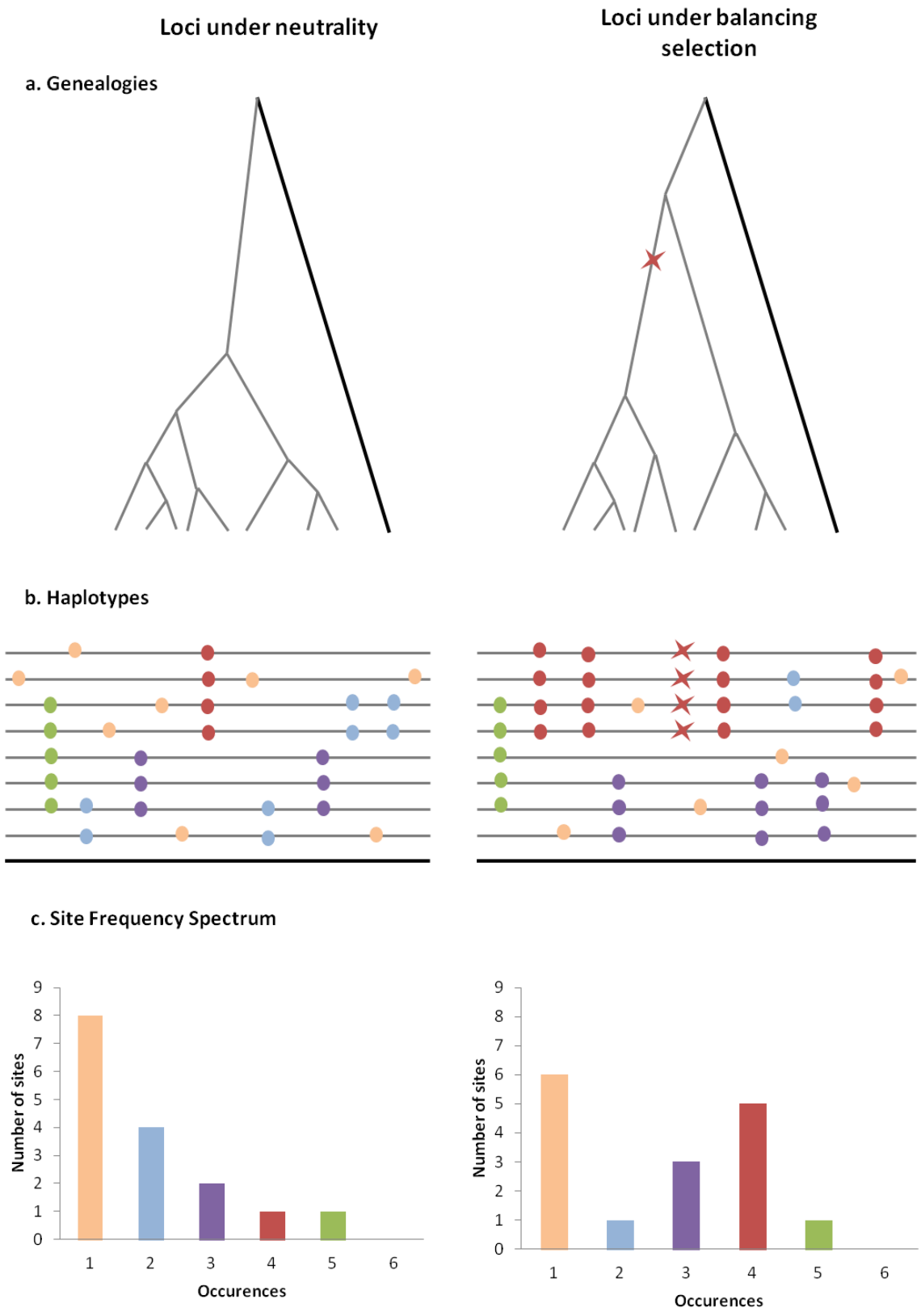

Loci under neutrality
Loci under balancing selection
a. Genealogies
b. Haplotypes
c. Site Frequency Spectrum
Number of sites
Occurences
0
1
2
3
4
5
6
7
8
9

Figure 1: Schematic representation of the signatures of balancing selection. a. The genealogies of loci under neutrality (left) and balancing selection (right) are represented. The grey lines correspond to the ingroup, the black line represents the outgroup and the red star represents the appearance of a selected variant in the genealogy. **b.** Haplotypes of each locus containing mutations that have accumulated in each genealogy. For each locus, the number of segregating sites is 16. **c.** The site frequency spectrum of each locus under neutrality (left) and balancing selection (right). Balancing selection can result in an excess of intermediate-frequency variants (purple and red).

1.2.2 Methods to detect balancing selection

The development of new technologies as well as the application of new statistical methods has allowed identifying genomic regions that appear to be shaped by natural selection (Bamshad and Wooding 2003; Nielsen *et al.* 2007; Vitti *et al.* 2013; Fijarczyk and Babik 2015). One of the easiest ways to detect balancing selection is to look for departures from a standard neutral model (Charlesworth and Charlesworth 2010). This can be assessed using several population genetics statistical tools to screen genomes to detect footprints of balancing selection.

The main characteristic of balancing selection is a high diversity around the selected locus and an excess of polymorphisms at intermediate frequency. We can use different statistics such as π or θ_w to detect region with high variability. The estimator π will calculate the average number of pairwise differences between two sequences chosen at random from a sample of sequences (Tajima 1983). The Watterson's estimator (θ_w; Watterson 1975) corresponds to the average number of segregating sites in a sample. Another method is the HKA test (Hudson *et al*. 1987), which compares the level of polymorphism within a species and the divergence between species for several loci. Under neutrality, the ratio of the level of polymorphism to divergence is expected to be similar for several regions. In case of balancing selection, the level of polymorphism should be higher than the divergence and significantly different from other regions of the genome.

The excess of polymorphisms at intermediate frequency can be detected with neutrality tests such as Tajima's *D* (Tajima 1989). This statistical test is based on the comparison of π and θ_w and measures deviations of the site frequency spectrum (SFS), which is the distribution of the number of times a derived allele is observed in a sample of DNA sequences, from the neutral expectation. A positive Tajima's *D* means that we have an excess of intermediate frequency alleles whereas under neutrality, it is supposed to be equal to zero. However, Tajima's *D* and the HKA test are unable to distinguish selection from the influences of demography and population structure, which can produce polymorphism patterns similar to those under balancing selection. Consequently, it is important to estimate a demographic model and use it as a null hypothesis (Ferrer-Admetlla *et al.* 2008; Andrés *et al.* 2009; Barreiro *et al.* 2009; Thomas *et al.* 2012, Quach *et al.* 2013) and look for deviations from the neutral model.

Thus, Andrés *et al*. (2009) analyzed 13,400 human genes using methods based on the HKA test and Tajima's *D* to detect evidence of balancing selection. They inferred a demographic history for their populations and used it as a null model. Then, neutrality tests were performed for each gene and genes significantly different from expectations under the null model were considered as candidates for balancing selection. This method is conservative and a very low number of false positives are expected (see Andrés *et al*. 2009). High recombination rates pose a further challenge in the application of these tests as they will confine selection signatures to a very narrow region around the target of selection potentially making them hard to detect. Recently, two likelihood-based methods have been developed to detect signatures of ancient balancing selection (DeGiorgio *et al*. 2014). The first method looks for the spatial distribution of polymorphisms and substitutions around a selective site. The second test is based on the allele frequency surrounding the polymorphic site.

To improve the performance and the detection of balancing selection, it is important to use multiple statistics and methods. We mention here a few additional methods and statistics which can help to confirm that a genomic region is effectively under balancing selection. Excess of nonsynonymous polymorphisms is also a signature of balancing selection. We can infer excess

of nonsynonymous polymorphism thanks to the McDonald-Kreitman test (MK test) (McDonald and Kreitman 1991). The MK test compares the ratio of nonsynonymous to synonymous differences of polymorphism compared to divergence. Under neutrality, these rates should be equal whereas a bigger value within species may suggest balancing selection (Vitti *et al.* 2013; Fijarczyk and Babik 2015).

Another footprint is low differentiation between populations. The F_{ST} (Wright's fixation index) is based on the variance of allele frequencies within and between populations. Small values of F_{ST} indicate that the locus being compared is homogenous across populations, which may be indicative of long-term balancing selection (polymorphisms shared between populations). Recently, a new method has been developed and implemented into the software "BayeScan" to determine outliers based on the F_{ST} values of polymorphisms. It is a good method to test specific hypotheses concerning individual genes or sets of genes, but it performs poorly in detecting balancing selection (Beaumont and Balding 2004).

Another well known method to identify genes under balancing selection is to look for TSP. Ancestral shared polymorphisms or trans-species polymorphisms are signatures of ancient balancing selection. A balanced polymorphism can be maintained for a very long time (i.e. before the speciation between two species) and be shared among species (Klein *et al.* 1998). This signature can be used to identify individual nucleotide sites sharing a polymorphism for a very long time (e.g. between humans and chimpanzees (Ségurel *et al.* 2012; Leffler *et al.* 2013).

Concerning recent balancing selection, linkage disequilibrium (LD) based methods could be used to identify this type of selection. Polymorphisms will be in LD (i.e. non-random association of alleles in haplotypes) around the selective site and haplotypes (i.e. the combination of polymorphisms in a genomic region) will cluster by allelic type rather than populations or species. However, this feature is comparable to the signal of an incomplete sweep and it will therefore be difficult to unambiguously identify balancing selection. Moreover, recombination

will rearrange the region around the selective site and it will be difficult to observe LD in species with high recombination rate as it is the case in *Drosophila*.

All these methods have some limitations to detect balancing selection such as low power and/or high false positive rates when they are used individually. Moreover, most of them have been designed to target ancient balancing selection. On the other hand, the detection of recent balancing selection is more difficult because the signatures it generates are subtle and this may lead to the underestimation of the frequency of genes under balancing selection. One of the possible solutions to detect candidate genes is to combine tests looking for various genetic patterns that are expected under balancing selection (Andrés *et al.* 2009; Nygaard *et al.* 2010; Ochola *et al.* 2010; Thomas *et al.* 2012).

1.3 Evidence of balancing selection

1.3.1 Balancing selection in immunity

Even if balancing selection seems to be rare, the majority of genes found to be subjected to this selection are involved in immunity. Several explanations have been proposed, including the coevolution between host and parasite. Indeed, immune genes are subjected to more constraint and have to evolve rapidly due to host-parasite interactions. Coevolution describes a process in which different species reciprocally affect the evolution of each other. For instance, parasites will act on resistance alleles in infected hosts and, in return, the parasite will try to escape recognition by the host (Woolhouse *et al*. 2002). Two major types of evolutionary dynamics have been described for host-parasite coevolution: arms race and trench warfare (Ebert 2008). Arms race induces fixation of alleles, and polymorphisms generated by mutation will be in a transient state until they go to fixation (Bergelson *et al*. 2001; Magwire *et al*. 2011; Bangham *et al.* 2007). Trench warfare dynamics maintain several alleles at intermediate frequencies in a population and coevolutionary cycles may be observed (Stahl *et al*. 1999; Gokhale *et al*. 2013;

Tellier *et al*. 2014). These two dynamics are driven by positive directional selection (arms race) or balancing selection (trench warfare).

When infection occurs, an immune response is triggered and the first defense is called innate immunity which is activated immediately after infection. Host receptors recognize non-self parasite molecules triggering different non-specific defense mechanisms. Innate immunity is present in plants (Jones and Dangl 2006) and animals (Kimbrell and Beutler 2001). In vertebrates, an additional immunity system is present, the adaptive (or acquired) immune system. In this case, the response against pathogens will be specific due to particular cells (B and T lymphocytes) that recognize a specific motif of the pathogen, and it is activated later (Medzhitov and Janeway 1997). Both immune systems of vertebrates evolve rapidly and are the target of selection due to the selective pressure of pathogens (Woolhouse *et al.* 2002). The system of recognition between host and parasite genes is suggested to follow a matching-allele model (Frank 1992; Little *et al.* 2006), in which the recognition allele in the host will match one parasite allele. Another model that has been particularly used in the plant literature is the gene-for-gene model (Thompson and Burdon 1992) where the host-parasite interactions will depend on the genotypes of the two species. These models will induce reciprocal changes in hosts and parasite populations.

Effects on genetic variation at immunity genes caused by host-parasite coevolution are well documented in vertebrates and particularly in humans (Bernatchez and Landry 2003; Eizaguirre *et al.* 2012; Spurgin and Richardson 2010).

1.3.2 Examples of balancing selection

Even if balancing selection is an important force driving the evolution of genes involved in immune function and host-pathogen interactions, its signatures remain uncommon in the genomes. For instance, genome-wide analyses (Bubb *et al*. 2006; Andrés *et al.* 2009; DeGiorgio

et al. 2014) have found that the number of genes under balancing selection is low and that many of them are related to immune genes.

One of the most famous examples of balancing selection is sickle-cell anemia, which is a case of heterozygote advantage (Hedrick 2011). The name of this disease comes from the sickle shape of red blood cells which is due to a mutation in the hemoglobin gene. This mutation causes a deficiency in the oxygen transport, but at the same time it confers resistance to malaria. Malaria is an infectious disease caused by the parasite *Plasmodium falciparum* (Hill *et al.* 1997). The heterozygote genotype will have an advantage in regions where malaria is common because it is less susceptible to malaria and leads to a less severe sickle cell disease. Consequently, heterozygote individuals will have higher fitness compared to homozygotes and will be selected for.

Another well-known examples are the Major Histo-Compatibility system (MHC) genes in vertebrates (Schierup, 2001; Kelley *et al.* 2005; Piertney and Oliver 2005; Eizaguirre *et al*. 2012; Spurgin and Richardson 2010) also known as the HLA system (Human Leukocyte Antigen) in humans. The MHC genes are involved in immune response in vertebrates. These genes encode for surface antigens which are involved in the first step of the immune response identifying the foreign proteins. It has been shown that several MHC loci are highly polymorphic, most likely because of their function. We know examples of MHC genes under overdominant and also under negative frequency-dependent selection. Moreover, these genes show evidence of ancient balancing selection by the presence of TSP in humans (Leffler *et al.* 2013; Andrés *et al.* 2009) but also in mammal, fish and bird species (Klein *et al.* 2007).

TSPs have not only been found in the MHC genes but also in other genes related to immunity such as pattern recognition receptors (Tesicky and Vickler 2015), cell migration genes (Fumagalli *et al.*, 2012), an autoimmunity-related gene LAD1 (Teixeira *et al.* 2015), two antiviral genes (ZC3HAV1, Cagliani *et al.* 2012, and TRIM5, Newman *et al.* 2006; Cagliani *et al.* 2010), host defense genes (Hollox and Armour 2008, Hellgren and Sheldon 2011), and ABO blood

group genes (Fumagalli *et al.* 2009; Segurel *et al.* 2012). As shown by these examples, polymorphisms can be maintained for millions of years due to the selective pressure of pathogens on the host.

The plant immune system is another rich source of loci under balancing selection (Michelmore and Meyres 1998; Holub 2001; Van der Hoorn *et al.* 2002; Meyers *et al.* 2005). For example, the R-genes loci, which are involved in pathogen recognition are under negative frequency-dependent selection in *Arabidopsis thaliana* (Stahl *et al.* 1999) and tomato (Hörger et al. 2012). In parasites, genes related to host-parasite interactions have been also found to be under balancing selection (Ochola *et al.* 2010; Amambua-Ngwa *et al.* 2012; Thomas *et al.* 2012). All these studies show that a disproportionate number of genes under balancing selection are involved in immune processes and host response to pathogens in plants.

Andrés *et al.* (2009) did a genome scan in two human populations (African and European Americans) and have found 60 out of 13,400 human genes which significantly rejected the neutral model and showed signatures of balancing selection. A large number of these genes are related to immunity such as genes involved in MHC functions. Several genes are involved in others functions such as genes encoding membrane channels or keratin genes. However, these genes may play a role during infection like controlling the response to infection. Other studies also found evidence of balancing selection in human immune genes and particularly in innate immunity such as interleukin genes (Ferrer-Admetlla *et al.* 2008; Fumagalli *et al.* 2009).

We have also to take into account that host defense genes are not only under balancing selection but some of them show signatures of positive selection. Several genome-wide analyses have been performed in humans to look for evidence of genes under natural selection (Sabeti *et al.* 2006; Nielsen *et al.* 2007; Akey *et al.* 2009). More than 300 genes are related to immunity showing signatures of positive directional selection (Barreiro and Quintana-Murci 2010). Evidence of positive selection has been found in MHC genes but also genes of innate immunity

such as the pattern recognition receptors (Sironi *et al.* 2015). Positive selection at these genes is likely due to a recent adaptation to pathogens.

Even if a large number of genes found under balancing selection are related to immunity, we also know examples of genes involved in others functions. For example, olfactory receptors have been proposed to evolve under balancing selection (Alonso *et al*. 2008) as well as the locus controlling the color vision in New World monkeys (Hiwatashi *et al*. 2010). In plants, genes involved in reproduction such as the self-incompatibility systems (S-locus) (prevents inbreeding in angiosperms; Wright 1969; Charlesworth 2006; Roux *et al.* 2013) have the presence of trans-species polymorphisms (Delph and Kelly 2013).

While we have some knowledge of balancing selection in humans, plants and parasites, very little is known about this type of selection in other model organisms. *D. melanogaster* has been a genetic model organism for one century. However, almost no analysis on balancing selection has been carried out in this species.

1.4 Balancing selection in *Drosophila melanogaster*

1.4.1 *Drosophila melanogaster* as model organism

Drosophila melanogaster is a model organism to study immunity and evolution in invertebrates and consequently many genetic and molecular tools are available for this species. This species has been studied for many years and its genome is well-characterized and full genomes for various populations are available (http://www.dpgp.org/). *D.* melanogaster has spread around the world and its demographic history is well known (Duchen et al. 2013; Laurent *et al*. 2011). Moreover, it is an animal with many molecular pathways and protein types similar to humans (Adams *et al.* 2000) and a complex immune system which is also well-known. Finally, it

can be infected by different kinds of parasites (viruses, bacteria, and fungi), which will induce different immune responses (Paparazzo *et al.* 2015).

In *Drosophila*, there are two kinds of immune systems: the humoral response and the cellular response (Hoffmann and Reichhart 2002; Hoffmann 2003; Lemaitre and Hoffmann 2007; Leulier and Lemaitre 2008). The immune response will be triggered by the detection of non-self molecules (pathogen molecules) by host pattern recognition receptor proteins. Following the type of pathogens, different pathways will be activated. Fungi and Gram-positive bacteria recognition triggers the activation of the Toll signaling pathway (Lemaitre and Hoffmann 2007) whereas Gram-negative bacteria recognition triggers the Imd signaling pathway. This induces the production of different antimicrobial peptides by the fat body. Furthermore, other immune genes are activated in response to an infection, such as the JAK/STAT and the JNK pathways, which seem to play additional roles (Boutros *et al.* 2002). The cellular immune system is characterized by the phagocytosis of microbes, coagulation at the wound and the cellular encapsulation of larger foreign material. Recently, another *Drosophila* innate immunity pathway was discovered: antiviral RNA interference (Wang *et al.* 2006; Obbard *et al.* 2009; Saleh *et al.* 2009). It has been shown that this pathway protects *Drosophila* from virus infection. Another kind of immunity is the epithelium barrier which is in contact with many microorganisms. For example the gut epithelium is in contact with the commensal flora and has to deal with bacterial tolerance and infection. In case of infection, it will produce reactive oxygen species and anti-microbial peptides. Several studies identified hundreds of genes whose expression changes after an infection with bacteria, viruses or fungi (De Gregorio *et al*. 2001, 2002; Irving *et al*. 2001; Carpenter *et al*. 2009; Cordes *et al*. 2013; Lu *et al*. 2015). The majority of these genes are involved in immunity, but many genes are unknown or involved in different functions, for example in cytoskeleton functions (Irving *et al*. 2002), behavioral traits or metabolic processes (Cordes *et al.* 2013: Lu *et al*. 2015). Even if the immune system of *Drosophila* is well characterized, many questions remain to be answered. One of them is: which are the selective forces shaping the evolution of immunity genes?

1.4.2 Evidence of balancing selection in *Drosophila*

Although *Drosophila melanogaster* is a model organism in biology, the evidence of balancing selection is still rare and very few studies have been done in this species. On the contrary, many examples of genes under positive selection have been observed in *D. melanogaster* (Jiggins & Kim 2006, 2007; Lazzaro *et al.* 2004; Lazzaro 2008; Schlenke & Begun 2003, 2005; Tinsley *et al.* 2006; Sackton *et al.* 2007; Obbard *et al.* 2009).

Positive selection has been observed in a lot of immune genes like in the Imd pathway (i.e. *Relish*, *Dredd*, Begun and Whitley 2000), RNAi genes (Obbard *et al*. 2006) and genes encoding recognition proteins (i.e. TEP genes, *eater*, Jiggins and Kim 2006; Sackton *et al*. 2007; Juneja and Lazzaro, 2010). Indeed, in *Drosophila*, immune genes evolve and adapt more rapidly than other kinds of genes. Moreover, evidence of adaptive evolution has been shown in several signaling and immune recognition genes (Lazzaro and Clark 2003; Schlenke and Begun 2003; Sackton *et al*. 2007; Lazzaro 2008; Obbard *et al*. 2009).

Furthermore, it has been shown that the immune response in *D. melanogaster* will be different following the kind of pathogens infecting the host but also it shows a certain degree of specificity against various viral species (Magwire *et al*. 2012). For example, polymorphisms have been identified in the genes *ref(2)P* and *CHKov1* that confer resistance to a sigma virus (Contamine *et al*. 1989; Magwire *et al*. 2011; Wilfert and Jiggins 2010). Each polymorphism seems to be associated with resistance to one virus. Consequently, these immunity genes seem to be under selective pressure due to interactions between pathogens and host.

However, contrary to vertebrates and plants that show many evidence of balancing selection and particularly in immune genes, this type of selection has rarely been detected in *Drosophila.* Hedrick (2012) suggests that only a small proportion of polymorphisms are maintained by heterozygote advantage in this species. The first example of a polymorphism

maintained by balancing selection in *D. melanogaster* was the alcohol dehydrogenase polymorphism (*Adh*) (van Delden *et al*. 1978; Hudson *et al*. 1987). Two divergent alleles are maintained at intermediate frequency (Begun *et al*. 1999). Due to their high diversity, few genes have been described as potentially under balancing selection (i.e. *Sod* and *Est- 6* locus). However, later studies showed that the observed patterns may also be explained by other types of selection or by demography (Peng *et al*. 1991; Ayala *et al*. 2002; Balakirev and Ayala 2003).

Recently, with the emergence of new methods, some studies have described genes potentially under balancing selection in *Drosophila*. Evidence of TSP between *D. melanogaster* and *D. simulans* has been found in 16 genes and most of them are involved in immunity (Langley *et al*. 2012). In another recent study, Comeron (2014) used a background selection model as a null model to detect signatures of recent selective sweeps and balancing selection. He found some candidate regions under balancing selection including genes related to immunity (*IM4* and the *CecA1/CecA2/CecB* genes). In addition to these genes, others candidate genes involved in other functions were found such as olfactory behavior genes (*Sema-5c*) or genes encoding cuticular proteins (*Cpr11A, Cpr62Bb* and *Cpr64Ec*). These results show that balancing selection may act on genes not directly related to immunity, but maybe having an indirect role in the defense to pathogens. Sato *et al*. (2016) found significantly elevated Tajima's *D* values in the core promoter regions of 7 genes. These genes are involved in neural and behavioral traits. Finally, Unckless *et al*. (2016) found phenotypic and molecular evidence of balancing selection in the *Diptericin* gene (an antimicrobial peptide) in a population of *D. melanogaster*. However, this gene does not show classical evidence of balancing selection (high Tajima's *D* and high diversity). Consequently, balancing selection might be underestimated in *D. melanogaster* when using basic population genetic statistics.

1.5 Aim of the project

Signatures of balancing selection have been thought to be very rare in the genome and observable only in a few classes of genes such as immunity genes. As mentioned above, Andrés *et al.* (2009) performed a genome-wide analysis in humans using methods that incorporate demography in neutrality tests. Using similar methods, how many genes do we find under balancing selection in *D. melanogaster*? Are they involved in immunity as it is the case in other species? To answer these questions, we performed a genome-wide scan for balancing selection in *D. melanogaster*. We used next generation sequencing data from an ancestral population from Africa (Rwanda) and from a derived population from Europe (the Netherlands and France). We look for characteristics of balancing selection such as a high level of polymorphism compared to neutral expectations and a distortion of the SFS toward intermediate frequency alleles. Two statistics were used to detect these footprints: Watterson's estimator θ_w (Watterson 1975) and Tajima's *D* (Tajima 1989). We performed coalescent simulations incorporating a demographic model to assess candidate genes under balancing selection in our two populations. Finally, we examine in more detail one of our best candidate gene and we look for further characteristics of balancing selection such as LD, haplotype structure and changes at the protein level of the gene.

CHAPTER 2

MATERIALS AND METHODS

2.1 Sequence data

Analysis was performed on full-genome sequences of *D. melanogaster* populations. These sequences were generated by Illumina next generation sequencing technology and are publically available at the *Drosophila* Population Genomics Project (DPGP) website (www.dpgp.org.Information). We used samples from an ancestral African and a derived European population whose demography is reasonably well known (Stephan and Li 2007) and where a sufficient number of lines is available. The African samples come from two locations in Rwanda, Gikongoro (22 lines) and Cyangugu (2 lines) (Pool *et al.* 2012). The European samples come from Lyon in France (four lines) (Pool *et al.* 2012) and from Leiden in the Netherlands (eight lines) (Voigt *et al.* 2015). These data were collected from haploid embryos as described in Langley *et al*. 2011, each obtained from an isofemale lines. Consequently, these genomes are considered haploid. Moreover, all lines used for analysis (12 in Europe and 22 in Africa) were

without admixture since we excluded lines with admixture after they were identified in a previous analysis, which tested for population substructure (A. Wollstein, unpublished results). At the end, seven lines out of originally 27 were removed from Gikongoro samples, two out of 10 lines from Leiden and four out of eight lines from Lyon. This procedure coincidentally also removed lines for which genomic blocks of identity-by-descent has previously been described (Pool *et al.* 2002). For the analysis of the joint effects of selection and demography (see part 2.3.1), we used an African sample of 20 lines from Gikongoro. We added the two lines from Cyangugu in Rwanda for the genome-scan analysis and the other analysis. For all the analyses, we used European samples of 12 lines (four lines from Lyon and eight lines from Leiden. A *Drosophila sechellia* reference strain was used for estimating divergence (Li *et al*., 1999; Kim and Stephan, 2002).

2.2 Window-size analysis

We used basic population genetic parameters to detect footprints of balancing selection such as a high level of polymorphisms using the Watterson's estimator θ_w (Watterson 1975) and an excess of polymorphism at intermediate frequency with Tajima's *D* (Tajima 1989). We used a similar method as Andrés *et al*. 2009 looking for similar characteristics of balancing, even though these estimators are slightly different from the statistics that they used. Moreover, the statistics that we used are easily computed and simulated.

In a first analysis, estimates of θ_w and Tajima's *D* were calculated for windows covering the complete genome of the African and the European populations using the program *VariScan* (Hutter *et al*. 2006). The sites with missing data were removed from the analysis. We then generated empirical distributions of both statistics separately for each of the five major chromosomal arms (X, 2L, 2R, 3L, 3R). Windows in which the two statistics jointly fell within the upper 95^{th} percentile of the distribution were considered candidates for balancing selection.

Since we cannot *a priori* know which window size is optimal, we performed our analysis for different window sizes (0.2, 0.5, 1, 2, and 5 kb).

We simulated balancing selection with the software *msms* (Ewing and Hermisson, 2010) to determine which window size is the best. We performed coalescent simulation under the estimated demographic model (see below part 2.3.2, Figure 2 for the estimates) for a selection model of heterozygote advantage and a neutral model. We set a selection coefficient for heterozygote advantage (*s*) of 0.1. We used a recombination rate of 0.5 cM/Mb and we set the start of selection to 2*Ne* (with *Ne* the current effective population size) generations backward in time. We compared the distributions of the Tajima's *D* and θ_w statistics for simulated values under neutrality and selection and we determined the percentage of overlap between the two distributions for each window size (0.2, 0.5, 1, 2, and 5 kb). Windows that show smaller overlap between distributions should have higher power to distinguish between selection and neutrality.

2.3 Coalescent simulations

2.3.1 Joint effects of selection and demography

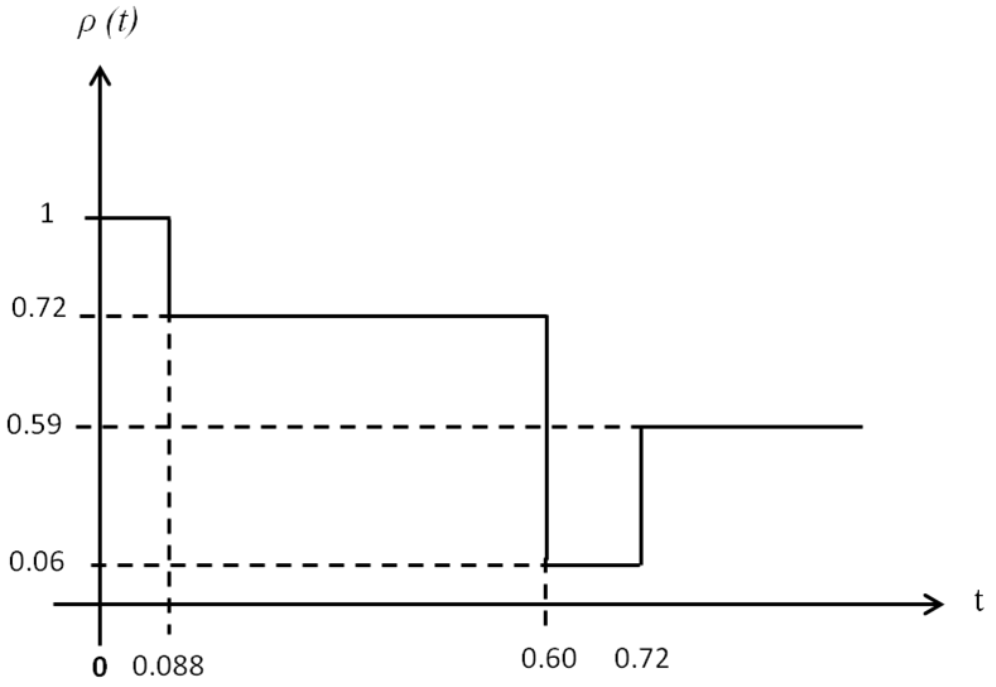
We performed an analysis to assess the joint effect of balancing selection and demography on the pattern of genetic variation. First, we created a demographic null model of a population from Africa (20 lines from Gikongoro, Rwanda). Previous demographic analyses of African populations have shown that a bottleneck model seems to be appropriate (Duchen *et al.* 2013). Consequently, we estimated parameters of a bottleneck model using an approach of Živković *et al.* (2015) that employs the SFS of neutrally evolving sites in a maximum-likelihood framework. We based our estimations on 2466 polymorphic sites located in small introns, which are thought to evolve neutrally (Parsch *et al.* 2010). The obtained estimates of the parameters are provided in Appendix A1.

Secondly, we used an algorithm similar to Zivkovic *et al*. (2015) to simulate a SFS under balancing selection and considering two demographic histories, one with a constant population size and one with a bottleneck model inferred for the African population. We used the dominance parameter *h* to simulate selection and 2*Nes* is the scaled selection of coefficient. The selective advantage of the favorable heterozygous allele pairs over the homozygous wildtype allele pair is given by 2*hs* and the selective advantage of the homozygous allele pair by 2*s*.

2.3.2 Genome-scan analysis

To conduct a test of balancing selection, we estimated the parameters of the demographic null models of the European (12 lines) and African populations (22 lines). We decided to estimate a new demographic model for the African population because we added two lines (from Cyangugu) to our African samples. Moreover, we used the method of A. Wollstein (unpublished results) to estimate the demographic history of the European population and consequently we wanted to use the same method to estimate the demographic model for the two populations. We used a similar method than previously (part 2.3.1) which is based on expectations of the SFS at neutral sites (Zivkovic *et al*. 2011). Demographic parameters were estimated for a model with instantaneous population size changes at varying time points. The demographic models that best fit the observed data were used for our analysis. The best-fit demographic models allow for a bottleneck in the European population and stepwise growth (with a shallow bottleneck) in the African population. Parameters were estimated for autosomal chromosomes (Figure 2) and X chromosome (Appendix A2) separately as autosomes and sex chromosomes might have different demographic histories (Hutter *et al*. 2007).

A.



B.

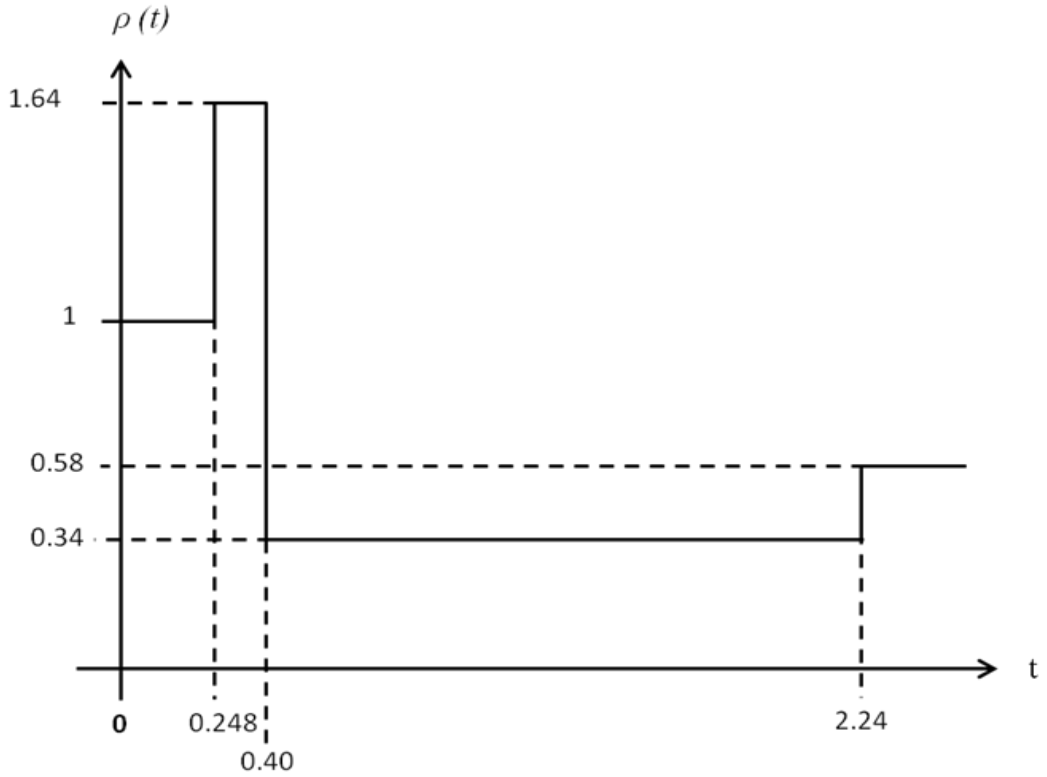


Figure 2: **Demographic models for the autosomal chromosomes for the European (A) and African (B) populations.** The x-axis represents the time t in *Ne* generations backwards in time and the y-axis represent the population size at the time t in *Ne*. *Ne* is defined as the current effective population size. Based on an estimated mutation rate of 1.5×10^{-9}, *Ne* is estimated to be equal to 1.09×10^{6} in the European population and 1.62×10^{6} in the African population.

We ran 1000 coalescent simulations for each window across the full genome using *ms* (Hudson, 2002). The local mutation rates were inferred based on divergence to *D. sechellia* (Li *et al*. 1999; Kim and Stephan 2002) for each window of 1-kb, which was deemed to be the optimal window size (see below). The local recombination rates were obtained using the *D. melanogaster* recombination rate calculator (Fiston-Lavier *et al*. 2010) based on the values of Comeron *et al*. (2012). We then compared the observed values of Tajima's *D* and θ_w for each window to the simulated neutral distributions. Only those windows for which the observed values of both statistics fell within the upper 95^{th} percentile of the simulations were kept as candidates. A *p*-value was estimated for each window for the θ_w and Tajima's *D* statistics based on the proportion of simulations for which θ_w and Tajima's *D* was greater than the observed value. When the *p*-value was equal to zero, we ran additional 10,000 coalescent simulations to obtain a more precise *p*-value. Benjamini-Hochberg multiple test correction (Benjamini and Hochberg, 1995) was applied to adjust the *p*-values. Windows with corrected *p*-values < 0.05 were retained as significant.

2.3.3 GO enrichment analysis

We decided to perform a gene ontology (GO) enrichment analysis to see if a function was overrepresented in our candidate genes. First, a list of genes located in candidate windows was determined for the African and European populations as well as for candidate regions and genes shared between the two populations. Then we applied the GO enrichment analysis to this list of

genes using Cytoscape version 3.2.0 (Shannon *et al*. 2003), in particular its plugin ClueGO version 2.2.5 (http://apps.cytoscape.org/apps/cluego) and CluePedia version 1.2.5 (Bindea *et al*. 2009, 2013) (http://apps.cytoscape.org/apps/cluepedia). We used Cohen's Kappa score (Cohen, 1968) of 0.7 as a threshold for the proportion of genes shared between enriched ontology and pathway terms to link the terms into GO networks (Bindea *et al*. 2009) and networks of KEGG (Kanehisa and Goto, 2000) and the Reactome (Croft *et al*. 2011) metabolic pathways. Using ClueGO and CluePedia we integrated enriched GO and pathway terms into networks. Enrichments and depletions of single terms were calculated using a two-tailed hypergeometric test. We applied the false-discovery-rate (FDR) correction (Benjamini and Hochberg, 1995) and retained the enriched terms with a FDR-corrected *p*-value of less than 0.05 that contained at least three candidate genes, or those whose candidate genes represented at least 4% of the total number of genes related to the term. In addition, we used the option *Fusion* to group the related terms that have similar associated genes.

2.3.4 Linkage disequilibrium analysis

We estimated the LD for SNP pairs for all the candidate genes for a region of 2 kb around each side of the candidate region for the European and African populations used for the genome-scan analysis. We calculated Hill and Robertson's r^2 (Hill and Robertson 1968) for each SNP pairs and we kept SNPs for which the allele frequency of the minor allele was above 10% and the site had less than 50% of missing data. We determined the significance of pairwise LD using Fisher's exact test (Weir 1996).

2.3.5 Trans-species polymorphisms

We used *D. simulans* as an outgroup to identify TSP. We used raw data (unmapped reads) of pooled sequences of four *D. simulans* populations from Queensland, Rhode Island, Tasmania and Florida (Sedghifar *et al*. 2016). The *D. simulans* polymorphism data were obtained by mapping the reads (alignment of the sequences) of our pooled sequences against the *D. simulans* reference genome (Hu *et al.* 2013) using bwa (Li and Durbin 2010). The alignment files were converted to SAM, and SAM files were filtered for reads mapped in proper pairs with a minimum mapping quality of 20 using SAMtools (Li *et al.* 2009). The filtered SAM files were converted into the pileup format. We computed the allele frequency of all polymorphisms in each population of *D. simulans* (Florida, Rhode Island, Queensland, and Tasmania) using custom Perl scripts. Then we aligned the *D. simulans* sequences with the *D. melanogaster* lines using ClustalW (Thompson *et al.* 1994) and searched for polymorphisms present both in all the *D. simulans* populations and all *D. melanogaster* lines.

2.4 Analysis of candidate genes

We calculated pairwise measures of linkage disequilibrium (LD) statistics (r^2) for the candidate genes *chm* and *CG15818* using the software Haploview (v. 4.2) (Barrett 2009). We excluded individuals with more than 50% of missing data. We used this software to identify structure in haplotype patterns such as haplotype blocks through an algorithm implemented in the software (Gabriel *et al.* 2002). We used the software to create a graphical representation of the LD and to define haplotype blocks in regions with strong LD.

We also studied both candidate genes at the protein level. The number of shared synonymous and non-synonymous polymorphisms within populations and between species (*D. melanogaster* and *D. simulans*) was calculated manually using DnaSP v5.10.02 (Librado and Rozas 2009). Functional information about the candidate genes was obtained from Flybase

(http://flybase.org/, version 5). We also performed the McDonald-Kreitman test (MK test, McDonald and Kreitman 1991) on our candidate regions. Fisher's exact tests were performed using R (R core team 2015). We also calculated the ratio of non-synonymous to synonymous divergence (Dn/Ds) and polymorphism (Pn/Ps) for our candidate genes and regions.

The protein structures of candidate genes were determined using the NCBI Structure and the Conserved Domains database CDD v3.11 (http://www.ncbi.nlm.nih.gov/Structure/cdd) (Marchler-Bauer *et al.* 2015). We also used the web-server Paircoil2 (McDonnell *et al.* 2006) to determine the presence of coiled-coil structures.

CHAPTER 3

RESULTS

3.1 Genome-scan for balancing selection

3.1.1 Choice of the method to detect balancing selection

We decided to use methods similar to Andrés et al. (2009) to detect evidence for balancing selection. We performed a window analysis on the full genome of *D. melanogaster* looking for a high variability (θ_w statistic) and an excess of alleles at intermediate frequency (Tajima's *D* statistic). Our goal is to detect windows with significantly high θ_w and Tajima's *D* statistics compared to neutrality. However, demography can mimic the effect that selection has on the SFS. For example, we know that a bottleneck will shift the Tajima's *D* statistic to more positive values. Consequently, balancing selection might be confounded with the population's demography. To investigate this problem, we compared the SFS for two demographic models

(one with constant population size and one with a bottleneck) under balancing selection (Figure 3).

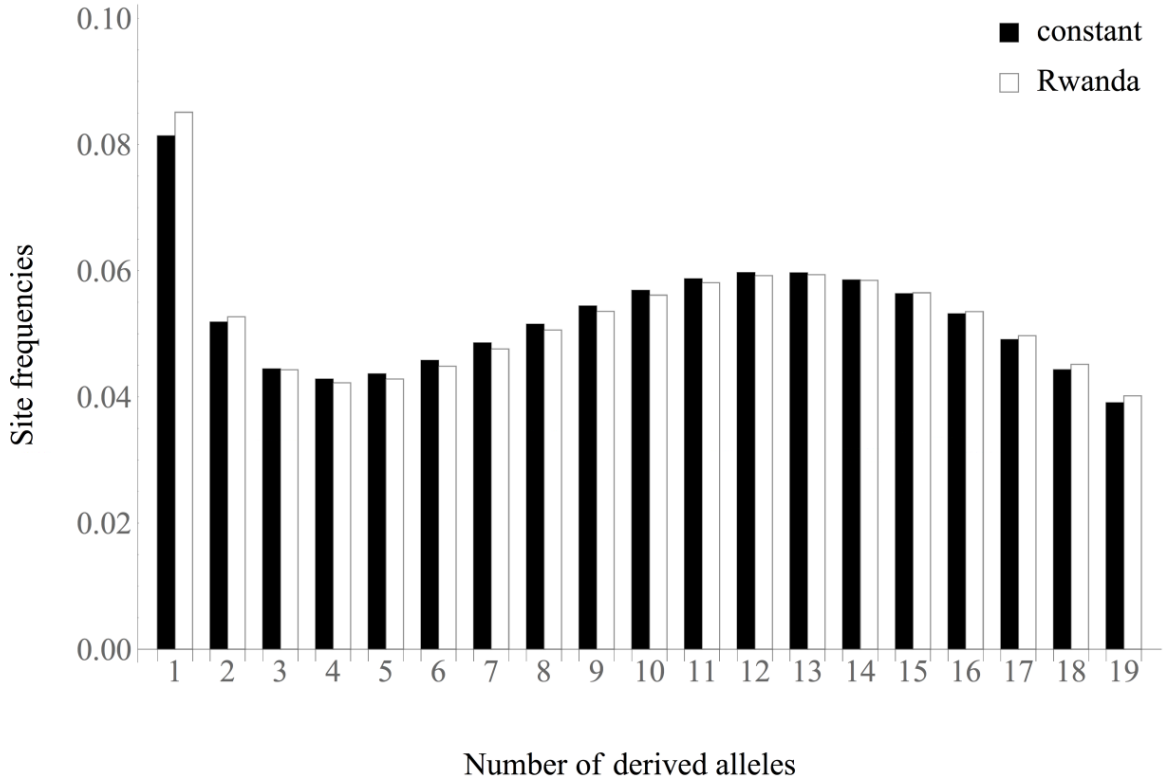


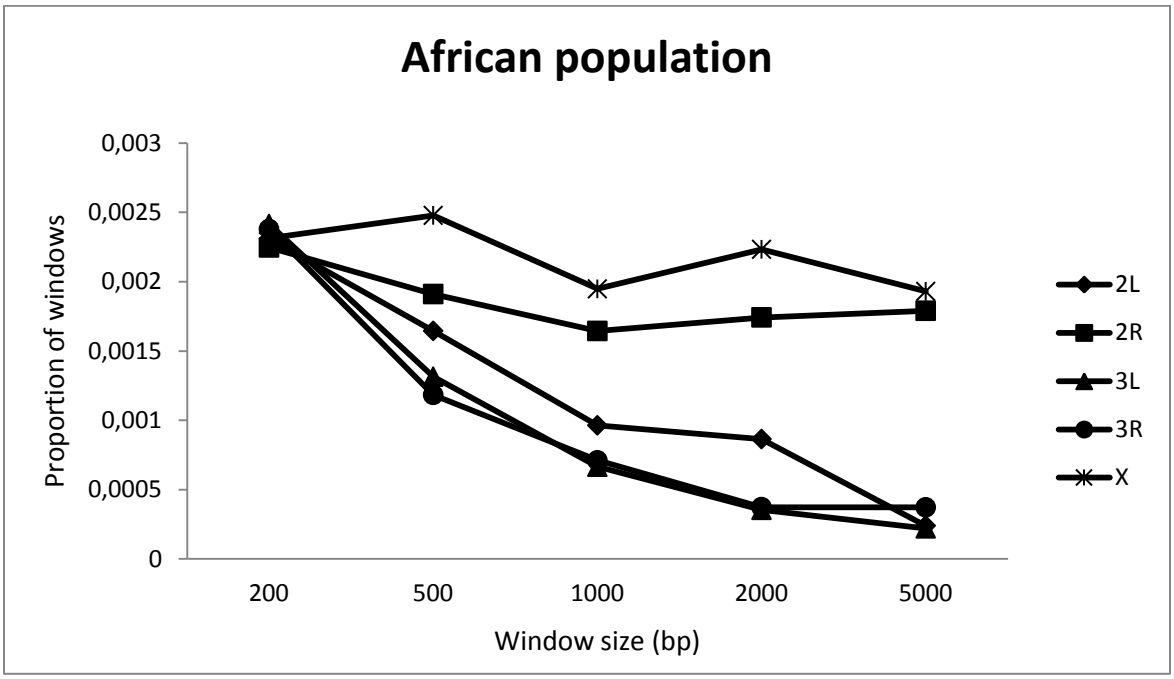
Figure 3: The site frequency spectrum (SFS) under balancing selection. We assume a constant population size (black) and the estimated African demography with a bottleneck (see Figure 2) (white), using a sample of size 20 individuals.

We can observe an excess of variants derived at intermediate frequency which is a typical signature of balancing selection, under both demographic models, which have a similar overall impact on the SFS (Figure 3). These observations are nearly independent of the strength of selection and the dominance coefficient. Therefore, we conclude that the flat Rwandan bottleneck

has not a strong effect on our procedure to identify signatures of balancing selection. However, note that other demographic histories may have much stronger effects on the signature of balancing selection. For instance, severe bottlenecks leading to an excess of low- and high-frequency derived variants (in comparison to a population of a constant size) may entirely obliterate the excess of variants derived at intermediate frequency. Moreover, the demographic models that we estimated do not represent exactly the history of our populations. Indeed their history is likely more complex but our models estimated fit the data sufficiently well to be used as a null model to reduce the number of false positives. Only windows significantly different from the overall patterns observed in the genome (taking into account the demographic history) will be candidates in our analysis.

3.1.2 Choice of the window-size for the genome scan

For the window-slide analysis, an appropriate window size has to be used to allow us to find evidence of balancing selection. Indeed, recombination might confine signals of selection to a very narrow region so the windows should be as small as possible. At the same time, the windows must be large enough to contain a sufficient number of polymorphisms to have reasonable estimates of the θ_w and Tajima's *D* statistics. To find an appropriate window size, we estimated both statistics for the full genome for different window sizes ranging from 200 to 5000 base pairs (bp) and looked at the proportion of windows in the 5% upper-tail of the distribution of both statistics which were considered as potential candidates for balancing selection. We did that for each chromosome arm in the African and European populations (Figure 4).



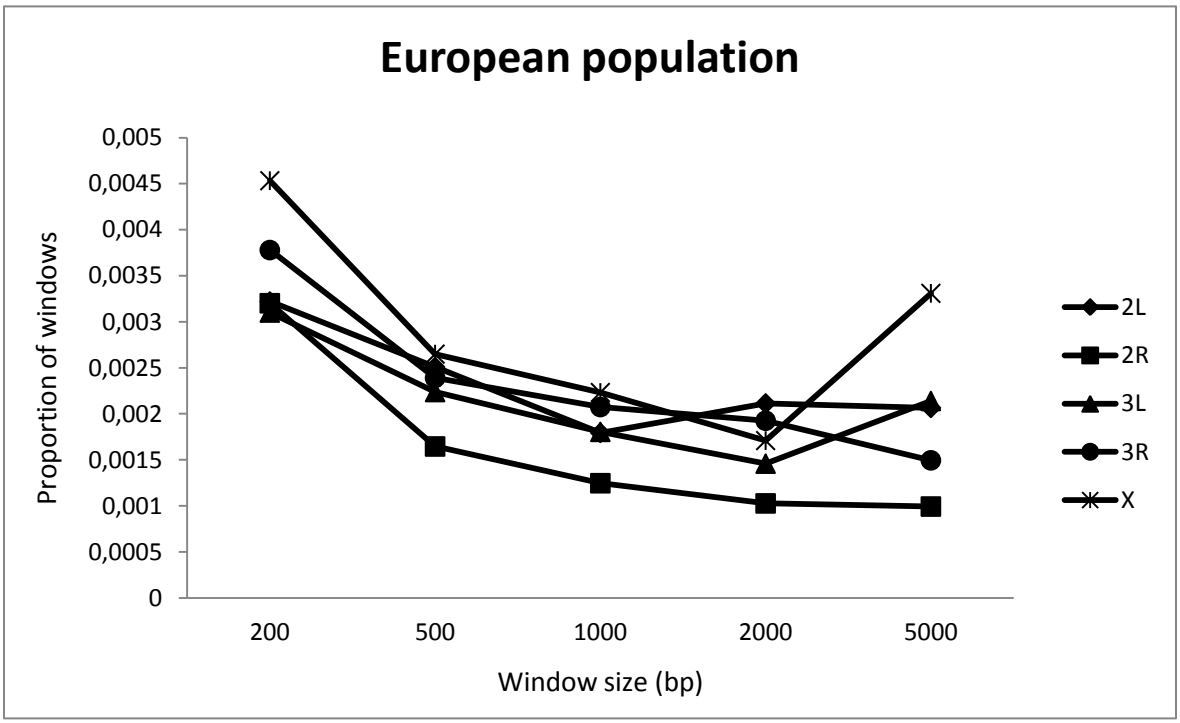


Figure 4: Proportion of candidate windows as a function of window size (in bp) for each chromosome arm (2L, 2R, 3L, 3R and X) in the African and the European populations.

The proportion of candidate windows identified on each chromosome differed depending on window size. In the two populations, the proportion of candidate windows decreased with increasing size. This decrease may be explained by linked recurrent positive (Stephan, 1995) or negative selection (Charlesworth *et al.* 1993) which is more pronounced in larger windows. However, the pattern for the chromosome arms 2R and X in the African population was not as clear since the proportion does not seem to decrease with larger window sizes (Figure 4). This may be explained by the higher average recombination rate on chromosome arms 2R and X compared to the other chromosome arms (Comeron *et al*. 2012). In the European population, the proportion of candidate windows decreases for all chromosome arms with the largest difference between 200 and 500 bp. Concerning, the chromosome X and 3L, we observed an increase for 5-kb windows which could be due to the fact that along the chromosome the overall number of 5-kb windows is low compared to smaller window sizes and consequently the proportion of candidate windows may be inflated purely to variance. Overall, the proportion of windows we identified as candidates is rather low (< 0.25% in Africa for all window sizes and from 500 bp on in Europe). This suggests that our approach may be conservative, which might be influenced by the fact that our two summary statistics (θ_w and Tajima's *D*) are numerically not independent.

We performed a second analysis in order to examine which window size has the highest power to detect balancing selection. We simulated sequence data under neutrality and balancing selection (see Methods part 2.2) and compared the overlap between the distributions of neutral and selected θ_w and Tajima's *D* values for different window sizes (Figure 5). The amount of overlap is inversely related to the power which means that the more the two distributions overlap, the less ability we have to distinguish selected from putatively neutral regions. We observed a larger overlap between the two distributions for larger window sizes and thus a lower power to distinguish selection from neutrality. The overlap for the 1-kb window is slightly larger than for the 0.2- and 0.5-kb windows and smaller than for the 2- and 5-kb ones. Moreover, the largest difference in power is between 1-kb and 2-kb. Consequently a window size of 1-kb seems to be a good choice. To choose the window size for subsequent analysis, we also take into account the

fact that in Figure 5 we show perfectly simulated data whereas in our genome scan data may be missing such that the windows are smaller than the corresponding simulated windows (on average, around 10% of the data are missing). Based on this power analysis and on the genome scan, we decided to continue our analyses with a window size of 1-kb.

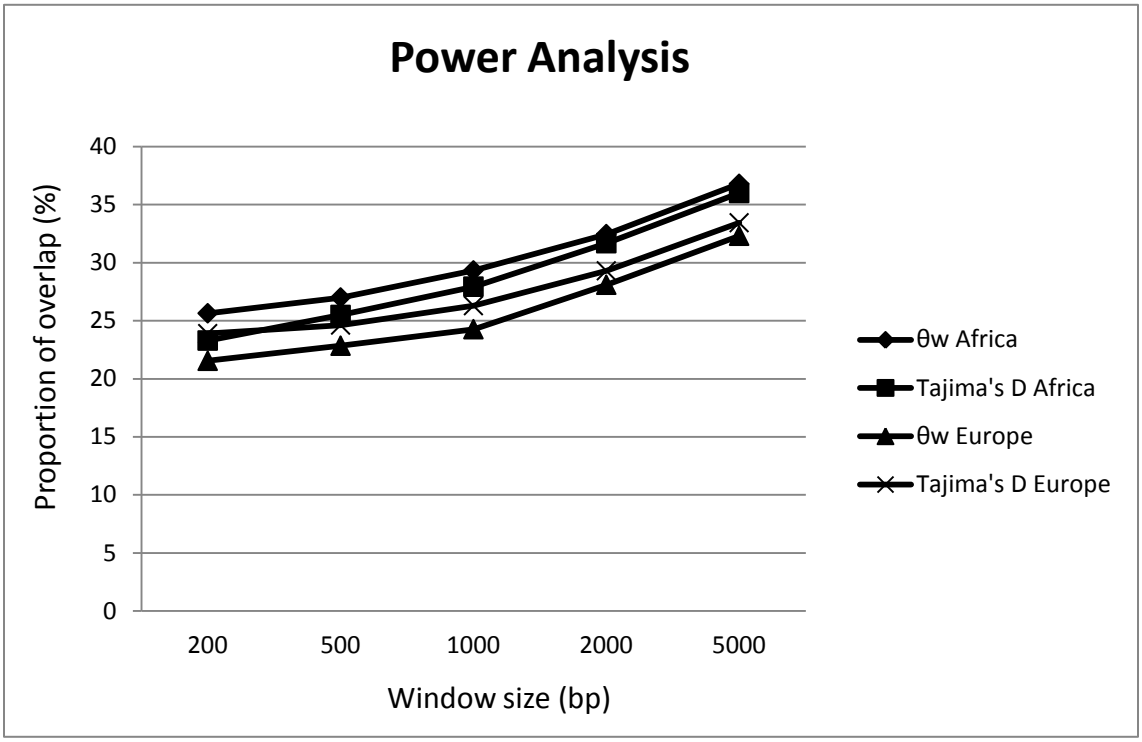


Figure 5: Power analysis for different window sizes (in bp). The two statistics θ_w and Tajima's *D* were estimated for the African and European populations. The overlap between the distributions of simulations with and without selection is represented on the y-axis.

3.1.3 Genome-scan analysis

When looking at the statistical values over all windows of 1-kb, we observed a mean θ_w of 0.0088 in Africa and 0.0033 in Europe (Table 1). The diversity (θ_w values) in the European population of *D. melanogaster* is reduced on each chromosome arm compared to the African

population, which agrees with what has been previously found (Pool *et al.* 2012). Mean Tajima's *D* averaged over all windows is -0.5605 in Africa and -0.4111 in Europe. However, compared to the autosomal chromosomes, the X chromosome has a reduced Tajima's *D* in Africa (Tajima's *D* = -0.8979), and on the contrary, an elevated Tajima's *D* in Europe (Tajima's *D* = -0.2968). Finally, as previously noticed by Glinka *et al*. 2003, the variance of Tajima's *D* is much higher in Europe than in Africa (Table 1), which indicates that the European population has been undergoing a stronger bottleneck than in Africa.

Table 1: Statistical values for the mean of θ_w and Tajima's *D* for each chromosome and population

Population	Chr.	θ_w			Tajima's D		
		5%	mean	95%	5%	mean	95%
Africa	2L	0.0028	0.0095	0.0173	-1.4842	-0.4714	0.6342
	2R	0.0022	0.0086	0.0167	-1.5752	-0.5905	0.4894
	3L	0.0017	0.0088	0.0174	-1.5104	-0.5035	0.6523
	3R	0.0015	0.0069	0.0141	-1.4119	-0.3390	0.8357
	X	0.0033	0.0100	0.0174	-1.7449	-0.8979	0.0263
	Average	0.0023	0.0088	0.0166	-1.2794	-0.5605	0.5980
Europe	2L	0.0013	0.0034	0.0073	-1.5795	-0.4262	1.1724
	2R	0.0010	0.0036	0.0082	-1.5058	-0.3851	1.1714
	3L	0.0007	0.0037	0.0087	-1.5849	-0.4517	1.1032
	3R	0.0007	0.0030	0.0071	-1.5830	-0.4957	1.1498
	X	0.0003	0.0030	0.0069	-1.7982	-0.2968	1.4504
	Average	0.0008	0.0033	0.0076	-1.6103	-0.4111	1.1075

3.1.4 Candidate genes

We searched for candidate windows with significantly elevated values of θ_w and Tajima's *D* compared to the distributions obtained by the neutral coalescent simulations under the demographic model that best fits the observed data for each population and for the autosomes and X chromosomes (A. Wollstein, unpublished results; see also Figure 2 and Appendix A2). We detected 171 candidate windows of 1-kb each for the European population and 60 for the African population with significant signatures of balancing selection. Interestingly, we found a large difference in the number of candidate windows on the X chromosome between Europe and Africa. In the European population we detected 77 candidate windows whereas in the African population only two candidate windows are on the X chromosome. Then, we identified the genes overlapping our candidate windows. Occasionally, we observed several genes (up to three genes for one window) which overlapped the same window. In the European population, 20 candidate windows have two genes present (and one with three genes), and we observed eight windows with two genes in the African population. In this case, it was difficult to identify the specific gene under balancing selection. We found 141 (Appendix B1) and 45 (Appendix B2) candidate genes in the European and African populations, respectively. Among these candidate genes, 43 genes in Europe and 16 genes in Africa are uncertain due to the fact that at least two genes are in the same candidate window.

We investigated this discrepancy in the number of candidate genes between both populations. In the European population the candidate genes are much larger than in the African population (the average size of the genes is 27.5 kb in Europe and 11.3 kb in Africa). To understand this observation, we studied the genomic distributions of the candidate genes. The European genes are restricted to regions of intermediate to high recombination rates, in which variation is less suppressed by linked selection. The 58 candidate genes on the X are distributed over about 20 Mb, whereas those on the autosome arms are located in narrower regions: 16 genes in about 9 Mb on 3R, 16 genes in 13.5 Mb on 3L, 28 genes in 15 Mb on 2R, and 23 genes in 12 Mb on 2L. This pattern may be explained to some extent by the higher average recombination

rate on chromosome arm 2R and X compared to the other chromosome arms (Comeron *et al*. 2012). The excess of large genes on the X compared to the autosome arms, however, cannot be explained by recombination ("large" is defined somewhat arbitrarily as >10 kb, but other definitions lead to similar conclusions). While 8-10 genes on each autosomal arm are large, 35 are large on the X. This suggests that the excess of large genes on the X in the European population may be due to false positives, which by chance hit longer genes more often than shorter ones. Protein-coding genes generally tend to be longer on the X chromosome compared to autosomes (with average lengths of 8.2 kb *vs.* 6.1 kb). This may partly explain the observed size distribution between X and autosomes.

The average size of the African candidate genes of 11.3 kb is also larger than the average gene length of *D. melanogaster* (which is 6.5 kb for protein-coding genes). This indicates that false positives may play a role in this dataset as well (although to a lesser extent, as only seven out of 45 genes are longer than 10 kb).

Three genes (*fry*, *chm* and *CG42389*) show signals of balancing selection in the European and African populations (Table 2). However, these signals were detected in two different regions (windows) of the genes (see e.g. Figure 6 for *chm*). Selection acting in both populations is characteristic for long-term balancing selection, which agrees with our expectation when selection predates the split of the two populations. Moreover, these genes might be under even older balancing selection (selection acting before the split of species). To look for that, we searched the presence of TSP in these three genes, but we did not find any evidence for TSP. Consequently these genes are not under ancient balancing selection. Concerning candidate genes with significant statistics only in one population, they have likely been under more recent balancing selection.

Table 2: List of candidate genes shared by the African and European populations. The values of the significant statistics observed (p-value < 0.05) for θ_w and Tajima's D are indicated in the brackets.

FBgn number	Gene name	Chromosome	Population	θ_w	Tajima's D
FBgn0016081	***fry***	3L	**Europe**	**0.0051** (0.0121)	**2.1338** (10^{-4})
			Africa	**0.0103** (10^{-4})	**1.8448** (0.0219)
FBgn0028387	***chm***	2L	**Europe**	**0.0042** (0.0285)	**2.2383** (0.0231)
			Africa	**0.008** (10^{-4})	**2.5511** (10^{-4})
FBgn0259735	***CG42389***	2L	**Europe**	**0.0025** (0.0469)	**2.3025** (10^{-4})
			Africa	**0.0202** (10^{-4})	**1.1303** (10^{-4})

The number of candidate genes detected in the two populations is very different: 45 in the African population and 141 in the European population. The differences between both populations are even more striking on the X chromosome where we found 58 candidate genes (overlapping with 77 windows) in Europe and only one candidate gene (overlapping with two windows) in Africa. In converse, it is important to notice that on the autosomes the total numbers are much closer: 44 in Africa and 82 in Europe. The disparity in the number of candidate genes between populations is unlikely strongly influenced by differences in statistical power as the proportion of overlap between simulated selected and neutral data in Africa for a 1-kb window is not much different from Europe (Figure 5). However, in Africa the proportion of overlap is slightly higher, which might indicate a lower power than in Europe. Many genes significant in Europe show high values of Tajima's D and θ_w in Africa as well, but they do not reach statistical significance in this population. In Europe, all the significant windows have also significant θ_w values in Africa, but their Tajima's D values are not significant. In the African population, we

observe 13 genes (same windows in Africa and Europe) with a Tajima's $D > 0$ (p-values = 1) for the X chromosome and 18 candidate genes with a Tajima's $D > 0.5$ (p-values ranging from 0.24 to 0.82) for the autosomal chromosomes.

To summarize, taking into account possible false positives the number of candidate genes on the X (without the excess of large genes) converges toward the numbers of candidate genes on the autosome arms in the European population. This is particularly the case for chromosome arm 2R. Furthermore, the overall number of candidate genes in the European population is no longer much greater than that of the African population and the numbers on the autosomes of both populations are more similar than reported above. On the other hand, the African X and the European X still differ greatly in the number of candidate genes, which might be due to an increase of false positives on the European X (see part 4.2.1).

3.1.5 GO terms

In order to know if balancing selection in *D. melanogaster* act on genes involved in specific function, we performed a GO terms analysis to see if and which functions are overrepresented. The GO analysis was performed on significant genes to determine the group of terms enriched for the European and the African populations. Groups are based on GO hierarchy or on the kappa score (Cohen, 1968), which is based on the overlapping genes (within categories). The name of the group is determined by the most significant term of the group (see Appendix B3).

The European population is enriched for many terms, 41 biological function categories are enriched and are grouped in eight groups (Table 3). We observed three large groups including many GO terms and consequently having a high number of genes (Table 3 and Appendix B3). The group called *cell morphogenesis involved in differentiation* encompasses 16 GO terms including terms related to behavior (sleep, circadian behavior, etc.), to development (e.g.

developmental growth) and neuronal terms (e.g. cell differentiation involved on neuron differentiation). In total 38 genes out of the 141 candidate genes are grouped under this term including genes present in several terms (e.g. the genes 5-HT1A is present in five GO terms; see Appendix B3). The term *regulation of stress fiber assembly* groups 14 GO terms including 16 genes and the category *central complex development* includes five terms and 11 genes.

We observed fewer molecular function categories: only three GO terms were enriched *(cation channel activity, protein homodimerization activity* and *transcription cofactor activity*) (Table 3 and Appendix B4). Three cellular component terms were also enriched in the European population: *microtubule, apical part of cell and plasma membrane region* (Table 3 and Appendix B4). Finally, we observed eight GO terms enriched for KEGG and Reactome pathways: *ECM-receptor interaction, cell-cell communication, EPH-Ephrin signaling, TGF-beta singnaling pathway, G alpha (s) signaling events, neuronal system, potassium channels* and *digestion of dietary lipid.* Many of the genes present in one GO terms are also found in others terms and groups. For example, the gene *mys* is present in 16 GO terms (Appendix B3 and B4).

However, since half of the European candidate genes are located on the X chromosome and there is evidence that these genes may contain an increased number of false positives we repeated our GO analysis with autosomal genes only (Appendix B5). With this reduced dataset we only found four enriched GO terms under biological process (*mushroom body development, regulation of circadian sleep/wake cycle, organophosphate metabolic process* and *nucleotide metabolic process*) and *transcription cofactor activity* under molecular process. All five of these terms were also significant in the original analysis.

In Africa, contrary to the European population, we found only two GO terms enriched (*aspartic-type endopeptidase activity* and *surfactant metabolism*) for all categories. However, this result might be an artifact since the three genes enriched for these terms (*CG31928, CG31926* and *CG33128*) are physically adjacent, which might explain why they collectively show a signal of balancing selection.

Table 3: List of enriched GO terms for the European population. The GO terms shaded in gray are the name of the group (most significant term).

Ontology	GO Group	GO term
Biological process	**Group 1**	**Cell morphogenesis involved in differentiation**
		locomotor rhythm
		circadidan behavior
		mating behavior
		sleep
		regulation of circadian sleep/wake cycle, sleep
		regulation of behavior
		detection of light stimulus
		axone extension
		modulation of synaptic transmission
		chemical synaptic transmission
		neuromuscular junstion development
		developmental growth
		developmental grpwth involved in morphogenesis
		axogenesis
		cell morphogenesis involved in neuron differentiation
	Group 2	**regulation of stress fiber assembly**
		actomyosin structure organization
		cell junction assembly
		regulation of cell migration
		actin filament bundle assembly
		heart development
		regulation of cytoskeleton organization
		regulation of cellular component movement
		regulation of locomotion
		regulation of cell morphogenesis
		regulation of anatomical structure morphogenesis
		regulation of cell morphogenesis involved in differentiation
		regulation of neuron differentiation
		regulation of dendrite morphogenesis

Table 3: continued

Biological process	**Group 3**	**central complex development**
		mushroom body development
		brain development
		neuron recognition
		synaptic development neuron recognition
	Group 4	**organophosphate metaboloic process**
		nuclrotide metabolic process
	Group 5	**heart process**
	Group 6	**establishment of localization by movement along microtubule**
	Group 7	**imaginal disc-derived wing hair organization**
	Group 8	**positive regualtion of developmental growth**
Molecular Function	**Group 1**	**Cation channel activity**
	Group 2	**protein homodimerization activity**
	Group 3	**transcription cofactor activity**
Cellular component	**Group 1**	**apical part of cell**
	Group 2	**plasma membrane region**
	Group 3	**microtubule**
KEGG and Reactome pathways	**Group 1**	**ECM-receptor interaction**
	Group 2	**Cell-Cell communication**
	Group 3	**EPH-Ephrin signaling**
	Group 4	**TGF-beta signaling pathway**
	Group 5	**G-alpha (s) signaling events**
	Group 6	**Neuronal system**
	Group 7	**Potassium channels**
	Group 8	**Digestion of dietary lipid**

Then, we examined in greater detail the more extreme candidate genes (p-value < 10^{-4} for θ_w and Tajima's D after multiple testing correction). In the European population, we found 17 genes (Table 4), which include genes involved in different functions such as cell migration (*klar*), circadian rhythm (*unc80*), neurogenesis and memory (*Tomosyn*), neuronal development (*lea* and *Ten-a*) and chemical synaptic transmission (*VGlut*). We also found genes related to immunity (*Nox, nub* and *tlk*) or involved in phagocytosis (*mvs* and *CHES-1-like)*. The genes *Nup133* and *cd* are located in the same region, *Nup133* is involved in nucleocytoplasmic transporter activity and *cd* in several processes such as response to oxidative stress. Interestingly, there is evidence that *Nup133* may also have undergone recurrent adaptive evolution in *D. simulans and D. melanogaster* (Presgraves and Stephan, 2007). Candidate genes with unknown functions have also been found (*CG2157, CG1637, CG1657* and *CG15744*). Concerning the African population, nine genes (Table 4) are highly significant (p-values < 10^{-4}). However, the genes *primo-1* and *primo-2* are located in the same region. Their proteins have the same function of dephosphorylation. The genes *CG15818* and c*hm* are also located in the same region with two significant adjacent windows. The gene *chm* is involved in 15 biological processes such as neuron differentiation, development (larvae, pupal and wing), histone acetylation and regulation of metabolic processes (see part 3.2). The gene *Cyp6a18* has an oxidoreductase activity. However, the function of the other genes remains unknown. Moreover, two of the best candidate genes (*chm* and *CG42389*) in Africa are also significant in Europe (Appendix B3). In addition, the gene *fry* is also shared by the two populations.

Table 4: List of the best candidate genes for the European and African populations with a *p*-value < 10^{-4} for θ_w and Tajima's *D*

Population	FBgn number	Gene name	Chromosome	θ_w	Tajima's *D*
EUROPE	FBgn0039004	*Nup133*	3R	0.0029	2.5160
	FBgn0263986	*cd*			
	FBgn0039536	*unc80*	3R	0.0028	2.0467
	FBgn0001316	*klar*	3L	0.0053	2.1501
	FBgn0265988	*mv*	3L	0.0046	2.5476
	FBgn0085428	*Nox*	2R	0.0064	2.1896
	FBgn0002543	*lea*	2L	0.0026	2.4560
	FBgn0031424	*VGlut*	2L	0.0040	2.2109
	FBgn0085424	*nub*	2L	0.0064	2.3335
	FBgn0086899	*tlk*	X	0.0044	2.2383
	FBgn0029504	*CHES-1-like*	X	0.0046	2.4145
	FBgn0030244	*CG2157*	X	0.0069	2.3996
	FBgn0030245	*CG1637*			
	FBgn0030286	*CG1657*	X	0.0041	2.0603
	FBgn0267001	*Ten-a*	X	0.0051	2.3024
	FBgn0030412	*Tomosyn*	X	0.0049	2.2223
	FBgn0030466	*CG15744*	X	0.0056	2.2758
			X	0.0058	2.3005
AFRICA	FBgn0040076	*primo-2*	3R	0.0105	2.2764
	FBgn0040077	*primo-1*			
	FBgn0039519	*Cyp6a18*	3R	0.0130	2.1106
	FBgn0036173	*CG7394*	3L	0.0103	2.1201
	FBgn0261853	*CG42782*	2R	0.0205	1.9016
	FBgn0031910	*CG15818*	2L	0.0121	1.8187
	FBgn0028387	*chm*	2L	0.0080	2.5511
	FBgn0028899	*CG31817*	2L	0.0144	1.8100
	FBgn0259735	*CG42389*	2L	0.0202	1.1303

3.2 Analysis of the candidate genes *chm and CG15818*

We decided to study two of our candidate genes in more detail in order to see if we find further features of balancing selection. We chose the genes *chm (chameau)* and *CG15818* for several reasons: first, the gene *chm* is one of the candidate genes found in both the African and European populations. Secondly, these two genes are the best candidate genes in Africa (Table 4). Finally, we observed strong LD in both populations in their candidate regions (Appendix B6). The function of the gene *CG15818* is unknown and the gene *chm* is involved in many functions including development and regulation of different genes expression.

As can be seen in the Figure 6, the candidate regions of this gene are different in both populations. For Africa, more than one candidate window was significant and so the region of interest is 2-kb while it is 1-kb in Europe. We can also observe that the values of Tajima's *D* are very high in the region of interest in each population. In Africa, the Tajima's *D* values are 1.82 and 2.55 (p-value < 10^{-4}) and the θ_w values are 0.012 and 0.008 (p-value < 10^{-4}) for the two windows of 1-kb between 7411.5 kb and 7413.5 kb (Table 4 and Figure 6). We can observe that the θ_w values are high in several regions of the gene *chm* but Tajima's *D* is high and significant only in our candidate region. In Europe, the Tajima's *D* and θ_w values are equal to 2.24 (p-value = 0.023) and 0.042 (p-value = 0.029), respectively (Table 2 and Figure 6). We observe also high values of Tajima's *D* and θ_w in other regions of the gene *chm*. However, the combination of the two statistics and their significance after correction allow us to define two candidate regions for each population.

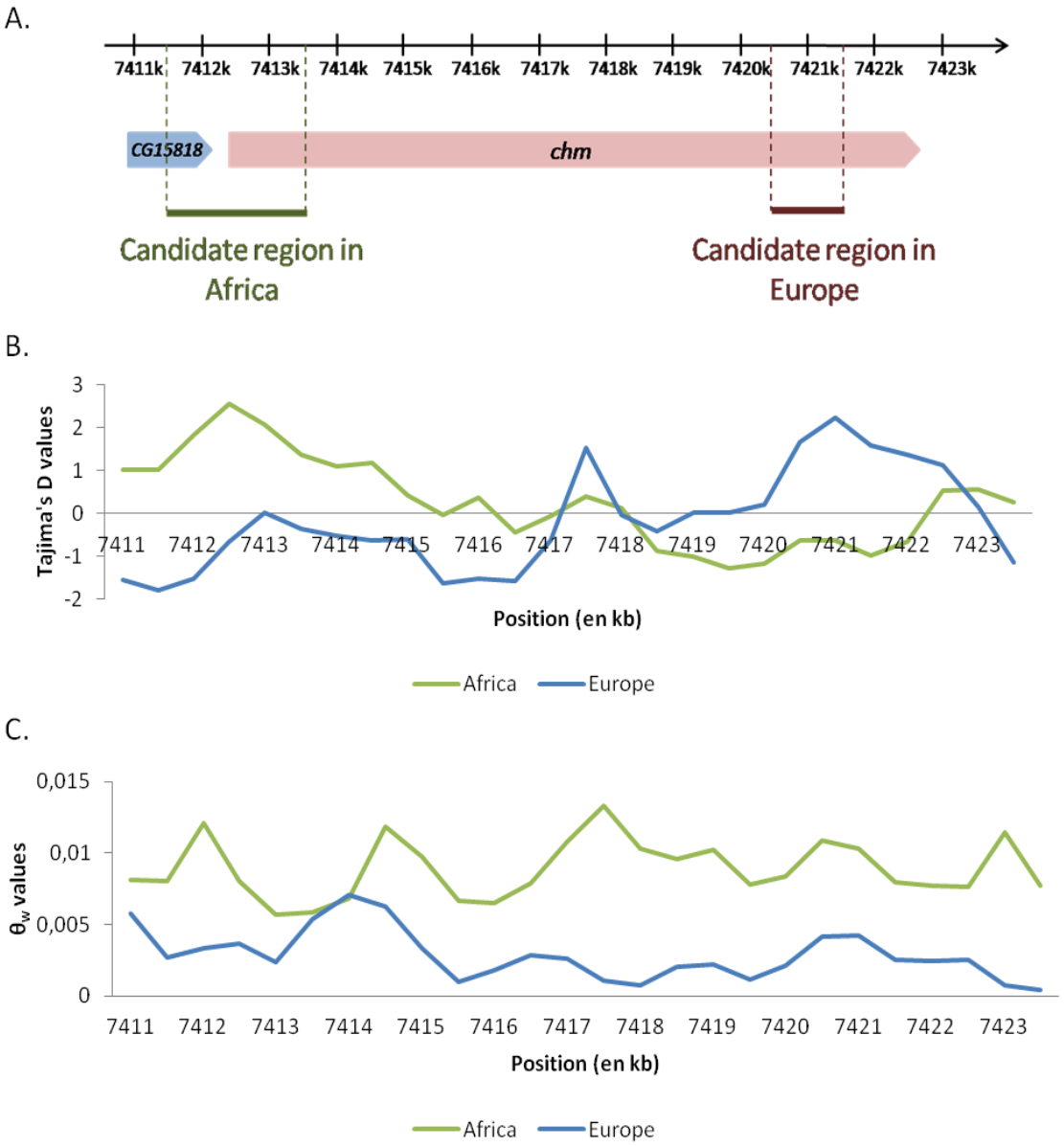


Figure 6: Map of the genes *CG15818* and *chm.* The green bar represents the region of interest in the African population and the red bar represents the region of interest in the European population. In the case of the African population, the region of interest is larger as two contiguous candidate windows are significant. The Tajima's *D* (B) and θ_w (C) values are plotted for a 1-kb sliding window across the genes *CG15818* and *Chm* for both Europe and Africa populations.

Interestingly, both genes have annotated protein domains in the candidate region (Figure 7). We identified a coiled-coil domain and a C-type-lectin / C-type lectin-like (CLECT) domain in the gene *CG15818*. A coiled-coil domain is a structural motif in proteins composed of several α-helices. This domain binds other molecules and is involved in many biological functions (Reddy and Etkin 1992; Mason and Arndt 2004), but it is unknown with which molecules this coiled-coil domain interacts. The CLECT domain binds protein molecules such as ligands, lipids and inorganic surfaces. Concerning the gene *chm*, two domains (Histone acetyltransferase (HAT) domain and a zinc finger domain) are in the gene and one domain (HAT domain) is in the candidate region found in the European population. The gene *chm* belongs to the MYST family which is the largest family of histone acetyltransferase. MYST proteins mediate many biological functions including gene regulation, DNA repair, development and cell-cell regulation.

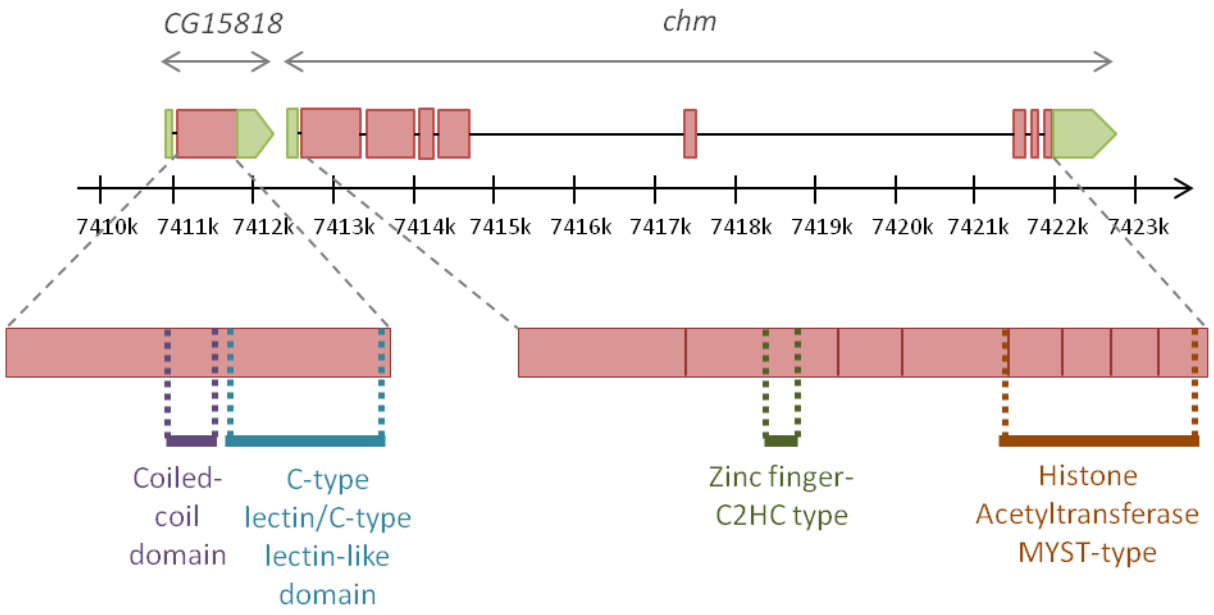


Figure 7: Schematic representations of the domains present in the proteins *CG15818* and *chm*. The green rectangles represent the 5'UTR regions, the green arrows correspond to the 3'UTR regions. Red rectangles represent the exons and the black lines the introns. The position of the protein domains are represented by different colored bars (purple, blue, green and orange).

We looked at these candidate genes at the protein level and we identified the synonymous and non-synonymous (NS) polymorphisms present in each of them. We observed several polymorphisms at intermediate frequency including NS polymorphisms in the two populations. In the gene *CG15818*, we found six NS and 13 synonymous polymorphisms in the European population. In the African population, we found four NS and 16 synonymous polymorphisms. Concerning the gene *chm*, the majority of the polymorphisms observed are found in the two first exons (from position 7412507 to 7414007, see figure 8) with four NS and 22 synonymous polymorphisms in Europe and three NS and 20 synonymous polymorphisms in Africa in these two exons (Figure 8). In the others exonic regions of the gene *chm*, no NS polymorphism were found. We performed a McDonald-Kreitman test on the genes *CG15818* and *chm*, but in both cases the MK test was not significant for either population even when we considered only the first two exons and each exon separately for the *chm* gene.

We also observed non-synonymous polymorphisms in the protein domains present in the gene *CG15818* (Figure 8). Two amino acids changes are in the coiled-coil domain including one polymorphism present in both populations. In the CLECT, three NS polymorphisms are observed, but only one is in both the European and the African population. Moreover, the two NS polymorphisms (one in Africa and one in Europe) in the coiled-coil domain are at intermediate frequency with 58% and 64% of the derived allele in the European and the African populations, respectively.

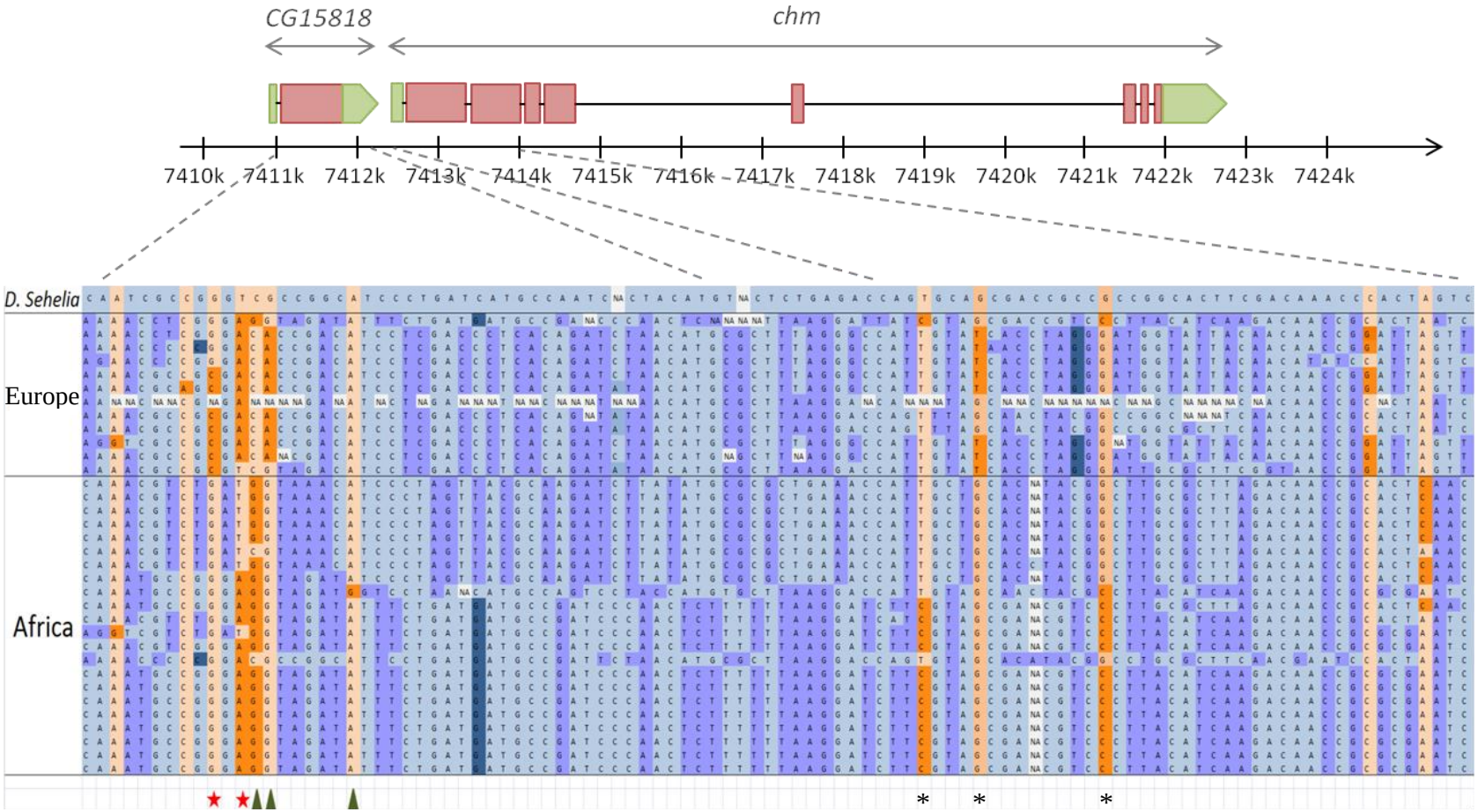


Figure 8: Polymorphism table of the candidate region in Africa. The blue color represents the synonymous polymorphisms and the orange color, the non-synonymous polymorphisms. The darker colors are the derived alleles and the light colors, the ancestral alleles based on the outgroup *D. sechellia*. The red stars indicate the non-synonymous polymorphisms present in the coiled-coil domain and the green rectangles, the polymorphisms in the C-type lectin like domain. The asterisks indicate the non-synonymous polymorphisms present in the first exon of the gene *chm*.

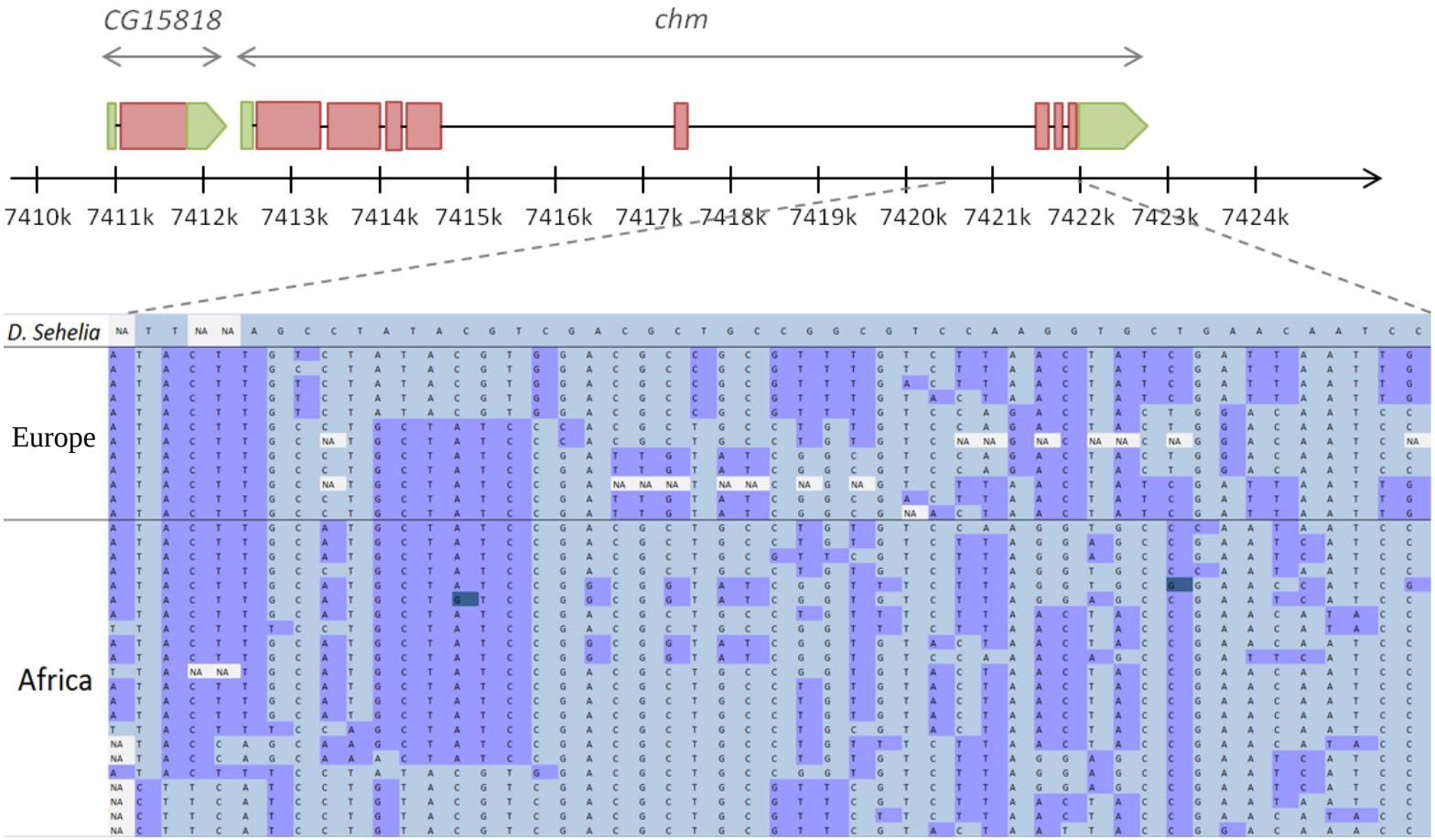
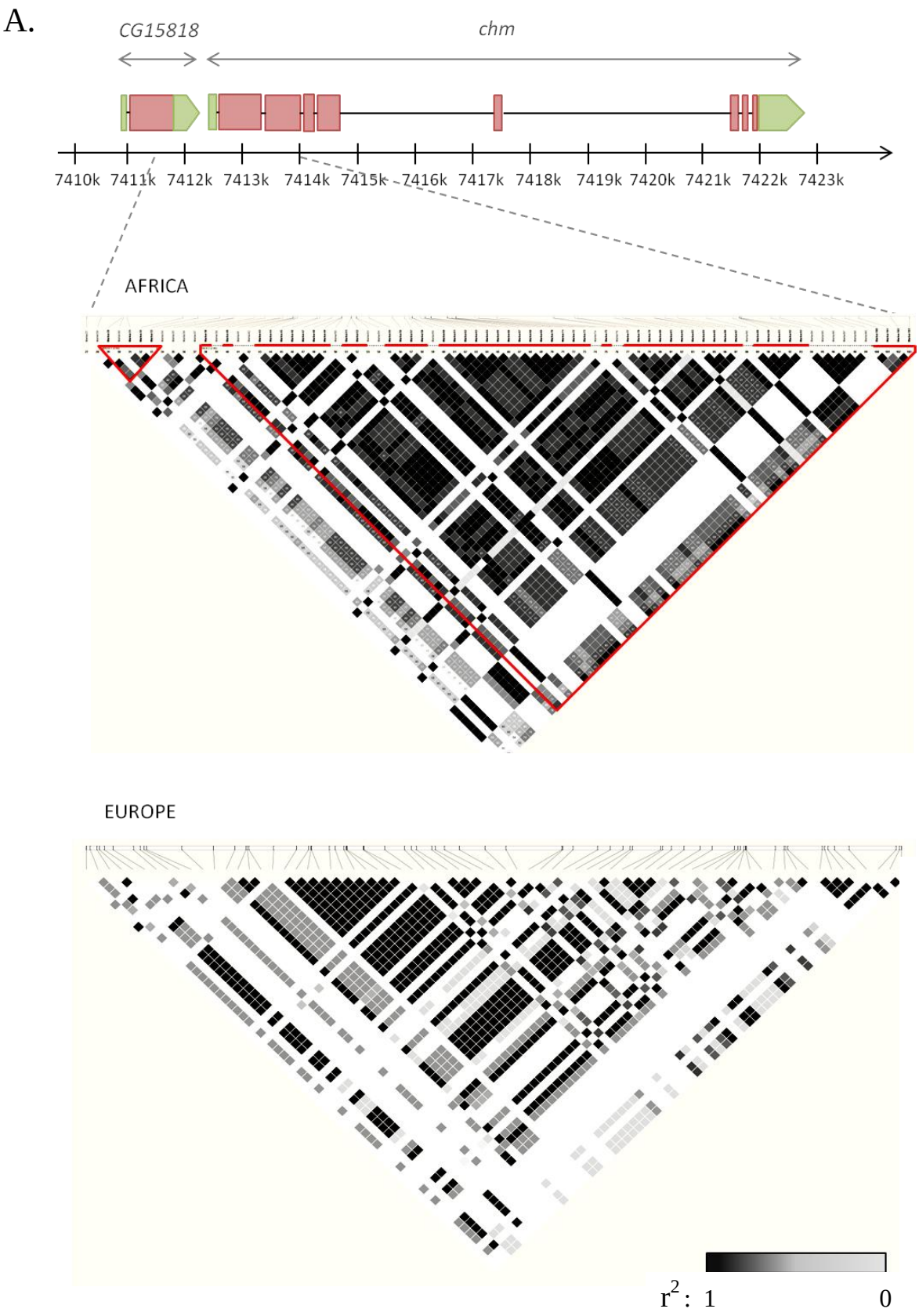


Figure 9: Polymorphism table of the candidate region in Europe. The blue color represents the synonymous polymorphisms. The darker colors are the derived alleles and the light colors, the ancestral alleles based on the outgroup *D. Sechelia*.

As mentioned before, we found evidence of LD in the candidate regions of the genes *chm* and *CG15818* for both populations (Appendix B7 and B8). In the first candidate region which corresponds to *CG15818* and part of *chm*, we found a block of strong LD (Figure 10A). In the African population, we observed a small block where we have the presence of a NS polymorphism in *CG15818* which is in LD with four synonymous polymorphisms. This NS SNP is in the coiled-coil domain of the gene and is almost fixed in the European population (11 lines out of 12 have the derived allele) whereas it is at intermediate frequency in the African population. A second large block of polymorphisms in strong LD is observed at positions 7 411 654 to 7 414 271 (see Figure 10A). This block includes three NS polymorphisms in the gene *chm* (two in the first exon and one in the second exon). These polymorphisms are at intermediate frequency (54.55%, 59.10% and 36.36% for the derived alleles) (Appendix B8A). Indeed, the two NS polymorphisms (one is a serine/proline replacement and the second is a arginine/proline replacement) in the first exon have already been describe by Levine and Begun (2008) as being in LD. Eight different haplotypes are found with two at intermediate frequency (45.5% and 27.3%), the six others are at a low frequency (4.5%). On the contrary, in Europe only two NS polymorphisms are observed and only one line has the derived allele. Finally, we observed that many polymorphisms in LD are located on the 3'UTR of the gene *CG15818* (14 SNPs) or they are intergenic (11 SNPs). In the European population, we did not observe any block of LD even though individual pairs of polymorphisms are in LD.

Concerning the second candidate region which is significant only in the European population, we also found strong LD and particularly in Europe. Indeed, we observed two haplotype blocks, one small (6 SNPs) and one larger (23 SNPs). However, no NS polymorphisms are present in the three exons in this region, but seven synonymous polymorphisms are located in the 5'UTR. In the small haplotype block, two haplotypes are observed with a frequency of 41.67% and 58.33% and in the large block, 6 haplotypes are found with a frequency of 36.4% for one, 18.2% for two and 9.1% for three haplotypes (Appendix B8B). In Africa, few SNPs in LD are significant and haplotype blocks contain only a small number of SNPs.


A.
CG15818
chm
7410k 7411k 7412k 7413k 7414k 7415k 7416k 7417k 7418k 7419k 7420k 7421k 7422k 7423k
AFRICA
EUROPE
r2: 1 0

B.

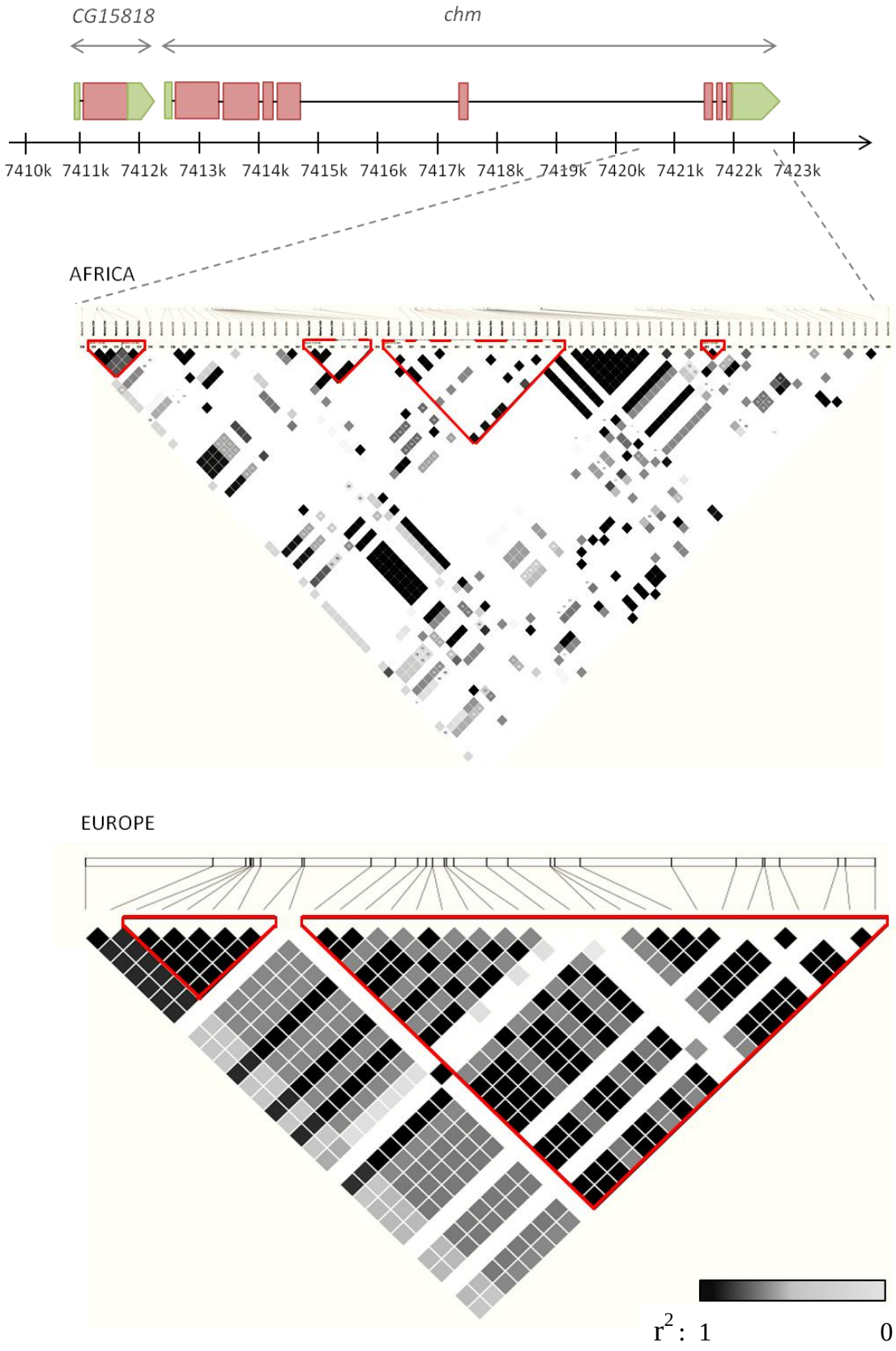

CG15818
chm
7410k 7411k 7412k 7413k 7414k 7415k 7416k 7417k 7418k 7419k 7420k 7421k 7422k 7423k
AFRICA
EUROPE
r2: 1 0

Figure 10: Representation of LD (r^2) for the two candidate regions. A. represents the LD in the African candidate region and B. LD for the European candidate region in both populations each. The magnitude of pairwise LD is given by the color shading; black and grey colors represent different levels of LD in descending order. Black/grey pairs are statistically significant, while white pairs are not. When $r^2 = 1$, the loci are in complete LD and when $r^2 = 0$, the loci are in perfect equilibrium. The red triangles represent the haplotype blocks.

In the candidate genes *chm* and *CG15818*, we found evidence of balancing selection acting on two different regions in the European and African populations. We observed the presence of LD and haplotypes at intermediate frequency which are patterns that we expect to find under balancing selection. In Europe, balancing selection is acting exclusively on the gene *chm* whereas in the African population, selection is acting on both genes. These results might indicate that due to the different environment of the two populations, balancing selection does not act on the same variant.

CHAPTER 4

DISCUSSION

4.1 Detection of footprints of balancing selection

In this thesis, we wanted to search for evidence of balancing selection in *D. melanogaster* which is an important model organism in biology. Effectively, balancing selection has been proposed as one of the mechanism maintaining variation in the genome of natural populations. However, few examples of this selection are known and only recently some studies have found evidence of genes under balancing selection in *D. melanogaster*. For instance, Sato *et al.* (2016) detected footprints of balancing selection in core promoter regions and Unckless *et al*. (2016) found evidence of alleles maintained by balancing selection in genes encoding antimicrobial peptides. Moreover, no genome-wide analysis for balancing selection has been done in *D. melanogaster*. One of the reasons for this was that until recently we did not have full-genome sequences. However, since several years several studies have sequenced the full genome of several species and population of *Drosophila* (Pool *et al*., 2012; Voigt *et al.*, 2015; Sedghifar *et*

al., 2016). Thanks to the availability of NGS data of good quality and new technologies, it is now possible to perform an analysis on the whole genome and to have accurate estimation of variation to detect balancing selection.

We performed a genome-scan in *D. melanogaster* to screen potential targets of balancing selection in two populations from Africa (ancestral) and Europe (derived). We decided to combine two common statistics which are the Watterson's theta (θ_w) (Watterson, 1975) and the Tajima's *D* (Tajima, 1989). These two estimates search for an excess of the number of SNPs and an excess of SNPs at intermediate frequencies. The combination of these two features is characteristic of balancing selection and cannot be confounded by other types of selection such as purifying and positive selection. Other studies have previously used a combination of tests looking for various features of balancing selection in others organisms (Andrés et al. 2009; Ochola et al. 2010; Thomas et al. 2012) and found some evidence of balancing selection. This method is conservative as it permits to detect only strong signals of balancing selection and we expected to have a low rate of false positive.

We used a sliding window approach on the full genome to estimate both statistics for each window. One important point was to define the best window size to detect footprints of balancing selection. We decided to perform our analysis for a window of 1-kb because this size seems to have a good power to distinguish between balancing selection and neutrality. Moreover this window size is sufficiently large to have good estimates of our statistics and at same time it is sufficiently small to detect the signature of balancing selection and avoid segments not under selection.

Furthermore, we accounted for demography in our methods. Thus we followed a similar approach as Andrés *et al*. (2009) rather than using a model-based method, such as DeGiorgio *et al*. (2014). Indeed, DeGiorgio *et al*. (2014) found that the aforementioned SFS-type tests of Andrés *et al*. (2009) have less power than their new method. We may conclude from this result that our approach may also lack power in detecting balancing selection in *D. melanogaster*. We

could apply the procedure of DeGiorgio *et al*. to the *Drosophila* data, but there is a *caveat*. Since the null model in the method of DeGiorgio *et al*. assumes neutrality just as the aforementioned methods, it is unclear to what extent it would improve our estimation. The reason can be seen in Figure 4, which indicates that larger fragments generally produce less candidate regions (with the exception of chromosome arms X and 2R in Africa and chromosome arms X and 3L in Europe). This observation can be explained by the action of linked positive or negative selection. Therefore a null model should incorporate both recurrent selection and demography in order to obtain unbiased estimates of the number of loci under balancing selection in *D. melanogaster*, which is at present only possible for unlinked sites (Živković *et al.* 2015).

4.2 Evidence of balancing selection in *D. melanogaster*

4.2.1 Candidate genes

In our first genome-wide analysis of an African and European populations in *D. melanogaster* we found 183 candidate genes under balancing selection out of 13.900 protein-coding genes, including 141 in the European population and 45 in the African one. Three genes were overlapping between both populations (*fry*, *chm* and *CG42389*). This overlap is much smaller than what Andrés *et al*. (2009) found in humans between a Europe-derived and an Africa-ancestral population. It may be explained by the much longer separation time (in generations) between the two fly populations. However, even if the *p*-values are not significant (especially in the African population), we observed many genes (31 genes) with high estimates of θ_w and Tajima's *D* shared between both populations. Perhaps our method is too conservative so that only genes under strong balancing selection are detected. Moreover, we performed rigorous testing corrections, which led to many overlapping genes losing significance after correction. Finally, a locus under balancing selection may show a strong signal in one statistic but weaker in the other. In this case, it is possible that we did not detect all the genes under balancing selection.

Although we fitted demographic models to the data for both the African and European populations, we found evidence for false positives (large genes > 10-kb) in our set of candidate genes. More false positives appear to be present in the European set of candidate genes (in particular on the X chromosome). Inaccuracies in the estimation of demographic parameters may be the primary reason for this problem. We estimated demography for the X and the autosomes separately, based on the SFS at neutral sites (Zivkovic *et al*. 2011; Parsch et al. 2010). Since the European X chromosome harbors the lowest amount of variability the estimated demography might have been less precise for the European X compared to the European autosomes and the African X and autosomes, leading to an elevation of false positives.

The discrepancy between the X chromosomes of both populations is particularly large. We observed only one candidate gene on the African X, but 20-30 on the European one (after correcting for the excess of large genes). As mentioned above, all significant windows on the X where we found our candidate genes in Europe have also significant θ_w values in Africa, but their Tajima's *D* values are not significant. The reason for this may be as follows. As already Glinka *et al*. (2003) noticed, the variance of the European X is higher than that under standard neutrality, and lower in Africa (see also Table 1). Therefore, scaling Tajima's *D* with the standard neutral variance may have led to too many candidates in Europe and/or too few in Africa.

We observed balancing selection in many genes in only one population. In the derived European population, balancing selection likely acted only recently (when *D. melanogaste*r arrived in Europe). *D. melanogaster* had to adapt to a new environment (different climate, pathogens, etc.) and consequently new genes were under balancing selection compared Africa. Another observation is that many of our candidate genes are under balancing selection only in Africa. Since the African population is ancestral, we might think that balancing selection is old and genes are also under this selection in Europe. This result may be explained by the fact that balancing selection is effectively old in Africa but the selective pressure on these candidate genes change in Europe due to the different environment such that these genes are no longer under balancing selection in Europe contrary to Africa.

Concerning the examples found previously in *D. melanogaster* (Comeron *et al*., 2014; Unckless *et al*., 2016; Sato *et al*. 2016), we did not confirm any of these examples although we observed some genes with similar functions such as oxidation-reduction process and olfactory behavior. The fact that we did not find overlap with other studies might be explained by the difference in the methods (Comeron *et al*., 2014) and the samples, which are not exactly the same. Moreover, in the studies Unckless *et al*., (2016) and Sato *et al*. (2016), the authors look for balancing selection only in a small part of the genome of *D. melanogaster*.

4.2.2 Function of the candidate genes

We decided to look in more details at our candidate genes and particularly their functions. We wanted to see if balancing selection is limited to a few classed of genes. We observed enrichment in many GO terms in the European population but in only two in Africa. Moreover, these two GO terms are probably not enriched as the three corresponding genes lie in the same genomic region. This low number of terms enriched in Africa compared to Europe may be explained by the fact that we have less candidate genes in this population than in Europe.

Concerning the functions of genes, we observed enrichment in many biological processes in the European population. When we repeated the analysis for autosomal genes only, we were, however, only left with five GO terms. This reduction might be because of an overrepresentation of certain functions on the X chromosome, but could also be purely due to reduced statistical power given the lower number of genes in the autosomal dataset. GO terms that were consistently detected include ones related to circadian behavior and the development of mushroom bodies. Mushroom bodies play a major role in olfactory learning and memory, but have also been shown to be involved in other behavioral traits and the regulation of sleep (Heisenberg, 2003; Joiner *et al*. 2006). Even though these GO terms seem to be closely related their statistical significance is driven by different sets of genes (Appendix B3, B4 and B5). Candidate genes related to neuronal development and behavior are particularly interesting, as evidence of balancing selection in genes

associated with neuromuscular junction development and behavior (Comeron, 2014) has previously been reported. For the African population, only two GO terms are enriched and the three corresponding genes lie in the same genomic region.

Having many candidate genes, we decided to look in more details at the best candidate genes with a *p*-value < 10^{-4} for θ_w and Tajima's *D* in each population. We found 17 extreme genes in the European population and 9 in the African population. Among these genes, some are involved in neural function, the gene *Ten-a* is involved in neuronal development and also in the establishment of neuron connectivity (Mosca and Luo, 2014). The gene *Tomosyn* plays a role in the regulation of behavioral plasticity and memory (Chen *et al.* 2011) and *VGlut* is involved in neuromuscular junctions. The genes *primo-1* and *primo-2* have both a function in dephosphorylation and play a role in different functions such as neurogenesis (Miller *et al.* 2000). Finally, *chm* enhances JNK signaling during metamorphosis and thorax closure and acts positively in the JNK-dependent apoptotic pathway (Miotto *et al.* 2006). This gene is also required for the maintenance of Hox gene silencing by PolyComb group proteins (Grienenberger *et al.* 2002). These results seem to confirm our previous observation that many GO terms enriched are involved in neuronal functions. These genes might be under balancing selection due to temporal changes in the environment like fluctuations between seasons (Bergland et al. 2014). However, as mentioned in the introduction, it has been thought that balancing selection acted on immune genes as many examples in immunity were found in the literature. We did not find GO terms enriched in immunity but among our extreme genes, some are related to immunity such as *tlk, nub, CHES-1-like* and *Cyp6a18* discussed next.

4.3 Balancing selection in immunity

Until now, many examples of genes under balancing selection were related to immunity in humans, plants and parasites (Andrés *et al.* 2009; Key *et al.* 2014; Delph and Kelly, 2013;

Amambwa-Ngwa *et al.* 2012). Indeed, genes involved in immune defense are assumed to often evolve under balancing selection. Effectively, it is thought that the host-parasite coevolution is one of the main forces maintaining diversity in the genome and driving immune genes to evolve under balancing selection (Schlenke and Begun, 2003; Obbard, 2009). For example, Andrés *et al*. found a relatively high number of candidate genes related to immunity. However, we did not find an enrichment of genes involved in immunity. Only a few candidate genes of our scan are involved in immunity, such as *Ser* gene (involved in melanization of pathogen) in Europe and *Dif* gene (it mediates an immune response in larvae) in Africa. We also detected four genes involved in wound healing (*Cad96Ca, Fhos, Rok* and *Hml*) in Europe. Among the 26 extreme candidate genes, some are related to immunity. The gene *tlk* has been reported to be involved in the humoral immune response (Kleino *et al.* 2005). The gene *Nox* has a role in both regulation of the gut microbiota and resistance to infection by inducing the generation of reactive oxygen species (Buchon *et al.* 2014). The gene *nub* is a negative regulator of antimicrobial peptide biosynthesis. It represses the expression of NF-κB-dependent immune genes and increases the tolerance to gut microbiota (Dantoft *et al.* 2013). The gene *CHES-1-like* is required for phagocytosis of the fungal pathogen *Candida olbicans* (Stroschein-Stevenson *et al.* 2006). *CG15818* has been shown to be down-regulated in flies infected by the Nora virus (Cordes *et al.* 2013). The gene *Cyp6a18* may play a role in the metabolism of insect hormones and in the resistance to insecticides. Furthermore, Comeron *et al*. (2014) found another gene from the cytochrome P450 family (*Cyp6a16*) as candidate gene for balancing selection in *D. melanogaster*.

Even if the majority of our candidate genes seem to be involved in other functions than immunity, they could also play a role during an infection. It has been shown that the immune system is linked to circadian rhythms (Tsoumtsa *et al.*, 2016). Clock genes may be involved in the fight against bacterial invasion (Shirasu-Hiza *et al*. 2007; Lee and Edery, 2008). For example, the ortholog of our candidate gene *cry* has been shown to up-regulate pro-inflammatory cytokine gene expression during an infection in mice (Narasimamurthy *et al*. 2012). Several studies in *D. melanogaster* found that genes induced by infection are not only involved in immunity but also in

other functions such as detoxification (Paparazzo *et al.* 2015), cell adhesion, calcium binding, etc (Irving *et al*. 2002). Recently, Lu *et al*. (2015) performed a genome screen to identify *Drosophila* genes affecting susceptibility to the pathogen *Metarhizium anisopliae.* In addition to classical immune genes, they identified many non-immune genes involved in several biological functions such as neurogenesis, metabolic processes, transcription regulation, and transport. Moreover, Andrés *et al*. (2009) also detected candidate genes involved in the extracellular matrix. Indeed, we found candidate genes involved in these different functions (see Appendix B3 and B4) Concerning the examples found previously in *D. melanogaster* (Comeron, 2014; Unckless *et al.* 2016; Sato *et al.* 2016), we did not confirm any of these examples although we observed some genes with similar functions such as oxidation-reduction process and olfactory behavior. There is evidence for a connection between the immune and the nervous system (including olfaction) in insects (Mallon *et al.* 2003). Consequently, several of the non-immune genes found to be under balancing selection may play a role in immune response.

The low number of immunity genes under balancing selection in *D. melanogaster* could be due to the difference in the immune system between *Drosophila* and other organisms. For example, in humans the immune system is more complex and it is comprised of an adaptive and an innate immunity system. Moreover, many genes under balancing selection in humans are in the MHC which is not present in *Drosophila* (Kelley *et al*. 2005; Piertney and Oliver 2005; Spurgin and Richardson 2010, Tesicky and Vinkler, 2015).

A theoretical argument for the lack of balancing selection detectable in genome studies and in immune genes is provided by Tellier *et al*. (2014). In this study, they considered host-parasite coevolutionary dynamics using a gene-for-gene model. This model is analyzed in a finite population. One version of the model is monocyclic (one parasite generation per host generation) and the other is polycyclic (two or more parasite generations per host generation). They looked for genomic footprints of balancing selection in host and parasite (i.e. excess of alleles at intermediate frequency in the SFS) after performing coalescent simulations. They observed signatures of balancing selection only in the parasite genomes and only in the polycyclic model.

Their explanation is that the equilibrium point (where alleles is maintained at intermediate frequency) in the host is closer to the boundary than in the parasite, and so it is hard to detect balancing selection in the host because of frequent fixations under drift (Tellier *et al*. 2014). Furthermore, balancing selection has to be very strong and act for a long time to be observed. These findings may suggest that old balancing selection events such as TSP are more readily detectable than short-lived ones (Leffler *et al.* 2013).

4.4 The candidate genes *chm* and *CG15818*

The gene *chm* has significantly high values of Tajima's *D* and θ_w in the European and African populations. This result seems to indicate that balancing selection is acting on this gene since a long time (before the separation of the two populations). However, we did not find any TSP which indicates that balancing selection is not ancient and does not predate the speciation of *D. melanogaster* and *D. simulans*. Moreover, the regions of interest are different in each population. In Africa, the candidate region overlaps the two genes (*CG15818* and *chm*) whereas in Europe, the candidate region is restricted to *chm*. This might indicate that due to the change of environment, balancing selection was no longer acting anymore on the same region as in the African population.

We decided to look in more details at these two genes to see if we observe patterns of balancing selection. First, we looked at the protein level in order to see if we find an excess of NS polymorphisms which is what we expect under balancing selection. We did not observe any excess of NS polymorphisms even if we found non-synonymous polymorphisms at intermediate frequency in the African candidate region. Indeed, five NS SNPs are in protein domains of the gene *CG15818* (two on a coiled-coil domain and three on a CLECT domain). One interesting observation is that one of the NS polymorphisms in the coiled-coil domain is in LD in Africa where it is at intermediate frequency. We might assume that this amino acid replacement could

modify the interaction with others molecules. However, we have little information about *CG15818*, it might be related to immunity as it has been shown that this gene is down-regulated in flies infected by *Nora* virus (Cordes *et al*. 2013). Moreover the CLECT domain is known to be involved functions such as extracellular matrix organization, endocytosis, complement activation, pathogen recognition, and cell-cell interactions. However, it is difficult to know how NS polymorphisms could modify the function of this gene or the interaction with other genes.

One interesting aspect was observed with respect to the two NS polymorphisms in the first exon of *chm* in the African population. First, we observed that they are in LD in the African (but not in Europe) and a haplotype structure is present in this candidate region at intermediate frequency which is characteristic of balancing selection. Indeed, these two SNPs have been observed in a study of Levine and Begun (2008). In this paper, the authors showed that the gene *chm* exhibits significant sequence differentiation between temperate and tropical populations from Australia and United States. Moreover, the higher sequence differentiation is restricted to a small region, at the 5' end of the gene which corresponds to our candidate region in Africa. They also observed in this region, a Tajima's *D* positive in the tropical population and negative in the temperate population in Australia. They found the two same NS polymorphisms of the first exon of *chm* and their results are similar to what we observed in the temperate European population and the tropical African population. The derived allele is at intermediate frequency in the tropical population whereas in the temperate population the ancestral allele is at high frequency (almost fixed). It seems to indicate that selection varies spatially depending on the environment (climate). In Europe, the derived allele could be disadvantageous and is therefore maintained at low frequency contrary to the African population.

The gene *chm* is a chromatin/histone remodeling gene so it is involved in the remodeling of the chromatin which contributes to transcriptional regulation of genes and plays an important role in many cellular events. The gene c*hm* maintains Hox gene silencing by other Polycomb group proteins (Grienenberger *et al.* 2002), it also cooperates with the JNK signaling pathway to promote transcriptional activity (Miotto *et al*. 2006). The JNK pathway regulates different

processes during *Drosophila* development including dorsal thorax closure during metamorphosis (Miotto *et al*. 2006). It will also contribute to stress response in Drosophila (Wang et al. 2003). The gene *chm* will interact with DFos and/or Djun to modulate the JNK pathway response to chemical and osmotic stress (Miotto and Struth, 2006). Indeed, the remodelling of DNA into proper chromatin structure is sensitive to environment such as temperature (Fauvarque and Dura, 1993) and to stress (Leibovitch *et al*. 2002; Wang and Brock, 2003, Smith *et al*. 2004). Following the environment, *Drosophila* will be under different stress such as cold temperatures, UV exposure, osmotic stress and/or heat shock. These stresses can induce a change in the expression of genes regulated by genes remodeling the chromatin such as *chm* (Berthiaume *et al*. 2006). Consequently, the gene *chm* has to adapt to a novel habitat in order to be less sensitive to stress and to maintain the transcription of genes such as developmental genes. Moreover, the adaptation to a new environment will induce different selective pressure and it might explain that balancing selection is not acting on the same region of this gene.

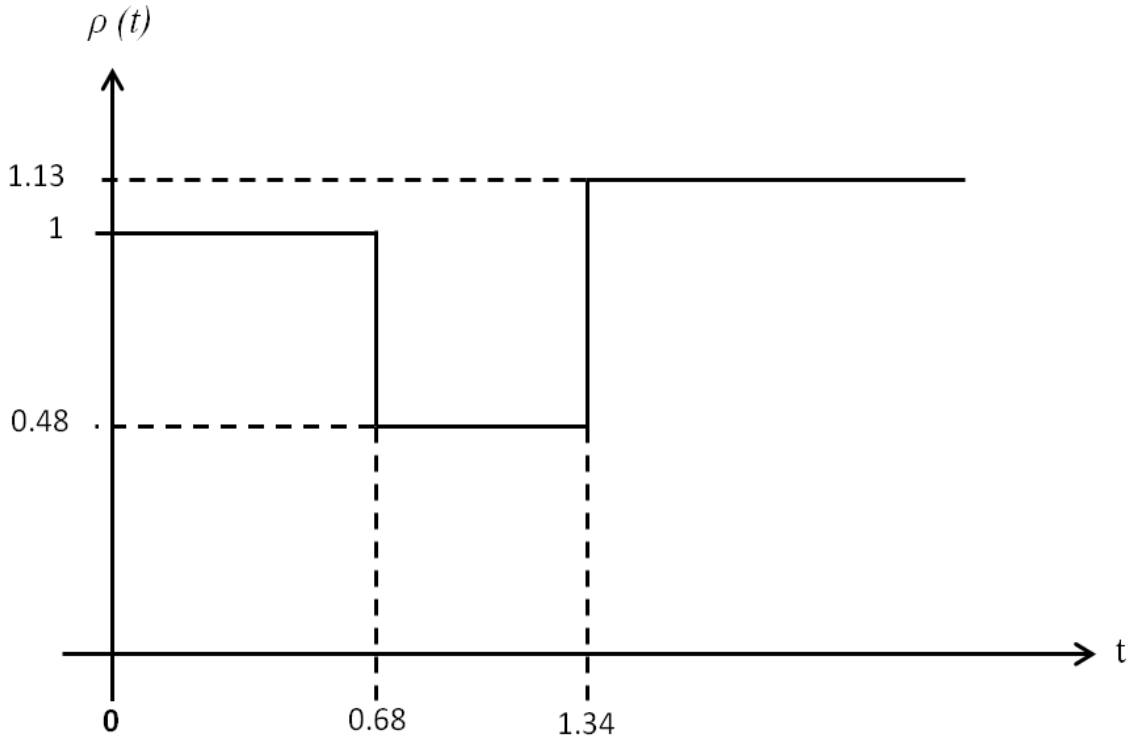
4.5 Conclusion and perspectives

We identified candidate genes under balancing selection in two populations of *D. melanogaster*: 141 in the European population and 45 in the African one. The difference between both populations is mainly due to an excess of candidate genes on the European X chromosome, which is likely due to false positives. Correcting for this effect reduces the difference between both populations considerably. Among the candidate genes detected in the European population there is an overrepresentation of genes involved in neuronal development and circadian rhythm. Other genes are involved in immunity including the top candidates. These top genes are also involved in behavioral plasticity, memory, neuromuscular junctions or neurogenesis. Moreover, when we looked in more details at two of our candidate genes (*chm* and *CG15818*), we observed patterns of balancing selection such as LD, haplotypes and NS polymorphisms at intermediate frequency. Moreover, selection seems to be environment-dependent as it acted on some gene

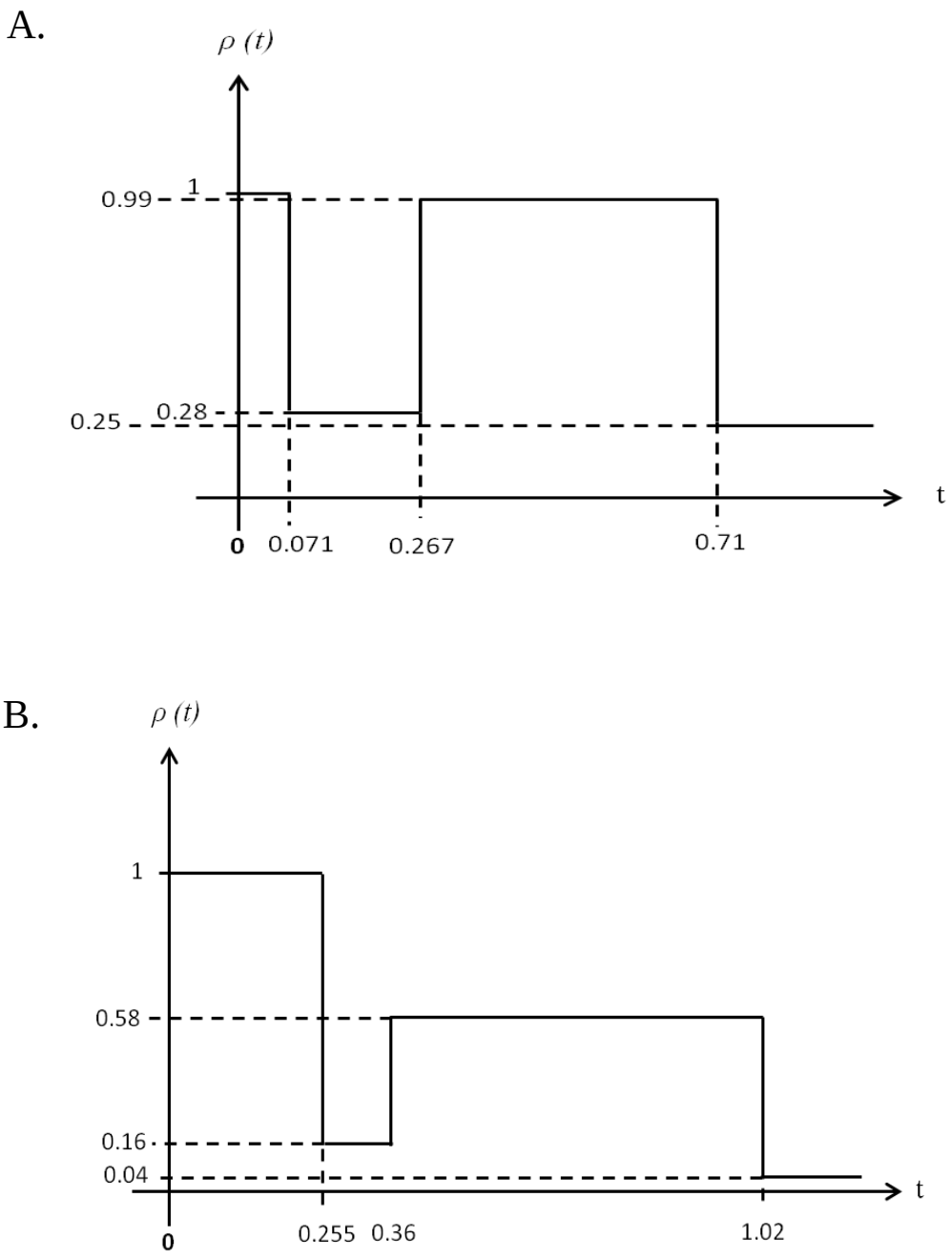
regions only in one of our populations but not the other. These results confirm that our method allowed us to detect genes under balancing selection in *D. melanogaster*. These candidate genes may serve as a starting point for a more detailed analysis. It will be useful to do the same kind of analysis that we did for *chm* and *CG15818* on the other candidate genes to confirm that they are under balancing selection in particular for the large genes which might be false positives.

APPENDIX A

MATERIALS AND METHODS



Appendix A1: Schematic representation of the demographic model for the African population. We used 20 lines from Gikongoro in Rwanda. The x-axis represents the time t in *Ne* generations backward in time and the y-axis represents the population size at the time t in *Ne*. *Ne* is defined as the current effective population size. Based on an estimated mutation rate of 1.5 x 10^{-9}, *Ne* is estimated to be equal to 3.4 x $10^{6.}$ We assume an instantaneous size changes and we estimate an ancestral size of 1.13 *Ne* until 1.34 *Ne* generations before sampling, when the population size dropped to about 0.48 *Ne* before again recovering at 0.68 *Ne* generations ago, to its current size.



Appendix A2: **Demographic models for the X chromosome for the European (A) and African (B) populations.** The x-axis represents the time t in *Ne* generations backwards in time and the y-axis represent the population size at the time t in *Ne*. *Ne* is defined as the current effective population size. Based on an estimated mutation rate of 1.5×10^{-9}, *Ne* is estimated to be equal to 1.09×10^{6} in the European population and 1.62×10^{6} in the African population.

APPENDIX B

RESULTS

Appendix B1: List of candidate genes for the European population and the values of the significant statistics observed (*p*-value < 0.05) for a 1-kb window. The *p*-values of the statistics are indicated in parentheses.

FBgn number	Gene name	Gene size (in bp)	Chr.	θ_w		Tajima's D	
FBgn0039004	*Nup133*	4165	3R	0.0029	(10^{-4})	2.5160	(10^{-4})
FBgn0263986	*cd*	3031					
FBgn0039536	*unc80*	14660	3R	0.0028	(10^{-4})	2.0467	(10^{-4})
FBgn0001316	*klar*	106531	3L	0.0053	(10^{-4})	2.1501	(10^{-4})
FBgn0265988	*mv*	13872	3L	0.0046	(10^{-4})	2.5476	(10^{-4})
FBgn0085428	*Nox*	9216	2R	0.0064	(10^{-4})	2.1896	(10^{-4})
FBgn0002543	*lea*	40367	2L	0.0026	(10^{-4})	2.4560	(10^{-4})
FBgn0031424	*VGlut*	19002	2L	0.0038	(0.0023)	2.0871	(0.0444)
				0.0040	(10^{-4})	2.2109	(10^{-4})
FBgn0085424	*nub*	40510	2L	0.0064	(10^{-4})	2.3335	(10^{-4})
FBgn0086899	*tlk*	69763	X	0.0044	(10^{-4})	2.2383	(10^{-4})
FBgn0029504	*CHES-1-like*	27047	X	0.0046	(10^{-4})	2.4145	(10^{-4})
				0.0028	(0.0374)	2.2258	(0.0107)
FBgn0030244	*CG2157*	1421	X	0.0069	(10^{-4})	2.3996	(10^{-4})
FBgn0030245	*CG1637*	7299			(10^{-4})		(10^{-4})
FBgn0030286	*CG1657*	8686	X	0.0042	(10^{-4})	1.9313	(0.0413)
				0.0041	(10^{-4})	2.0603	(10^{-4})
FBgn0267001	*Ten-a*	291516	X	0.0051	(10^{-4})	2.3024	(10^{-4})
FBgn0030412	*Tomosyn*	21837	X	0.0049	(10^{-4})	2.2223	(10^{-4})
FBgn0030466	*CG15744*	6382	X	0.0056	(10^{-4})	2.2758	(10^{-4})
				0.0046	(0.0005)	2.1004	(0.0107)
				0.0058	(10^{-4})	2.3005	(10^{-4})
FBgn0086680	*vvl*	4783	3L	0.0016	(0.0004)	1.8337	(10^{-4})
FBgn0013765	*cnn*	11161	2R	0.0052	(0.0004)	2.0192	(10^{-4})
FBgn0050062	*CG30062*	609					

FBgn0024315 FBgn0024319	*Picot* *Nach*	12352 1841	2R	0.0075	(0.0004)	2.1610	(10^{-4})
FBgn0024992	*CG2658*	3819	X	0.0055	(0.0005)	2.2549	(10^{-4})
FBgn0052791 FBgn0052792	*CG32791* *ppk8*	64404 14188	X	0.0069	(0.0005)	2.2098	(10^{-4})
FBgn0030613	*Rab3-GEF*	24062	X	0.0029 0.0037 0.0039	(0.0399) (0.0005) (0.0011)	2.4475 2.3121 2.3986	(10^{-4}) (10^{-4}) (10^{-4})
FBgn0260486	*Ziz*	23794	2L	0.0041 0.0034	(0,0006) (0.0375)	2.2383 2.3238	(10^{-4}) (10^{-4})
FBgn0261041	*stj*	13777	2R	0.0095	(0.0009)	1.8089	(10^{-4})
FBgn0030081	*CG7246*	2234	X	0.0042	(0.0011)	2.3344	(10^{-4})
FBgn0034599	*hng1*	1080	2R	0.0068	(0.0013)	2.4741	(10^{-4})
FBgn0029896	*CG3168*	19982	X	0.0048	(0.0017)	2.3344	(10^{-4})
FBgn0261260	*mgl*	141633	X	0.0036 0.0060	(0.0017) (0.0047)	2.4179 2.0536	(10^{-4}) (0.0107)
FBgn0000028	*acj6*	28825	X	0.0074	(0.0017)	2.2739	(10^{-4})
FBgn0023506	*Es2*	1870	X	0.0031	(0.0415)	2.2556	(10^{-4})
FBgn0023506 FBgn0014032	*Es2* *Sptr*	1870 1142	X	0.0042	(0.0028)	2.4465	(10^{-4})
FBgn0025378	*CG3795*	1376	X	0.0023	(0.0040)	2.3519	(10^{-4})
FBgn0001624	*dlg1*	40111	X	0.0036 0.0033	(0.00457) (0.0090)	2.6062 2.4413	(10^{-4}) (10^{-4})
FBgn0051145	*CG31145*	63908	3R	0.0050	(0.0045)	2.3290	(10^{-4})
FBgn0262733	*Src64B*	33749	3L	0.0039	(0.0053)	2.4556	(10^{-4})
FBgn0025833	*CG8910*	20928	2R	0.0082	(0.00553)	1.7806	(10^{-4})
FBgn0050263	*CG30263*	10099	2R	0.0039	(0.0059)	2.3238	(10^{-4})
FBgn0030884	*CG6847*	12701	X	0.0035	(0.0066)	2.3050	(10^{-4})
FBgn0263111	*cac*	53807	X	0.0031 0.0032	(0.00902) (0.0085)	2.1962 2.3175	(10^{-4}) (10^{-4})

FBgn0262872	*milt*	18187	2L	0.0027	(0.00881)	2.3792	(10^{-4})
FBgn0020653	*Trxr-1*	5505	X	0.0035	(0.0090)	2.4556	(10^{-4})
FBgn0004168	*5-HT1A*	54803	2R	0.0033	(0.0104)	2.5858	(10^{-4})
FBgn0050295	*Ipk1*	8354	2R	0.0045	(0.0104)	2.1563	(10^{-4})
FBgn0038880	*SIFaR*	17334	3R	0.0049	(0.0112)	2.6636	(10^{-4})
FBgn0016081	*fry*	47269	3L	0.0051	(0.0120)	2.1338	(10^{-4})
FBgn0051774	*fred*	88785	2L	0.0043	(0.01230)	2.4333	(10^{-4})
FBgn0028369	*CG3603*	1086	X	0.0074	(0.00057)	2.1427	(0.0413)
FBgn0028369	*kirre*	394148					
FBgn0028369	*kirre*	394148	X	0.0042	(0.01250)	2.1582	(10^{-4})
FBgn0003463	*sog*	30151	X	0.0046	(10^{-4})	2.0855	(0.0413)
				0.0031	(0.0137)	2.3381	(10^{-4})
FBgn0027093	*Aats-arg*	2376	X	0.0031	(0.0141)	2.5032	(10^{-4})
				0.0042	(0.0343)	2.6227	(10^{-4})
FBgn0262719	*CG43163*	34339	3L	0.0043	(0.0148)	2.1742	(10^{-4})
FBgn0017561	*Ork1*	13196	X	0.0031	(0.0149)	2.4048	(10^{-4})
FBgn0267253	*CG32700*	48012	X	0.0046	(0.0152)	2.3665	(10^{-4})
FBgn0004197	*Ser*	22177	3R	0.0059	(0.0156)	2.2602	(10^{-4})
FBgn0034974	*CG16786*	4505	2R	0.0055	(0.0164)	2.2882	(10^{-4})
FBgn0086129	*snama*	5410					
FBgn0029167	*Hml*	13947	3L	0.0051	(0.00467)	2.3466	(0.0414)
				0.0043	(0.0173)	2.6227	(10^{-4})
FBgn0053223	*CG33223*	18195	X	0.0070	(0.01836)	2.5690	(10^{-4})
FBgn0030041	*CG12116*	1569					
FBgn0025741	*PlexA*	17206	X	0.0079	(0.0187)	2.4014	(10^{-4})
				0.0030	(0.0454)	2.4018	(10^{-4})
FBgn0027505	*Rab3-GAP*	4880	2L	0.0034	(0.0197)	2.2297	(10^{-4})
FBgn0263512	*Vsx2*	31256	X	0.0032	(0.0198)	2.4703	(10^{-4})
FBgn0030055	*CG12772*	3849	X	0.0033	(0.0226)	2.4413	(10^{-4})
FBgn0086757	*cbs*	2807	2R	0.0026	(0.0240)	2.1234	(10^{-4})

FBgn0024973	*CG2701*	2689	X	0.0044	(0.0240)	2.4367	(10^{-4})
FBgn0023458	*Rbcn-3A*	14742	X	0.0020	(0.0241)	2.2509	(10^{-4})
FBgn0031263	*CG2789*	911	2L	0.0062	(0.0243)	2.4113	(10^{-4})
FBgn0085446	*CG34417*	48402	X	0.0034	(0.0248)	2.4620	(10^{-4})
FBgn0039225	*Ets96B*	9901	3R	0.0052	(0.0249)	2.6073	(10^{-4})
FBgn0030011	*Gbeta5*	1544	X	0.0038	(0.0251)	2.3803	(10^{-4})
FBgn0261509 FBgn0051935	*haf* *CG31935*	53275 16335	2L	0.0039	(0.0271)	2.4703	(10^{-4})
FBgn0027496	*epsilonCOP*	1104	2L	0.0043	(0.0271)	2.1422	(10^{-4})
FBgn0038727	*CG7432*	5820	3R	0.0025	(0.0272)	2.0723	(10^{-4})
FBgn0261931	*CG42797*	14297	X	0.0049	(0.0279)	2.4626	(10^{-4})
FBgn0033766	*CG8771*	6212	2R	0.0034	(0.0283)	2.2862	(10^{-4})
FBgn0003380	*Sh*	138938	X	0.0045	(0.0289)	2.5907	(10^{-4})
FBgn0000259	*CkIIbeta*	9054	X	0.0042	(0.0319)	2.3050	(10^{-4})
FBgn0029922 FBgn0052732	*CG14431* *CG32732*	60552 10718	X	0.0044	(0.0323)	2.4465	(10^{-4})
FBgn0261985 FBgn0035131	*Ptpmeg* *mthl9*	27953 2487	3L	0.0044	(0.0336)	1.9364	(10^{-4})
FBgn0050395	*CG30395*	4707	2R	0.0052	(0.0363)	1.9662	(10^{-4})
FBgn0023524	*CG3078*	5460	X	0.0024	(0.0369)	2.3025	(10^{-4})
FBgn0000064	*Ald*	7522	3R	0.0065	(0.0374)	2.1660	(10^{-4})
FBgn0003301	*rut*	38595	X	0.0030	(0.0390)	2.3333	(10^{-4})
FBgn0011589	*Elk*	52756	2R	0.0039	(0.0432)	2.0525	(10^{-4})
FBgn0265597	*rad*	90455	X	0.0028	(0.0448)	2.5388	(10^{-4})
FBgn0259735	*CG42389*	118202	2L	0.0025	(0.0469)	2.3025	(10^{-4})
FBgn0011653	*mas*	5798	3L	0.0022	(0.0485)	2.1484	(10^{-4})
FBgn0004657	*mys*	8592	X	0.0036	(0.0486)	2.2485	(10^{-4})
FBgn0260439	*Pp2A-29B*	4052	2L	0.0031	(0.0494)	2.2969	(10^{-4})
FBgn0034598	*CG4266*	6946	2R	0.0051	(0.0499)	2.3859	(10^{-4})
FBgn0030274	*Lint-1*	3269	X	0.0052	(0.0017)	2.2453	(0.0107)

FBgn0000479	*dnc*	167327	X	0.0044	(0.0489)	2.0942	(0.0107)
FBgn0029881	*pigs*	52953	X	0.0054	(10^{-4})	2.0352	(0.0211)
FBgn0026181	*Rok*	13402	X	0.0040	(10^{-4})	2.1217	(0.0211)
FBgn0004045	*Yp1*	1687	X	0.0056	(0.0005)	2.1276	(0.0211)
FBgn0030174	*CG15312*	15239					
FBgn0266350	*CG12535*	60999	X	0.0042	(0.0223)	2.1920	(0.0211)
FBgn0264979	*CG4267*	2896	2L	0.0044	(0.0058)	2.0781	(0.0230)
FBgn0260933	*rempA*	6568	2L	0.0085	(0.0155)	2.2570	(0.0230)
FBgn0032085	*CG9555*	1897	2L	0.0078	(0.0186)	1.9049	(0.0230)
FBgn0013746	*alien*	3561					
FBgn0032086	*CG17906*	1226					
FBgn0028387	*chm*	10319	2L	0.0042	(0.0284)	2.2383	(0.0230)
FBgn0034229	*CG4847*	2342	2R	0.0054	(0.0039)	1.9804	(0.0241)
FBgn0034230	*CG4853*	3228					
FBgn0034776	*CG13527*	1716	2R	0.0081	(0.0039)	1.5816	(0.0241)
FBgn0259145	*CG42260*	57290					
FBgn0028473	*CG8801*	2469	2R	0.0035	(0.0188)	2.1318	(0.0241)
FBgn0033408	*CG8800*	876					
FBgn0034282	*Mapmodulin*	5837	2R	0.0066	(0.0246)	2.1790	(0.0241)
FBgn0039234	*nct*	3268	3R	0.0072	(10^{-4})	2.1898	(0.0308)
FBgn0025680	*cry*	3288	3R	0.0047	(0.0112)	2.2047	(0.0308)
				0.0075	(0.0015)	2.1369	(0.0308)
FBgn0038660	*CG14291*	1869	3R	0.0083	(0.0022)	2.0666	(0.0308)
FBgn0261262	*CG42613*	48892	3R	0.0024	(0.0197)	2.0147	(0.0308)
FBgn0263983	*CG43732*	26361					
FBgn0022800	*Cad96Ca*	10712	3R	0.0051	(0.0374)	2.3091	(0.0308)
FBgn0039290	*CG13654*	14437					
FBgn0029688	*lva*	9784	X	0.0045	(10^{-4})	2.0109	(0.0313)
FBgn0266199	*CG43902*	36903	X	0.0045	(10^{-4})	2.0621	(0.0313)
FBgn0000709	*fliI*	5347	X	0.0039	(0.0265)	2.1920	(0.03138)

FBgn0029939	*CG9650*	104209	X	0.0036	(0.0116)	2.1544	(0.0413)
FBgn0036022	*CG8329*	909	3L	0.0062	(10^{-4})	2.1894	(0.0414)
FBgn0262524 FBgn0004926	*ver* *eIF-2beta*	1163 *1638*	3L	0.0049	(0.0008)	2.1153	(0.0414)
FBgn0010825	*Gug*	22747	3L	0.0042	(0.0035)	2.1390	(0.04144)
FBgn0052062	*A2bp1*	112600	3L	0.0080	(0.0095)	2.2025	(0.0414)
FBgn0266084	*Fhos*	45100	3L	0.0032	(0.0245)	2.2143	(0.0414)
FBgn0264815	*Pde1c*	114416	2L	0.0061	(10^{-4})	2.2595	(0.0444)
FBgn0032382	*Mal-B2*	2923	2L	0.0050	(10^{-4})	2.1477	(0.0444)
FBgn0032036 FBgn0262001	*CG13384* *CG42819*	4349 603	2L	0.0042	(0.0023)	2.0621	(0.0444)
FBgn0016059	*Sema-1b*	10660	2R	0.0055	(10^{-4})	1.8828	(0.0471)
FBgn0003545	*sub*	2578	2R	0.0085	(10^{-4})	1.9183	(0.0471)
FBgn0010434	*cora*	14688	2R	0.0049	(10^{-4})	2.0301	(0.0471)

Appendix B2: List of candidate genes for the African population and the values of the significant statistics observed (*p*-value < 0.05) for a 1-kb window. The *p*-value of the statistics is indicated in parentheses.

FBgn number	Gene name	Gene size (in bp)	Chr.	θ_w		Tajima's D	
FBgn0040076	*primo-2*	1818	3R	0.0105	(10^{-4})	2.2764	(10^{-4})
FBgn0040077	*primo-1*	1818					
FBgn0039519	*Cyp6a18*	4229	3R	0.0130	(10^{-4})	2.1106	(10^{-4})
FBgn0036173	*CG7394*	2316	3L	0.0103	(10^{-4})	2.1201	(10^{-4})
FBgn0261853	*CG42782*	281	2R	0.0205	(10^{-4})	1.9016	(10^{-4})
FBgn0031910	*CG15818*	1320	2L	0.0121	(10^{-4})	1.8187	(10^{-4})
FBgn0028387	*chm*	10319	2L	0.0080	(10^{-4})	2.5511	(10^{-4})
FBgn0028899	*CG31817*	7325	2L	0.0144	(10^{-4})	1.8100	(10^{-4})
FBgn0259735	*CG42389*	118202	2L	0.0202	(10^{-4})	1.1303	(10^{-4})
FBgn0036817	*CG6865*	1867	3L	0.0133	(0.0001)	1.8804	(10^{-4})
FBgn0001258	*ImpL3*	3589	3L	0.0156	(0.0003)	1.6713	(10^{-4})
FBgn0051469	*CG31469*	780	3R	0.0077	(0.0004)	2.3375	(10^{-4})
FBgn0033732	*CG13157*	2187	2R	0.0185	(0.0030)	1.3831	(10^{-4})
FBgn0033480	*mRpL42*	592	2R	0.0050	(0.0038)	2.4522	(10^{-4})
FBgn0013435	*cdc2rk*	1485					
FBgn0038938	*CG7084*	5615	3R	0.0084	(0.0043)	1.9567	(10^{-4})
FBgn0042111	*CG18766*	2914	3R	0.0089	(0.0043)	1.9087	(10^{-4})
FBgn0025678	*CaBP1*	1849	2L	0.0053	(0.0091)	2.5415	(10^{-4})
FBgn0025621	*CG16989*	3914	X	0.0038	(0.0195)	1.8050	(10^{-4})
FBgn0011274	*Dif*	19913	2L	0.0122	(0.0209)	1.8945	(10^{-4})
FBgn0050049	*CG30049*	3416	2R	0.0087	(0.0253)	1.6336	(10^{-4})
FBgn0028506	*CG4455*	1605	2L	0.0060	(0.0339)	2.2241	(10^{-4})
FBgn0038087	*beat-Va*	7859	3R	0.0088	(0.0351)	2.1085	(10^{-4})
FBgn0038653	*CG18208*	49944	3R	0.0075	(0.0351)	1.2693	(10^{-4})
FBgn0020503	*CLIP-190*	24998	2L	0.0075	(0.0442)	1.9602	(10^{-4})
FBgn0266064	*GlyS*	5688	3R	0.0036	(10^{-4})	2.0604	(0.0183)

FBgn0261984	*Ire1*	7383	3R	0.0076	(0.0468)	1.6305	(0.0183)
FBgn0038737	*CG11447*	990					
FBgn0016081	*fry*	47269	3L	0.0103	(10^{-4})	1.8448	(0.0218)
FBgn0036024	*CG18180*	900					
FBgn0036489	*CG7011*	2752	3L	0.0079	(-0.0096)	1.8362	(0.0218)
FBgn0036488	*CG6878*	653					
FBgn0036680	*Cpr73D*	3849	3L	0.0099	(0.0234)	2.0043	(0.0218)
FBgn0261999	*CG42817*	10052	2L	0.0135	(10^{-4})	1.4012	(0.0251)
FBgn0027094	*Aats-ala*	3502	2L	0.0057	(0.0003)	2.1105	(0.0251)
FBgn0051928	*CG31928*	1564	2L	0.0103	(0.0005)	2.1035	(0.0251)
FBgn0053128	*CG33128*	1386					
FBgn0261597	*RpS26*	1942	2L	0.0154	(0.0160)	1.2594	(0.0251)
FBgn0051926	*CG31926*	1611	2L	0.0075	(0.0236)	2.1898	(0.0251)
FBgn0262024	*CG42835*	351	3R	0.0128	(0.0043)	1.9596	(0.0361)
FBgn0038509	*CG14332*	512					
FBgn0262869	*Gfrl*	106130	3R	0.0065	(0.0136)	2.0865	(0.0361)
FBgn0015338	*CG5861*	1497	2L	0.0083	(0.0010)	1.6931	(0.0491)
FBgn0011708	*Syx5*	1928					
FBgn0002023	*Lim3*	29171	2L	0.0081	(0.0176)	1.4930	(0.0491)

Appendix B3: List of enriched GO terms and the genes grouped under this term for the European population in biological processes. The GO terms shown highlighted in gray are the name of the group (the most significant term of the group).

GO Group	GO term	Genes	*p*-value
Group 1	**cell morphogenesis involved in differentiation**	***Fhos, Ptpmeg, Sh, Src64B, acj6, chm, cnn, dnc, eIF-2beta, fry, haf, lea, lva, mys, nct, nub, pigs, plexA, rok, rut, vvl***	**0.0002**
	cell morphogenesis involved in neuron differentiation	*Ptpmeg, Sh, Src64B, acj6, chm, cnn, dnc, eIF-2beta, fry, haf, lea, lva, mys, nct, nub, plexA, rok, rut, vvl*	0.0004
	chemical synaptic transmission	*Sh, VGlut, cac, dlg1, dnc, rab3-GAP, rut, stj, tomosyn*	0.0021
	axogenesis	*Ptpmeg, Sh, Src64B, acj6, dnc, eIF-2beta, haf, lea, mys, plexA, rok, rut, vvl*	0.0029
	developmental growth	*CG7246, Ptpmeg, Sh, Src64B, Ten-a, cac, dlg1, dnc, mys, rok, rut, stj, tlk*	0.003
	circadian behavior	*5-HT1A, CkIIbeta, Ork1, Sh, cry, dlg1*	0.0037
	developmental growth involved in morphogenesis	*Ptpmeg, Sh, dnc, mys, rok, rut*	0.0038
	regulation of circadian sleep/wake cycle, sleep	*5-HT1A, Sh, cry*	0.0073
	axon extension	*Ptpmeg, Sh, dnc, rut*	0.0079
	sleep	*5-HT1A, CG42613, CG8329, Sh, cry, rut*	0.0128
	locomotor rhythm	*CkIIbeta, Ork1, cry, dlg1*	0.0165
	mating behavior	*5-HT1A, Pde1c, Sh, cac, dlg1*	0.0181
	regulation of behavior	*5-HT1A, Sh, cry, rut*	0.0226
	detection of light stimulus	*Sh, cac, cry, milt*	0.0241
	neuromuscular junction development	*Src64B, Ten-a, cac, dlg1, rut, stj*	0.0249
	modulation of synaptic transmission	*Sh, cac, rab3-GAP, tomosyn*	0.0271

GO Group	GO term	Genes	*p*-value
Group 2	**regulation of stress fiber assembly**	***Fhos, mys, rok***	**0.0013**
	actomyosin structure organization	*Fhos, Src64B, flil, mys, rok*	0.0022
	regulation of cell morphogenesis involved in differentiation	*Fhos, fry, lva, plexA, rok, vvl*	0.003
	actin filament bundle assembly	*Fhos, Src64B, mys, rok*	0.0031
	regulation of cellular component movement	*Fhos, lea, mys, plexA*	0.0077
	regulation of anatomical structure morphogenesis	*Fhos, fry, kirre, lva, mys, nct, plexA, rok, tlk, vvl*	0.0078
	regulation of cell morphogenesis	*Fhos, fry, lva, mys, plexA, rok, tlk, vvl*	0.0079
	regulation of locomotion	*Fhos, lea, mys, plexA*	0.0083
	regulation of cell migration	*Fhos, lea, mys*	0.0086
	regulation of cytoskeleton organization	*Fhos, Src64B, cnn, mys, rok*	0.0123
	regulation of dendrite morphogenesis	*fry, lva, vvl*	0.0158
	regulation of neuron differentiation	*fry, lva, nct, plexA, rok, vvl*	0.0165
	cell junction assembly	*Src64B, kirre, rok*	0.0306
	heart development	*CHES-1-like, cora, lea, mys*	0.0425
Group 3	**central complex development**	***Ptpmeg, Ten-a, lea***	**0.0013**
	brain development	*CG4853, CkIIbeta, Ptpmeg, Src64B, Ten-a, lea, vvl*	0.0031
	mushroom body development	*CG4853, CkIIbeta, Ptpmeg, Src64B, lea*	0.0083
	neuron recognition	*Ten-a, acj6, eIF-2beta, fry, lea, plexA*	0.009
	synaptic target recognition	*Ten-a, acj6, lea*	0.0315
Group 4	**organophosphate metabolic process**	***Ald, Ipk1, Pde1c, rut***	**0.0315**
	nucleotide metabolic process	*Ald, Pde1c, rut*	0.0417
Group 5	**establishment of localization by movement along microtubule**	***klar, milt, rempA***	**0.0289**
Group 6	**heart process**	***Ork1, cac, cora***	**0.0306**
Group 7	**imaginal disc-derived wing hair organization**	***cora, fry, rok***	**0.042**
Group 8	**positive regulation of developmental growth**	***CG7246, cac, dlg1, tlk***	**0.0477**

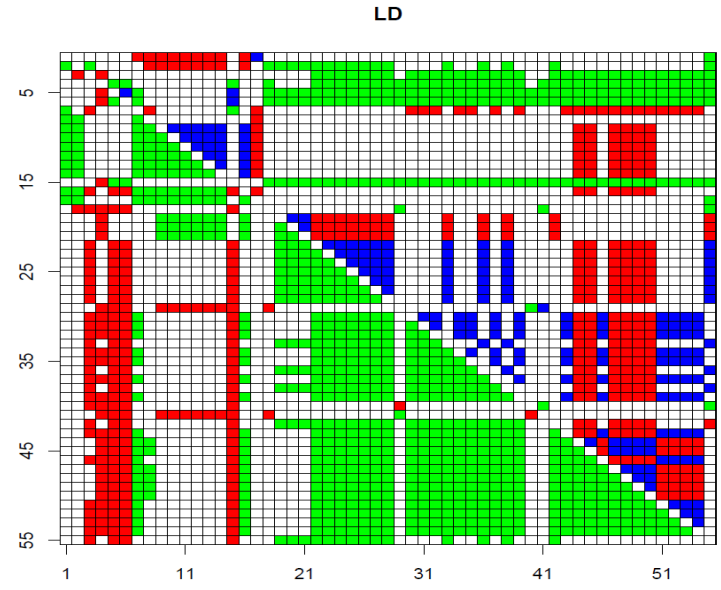
Appendix B4: List of enriched GO terms and the genes grouped under this term for the European population for molecular function, cellular component and KEGG and Reactome pathways.

GO term	Genes	*p*-value
Molecular Function		
cation channel activity	*Sh, cac, Ork1*	0.0216
transcription cofactor activity	*alien, chm, Gug, acj6*	0.0287
protein homodimerization activity	*Ork1, Hml, Trxr-1, lea, alien*	0.0385
Cellular component		
apical part of cell	*rok, fry, Ser, cac, Cad96Ca, Megalin, dlg1*	0.0160
plasma membrane region	*Megalin, Cad96Ca, cac, Ser, mys, dlg1, Ten-a*	0.0236
microtubule	*klar, pigs, sub*	0.0495
KEGG and Reactome		
ECM-receptor interaction	*Hml, CG3168, mys*	0.0045
neuronal system	*rut, dlg1, Vglut, elk, Ork1, Sh*	0.0088
G-alpha (s) signaling events	*dnc, Pde1c, rut*	0.0108
potassium channels	*elk, Ork1, Sh*	0.0127
cell-cell communication	*mys, kirre, Src64B*	0.0138
TGF-beta signaling pathway	*rok, sog, Pp2A-29B*	0.0221
EPH-Ephrin signaling	*Src64B, nct, rok*	0.0353
digestion of dietary lipid	*CG6847, CG4267, Yp1*	0.0434

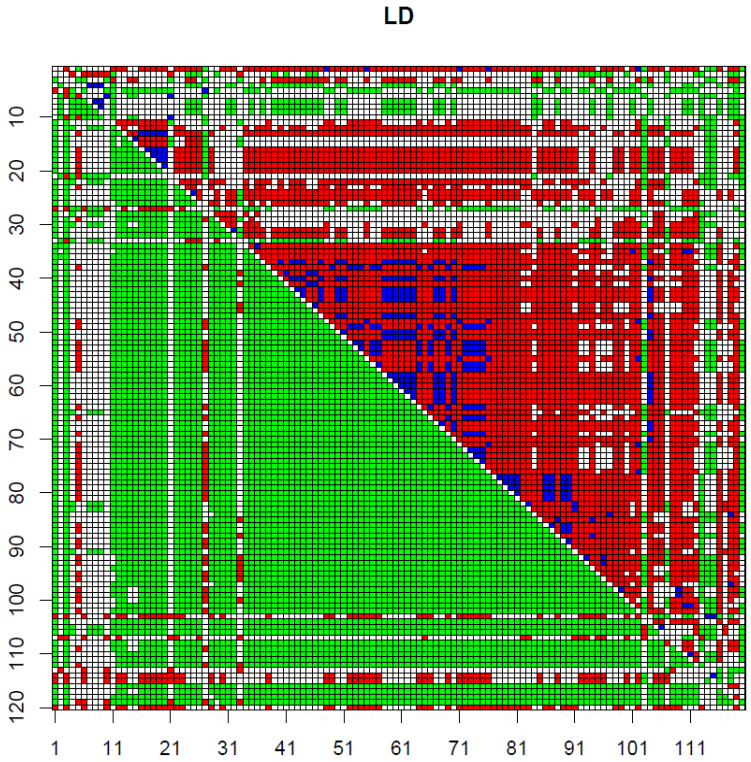
Appendix B5: List of enriched GO terms and the genes grouped under this term for the autosomal genes in the European population.

GO terms	Genes	*p*-value
Biological process		
regulation of circadian sleep/wake cycle, sleep	*5-HT1A, Ets96B, cry*	0.002
mushroom body development	*CG4853, Ptpmeg, Src64B, lea*	0.004
organophosphate metabolic process	*Ald, Ipk1, Pde1c, rut*	0.007
nucleotide metabolic process	*Ald, Pde1c, rut*	0.012
Molecular process		
transcription cofactor activity	*Gug, alien, chm*	0.005

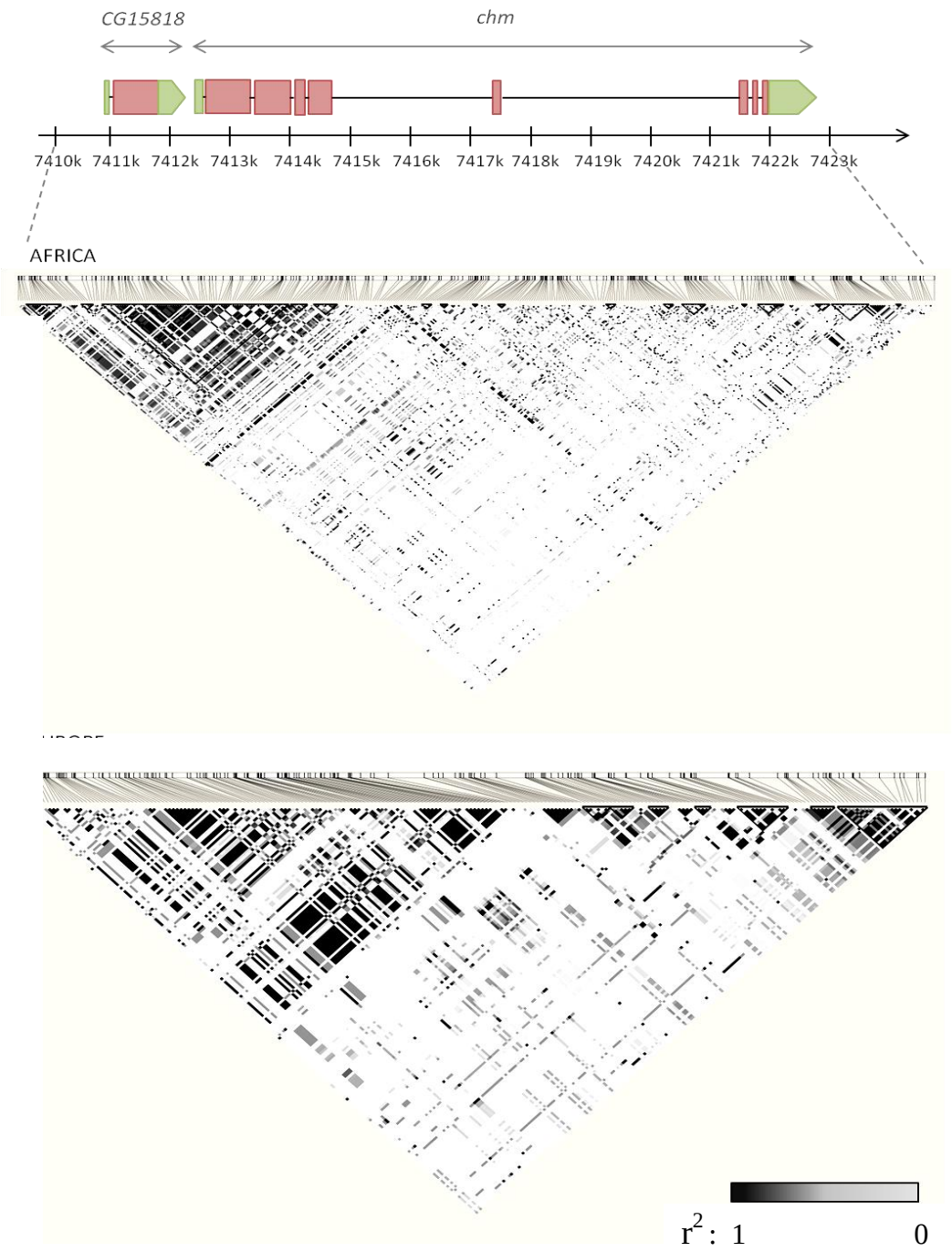
A.



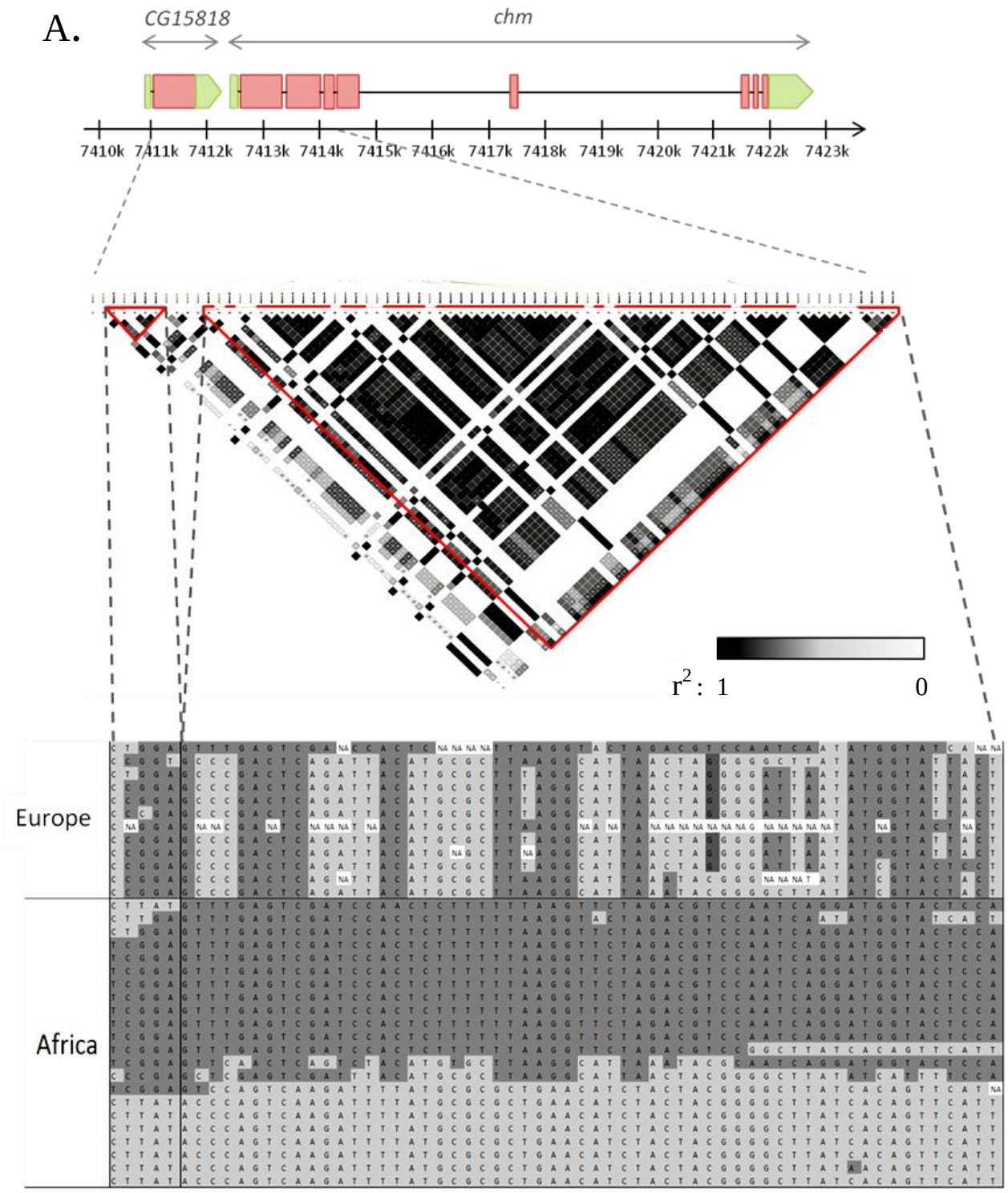
B.

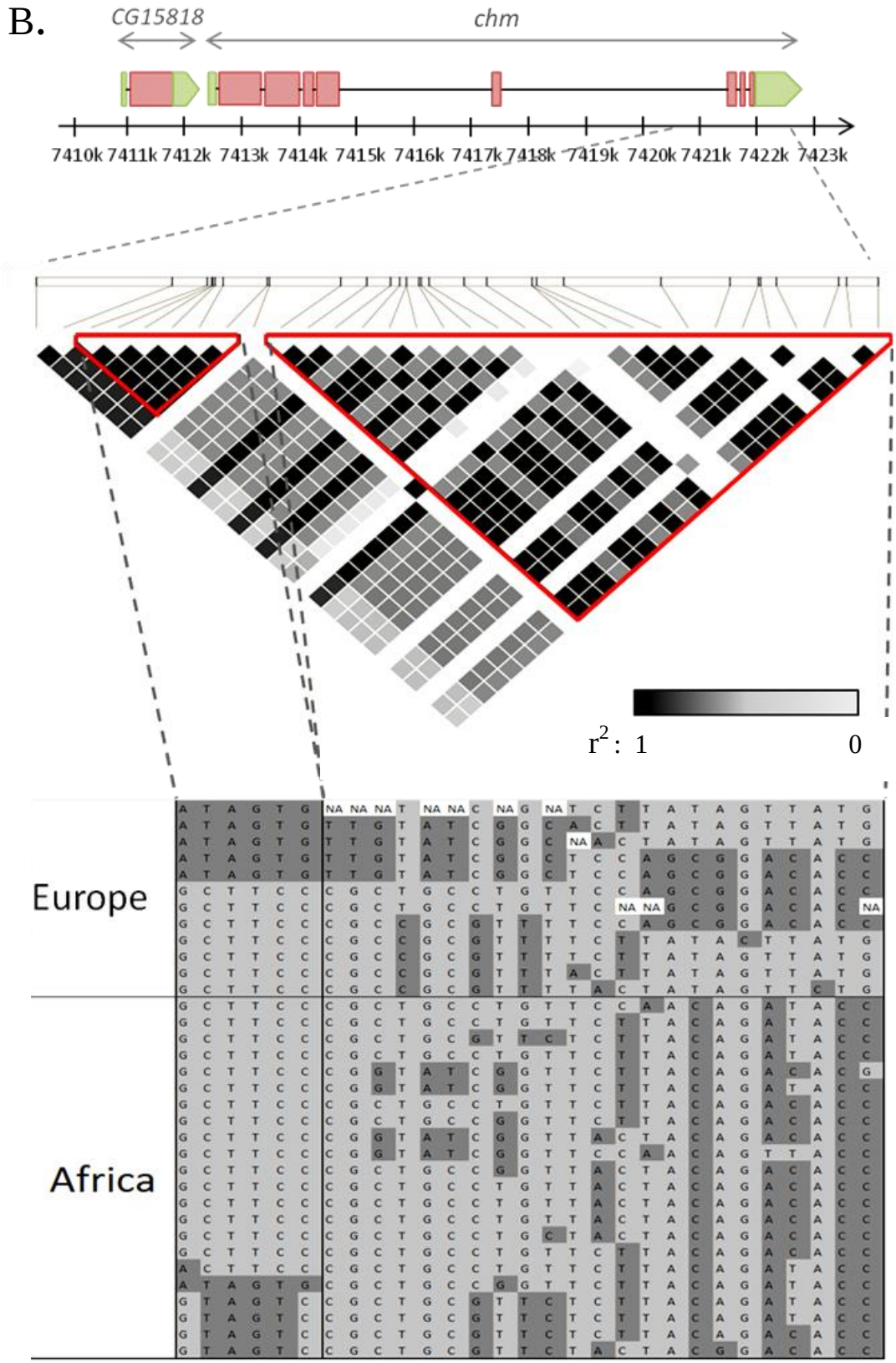


Appendix B6: Linkage disequilibrium (LD) matrix of SNPs in 5 kb region (2 kb around the candidate region) for the gene *chm* and *CG15818*. A. shows the candidate region in Europe for the European population and B. the candidate region in Africa for the African population. Patterns of LD (r^2) are shown above the diagonal and *p-values* from Fisher's exact test below the diagonal. The green color corresponds to a value from 0 to 0.05, the white color is from 0.051 to 0.5, the red color is from 0.51 to 0.9 and blue color is from 0.91 to 1.



Appendix B7: Representation of LD (r^2) for the candidate genes *CG15818* and *chm*. LD patterns are presented for the African population (top) and the European population (bottom) across the complete region. The magnitude of pairwise LD is given by the color shading; black and grey colors represent different levels of LD in descending order. Black/grey pairs are statistically significant, while white pairs are not.


A.
CG15818
chm
7410k 7411k 7412k 7413k 7414k 7415k 7416k 7417k 7418k 7419k 7420k 7421k 7422k 7423k
r2: 1
0
Europe
Africa


B.
CG15818
chm
7410k 7411k 7412k 7413k 7414k 7415k 7416k 7417k 7418k 7419k 7420k 7421k 7422k 7423k
r2 : 1
0
Europe
Africa

Appendix B8: Haplotype blocks (clusters of high-LD variants) as defined by Haploview. (A) represents the candidate region in Africa. The LD plot shows the patterns within the African population. The candidate region in the European population is represented in (B). The LD plot shows the patterns within the European population. In (A) and (B), the magnitude of pairwise LD is given by the color shading; black and grey colors represent different levels of LD in descending order. Black/grey pairs are statistically significant, while white pairs are not. The red triangles represent the haplotype blocks. In the two polymorphisms tables, the two different shades of grey represent a different allele for each site.

BIBLIOGRAPHY

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000) The genome sequence of Drosophila melanogaster. Science, 287, 2185-2195.

Akey, J. M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res, 19, 711-722.

Alonso, S., Lopez, S., Izagirre, N. & de la Rua, C. (2008) Overdominance in the human genome and olfactory receptor activity. Mol Biol Evol, 25, 997-1001.

Amambua-Ngwa, A., Tetteh, K. K., Manske, M., Gomez-Escobar, N., Stewart, L. B., Deerhake, M. E., et al. (2012) Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. PLoS Genet, 8, e1002992.

Andres, A. M., Hubisz, M. J., Indap, A., Torgerson, D. G., Degenhardt, J. D., Boyko, A. R., et al. (2009) Targets of balancing selection in the human genome. Mol Biol Evol, 26, 2755-2764.

Asthana, S., Schmidt, S. & Sunyaev, S. (2005) A limited role for balancing selection. Trends Genet, 21, 30-32.

Ayala, F. J., Balakirev, E. S. & Saez, A. G. (2002) Genetic polymorphism at two linked loci, Sod and Est-6, in Drosophila melanogaster. Gene, 300, 19-29.

Balakirev, E. S. & Ayala, F. J. (2003) Nucleotide variation of the Est-6 gene region in natural populations of Drosophila melanogaster. Genetics, 165, 1901-1914.

Bamshad, M. & Wooding, S. P. (2003) Signatures of natural selection in the human genome. Nat Rev Genet, 4, 99-111.

Bangham, J., Obbard, D. J., Kim, K. W., Haddrill, P. R. & Jiggins, F. M. (2007) The age and evolution of an antiviral resistance mutation in Drosophila melanogaster. Proc Biol Sci, 274, 2027-2034.

Barreiro, L. B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J. K., et al. (2009) Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet, 5, e1000562.

Barreiro, L. B. & Quintana-Murci, L. (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet, 11, 17-30.

Barrett, J. C. (2009) Haploview: Visualization and analysis of SNP genotype data. Cold Spring Harb Protoc, 2009, pdb ip71.

Beaumont, M. A. & Balding, D. J. (2004) Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol, 13, 969-980.

Begun, D. J., Betancourt, A. J., Langley, C. H. & Stephan, W. (1999) Is the fast/slow allozyme variation at the Adh locus of Drosophila melanogaster an ancient balanced polymorphism? Mol Biol Evol, 16, 1816-1819.

Begun, D. J. & Whitley, P. (2000) Adaptive evolution of relish, a Drosophila NF-kappaB/IkappaB protein. Genetics, 154, 1231-1238.

Benjamini Y., Hochberg Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological), 57(1):289-300.

Bergelson, J., Dwyer, G. & Emerson, J. J. (2001) Models and data on plant-enemy coevolution. Annu Rev Genet, 35, 469^{-4}99.

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S. & Petrov, D. A. (2014) Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. PLoS Genet, 10, e1004775.

Bernatchez, L. & Landry, C. (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol, 16, 363-377.

Berthiaume, M., Boufaied, N., Moisan, A. & Gaudreau, L. (2006) High levels of oxidative stress globally inhibit gene transcription and histone acetylation. DNA Cell Biol, 25, 124-134.

Bindea, G., Galon, J. & Mlecnik, B. (2013) CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. Bioinformatics, 29, 661-663.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics, 25, 1091-1093.

Boutros, M., Agaisse, H. & Perrimon, N. (2002) Sequential activation of signaling pathways during innate immune responses in Drosophila. Dev Cell, 3, 711-722.

Brown, J. K. & Tellier, A. (2011) Plant-parasite coevolution: bridging the gap between genetics and ecology. Annu Rev Phytopathol, 49, 345-367.

Bubb, K. L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., et al. (2006) Scan of human genome reveals no new Loci under ancient balancing selection. Genetics, 173, 2165-2177.

Buchon, N., Silverman, N. & Cherry, S. (2014) Immunity in Drosophila melanogaster--from microbial recognition to whole-organism physiology. Nat Rev Immunol, 14, 796-810.

Cagliani, R., Fumagalli, M., Biasin, M., Piacentini, L., Riva, S., Pozzoli, U., et al. (2010) Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene. Hum Genet, 128, 577-588.

Cagliani, R., Guerini, F. R., Fumagalli, M., Riva, S., Agliardi, C., Galimberti, D., et al. (2012) A trans-specific polymorphism in ZC3HAV1 is maintained by long-standing balancing selection and may confer susceptibility to multiple sclerosis. Mol Biol Evol, 29, 1599-1613.

Carpenter, J., Hutter, S., Baines, J. F., Roller, J., Saminadin-Peter, S. S., Parsch, J., et al. (2009) The transcriptional response of Drosophila melanogaster to infection with the sigma virus (Rhabdoviridae). PLoS One, 4, e6838.

Charlesworth, B., Morgan, M. T. & Charlesworth, D. (1993) The effect of deleterious mutations on neutral molecular variation. Genetics, 134, 1289-1303.

Charlesworth, D. (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet, 2, e64.

Charlesworth, D. & Charlesworth, B. (2010) Evolutionary biology: the origins of two sexes. Curr Biol, 20, R519-521.

Chen, K., Richlitzki, A., Featherstone, D. E., Schwarzel, M. & Richmond, J. E. (2011) Tomosyn-dependent regulation of synaptic transmission is required for a late phase of associative odor memory. Proc Natl Acad Sci U S A, 108, 18482-18487.

Cohen, J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull, 70, 213-220.

Comeron, J. M. (2014) Background selection as baseline for nucleotide variation across the Drosophila genome. PLoS Genet, 10, e1004434.

Comeron, J. M., Ratnappan, R. & Bailin, S. (2012) The many landscapes of recombination in Drosophila melanogaster. PLoS Genet, 8, e1002905.

Contamine, D., Petitjean, A. M. & Ashburner, M. (1989) Genetic resistance to viral infection: the molecular cloning of a Drosophila gene that restricts infection by the rhabdovirus sigma. Genetics, 123, 525-533.

Cordes, E. J., Licking-Murray, K. D. & Carlson, K. A. (2013) Differential gene expression related to Nora virus infection of Drosophila melanogaster. Virus Res, 175, 95-100.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res, 39, D691-697.

Dantoft, W., Davis, M. M., Lindvall, J. M., Tang, X., Uvell, H., Junell, A., et al. (2013) The Oct1 homolog Nubbin is a repressor of NF-kappaB-dependent immune gene expression that increases the tolerance to gut microbiota. BMC Biol, 11, 99.

De Gregorio, E., Spellman, P. T., Rubin, G. M. & Lemaitre, B. (2001) Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. Proc Natl Acad Sci U S A, 98, 12590-12595.

De Gregorio, E., Spellman, P. T., Tzou, P., Rubin, G. M. & Lemaitre, B. (2002) The Toll and Imd pathways are the major regulators of the immune

response in Drosophila. EMBO J, 21, 2568-2579.

DeGiorgio, M., Lohmueller, K. E. & Nielsen, R. (2014) A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet, 10, e1004561.

Delph, L. F. & Kelly, J. K. (2014) On the importance of balancing selection in plants. New Phytol, 201, 45-56.

Dobzhansky, T. (1955) A review of some fundamental concepts and problems of population genetics. Cold Spring Harb Symp Quant Biol, 20, 1-15.

Duchen, P., Zivkovic, D., Hutter, S., Stephan, W. & Laurent, S. (2013) Demographic inference reveals African and European admixture in the North American Drosophila melanogaster population. Genetics, 193, 291-301.

Ebert, D. (2008) Host-parasite coevolution: Insights from the Daphnia-parasite model system. Curr Opin Microbiol, 11, 290-301.

Eizaguirre, C., Lenz, T. L., Kalbe, M. & Milinski, M. (2012) Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. Nat Commun, 3, 621.

Ewing, G. & Hermisson, J. (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics, 26, 2064-2065.

Fauvarque, M. O. & Dura, J. M. (1993) polyhomeotic regulatory sequences induce developmental regulator-dependent variegation and targeted P-element insertions in Drosophila. Genes Dev, 7, 1508-1520.

Ferrer-Admetlla, A., Bosch, E., Sikora, M., Marques-Bonet, T., Ramirez-Soriano, A., Muntasell, A., et al. (2008) Balancing selection is the main force shaping the evolution of innate immunity genes. J Immunol, 181, 1315-1322.

Fijarczyk, A. & Babik, W. (2015) Detecting balancing selection in genomes: limits and prospects. Mol Ecol, 24, 3529-3545.

Fiston-Lavier, A. S., Singh, N. D., Lipatov, M. & Petrov, D. A. (2010) Drosophila melanogaster recombination rate calculator. Gene, 463, 18-20.

Frank, S. A. (1992) Models of plant-pathogen coevolution. Trends Genet, 8, 213-219.

Fumagalli, M., Cagliani, R., Pozzoli, U., Riva, S., Comi, G. P., Menozzi, G., et al. (2009) Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res, 19, 199-212.

Fumagalli, M., Fracassetti, M., Cagliani, R., Forni, D., Pozzoli, U., Comi, G. P., et al. (2012) An evolutionary history of the selectin gene cluster in humans. Heredity (Edinb), 109, 117-126.

Fumagalli, M., Pozzoli, U., Cagliani, R., Comi, G. P., Riva, S., Clerici, M., et al. (2009) Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J Exp Med, 206, 1395-1408.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J.,
Blumenstiel, B., et al. (2002) The structure of haplotype blocks in the human genome. Science, 296, 2225-2229.

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

Gokhale, C. S., Papkou, A., Traulsen, A. & Schulenburg, H. (2013) Lotka-Volterra dynamics kills the Red Queen: population size fluctuations and associated stochasticity dramatically change host-parasite coevolution. BMC Evol Biol, 13, 254.

Grienenberger, A., Miotto, B., Sagnier, T., Cavalli, G., Schramke, V., Geli, V., et al. (2002) The MYST domain acetyltransferase Chameau functions in epigenetic mechanisms of transcriptional repression. Curr Biol, 12, 762-766.

Hedrick, P. W. (2011) Population genetics of malaria resistance in humans. Heredity (Edinb), 107, 283-304.

Hedrick, P. W. (2012) What is the evidence for heterozygote advantage selection? Trends Ecol Evol, 27, 698-704.

Hein, J., Schierup, M. H. & Wiuf, C. (2005) Gene genealogies, variation and evolution : a primer in coalescent theory. Oxford University Press, Oxford ; New York.

Heisenberg, M. (2003) Mushroom body memoir: from maps to models. Nat Rev Neurosci, 4, 266-275.

Hellgren, O. & Sheldon, B. C. (2011) Locus-specific protocol for nine different innate immune genes (antimicrobial peptides: beta-defensins) across passerine bird species reveals within-species coding variation and a case of trans-species polymorphisms. Mol Ecol Resour, 11, 686-692.

Hermisson, J. & Pennings, P. S. (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics, 169, 2335-2352.

Hill, A. V., Jepson, A., Plebanski, M. & Gilbert, S. C. (1997) Genetic analysis of host-parasite coevolution in human malaria. Philos Trans R Soc Lond B Biol Sci, 352, 1317-1325.

Hill, W. G. & Robertson, A. (1968) Linkage disequilibrium in finite populations. Theor Appl Genet, 38, 226-231.

Hiwatashi, T., Okabe, Y., Tsutsui, T., Hiramatsu, C., Melin, A. D., Oota, H., et al. (2010) An explicit signature of balancing selection for color-vision variation in new world monkeys. Mol Biol Evol, 27, $453^{-4}64$.

Hoffmann, J. A. (2003) The immune response of Drosophila. Nature, 426, 33-38.

Hoffmann, J. A. & Reichhart, J. M. (2002) Drosophila innate immunity: an evolutionary perspective. Nat Immunol, 3, 121-126.

Hollox, E. J. & Armour, J. A. (2008) Directional and balancing selection in human beta-defensins. BMC Evol Biol, 8, 113.

Holub, E. B. (2001) The arms race is ancient history in Arabidopsis, the wildflower. Nat Rev Genet, 2, 516-527.

Horger, A. C., Ilyas, M., Stephan, W., Tellier, A., van der Hoorn, R. A. & Rose, L. E. (2012) Balancing selection at the tomato RCR3 Guardee gene family maintains variation in strength of pathogen defense. PLoS Genet, 8, e1002813.

Hu, T. T., Eisen, M. B., Thornton, K. R. & Andolfatto, P. (2013) A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. Genome Res, 23, 89-98.

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

Hudson, R. R., Kreitman, M. & Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. Genetics, 116, 153-159.

Hutter, S., Li, H., Beisswanger, S., De Lorenzo, D. & Stephan, W. (2007) Distinctly different sex ratios in African and European populations of Drosophila melanogaster inferred from chromosomewide single nucleotide polymorphism data. Genetics, 177, 469^{-4}80.

Hutter, S., Vilella, A. J. & Rozas, J. (2006) Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics, 7, 409.

Irving, P., Troxler, L., Heuer, T. S., Belvin, M., Kopczynski, C., Reichhart, J. M., et al. (2001) A genome-wide analysis of immune responses in Drosophila. Proc Natl Acad Sci U S A, 98, 15119-15124.

Irving, P., Troxler, L., Heuer, T. S., Belvin, M., Kopczynski, C., Reichhart, J. M., et al. (2001) A genome-wide analysis of immune responses in Drosophila. Proc Natl Acad Sci U S A, 98, 15119-15124.

Jiggins, F. M. & Hurst, G. D. (2003) The evolution of parasite recognition genes in the innate immune system: purifying selection on Drosophila melanogaster peptidoglycan recognition proteins. J Mol Evol, 57, 598-605.

Jiggins, F. M. & Kim, K. W. (2006) Contrasting evolutionary patterns in Drosophila immune receptors. J Mol Evol, 63, 769-780.

Jiggins, F. M. & Kim, K. W. (2007) A screen for immunity genes evolving under positive selection in Drosophila. J Evol Biol, 20, 965-970.

Joiner, W. J., Crocker, A., White, B. H. & Sehgal, A. (2006) Sleep in Drosophila is regulated by adult mushroom bodies. Nature, 441, 757-760.
Jones, J. D. & Dangl, J. L. (2006) The plant immune system. Nature, 444, 323-329.

Juneja, P. & Lazzaro, B. P. (2010) Haplotype structure and expression divergence at the Drosophila cellular immune gene eater. Mol Biol Evol, 27, 2284-2299.

Kanehisa, M. & Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 28, 27-30.

Kelley, J., Walter, L. & Trowsdale, J. (2005) Comparative genomics of natural killer cell receptor gene clusters. PLoS Genet, 1, 129-139.

Key, F. M., Teixeira, J. C., de Filippo, C. & Andres, A. M. (2014) Advantageous diversity maintained by balancing selection in humans. Curr Opin Genet Dev, 29, 45-51.

Kim, Y. & Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics, 160, 765-777.

Kimbrell, D. A. & Beutler, B. (2001) The evolution and genetics of innate immunity. Nat Rev Genet, 2, 256-267.

Kimura, M. (1968) Evolutionary rate at the molecular level. Nature, 217, 624-626.

Klein, J. (1987) Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. Hum Immunol, 19, 155-162.

Klein, J., Sato, A. & Nikolaidis, N. (2007) MHC, TSP, and the origin of species: from immunogenetics to evolutionary genetics. Annu Rev Genet, 41, 281-304.

Kleino, A., Valanne, S., Ulvila, J., Kallio, J., Myllymaki, H., Enwald, H., et al. (2005) Inhibitor of apoptosis 2 and TAK1-binding protein are components of the Drosophila Imd pathway. EMBO J, 24, 3423-3434.

Kreitman, M. (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature, 304, 412^{-4}17.

Langley, C. H., Crepeau, M., Cardeno, C., Corbett-Detig, R. & Stevens, K. (2011) Circumventing heterozygosity: sequencing the amplified genome of a single haploid Drosophila melanogaster embryo. Genetics, 188, 239-246.

Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C., Schrider, D. R., Pool, J. E., et al. (2012) Genomic variation in natural populations of Drosophila melanogaster. Genetics, 192, 533-598.

Laurent, S. J., Werzner, A., Excoffier, L. & Stephan, W. (2011) Approximate Bayesian analysis of Drosophila melanogaster polymorphism data reveals a recent colonization of Southeast Asia. Mol Biol Evol, 28, 2041-2051.

Lazzaro, B. P. (2008) Natural selection on the Drosophila antimicrobial immune system. Curr Opin Microbiol, 11, 284-289.

Lazzaro, B. P. & Clark, A. G. (2003) Molecular population genetics of inducible antibacterial peptide genes in Drosophila melanogaster. Mol Biol Evol, 20, 914-923.

Lazzaro, B. P., Sceurman, B. K. & Clark, A. G. (2004) Genetic basis of natural variation in D. melanogaster antibacterial immunity. Science, 303, 1873-1876.

Lee, J. E. & Edery, I. (2008) Circadian regulation in the ability of Drosophila to combat pathogenic infections. Curr Biol, 18, 195-199.

Leffler, E. M., Gao, Z., Pfeifer, S., Segurel, L., Auton, A., Venn, O., et al. (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science, 339, 1578-1582.

Leibovitch, B. A., Lu, Q., Benjamin, L. R., Liu, Y., Gilmour, D. S. & Elgin, S. C. (2002) GAGA factor and the TFIID complex collaborate in generating an open chromatin structure at the Drosophila melanogaster hsp26 promoter. Mol Cell Biol, 22, 6148-6157.

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

Leulier, F. & Lemaitre, B. (2008) Toll-like receptors--taking an evolutionary approach. Nat Rev Genet, 9, 165-178.

Levine, M. T. & Begun, D. J. (2008) Evidence of spatially varying selection acting on four chromatin-remodeling loci in Drosophila melanogaster. Genetics, 179, 475^{-4}85.

Lewontin, R. C. & Hubby, J. L. (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics, 54, 595-609.

Lewontin, R.C. (1974) The genetic basis of evolutionary change. Columbia University Press, New-York.

Li, H. & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078-2079.

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

Librado, P. & Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics, 25, 1451-1452.

Little, T. J., Watt, K. & Ebert, D. (2006) Parasite-host specificity: experimental studies on the basis of parasite adaptation. Evolution, 60, 31-38.

Lu, H. L., Wang, J. B., Brown, M. A., Euerle, C. & St Leger, R. J. (2015) Identification of Drosophila Mutants Affecting Defense to an Entomopathogenic Fungus. Sci Rep, 5, 12350.

Magwire, M. M., Bayer, F., Webster, C. L., Cao, C. & Jiggins, F. M. (2011) Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication. PLoS Genet, 7, e1002337.

Magwire, M. M., Fabian, D. K., Schweyen, H., Cao, C., Longdon, B., Bayer, F., et al. (2012) Genome-wide association studies reveal a simple genetic basis of resistance to naturally coevolving viruses in Drosophila melanogaster. PLoS Genet, 8, e1003057.

Mallon, E. B., Brockmann, A. & Schmid-Hempel, P. (2003) Immune response inhibits associative learning in insects. Proc Biol Sci, 270, 2471-2473.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., et al. (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res, 43, D222-226.

Mason, J. M. & Arndt, K. M. (2004) Coiled coil domains: stability, specificity, and biological implications. Chembiochem, 5, 170-176.

McDonald, J. H. & Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature, 351, 652-654.

McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics, 22, 356-358.

Medzhitov, R. & Janeway, C. A., Jr. (1997) Innate immunity: impact on the adaptive immune response. Curr Opin Immunol, 9, 4-9.

Meyers, B. C., Kaushik, S. & Nandety, R. S. (2005) Evolving disease resistance genes. Curr Opin Plant Biol, 8, 129-134.

Michelmore, R. W. & Meyers, B. C. (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res, 8, 1113-1130.

Miller, D. T., Read, R., Rusconi, J. & Cagan, R. L. (2000) The Drosophila primo locus encodes two low-molecular-weight tyrosine phosphatases. Gene, 243, 1-9.

Miotto, B., Sagnier, T., Berenger, H., Bohmann, D., Pradel, J. & Graba, Y. (2006) Chameau HAT and DRpd3 HDAC function as antagonistic cofactors of JNK/AP-1-dependent transcription during Drosophila metamorphosis. Genes Dev, 20, 101-112.

Miotto, B. & Struhl, K. (2006) Differential gene regulation by selective association of transcriptional coactivators and bZIP DNA-binding domains. Mol Cell Biol, 26, 5969-5982.

Mosca, T. J. & Luo, L. (2014) Synaptic organization of the Drosophila antennal lobe and its regulation by the Teneurins. Elife, 3, e03726.

Muller H.J. (1950) Our load of mutations. Am J Hum Genet, 2: 111–176.

Narasimamurthy, R., Hatori, M., Nayak, S. K., Liu, F., Panda, S. & Verma, I. M. (2012) Circadian clock protein cryptochrome regulates the expression of proinflammatory cytokines. Proc Natl Acad Sci U S A, 109, 12662-12667.

Newman, R. M., Hall, L., Connole, M., Chen, G. L., Sato, S., Yuste, E., et al. (2006) Balancing selection and the evolution of functional polymorphism in Old World monkey TRIM5alpha. Proc Natl Acad Sci U S A, 103, 19134-19139.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. G.

(2007) Recent and ongoing selection in the human genome. Nat Rev Genet, 8, 857-868.

Nygaard, S., Braunstein, A., Malsen, G., Van Dongen, S., Gardner, P. P., Krogh, A., et al. (2010) Long- and short-term selective forces on malaria parasite genomes. PLoS Genet, 6, e1001099.

Obbard, D. J., Jiggins, F. M., Halligan, D. L. & Little, T. J. (2006) Natural selection drives extremely rapid evolution in antiviral RNAi genes. Curr Biol, 16, 580-585.

Obbard, D. J., Welch, J. J., Kim, K. W. & Jiggins, F. M. (2009) Quantifying adaptive evolution in the Drosophila immune system. PLoS Genet, 5, e1000698.

Obbard D.J., K.H. Gordon, A.H. Buck, F.M. Jiggins (2009) The evolution of RNAi as a defence against viruses and transposable elements. Philos Trans R Soc Lond B Biol Sci. 364(1513):99-115.

Ochola, L. I., Tetteh, K. K., Stewart, L. B., Riitho, V., Marsh, K. & Conway, D. J. (2010) Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in Plasmodium falciparum. Mol Biol Evol, 27, 2344-2351.

Paparazzo, F., Tellier, A., Stephan, W. & Hutter, S. (2015) Survival Rate and Transcriptional Response upon Infection with the Generalist Parasite Beauveria bassiana in a World-Wide Sample of Drosophila melanogaster. PLoS One, 10, e0132129.

Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M. & Andolfatto, P. (2010) On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. Mol Biol Evol, 27, 1226-1234.

Peng, T. X., Moya, A. & Ayala, F. J. (1991) Two modes of balancing selection in Drosophila melanogaster: overcompensation and overdominance. Genetics, 128, 381-391.

Piertney, S. B. & Oliver, M. K. (2006) The evolutionary ecology of the major histocompatibility complex. Heredity (Edinb), 96, 7-21.
Pool, J. E., Corbett-Detig, R. B., Sugino, R. P., Stevens, K. A., Cardeno, C. M., Crepeau, M. W., et al. (2012) Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genet, 8, e1003080.

Presgraves, D. C. & Stephan, W. (2007) Pervasive adaptive evolution among interactors of the Drosophila hybrid inviability gene, Nup96. Mol Biol Evol, 24, 306-314.

Quach, H., Wilson, D., Laval, G., Patin, E., Manry, J., Guibert, J., et al. (2013) Different selective pressures shape the evolution of Toll-like receptors in human and African great ape populations. Hum Mol Genet, 22, 4829-4840.

Quintana-Murci, L. & Clark, A. G. (2013) Population genetic tools for dissecting innate immunity in humans. Nat Rev Immunol, 13, 280-293.

Reddy, B. A., Etkin, L. D. & Freemont, P. S. (1992) A novel zinc finger coiled-coil domain in a family of nuclear proteins. Trends Biochem Sci, 17, 344-345.

Roux, C., Pauwels, M., Ruggiero, M. V., Charlesworth, D., Castric, V. & Vekemans, X. (2013) Recent and ancient signature of balancing selection around the S-locus in Arabidopsis halleri and A. lyrata. Mol Biol Evol, 30, 435-447.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., et al. (2006) Positive natural selection in the human lineage. Science, 312, 1614-1620.

Sackton, T. B., Lazzaro, B. P., Schlenke, T. A., Evans, J. D., Hultmark, D. & Clark, A. G. (2007) Dynamic evolution of the innate immune system in Drosophila. Nat Genet, 39, 1461-1468.

Saleh, M. C., Tassetto, M., van Rij, R. P., Goic, B., Gausson, V., Berry, B., et al. (2009) Antiviral immunity in Drosophila requires systemic RNA interference spread. Nature, 458, 346-350.

Sato, M. P., Makino, T. & Kawata, M. (2016) Natural selection in a population of Drosophila melanogaster explained by changes in gene expression caused by sequence variation in core promoter regions. BMC Evol Biol, 16, 35.

Schierup, M. H., Mikkelsen, A. M. & Hein, J. (2001) Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. Genetics, 159, 1833-1844.

Schlenke, T. A. & Begun, D. J. (2003) Natural selection drives Drosophila immune system evolution. Genetics, 164, 1471-1480.

Schlenke, T. A. & Begun, D. J. (2005) Linkage disequilibrium and recent selection at three immunity receptor loci in Drosophila simulans. Genetics, 169, 2013-2022.

Sedghifar, A., Saelao, P. & Begun, D. J. (2016) Genomic Patterns of Geographic Differentiation in Drosophila simulans. Genetics, 202, 1229-1240.

Segurel, L., Thompson, E. E., Flutre, T., Lovstad, J., Venkat, A., Margulis, S. W., et al. (2012) The ABO blood group is a trans-species polymorphism in primates. Proc Natl Acad Sci U S A, 109, 18493-18498.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 13, 2498-2504.

Shirasu-Hiza, M. M., Dionne, M. S., Pham, L. N., Ayres, J. S. & Schneider, D. S. (2007) Interactions between circadian rhythm and immunity in Drosophila melanogaster. Curr Biol, 17, R353-355.

Sironi, M., Cagliani, R., Forni, D. & Clerici, M. (2015) Evolutionary insights into host-pathogen interactions from mammalian sequence data. Nat Rev Genet, 16, 224-236.

Smith, M., Bhaskar, V., Fernandez, J. & Courey, A. J. (2004) Drosophila Ulp1, a nuclear pore-associated SUMO protease, prevents accumulation of cytoplasmic SUMO conjugates. J Biol Chem, 279, 43805^{-4}3814.

Spurgin, L. G. & Richardson, D. S. (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci, 277, 979-988.

Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J. (1999) Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature, 400, 667-671.

Stephan, W. (1995) An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. Mol Biol Evol, 12, 959-962.

Stephan, W. & Li, H. (2007) The recent demographic and adaptive history of Drosophila melanogaster. Heredity (Edinb), 98, 65-68.

Stroschein-Stevenson, S. L., Foley, E., O'Farrell, P. H. & Johnson, A. D. (2006) Identification of Drosophila gene products required for phagocytosis of Candida albicans. PLoS Biol, 4, e4.

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics, 105, 437^{-4}60.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123, 585-595.

Teixeira, J. C., de Filippo, C., Weihmann, A., Meneu, J. R., Racimo, F., Dannemann, M., et al. (2015) Long-Term Balancing Selection in LAD1 Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and Bonobos. Mol Biol Evol, 32, 1186-1196.

Tellier, A., Moreno-Gamez, S. & Stephan, W. (2014) Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics. Evolution, 68, 2211-2224.

Tesicky, M. & Vinkler, M. (2015) Trans-Species Polymorphism in Immune Genes: General Pattern or MHC-Restricted Phenomenon? J Immunol Res, 2015, 838035.

Thomas, J. C., Godfrey, P. A., Feldgarden, M. & Robinson, D. A. (2012) Candidate targets of balancing selection in the genome of Staphylococcus aureus. Mol Biol Evol, 29, 1175-1186.

Thompson, J.N., Burdon J.J (1992) Gene-for-gene coevolution between plants and parasites. Nature 360: 121-125.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (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, 4673^{-4}680.

Tinsley, M. C., Blanford, S. & Jiggins, F. M. (2006) Genetic variation in Drosophila melanogaster pathogen susceptibility. Parasitology, 132, 767-773.

Tsoumtsa, L. L., Torre, C. & Ghigo, E. (2016) Circadian Control of Antibacterial Immunity: Findings from Animal Models. Front Cell Infect Microbiol, 6, 54.

Unckless, R. L., Howick, V. M. & Lazzaro, B. P. (2016) Convergent Balancing Selection on an Antimicrobial Peptide in Drosophila. Curr Biol, 26, 257-262.

van Delden, W., Boerema, A. C. & Kamping, A. (1978) The alcohol dehydrogenase polymorphism in populations of Drosophila melanogaster. I. Selection in different environments. Genetics, 90, 161-191.

Van der Hoorn, R. A., De Wit, P. J. & Joosten, M. H. (2002) Balancing selection favors guarding resistance proteins. Trends Plant Sci, 7, 67-71.

Vitti, J. J., Grossman, S. R. & Sabeti, P. C. (2013) Detecting natural selection in genomic data. Annu Rev Genet, 47, 97-120.

Voigt, S., Laurent, S., Litovchenko, M. & Stephan, W. (2015) Positive Selection at the Polyhomeotic Locus Led to Decreased Thermosensitivity of Gene Expression in Temperate Drosophila melanogaster. Genetics, 200, 591-599.

Wang, M. C., Bohmann, D. & Jasper, H. (2003) JNK signaling confers tolerance to oxidative stress and extends lifespan in Drosophila. Dev Cell, 5, 811-816.

Wang, X. H., Aliyari, R., Li, W. X., Li, H. W., Kim, K., Carthew, R., et al. (2006) RNA interference directs innate immunity against viruses in adult Drosophila. Science, 312, 452^{-4}54.

Wang, Y. J. & Brock, H. W. (2003) Polyhomeotic stably associates with molecular chaperones Hsc4 and Droj2 in Drosophila Kc1 cells. Dev Biol, 262, 350-360.

Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol, 7, 256-276.

Weir, B. S. (1996) Genetic data analysis II : methods for discrete population genetic data. Sinauer Associates, Sunderland, Mass.

Wilfert, L. & Jiggins, F. M. (2010) Host-parasite coevolution: genetic variation in a virus population and the interaction with a host gene. J Evol Biol, 23, 1447-1455.

Woolhouse, M. E., Webster, J. P., Domingo, E., Charlesworth, B. & Levin, B. R. (2002) Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet, 32, 569-577.

Zivkovic, D., Steinrucken, M., Song, Y. S. & Stephan, W. (2015) Transition Densities and Sample Frequency Spectra of Diffusion Processes with Selection and Variable Population Size. Genetics, 200, 601-617.

Zivkovic, D. & Stephan, W. (2011) Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. Theor Popul Biol, 79, 184-191.

ACKNOWLEDGEMENTS

I would like to thank Prof. Wolfgang Stephan for giving me the opportunity to join his group and to work on this project. I am very grateful for his advice and support.
A great thank you to Dr. Stephan Hutter for all the discussions, his help and advice throughout the years.
I thank Daniel Živković and Andreas Wollstein for their help and advice on the project and on my manuscripts.

I thank all the people of the evolutionary biology group for the friendly and good atmosphere but also for enriching discussion. It was great to be part of this group.
I would like particularly thank Vedran Božičević who helps me for the GO analysis, but also Soumya Ranganathan for helping me with R codes and for all our discussions. I thank Pablo Duchén for his help and for providing me codes and helping me with the analysis of the *D. Simulans* sequences and Alisa Sedghifar for providing me the sequences of *D. Simulans*.
I also would like to thank Ingrid Kroiss for her help with bureaucratic issues.
Special thank for my friend Isidora for the coffee break.

Finally, I would like to thank my family for their support during all these years.