Evolution of genes related to temperature adaptation in *Drosophila melanogaster* as revealed by QTL and population genetics analyses

Dissertation

an der Fakultät für Biologie

der Ludwig-Maximilians-Universität München

vorgelegt von

Ricardo Wilches

aus Bogotá D.C.

München, den 28.04.2014

Dekan: Prof. Dr. Heinrich Leonhardt
1. Gutachter: Prof. Dr. Wolfgang Stephan
2. Gutachter: Prof. Dr. Niels Dingemanse
Dissertation eingereicht am: 28.04.2014
Datum der Disputation: 21.07.2014

ERKLÄRUNG

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

EIDESSTATTLICHE ERKLÄRUNG

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

Ricardo Wilches

München, 28.04.2014

a Caro, Lili,
mamá y papá

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

STATEMENT OF CONTRIBUTIONS

In this thesis, I present my doctoral research, all of which has been done by myself except for the following: Gene expression assays with qPCR on candidate genes presented in section 2.2 were done together with S. Voigt. P. Duchén wrote scripts to handle next-generation sequence datasets and calculate summary statistics, which were used in all projects included in this dissertation. He also conducted an ancestral state reconstruction analysis with Drosophilid sequences of the gene *CG1677*. S. Laurent conducted CLR tests (sections 2.3 and 2.4) and P. Pavlidis conducted the omega statistic tests in section 2.1. A. Steincke, H. Lainer and S. Lange carried out the sequencing of the region between genes *unc-119* and *brk* (section 2.3).

Publications derived from the present work:

Section 2.1 is a contribution to the published work: Svetec N, Werzner A, Wilches R, *et al.* (2011) Identification of X-linked quantitative trait loci affecting cold tolerance in *Drosophila melanogaster* and fine mapping by selective sweep analysis. Molecular Ecology 20, 530-544.

Sections 2.2 and 2.3 together belong to the published work: Wilches R, Voigt S, Duchen P, *et al.* (2014) Fine-mapping and selective sweep analysis of QTL for cold tolerance in *Drosophila melanogaster*. G3-Genes Genomes Genetics. doi:10.1534/g3.114.012757.

SUMMARY

The fixation of beneficial variants leaves genomic footprints characterized by a reduction of genetic variation at linked neutral sites and strong, localized allele frequency differentiation among subpopulations. In contrast, for phenotypic evolution the effect of adaptation on the genes controlling the trait is little understood. Theoretical work on polygenic selection suggests that fixations of beneficial alleles (causing selective sweeps) are less likely than small-to-moderate allele frequency shifts among subpopulations. This thesis encompasses three projects in which we have experimentally addressed the issue of selective sweeps *vs.* allele frequency shifts in the context of polygenic adaptation. We studied three X-linked QTL underlying variation in chill coma recovery time (CCRT), a proxy for cold tolerance, in *Drosophila melanogaster* from temperate (European) and tropical (African) environments. The analysis of these QTL was performed by means of selective sweep mapping and quantitative complementation tests coupled with expression assays.

While the results of the selective sweep mapping approach identified a gene (*CG4491)* that is unlikely to be affecting CCRT, quantitative and gene expression analyses revealed two linked candidate genes (*brk* and *CG1677*) that appear to differ in their evolutionary histories. We found that the difference in expression of the gene *brk* between populations affects CCRT variation. Cold tolerant flies from the temperate zone have a lower expression of this gene than cold sensitive flies from the tropics. We found that a likely cause of this difference is variation in a *cis*-regulatory element in the *brk* 5' enhancer region. Sequence variants in this element exhibit moderate frequency differences between populations from temperate and tropical environments, forming two latitudinal clines: one from the equator to the north and another one in opposite direction to the south. In contrast, the other gene within the same QTL (*CG1677*), which is linked to *brk*, showed no measurable effect on cold tolerance but is a likely target of strong positive selection leading to a selective sweep in the European population.

These results are consistent with the aforementioned theoretical predictions about footprints of selection in polygenic adaptation. They are also proof of the conceptual bias incurred when identifying candidate genes within a QTL via selective sweep mapping, at least in naturally evolving populations. The challenge for the evolutionary genetics community in the coming years is to develop statistical tools that are as powerful and

robust as those already available to map selective sweeps to identify sites in the genome where allele frequency shifts have occurred due to adaptive evolution at the phenotypic level.

Finally, the last section of the results is a report of a new population genetics dataset. It consists of a collection of 80 inbred lines from a natural *D. melanogaster* population in Sweden and 19 full genome sequences derived from this sample. We hope this material will provide us with further insight into the processes underlying adaptation to novel and stressful environments.

I – INTRODUCTION

1.1 VIEWS OF ADAPTATION

Adaptation is defined as the movement of a population towards a phenotype that best fits the present environment (Orr 2005). According to this definition, adaptation is a process distinguishable from other types of evolutionary change by the benefit it confers. Adaptation is typically documented by following the change of a trait over time or by inferring how it changed. To be considered adaptive, this change has to be accompanied by an increase in biological fitness. In practice, however, biological fitness cannot be directly scored; empiricists rely on measuring its components, such as viability, growth rate, fecundity, or reproductive success. Although seemingly straightforward, this approach may not be applicable to all species or ecological contexts. Thus, the assessment of fitness depends heavily on the type of organism being studied and whether the researcher can measure relevant proxies for fitness. In order to understand adaptation, it is important to understand how it happens. Existing models and empirical evidence gathered in the last 100 years have provided great insight into the matter, but are not yet flexible enough to account to for the complexities of biological systems. Thus, the study of adaptation is still a challenge for evolutionary biologists.

This dissertation responds to this challenge and may be seen as a brick freshly laid atop the growing edifice of knowledge constructed with the purpose of understanding evolution. A look at the foundations whereupon it lies is useful to understand its location in the developing structure[1]. This introductory chapter provides an overview of the concepts and approaches in evolutionary biology that were central to the development of this thesis. Without diving too deeply into the theories and evolutionary models, I provide a chronological view of the main developments of the study of adaptation. This is followed by an overview of the ideas and tools that population and quantitative genetics have created to study adaptation. This chapter ends with the specific research question that motivated this doctoral dissertation.

The notion of biological adaptation has long been present in the perception of the natural word. In the middle of the 19th century, adaptation was regarded as the inevitable consequence of evolution. The view of evolution championed by J.B. Lamarck posed that adaptation was the perfected state of form with respect to a given environment. It relied, however, on the idea of supernatural intervention to ignite the process leading to adaptation (Koonin 2009). The Darwinian conception of evolution went beyond this mere statement and introduced natural selection as the mechanism whereby adaptation was possible. Darwinism explained how change, via natural selection, was able to mold biological structures (phenotypes). It posed that phenotypes would gradually change building on preexisting slight successive variants. This view came to be recognized as the micro-mutational model, and constituted one of the cornerstones of quantitative genetics (Orr 2005). Neither Lamarckism nor Darwinism had the right conception of the mechanisms of inheritance, which made them fail when explaining how beneficial changes could be passed on to subsequent generations. In spite of this limitation, by the end of the 19th century the Darwinian view of adaptation was the only solid framework to study the diversity of forms, function and behavior existing in nature. The idea that a process such as natural selection could shape the interactions between organisms and their environments constituted the first truly rational hypothesis in biology (see for example Nilsson (1998) for an example of plant-pollinator coevolution).

[1] This metaphor refers to that in the correspondence C. Darwin and the Scottish surgeon and paleontologist H. Falconer in which Darwin's ideas of evolution are equated to the foundations of a building, namely Milan's cathedral (See Gould 2002, p. 3). Gould himself is the author of another architectural metaphor that illustrates his criticism of the adaptationist program (Gould and Lewontin 1979).

The 20th century started with the missing piece of puzzle, when G. Mendel's laws of inheritance were made available to the broad scientific community (Von Tschermak-Seysenegg 1951). This propelled developments in mathematical biology, which provided the first formal approach to evolution. Mendel's work also shed new light to understand the function of already-known nucleic acids and chromatin as the substance of inheritance (Moore 1983). With a solid view of inheritance a tractable way to understand adaptation emerged; the essence of evolution was to be distilled at the genetic level (Lewontin 1997).

The source of the variation that Darwinism considered as the starting point for evolution was attributed to mutation and what followed for evolutionary models was to explain the fate of new mutations within populations. Three outcomes were envisioned for these mutations, (i) they appear and get lost, (ii) they remain in the population at intermediate frequency, or (iii) they reach fixation. These fates are dependent on parameters such as population size, reproductive strategies and, most importantly, on whether a new mutation confers any benefit to its bearer. Theoreticians came to realize that chance alone could be sufficient to govern the destiny of new mutations; for instance, in small populations mutations can become fixed by chance (genetic drift).

Around 1940, and the subsequent three decades, the evolutionary biology community worked actively under a hardened neo-Darwinian (adaptationist) paradigm. In their view of adaptation all new beneficial mutations were to be seen by selection and then taken to fixation. Other mechanisms such as purifying selection (whereby deleterious mutations are eliminated so that unfit phenotypes are purged from the population) were described but considered merely as accessory to adaptation. Moreover, any role of genetic drift was neglected. There was a sound reticence to admit that stochasticity could have any important role in evolution (see Gould 2002 on Dobzhansky's 1951 edition of Genetics and the origin of species).

The adaptationist view of evolution, as depicted by Darwinism, guided biological research for almost 70 years; until strong criticism of this research program was eventually heard (Gould and Lewontin 1979). If adaptation is to be invoked as the phenomenon leading the evolution of form, function and behavior, sufficient evidence has to be provided. Gould and Lewontin's (1979) reaction against the adaptationist view was also a call for scientists to consider alternative possibilities other than adaptation, *e.g.* selection could have indirectly shaped the putative adaptive trait because of correlation with the actual target of selection or the trait having evolved by chance. Adaptive hypotheses

should be treated as susceptible to being falsified. This marked the beginning of an era of statistical developments in the study of evolution (Barrett and Hoekstra 2011). At the phenotype level, the task of finding ways to disentangle the effect of selection and drift was undertaken by Lande (1983). However, the best approach to the study of adaptive evolution remained the careful observation and documentation of change in natural populations over several generations, an endeavor successful in a limited number of cases (see Grant and Grant 2002).

Gould and Lewontin's 1979 paper accompanied a challenging period for the study of adaptation, especially in the field of molecular evolution. Already a decade earlier, Kimura's work on the role of genetic drift in evolution generated a new research program, which focused on neutral evolutionary processes and almost obliterated the interest for the study of selection (Jukes 2000; Koonin 2009). The neutral theory maintained that the majority of mutations that are fixed in the course of evolution are selectively neutral (or nearly neutral), so their fixation occurs via random drift. However, if mutations do have a negative effect on fitness, for instance those occurring in functionally constrained parts of a protein, purifying selection will operate to maintain the functional *status quo*. The statement of this theory that challenged the reign of adaptationism was that beneficial mutations were so rare that the contribution of positive selection to molecular evolution could be neglected.

The central premise of the neutral theory is in itself a prediction; if neutral evolution was that pervasive, then one should expect more differences between species in functionally less important sequences. This prediction has been confirmed by protein coding gene comparisons among mammalian species. Non-coding parts of the genes and synonymous sites evolve faster than non-synonymous ones (Kreitman 1996; Makalowski and Boguski 1998). However, not all neutral expectations have been met; for instance, the existence of codon bias in *Escherichia coli*, *Sacharomyces cereviciae* and *Drosophila* (Akashi 1995) was a finding against pan-neutrality. The fact that a codon type is preferred over the other possible ones suggests that some level of selection operates at supposedly neutral synonymous sites (Kreitman 1996). Another aspect of protein evolution that contradicts neutral evolution is the existence of unequal ratios of non-synonymous to synonymous divergence with respect to non-synonymous to synonymous polymorphism. This observation was made by MacDonald and Kreitman (1991) when studying the alcohol dehydrogenase (*Adh*) gene in three *Drosophila* species, leading to the conclusion that

positive selection has driven the evolution of this protein.

The work of MacDonald and Kreitman (1991) provided one of the first molecular tests of the neutral model. Its application to larger gene sequence datasets has revealed that adaptive evolution in the genome is more common than argued by the neutral theory; for instance 35% of the assessed amino acid substitutions between humans and old world monkeys have been driven by positive selection, while up to 45% of amino acid substitutions were driven by positive selection in the divergence between Drosophilid species (Bierne and Eyre-Walker 2004). These estimates account for spurious false positives due to the effect of weak selection and codon bias. The statistical tractability of the predictions of the neutral theory made it the ideal null hypothesis to test adaptive evolution at the molecular level. A rich statistical toolkit developed in the last 30 years provides evolutionary geneticists with the means to accept or reject adaptive hypotheses (Hudson *et al.* 1987; Tajima 1989; Fay and Wu 2000; Kim and Stephan 2002).

In the history of evolutionary genetics, the study of adaptation from DNA sequence data has primarily focused on coding DNA. However the fast evolving fields of functional genomics and transcriptomics have attested to the relevance of non-coding DNA in evolution. One of the functions of non-coding DNA is to tune the level of expression of nearby genes. Sequences with this function are known as *cis*-regulatory elements Because non-coding DNA is also a source of phenotypic variability, its role in evolution has been well documented (see for example King and Wilson 1975; Carroll 2005; Gompel *et al.* 2005; Rubinstein and De Souza 2013). In fact, *cis*-elements have been identified as targets of positive selection in humans (Enattah *et al.* 2007) and *Drosophila* (Saminadin-Peter *et al.* 2012; Glaser-Schmitt *et al.* 2013).

Currently the fields of functional genomics, transcriptomics and systems biology provide new tools to reveal the mechanisms that bridge genetic and phenotypic variation in natural populations. A major goal of evolutionary biology is to extend the existing theoretical framework to study adaptation allowing for more realistic models in which new data can be sensibly explained. In the following two of the main approaches to study adaptation will be described. Each approach has been developed within the two central disciplines of evolutionary biology: population and quantitative genetics.

1.2 POPULATION GENETICS AND ADAPTATION

1.2.1 Adaptation in sequence space

The theoretical approach to study adaptation dates back to S. Wright, who in 1932 depicted his ideas about adaptation in the form of a fitness (or adaptive) landscape (Dietrich and Skipper 2012). Wright's adaptive landscape represents the genetic constitution of a population. It could be represented by points in an *n*-dimensional space, with each point representing a unique genetic combination of alleles at *n* loci. The biological fitness associated with each particular genetic combination was then represented by a value on a further dimension, so that these fitness values form a surface with valleys and peaks, where the latter represent areas of high fitness and the former depict the least fit (disadvantageous) genetic compositions. During adaptation, selection will push average population fitness values to those situated on the peaks of the landscape. Once populations are sitting on the peaks of higher fitness, the process is completed. However, when new environmental conditions arise and new fitness maxima are created, a new bout of adaptation will start, and the population will climb up a new hill (Orr 2005).

Population genetics aims at understanding the processes that shape genetic variability. Therefore population geneticists study evolution in sequence space. Although the reign of the neutral theory represented something similar to the dark ages for the study of adaptation, Wright's adaptive landscapes were reframed to study the dynamics of adaptation using protein sequence data, already available in the 1960s (Maynard Smith 1970). Almost a decade later the same model was reinterpreted using DNA sequences (Gillespie 1984). Maynard Smith and Gillespie's reinterpretation of adaptive landscapes led to the development of so-called mutational landscapes (Orr 2005; Dietrich and Skipper 2012).

Mutational landscapes arise in a space of nucleotide sequences, say genes, in which sequences are arranged in such a way that sequences that differ from one another by a single mutation are adjacent while highly divergent sequences are found far from one another. The landscape arises when each sequence is given a fitness value, which is plotted onto a new dimension. The process of adaptation explained by this mutational landscape follows roughly the same idea as fitness landscapes. Imagine a population that

is fixed for a given DNA sequence of length L (L being the number of nucleotides). Due to a change in the environment or the colonization of new habitats the wild type sequence, which has been so far the fittest, no longer has the highest fitness value. Adaptation will occur when new fitter mutants of this wild type appear.

Gillespie stated that a maximum of 3^L new sequences might appear if all sites mutate, because each base in the wild type sequence has only three other possibilities to mutate. Gillespie also pointed out that mutants that matter occur at single bases. Double or triple mutations were extremely rare and could be safely disregarded (Orr 2008; Gillespie 1984). Another important aspect of this model is that only a small fraction of the possible new alleles will be beneficial, while the vast majority will have neutral or deleterious effects on fitness. Their destiny will be dictated by genetic drift and purifying selection, if population sizes allow so.

The empirical bases to study adaptation in population genetics came from experimental bacterial evolution first performed in the 1950s, providing substantial information on how selection shapes genetic diversity in bacterial populations. The work by Atwood *et al.* (1951) on *E. coli* showed that among populations of auxotrophic bacteria rare beneficial mutants (that originate from prototrophic bacteria[2]) appear and rapidly become the dominant type in the culture; in other words, new advantageous mutants become fixed via selective sweeps. The process is completed upon fixation and resumes only when a new fitter mutant occurs in the current wild type background. This situation is described as the sequential fixation of new mutants. Through time, these fast fixation events alternate with periods in which new beneficial mutants occur. This periodic selection scenario was central for the development of Maynard Smith and Haigh's (1974) ideas on the fixation of a new beneficial mutation in a sexually reproducing organism, *i.e.* in the presence of recombination. This led to the selective sweep model (Kaplan *et al.* 1989; Stephan *et al.* 1992; Barton 2000), the simplest and best studied way to describe positive selection in sequence space.

Selective sweep theory states that neutral variants in proximity to the target of selection will increase in frequency or fix along with the beneficial allele, a phenomenon known as genetic hitchhiking. Consequently genetic variants are substantially reduced in

[2] Strains that possess the enzymatic machinery to synthesize a given metabolite are called prototroph. By contrast, an auxotroph is a strain that by mutation lost the mechanism to synthesize the metabolite.

the neighborhood of the selected site, unless recombination is frequent enough to mitigate the loss of linked neutral variants by creating haplotypes that encompass the selected allele. As expected, the larger the distance to the target of selection the higher the chance of recombination events. This creates a signal of selective sweeps characterized by valleys of genetic polymorphism centered on the target of selection (Kim and Stephan 2002).

1.2.2 Selective sweeps and the site frequency spectrum

The study of the properties of the signal of selective sweeps has facilitated the development of statistical approaches to identify instances of positive selection in sequence data from natural populations, reviewed in Pavlidis *et al.* (2008) and Stephan (2010). The feature of the selective sweep model that has been most intensely studied is its effect on neighboring genetic variation. As a consequence of hitchhiking the expected reduction of heterozygosity generates a new distribution of allele classes (the site frequency spectrum, SFS[3]), compared to cases where no selection has occurred (Kim and Stephan 2002; Jensen *et al.* 2005; Thornton *et al.* 2007).

A distorted SFS is characterized by showing an excess of rare (low frequency) variants (Tajima 1989) as well as an excess of fixed or nearly fixed derived variants (Fay and Wu 2000). The first statistical tests for sequence data made use of this expected differences between allele classes under selective and neutral scenarios (Tajima 1989). Tajima's *D* statistic, for example, compares two estimators of genetic diversity parameter (θ), one that reflects the average number of nucleotide differences among two sequences, $\theta\pi$ (Tajima 1983), and the other based on the number of polymorphic sites, θ_W (Watterson 1975).

As mentioned before, positive selection increases the proportion of rare variants, thus inflating the value of θ_W relative to $\theta\pi$. When Tajima's *D* is calculated, a negative value reflects the excess of rare variants rejecting the neutral hypothesis in favor of positive selection. Subsequently, Fu and Li (1993) as well as Fay and Wu (2000) based their tests on the effect of the excess of high-frequency derived variants, thus increasing the power to distinguish selection from neutral scenarios. However, even if such power is achieved, another problem arises when accounting for spurious signals of positive selection. Genetic drift and selection are not the only forces that shape genetic variation in natural

[3] The site frequency spectrum is one the summary statistics it captures all mutation classes that segregate in population. This spectrum is represented by a histogram of frequencies.

populations, but demographic events (population history events) also do. For instance, recent population size changes, such as bottlenecks, also reduce genetic diversity and can reshape the SFS in ways very similar to that achieved by selection (Thornton and Jensen 2007). However, demographic events are expected to affect the entire genome, while selection has a localized effect.

The following generation of tests for selection focused on the whole breadth of the SFS. For example Kim and Stephan (2002) developed a test based on the ratio between the likelihood of a null (neutral evolution model) and the alternative (selective sweep) hypotheses. In their approach, false positives due to demographic events can be detected when this confounding factor is accounted for in the null hypothesis. However, this is only possible if the right demographic history of the population under study is known (Thornton & Jensen 2007). Unfortunately, this information is available only for some populations of model species such as *Arabidopsis thaliana* (François *et al.* 2008), *Drosophila melanogaster* (Stephan and Li 2007; Laurent *et al.* 2011; Duchen *et al.* 2013), and humans (Excoffier *et al.* 2013).

Hence, it was necessary to develop tests that account for demography even if the correct history is unknown. The method of Nielsen and colleagues, implemented in SweepFinder (Nielsen *et al.* 2005), fulfilled this need. It is a likelihood ratio test, similar to the Kim-Stephan test. However, the method derives the null hypothesis using the background pattern of variation of the data itself. As larger genome-wide sequence datasets accumulate, greater computational power and efficient algorithms are required to conduct satisfactory analyses and compare for instance data sets from a decade ago (Glinka *et al.* 2003) with current ones (Langley *et al.* 2012; Mackay *et al.* 2012). Fully aware of this need, Pavlidis and colleagues (2013) presented their improved, more stable, and scalable implementation of SweepFinder. It is the state-of-the-art method to detect selective sweeps on genomic data.

1.2.3 Selective sweeps and linkage disequilibrium

Another consequence of genetic hitchhiking is the pattern of linkage disequilibrium (LD) that is left after the sweep is completed (Kim and Stephan 2002; Przeworski 2002). Such an LD pattern emerges from the action of recombination during the early phases of the sweep generating haplotype structure (Pfaffelhuber *et al.* 2008). Kim & Nielsen (2004) performed a study of the genealogy of the sites adjacent to the target of selection, reaching

the following conclusions: (i) a high level of LD is expected in regions close, but not immediately adjacent, to the site where the fixation of the beneficial allele occurred. (ii) Once the chromosomal fraction under study is divided by the location of the beneficial mutation, a high level of LD is expected within each side but not across the two sides. (iii) The probability of observing a high frequency of derived alleles in the sample is greater in regions where LD is strong. A composite likelihood method based on the statistic (ω) that captures these three features of LD was developed (Kim and Nielsen 2004). With this approach, if the two initial conditions are met, the calculated ω value should be maximized signaling the occurrence of a selective sweep. It has been shown that the ω statistic has good power to detect genuine signals of selection in populations that experienced population size bottlenecks (Jensen *et al.* 2007). This motivated Pavlidis and colleagues (2010) to explore the utility of applying the ω statistic together with the SFS-based method of SweepFinder, to gain accuracy in the detection of selective sweep targets in populations that have experienced recent bottlenecks. An approach based on the combination of these two aspects of genetic hitchhiking is regarded as a promising way to capture footprints of selection in whole-genome sequence datasets.

1.2.4 Selective sweeps and population differentiation

A third feature of positive selection in spatially structured populations is locus-specific allele frequency differentiation (Lewontin and Krakauer 1973; Beaumont 2005). The distribution of allele frequency differentiation values between two or more populations results from the interplay of random processes and selection across the entire genome. Sites that show above-average allele frequency differences, measured by estimates of the parameter F_{ST} (Weir and Cockerham 1984), are likely subjects of positive directional selection, while sites with substantial below-average F_{ST} values are evolving under balancing or strong purifying selection. Relying on this rationale, Lewontin and Krakauer (1973) developed the first F_{ST} based method to identifying loci evolving under selection. The approach, however, was strongly criticized (see Nei and Maruyama 1975) because of its unrealistic demographic assumptions that all subpopulations have the same splitting time from the ancestral source population and that the number of migrants exchanged among them was the same. In addition, the fact that the neutral F_{ST} distribution used to identify outliers depended on demographic assumptions was also a subject of strong

debate (see Beaumont 2005 and recently Bierne *et al.* 2013).

The reassessments of the method, however, yielded success from the mid-1990s until the present. These focused on overcoming the limitations imposed by demography (Beaumont and Nichols 1996; Beaumont and Balding 2004; Riebler *et al.* 2008). The work of Beaumont and Balding (2004) considers a string of $\mathcal{J}$ loci sampled within an array of I populations, each unique ij combination corresponds to a model-based $F_{ST}{}_{ij}$ coefficient. In turn, each $F_{ST}{}_{ij}$ is decomposed into locus and population effects represented by variables (α) and (β), respectively. The locus effect α_i is shared across populations and all sites within a given population share the population effect β_j. If selection is responsible for any given F_{ST}, it is represented by the locus-specific variable α, while demographic aspects are accounted for by β, which affects all loci within one population. Even though this reassessment provides an elegant and intuitive way to account for demography, relaxing many of the assumptions, there is still the concern that the approach does not provide a rigorous way to test the hypothesis that a locus is subject to selection (Foll and Gaggiotti 2008).

Riebler *et al.* (2008) approached the issue by introducing a locus-specific auxiliary variable (δ_i), which indicates that a locus is under selection if its posterior probability is larger than a threshold value obtained by simulation. Foll and Gaggiotti's (2008) way to address the problem is also Bayesian. However, they determine which locus is under selection by estimating the posterior probability of two models, one that invokes selection, while the other excludes it. The decision of which model best explains the data is made based on the obtained ratio of these two posterior probabilities, *i.e.* Bayes factors. In addition, the problem of multiple testing is addressed by calculating false discovery rates (Foll and Gaggiotti 2008).

1.2.5 Positive selection in the genome

The preceding sections dealt with approaches to identify positive selection on genetic data based on the features of the selective sweep model. A common aspect of the extensive research done in this area is the need for improvement of the tests in order to keep up with the increasing complexity of the data and the need to increase statistical power while minimizing error, particularly false positives. It is important to emphasize that these methods are based on the selective sweep model. There are other models of positive

selection that relax some of the major assumptions of the selective sweep model, for example the soft sweep model does not require that the beneficial alleles are new mutations or low-frequency migrants. Instead, standing variants with neutral or slightly deleterious effects on fitness can become advantageous upon sudden environmental changes and may go to fixation (Hermisson and Pennings 2005). Positive selection scenarios like this are thought to be as important as classical selective sweeps for adaptation (Pritchard and Di Rienzo 2010; Pritchard *et al.* 2010). However, since they do not leave strong characteristic signals in the genome when they occur (Przeworski *et al.* 2005), they are virtually impossible to detect with the methods previously described.

A compelling example of adaptation in humans is provided by lactase, the enzyme that hydrolyzes lactose, the main sugar in milk. The expression of the lactase gene evolved such that individuals beyond weaning age can use milk as an energy source. This trait has evolved in parallel in dairy farming populations in Europe, East Africa, and the Middle East (Bersaglieri *et al.* 2004; Enattah *et al.* 2008; Tishkoff *et al.* 2007). Another example of positive selection in human populations is found at the gene SLC24A5, this gene is one of those responsible for skin pigmentation. A derived, light-skin variant of this gene has been fixed in European populations (Lamason *et al.* 2005). Its fixation has been correlated with improved vitamin D synthesis in Caucasian populations (Jablonski and Chaplin 2012). In model organisms, such as *Drosophila*, a considerable effort has been put into characterizing genome-wide sequence variation with the aim of pinpointing patterns that can be unequivocally assigned to the action of positive selection. These patterns have been found within or around several genes, *e.g.* the *php-p* gene region and the *diminutive* and *timeless* genes (Jensen *et al.* 2007; Tauber *et al.* 2007; Beisswanger and Stephan 2008).

In all of these cases *a priori* knowledge of gene function paved the way to connect selection at the molecular level with an adaptive phenotype, for instance, diapause onset in the case of *timeless*, a trait that is tightly linked with latitudinal adaptation. In cases such as the *unc-119/brk* gene region (Glinka *et al.* 2006) or *HDAC6* (Svetec *et al.* 2009) the phenotypic link remains obscure. Such situations constitute the ground for the criticism to the population genetics approach to the study of adaptation. The identification of sites in the genome that have experienced positive selection can provide estimates of how important selection is in evolution, even when the actual change in fitness is not assessed. However, by overlooking the phenotypic side of adaptation, no complete picture of the process of adaptation can be obtained.

1.3 QUANTITATIVE GENETICS AND ADAPTATION

1.3.1 Polygenic traits

Phenotypic variation occurs naturally within populations. All traits exhibit a certain degree of variation. The distribution of the traits can be either discrete or continuous. Phenotypes of the former category are, for instance, those observed by G. Mendel in *Pisum sativum* such as seed color (yellow or green), seed shape (round or wrinkled), flower position (axial or terminal) or in human diseases, such as alkaptonuria as described by A. Garrod in 1902 (Scriver 2008). Continuous traits include body size, body fat content, photosynthetic rate, growth rate, flowering time, etc. Single genes underlie discrete traits, such as those mentioned above. For historical reasons, traits with this simple genetic architecture were named Mendelian traits.

Around 1900, although animal breeders and physicians were well aware that continuous traits were highly heritable, the exact genetic mechanisms underlying these traits remained a mystery. The British Biometric School, led by F. Galton, K. Pearson and W.F. Weldon, set to study continuous traits. They developed a considerable amount of statistical tools to trace their inheritance and study their role in evolution, which was based on Darwinian micro-mutational ideas (Lynch and Walsh 1998; Barton and Keightley 2002; Orr 2005). It was difficult at that time to consider that the continuous and Mendelian traits shared the same hereditary and evolutionary properties.

R. Fisher's theoretical contribution to the study of quantitative traits closed the gap between Mendelian and continuous, or complex, traits. Fisher showed in 1918 that a continuous phenotypic distribution could result from the effect of several loci. The more genes involved, the smother the distribution. It was necessary to assume that the effects of each locus were small and purely additive. This result, however, did not prompt an interest in investigating mutation rates and fitness effects at each of the loci affecting a continuous trait. It was necessary to know what loci to look at in the first place. Gene mapping techniques were already available at that time, but they were more useful to researchers working on monogenic traits than continuous ones (see Stutervant 1913).

Further developments of the theory of adaptation, also contributed by Fisher, partly justified the lack of interest in single locus effects (Orr 2005): if the trait of interest is governed by an infinite number of genes, all of them exerting an equally small effect, the most effective way to study their role in evolution is by considering their aggregate effect

on fitness. Fisher found out that new mutations with infinitesimally small phenotypic effects are more likely to be beneficial than those of larger effect. A conclusion drawn from this result is that small mutations are the genetic basis of adaptation (this theoretical finding is at the core of his so called geometric model, see Orr 2005 for a review). Later, Kimura would reevaluate Fisher's results to conclude that alleles of medium effect size were most important in adaptation (Orr 2005).

Data from artificial selection experiments seemed to provide support for the geometric model. For instance, a sustained response to selection has been reported in plants and animals. Selection for increased abdominal bristle number in *Drosophila* steadily continued for up to 90 generations (Yoo *et al.* 1980), and selection for oil content in maize kernels was possible for about 20 years (a period of time that corresponds to approximately 70 maize generations). In the latter experiment, selected lines changed from an original 4.7% to a current 20% oil content (Laurie *et al.* 2004). One can easily imagine that with high heritability and a trait architecture based on thousands of loci with very small effects, such long-term response to selection can be maintained.

The refinement of gene-mapping techniques and the development of marker-based genetic maps in model species, as well appropriate statistical methods, allowed the localization of genetic factors affecting continuous traits, best known as quantitative trait loci (QTL) (Lander and Botstein 1989; Falconer and Mackay 1996). The accumulating body of QTL studies soon revealed that complex traits are governed by a finite number of loci, and that not all loci have the same allele effects and that these effects tend to be exponentially distributed (Mackay 2001). In *Drosophila*, traits such as sensory bristle number (Gurganus *et al.* 1999; Mackay and Lyman 2005), wing shape (Weber 1999), and longevity (Nuzhdin *et al.* 1997; Valenzuela *et al.* 2004; Wilson *et al.* 2006) showed this genetic architecture. Such evidence, at odds with Fisher's geometric model, suggests that mutations with big effects are also important for adaptation (Orr 1998). Ingenious developments of the methods to reveal the genetic composition of complex traits have made possible the fine mapping of QTL down to quantitative trait genes (QTGs) and even to SNPs, giving rise to the concept of a quantitative trait nucleotide (QTN). Thanks to these approaches we are learning that complex traits in humans and other species are even more "polygenic" than suggested by classical QTL mapping approaches (Risch and Merikangas 1996; Mackay 2001; Mackay *et al.* 2009).

1.3.2 Polygenic adaptation

The possibility of identifying the genes and nucleotide variants that affect complex traits opens new avenues to study how these traits evolve, especially how adaptation proceeds when several genes affect a fitness-associated trait (Rockman 2012). An attractive approach to do so is to extend current theory of positive selection on single loci to the polygenic case (Chevin and Hospital 2008).

In contrast to the ideas of adaptation at a single locus, in which a beneficial allele is driven to fixation by positive selection (see section 1.2), polygenic adaptation does not require new beneficial mutants and may not lead to the fixation of beneficial alleles. Selection acts on standing genetic variation at all involved loci (Barton and Turelli 1989; Falconer and Mackay 1996; Chevin and Hospital 2008; Messer and Petrov 2013). Thereby populations adapt by allele frequency shifts at many loci if environmental changes result in a new phenotypic optimum. Once the average phenotype in the population matches the new optimum the intensity of selection will decrease. While this process could allow very rapid adaptation, no conspicuous footprints will be left on linked neutral variants making this selective event difficult to detect with current population genetic methods.

Nevertheless the chance for the fixation of beneficial quantitative alleles is not negligible, at least in theory. Recently, Pavlidis *et al.* (2012) analyzed a model with n loci controlling a trait under stabilizing selection. They conclude that multilocus response to selection may in some cases prevent selective sweeps from being completed, but that conditions causing this to happen strongly depend on the genetic architecture of the trait. For instance, the probability of fixation of selected mutations decreases with the number n of loci involved and also depends on their effect sizes. Fixations are more common when the effects are about equal (in absolute size). This could partly explain the relative success that selective sweep (or hitchhiking) mapping approaches have achieved as QTL mapping tools (Nuzhdin and Turner 2013).

1.4 COLD TOLERANCE IN *D. MELANOGASTER*: A CASE STUDY OF ADAPTATION

Temperature is one the most important abiotic factors that determines species abundance and distribution (Hoffmann and Blows 1994; Hoffmann *et al.* 2005; Geber 2011). Its effect on life history traits such as growth (developmental time); sexual maturation and reproduction has been well documented in insects (Régniere *et al.* 2012). Cold temperatures are inherent features of thermally heterogeneous habitats, becoming more pronounced with latitude or altitude. Cold represents one of the sources of environmental stresses to which animals have evolved numerous physiological strategies to counteract its negative effects (Ayrinhac *et al.* 2004; Hoffmann *et al.* 2005; Koštál *et al.* 2007). A key quest is to understand how organisms cope with cold, identifying the mechanisms that allow them to survive and reproduce successfully under sustained cold conditions.

D. melanogaster is a suitable organism to study cold adaptation. Since its origin in Sub-Saharan Africa, the fruit fly has spread all over the world and adapted to diverse climatic conditions (David and Capy 1988; Pool and Aquadro 2006). Although *D. melanogaster* diverged from its closest relative *Drosophila simulans* around 2.3 Mya (million years ago) (Li *et al.* 1999), the first out-of-Africa migration event took place around 19,000 ya, when the African and non-African populations first separated. Subsequently, multiple colonization events happened in the last 10,000 years including the colonization of Asia ~5000 ya (Laurent *et al.* 2011), Australia ~1000 ya (Lachaise *et al.* 1988), and North America ~300 ya (Keller 2007; Duchen *et al.* 2013).

Two aspects are relevant from *D. melanogaster*'s colonization history. First, the time spent in Africa by this organism, from its origin until the first out-of-Africa migration event, was long enough to generate sufficient genetic diversity on which selection can act. Second, the fly has successfully colonized multiple environments, which include high latitude zones with extreme ambient temperature fluctuations. For instance, viable populations have been reported in Scandinavia (Bächli *et al.* 2005), North America (Keller 2007), southern Australia and Tasmania (Hoffmann and Parsons 1989). With this background, it is clear that such a successful establishment in temperate zones has been aided by the development of strategies to cope with cold stress (Kimura 1988; Izquierdo 1991; Goto *et al.* 1999; Gibert *et al.* 2001).

For outdoor *D. melanogaster* populations, cold stress tolerance is positively correlated with latitude (Hoffmann *et al.* 2001; Kimura 2004). This pattern has been observed in the two hemispheres, namely along the Australian and North American east coasts. Flies collected in high latitude locations of these continents consistently exhibit more cold stress resistance than their subtropical and subtropical/tropical conspecifics (Hoffmann *et al.* 2002; Schmidt *et al.* 2005; Svetec *et al.* 2011). There are several methods to assess cold stress resistance in *D. melanogaster* (see Hoffmann *et al.* 2003b for a review). Briefly, cold tolerance assessments are done by exposing flies to either freezing (< -5 ºC) or chilling (~0 ºC) temperatures for a given amount of time. In the case of exposure to below-zero temperatures, individual survival is scored, while under less life-threatening chilling conditions that induce a coma-like state, the time to recover from the chill-induced coma is recorded as measure of stress resistance. Flies that take less time to return to an upright position are regarded as more cold resistant. Although cold shock resistance and chill coma recovery time (CCRT) are proxies for cold tolerance, the physiological changes triggered by these stimuli may have different underlying mechanisms (Macmillan and Sinclair 2011).

Currently, the available full-genome sequences from different worldwide populations of *D. melanogaster* constitute the most complete whole-species range catalogue of genetic variants. This variation can be studied to identify latitudinal and attitudinal patterns that lead to the inference of adaptation to cold environments. Tropical and temperate populations are already part of the datasets. Thus far, Africa is the best-sampled continent (Pool *et al.* 2012), followed by North America (Mackay *et al.* 2012, DGRP) and to a lesser extent, Europe (Pool *et al.* 2012). Given the great potential of these datasets for conducting comprehensive population and quantitative genetic analyses on a species-wide scale and considering the current bias towards tropical/ancestral populations, it is necessary to direct further sampling efforts to higher latitudes. The inclusion of alleles naturally selected in extreme environments can help us understand where in the genome of the fly reside the keys to its successful colonization of the world.

The study of variation in CCRT in *D. melanogaster* has revealed a significant heritable component, leading to the conclusion that cold stress resistance is indeed an evolvable trait (Hoffmann *et al.* 2001; Anderson *et al.* 2005). The availability of a suitable genetic toolkit for this species has served as a means to dissect the molecular basis of CCRT. By means of recombination mapping, QTL have been mapped onto chromosome 3R

(cytological intervals 76B-87B and 73A–90B). These QTL explain 10–35% of the total variation in CCRT. QTL on the X chromosome have also been identified and explain up to 14% of the variation for CCRT between European and African strains of *D. melanogaster* (Svetec *et al.* 2011). The main drawback of this mapping strategy is its lack of resolution. The aforementioned QTL intervals are on the order of 1Mb in length and often contain hundreds of annotated or computationally predicted genes that can be potentially associated with the phenotype, making downstream gene validation analysis a laborious endeavor.

The implementation of genetic complementation tests in the quantitative genetics context has served as a way to reduce the size of QTL intervals, hence reducing the number of likely associated genes (Mackay 2001). This is achieved by comparing the effects on the phenotype of chromosomal aberrations, chromosomes with deleted fragments or deficiencies *vs.* those of complete chromosomes. In classical genetics, complementation testing has been used to find the role of a given gene by comparing phenotype effects of wild type allele *vs*. a recessive mutant (Mackay 2001). In the case of quantitative deficiency complementation tests, chromosomal deletions assume the role of the mutation, and a deletion that reveals effects similar to those of the entire QTL are regarded as complementation failures (Pasyukova *et al.* 2000). However, the interpretation of a quantitative failure to complement is subject to several caveats (Mackay 2001; Service 2004). So far, quantitative complementation tests have been used to achieve high-resolution mapping of a QTL affecting lifespan (Pasyukova *et al.* 2000; Wilson *et al.* 2006), leading to the identification of candidate QTGs for this trait (De Luca *et al.* 2003; Pasyukova *et al.* 2004). In the case of CCRT, successful fine mapping of a QTL to a causative QTG has been done on the basis of previous knowledge of the genes' association with stress resistance, as in the case of *Smp-30* and *Frost* on chromosome 3R (Clowers *et al.* 2010).

Available full genome sequence data of fly panels with reported variation for CCRT allows for the association of genetic variants at the nucleotide level, single nucleotide polymorphism (SNPs), with a given score for CCRT. This approach constitutes a huge leap in mapping resolution compared to QTL mapping. Mackay *et al* (2012) conducted such genome wide association study (GWAS) and found significant association ($P < 10^{-5}$) of 295 SNPs (out of ~2.5 million) with variation for CCRT in a sample of 168 inbred strains of the *Drosophila melanogaster* genetic reference panel (DGRP). These significant 295 SNPs,

interspersed across the genome, can explain up to 78% of the observed phenotype variation. However, as pointed out by the authors, although this list of candidate SNPs is likely to be enriched for true QTNs, the actual QTN can be in linkage disequilibrium with the trait-associated SNPs and in some instances the association with the phenotype may be spurious; that is the candidate SNP has no effect whatsoever on CCRT. Furthermore, if the associated SNP is not part, or at least in proximity, of any annotated gene, then its functional validation turns into a genome annotation endeavor at a local scale. Systems biology approaches can contribute to uncover of variants affecting CCRT (Ayroles *et al.* 2009)

1.5 OBJECTIVE AND STRUCTURE OF THIS THESIS

This work represents an experimental approach to study positive selection at polygenic traits. The main research question addressed in this study is whether loci underlying traits that experienced adaptive evolution show footprints of positive selection and, if so, what sort of positive selection signal do they exhibit. In the same spirit as Svetec (2009) and Werzner (2011), we used cold tolerance as the complex trait that has evolved adaptively in temperate populations of *D. melanogaster*. The QTL that affect this trait were previously identified by Svetec *et al* (2011). Here we studied three of these QTL with the aim of increasing their mapping resolution so that the mode of evolution of these genomic intervals can be studied via population genetic analyses. This approach provides empirical data to test emerging models of polygenic selection. The way we undertook this task was through three different projects that are presented in sections 1 through 3 of chapter 2 (Results).

In section 2.1, we use a selective sweep mapping approach to study the genes of the X-linked QTL encompassing cytological regions 13E-20E (Sevetc 2009; Werzner 2011; Svetec *et al.* 2011). In doing so, we establish a link to a previously reported selective sweep in this region (Li and Stephan 2006) analyzing in greater detail the evolutionary history of the cytological interval 15E.

In section 2.2, we focus on a combined quantitative and functional approach to fine map the broadest QTL affecting cold tolerance mapped by Svetec (2009) and Werzner (2011), spanning the cytological region 6C-11D. This QTL is by far the one with the largest effect on the trait.

Section 2.3 revisits the population genetics of a QTL. In this part, we use SFS and F_{ST}-based methods to better understand the evolutionary history of the chromosomal region containing the candidate QTGs identified in section 2.2. These analyses document one of the strongest signals of positive selection in European *D. melanogaster*.

Section 2.4 reports on a satellite project that is relevant because of its long-term implications for the study of molecular evolutionary biology in our research group. In this chapter we introduce a new full-genome sequence dataset of a Scandinavian *D. melanogaster* population and gauge its potential to study adaptation to high latitude ecological conditions in a natural population.

Finally, chapter 3 consists of a discussion of the findings reported in the preceding chapter and chapter 4 presents a unified Materials and Methods section.

II – RESULTS

2.1 SELECTIVE SWEEP MAPPING OF A QTL FOR COLD STRESS TOLERANCE

2.1.1 Co-localized QTL and valleys of genetic variation

Positive selection drives the evolution of ecologically relevant traits such as cold stress tolerance in *D. melanogaster*. There is growing evidence that supports the polygenic nature of this phenotype (Mackay *et al.* 2012). The question that arises from an evolutionary perspective is whether the underlying loci bear footprints of positive selection, and if so, to what extent they do. Currently, the selective sweep model is a well-studied case of positive selection in sequence space describing the effect that the fixation of a beneficial mutation leaves on neighboring neutral variation (see section 1.2). The characteristic signal of a selective sweep can be used to assist the fine mapping of QTL affecting an adaptive trait (Nuzhdin *et al.* 2007). Recently, Svetec (2009) and Werzner (2011) identified X-linked genetic factors underlying the difference in cold stress tolerance between two natural populations of *Drosophila melanogaster,* one from the Zimbabwean shore of Lake Kariba and the other from the seaside locality of Leiden, in the Netherlands. The results of their work revealed a series of QTL along the entire X chromosome.

These QTL map to the following cytological intervals in both sexes: 6C-10B, 8E-11D and 13E-20E. Because of the breadth of these intervals, up to 1,000 genes are potential QTGs for cold stress tolerance. Moreover, the lack of *bona fide* candidate genes located on the X does not facilitate the sorting of our potential QTGs by functional relevance. The interval at 13E-20E has an approximate length of 6.7 Mega base pairs (Mb) and co-localizes with a previously identified selective sweep exclusive to the Netherlands population (Li and Stephan 2006). Taking advantage of the co-occurrence of these two signals in the same chromosomal interval, we explored the possibility to conduct a fine scale reassessment of this region using a selective sweep mapping strategy. Our approach was viable because: (i) the QTL is significant in both male and female flies and does not exhibit QTL–sex interactions and, (ii) the sweep is specific to the Dutch population. This means that the beneficial allele involved in adaptation to the new environmental conditions is more likely to be observed in temperate populations. The sweep region in question encompasses ~86 kb, located exactly at cytological band 15E. Li and Stephan (2006) determined the sweep-like features of this window by studying the SFS of three 500-bp fragments scattered along this interval. They noticed that the lowest point of genetic variation in the Netherlands was seen in the fragment located in the 3.5-kb long region between genes *CG16700* and *CG4991*.

Further characterization of this sweep region requires more detailed sequence data from both the Netherlands and Zimbabwean populations. We therefore increased the amount of SNPs by sequencing 12 additional fragments along the 86 kb region. These fragments were placed in non-coding regions, such that the average distance between them was 10 kb. In addition we also sequenced the entire *CG16700 – CG4991* intergenic region. With this new dataset we calculated a suite of summary statistics including genetic diversity estimates θ_W and $\theta\pi$, haplotype diversity (*H*) and average LD (Z_{nS}). Since the two genetic diversity estimates are explicitly related to the SFS of the region and this was previously explored (Li and Stephan 2006), we show them together with the other summary statistics in Appendix A. Alternatively, haplotype diversity and average LD estimates per fragment, are loosely associated with the SFS and convey information that we want to explore further in this section. The profile obtained for Z_{nS} within each fragment in both populations is shown in Figure 1A. This profile had a distinctive pith-like shape in the Netherlands with above-average Z_{nS} values on both sides of the selective sweep at relative positions 32,900 and 35,900.

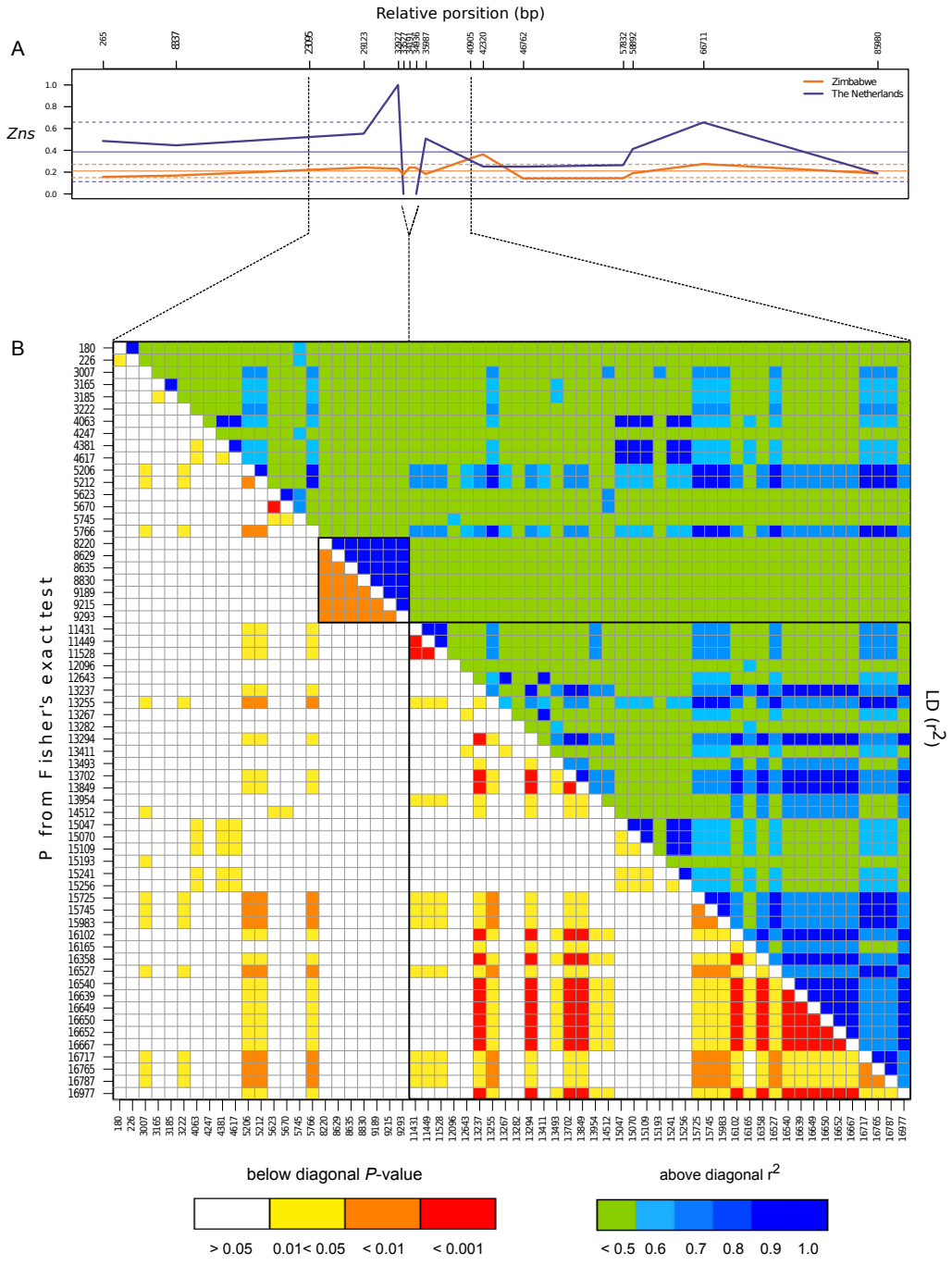


Figure 1. LD patterns at 15E.

A) LD (Z_{nS}) profile for a set of 12 fragments located in the 86 kb region of interest as 15E. Note the pit-like behavior of Z_{nS} around relative position 34,000 in the Netherlands population. B) LD matrix for 69 relevant SNPs in the approximately 18-kb long *CG16700 – CG4991* gene region. Patterns of LD (r^2) are shown above the diagonal and *P* values from Fisher's exact test below the diagonal. Notice the LD block structure of this dataset, the SNPs 8,220 to 9,293 form an LD block that is not LD with the other adjacent SNPs.

The two flanks of the 3-kb long pit-like reduction in sequence diversity show haplotype diversity values above 50%. This pattern is reminiscent of the theoretical expectation of LD blocks at the flanks of selective sweep regions (Kim and Nielsen 2004).

We conducted a fine scale LD analysis for this region available next generation sequence data from the Netherlands population (see Material and Methods, Table 8). We obtained pairwise LD estimates for all LD informative SNPs within a 17 kb window centered on the *CG16700* - *CG4991* gene region. The resulting LD matrix (Figure 1B) comprises a total of 62 SNPs that do not include singletons. Interestingly, the two sets of SNPs at both sides of the sweep form two noticeable LD clusters (framed within the matrix.) We noticed that the right-hand side LD cluster is composed of three SNPs within a range of 100 bp, which seems to be part of a yet broader cluster in course of erosion. With this LD matrix we could also observe how LD is broken across the two main LD blocks.

2.1.2 Positive selection in the *CG16700* - *CG4991* region

In order to quantify the extent to which this LD pattern agrees with the expected selective sweep scenario, we subjected the 6 kb region comprising the coding region of *CG4991* and its 5' UTR to a formal LD analysis using the ω_{MAX} statistic (Kim and Nielsen 2004). The profile of the ω statistic in this region for the Netherlands is shown in Figure 2A. The dashed line indicates the 95th percentile of ω_{MAX}, which we obtained by neutral simulations (see Materials and Methods). With a *P*-value of 0.037 the observed maximum of the distribution of ω values is located within the 3 kb between the coding regions of *CG16700* and *CG4991*, most precisely between relative positions 1.3 to 2 kb. While this statistical approach confirmed the presence of a selective sweep immediately upstream of the coding region of *CG4991*, the question still remains as to what the exact location of the selected variant is. The approximately 700 bp that encompass the target of selection are devoid of any variation in the Netherlands, so in principle any of these 700 sites could bear the beneficial allele. To address this question we made use of available full genome sequence data from other African *D. melanogaster* populations (Pool *et al.* 2012) with which we explored patterns of allele frequency differentiation per site.

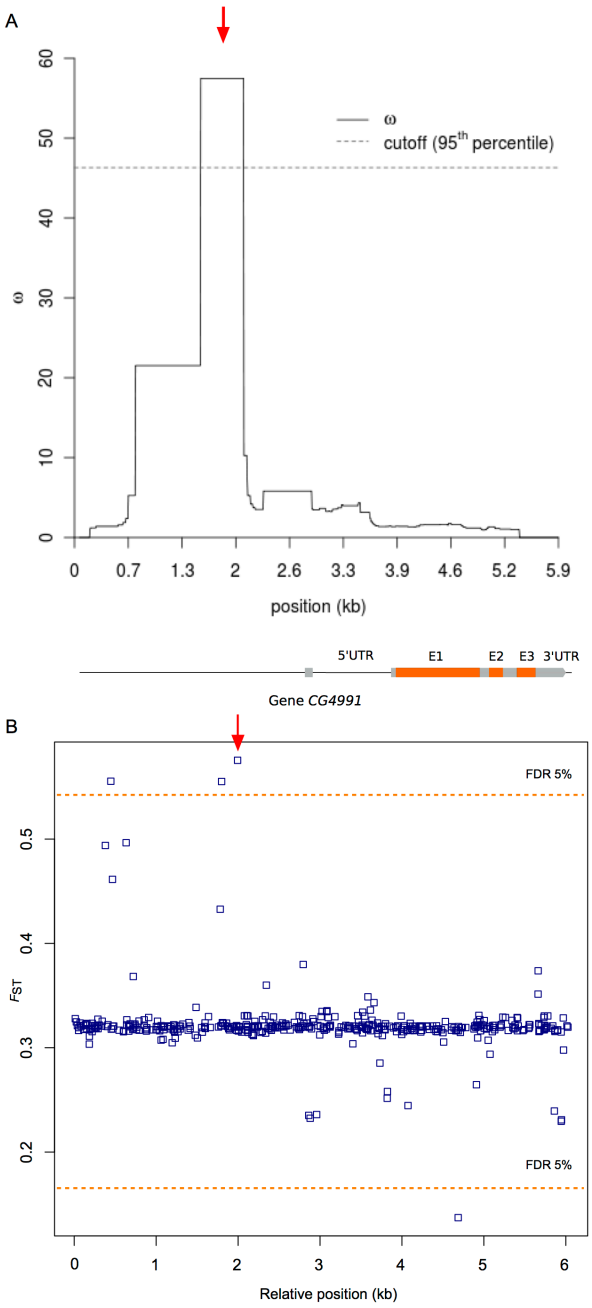


Figure 2. Selective sweep mapping at cytological interval 15E.

The tests were conducted on the 6 kb encompassing the interval between genes *CG16700* and *CG4991* (in this case it is also the 5' UTR) and the entire coding region of the latter. A) Omega statistics (ω) maximizes between positions 1.6 and 2 kb from the relative point of origin. B) Model-based F_{ST} values for 422 SNPs from a dataset including two European and five of African samples (see Materials and Methods). Notice the outlier SNP positions with above and below average F_{ST} values, in particular those showing high differentiation enriched in 5'UTR of *CG4991* (indicated by a red arrow). FDR of 5% thresholds are marked as orange dashed lines.

Positive selection may fix beneficial alleles in local populations increasing the level of allele frequency differentiation among localities within the entire population range (Lewontin and Krakauer 1973). We subjected the SNP dataset derived from the Netherlands and five African populations to an F_{ST} analysis implemented in BayeScan (Foll and Gaggiotti 2008). We obtained F_{ST} coefficients for a total of 422 SNPs, with an average F_{ST} coefficient of 0.3213. F_{ST} values per SNP are shown in Figure 2B. The handful of SNPs with F_{ST} values reflecting the highest differentiation across populations are well above 0.55, and are considered significant outliers at a FDR rate of 5% (orange dashed line). The three top significant outlier SNPs are located at relative positions 1,994 (F_{ST}=0.5755, q-value=0.0062), 1,799 (F_{ST}=0.5554, q-value=0.0050), and 447 (F_{ST}=0.5551, q-value=0.0085). Interestingly, the first two of these SNP are located within the 700-bp target of selection identified with the ω statistic (Figure 2A,B). In addition, BayeScan also detected a significant outlier SNP at 4,686 with an F_{ST} value of 0.1371 (q-value=0.0536), located in exon 1 of *CG4991*. By inspection of the corresponding polymorphism table (data not shown), the same allele is kept at low frequency in all populations; therefore this site is a suspected target of purifying selection.

We gained great insight characterizing this selective sweep and identified a cluster of SNPs in the region between *CG16700* and *CG4991* as putative targets of positive selection. A crucial question remains, whether these SNPs are also affecting CCRT. Our first thought was that, because of their location (immediately upstream of *CG4991*) these SNPs could be affecting the expression of *CG4991*. We conducted gene expression assessments via qPCR in lines from Zimbabwe and the Netherlands and observed that expression of this gene is on average higher in European flies than in African ones. Furthermore, we also noticed that there is more variation in the level of expression of this gene among Dutch lines than among Zimbabwean ones (Figure 3). Although we measured gene expression in flies of both sexes separately we did not observe any sex-specific expression difference.

The exact opposite pattern of expression was observed for *CG16700*. For this gene we detected a ~1.5 fold difference in average expression between populations, however it was not significant. The difference between sexes was significant only in the Netherlands (P>0.05). Interestingly for *CG16700* variation among Dutch lines is ostensibly lower than among Zimbabwean flies (Figure 3).

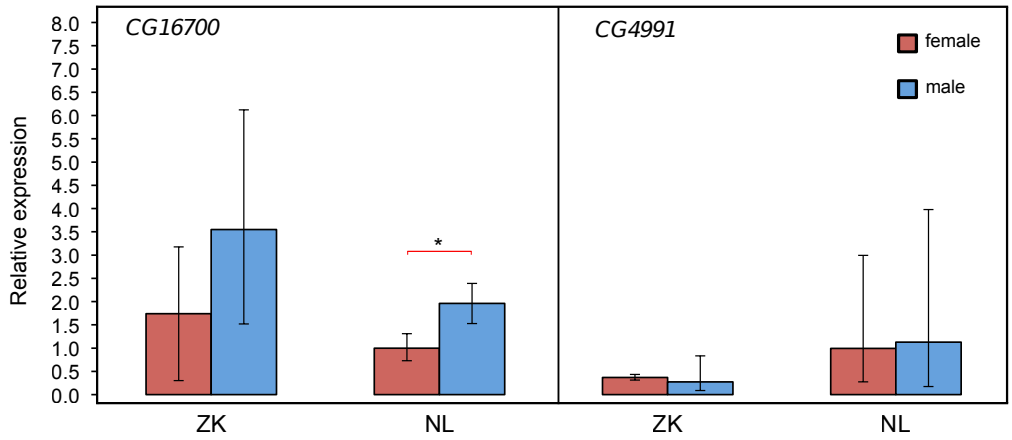


Figure 3. Expression assays of candidate genes at interval 15E.

Expression of genes *CG16700* and *CG4491* was measured at the constitutive level in flies of both sexes from a temperate population the Netherlands (NL) and a tropical location Zimbabwe (ZK). Expression level of these candidate genes was normalized respect to that of ribosomal gene *RpS20*. The height of the bars represents the mean of three calibrated normalized relative quantities (CNRQ) per population per sex, per gene rescaled to that of the NL females. Error bars also represent rescaled confidence intervals. Levels of significance for tests of differences in expression levels among treatments within and between populations are indicated with asterisks, P<0.05 (*), P<0.01(**) and P<0.001(***).

This selective sweep-based strategy revealed an intergenic region targeted by selection with a still unclear functional role. Since the target of selection is located upstream of *CG4991*, it is sensible to think that the expression of this gene is the phenotype being tuned by selection. Interestingly, our qPCR assays (Figure 3) evidenced a non-buffered (*i.e.* variable) *CG4991* expression pattern among Dutch lines. This result may suggest that the aim of selection here is to promote plasticity in gene expression, rather than reducing its variation. However, before formulating any hypothesis on this assumption, it is important to determine the functional role of this intergenic target of selection and establish how it affects the expression of its two flanking genes; *CG4991* and *CG16700* in European and African populations. Only then it should be possible to evaluate the possibility of a link between this selective event and cold stress tolerance. Further discussion of this selective sweep mapping approach is provided in section 6.2.2.

2.2 FINE MAPPING OF A QTL FOR COLD TOLERANCE

2.2.1 Quantitative complementation mapping

The use of selective sweep mapping to aid the characterization of QTL affecting adaptive traits is justified by the expectation that the loci influencing an adaptive phenotype (that has been changed by directional selection) also show footprints of positive selection. However, because of the hard sweep model assumptions upon which the methods are built, only strong signals of selection can be detected. Due to this conceptual bias we might overlook other selection events where the fixation of the beneficial allele did not occur. In fact, it has been suggested that this scenario is more common than hard sweeps in the case of polygenic selection (Pritchard and Di Rienzo 2010). An additional weakness of the approach is that it does not provide conclusive evidence to establish a functional connection between the site (or gene) target of selection and the studied phenotype. This is especially true when mapped QTL are broad and contain several candidate genes. In this section we undertook a quantitative genetics approach that avoids the biases inherent to selective sweep mapping and explores the functional link between the genotype and the phenotype. The aim here is to find candidate QTGs for cold tolerance within the QTL mapped by Svetec *et al.* (2011). The approach presented here is based on two quantitative variants of genetic complementation tests (Pasyukova *et al.* 2000; Mackay 2001) assisted by gene expression assays.

We started by dissecting two X-linked QTL that affect the difference in chill comma recovery time (CCRT) between African and European populations of *D. melanogaster*, as reported by Svetec *et al.* (2011). These QTL encompass to the cytological interval at 6C-11D (of approximately 6.2 Mb in length). We dissected this interval via quantitative complementation tests with 24 overlapping chromosomal deletions spanning 94% of this interval. The chromosome fractions making up the remaining 6% of this interval were left untested due to lack of suitable deletions. The Dutch and Zimbabwean versions of the X chromosome used in these tests are contained in fly lines A* and E*, created by introgression of one X chromosome from a population of the Netherlands and one from Zimbabwe into a common laboratory strain (Svetec *et al.* 2011). Hence, these two lines bear different X-linked alleles while the rest of the major nuclear chromosomes and mitochondrial DNA is the same. These two lines are the parents of the X-recombinant population employed to map the QTL that concerns us in this project (see Svetec 2009).

With a set of 24 deficiencies (Figure 4) we could potentially uncover the effect of line-specific alleles (line-specific refers to the type of X chromosome involved in the test, which is either African or European) at 588 (95%) of the 622 annotated and computationally predicted genes within the interval. A total of 14 of the 24 tested deficiencies showed significant line effects at the 5% level, whereas 9 of 24 showed a significant effect of the genomic background on CCRT scores (with the term "genomic background" we refer to the involved deletion and balancer chromosomes; see Materials and Methods). We observed failure to complement in 4 of the 24 tested deletions (Table 1, Figure 4). Failure to complement implied both a significant effect of line (*L*) and a significant 'line by genomic background' interaction (*L x G*) as long as the differences in CCRT followed the expected direction. That is, shorter CCRT times for flies bearing the E* X chromosome in the presence of the deletion compared to the corresponding flies bearing the A* X chromosome in the presence of the same deletion, while showing no difference between the CCRT of the flies bearing the E* and A* X chromosomes in the presence of the balancer chromosome.

Deletion *Df(1)ED6906* encompasses a fragment of 210 kb; this deletion was one of the two that revealed a highly significant failure to complement (Table 1, Figure 4). The difference between the means of the CCRT scores for the flies bearing this deletion is 9.18 minutes, whereas that of the flies harboring the balancer chromosome is 1.82 minutes (Table 1). In other words, E*/ *Df(1)ED6906* flies woke up on average 4.87 minutes faster than their respective balancer counterparts and that A*/ *Df(1)ED6906* flies recovered on average 4.37 minutes later than flies in the respective balancer background. Deletions *Df(1)C128*, *Df(1)BSC592* and *Df(1)BSC537* also failed to complement as revealed by the significant line and *L x G* effects. However, for the two last deletions these effects were marginally significant. Their respective differences in average CCRT between the E* and A* X chromosomes in the deficiency and balancer backgrounds can be seen in Table 1.

The fact that we used a set of overlapping deficiencies allowed us to better define the stretch that revealed quantitative failure to complement. With respect to the 210-kb long span of deletion *Df(1)ED6906*, the 67.15 kb overlapping with deletion *Df(1)BSC536* were subtracted from the stretch of interest (Figure 4). Furthermore, the results of the complementation tests with yet another overlapping deficiency at the same end (*Df(1)BSC711*) allowed us to subtract additional 19.64 kb from the 210 kb encompassing

Df(1)ED6906 (Table 1, Figure 4). At the other end of deletion *Df(1)ED6906*, its overlap with deletion *Df(1)HA32* is not known at the base pair level. Thus, the redefined fraction of interest under deletion *Df(1)ED6906* encompasses a total of 124 kb (between coordinates 7,089,000 and 7,212,999). We determined, in a similar way, the fraction of interest under deletions *Df(1)C128*, *Df(1)BSC592* and *Df(1)BSC537*.

This fine mapping approach, based on overlapping deletions, has allowed us to reduce the number of initial candidate genes within the QTL under study, from 622 to a subset of 89. A total of 7 genes are located within the 124 kb uncovered by deletion *Df(1)ED6906*, a total of 14 genes were uncovered by deletion *Df(1)HA32*, 19 genes by *Df(1)BSC592*, and 49 genes by *Df(1)BSC537*. This is remarkable given the substantial fraction of uncharacterized genes in the 6.2 Mb of the QTL defined by Svetec *et al.* (2011) and the absence of known *a priori* candidate genes for CCRT in this chromosomal region.

Table 1. Deficiency analysis of QTL at 6C-11D affecting CCRT in female flies (next page).

This table summarizes quantitative deficiency tests performed with the listed deletions. Δdef is the difference between the average CCRT of flies bearing E* and A* chromosomes in the presence of a given deletion. Negative differences suggested the presence of CCRT reducing alleles at sites potentially uncovered by the deletion. Δbal is the difference between the average CCRT of flies bearing E* and A* chromosomes in the presence of a given balancer chromosome. Note that deletions held with the same balancer show the same the Δbal values. P_{LINE} is the value for the line effect (E* or A*) from two-way Anova analysis. P_{GEN} is the value for the genomic background effect (deletion or balancer) effect from two-way Anova analysis. $P_{LINE \times GEN}$ is the value for the interaction between the two above-mentioned variables.

Table 1. Deficiency analysis of QTL at 6C-11D affecting CCRT in female flies

Deletion	Balancer	Mean CCRT (SD) in minutes				Δ*def*	Δ*bal*	*P* L	*P* G	*P* L *x* G
		E/deletion*	*E*/balancer*	*A*/deletion*	*A*/balancer*					
Df(1)BSC351	*FM7h*	31.70 (7.99)	30.91 (7.88)	31.46 (9.42)	32.73 (8.72)	0.24	-1.82	0.070057	0.838897	0.343401
Df(1)BSC882	*FM7h*	29.39 (9.94)	30.91 (7.88)	32.69 (12.41)	32.73 (8.72)	-3.3	-1.82	0.018773	0.149609	0.574953
Df(1)HA32	*FM7c*	37.75 (8.24)	32.16 (9.01)	41.61 (10.42)	33.25 (9.28)	-3.86	-1.09	0.177930	0.000001	0.454433
Df(1)ED6906	*FM7h*	26.93 (5.66)	30.91 (7.88)	36.28 (8.52)	32.73 (8.72)	-9.35	-1.82	0.000103	0.779708	0.000289
Df(1)BSC711	*FM7h*	35.73 (6.52)	30.91 (7.88)	34.03 (8.45)	32.73 (8.72)	1.7	-1.82	0.944120	0.011307	0.121501
Df(1)BSC536	*FM7h*	36.27 (9.81)	30.91 (7.88)	36.43 (11.34)	32.73 (8.72)	-0.16	-1.82	0.055280	0.002486	0.473532
Df(1)BSC622	*Binsinscy*	33.37 (9.22)	35.59 (8.83)	34.50 (8.65)	36.60 (9.59)	-1.13	-1.01	0.386305	0.121365	0.820123
Df(1)C128	*FM6*	26.97 (6.16)	30.11 (8.63)	37.44 (8.11)	32.63 (7.74)	-10.48	-2.52	0.000012	0.606252	0.000885
Df(1)BSC366	*Binsinscy*	36.26 (8.16)	35.59 (8.83)	37.46 (11.06)	36.60 (9.59)	-1.21	-1.01	0.480645	0.501206	0.850165
Df(1)BSC662	*Binsinscy*	40.29 (8.98)	35.59 (8.83)	39.89 (9.70)	36.60 (9.59)	0.39	-1.01	0.330040	0.002076	0.615274
Df(1)BSC592	*Binsinscy*	31.42 (7.47)	35.59 (8.83)	37.92 (9.16)	36.60 (9.59)	-6.51	-1.01	0.063548	0.639100	0.031094
Df(1)Exel6241	*FM7c*	31.57 (7.29)	32.16 (9.01)	30.29 (8.76)	33.25 (9.28)	1.29	-1.09	0.685768	0.224516	0.294193
Df(1)ED6957	*FM7j*	27.90 (7.68)	32.16 (9.01)	29.21 (6.77)	33.25 (9.28)	-1.31	-1.09	0.261153	0.001693	0.795931
Df(1)BSC537	*FM7h*	29.64 (7.36)	30.91 (7.88)	35.36 (8.17)	32.73 (8.72)	-5.71	-1.82	0.004959	0.528960	0.099098
Df(1)BSC712	*FM7j*	40.21 (8.39)	35.59 (8.83)	39.66 (9.36)	36.60 (9.59)	0.55	-1.01	0.623921	0.005147	0.565518
Df(1)BSC539	*Binsinscy*	31.50 (7.07)	35.59 (8.83)	34.30 (7.52)	36.60 (9.59)	-2.8	-1.01	0.239470	0.026941	0.414266
Df(1)ED7005	*FM7h*	29.63 (6.45)	30.91 (7.88)	29.21 (6.46)	32.73 (8.72)	0.43	-1.82	0.060255	0.075504	0.350039
Df(1)BSC755	*Binsinscy*	30.36 (7.49)	30.11 (8.63)	33.33 (6.92)	32.63 (7.74)	-2.98	-2.52	0.009057	0.547656	0.88738
Df(1)BSC540	*FM7h*	32.20 (7.51)	30.91 (7.88)	34.73 (9.31)	32.73 (8.72)	-2.53	-1.82	0.022620	0.141156	0.845784
Df(1)BSC572	*FM7h*	31.11 (6.04)	30.91 (7.88)	37.05 (7.54)	32.73 (8.72)	-5.94	-1.82	0.009129	0.070999	0.152617
Df(1)BSC287	*Binsinscy*	36.11 (9.57)	35.59 (8.83)	35.23 (8.06)	36.60 (9.59)	0.87	-1.01	0.596120	0.852371	0.643564
Df(1)ED7067	*FM7h*	28.89 (7.99)	30.91 (7.88)	29.58 (7.80)	32.73 (8.72)	-0.68	-1.82	0.039538	0.029350	0.702833
Df(1)Exel6242	*FM7c*	33.96 (7.48)	32.16 (9.01)	32.93 (7.68)	33.25 (9.28)	1.03	-1.09	0.585559	0.397537	0.406384
Df(1)ED7147	*FM7h*	31.43 (7.44)	30.91 (7.88)	34.96 (8.42)	32.73 (8.72)	-3.52	-1.82	0.010630	0.272976	0.497331
Df(1)BSC543	*FM7h*	30.88 (6.43)	30.91 (7.88)	33.75 (6.67)	32.73 (8.72)	-2.88	-1.82	0.012423	0.689596	0.674376
Df(1)ED7153	*FM7h*	29.80 (6.30)	30.91 (7.88)	30.47 (6.82)	32.73 (8.72)	-0.67	-1.82	0.028176	0.134114	0.603622

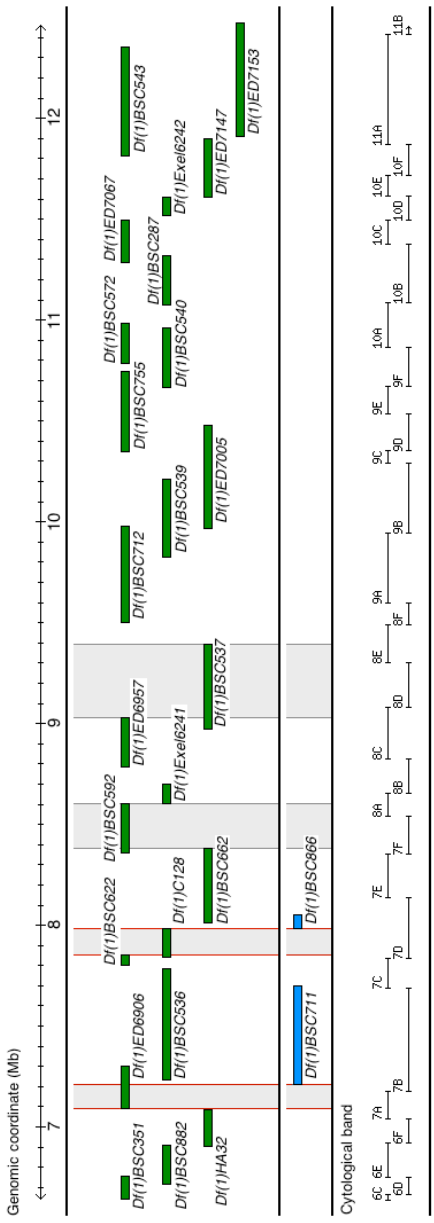


Figure 4. Map of tested deletions within the QTL interval under study.

All deletions, represented by green or blue bars, span a known fraction of X chromosome. Deletion breakpoints at the base pair level are known for all deletions except *Df(1)HA32* and *Df(1)C128*, for which only cytological bands are reported. Both physical and cytological coordinates are provided. The 24 deletions represented in green represent the minimum set spanning the 5.8 Mb QTL interval, deletions in blue were tested upon failure to complement of one of the overlapping deletion in green. Fractions of the QTL interval with light grey shading indicate regions of interest under deletions that show failure to complement. Red borders of this grey background indicate highly significant failure to complement $P<<0.01$

While most of the known QTL for this phenotype in *D. melanogaster* have been found on chromosomes 2 and 3 with a vague suggestion of sex-specific effects of the X chromosome (Morgan & Mackay, 2006; Norry *et al.* 2008) compelling evidence of the existence of X-linked QTL affecting cold tolerance in natural populations has been provided only by Svetec *et al.* (2011). Our set of 89 candidate genes is heterogeneous regarding the types of existing functional annotations and only few of them have a reported association with cold tolerance. We reduced this set of genes to those encompassed by the most highly significant deletion *Df(1)ED9606*. There is also evidence of association with chill coma recovery time for some genes encompassing this deletion (Ayroles *et al.* 2009; Mackay *et al.* 2012).

To test the effect of the candidate genes uncovered by deletion *Df(1)ED9606* on CCRT we conducted single-gene based quantitative complementation tests. We made use of commercially available fly strains with P-element insertions disrupting the expression of three candidate genes; *CG1677*, *unc-119* and *brk*. The three tests, carried out in an analogous way as with the deletions revealed one case of quantitative failure to complement (Table 2). The tested P-element insertion disrupting the expression of gene *brk* showed a significant difference in the CCRT of the in the E* and A* background respect to that of the balancers. The effect excreted by this insertion in is a reduction of CCRT of about 5 minutes in the presence of the European-derived X chromosome.

Table 2. P-element analysis of candidate genes affecting CCRT in female flies.

Target gene	Balancer	Mean CCRT (SD) in minutes				Δ_{mut}	Δ_{bal}	P L	P G	P L x G
		E/mutation*	*E*/balancer*	*A*/mutation*	*A*/balancer*					
brk	*FM7c*	28.18 (8.53)	32.79 (8.59)	32.91 (10.03)	33.04 (9.89)	-4.74	-0.25	0.00189	0.00210	0.00511
CG1677	*FM7a*	29.28 (8.92)	31.32 (9.48)	29.96 (8.73)	32.44 (10.16)	-0.67	-1.12	0.38070	0.02130	0.83404
unc-119	*FM7a*	27.99 (7.83)	31.32 (9.48)	31.78 (9.28)	32.44 (10.16)	-3.79	-1.12	0.01166	0.02542	0.15217

This table summarizes quantitative deficiency tests performed with the listed P-element insertions. Δmut is the difference between the average CCRT of flies bearing E* and A* chromosomes in the presence of a given P-element. Negative differences suggest the presence of CCRT-reducing alleles at the gene affected by the tested P-element. The other symbols are defined in Table 1.

2.2.2 Candidate gene expression analyses coupled to CCRT

We conducted expression analyses for six of the candidate genes under deletion *Df(1)ED9606*. qPCR assays were performed on cDNA prepared from pools of female flies from the Netherlands and Zimbabwe (see Materials and Methods). Expression of candidate genes was measured at two moments after cold stress exposure as well as under control conditions. The two post-cold stress time points were; 10 minutes immediately after the end of cold stress and 15 minutes after flies recovered from chill coma. Controls consisted of flies of the same lines that were not subjected to cold stress

Of the six genes, *CG1958* and *brk* showed significant differences in constitutive expression levels between the Netherlands pool and the Zimbabwean pool ($P<0.01$). In both cases the genes were over-expressed in Zimbabwe. The same general trend was also observed for the other four genes (Figure 5). Average expression level appeared to be unaffected by cold stress within pools at 10 minutes during recovery from chill coma. At this time point, the only highly significant difference between pools was observed at *brk* ($P<0.001$). Expression levels measured at 15 minutes after recovery from chill coma revealed one significant difference within pools: *brk* was significantly over-expressed with respect to controls in the Netherlands pool ($P<0.05$). Between-pool contrasts at 15 minutes after recovery from chill coma revealed only a significant difference for *brk* ($P<0.01$).

Figure 5. Expression assays of candidate genes at interval 7A3-7B1 (next page).

Expression of genes located in the significant fraction under deletion *Df(1)ED6906* was measured under two cold stress and control conditions in pools of flies from a temperate population the Netherlands (NL) and a tropical location Zimbabwe (ZK). Expression level of these candidate genes was normalized respect to that of ribosomal genes *RpS20* and *RpL32*. The height of the bars represents the mean of three calibrated normalized relative quantities (CNRQ) per pool per gene rescaled to that of the corresponding ZK control. Error bars also represent rescaled confidence intervals. Levels of significance for tests of differences in expression levels among treatments within and between populations are indicated with asterisks, P<0.05 (*), P<0.01(**) and P<0.001(***).

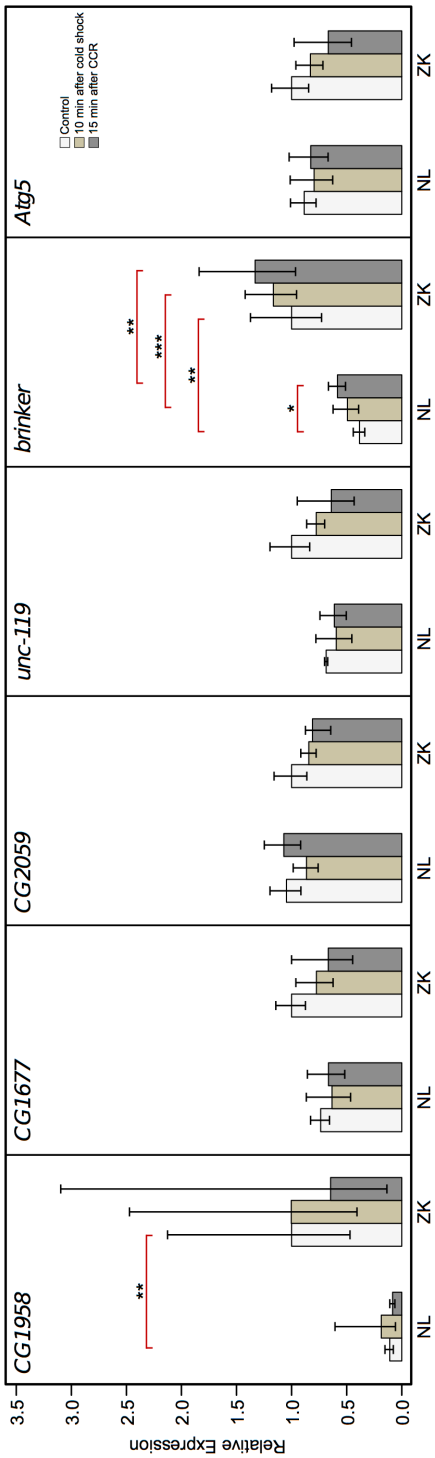


Figure 5. Expression assays of candidate genes at interval 7A3-7B1

2.2.3 Genetic variation at *brk* enhancer region

The results from the quantitative complementation tests and candidate gene expression analyses suggest that *brk* is involved in the CCRT difference between E* and A* lines. This difference in expression levels was also observed between the populations from which the E* and A* X chromosomes were derived. Therefore we suspected that a quantitative trait variant or a QTN was located upstream of *brk*, in its reported enhancer region (Pyrowolakis *et al.* 2004; Yao *et al.* 2008), thereby acting a *cis*-regulatory element of its expression.

We employed the E* and A* lines to sequence the ~16.6 kb encompassing the intergenic region upstream of *brk* (between coordinates 7,185,337 and 7,201,972). The alignment of the two sequences revealed a total of 241 nucleotide differences (see Figure 6A) and 76 structural variants (indels). From these indels 12% were found to be due to loss or gain of short tandem repeats. The other 88% of the variants encompasses single nucleotide indels (40 of 67 cases), followed in number by dinucleotide indels in 6 of the 67 cases. Indels of sizes ranging from three to less than 10 nucleotides make up 21% (14 of 67) of the cases while indels of 10 or more nucleotides add to the remaining 10% (7 of 67). All relative positions (considering the first position of *brk* 5' UTR as point zero) of these differences between the two lines are depicted in Figure 6A-B.

In the absence of annotated transcription enhancers in the 16.6 kb upstream of *brk*, we used community resource-based information to narrow the list of observed nucleotide differences and structural differences between E* and A* to a handful of putative *cis*-elements responsible for the observed *brk* expression pattern and the detected difference in CCRT between the A* and E* lines. First, we looked at reported SNPs associated with CCRT in the *D. melanogaster* genetic reference panel (DGRP). One such SNP, located 3,268 bp upstream of *brk* (*-3,268A/T*) was shown to have a highly significant association with CCRT in the DGRP (Figure 6C). Second, we looked for enrichment of DNA-protein interaction sites in the Encyclopedia of DNA elements (modeENCODE).

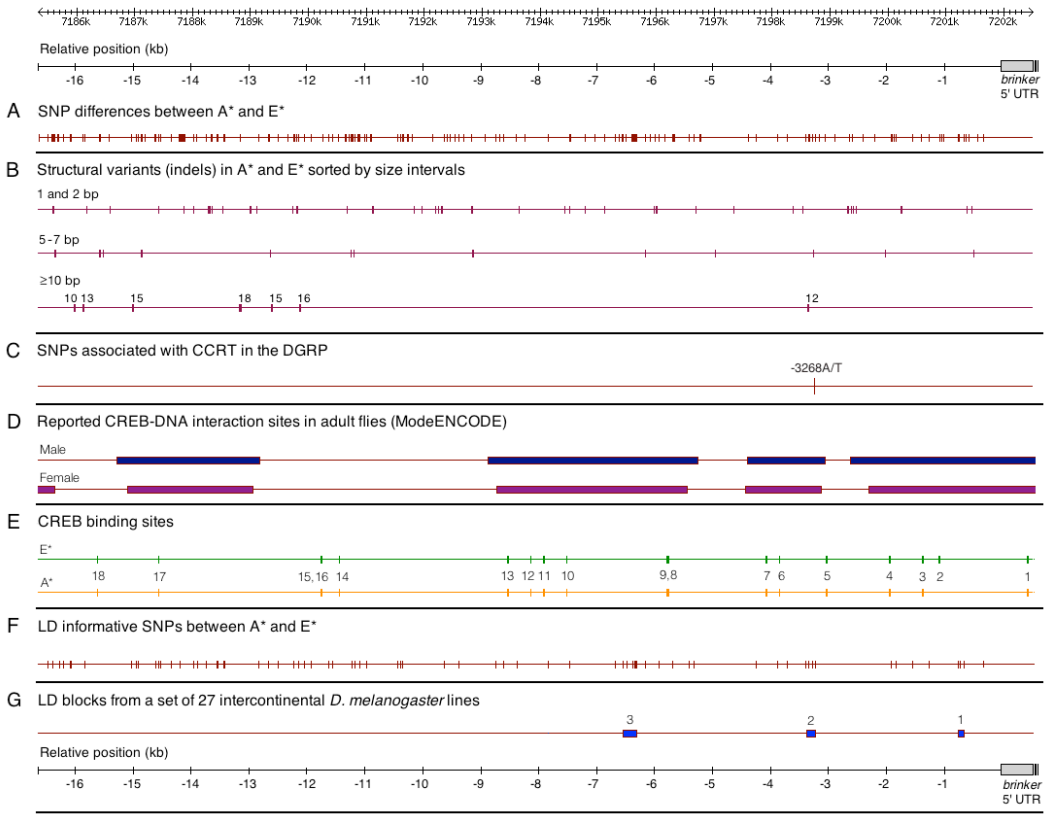


Figure 6. Catalogue of sequence variants along the enhancer region of *brk*.

A) Differences detected between A* and E* sequences. B) Structural variants (indels) from single to dinucleotide indels, 5 to 7 bp indels and higher than 10 bp. C) Location of a SNP associated with chill coma recovery time in a North American *D. melanogaster* population. D) Fractions of the interval (depicted as solid bars) with reported interactions with one transcription factor (CREB) in adult flies of both sexes. E) Predicted binding sites for CREB along the A* and E* sequences. F) Set of 89 SNPs, with contrasting alleles between A* and E* sequence, and segregating with frequency higher than 10% in a sample of 27 European and African lines for which CCRT is known. LD estimates between these pairs of these sites were obtained to identify G) blocks of three or more consecutive sites in with LD higher than 0.8, depicted as blue numerated bars. All sites presented in relative distance respect to the transcription origin of brinker as point zero.

This search revealed that the 9 kb upstream of *brk* are enriched for DNA-protein interaction sites as reported by chromatin immunoprecipitation-on-chip (ChIP-on-chip) assays on adult fly material using CREB-binding protein (CBP) antibodies to precipitate the DNA-protein complexes (Figure 6D). CBP is a known transcription co-regulator encoded in *Drosophila* by the gene *nejire*. The corresponding protein CBP (or Nejire) is recruited by CREB. The resulting protein complex regulates the transcription of downstream genes. This piece of information prompted us to look for differences in CREB binding motifs between the E* and A* sequences. Using a position-weight-matrix approach we predicted a total of 17 transcription factor binding sites (TFBS) for CREB in the 16.6 kb upstream of *brk* plus one in its 5' UTR in the E* sequence, while in the A* sequence only 16 CREB binding sites were found upstream of *brk* and one in its 5' UTR. The locations of the predicted TFBS are shown in Figure 6E. The presence/absence of TFBS No. 2 (between relative positions -1,225 and -1,230) represents the only difference between the two sequences.

In addition to the community-resource search for putative TFBS upstream of *brk*, we conducted a linkage disequilibrium (LD) assisted search for sets of three or more consecutive SNPs forming haplotypes with presumed functional roles. Following De Luca *et al.* (2003) and Clowers *et al.* (2010), we calculated pair-wise LD tests between SNPs from an alignment of next-generation sequence data of 27 *D. melanogaster* lines of European and African origin that include the parents of the E* and A* strains for which we previously scored CCRT. After excluding sites with minor allele frequency of 10% (or less) as well as sites with less than 50% data, we retained a total 89 LD informative SNPs (Figure 6F). The resulting LD matrix (Appendix B) depicts association levels around 0.5 between sites scattered along the 16.6 kb.

It was not surprising to observe this pattern of LD since the employed set of 27 lines was not sampled from the same population instead they constitute a pool of African and European lines (in similar numbers). These two populations have been previously shown to have demographic histories that have led to SNP allele frequency differentiation. Moreover, it has also been reported that the European fraction of the pool exhibits a reduction in genetic polymorphism consistent with a footprint of positive selection. Hence this pooling strategy led to an artificial creation of LD. To overcome this limitation, we defined an LD threshold of $r^2 \geq 0.8$ for the sets of at least three consecutive SNPs to define LD blocks of interest. We prioritized all reported differences between E* and A* based

on the outcomes of the previous approaches. We defined the top two relevant fragments upstream of *brk*. These correspond to relative positions -784 to -1,243 (including the CREB binding site difference and LD block 1) and positions -3,016 to -3,553 (including the CCRT associated SNP *-3,268A/T* and LD block 2) (Figure 6G).

We re-sequenced these two fragments in European and Southeast African inbred lines. This set of lines included those previously used to assay candidate gene expression. Upon inspection of the alignments of the 460-bp fragment between relative positions -784 to -1,243 (Figure 7A), we found that the predicted CREB binding motif "GACGT" is present in 90% of the European lines, while in the Southeast African lines it occurs in 33% of the cases. Absence of this predicted binding sites in Africa could be due to two mutational events: allele variants at SNP -1,226 or a 73-bp long deletion encompassing relative position -1,230 to -1,158. This structural variant is in 20% in the Zimbabwean sample. There is no evidence, however, that the absence or presence of this motif causes the pattern of *brk* expression observed in the wild type lines. Intuitively, the presence of an extra TFBS for CREB in Europe should lead to higher expression of *brk*, but expression is higher in the Zimbabwean pool.

Next we turned our attention to the 534-bp fragment between positions -3,000 to -3,553 (Figure 7B). This fragment depicts a pattern of SNPs and deletions more complex than that revealed by next-generation sequence data. In this fragment, the SNPs belonging to LD block 2, *-3,503T/C*, *-3,479G/C*, are in LD with two adjacent 6-bp long indels at relative positions -3,474 and -3457. These two deletions correspond to the 12-bp indel initially observed when comparing the sequences of A* and E* lines only. The other two SNPs of LD block 2, *-3,409T/C* and *-3,355T/C*, show a clear association in the European lines with the 7-bp deletion at position -3,298, also observed when comparing the A* and E* lines. Another important finding of this re-sequencing approach is that a low frequency 8-bp indel at position -3,271 encompasses the previously reported SNP associated with CCRT at position -3,268 (Mackay *et al.* 2012). However, our data indicate that this polymorphism corresponds to an indel and not to a SNP.

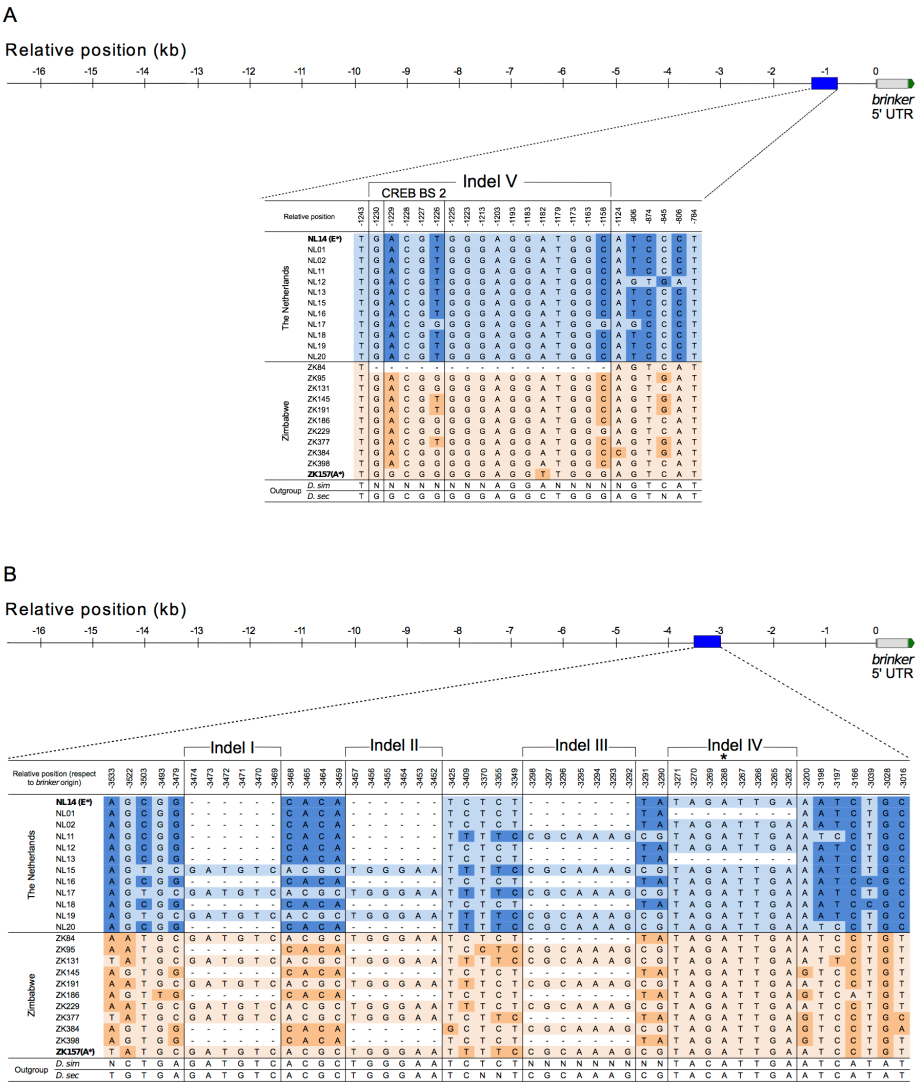


Figure 7. Polymorphism tables of chosen fragments upstream of *brk*.

A) Between relative positions -784 and -1,243 from the origin of *brk*'s 5' UTR a fragment of interest was resequenced in lines from the Netherlands (NL) and Zimbabwe (ZK). The table depicts nucleotide (SNP) and structural variants (indels) of two *D. melanogaster* population, including E* (top line) and A* (bottom line) plus two out-groups (*D. simulans* and *D. sechellia*). Light blue and orange indicate the inferred ancestral state of the SNP considering the two outgroups in NL and ZK respectively, whereas darker tones of the same color represent the derived allele. Deletions are indicated in white background. Note the presence of the CREB binding site (BS) motif "GACGT" in NL, which is less frequent in ZK due to mutation or loss because of deletion (Indel V). The acquisition of this motif occurred in the *melanogaster* lineage based on its absence in the out-group species. B) Polymorphism table of a 534-bp fragment between relative positions -3,000 to -3,553 upstream of *brk*. As in panel A, the table depicts SNPs and structural variants (indels) in NL and ZK, including E* (top line) and A* (bottom line) plus two out-groups (*D. simulans* and *D. sechellia*). Deletions are indicated in white background. Relative position -3,268 marked with an asterisk is highly associated with CCRT in the Raleigh population.

2.2.4 Frequency shift of variants likely associated with CCRT

We rely on experimental evidence obtained by Yao and colleagues (2008) that several elements upstream of *brk* affect its expression levels in developmental stages. However, we should make clear that we ignore whether this is also the case for adult flies (and under stress conditions). For the sake of our investigation, it is important to underline a point originated from the observation of the polymorphism tables in Figure 8A. The frequencies with which these deletions (and associated SNPs) occur in the Netherlands and Zimbabwe suggest a pattern of intercontinental differentiation.

We studied the frequencies of the haplotypes defined solely by the number of deletions (Figure 8B) and observed that the haplotype group represented by the E* line (ND haplotype) encompasses three deletions and is found in 60% of the Dutch flies, while it is seen in less than 30% in the Zimbabwean (SEA) flies. Moreover, this ND haplotype exhibited a frequency shift across continents (Figure 8C). The fact that the frequencies of this haplotype are also correlated with the latitude (an important geographical variable that is in turn associated with climate) is a strong motivation to further evaluate the effect of these structural variants on CCRT. In fact, we initiated the study of the potential effect of the ND haplotype on both *brk* expression levels and CCRT with a simple qPCR assay using lines from Europe and Africa (see Appendix C). The results were not conclusive enough to substantiate (or rule out) the claim of involvement of the haplotype on the two traits of interest. However, this is a first step in the examination of *brk* as one of the QTGs underlying the CCRT and whether this detected allele frequency shift is a genuine footprint of polygenic adaptation.

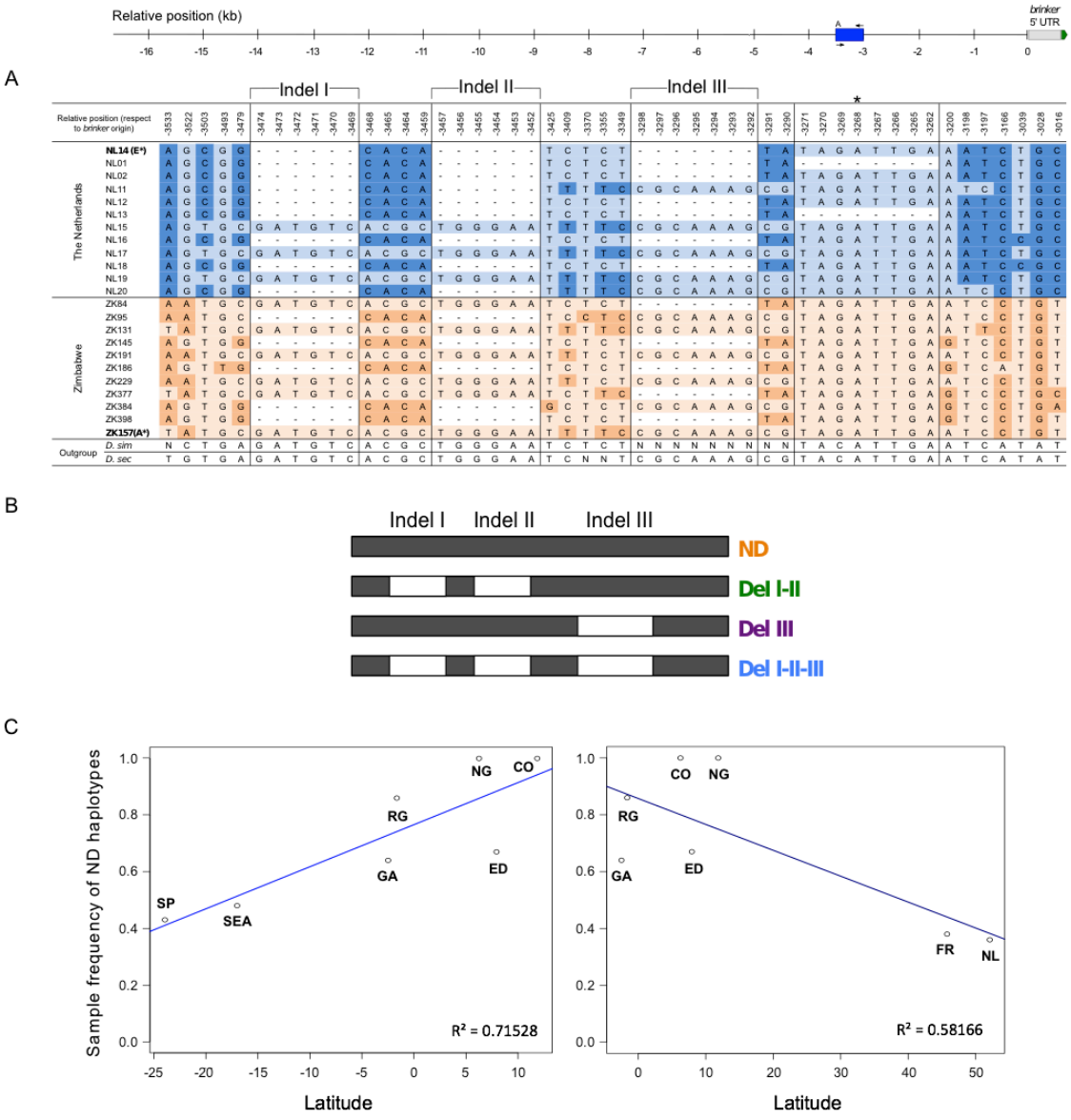


Figure 8. Allele frequency shift at a putative *cis*-regulatory element for *brk*.

A) Four haplotypes defined by the presence/absence of deletions and their numbers in the fragment. B) Frequencies of the haplotypes at the extremes along a latitudinal gradient of *D. melanogaster* populations; EUR is a pool of NL and FR (French) lines, RG denotes Rwandan lines, and SEA the groups ZK and ZI (Zambian) lines.

2.3 POPULATION GENETICS REVISITED

2.3.1 Patterns of variation under deletion *Df(1)ED6906*

The results of the QTL fine mapping strategy in the preceding section allowed us to define small chromosomal intervals that containing tractable numbers of candidate QTNs affecting CCRT. One of these fragments was uncovered by deletion *Df(1)ED6906*. Having deified this region free of population genetics biases we then investigated its pattern of genetic variation using a comprehensive array of *D. melanogaster* SNPs. We were particularly interested in identifying conspicuous signals of non-neutral evolution. For this purpose we calculated a set of summary statistics on a 2-kb, non-overlapping window basis using next-generation sequence data from the two European (the Netherlands and France) and two African (Rwanda and Southeast Africa) populations. The Netherlands population and a set of Southeast African lines represent the gene pools from which the E* and A* lines were derived. The additional two populations consisted of French and Rwanda sequence data from the DPGP (Pool *et al.* 2012). These four populations allowed us to draw conclusions about the patterns of variation in temperate and tropical populations.

For each population we obtained nucleotide diversity estimates measured by the average number of pair-wise differences (θ_π) and Watterson's estimator (θ_W). Both European populations showed a three to four-fold reduction in nucleotide diversity with respect to the African populations. For instance, average θ_π (± SD) were 0.0010 (± 0.0007) and 0.0008 (± 0.0007) for the Netherlands and France, respectively, while the values for Rwanda and the Southeast African pool were 0.0031 (± 0.0011) and 0.0034 (± 0.0011), respectively. The averages of the two genetic variability estimates are shown in Table 3. The entire profile of variation can be observed in Figure 9. This figure depicts the values of the 2-kb windows along the region of 124 kb in the four populations. Interestingly the plots for the Netherlands and France reveal a 40-kb long fragment with θ_π values as low as 1 SD from their respective averages over the entire 124-kb region. This low polymorphism region spans the windows from 66 to 106 kb. This pattern is in contrast to that observed in the two African populations for the same coordinates, for which nucleotide diversity values tend to be above their respective population averages.

This profile of genetic variability prompted us to study the pattern of population differentiation and divergence. The average population differentiation, as revealed by F_{ST} (Nei and Li 1979), clearly reflected the continental groupings (Table 4). The lowest F_{ST} values were reported between the Netherlands and France, as well as between Rwanda and the Southeast African pool. The behavior of window-based F_{ST} estimates between each of the European populations and the Southeast African pool revealed a trend to group high F_{ST} values along the fragment with the lowest values of genetic diversity in the European populations (between relative positions 95,000 and 109,000; Figure 9). In addition, the values for genetic divergence with respect to *D. simulans* are shown in Table 3 and plotted in Appendix D.

Table 3. Summary statistics at 7A3-7B1 in four *D. melanogaster* populations.

Population	Diversity estimators (mean ±SD)		D_{xy} (mean ±SD)	Tajima's *D* (mean ±SD)
	θ_π	θ_W		
NL	0.0010 ±0.0007	0.0010 ±0.0006	0.0953 ±0.0352	-0.4995 ±1.0403
FR	0.0008 ±0.0007	0.0008 ±0.0006	0.1042 ±0.0382	-0.1010 ±0.8699
RG	0.0031 ±0.0011	0.0041 ±0.0013	0.1286 ±0.0422	-0.9671 ±0.4024
SEA	0.0034 ±0.0011	0.0042 ±0.0012	0.1248 ±0.0482	-0.7834 ±0.3875

Table 4. F_{ST} at 7A3-7B1.

	F_{ST} (mean ±SD)		
	FR	RG	SEA
NL	0.1180 ±0.1446	0.3148 ±0.1425	0.3223 ±0.1402
FR		0.2733 ±0.1346	0.2676 ±0.1420
RG			0.0998 ±0.0929

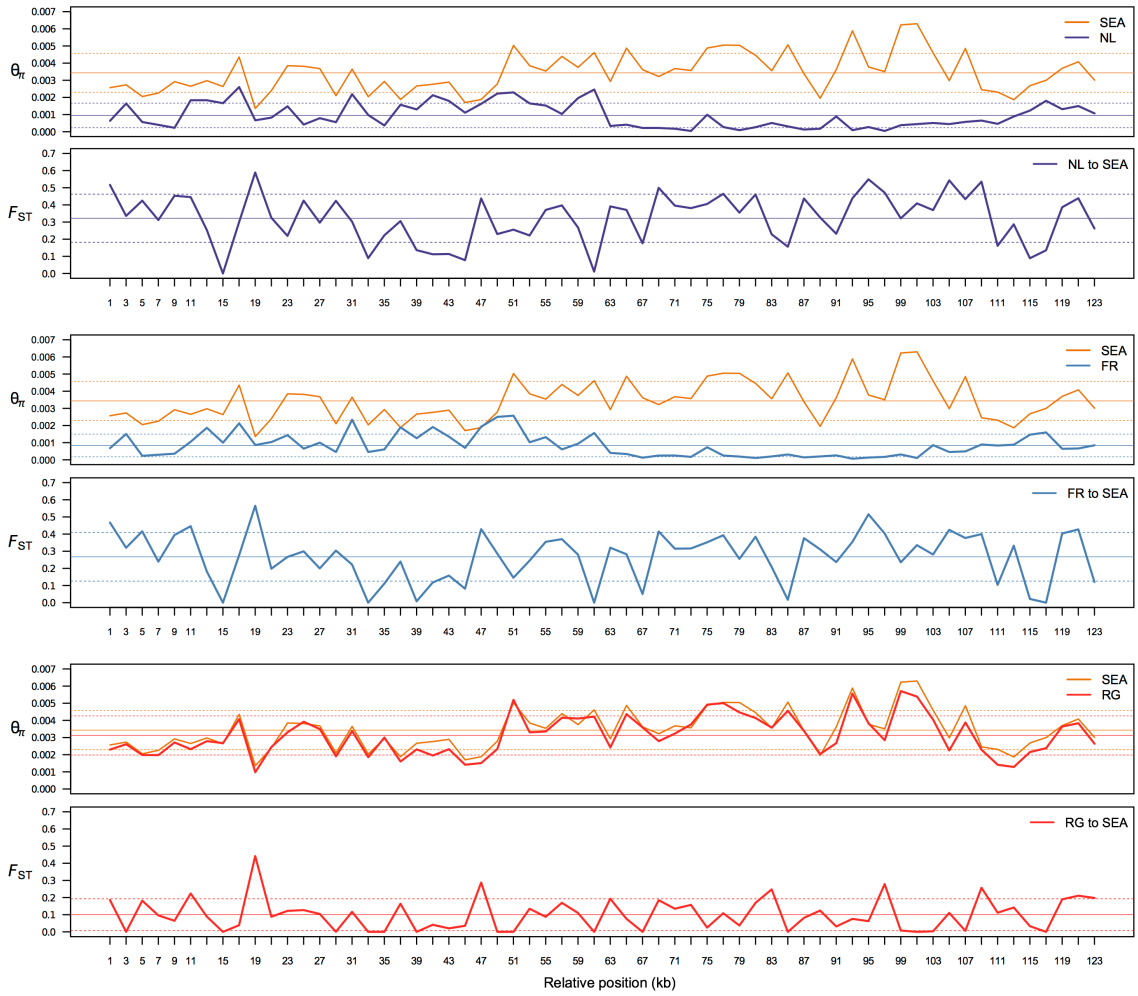


Figure 9. Polymorphism and between-population differentiation at 7A3-7B1.

Nucleotide diversity (θπ) was obtained for consecutive 2-kb long windows in four different populations: the Netherlands (NL), France (FR), Rwanda (RG) and a pool of Southeast African (SEA) lines sampled around Lake Kariba in Zimbawe and Zambia. The SEA profile is shown in all three θπ panels for sake of comparison, because SEA is thought to be the ancestral range of *D. melanogaster*. Below each θπ panel, inter population differentiation profiles are shown. Differentiation (F_{ST}) was calculated as normalized distance of Nei. Thin, continuous lines represent the average value for each summary statistic across the 62 windows, dashed lines represent 1 SD above and below the corresponding summary statistic mean.

2.3.2 Testing neutrality

Thus far, our findings have suggested the effect of evolutionary forces leading to a conspicuous reduction in genetic variability along the chromosomal interval at 7A3-7B1 in the two European populations. This feature was accompanied by enrichment, within the same interval, for fragments with above-average genetic differentiation between temperate and tropical (African) populations. In the light of these observations it is important to establish whether this pattern of variation in Europe was created by positive selection or is a consequence neutral process such as genetic drift or the complex demographic history of non-African *D. melanogaster* flies (Laurent *et al.* 2011). The first step we made to clear this question was to explore the relationship between the values of θ_π, and θ_W, as measured by Tajima's *D* statistic (Tajima 1989). Under neutral evolution, the quantities obtained for these two estimators should be statistically the same. If differences are detected these can be attributed to non-neutral processes such as demography, selection or the combination of the two.

Tajima's *D* values (shown in Table 3 and plotted in Figure 10) were overall negative in the four populations. Among the four populations, France showed the least negative average *D* value, while Rwanda exhibited the top negative average value (Table 3). We did not explicitly tested whether the window-based *D* values per population were statistically different from zero; instead we used the distribution of obtained values to identify outlier windows above and below 1 SD of the reported mean. A total of 4 windows (at relative positions 35 kb, 77-79 kb and 95 kb) showed outlier values below 1 SD in both European populations. Above 1 SD of the respective means, another four windows were common outliers to the Netherlands and France (located at 37 kb, 53 kb, 87 kb and 109 kb). Figure 10 also shows, for the two European populations, a stretch of below-average *D* values that extends for ~40 kb (windows 66-106 kb) interrupted by a positive peak at 87 kb. In contrast, the two African populations show a less variable Tajima's *D* pattern. The top negative peaks shared by both populations are centered on are positions 19-89 kb, 91 - 109 kb and 113 kb. The highest Tajima's *D* values (close to zero) are seen from windows at 51 to 65 kb.

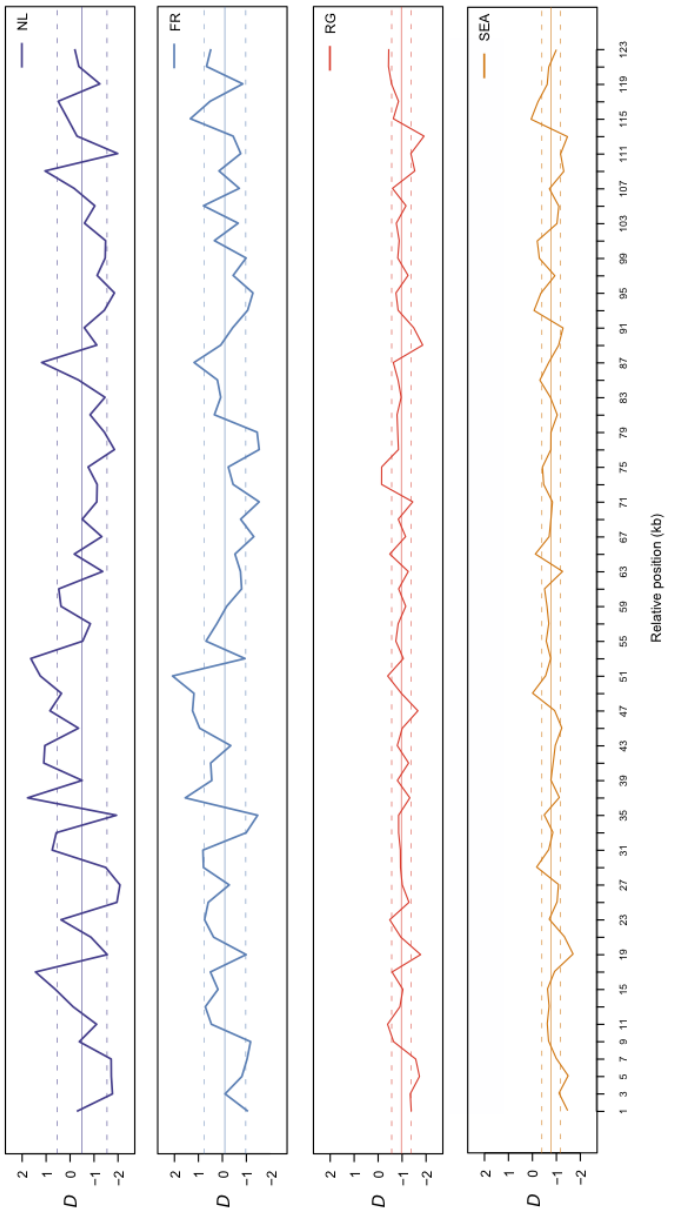


Figure 10. Tajima's *D* at cytological region 7A3-7B1.

Tajima's *D* (*D*) profiles, along the 124 kb of interest are shown in paired panels for each scrutinized population: the Netherlands (NL), France (FR), Rwanda (RG), and Southeast Africa (SEA). Thin, continuous lines represent the average value for each summary statistic across the 62 windows, dashed lines represent 1 SD above and below the corresponding summary statistic mean.

2.3.3 CLR and F_{ST} scans for positive selection

The patterns of polymorphism observed in the region of interest in the European and African populations revealed a conspicuous reduction of genetic variability and a negative Tajima's *D* in both European populations that extends for approximately 40 kb in the 124-kb region. This reduction has been already identified as a footprint of positive selection in non-African populations (Glinka *et al.* 2006; Langley *et al.* 2012). In this work, motivated by the link with a QTL for chill coma recovery, we conducted an exhaustive analysis of this region to further document the historical footprint left by positive selection in this chromosomal region.

We studied the site frequency spectrum (SFS) of the region for the available European sample (pooling the Netherlands and French lines) and subjected the SNP dataset to the composite likelihood ratio (CLR) test implemented in the program SweeD (Pavlidis *et al.* 2010). This likelihood ratio was computed between a selective sweep model and a neutral model that is calibrated with the genomic background frequency spectrum. The background SFS was obtained from 20 Mb of the X chromosome, excluding the telomere and centromere regions (see Materials and Methods). In our region of interest the fragment between relative positions 63,000 and 107,000 exhibits a SFS that is in contrast to that of the genomic background and is better described by a selective sweep model (Figure 11A). The CLR values obtained for this interval (Λ_{CLR}>300) are above the significance threshold of 72 that corresponds to the 95th quantile of the top CLR values of 100 simulated sub-genomic regions of 5 Mb. This value did not increase when larger genomic regions were simulated (Appendix E). Simulations were based on our current understanding of the demographic history of European populations (Laurent *et al.* 2011). Furthermore the observed CLR peak falls within the top 1% of CLR values along the entire X chromosome (Appendix F)

A remarkable aspect of the fragment with the highest CLR values is its absence of variation in the coding regions of genes *CG1958*, *CG1677*, *CG5059*, and *unc-119* (see their location in Figure 11). However, this feature greatly limited the power of the test in identifying targets of selection. To circumvent this problem, we explored another feature of positive directional selection, namely that this type of selection may cause allele frequency shifts at target loci in structured populations.

Therefore measures of population differentiation (F_{ST}) can be used to look for targets of selection in genome-wide SNP datasets. We obtained model-based F_{ST} coefficients for each SNP within the QTL of interest in an intercontinental group of *D. melanogaster* populations (see Materials and Methods). We considered SNP data from seven populations along a south-north gradient across Africa and Europe: South Africa, Southeast Africa, Rwanda, Cameroon, Ethiopia, France, and the Netherlands. Using BayeScan (Foll and Gaggiotti 2008), we obtained F_{ST} values from a total of 7,316 SNPs with and an average F_{ST} of 0.2621 and revealed four outlier SNPs that showed the highest differentiation across populations at a FDR of 5% (Figure 12B). These four SNPs are located within the 40-kb long fragment enriched for SNPs showing significant CLR values between positions 65,000 and 105,000. The 65-kb and 19-kb long flanking regions to the left and right of the 40 kb fragment are enriched for SNPs showing below- average F_{ST} values (Figure 12B). However, none of these SNPs with low differentiation across populations is significant at the 5% FDR.

Figure 11. Evidence of positive selection at 7A3-7B1 (next page).

A) Likelihood (CLR) profile along the 124 kb at 7A3-7B1 using SNP data of two pooled European *D. melanonagster* from the Netherlands and France. Two significance thresholds are depicted. The solid line corresponds to the average of the top 1% CLR values for the X chromosome in Europe and the dashed red line represents the significance threshold from simulations of equivalent sub-genomic regions. Note that the CLR profile does not overlap with the putative *cis*-acting element upstream of *brk* that is likely to cause expression and cold tolerance differences between tropical and temperate populations and was described in section 2.2. This element encompasses relative positions 109,442 to 109,976. B and C) Model-based F_{ST} values for 7,364 SNPs from a 6-population dataset: the Netherlands and France as one single population, Ethiopia, Cameroon, Rwanda, Southeast Africa and South Africa. The top SNPs above the false discovery rate of 5% are indicated and constitute candidate positions for directional selection across the intercontinental dataset. D) Model-based F_{ST} coefficients for 7,095 SNPs from a 5-population dataset where only the African populations in panel A were considered. For both panels the dashed line corresponds to the F_{ST} value of a false discovery rate (FDR) of 5%.

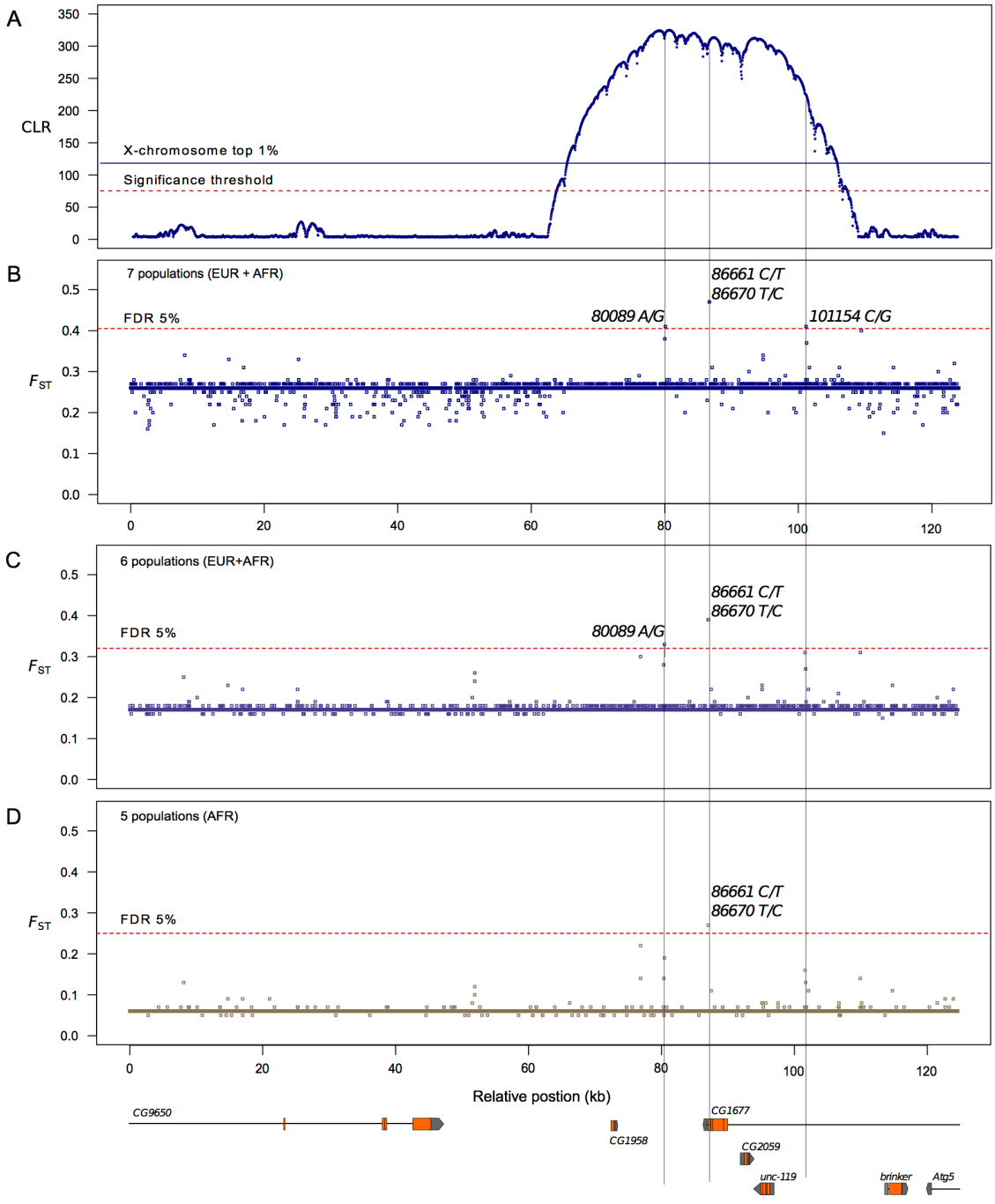


Figure 11. Evidence of positive selection at 7A3-7B1.

The demographic models of the populations used in the analysis could have an effect on the results. For instance, the fact that the two European populations were derived from the same ancestral bottleneck constitutes a violation of BayeScan's demographic model. Therefore, we ran Bayescan on the same dataset treating the two European populations as one. The results show a substantial decrease in the number sites with below-average F_{ST} values but the same outliers of high differentiation remained (Figure 11C). Interestingly, the exclusion of European populations from the analysis did not change the pattern of highly differentiated, outlier SNPs (Figure 11D). This suggests that the phenomenon of allele frequency differentiation at candidate SNPs had already started within the African continent.

Among the outlier SNPs that show high differentiation across the entire intercontinental dataset, the top ones are *86,661C/T* (F_{ST}=0.4697, q-value=0.0024) and *86,670T/C* (F_{ST}=0.4653, q-value=0.0042). These two non-synonymous SNPs are located in exon 5 of the computationally predicted gene *CG1677* and show alleles in perfect LD (Figure 12A). The TT haplotype (*86,661T – 86,670T)* is in high frequency in the Southeast African samples and intermediate in Rwanda; its frequency decreases with increasing latitude to be replaced in the European populations by the CC haplotype (Figure 12B). Both SNPs predict changes in the amino acid sequence of the protein. The common Southeast African form of the protein codes for a threonine (Thr) and an asparagine (Asn) at residues 936 and 939, while the cosmopolitan form has an alanine (Ala) and aspartic acid (Asp) at these two positions.

The third highly significant SNP is *80,089A/G* (F_{ST}=0.4145, q-value=0.0313) located between genes *CG1958* and *CG1677*. It is an outlier because of the contrasting differences in allele frequencies between Southeast African populations and West Africa (Cameroon) (Figures 12B). The putatively ancestral allele is most frequent in South Africa, Ethiopia and Europe, intermediate in Rwanda, and in Congo the derived allele G is fixed. Our fourth top SNP *101,154C/G* (F_{ST}=0.4067, q-value=0.0480) that is located 5 kb upstream of gene *unc-119* shows allele frequency changes following a south to north gradient (Figures 12B and 13B). The Southeast African populations exhibit the derived variant G in low frequency while in Rwanda this allele has higher frequency, but is still considered low to intermediate. In western Africa and Ethiopia the variant was fixed. Non-African lines are also fixed for allele G at this position. There is no evidence for LD of this SNP with adjacent SNPs.

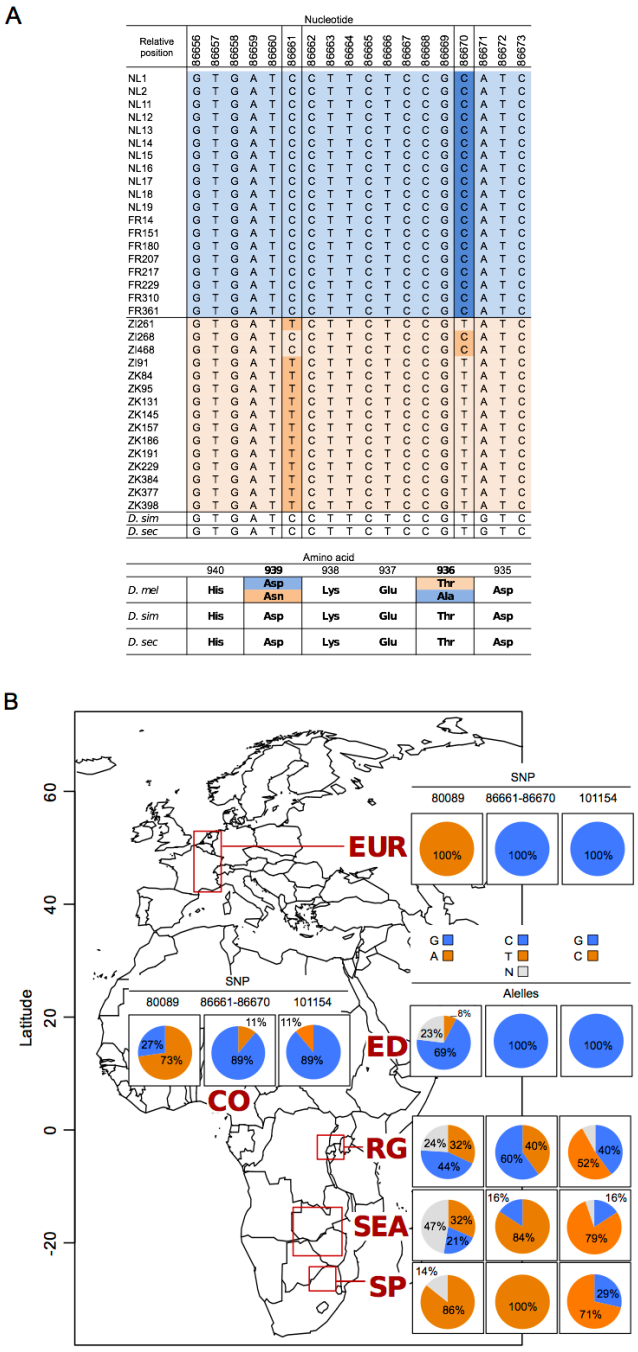


Figure 12. Allele frequency change at highly differentiated SNPs at 7A3-7B1.

A) European and Southeast African *D. melanogaster* haplotypes for the two non-synonymous SNPs (86,661-86,670) in intron 5 of gene *CG1677*. These two SNPs correspond to amino acid positions 939 and 936. B) Allele frequencies of the four top differentiated SNPs across seven different populations along a latitudinal gradient. Populations are as follows: the Netherlands and France (EUR), Ethiopia (ED), Cameroon (CO), Rwanda (RG), Southeast Africa (SEA), and South Africa (SP) (see Materials and methods).

2.3.4 A likely case of compensatory evolution at *CG1677*

Insights into the evolution of the four significantly differentiated polymorphisms described above may be inferred from the two coding SNPs that cause the amino acid differences between the African and cosmopolitan versions of peptide CG1677. Two aspects are especially interesting about this discovery. First, the close proximity of the two involved amino acid positions (936 and 939) and second the pattern with which the amino acid combinations are observed across continents. In Southeast Africa both combinations Thr-Asn and Ala-Asp, are present at sites 936 and 939 respectively, where the former is more common and no other combinations exist. In Europe, however, Ala-Asp is fixed (Figure 12A). These facts could reflect a compensatory evolution scenario. However, it is difficult to study in detail this hypothesis in the absence of tertiary structure of protein CG1677. After subjecting the primary sequence of the protein, in its two versions (E14 and ZK157), to a structure prediction program (Kelley and Sternberg 2009), we observed that both amino acid positions are part of an α-helix; *i.e.*, they are located on neighboring helix turns and can therefore interact. Interestingly, Thr and Asn can form one hydrogen bond between their side-chains more than Ala-Asp. Because of a homology search at Uniprot (Uniprot-Consortium 2014) we could also establish that residues 415 to 450 likely comprise its zinc finger domain, presumably the active site of this protein.

This likely case of compensatory evolution could explain the observed selective sweep. It is interesting to reconstruct the mutational events at the codons corresponding to residues 936 and 939. First, we looked at patterns of background variation of the TT and CC haplotypes in the Sub-Saharan African populations where these are found. We observed that variation in the TT background is higher, which attests for its ancestral status within the *D. melanogaster* lineage. However, the presence of a CT haplotype in the out-groups suggests that a transition C → T at position 86,661 occurred early in *D. melanogaster*. Subsequently the reverse transition (T → C) occurred at 86,670, which in turn was compensated by back mutation (T → C) at position 86,661. This succession of point mutations may have created the CC haplotype currently observed in cosmopolitan *D. melanogaster*. We support this conclusion based on an ancestral state reconstruction analysis conducted with all available Drosophilid sequences of the gene *CG1677* (data not shown).

2.4 SWEDISH FLIES

2.4.1 The edge of the *D. melanogaster* habitat range

For several *Drosophila* species habitat range borders usually entail high latitudinal and altitudinal locations with heterogeneous climate conditions (Kellerman *et al.* 2009). Climate conditions such as temperature experience daily and seasonal fluctuations, which become more pronounced as latitude increases. The low temperatures that characterize high latitude locations are stressful for *D. melanogaster* and therefore set the edge of its habitat range. Both, cold stress tolerance and reproductive diapause are traits associated with overwintering behavior that have been well documented in flies from high latitudes (Izquierdo 1991; Goto *et al.* 1999; Hoffmann *et al.* 2003a). These traits have evolved as adaptations in these particular environments (Hoffmann *et al.* 2003a; Ayrinhac *et al.* 2004; Kimura 2004). While some researchers maintain that edge populations are ideal systems to study adaptation to novel habitats (Hoffmann and Blows 1994), others have suggested that the demographic conditions of these edge populations may not facilitate the operation of positive directional selection (Kawecki 2008; Sexton *et al.* 2009).

Populations in marginal habitats have lower population densities relative to core habitats and are more fractionated in space (Vucetich and Waite 2003). Furthermore, edge populations experience high connectivity with core populations characterized by a highly asymmetrical dispersal (Kawecki 2008). Because core populations produce more individuals these will be more likely to migrate towards edge populations than the other way round. With small population sizes the destiny of beneficial mutations, if they get to occur *in situ*, is primarily dictated by genetic drift. Beneficial mutations can also arrive from core populations. Even then, their benefit has to be high enough to escape loss to drift. The general view is that the majority of mutations brought by migrants from core populations will be maladaptive in edge environments (Kirkpatrick and Barton 1997; Kawecki 2008). However, this detrimental effect of migration is only expected when there is a steep transition of environmental conditions between core and edge locations. If the change of ecological conditions across habitats is smoother, migration from core populations to the edges may bring a higher proportion of alleles beneficial in edge localities (Kirkpatrick and Barton 1997). Likewise, further inclusion of ecological dynamics at habitat edges, such as interspecific competition, may increase the chances of

adaptive evolution (see Sexton e*t al.* 2009 for a review). Although edge populations may be regarded as genetically depauperate (due to their low levels of genetic variation) they may still respond to selection on ecological traits (see selection on desiccation resistance in *Drosophila serrata*, Blows and Hoffmann 1993). This implies that the adaptive potential of edge population may be unaffected by their demographic situation (Willi *et al.* 2006).

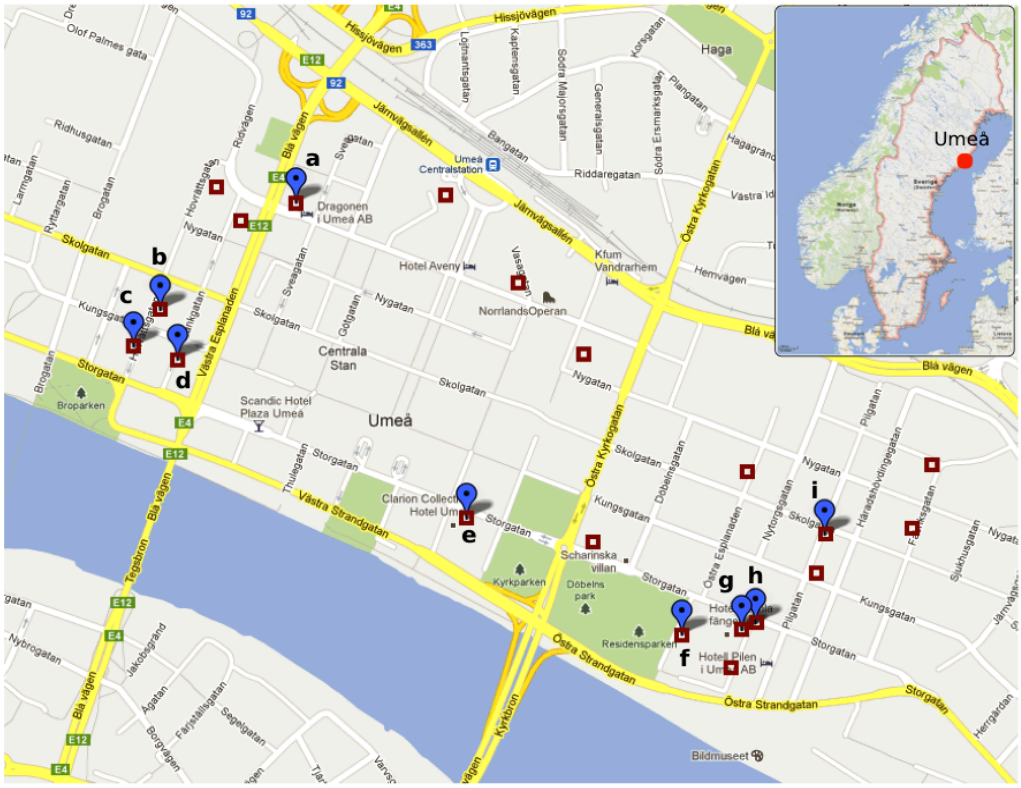
The latitudinal edges of *D. melanogaster*'s current distribution are not known with geographic precision. However, these are expected to be variable and strongly dependent on both, insect own dispersal capacity and the speed with which global warming creates favorable conditions for successful northward and southward colonization of new territories (Régnière *et al.* 2012; Rodríguez-Trelles *et al.* 2013). A fairly good approximation of how far *D. melanogaster* has gone and of where edge populations are located can be derived from reports on natural populations worldwide. The east coasts of Australia and Tasmania are by far the best-sampled locations with a latitudinal extreme reported as far as 43º S (Hoffmann *et al.* 2002). In the Americas the reported southernmost population was in central-south Argentina at 38º S (Fallis *et al.* 2011), while in North America, flies have been sampled at latitudes between 44º and 45º N (Capy *et al.* 1993; Schmidt and Paaby 2008). In Europe, *D. melanogaster* has been sampled in Scandinavia at latitudes between 55º and 60º N (Hale and Singh 1991). Moreover, natural populations have been reported in the warmest periods of the year at higher latitudes in Sweden and Finland (Bächli *et al.* 2005); A. Saura personal communication).

As a first step to conduct studies aimed at understanding the evolutionary dynamics that take place in edge populations, particularly whether they have successfully adapted to local conditions, we sampled *D. melanogaster* in the urban area of Umeå (in northern Sweden). These flies were parental to a collection of 80 inbred lines, currently kept in our laboratory. This section is meant to introduce this Umeå collection, reporting on aspects of the collection and the generation of sequence data at a genome-wide scale. While a thorough description of genome variability patterns in this population is beyond the intended scope for this section, we revisited chromosomal region 7A3-7B1 (of interest in section 2.3) and provide a glance of how variation appears at this QTL region in the Swedish population.

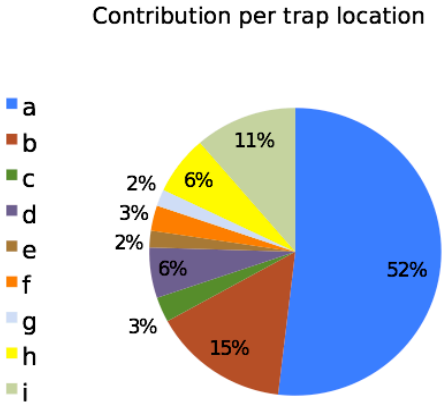
2.4.2 The Umeå collection and its 19 genomes

Towards the end of the warm season in the second half of August 2012 we collected a total of 106 gravid females in the lapse of one week in the urban area of Umeå (63° 49' N, 20° 15' E). The exact sampling places are indicated in Figure 13A. These trap places were primarily chosen because of their likelihood of being visited by flies as well as by the intention of evenly sampling Umeå's city center. Sampling on a single spot would increase the chance of bias towards few fly families, which could lead to wrong conclusions upon genetic variability analyses. Trap site yield was highly heterogeneous across the entire sampling area. We obtained flies from 9 of the 20 trap sites and observed that not all sites contributed equal numbers of flies to the whole collection. For instance, 78% of the final number of female flies came from 3 of them (Figure 13A,B). In situ inspection of trapped flies revealed that only flies of *D. melanogaster* or *D. simulans* were collected. The next step after collection was the species diagnosis of the collected females.

A



B



C

Sequenced lines per location

a Västra Norrlandsgatan 5 (HD)	SU 46 SU 58 SU 81 SU 93
b Hovrättsgatan (covered parking lot)	SU 5 SU 7 SU 8 SU 11
c Hovrättsgatan 1	SU 37
d Bankgatan 6 (open parking lot)	SU 75
e Storgatan 52 (HC)	SU 96
f Storgatan 60 (Goeteborg R)	SU 21 SU 77
g Storgatan 62 (Droskan R)	SU 94
h Storgatan 64	SU 20 SU 23
i Pilgatan 14(Bookshop)	SU 2 SU 25 SU 26 SU 29

Figure 13. Details of Umeå fly sampling and sequencing project.

A) Location of traps in Umeå city center (marked with brown squares). Traps that flies were found are marked with an additional blue pin. B) A total of 106 female flies were collected by the end of the week. The pie chart shows the fraction of flies contributed by each sampling location. C) List of *D. melanogaster* lines from the Umeå collection that were sequenced. The list is sorted by collection site.

Using a molecular species diagnosis protocol we determined that the total of 96 surviving females brought from Sweden belong to the species *D. melanogaster*. We confirmed this diagnosis by conducting a batch of crosses with *D. simulans* males and monitoring offspring viability (see Materials and Methods). Subsequently we initiated a full sibling inbreeding protocol that we maintained for 10 generations to obtain a total 80 inbred lines. These lines are currently maintained in our stock collection. All inbred lines were given names consisting of the abbreviation "SU" that stands for Sweden and Umeå, followed a given line number.

This Umeå collection is to our knowledge the largest set of wild type inbred lines from a high latitude location. The phenotypic characterization of these lines for phenotypes relevant to high latitude locations such us cold stress tolerance was already initiated in our laboratory. Swedish flies have showed on average the shortest recovery times from chill induced coma (K. von Heckel, unpublished results). The goal, however is to obtain a view of genome wide variability patters for this population.

We prepared a total of 20 lines for full genome sequencing using Illumina technology. These 20 lines were chosen across different sampling spots (Figure 13C). Aware of the fact that localized residual heterozygosity found in genomes of inbred lines complicates the course of population genetics analyses, we used genomes of haploid embryos as material for the sequencing process. Haploid embryos from each chosen line were generated following the method of Langley *et al.* (2011) and their genomic material fully sequenced. We obtained 19 full genomes from this sequencing effort (Table 5). The embryo of line SU20 was virtually devoid of any nuclear DNA, despite having passed embryo DNA assessments (see Materials and Methods).

Within the 19 Umeå genomes there is substantial variation regarding sequence depth and coverage (see Table 5). A total of 14 genomes can be regarded as the "high quality" sequence set. Across this high quality group, the mean sequence depth was 59X (±22.5X) with an average value of median sequence depth of 51.2X (±22X). Sequence coverage in this group is homogenous with a mean of 68.84% (±0.96). The remaining 5 genomes make up the low quality set with both average mean and median depth of 4X and coverage of 51.3 % (±13.97). We observed no correlation between mean depth and genomic coverage (Spearman ρ= 0.51971, P>0.01).

Table 5. Assembly statistics of the 19 Umeå genomes.

No.	Lines	Nuc gen reads (%)	Heterozygous sites (%)	1st Q. depth	Median depth	Mean depth	3rd Q. depth	Coverage (%)
1	SU25n	80.86	0.0000877	57	88.00	91.31	121	67.46
2	SU05n	71.57	0.0000762	49	80.00	89.62	118	68.05
3	SU37n	77.25	0.0000957	51	75.00	80.50	103	68.05
4	SU93n	68.52	0.0001042	43	68.00	80.77	103	68.64
5	SU26n	79.37	0.0000928	42	62.00	66.82	85	68.64
6	SU58n	71.96	0.0000744	39	58.00	64.90	83	69.23
7	SU75n	81.82	0.0000714	33	58.00	66.74	89	69.23
8	SU81n	75.00	0.0001143	28	45.00	52.63	67	71.01
9	SU02n	83.77	0.0000783	24	43.00	59.42	74	67.26
10	SU94	61.71	0.0000540	20	37.00	53.17	65	69.23
11	SU29	73.56	0.0006757	19	34.00	42.27	55	68.64
12	SU07n	71.57	0.0000775	19	29.00	33.02	41	69.82
13	SU08	55.81	0.0000034	12	20.00	25.84	33	69.23
14	SU21n	81.67	0.0000000	13	20.00	21.81	29	69.23
15	SU46n	24.27	0.0006921	1	3.00	17.51	11	39.05
16	SU11n	39.13	0.0004643	2	4.00	10.30	9	33.73
17	SU77	16.74	0.0002205	3	5.00	8.13	10	63.91
18	SU23	41.93	0.0000854	2	4.00	8.25	7	61.54
19	SU96n	22.77	0.0001282	2	4.00	7.49	8	58.93
20	SU20	0.00	NA	NA	NA	NA	NA	NA

In this table, 'Nuc gen reads' shows the fraction of the reads per embryo that aligned to nuclear *D. melanogaster* genome reference assembly (see Material and Methods). Heterozygous sites indicate the fraction of position where more than one base was observed. The four columns related to sequence depth provide an idea of the distribution of values for this variable per embryo. Note that all positions of the reference genome (including the mitochondrial genome) are taken into account. Coverage reports the fraction of the reference assembly that was covered by reads derived from each sequenced embryo.

2.4.3 A preview of selection patterns in Umeå

After having generated this valuable Swedish *D. melanogaster* dataset, it is impossible to refrain from looking into variability patterns across regions (and even entire chromosomes) of interest. We examined some genomic regions in this Swedish population and briefly report our findings. Because of the interest generated by the results presented in sections 2.2 and 2.3, we looked at the X-linked 124-kb long QTL region (at 7A3-7B1). We retrieved the 14 core sequences of the Swedish set for this region from our server and run our pipeline to obtain standard summary statistics (see Materials and Methods). It was not surprising to learn that average values of sequence diversity and divergence with *D. simulans* were on the range reported for other European populations (Table 6) and that the approximately 40 kb reduction of heterozygosity that characterizes the European selective sweep was also present in this Swedish population (Figure 14A).

A remarkable feature revealed by pair-wise F_{ST} estimates among European and African populations is that, for this region, Sweden is on average the least differentiated of the three European populations with respect to the Africans, and among Europeans Sweden and France are the closest pair (Table 7). The immediate scenario that comes to mind when looking this pattern of population differentiation is that Sweden exchanges more migrants with France and Africa than with its closest geographic population the Netherlands, in fact it seems as if the Netherlands were the most isolated of all five populations.

While there are several hypotheses that emerge to provide explanations to this counterintuitive observation, we have to remember that this fragment represents only a small part of the genome and that conclusions drawn here should be treated with caution. A detailed modeling of gene flow among these populations using the available genome wide pattern of variation will certainly shed light on this matter. Another reason for this observed pattern of genetic differentiation might be that the Dutch sample does not reflect the *status quo* of this population. Unlike the Swedish or French collections that took place within the last 5 years (Pool *et al.* 2012), Dutch flies were collected ~14 years ago. Such 10-year lapse between samplings may have captured different population dynamics marked by recent, improved dispersal (because of intensified fruit trade) and accentuated by current global warming.

Table 6. Summary statistics at 7A3-7B1 in three European *D. melanogaster* populations.

Population	θ_W (mean ±SD)		θ_π (mean ±SD)		D_{xy} (mean ±SD)	
SU	0.0008	±0.0005	0.0008	±0.0006	0.11160	±0.04213
NL	0.0010	±0.0006	0.0010	±0.0007	0.09529	±0.03522
FR	0.0008	±0.0006	0.0008	±0.0007	0.10423	±0.03817

Table 7. F_{ST} at 7A3-7B1 among five *D. melanogaster* populations.

	F_{ST} (mean ±SD)							
	SU		NL		FR		RG	
NL	0.1472	±0.1462						
FR	0.1023	±0.1796	0.1180	±0.1446				
RG	0.2464	±0.1480	0.3148	±0.1425	0.2733	±0.1346		
SEA	0.2312	±0.1487	0.3223	±0.1402	0.2676	±0.1420	0.0998	±0.0929

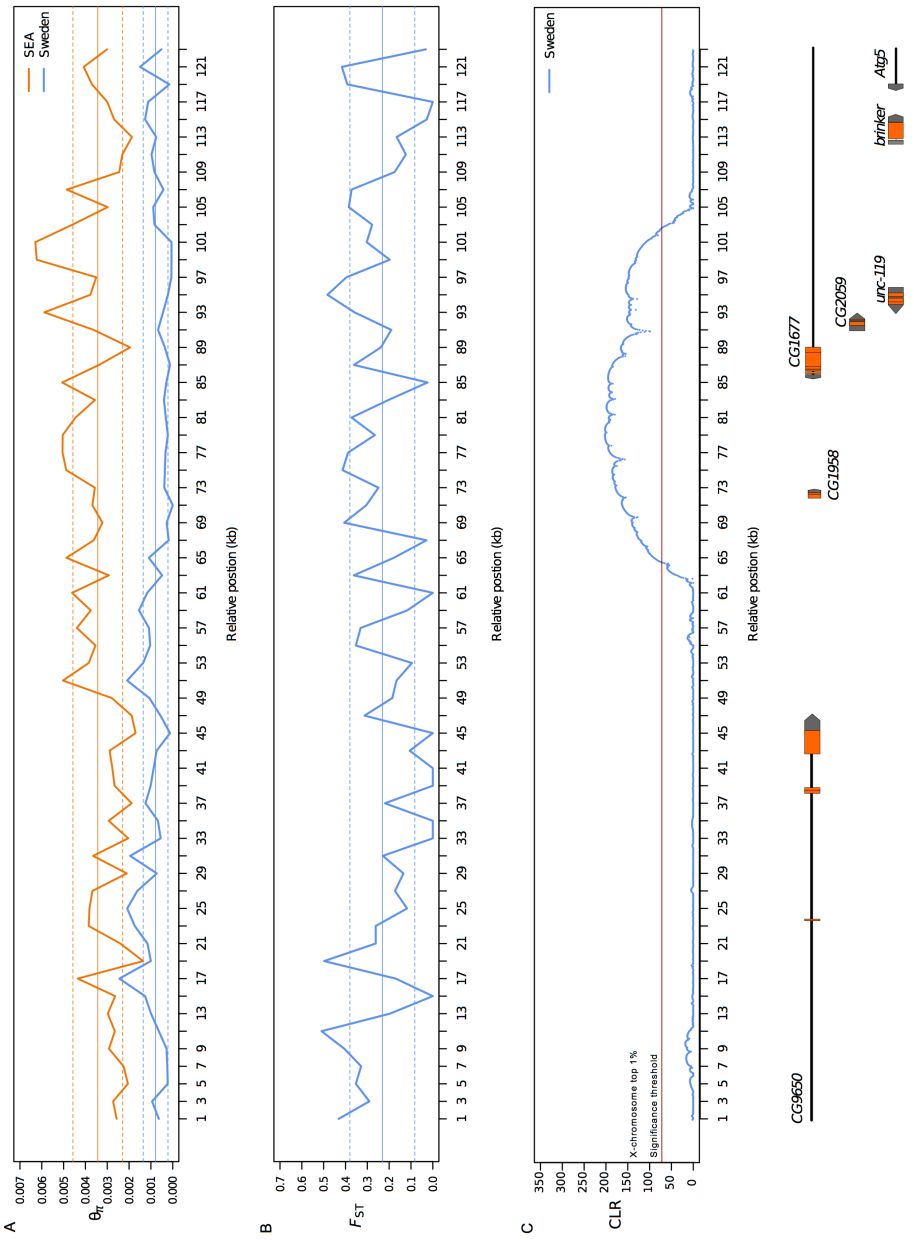


Figure 14. Patterns of genetic variation at 7A3-7B1 in Sweden.

A) Nucleotide diversity ($\theta\pi$) obtained for consecutive 2-kb long windows in Sweden (SU), light blue, and a pool of Southeast African (SEA) lines sampled around Lake Kariba in Zimbawe and Zambia in orange. B) Differentiation (F_{ST}), between SU and SEA calculated as normalized distance of Nei. C) Likelihood (CLR) profile along the 124 kb at 7A3-7B1 using SNP data of Sweden, with significance thresholds as in Figure 11. In panels A and B, thin, continuous lines represent the average value for each summary statistic across the 62 windows, dashed lines represent 1 SD above and below the corresponding summary statistic mean.

Finding evidence for positive selection in the genome is another intended goal of the analysis of this Swedish dataset. We conducted a X chromosome wide search for footprints of selective sweeps based on the analysis of the SFS, using the program SweeD (Pavlidis *et al.* 2013), just as we did with the X chromosome dataset from the Netherlands and France in the preceding section. The sweep-like CLR profile, observed in the Netherlands and France at 7A3-7B1 was also detected in our Swedish population (Figure 14C). Although the demographic modeling of this population has not yet been completed, which leaves us without a proper scenario to test this selective hypothesis; this selective sweep likely reflects an authentic signal of positive selection, as revealed by its height along the X chromosome scan (Figure 15).

The comparison of the Swedish CLR profile with that of the Netherlands and France (as one population) reveals strong similarities in the location of selective sweeps, yet the heights of the corresponding peaks might be different. This is a likely reflection of the common evolutionary history of these European populations, for instance the peak under deficiency *Df(1)ED6906* at 7.2 Mb is the highest among all observed in both datasets. Interestingly, the 0.5 Mb neighborhood where this peak is located appears as the most conspicuous selective sweep hotspot along the X chromosome. There are also other likely cases of selective sweeps private to the Swedish population, for example a series of peaks at 14.2 Mb that are not observed in the Netherlands or France. By obtaining CLR profiles along broad genomic regions, such that presented here, we can now locate selective sweeps with unprecedented precision. The task that still remains a challenge is to functionally characterize these selective sweeps and put these results in the light of trait evolution (Ober *et al.* 2012).

Figure 15. CLR profile along the X chromosome in three European populations of *D. melanogaster* (opposite page).

CLR values above significance line at 72 (see chapter 4 and methods section for SweeD) mark putative selective sweep regions. Note how these regions are shared by both European datasets. The red arrow at 7.2 Mb indicates the CLR of the region under deficiency *Df(1)ED6906*.

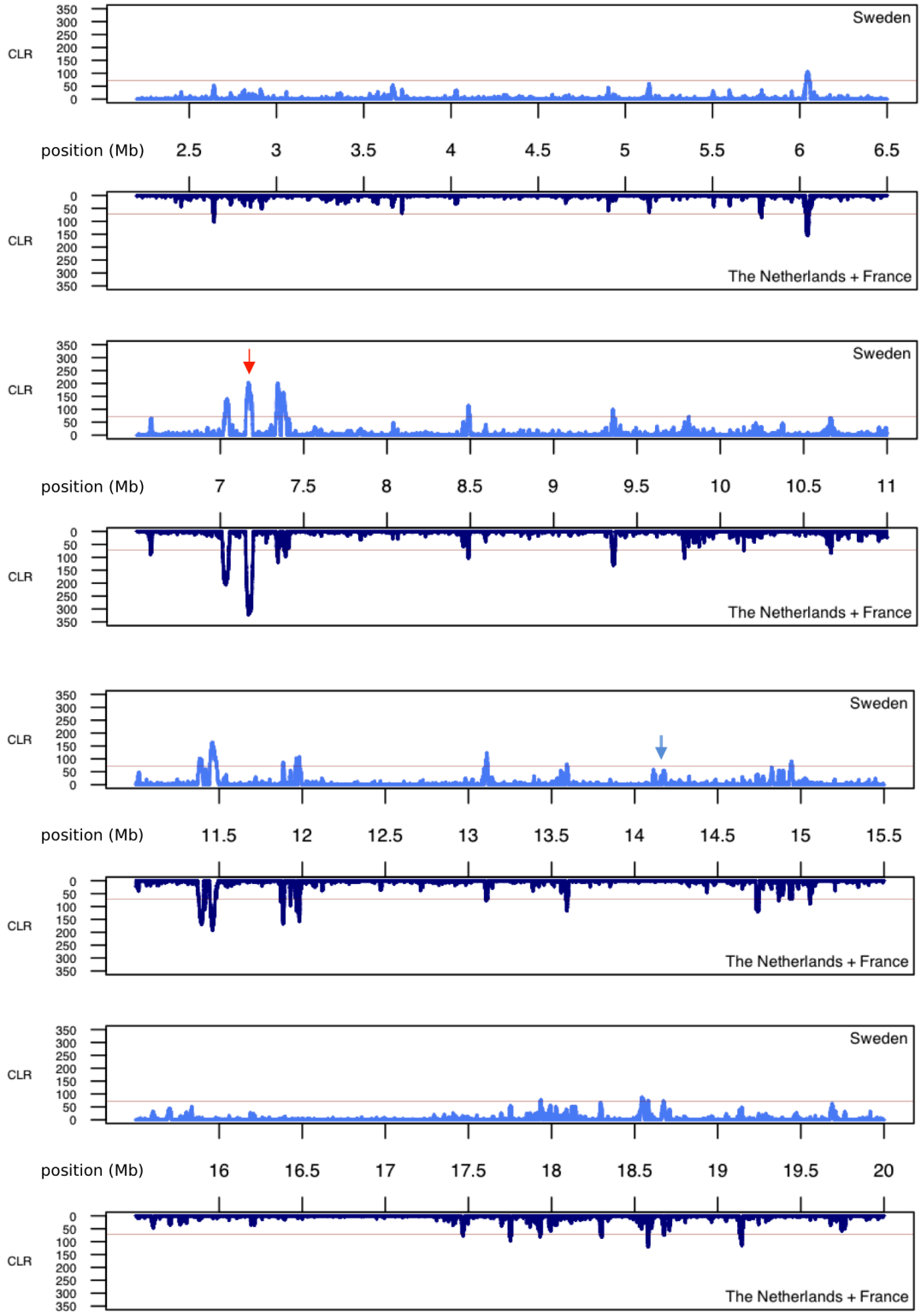


Figure 15. CLR profile along the X chromosome in three European populations of *D. melanogaster*.

III – DISCUSSION

3.1 THE GENOMIC BLUEPRINT OF COLD TOLERANCE

3.1.1 X-linked variation affecting CCRT: from QTL to QTGs

The interest in revealing the genetic basis of complex traits has, thus far, been mostly limited to the field of quantitative genetics. For an evolutionary quantitative geneticist, the interest goes beyond cataloguing genes (or their variants) that affect a given phenotype. The aim is to use this knowledge to understand the direction of evolution in traits and the constraints that are placed on them (Lynch and Walsch 1998). As population geneticists, we have joined this "gene mining endeavor" (see Rockman 2012) with a more specific goal in mind, namely to investigate whether selection on these traits is also reflected in its underlying genes. In the current study, we were particularly interested in knowing whether these genes also bear the footprints of selective sweeps. We started by defining our phenotype of interest: cold tolerance. This trait is assessed in *D. melanogaster* by scoring the time to recover after chill-induced coma. Subsequently, we developed an experimental design to identify the genes (and their variants) that affect this phenotype in order to define the appropriate genomic regions for population genetics analyses. Although the entire genome is expected to harbor variants of interest, here we restricted our search to the X chromosome where no cold tolerance genes have been reported up to this point (Svetec *et al.* 2011).

Our results confirm recent findings that the X chromosome harbors variants that affect CCRT. Initial evidence of substantial X-linked variation associated with CCRT came from transcriptome analyses by (Ayroles *et al.* 2009) and (Telonis-Scott *et al.* 2009). Subsequently, we revealed the first set of X-linked QTL for cold tolerance (Svetec *et al.* 2011). A total of six QTL with significant additive effects on CCRT were identified. These QTL encompass broad chromosomal regions, each containing hundreds of genes that can, in principle, be regarded as candidate QTGs. Furthermore, these QTL were affected by sex-specific and QTL - QTL (epistatic) effects. This is a fact that complicates the elucidation of how these QTL may influence the phenotype (Phillips 2008; Mackay 2014). For instance, the QTL at 0 cM is only detected in male flies (and explains up to 9.30 % of the phenotype variance) and the QTL at 9 cM is only detected in female flies (explaining 5.63% of the phenotype variance). These sex-specific effects can arise from actual sex differences in the architecture of the trait or merely represent a lack statistical power to reveal its effect in the other sex (Mackay 2001; Svetec *et al.* 2011). Epistatic interactions in the context of QTL mapping, if not considered carefully, may promote misleading interpretations. For instance, chromosomal intervals may appear as QTL with spurious additive effects. These cases usually occur when there are interactions among two or more factors, which on their own do not have any effect on the phenotype. In the work of Svetec and colleagues (2011) the level of resolution within each QTL was mainly limited by the density of markers employed to define mapping intervals. The question still remains, however, whether each QTL interval represents a single QTL or arises from the interaction of another smaller QTL nestled within.

Dissecting these QTL is the only possible way to identify the underlying genes and reveal whether sex-specific QTL or epistatic interactions are responsible for the assigned effect on the phenotype. In this work, we opted for two parallel approaches to carry out the dissection of the three X-linked QTL for CCRT reported by Svetec *et al.* (2011). One approach consisted of selective sweep mapping (also known as hitchhicking mapping, see Nuzhdin and Turner 2013) applied to the QTL interval at 56 cM (cytological interval 13E-20E). The other method was deficiency-based quantitative complementation mapping (Pasyukova *et al.* 2000; Mackay 2001), carried our on the QTL at 17 cM (cytological interval 6C-10B) and at 24 cM (cytological interval 8E-11D). In the following section, we discuss the results of the quantitative deficiency mapping approach. The results obtained with the selective sweep mapping method are discussed in section 6.2.2.

The QTL intervals at 17 and 24 cM were chosen because they have substantial overlap with each other and exert individual additive effects of CCRT, while exhibiting epistatic interactions in flies of both sexes (Svetec *et al.* 2011). This is advantageous, because quantitative complementation tests for X-linked QTL in *D. melanogaster* can only be conducted in the homogametic sex. We used a minimal set of 24 deficiencies that we tested in female flies revealing potential allelic (or additive) effects on CCRT at the following cytological intervals: 7A3-7B1 (*Df(1)ED6906)*; 7D1-7D5 (*Df(1)C128)*; 7F1-8A2 (*Df(1)BSC592)*; 8C-8E (*Df(1)BSC537)*. Of these four intervals, the first two had highly significant, likely additive effects. The comparison of the location of these new intervals with respect to the span of the initial QTL reveals that (i) the QTL at 17 cM is split into two significantly linked intervals and (ii) the initial QT at 24 cM does not seem to be represented by any of the deficiencies that failed to complement except for *Df(1)BSC537* at the interval 8C-8E.

The apparent differences between the original QTL intervals and those revealed by our deficiency mapping approach can be readily explained. For example, the fact that the QTL at 17 cM is now split in smaller intervals, each with additive effects on the trait in question, is a common feature of QTL fine mapping methods. This has been seen elsewhere within QTL for longevity (Pasyukova *et al.* 2000), olfactory behavior (Fanara *et al.* 2002), starvation resistance (Harbison *et al.* 2004), and locomotor behavior (Jordan *et al.* 2006). The commonly accepted explanation for this observation is that many broad intervals detected by recombination mapping methods represent the compound signal of multiple linked factors. Successful confirmation of the individual effects of the smaller QTL by functional tests has been achieved in several cases (Mackay 2001). Likewise, the overlap of several linked SNPs, identified via GWA studies, with known QTL for the same trait is further evidence of this explanation, and it emphasizes the underestimation of the associated factors while overestimating its effects (Mackay *et al.* 2009) as a general feature of recombination mapping as a method. On the other hand, the lack of correspondence between the deficiencies within the cytological interval at 24 cM should not be used as an argument to rule out the existence of this QTL. A parallel approach, such as GWA study, may substantially help with the resolution of these conflicting results. In the next section, we will consider the greatest weakness of quantitative complementation testing, namely that it cannot distinguish between additive and epistatic failures to complement.

3.1.2 Ubiquitous epistasis

Variation in quantitative traits arises by the interaction between genes, non-coding functional and sites throughout the genome. This may provide the elements for charting traits in the genome in order to understand their potential to evolve (Mackay 2014). Epistasis is a term coined by Bateson (1909) to describe the ability of a gene to "mask" the influence of mutations at another gene on a given phenotype (Cordell 2002). The same phenomenon, but from a population genetics perspective, was described by Fisher in 1918 under the term 'epistacy'. Specifically, epistacy refers to the statistical deviation of multilocus genotype values from an additive model for the value of a phenotype (Phillips 2008). Although the term epistacy is seldom used in the contemporary lexicon, both ideas of epistasis pertain to quantitative genetics, and are of particular importance in QTL mapping approaches (Lynch and Walsh 1998; Mackay 2001, Phillips 2008).

As previously discussed, the X-linked QTL reported by Svetec *et al.* (2011) at 17 cM and at 24 cM revealed a significant epistatic interaction. In order to determine which genes within each interval are likely to be interacting, we dissected these two QTL with a deficiency complementation test approach (chapter 3). Although we initiated the discussion of these results in the preceding section, here we focus exclusively on the effect that epistasis may have contributed to our results. Using quantitative complementation tests with deficiencies that served as the genomic background of two X chromosomes derived from European and African populations (Svetec *et al.* 2011), we observed that the intervals at 7A3-7B1 (*Df(1)ED6906*) and 7D1-7D5 (*Df(1)C128*) showed a significant failure to complement. However, this result does not have a unique interpretation. A failure to complement may arise from purely additive effects of alleles uncovered by the deficiency (the expected explanation) or by interactions between these uncovered alleles and other sites in the genome (Falconer and Mackay 1996; Pasuykova *et al.* 2000, Mackay 2001).

Our results are likely cases of allelic failure to complement, though we cannot fully discount the concomitant epistatic effects that became evident in the ANOVA analyses (Table 1). We noticed that for both deletions, the expected *L x G* effect was also accompanied by a highly significant *L* effect. By looking back at our experimental design, we can find the likely source(s) of this epistatic effect. Since we used the E* and A* lines in deficiency tests and not their wild type progenitors, NL14 and ZK157, we may arguably be able to restrict the source of epistasis to the X chromosome. There is still uncertainty as

to how much of the two observed failures to complement are due to epistasis. Although we lack additional information to provide an answer to this question, we suggest that one way to disentangle any possible epistatic and additive effects on CCRT is by testing the same deficiencies with lines where smaller intervals of the two wild type derived X chromosomes are introgressed on the same genetic background. Doing so would control for the effects due to the genetic background. We want to emphasize, however, that such an approach is beyond the scope of this thesis and that the impossibility to rule out any effect of epistasis in our findings does not interfere with our major goals in this work. Furthermore, all of the other methods thus far employed to identifying candidate genes for cold stress tolerance in the literature (*e.g.* GWAS (Mackay *et al.* 2012), transcriptome analyses on artificially selected populations (Telonis-Scott *et al.* 2009) or in naturally evolved populations (Ayroles *et al.* 2009)) are also sensitive to the confounding effects of epistasis (Huang *et al.* 2012). Since epistasis is inherent to biological systems (Remold and Lenski 2004; Shao *et al.* 2008), the best course of action is to treat it as an asset or a tool, instead of a nuisance, when mapping the QTL or QTNs that affect complex traits (see Verhoeven *et al.* 2010). Consequently, in the remainder of this chapter, epistasis is invoked to describe the ways in which our candidate genes might be contributing to variation in cold stress tolerance.

In spite of the unpredictable nature of epistasis, a handful of candidate genes for CCRT and other related proxies of cold stress tolerance appear consistently among previous investigations (see Hoffmann *et al.* 2003b for a review). A common function among these genes is their role in known physiological processes that lead to cold-induced coma or that are able to reverse this state (reviewed in Macmillan and Sinclair 2011). For instance, the involvement of heat shock protein genes, such as *hsp70* in CCRT is easy to understand, considering the protective role of these proteins against the damaging effects of temperature extremes (Hoffmann *et al.* 2003b; Colinet *et al.* 2009). Another important example is *Smp-30*, a gene that encodes as 33kD protein involved in Ca^{2+} ion homeostasis, since loss of ion homeostasis is one of the well-studied physiological disturbances induced by cold stress in insects (Macmillan *et al.* 2012). *Smp*-30 is currently the best-studied candidate gene for CCRT, even in the context of population genetics (see Clowers *et al.* 2010). In this work, we gathered initial evidence suggesting that *brk* is the first X-linked gene that might be involved in CCRT and cold tolerance, this is discussed in the following section.

3.1.3 Candidate QTG: *brk*

Our current understanding of the molecular role of *brk* comes from the field of developmental biology. Brk is a transcription factor with DNA-binding properties. During early development this protein targets and represses the expression of genes that belong to the Decapentaplegic (Dpp) network (Kirkpatrick *et al.* 2001). Both Brk and Dpp are necessary for successful patterning of embryonic and larval structures. In the developing wing, Brk triggers an apoptosis cascade in cells with impaired Dpp signaling (Minami *et al.* 1999; Moreno *et al.* 2002; Ziv *et al.* 2009), however, the mechanisms behind this apoptotic pathway remain poorly understood (Suissa *et al.* 2011). The expression of *brk* is itself negatively regulated by the presence of Dpp. Dpp signals the formation of yet another expression repressor protein complex called SMM (so named for its three constituent proteins: Schnurri, Mad and Medea), which binds to multiple modular enhancers along the 16 kb-long promoter region of *brk* to repress its expression (Pyrowolakis *et al.* 2004; Yao *et al.* 2008).

To our knowledge, our work is the first suggesting the involvement of *brk* in an adult phenotype. The results from a series of expression analyses suggest that the difference in the level of *brk* expression between Zimbabwean and Dutch flies is likely associated with the average CCRT observed between these two populations (Figure 5). The actual challenge is identifying the genetic variants that could bridge these two phenotypes. In addressing this challenge, we found a haplotype consisting of a set of SNPs and deletions located 3 kb upstream of *brk*, exactly where transcription factor binding sites affecting *brk* expression were identified (Yao *et al.* 2008). We hypothesized that this haplotype could help us explain, at least to some extent, the different expression levels of *brk* between populations, as well as a fraction of the quantitative variation in CCRT. We did not find clear evidence to support these hypotheses. The inbred lines that were employed to assess correlations with the presence or absence of the haplotype did not reveal significant differences in *brk* expression between populations or between treatments. However, a trend among the lines of African origin was observed (Appendix C). When these observations are taken together (Figure 5, and see Hutter *et al.* 2008), the case can be made that *brk* is strongly influenced by its genetic background (Mackay 2001, Huang *et al.* 2013 Mackay 2014). Moreover, the effect of genetic background might also help us understand why a SNP located approximately 3kb upstream of *brk* shows substantial association with CCRT (Mackay *et al.* 2012) in the North American population of

Raleigh, which is a product of admixture of African and European gene pools (Duchen *et al.* 2013). This may also explain why the same site may have an effect on the trait in African flies and not in European ones.

Detected QTL-QTL interactions are likely reflections of underlying gene-gene interactions for a given phenotype. This may be an explanation for the interaction revealed by Svetec *et al.* (2011) between QTL at 17 and 24 cM and the potential epistatic effect of the intervals at 7A3-7B1 and 7D1-7D5 with other functional elements located elsewhere on the X chromosome. Our candidate gene *brk* could be one of these interacting factors. Reviewing literature in which *brk* is mentioned, we found a rich list of transcription factors and cofactors with binding sites within the 9 kb upstream of *brk*'s 5' UTR (Anderson *et al.* 2005b; Yao *et al.* 2008; Negre *et al.* 2011). Since the phenotype of interest here is measured in adult flies, we reduced the list of interacting transcription factors to those with experimental evidence of expression in adult flies. There were three proteins of top interest that conformed to our criterion: Nejire (*CG15319*), CrebA (*CG7450*), and CrebB (*CG6103*).

It is possible that Nejire and CrebA/B are members of a transcriptional network that includes *brk* and are also activated in response to cold stress. Current experimental evidence suggests that these three genes have a moderate response to cold stress. However, considering that gene expression is a trait that is highly dependent on genetic background, it is better to take this information with caution and study these genes within each of our fly populations. Furthermore, it might be relevant to mention that *nejire* is just one of the many X-linked genes located in the cytological interval spanned by the QTL at 24 cM.

Nevertheless, experimental validation of these gene-gene interaction hypotheses is required to check whether they actually have a role on cold stress tolerance variation or any other trait related to environmental adaption. The relationship of *brk* expression to variation in CCRT may be further addressed by following a mapping strategy for variants affecting its expression, for instance as part of an "expression QTL" (eQTL) mapping effort (Sun and Schliekelman 2010; Zichner *et al.* 2013), by directly studying the candidate eQTG for the genes of interest (*nejire* and *CrebA/B*), or by performing protein immunoprecipitation assays under cold stress conditions. The latter approach could provide an insight into the transcriptional networks that are activated during stress response, revealing the set of transcription factors and their respective binding sites that

are active under cold stress. Moreover, performing these analyses in lines with contrasting CCRT such as NL14 and ZK157 and/or their X chromosome introgressed derivate E* and A* can tell us about the transcriptional differences that underlie their cold stress tolerance differences and to what extent X-linked allelic differences are reflected in this difference or are masked by the isogenic background.

3.2 THE MEANING OF SELECTIVE SWEEPS AT A QTL

At the interval 7A3-7B1, when *Df(1)ED6906* was deleted, we detected and characterized one of the strongest signatures of positive selection in the *D. melanogaster* genome. In fact, it appeared as the single strongest selection signature on the entire X chromosome (see Figure 11 and Appendix F). Although we have not estimated its selection coefficient, we the CLR values reveal that the SFS of this region is exceptionally biased towards fixed and high frequency derived variants, in comparison to the rest of the X chromosome in European *D. melanogaster*. This is an unequivocal sign of the effect of positive selection in molecular space. The sweep region encompasses about 40 kb with boundaries that are sharply defined. Of the seven candidate genes mapped to this cytological interval, the coding region of four of them (*CG1958*, *CG1677*, *CG2059*, and *unc-119*) are found within the selective sweep. We observed that *brk* and approximately 11 kb of its upstream enhancer region are left outside this selective sweep.

After having found such a strong signal of positive selection involving four different coding regions, it was necessary to determine the exact target of selection. To achieve this, we studied patterns of allele frequency differentiation in a panel of European and African (derived and ancestral) populations. This is an approach that proved to be of enormous utility, since the information offered by the European population alone would have been limited, due to their virtual absence of variation along the 40 kb sweep interval. With these F_{ST}-based analysis (Foll and Gaggiotti 2008), we observed that highly differentiated SNPs are enriched within the sweep area (Figure 11), and that the top SNPs were found within the coding region of *CG1677*, in the intergenic space between *CG1958* and *CG1677* and upstream of *unc-119*.

3.2.1 Candidate gene: *CG1677*

The most interesting candidate target of selection is composed of two SNPs at relative positions 86,661 and 86,670, within exon 5 of *CG1677*. Recent work on correlated evolution of tightly linked sites in the *Drosophila* genome, suggests that such cases are well explained by compensatory evolution (Callahan *et al.* 2011). In our case, each of these SNPs codes for amino acid differences at residues 936 and 939 of the peptide CG1677. In the Southeast African sample, the amino acid combinations Thr-Asn and Ala-Asp are both present, while in Europe the combination Ala-Asp is fixed (Figure 12A). An exciting result came with the model of the secondary structure of CG1677 (Kelley and Sternberg 2009) where we observed that both amino acid positions seem to be part of an α-helix, a configuration that implies an interaction of the amino acids at these positions.

Interestingly, Thr and Asn can form one hydrogen bond more between their side-chains more than Ala-Asp. The combination Thr-Asn may therefore make the protein more heat stable than Ala-Asp (Perl and Schmid 2002), which could be advantangeous in tropical Africa, given that ambient temperature is an important variant affecting life history traits in fruit flies. Conversely, the combination Ala-Asp may lead to a less rigid structure and thus a possibly more efficient protein, which may be an advantage in the temperate climate of Europe. Ancestral state reconstruction (Lewis 2001) shows that the Thr-Asn combination represents the ancestral state with high probability and that Ala-Asp arose through two point mutations within the *D. melanogaster* lineage (section 2.3). Since the intermediate states are not observed in the European and African population samples, the transition from Thr-Asn to Ala-Asp probably follows a compensatory evolution model (Kimura 1985; Innan and Stephan 2001) in which the intermediates are assumed to be strongly deleterious.

There are several possible explanations for why selection may have targeted these two amino acid positions in the peptide CG1677. The gene is homologous to the human spliceosomal protein ZC3H18, with a Zinc finger domain for binding with RNA molecules (Andersen *et al.* 2013). By alignment with vertebrate homologous proteins sequences, we determined that the sites we believe were targeted by selection do not belong to the Zinc finger domain, which is likely the active site of the protein. Experimental evidence from Herold *et al.* (Herold *et al.* 2009) revealed that *CG1677* is one of the spliceosomal proteins in the fly that physically interacts with at least 18 other

proteins while performing its RNA editing task. In the absence of a resolved tertiary structure of the protein that would allow us to determine where, in space, residues 936 and 939 are located, we can only speculate that these residues may be involved in spliceosomal assembly by directly participating in protein-protein interaction. It is plausible that compensatory evolution had taken place at these two positions to maintain (or optimize) the stability of the spliceosome in new environmental conditions, for instance at lower average temperatures. Finally, we should mention that although 'regulation of gene expression' is not a common biological category among genes associated with cold stress, two other genes on chromosome 3 (recently identified as candidate genes for CCRT) *Taf5,* and *lsm10*, also belong to this functional group. This is especially true for Lsm10, which is a part of the U7 small nuclear ribonucleoprotein complex (U7 snRNP), and plays an essential role in pre-mRNA processing (Fallis 2012).

The other two significantly differentiated SNPs in the sweep region 7A3-7B1 occur in noncoding locations at relative position 80,089, between genes *CG1958* and *CG1677,* and at relative position 101,154, within the large intron of *CG1677* (see gene model below Figure 11D). Considering the distance between these two sites one important question arises here: namely whether these SNPs are hitchhiking with the two sites in exon 5 of *CG1677*, or are themselves independent targets of selection. Our BayeScan results in Figure 11D and Figure 12B already suggest that a strong process of allele frequency differentiation might have been occurred within the African continent. For example, it seems that SNP 80,089 might constitute, or be linked to, yet another target of selection perhaps involving *CG1958,* which, has been reported as a rapidly evolving, strongly male-biased gene in flies of European origin (Hutter *et al.* 2008; Graveley *et al.* 2011). In addition, our qPCR assays (Figure 5) indicate that the expression of this gene in female African flies is higher and more variable than in Europe. Based on our data, it is impossible to determine whether the sex-biased pattern is reversed in Africa. In the African population, we studied expression for this gene in female flies only. Nevertheless, this may be a point of importance for future study, specifically looking at the meaning of such a strong selective sweep signal within a genomic region for relevance to cold stress tolerance.

3.2.2 Targets of positive selection and candidate QTG?

Thus far, we have discussed the role of one of the candidate genes as a likely target of selection, responsible for the selective sweep that was documented at the interval *7A3-7B1* under deletion *Df(1)ED6906*. This interval, as well as that at 15E, within the QTL at 56 cM (Svetec *et al.* 2011), contains other genes that may also be targets of selection. Of all the possible causes of positive selection, we are currently interested in climate-driven selection, which we investigated with QTL for CCRT. Furthermore, because of co-localization of these sweeps within the QTL for CCRT, we expected a causal association between the selective event and the evolution of the trait under consideration. In other words, we used selective sweep mapping as a criterion to prioritize candidate QTGs affecting CCRT. Our results, however, do not suggest that the genes under the investigated sweep regions are strongly related to the cold-stress phenotype. As revealed by quantitative complementation tests (Table 2) and gene expression analyses (Figure 5), *brk* is the only candidate gene that was induced by cold stress and it is located outside the sweep region. The remaining genes within the sweep region (*CG1958, CG1677, CG2059* and *unc-119)* did not show cold related changes in their expression levels (Figure 5). Although *CG1958* is differentially expressed at the constitutive level between Dutch and Zimbabwean flies, this difference remains unaltered with stress (Figure 4). Furthermore, we have carried out quantitative complementation tests on two of the four genes under the sweep (*CG1677* and *unc-119*) where P-element insertions disrupting the expression of the genes were available. None of these tests (performed in the same way as with the deletions) revealed quantitative failure to complement (Table 2). Concordant results have been obtained when assessing expression levels of these genes at *7A3-7B1* in a *D. melanogaster* laboratory strain after cold stress (Graveley *et al.* 2011).

A similar situation can be seen with candidate genes *CG4491* and *CG16700,* flanking the sweep at 15E. However, one exception should be noted. Ayroles *et al.* (2009) showed an association between *CG16700* transcript variation and CCRT in a North American fly population. This suggests, once again, that in the cytological region 15E, the likely target of the selection (*CG4991*) is different from the gene that might be influencing cold stress responses. However, the fact that *CG4991* is the target of selection (and not *CG16700*) is not clearly established by our results. Note that (in Figure 3) the expression level of *CG4991* does not differ between the Netherlands and Zimbabwe, and that the high levels of gene expression variation within the former population disagree with the expected

effect of selection as a buffering agent of phenotypic variation (Zhou *et al.* 2012). The case of gene *CG9509*, also a detected target of selection in Non-African *D. melanogaster*, is a clearer example of canalization of gene expression (Glaser-Schmitt *et al.* 2013).

Does this mean that selective sweep (or hitchhiking) mapping is a flawed approach to fine mapping of QTL affecting selection-driven traits? Hitchhiking mapping was envisioned as a QTL mapping tool in the late 1970s, however only two decades later did it start to be applied in model species for a variety of phenotypes (Keightley and Bulfield 1993; Keightley 1998; Nuzhdin *et al.* 2007; Turner *et al.* 2011; Remolina *et al.* 2012). A common feature of these studies is that selection for the trait of interest was carried out experimentally, using populations that were also established in the laboratory. Under these conditions, initial allele frequencies at marker loci were known from the onset as a baseline. Therefore, at the end of the selection regime, the detected differences in allele frequencies between control and treatment populations, like those predicted under selective sweep theory, can be attributed to the applied selective pressure. This observation raises two contrasting points with respect to our approach. The first point is that we conducted our fine mapping protocol in a naturally evolved population from a temperate location. Although we only focused on cold stress tolerance, we are aware that in natural conditions, in addition to this phenotype, other traits were naturally selected (see Werzner 2011). This explains the fact that our results show co-localization of selective sweeps and QTL for CCRT but do not reveal a causal correlation.

The second point is that in these artificial selection experiments (*e.g.* Nudzhin *et al* 2007), the methods that were used to identify the divergent genomic regions between selected and control populations could not assess the final allele frequencies after the selection regime was completed. Thus, the fraction of the identified differentiated genomic regions (or QTL) in which fixation of beneficial alleles actually occurred remains a mystery. This is important because in our selective sweep mapping approach we can only identify regions of the genome where the fixation of alleles has occurred. Interestingly, recent implementation of next generation sequencing technology within hitchhiking mapping protocol sheds light on this matter. Turner *et al.* (2011) selected for body size *in D. melanogaster* (in both directions) for 100 generations and determined genome wide levels of allele frequency differentiation among evolved and control populations once selection ceased. On average, 12% of the SNPs in which significant frequency changes were seen correspond to differences >0.95 (*i.e.* fixation of, arguably,

beneficial and linked neutral alleles).

It seems that hard sweeps represent only a small, yet non-negligible, fraction of the changes in allele frequencies driven by selection on multilocus traits. The fixation of beneficial alleles has been documented at genes associated with the domestication and development of breeds in economically important organisms such as chickens (Rubin *et al.* 2010), dogs (Axelsson *et al.* 2013), and pigs (Rubin *et al.* 2012). But, this has also been documented in natural populations, such as coat color in deer mice (*Peromyscus polionotus*) (Linnen *et al.* 2013) or lateral plate armor in three-spine sticklebacks (*Gasteroseus aculeatus*) (Cano *et al.* 2006; Barrett and Hoekstra 2011). One can argue, however, that reduced pleiotropic effects of these genes on other ecological/life history traits frees them from any selective interference, therefore their response to directional selection is akin to that on monogenic traits. In conclusion, the use of population genetics approaches, focused on hard selective sweep mapping, to characterize QTL affecting polygenic traits such as cold stress tolerance, may not be sufficient to reveal the loci that harbor the fraction of genetic variation that matters for the evolution of the trait.

3.2.3 Allele frequency shifts at a QTL

The key changes in polygenetic adaptation are likely to occur in the parts of the allele differentiation spectrum that represent intermediate allele frequency changes due to selection. In other words, the allele frequency shifts between populations. While theoretical results are succeeding in promoting the paradigm change from viewing adaptation as occurring mostly by the fixation of new variants (hard sweeps), there is also increasing acknowledgement of situations where selection occurs on standing genetic variation that does not lead fixation of beneficial alleles (Chevin and Hospital 2008; Pritchard and Di Rienzo 2010; Pritchard *et al.* 2010). We need to be able to detect these shifts at loci affecting adaptive traits in natural populations to gain a better understanding of this phenomenon.

In the present work, we conclude that *brk* is the prime candidate gene affecting variation in cold tolerance (among the seven genes uncovered by deletion *Df(1)ED6906*). In all the assessed *D. melanogaster* populations, we observed that genetic variation around *brk* is not lost, and particularly in European populations it is much higher than in the adjacent sweep region (see Figures 10, 11A). We defined a series of indel polymorphisms in the enhancer region of *brk* (Figures 7 and 8) at relative positions 109,442 to 109,976 (*i.e.* about 3 kb upstream of *brk's* 5' UTR) and thus also outside the sweep region. Using an extended sample of populations from the DGPG2 project (Pool *et al.* 2012), we investigated the frequencies of this indel polymorphism in these populations. We classified the indel polymorphism into non-deletion haplotypes and three classes of deletions (see Figure 8). Based on linear regression analysis of the frequencies of the non-deletion haplotypes, we detected two antiparallel latitudinal clines where one spans from the populations near the equator (Rwanda, Gabon, Cameroon, Ethiopia, and Nigeria) to the north (France and the Netherlands) and another one from the equator to the south (Southeast Africa and South Africa) ($P<0.05$ in both cases; Figure 7C). Unlike well-studied cases of clinal variation in metabolic genes in *Drosophila* (*e.g.* the *Adh* gene) (Kreitman 1983; Hoffmann and Weeks 2007), both the functional and selective significance of the geographic correlation reported here for variants at the enhancer region of *brk* has yet to be established. There is the outstanding question about whether this search can be done at genomic scales.

Allele frequency shifts associated with environmental adaptation have been already reported in humans (Hancock *et al.* 2010; Pritchard *et al.* 2010; Hancock *et al.* 2011). These studies made use of geographic or climatic variables to support the hypothesis that observed allele frequency shifts at tested SNPs were driven by selection. The underlying idea of their approach was to find loci where the allele frequencies showed unusually strong correlations with the environmental variable (Coop *et al.* 2010; Günther and Coop 2013). Clearly F_{ST} is the summary statistic that can capture frequency shifts across populations (Hancock *et al.* 2010), but the successful implementation of this method requires the development of a null model in which the confounding effects of shared ancestry, gene flow, and genetic drift are considered (Beaumont 2005; Coop *et al.* 2010). Moreover, sampling error or uneven sample sizes among populations can also create false correlations between allele frequencies and environmental variables.

Given these issues, there are two main conclusions to be drawn from our experience with the *brk* enhancer region. Namely, (i) Although all the SNPs within the entire enhancer region of *brk* were subjected to an F_{ST}-based test of selection (Foll and Gaggiotti 2008; Gaggiotti and Foll 2010), the fact that the allele frequency shift that we reported by other methods, had remained undetected by this F_{ST} analysis, attests to the insensitivity of many of the most powerful approaches in population genetics to identify cases of polygenic selection and (ii) however significant, the correlation between latitude and frequencies of insertion-deletion haplotype in *brk* enhancer region maybe spurious. Any future effort to find local adaptation in *Drosophila* should follow the considerations of Coop *et al* (2010) and integrate all available genomic resources such that more continents and geographic regions are better represented in the analyses.

3.3 OLD QUESTIONS IN THE LIGHT OF NEW DATA

For a long time, evolutionary biologists have been interested in understanding the effects of mutation, recombination, genetic drift, selection, demographic dynamics, migration, etc. on the genetic composition of organisms and populations. However, these questions have been chiefly addressed at a theoretical level, and only in the last 50 years, in an empirical manner, using genetic markers such as allozymes, RFLPs, microsatellites or a handful of genes. Although these studies represent a substantial deal of effort by the scientific community, the level of understanding of the initial phenomena is still incomplete. High throughput DNA sequencing technologies were developed with the promise of providing scientists with access to the genomes and its secrets. For evolutionary biologists this has also meant an opportunity to grasp the whole breadth of variation contained within populations. Nevertheless, several technical difficulties have had to be overcome before these technologies could properly assist *Drosophila* population geneticists in their research.

Thus far, the DPGP is the leading *D. melanogaster* sequencing effort with an unprecedented sampling program on the African continent (Pool *et al.* 2012). By the end of its second phase, it has yielded a number of 139 genomes representing 21 African populations and 1 European population. Efforts undertaken by other laboratories have sequenced populations from around the globe. For instance, Kolaczkowski *et al.* (2011) sequenced populations at the extremes of a latitudinal gradient in Australia; Fabian *et al.*

(2012) concentrated on three populations along the east coast of the United States; Campo *et al.* (2013) focused on the North American west coast. With a total of 168 fully sequenced inbred lines from Raleigh (North Carolina), the DGRP represents the largest data collection from a natural population (Mackay *et al.* 2012). *D. melanogaster* populations from the Middle East (Hübner *et al.* 2013) and Malaysia (W. Stephan, unpublished data) have also been completed. Our own sequencing effort is, to our knowledge, the only one that is meant to investigate the genetic composition of a high latitude population of *D. melanogaster*.

In spite of its 19 sequenced lines, representing a middle-of-the-ground sequencing effort, the sequence quality of the Umeå dataset is high. First, we dealt with residual heterozygosity by sequencing haploid embryos (Langley *et al.* 2011). As expected, the reported fraction of heterozygous positions after completing the sequence reads to the reference genome was on average 0.00015%. Second, our average sequence depth of 59.4X is 3 times higher that achieved by DGRP and 6-fold higher than that of the Australian east coast, and the North American east and west coasts. This increases our power to determine which of our SNPs are likely representing true natural variants and which are artifacts of the sequencing process. This, therefore, enhances the reliability of the conclusions that can be drawn from our dataset.

The first population genetics question that our Swedish dataset will help investigate, together with DPGP2 dataset, deals with gene flow (Wright 1931). The immediate goal is to be obtain the estimates of migration rates between continents, a task that is currently being undertaken within an approximate Bayesian computation framework (Duchen 2013). If these estimates can be successfully obtained, we would able to understand the population dynamics taking place near the latitudinal border of *D. melanogaster*'s habitat range. We would be able to know, for instance, what European (or Mediterranean) populations exchange migrants with Scandinavia. Furthermore, it is still unknown whether Northern European populations overwinter *in situ* or whether they die out by the end of the warm season (Izquierdo 1991). If the latter scenario is true, Northern European populations are likely to be reestablished in late spring by migrants from warmer areas in the south.

A long term goal, however, is to develop methods to understand how much of the genetic variation that is shared by two or more populations can be unequivocally assigned to the homogenizing effect of gene flow and that of shared ancestry. If we could

determine which variants are new in the population because of gene flow, we would be able to proceed with testing hypotheses about the maladaptive effects of gene flow from habitat cores towards habitat edges (Kirkpatrick and Barton 1997; Kawecki 2008). This is a project for which our Umeå collection is suitable. Of course, knowing how allele frequencies change due to migration is not enough to fully address such questions. It is also necessary to know which alleles are potentially maladaptive, for instance those that affect life history traits and those traits, which increase survival in temperate environments. This requires knowledge of the genetic architecture of these traits in northern European populations, an endeavor we have just started with the present work (see also Svetec *et al.* 2011).

Regarding the genetic basis of phenotypic variation, the level of understanding achieved with the DGRP population has no parallel. The DGRP has served to associate standing genetic variation with organismal phenotypes such as time to develop, lifespan, starvation resistance and CCRT (Mackay *et al.* 2012). Genetic variation has also been associated with transcriptome variation, by mapping *cis*-acting elements (or *cis*-eQTL) (Massouras *et al.* 2012). These studies have paved the way to address more complicated questions in systems biology, such as: how stable the transcriptome is to environmental disturbances (Zhou *et al.* 2012)? And how does epistasis governs phenotype variation (Huang *et al.* 2012)? From this last study we have learned an important lesson. The effect of any genetic variant is strongly dependent on the composition of its genetic background. New genetic backgrounds (presence or absence of alleles at other sites in genome) give rise to new interactions and therefore to new ways of building phenotypes (Mackay 2014). By extension, this is true for the genes, functional elements, or QTL where the variant resides. This has important implications if we intend to use our Umeå collection to study adaptation to temperate environments using candidate gene lists reported elsewhere (Mackay *et al.* 2012). We can, for instance, consider the variants known to be affecting CCRT in the DGRP as our candidate QTNs, but we need to prove that these variants also have the expected effect to avoid erroneous conclusions. Alternatively we could conduct an association study and, in the end, compare the reproducibility of the results between studies. However, because of the reason presented above, the reproducibility is expected to be low (Huang *et al.* 2012; Rockman 2012; Mackay 2014).

The increasing complexity of the datasets used by evolutionary geneticists to test their hypotheses facilitates the desire of getting an accurate picture of what occurs in nature, but with the possibility of controlling as many variables as possible. The roughly 100 years that have elapsed between the dihybrid crosses conducted by Bateson (1909) and Huang and colleague's (2013) use of the DGRP to study complex phenotypes, have served to make us appreciate the pervasiveness of epistatic interactions. However, we are still far from being able to predict its occurrence. This, in turn, may be a consequence of how little we understand the way genomes work. The challenge for evolutionary geneticists is to link the study of genetic variability present within populations with its phenotypic dimension.

3.4 CONCLUSIONS AND PERSPECTIVES

The interest in understanding the genetic basis of cold stress tolerance in *D. melanogaster* is the common motivation to the different projects that comprise this dissertation. Cold tolerance is a phenotype that reflects the adaptation of the fly to temperate environments. We focused on the contribution of genetic loci on the X chromosome and reported the first X-linked candidate genes for this phenotype based on quantitative and population genetic approaches, in combination with expression studies. We summarize the emerging picture below.

The pattern of expression of the candidate gene *brk* is highly population-dependent and is governed by both *cis* and *trans* factors. We also observed that one of the putative *cis*-elements of *brk* shows a moderate change in frequency between natural populations along a latitudinal cline from tropical to temperate regions. Such frequency shifts are considered signatures of positive selection affecting traits controlled by many loci.

The case of *CG1677* is different and likely unrelated to CCRT. Although the gene is located near *brk*, *CG1677* exhibits some of the strongest signatures of a selective sweep on the entire X chromosome observed thus far in European *D. melanogaster* populations: (i) the largest CLR values and (ii) significant differentiation at two non-synonymous SNPs between temperate and tropical populations. The fact that the two non-synonymous SNPs are part of a α-helix such that the encoded amino acids can interact and form different numbers of hydrogen bonds between their side-chains provides insights into the

functional significance of our finding. However, we found no evidence that *CG1677* affects CCRT, and the selective sweep observed at *CG1677* does not overlap with the observed allele frequency shift in the 5' UTR of *brk*. Thus, since cold stress tolerance is a quantitative trait controlled by many QTL and *brk* is one of the genes involved (in contrast to *CG1677*), the signatures of selection we detected at these two genes are consistent with the theoretical predictions.

The conclusions we may draw from the involvement of genes *CG4991* and *CG16700* in CCRT are similar to the case of *CG1677*. Although these two genes are involved in in different selective sweeps in European populations of *D. melanogaster*, they also co-localize with a QTL affecting the CCRT. We have no further substantive evidence for the involvement of either of these genes in CCRT. Selective sweep mapping is theoretically a sensible approach to identifying genes associated with adaptive phenotypes. However, when used as a mapping tool in naturally evolved populations, the results should be treated with caution and substantiated with functional assays.

With the aim of extending available resources both at the genetic and phenotypic level that allow us investigate the evolution of cold tolerance mechanisms in natural populations, we collected *D. melanogaster* from a Scandinavian population, established a panel of 80 inbred lines and fully sequenced 20 of them. We have just begun to initiate the genetic characterization of this population, analyzing genome wide variation with population genetic approaches. We expect this resource will soon be used for quantitative genetics work. Although we did not aim to study epistasis, we realized how pervasive its effects are on phenotypic evolution. In the future we expect to make use of epistasis as tool to understand how phenotypic variation arises and not as a nuisance that should be excluded or minimized in the models to study complex traits.

In this work, by bringing together quantitative and population genetics approaches to address the question whether genes underlying complex traits show footprints of positive selection, we have made evident that the current models to study adaptation in sequence space (selective sweeps) might be able to explain only a small fraction of the evolutionary dynamics that underlie the phenomenon of polygenic adaptation. The challenge for the evolutionary genetics community in the following years is to develop statistical tools to identify and characterize these other instances of selection (*i.e.* allele frequency shifts) whereby adaptation at the phenotypic level can be achieved.

IV – MATERIALS AND METHODS

4.1 POPULATION GENETICS ANALYSES

4.1.1 Wild type fly stocks, next generation sequence data

Throughout the chapters that comprise this dissertation different *D. melanogaster* sequence datasets were employed in the analyses. Sequences were produced either by sanger method using the fly stocks maintained in our laboratory or obtained with Illumina technology of the same stocks as well as others derived to other African and European natural populations. These full genome datasets were completed as part of collaborative efforts with the lab of Prof. Charles Langley at UC Davis. (Langley *et al.* 2011; Pool *et al.* 2012). Publically available whole genome sequences generated by Illumina NGS technology for African and European *D. melanogaster* populations reported in Table 8 were retrieved from the DPGP (www.dpgp.org.Information). In addition some of the fly stocks were kindly donated to our fly collection by Professor C. Langley.

Table 8. Catalogue of wild type lines and NGS datasets used in this study.

Region	Country	Code(s)	Collector/reference	Lines	
				NGS dataset	Fly stocks
Europe	Sweden	SU	R. Wilches (2012) unpublished dataset	SU02N, SU05N, SU07N, SU08, SU21N, SU25N, SU26N, SU29, SU37N, SU58N, SU75N, SU81N, SU93N, SU94.	Up to 80 isofemale lines
	The Netherlands	NL	stocks donated by C. Langley	NL01, NL02, NL11, NL12, NL13, NL14, NL15, NL16, NL17, NL18, NL19.	NL01, NL02, NL11, NL12, NL13, NL14, NL15, NL16, NL17, NL18, NL19, NL20.
	France	FR	Pool *et al.* 2012	FR14, FR151, FR180, FR207, FR217, FR229, FR310, FR361.	-
Eastern Africa	Ethiopia	ED-EZ	Pool *et al.* 2012	ED2, ED3, ED5N, ED6N, ED10N, EZ2, EZ5N, EZ9N, EZ25.	-
Central Africa	Rwanda	RG	Pool *et al.* 2012	RG2, RG3, RG4N, RG5, RG6N, RG7, RG8, RG9, RG10, RG11N, RG13N, RG15, RG18N, RG19, RG21N, RG22, RG24, RG25, RG28, RG2, RG32N, RG33, RG36, RG37N, RG38N.	-
Western Africa	Cameroon	CO	Pool *et al.* 2012	CO1, CO2, CO4N, CO8N, CO9N, CO10N, CO13N, CO14, CO15N, CO16.	-
South East Africa	Zambia	ZI-ZO	Pool *et al.* 2012, stocks donated by C. Langley	ZI91, ZI261, ZI268, ZI468, ZO12, ZO65.	ZI160, ZI173, ZI186, ZI261, ZI273, ZI403, ZI418, ZI468, ZI507, ZI514
	Zimbabwe (Lake Kariba)	ZK	Begun & Aquadro 1993	ZK84, ZK131, ZK186.	ZK 84, 95, ZK133, ZK145, ZK157, ZK186, ZK191, ZK229, ZK377, ZK384, ZK398
	Zimbabwe	ZS	Pool *et al.* 2012	ZS5, ZS11, ZS56.	-
	Malawi	MW	Pool *et al.* 2012	MW6, MW11, MW28, MW38, MW46, MW63.	-
South Africa	South Africa	SP	Pool *et al.* 2012	SP80, SP173, SP188, SP221, SP235, SP241, SP254.	-

The analyses we conducted in chapter 2 we done on sequence fragments with the Sanger method in our laboratory. Study patters of variation via summary statistics (see below) in the Netherlands and Zimbabwean Lake Kariba populations, within in a region of ~86 kb with coordinates (16,960,095 to 17,046,326). The exact location of fragments within this region as well as the information about their primers is provided in Appendix G. In Chapter 4, we used exclusively next generation sequence data of European populations the Netherlands and France as well as two African Rwanda and Southeast Africa pooling data from geographically close sampling sites. With these datasets we obtained summary statistics for the X chromosome regions under the QTL for CCRT uncovered by the deletion *Df(1)ED9606*; *i.e.*, a total of 124 kb between coordinates 7,089,000 and 7,212,999.

The following quality control steps during the initial handling of the sequence data were used: (i) nucleotides with a PHRED score lower than 21 were set to 'N'. Unless otherwise stated, this quality criterion was applied to all analyses in which DPGP2 sequence data were used. (ii) If a given polymorphic site in the alignment showed a frequency of N higher than 10% it was excluded from the analysis. The following summary statistics were then computed on 2-kb long non-overlapping windows (or in fragments of ~570 pb in length for Chapter 2): θ_π (Tajima 1983), θ_W (Watterson 1975), and divergence (D_{xy}) to the out-group (Nei 1987). Haplotype diversity *H* (Depaulis and Veuille 1998) and average allelic association estimates based on LD, (Z_{nS}) (Kelly 1997) In addition pairwise F_{ST} were calculated as normalized Nei (Nei and Li 1979) and Tajima's *D* (1989) and were also obtained with a summary statistic calculator written by P. Duchen.

4.1.2 Composite likelihood ratio test for positive selection

To investigate whether the observed SFS in the region of interest is compatible with the one expected after a selective sweep we calculated the composite likelihood ratio (CLR) statistic (Kim and Stephan 2002; Nielsen *et al.* 2005; Pavlidis *et al.* 2010) as it is implemented in the software SweeD (Pavlidis *et al.* 2013). This likelihood ratio test statistic compares a selective sweep model and a neutral model that is calibrated with the genomic background frequency spectrum. We used the parallel version of the software (SweeD-P) to calculate the CLR statistic along the complete X chromosome in our European samples (19 lines from the Netherlands and France in chapter 4 and 14 Swedish lines in chapter 5). In addition to the classes of the SFS (*i.e.* 1 to *n*-1, where *n* is the sample size),

we considered two additional site classes consisting of monomorphic sites in the European sample and polymorphic in the Rwandan sample. Extending the SFS in this way was shown to improve the power of the method to detect selective sweeps (Nielsen *et al.* 2005). SweeD was run on a 16-core CPU using the command line option "- -monomorphic" with 500,000 grid points. The background SFS was taken from the complete X chromosome. However, following Pool *et al.* (2012) we excluded from the analysis telomere and centromere regions of the X chromosome due to their very low recombination rate. The coordinates of the excluded regions range from the origin until position 2,222,391 for the telomere and from position 20,054,556 to the end for the centromere region. Finally we compared the CLR profile of our region of interest to the profile calculated for the complete chromosome.

The significance level of the CLR-test statistic was calculated by simulating large genomic regions with the coalescent simulator fastsimcoal2 (Excoffier and Foll 2011) under a neutral model that takes into account our current knowledge of the demography of European populations of *D. melanogaster* (Laurent *et al.* 2011). For every one of the simulated datasets we computed the CLR-test statistic in the same way as we did for the observed dataset and recorded the maximum CLR value. We used the 95th quantile of the distribution of top CLR values as our significance threshold. Since this analysis becomes computationally intensive as the size of the simulated genomic region increases, we investigated the relation between the threshold value and the size of the simulated region. We simulated batches of 100 datasets or increasing size from 50 to 5000 kb in length and took the asymptotic value reached as the chromosomal threshold (Appendix E).

4.1.3 LD-based test: the ω statistic

Of the selective sweep map of the X chromosome inferred by Li and Stephan (2006) that consisted of 54 and 55 putative 100-kb long fragments in the Zimbabwean and Dutch populations, respectively, where sweeps were identified, we investigated one region in detail (window 55 in the Netherlands, cytological position 15E). We chose this region because it fulfills necessary conditions for the occurrence of a complete sweep in the Netherlands and may be contributing to the reduced CCRT observed in this population: (i) the sweep co-localizes with a QTL, namely the QTL at position 56 cM (cytological

position 13E-20E, Figure 3), which is significant in both males and females and is lacking QTL–sex interactions; (ii) the sweep is specific to the Netherlands. Because the sweep is observed in this population, but not in Zimbabwe, allelic differences at the gene(s) affecting the trait may therefore be found.

Previously, Li and Stephan (2006) identified this sweep region in the Netherlands using the site frequency spectrum of SNPs averaged with three of 500-bp long fragments that span a 100-kb window. Here, we increased the amount of sequence data in window 55 in two steps as described by Svetec *et al.* (2009): first, 12 additional fragments of 500-bp length were PCR-amplified and re-sequenced in both the Netherlands and Zimbabwean samples used by Li and Stephan (2006) (all available ZK and NL stocks Table 8), and second, a region of 6.4 kb (between absolute positions 16,992,569 and 16,998,925; release 5.29 of Flybase, http://flybase.org) was completely re-sequenced in these 23 lines. This fine-scale analysis allowed us to determine the target of selection very precisely (*i.e.* down to the level of individual genes)

This data set was then subjected to an analysis of linkage disequilibrium (LD) using the w statistic (Kim and Nielsen 2004). Elevated values of w provide evidence of a selective sweep, and the peak of this statistic (ω_{MAX}) indicates the location of the target of selection in the genome (Pavlidis et al. 2010). Positions containing insertions or deletions (indels) were excluded. To assess the statistical significance of the maximum value $\boldsymbol{\omega}_{MAX}$, we ran 10,000 neutral simulations with the 'ms' software (Hudson 2002). The demographic scenario of the Netherlands (Li & Stephan 2006) was considered as the null hypothesis. The mutation rate ($1.47 \cdot 10^{-9}$) was estimated from the observed number of polymorphisms in the African population using the method of (Živković and Wiehe 2008). Thus, the African population was used as a proxy for selective neutrality. The recombination rate ($3.6 \cdot 10^{-8}$) was obtained from the *D. melanogaster* recombination rate calculator (Fiston-Lavier *et al.* 2010). Only the European subset of each simulation was used to assess the significance of $\boldsymbol{\omega}_{MAX}$.

4.1.4 F_{ST}-based scan for positive selection

For the set of F_{ST} analyses performed with BayeScan (Foll and Gaggiotti, 2008), input files were prepared following the author's instructions. The different runs were done using default parameters. Sequence data was obtained from the DPGP samples of the

Netherlands, France, Ethiopia, Cameroon, Rwanda, Southeast Africa and South Africa (Table 8). SNP exclusion criteria were as follows: positions showing more than two segregating alleles as well as sites with less than 50% base calls in one population were excluded from the analysis.

4.2 QUANTITATIVE GENETICS AND GENE EXPRESSION EXPERIMENTS

4.2.1 Mutant fly stocks

In order to conduct quantitative complementation tests on chromosomal deletions and P-element insertions targeted to disrupt candidate genes, a set of available deficiency/gene disruption lines were ordered at the Bloominton stock center. The set of tested X-linked deletions spans a chromosome fraction of 5.8 Mb, between coordinates 6,642,419 and 12,461,494 that correspond to cytological bands 6C to 11B. A total of 24 overlapping deletions with known breakpoints at the sequence level in 92% of the cases constituted the minimum number of available deletions covering the 5.8 Mb of interest. Additional deficiencies were tested if deemed necessary. In the cases in which gene disruption was tested the following P-element insertions were used: *P[EPgy2]CG1677*^{EY06475} disrupting gene *CG1677*, *P[SUPor-P]brk*^{KG08470} for *brinker* and *P[w[+mC],y[+mDint2]=EPgy2]unc-119*^{EY20221} for *unc-119* . Bloomington identification numbers of each line are provided in Appendix H.

Prior to CCRT scoring experiments, virgin female flies bearing the deficiency chromosome/P-element insertion and the respective balancer were mated with males of the A* and E* lines, respectively. Eggs were allowed to develop in the same medium in which they were laid at 23 ºC. Female F1 were sorted upon hatching by phenotype as balancer or deletion/P-element insertion bearer. Since all balancer types used to maintain the deletions have a dominant mutation for eye shape at locus *Bar* (B^1), F1 flies exhibiting the B^1 mutant phenotype were considered as balancer bearers, while wild type appearance was indicative of bearing the deletion/P-element insertion. Sorted flies were kept at same room temperature until CCRT scoring on their 4-6th day of life.

4.2.2 CCRT scoring

Once flies reached 4-6 days of age they were scored for CCRT following the protocol as in Svetec *et al* (2011). Briefly, flies were transferred to glass vials without anesthesia and placed in an ice-water bath of 0 °C for 7 hours. At the end of this time period flies were brought back to room temperature (23 °C) and observed in time intervals of 1 minute. The minute in which a fly was standing on its feet was recorded as its CCRT.

4.2.3 Quantitative complementation tests on deficiencies and P-element insertions

On average 35 female flies per each of the four resulting genotypes *E*/def, A*/def* (or *E*/mut, A*/mut,* for P-element insertions), *E*/bal,* and *A*/bal* were scored. For ANOVA analysis on log-transformed CCRT scores per genotype, line (*L*) and genomic background (*G*) were kept as fixed effects. We focused on the significance of the interactions of these two factors (*L x G)* as well as on the following two conditions to call the procedure quantitative failure to complement: (i) the differences in CCRT for the genotypes bearing the balancer had to be small or negligible compared with that of the genotypes bearing the deletion/P-element insertion. (ii) In the latter case the *E*/def* (or *E*/mut*) flies should show a reduced CCRT with respect to the *A*/def* (or *A*/mut*) genotypes. These conditions should be satisfied in order to control for false positives arising from epistatic interactions between alleles at loci other than the one under consideration. Bonferroni correction was applied to control for multiple testing.

4.2.4 Gene expression assays

Assessments of expression levels of candidate genes coupled with chill stress lines were conducted using 4-6 days old female flies belonging to the Netherlands population (isofemale lines: NL01, NL12, NL14, NL15, NL16, NL18, NL19, NL20) and the Zimbabwean population (isofemale lines: ZK84, ZK131, ZK145, ZK157, ZK186, ZK229, ZK377, ZK398). Flies were reared at 23 ºC and subjected to cold stress in the same manner as reported for CCRT scoring. For all treatments and controls we used three flies per line. The same amount of flies per line were snap frozen in liquid nitrogen at 10 minutes after being brought to room temperature while three remaining flies per line were scored for their CCRT and frozen 15 minutes after the minute in which they were reported as recovered. Control flies, which remained at 23 °C in glass vials during

the seven hours of treatment, were also snap frozen at the end of this time period. Frozen material was stored at -80 ºC until RNA extractions were performed. Population pools per line/treatment were made prior to RNA extraction. Each population pool per treatment consisted of eight flies of the same population. Three population pools per treatment were made for both the Netherlands and Zimbabwe.

In section 2.1, however, flies of both sexes were used to study the expression of genes *CG4991* and *CG16700* only at the constitutive level (*i.e.* flies were not cold-stressed). In addition, as reported in Appendix C, we conducted expression assays especially for gene *brk* were conducted with 4-6 days old female flies of the following lines NL14, NL20, NL19, ZI507, ZK84, ZK145, ZK157. Assessments coupled with cold stress were conducted as above, except that the number of flies used per line per treatment was 12. This allowed us making three pools per line per treatment with each pool of four flies each.

.

4.2.5 RNA extraction and cDNA synthesis

RNA was extracted from pools within line (see Appendix C) or population using the MasterPure RNA Purification Kit (Epicentre Biotechnologies, Madison, WI), followed by DNAse treatment. Purified RNA was quantified with a nanodrop apparatus and tested for genomic DNA contamination based on a PCR (Phusion) protocol using a primer pair binding in non-transcribed regions of the X chromosome. (Primer code: X-1435, see Appendix G for further details). Samples tested positive for genomic DNA were excluded from downstream protocols. cDNA synthesis was performed with SuperScript III Reverse Transcriptase (Invitrogen, Carlsbad, CA) on 1400 ng of RNA per reaction.

4.2.6 RT-qPCR and expression level analyses

RT-qPCR assays for candidate genes under cytological region 7A: *CG1958, CG1677, CG2059, unc-119, brk* and *Atg5* were done with primers designed using the online tool QuantPrime (www.quantprime.de) (Appendix G) to match all possible transcript types per candidate gene. For candidate genes at 15E: *CG16700* and *CG4991* QuantiTect Primer Assays*(Dm_CG16700_1_SG* and *Dm_CG4991_1_SG*, Qiagen Carlsbad, CA) were employed. The ribosomal genes *RpS20* and *RpL32* were taken as reference genes, against which relative gene expression levels of our genes of interest were normalized. RT-qPCR

assays consisted of a total of three biological replicates each run in triplicate and were conducted on a Real-Time thermal cycler CFX96 platform (BioRad, Hercules, CA). Each reaction was taken to a final volume of 10µl using iQ™ SYBR® Green Supermix (BioRad, Hercules, CA). Further details of the experimental setup, such as amplification efficiencies assessments with dilution series can be provided upon request.

Obtained Cq values per replicate within line (or pool) and treatment were transformed to calibrated normalized relative quantities (CNRQ) following Hellemans *et al.* (2007). Log transformed CNRQs were then used to test the hypothesis of expression differences between pairs of lines (or pool) within and between treatments. For this purpose Welch two-sample t-tests were performed on comparisons with fold differences above a threshold (defined by variance between technical replicates). Benjamini and Hochberg's (1995) *P*-value correction was applied to control for false positives, due to the high number of simultaneous tests performed.

4.2.7 Transcription factor binding site prediction

Genomic locations enriched with specific transcription factors (TF) at the adult stage in the 16.6 kb region upstream of *brk* were identified using the modENCODE database (Negre *et al.* 2011). The only TF that was found to be enriched in this region is Nejire (also called CREB-binding protein or CBP), which is known to interact physically with the DNA-binding TF cAMP response element binding (CREB). TFBS for CREB were then predicted in the E* and A* lines using a position-weight-matrix (PWM) approach implemented in the program Patser (Hertz and Stormo 1999). The PWM used to describe the binding motif of CREB was taken from FlyFactorSurvey (www.pgfe.umassmed.edu/ffs). CREB binding site profiles were compared between E* and A* to identify gain or loss of TFBS in these two lines. Finally, allelic frequencies of mutations that are responsible for modifications of CREB TFBS were calculated in our population sample.

4.2.8 Resequencing of the 16.6 kb upstream of gene *brk*

In order to catalogue all nucleotide and structural differences between the E* and A* lines that may be associated with the phenotype difference between these two fly lines, the entire 16.6 kb corresponding to the intergenic region upstream of *brk* were sequenced in

these two lines. Primers for PCR amplification of overlapping fragments of approximately 700 bp spanning the intergenic region plus the 5' UTR of *brk* were designed with the program Primer 3' using the sequence reference between coordinates 7,185,337 and 7,201,971 of *D. melanogaster* genome sequence release 5.53 (Flybase). Genomic DNA was prepared from single E*(NL14) and A*(ZK157) males, PCR amplified and sequenced with the Sanger method on both strands. Alignment of the two E* and A* sequences and identification of differences were done with the aid of the program Seaview (Gouy *et al.* 2010). All reported differences were also ascertained by inspection of the alignment.

The fractions of this 16.6 kb containing polymorphisms that can explain the difference in gene expression levels between Europe and Africa were likewise resequenced in the following populations the fly stocks Netherlands, Zimbabwe (Lake Kariba) and Zambia (ZI), (Table 8)

4.2.9 Linkage disequilibrium matrix

LD estimations for SNP pairs within the 16.6 kb upstream of *brinker* were obtained using Hill and Robertson's r^2 (Hill and Robertson 1968). LD informative SNPs were sites for which minor frequency alleles were above 10% and the site had less than 50% N. Significance of pairwise LD was determined using Fisher's exact test (Weir 1996). The pooled *D. melanogaster* set used for these LD estimations included lines from the following DPGP populations: NL=10, FR=3, RG=8, ZI=1, ZK=3 and the sequences of the E* and A* lines.

4.2.10 Prediction of the protein structure of *CG1677*

To gain insight into the functional significance of the detected amino acid substitutions at gene *CG1677*, we subjected the primary protein sequences of the lines A* and E* to a homology-based secondary structure prediction using the tool Phyre^2 (Kelley and Sternberg 2009) and conducted a Blast-based search for homologue sequences from invertebrate and vertebrate accessions at (www.uniprot.org) (Uniprot-Consortium 2014). In addition, we conducted an ancestral state reconstruction analysis (Lewis 2001) with all available Drosophilid sequences of this gene to reconstruct the sequence of mutational events behind the amino acid substations observed in our *D. melanogaster* populations.

4.3 SWEDISH FLIES COLLECTION AND SEQUENCING

4.3.1 Fly collection

The Swedish city of Umeå (63° 49' 30" N, 20° 15' 50" E) was visited between August 17th and 23rd 2012 to collect local *D. melanogaster*. The sampling was made following Pool's (2009) guidelines. In order to establish our sampling sites we took advantage of the grid-like layout of the city center to locate points where *D. melanogaster* was likely to be found (Figure 13). These included backdoors of public establishments such as cafés and restaurants, as well as protected sites with trash containers usually located near parking lots. Traps were prepared *in situ*, using empty plastic bottles of juice with banana inside as bait as suggested by Pool (2009). Once traps were located, they were checked twice a day and harvested the following day. Traps were emptied after having anaesthetized flies with FlyNap (Carolina Biological Supply. Burlington, NC). Females and males were separated and the former relocated in coded food vials (one individual per vial). By the end of the week, on August 24th, a total of 106 vials were kept in an incubator at the Centre for Evolutionary Biology at Uppsala University and shipped to our lab in Munich the following week.

4.3.2 Species diagnosis and isofemale lines

Of all possible Drosophilids that could have been sampled in Umeå, *D. melanogaster* and *D. simulans* are the most difficult to tell apart in the field. The shape of the male genital arch is the only reliable morphological diagnostic character (Davis *et al.* 1996) however; molecular diagnosis of the species status is an efficient and straightforward alternative. We designed a molecular diagnosis assay based on differences between enzyme digestion patterns for *D. melanogaster* and *D. simulans* at the mitochondrial gene Cytochrome C oxidase subunit I (*COI)*. With a set of appropriate primers (Appendix G) a fragment of *COI* gene was PCR amplified from DNA of daughters from collected females. We also obtained DNA from known *D. melanogaster* and *D. simulans* lines kept in our laboratory. The approximately 300 bp long amplicons were digested with endonucleases *MnII* and *AciI* (New England Biolabs, Frankfurt, Germany). Within the amplicon, the target sequences of these restriction enzymes include fixed differences between the two species. *D. melanogaster* amplicons were digested by *MnII* generating two fragments of 110 and 200

bp, respectively. In the presence of the enzyme *AciI*, only the *D. simulans* versions of the amplicon were digested, generating two fragments of similar sizes. Digested products were run on 1.5% agarose gels and inspected to assign species status of the lines.

In addition to this molecular species diagnosis we set up crosses between female virgin daughters from collected females with *D. simulans* males. After mating females were allowed to lay eggs for two days then all adults were removed. The content of each vial was monitored for the presence of adults of both sexes. The absence of male offspring during within each vial was taken as an indication that mothers were *D. melanogaster* (Davis *et al.* 1996). Once the species status of the collection was determined, the inbreeding scheme was initiated. Full sibling-inbreeding scheme was initiated with daughters and collected females. This inbreeding scheme was maintained for 10 generations.

4.3.3 Haploid embryo preparation

Aware of the complications caused by residual heterozygosity in population genetics analyses of full genome sequencing datasets obtained from DNA of inbreed lines we prepared chosen line's genomic material following Langley *et al* (2011). This protocol requires the generation of haploid embryos from which genomic DNA is obtained and prepared for sequencing. Haploid embryos were derived from crosses of virgin female flies belonging to 20 chosen Umeå isofemale lines with mutant males homozygous for the mutation *ms(3)K81*^{1} (Bloomington stock number 5252). This mutation was isolated from nature by Fuyama (1984), and genetically characterized by (Yasuda *et al.* 1995). The 710 bp long locus *ms(3)K81*, mapped to 97D4 encodes a paternal-specific protein essential for the development of the male pronucleus before the first zygotic division. The sperm of males homozygous for the mutation ms*(3)K81*^{1} has normal motility and is able to enter the egg at fertilization. However, eggs fertilized by bearers of this mutation undergo arrested embryogenesis. Around 75% of these embryos stop development after several nuclear divisions, while the remainder develops beyond blastoderm but does not hatch as larvae. These late-arrestment fraction of the embryos has haploid karyotypes. The haploid genome that is present within these embryos is derived from the maternal pronucleus (Yasuda *et al.* 1995).

A total of 20 isofemale lines were chosen for whole genome sequencing using Illumina

technology (Figure 13C, Table 5). This set of 20 lines comprises a representative sample of the Umeå collection. We chose these lines so that every sampling location would be fairly represented in the final sequence set, thereby lowering the chance of obtaining population genetics information on a single family. Haploid embryos per line were obtained as follows. Four virgin female files of each chosen isofemale line were allowed to mate with an equal number of males homozygous for the *ms(3)K81*^{1} genotype, keeping each mating pair in its own vial for 24 hours. Mated females were then transferred to molasses agar plates, as oviposition substrate. Plates were sealed and kept in the incubator at 26 ºC for another 24 hours. Subsequently, females were transferred back to the mating vials for 1.5 weeks to monitor presence of larvae. Have these been observed it would have indicated that females were not virgin or that the mutant male phenotype was misclassified. Once 24 hours elapsed, all eggs in the agar plates were harvested and dechorionated under the stereomicroscope. All dechorionated eggs were screened for signs of development beyond gastrula stage (Langley *et al.* 2011). Embryos that met this criterion were individually transferred to sterile 200 ml PCR reactions tubes with 3 µl of 1 % PBS and frozen at -80 ºC, until next use

Lysis of embryos is a required first step in the process of whole genome amplification with the QIAGEN REPLI-g Midi kit (QIAGEN, Hilden, Germany). However, before lysates were further processed we took a 2 µl aliquot, and preserved the reminder at -30 ºC. Each aliquot was diluted in water and employed to assess the relative proportions of nuclear, mitochondrial and *Wolbachia sp.* DNA within each embryo. This step was necessary to double check the screening of embryos and maximize the yield of nuclear DNA sequence reads. With a standard qPCR assay, we compared Cq values of each DNA target and from there we inferred their respective proportions per embryo. Each qPCR reaction was taken to a final volume of 10µl using iQ™ SYBR® Green Supermix (BioRad, Hercules, CA), the corresponding primer pairs for each DNA type (Appendix G) and the diluted embryo lysate. Reactions were run on a Real-Time thermal cycler CFX96 platform (BioRad, Hercules, CA). Annealing temperatures of the three primer pairs were optimized so that the three assays per lysate could be run simultaneously. Additional details are provided in Appendix I. This protocol was modified from that kindly provided by C. Cardeno at C. Langley's laboratory at UC Davis.

Embryo lysates in which we detected mitochondrial and nuclear DNA within expected Cq-value ratios while exhibiting the lowest content of *Wolbachia sp.* DNA (see Appendix I)

were used as starting material for whole genome amplification. This step was done using the QIAGEN REPLI-g Midi kit (QIAGEN, Hilden, Germany) with the 8 µl of embryo lysates that were spared at -30. After protocol completion, 3 µl of a 1:100 dilution from the amplified material were run on a 1% agarose gel to visualize the amplification product. This should be seen as an intense band above 10 kb.

4.3.4 Illumina sequencing and assembly of reads

We determined DNA concentration of the randomly amplified DNA from haploid embryos using the Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE) with 1 µl of a 1:100 dilution of amplified DNA in molecular grade water. Up to 10 µg (~ 200 ng/µl) of DNA, were employed to construct standard genomic libraries for paired-end sequencing with Illumina technology. Both library construction and sequencing took place at GATC Biotech (Konstanz, Germany). Sequencing of the resulting 20 libraries was performed on an Illumina HiSeq 2000 machine. The sequencing method was paired end with reads of 100 bp in length.

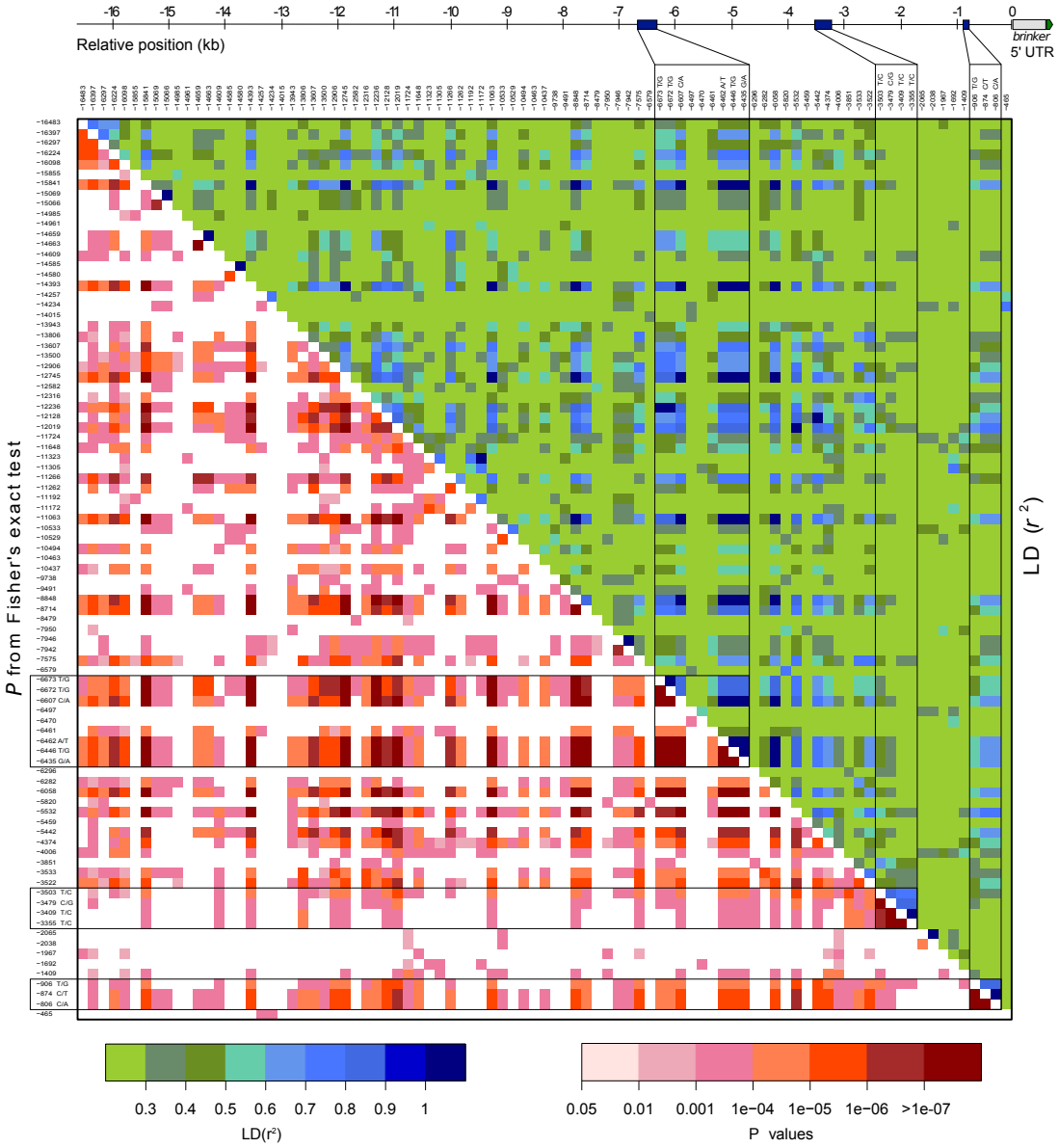
Paired end reads were aligned to the *D. melanogaster* reference genome (Release 5 of the Berkley *Drosophila* Genome Project). For this purpose we used the program BWA version 0.59 (Li *et al.* 2008) with default settings and the "-I" flag. Reads with BWA mapping quality scores less than 20 were excluded from the assembly. This measure reduces coverage biases due to ambiguously mapping reads. Once alignment to the reference genome was completed we used the program SAMtools version 0.1.16 (Li *et al.* 2009), to generate the consensus sequence of each assembly. All putatively heterozygous sites were masked to 'N', as well as sites within 10 bp of a consensus structural variant (*i.e.* indels). Only the euchromatic fraction of *D. melanogaster* major chromosomes (X, 2L, 2R, 3L and 3R) was considered after assembly. The resulting fastq files were subsequently deposited in the next generation sequence server of our department. When retrieving any required genomic fragment or full chromosome a quality threshold (PHRED score) has to be provided. Nucleotide positions with PHRED scores lower than the specified value is automatically masked to 'N'.

APPENDIX A

Population	FIN	*x*	*L*	S	θW	θπ	*D*	*H*	Zns	D*xy*
the Netherlands	1282	265	516	9	0.007	0.010	1.447	0.933	0.486	0.084
	406	8337	671	13	0.032	0.003	0.669	0.864	0.446	0.056
	1283	29123	539	9	0.006	0.004	-1.459	0.455	0.553	0.061
	1285	32927	456	4	0.003	0.004	0.828	0.409	1.000	0.048
	1327	33527	599	2	0.001	0.001	-0.382	0.530	0.000	0.036
	407	34191	618	0	0.000	0.000	NA	0.000	NA	0.099
	1328	34936	644	1	0.001	0.000	-1.141	0.167	0.000	0.055
	1287	35987	656	11	0.006	0.005	-0.561	0.591	0.508	0.080
	1296	42320	565	4	0.002	0.002	-1.322	0.491	0.252	0.131
	1323	46762	464	7	0.005	0.005	-0.106	0.709	0.250	0.056
	1297	57832	646	10	0.006	0.005	-0.857	0.864	0.264	0.042
	1298	58892	538	9	0.006	0.007	0.502	0.697	0.412	0.023
	1288	66711	574	2	0.001	0.002	1.176	0.564	0.656	NA
	1299	85980	502	5	0.003	0.003	-0.313	0.818	0.189	0.035
Zimbabwe	1282	265	516	29	0.021	0.018	-0.735	1.000	0.156	0.082
	406	8337	671	23	0.012	0.000	0.162	0.985	0.169	0.062
	1283	29123	539	15	0.009	0.011	0.552	0.985	0.242	0.063
	1285	32927	456	23	0.020	0.018	-0.366	1.000	0.231	0.053
	1327	33527	599	21	0.014	0.011	-0.847	1.000	0.177	0.032
	407	34191	618	33	0.020	0.016	-0.900	1.000	0.244	0.092
	1328	34936	644	18	0.010	0.010	0.133	1.000	0.239	0.052
	1287	35987	656	32	0.018	0.017	-0.321	0.985	0.183	0.077
	1296	42320	565	23	0.015	0.013	-0.359	0.970	0.364	0.135
	1323	46762	464	11	0.009	0.007	-0.970	0.945	0.142	0.055
	1297	57832	646	18	0.012	0.011	-0.603	0.985	0.144	0.010
	1298	58892	538	13	0.008	0.008	-0.076	0.939	0.190	0.023
	1288	66711	574	25	0.017	0.014	-0.949	0.972	0.275	NA
	1299	85980	502	5	0.004	0.003	-0.229	0.844	0.188	0.033

Selected summary statistics table for the sweep region at 15E
FIN is the fragment name. *x* is the relative position of the mid point of the fragment (used also in the x axis in Figure 1A). (*L)* Fragment length, (S) number of segregating sites per fragment, (θ_W) Watterson's estimator of genetic diversity, (θπ). (*D)* Tajima's *D.* (H) haplotype diversity index, Z_{nS} average LD per fragment.

APPENDIX B



Linkage disequilibrium (LD) matrix of 89 informative SNPs in 16.6 kb upstream of *brinker* (*brk)* in a panel of 27 *D. melanogaster* lines. All positions are numbered with respect to the origin of *brk*'s 5' UTR at 7,201,972 as position zero. Patterns of LD (r^2) are shown above the diagonal and *P* values from Fisher's exact test below the diagonal. LD block 1 is proximal to the transcription start of *brk*. It is formed by three SNPs at relative positions *-806A/C*, *-874T/C* and *-906G/T*, and spans a total of 100 bp. The second LD block spans a total of 150 bp and is made up of four SNPs: *-3,503T/C*, *-3,479C/G*, *-3,409T/C*, and *-3,355T/C*. The third block encompasses 240 bp and includes six SNPs distributed in two groups of consecutive sites in LD. SNPs *-6,673G/T*, *-6,672G/T* and *-6,607C/A* form the first sub-block to the left and SNPs *-6,462A/T*, *-6,446T/A* and *-6,435G/A* constitute the second sub-block to the right.

APPENDIX C

Results gene expression with *brk* based on defined haplotypes

To test whether the observed pattern of deletion frequencies may partly explain the expression difference for *brk* between the Netherlands and Zimbabwe (Figure AC1), we hypothesized that an increased number of deletions may result in lower DNA-transcription factor interaction and therefore reduce the amount of *brk* transcripts in comparison with sequences in which deletions are absent. To evaluate this prediction we conducted a new round of expression assays for *brk* in lines, from each continent, harboring the haplotypes without deletions (ND), the haplotype with the three deletions (I-II-III), as well as the intermediate haplotypes (deletions I-II) and (deletion III) (Figure AC1 A).

The observed constitutive expression levels and also those 15 minutes after recovery showed a trend towards lower average expression with increasing number of deletions in the African lines, which is consistent with the above hypothesis (Figure AC1 B). However, in the European sample (in which the haplotype with deletion III is absent) this trend was not observed. The expression level was roughly the same regardless of the number of deletions.

The emerging picture from these results suggests that *brk* expression levels and its potential role in cold tolerance are population dependent, being tuned by the interaction of putative *cis*-regulatory elements, such that we defined here, and not yet identified *trans*-acting factors. These results highlight the prominent role of gene-by-gene interactions in both transcriptional variance and complex phenotypes. Functional work is needed to analyze these hypotheses in greater detail.

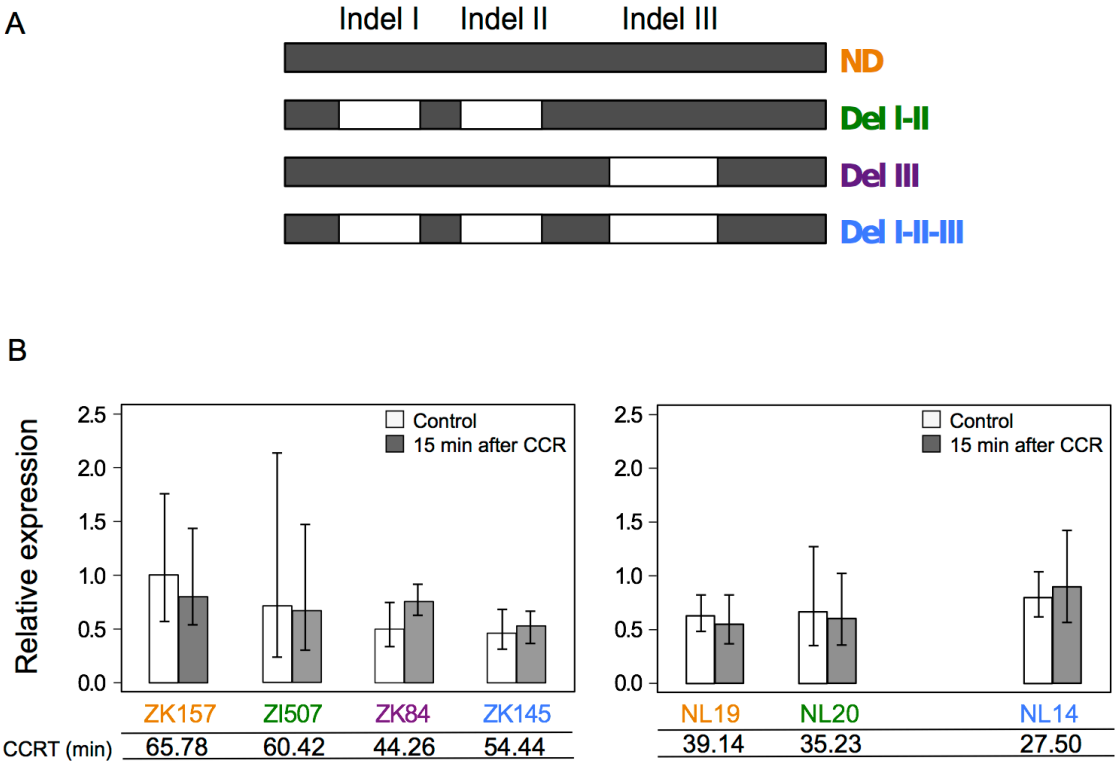
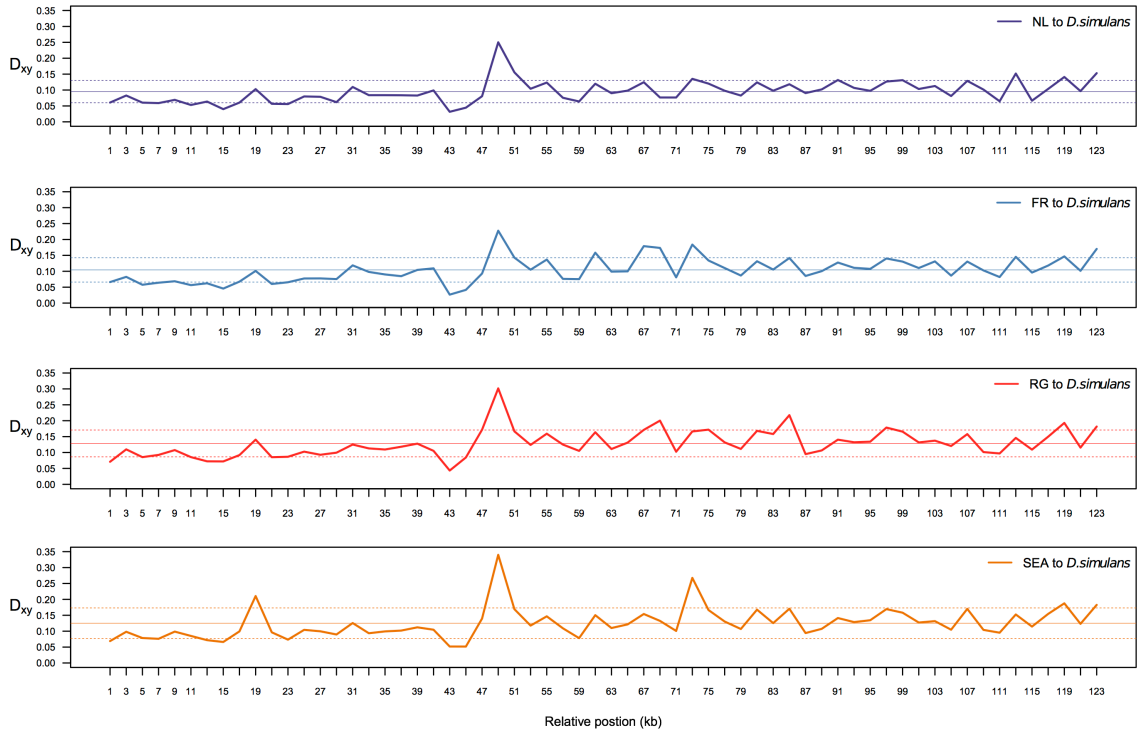


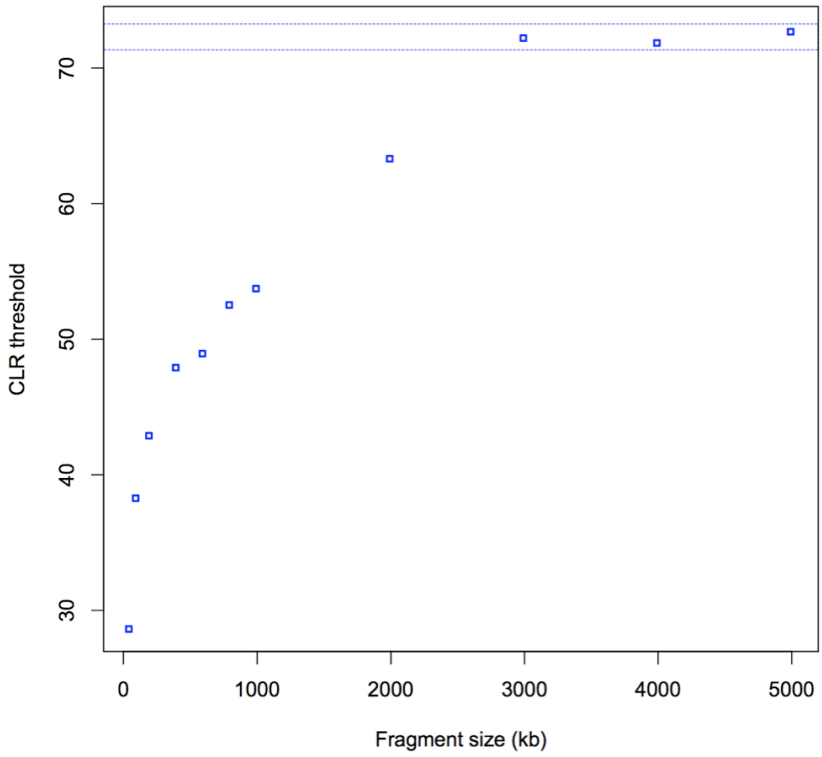
Figure 1 of Appendix C (AC1). A) Line-specific *brk* expression profiles. European and African lines were chosen based on the defined haplotypes (notice matching colors of the labels). Gene expression assays were done per line in control and cold stress conditions (15 minutes after recovery from chill coma). Average CCRT among replicates is reported below each line label. Expression levels of brinker were normalized with respect to that of ribosomal genes *RpS20* and *RpL32*. The height of the bars represents the mean of three calibrated normalized relative quantities (CNRQ) per line rescaled to that of the corresponding ZK157 control. Error bars also represent rescaled confidence intervals.

APPENDIX D



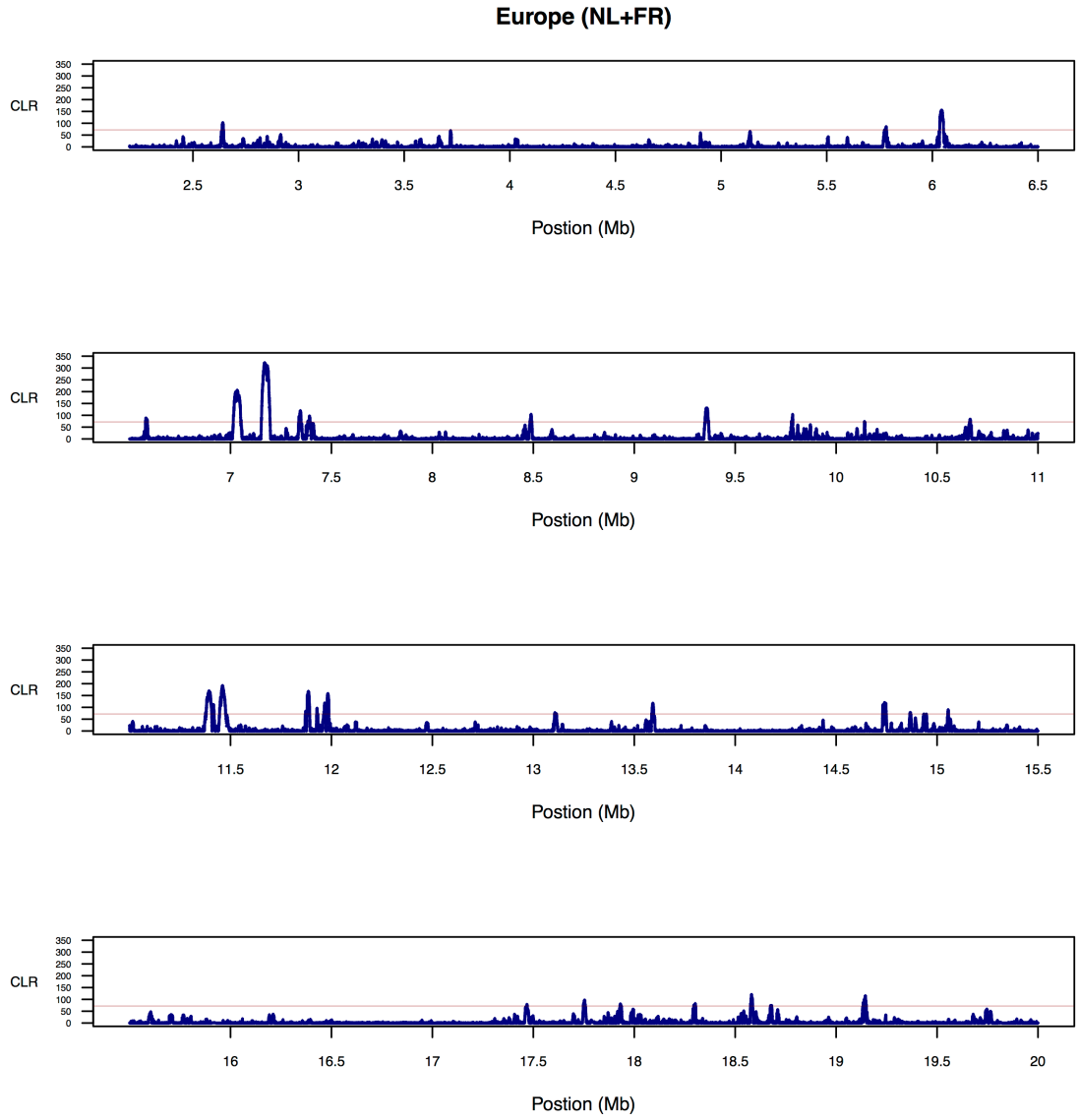
Divergence (D*xy*) along the 124 kb of interest calculated for each studied population: the Netherlands (NL), France (FR), Rwanda (RG) and a pool of Southeast African (SEA) lines with respect to an homologous sequence of *D. simulans*. Calculations were done for 62 consecutive windows of 2 kb each.

APPENDIX E



Composite likelihood ratio (CLR) thresholds *vs.* simulated fragment size. CLR thresholds, *i.e.* the top 5% CLR values of 100 simulated fragments of lengths from 5 to 5000 kb reach an asymptotic value around 72 at fragment size ≥3000 kb.

APPENDIX F



Composite likelihood ratio (CLR) test results for 18 Mb of a sample of 19 European (the Netherlands and French) *D. melanogaster* X chromosomes. For this chromosome-wide test all categories (0 to *n*) of the SFS were included. Significance threshold at CLR 72 was obtained from simulated subgenomic datasets (see text and Figure 8.3.2). The CLR peak above 300 between positions 7.0 and 7.5 Mb corresponds to that depicted in detail in Figure 11.

APPENDIX G

List of all primers employed in this dissertation

	Chr.	Rel. pos1	Abs. pos1	AL (bp)	Forward (5'-3')	Reverse (5'-3')	AT (°C)	Name
Primers for Sanger sequencing								
Enhancer region *brk*	X	67	7,185,066	594	CCCTTTTCGGTTCGCGTCC	GGCGAGTGACGGGGGTATATG	54.7	X-01475
	X	523	7,185,522	797	CGAAAGTAAGCGAACGTTTGC	TGGCCCCTGAAACTTAACACT	54.1	X-01476
	X	1,177	7,186,176	693	AGTGCGAGACGAAAGTACTTG	CTGGGCGAATAAAATTATCCT	53.7	X-01477
	X	1,689	7,186,688	769	TATCAGCGCAAATGCCATAGG	CGGCGCGCATAACGTC	54.2	X-01478
	X	2,273	7,187,272	684	TGCATCTAATAAATGGGATCT	CTGCCCACTGTTGTTGTA	52.8	X-01479
	X	2,768	7,187,767	765	AACTGCGTGCACTCTATGAA	CTGGCACACCTACGTTCTTAG	54.2	X-01480
	X	3,214	7,188,213	712	TTGGCCTGCCTTTATACATGG	CACTCCGGGACTTGAGCAC	54.6	X-01481
	X	3,723	7,188,722	751	AAAACCGAAACTCAAACTCTA	TCCCCACCTCTTAAAGC	51.8	X-01482
	X	4,248	7,189,247	796	TTGTTCACCTCCTCTCGAC	CGCACCCGAAATACTAATTAT	53.2	X-01483
	X	4,888	7,189,887	715	TCTGCCTTTCTCTTTCGTATA	CGTGTGGCAATTGTAGTG	53.2	X-01484
	X	5,295	7,190,294	795	GGGGGTGCGTATCTAATGAGG	CCGTCCGGCTAATTGTAATCG	54.0	X-01485
	X	5,895	7,190,894	692	TGGTCGGAAAATACCTGC	AGCCCATTTTGATCTTAACAG	54.9	X-01486
	X	6,359	7,191,358	767	TTAGGACCGATACAGGTTGA	CCATCCCAGTTAATTGTGTAA	52.4	X-01487
	X	6,808	7,191,807	682	GGGCTTTCGTTGCACAACT	ATTCCCCACTGCACTACCCTC	53.3	X-01488
	X	7,223	7,192,222	666	AAAACCGAATCGATATGG	ATGCCCTATTTGTTAACACAT	50.1	X-01489
	X	7,643	7,192,642	773	GCCATCGCGAATACACA	ATGCCACCCCATGTATATACC	51.6	X-01490
	X	8,222	7,193,221	751	CCCTTAAGTGGAAAAGTACCG	TTCATCCACATTCATCGCT	54.4	X-01491
	X	8,769	7,193,768	689	TGGCGGCGTAGCAAGA	GCTGCACACAGGAAATAGCTG	57.5	X-01492
	X	9,301	7,194,300	753	GCAGACGCAGACATAGAGG	CTCGCTCTCACTCTATGCATT	56.1	X-01493
	X	9,889	7,194,888	673	GAGCACGAGTGCGAGAGT	ATTGCAAGCATTACGAAATTC	56.1	X-01494
	X	10,369	7,195,368	672	CGCCAAGAAAATATAACCGTT	CGGTTGCACTTTTCGGAC	53.2	X-01495
	X	10,881	7,195,880	756	GGGGCTCTATGGATGCATGT	AGGGCTGCACTGTGATAAACA	55.6	X-01496
	X	11,499	7,196,498	788	TATGCGTACAAAAATAGATAC	TTTTCGAAGTACTGGTG	52.3	X-01497
	X	12,087	7,197,086	788	CTTGCGAGTCGATTCTT	TTGAAGGCCTAAATGATATAG	51.0	X-01498
	X	12,611	7,197,610	672	GGCAGCTAGACTTTCCAATAG	GCCAATAAAAGTCGATTCTG	52.6	X-01499
	X	13,012	7,198,011	717	CCCATAAAGATACCGAAGT	GGATACCGAAAACTATACCAA	51.0	X-01500(*)
	X	13,517	7,198,516	734	TTGGGAAGATAGAAGACTGTC	CCGTCGGCTTCTAAATC	52.1	X-01501(*)
	X	14,024	7,199,023	634	TTTTCCTTTTTCGTCGG	AAGGAGCTCTAAAATCGGTAA	49.2	X-01502(*)
	X	14,443	7,199,442	697	GGTTTTATCTTTGGCGTCTTA	CTTCGCTTGCGTGTGTA	51.5	X-01503
	X	14,903	7,199,902	791	CTGAACCCCAGATATCTATGT	AACGAACTTGTTGAGCG	51.9	X-01504
	X	15,504	7,200,503	657	CCCACGGCCATAACAA	TAAACCCAAAGTCGACAAATG	55.7	X-01505(**)
	X	15,941	7,200,940	605	AGAGCCAGCGAATTTGTAGCG	AATGGACAACATGGGACGGAG	54.3	X-01506(**)
	X	16,227	7,201,226	738	CGCGTTTCACTTATTGACTC	GGCGTTCTAGTTCGAAGATAC	55.2	X-01507
	X	16,789	7,201,788	733	CGCACACGAACACCCA	GCTGCTGCTATCCATGATTAT	54.2	X-01508
	X	17,205	7,202,204	505	CCAAAAAGTATCAGAGCGAGT	CCCTTGCAATCATTATCATTC	53.6	X-01509
Selective sweep at 15E	X	1	16,960,095	516	CAGCCTCAAAGTGCATTCAA	GTTTCGAAAACGCCGAGTAA	60.0	X-01282
	X	28,836	16,988,930	539	GATTCACTTGGACCCCTTCA	AAGTGGCCTTTCGATGACAC	60.0	X-01283
	X	32,723	16,992,817	456	CTCTGCGCTTCTCAATCTAT	ATTTGCCAGATCTTAACGAA	51.0	X-01285
	X	33,180	16,993,274	599	CATGCCACAAGAGGAATCG	AATGTTGGCATCGTAACGC	57.0	X-01327
	X	34,544	16,994,638	644	AACCACTGATCCATGACGG	CACGGCAAGAACTCAAACC	58.8	X-01328
	X	35,583	16,995,677	656	AGAGCGCGACAGTAATCAGT	TCTCTCGTCGGCTCTAGAAG	50.7	X-01287
	X	42,007	17,002,101	565	TCGACGAGAACTGTGACAC	TTGTTGGCTGTCTTTAAGGT	50.0	X-01296
	X	46,550	17,006,644	464	CCCATAACCATAAACGCAC	CGAGCGCGTATCTAACCT	56.4	X-01323
	X	57,438	17,017,532	646	GGCCGAATACGTTATGAACC	CAACAGCCACAGTGATACGA	53.9	X-01297
	X	58,606	17,018,700	538	GCGGCTTTATAGTTGACC	CGCTCGATTGAATACCAG	51.2	X-01298
	X	66,389	17,026,483	574	TGGATTTCATCGTCATCCAG	CAGTAAATCCGCCTTTTCCA	59.0	X-01288
	X	85,730	17,045,824	502	CCTGCGTTCGAGACATTC	AGTCCCGGGAAAGTCTTG	51.7	X-01299
Primers for qPCR								
	Chr.	Rel. pos1	Abs. pos1	AL (bp)	Forward (5'-3')	Reverse (5'-3')	AT (°C)	Name
Gene expression assays	X	-	-	79	AATGCCCATACATCGCGAGAGC	TTGCATCGGTGTGTTGTTGTGG	63.0	*CG1958*
	X	-	-	80	GCCAGCACCCAAGAAGTTTGAC	ACGACGTATCCGATGAGTCTGAG	62.0	*CG1677*
	X	-	-	77	ACGCGATCACCTTACATGAGACG	AGCCAGTGGATCTCAGAATTGGG	63.0	*CG2059*
	X	-	-	115	AACCTTCCACCCGACCTAGTTG	AGGCGTAGTCGGCTTTGTTGTG	64.0	*unc-119*
	X	-	-	65	TGCGAGGACATCATCCGTCAAC	TCAGGTTTGTGGGCGCAGTATC	63.0	*brk*
	X	-	-	144	CTCGTCAAGCTCAACTCCAAGG	GTTGACCAATCCCAGCCAAAGC	62.0	*Atg5*
	3R	-	-	76	TTCGCATCACCACCCGTAAGAC	TTGTGGATTCTCATCTGGAAGCG	62.0	*RpS20*
	3R	-	-	77	ATCGTGAAGAAGCGCACCAAGC	TTGCGCCATTTGTGCGACAG	62.0	*RpL32*
Embryos	X	-	2,034,560	711	TGCGAAACAGGTACAAGT	GGATTCGTGAACGGGAAA	50.0	X-01435(§)
	3R	-	21,993,006	246	TCTGACCCACTCTCCACTTG	TCGAATGATATACGAAGCGTTTAC	58.0	Nuc
	Mit	-	8,559	374	CTTCGACTTCCAAGACGTTC	CCTAAAGCTCATGTTGAAGCTC	58.0	Mit
	Wol	-	350,752	350	CCATATGTTGGTATTGGTGCAG	ATTCAACACGTGCAGTTTCATC	58.0	Wol
Primer for *D. melanogaster* species diagnosis								
	Mit	1	1,588	311	TTAGGACATCCTGGAGCA	TCAACTGAAGCTCCA	55.0	*COI*

Appendix G. Primer information for resequencing of 16.6 kb upstream of *brk*

AL is the length of the amplicon in base pairs
TA is annealing temperature
Primer pairs with (*) were used to resequenced the fragment of interest between relative positions -3,000 to -3,553 or LD block 2 (Figure 7B)
Primer pairs with (**) were used to resequenced the fragment of interest between relative positions -784 and -1,243 or LD block 1 (Figure 7A)
Primer pair (§) was used to check for the quality of DNA digestions during RNA purification protocol.

APPENDIX H

List of all mutant lines from Bloomington (www.flystocks.bio.indiana.edu)

Type of mutant	Name	Balancer	Bloomington stock number
Deficiency	*Df(1)BSC351*	*FM7h*	24375
	Df(1)BSC882	*FM7h*	30587
	Df(1)HA32	*FM7c*	947
	Df(1)ED6906	*FM7h*	8955
	Df(1)BSC711	*FM7h*	26563
	Df(1)BSC536	*FM7h*	25064
	Df(1)BSC622	*Binsinscy*	25697
	Df(1)C128	*FM6*	949
	Df(1)BSC866	*Binsinscy*	29989
	Df(1)BSC662	*Binsinscy*	26514
	Df(1)BSC592	*Binsinscy*	25426
	Df(1)Exel6241	*FM7c*	7715
	Df(1)ED6957	*FM7j*	8033
	Df(1)BSC537	*FM7h*	25065
	Df(1)BSC712	*FM7j*	26564
	Df(1)BSC539	*Binsinscy*	25067
	Df(1)ED7005	*FM7h*	9153
	Df(1)BSC755	*Binsinscy*	26853
	Df(1)BSC540	*FM7h*	25068
	Df(1)BSC572	*FM7h*	25391
	Df(1)BSC287	*Binsinscy*	23672
	Df(1)ED7067	*FM7h*	9154
	Df(1)Exel6242	*FM7c*	7716
	Df(1)ED7147	*FM7h*	9171
	Df(1)BSC543	*FM7h*	25071
	Df(1)ED7153	*FM7h*	8953
P-element insertion	*P{SUPor-P}brkKG08470*	*FM7c*	14953
	P{EPgy2}CG1677EY06475	*FM7a*	17545
	P{EPgy2}unc-119EY20221	-	22375
Mutation	*ms(3)K81[1]*	*TM3, Sb[1] Ser[1]*	5352

APPENDIX I

Protocol Haploid Embryo

Lysing Embryo

(Qiagen REPLI-g reagents)

1. Thaw Qiagen reagents (REPLI-g Midi, 150045) – PBS, Buffer D2 (denaturation buffer), and Stop Solution.

- Prepare sufficient Buffer D2; 3.5 μl per embryo
    - D2 = 5 μl DTT + 55 μl Reconstituted Buffer DLB

Use microscope for steps 3-6.

3. Add 3.5 μl of Buffer D2.
6. Crush/Pop/Grind-up embryo with pipet tip.

- TIP: Try putting embryo into a bubble of D2+PBS (6.5 μl total) on side of tube.

7. Vortex/spin-down and incubate on ice for 10 minutes.

- If lysing more than one embryo, make sure that each embryo is on ice for only 10 minutes.

8. Add 3.5 μl Stop Solution. Vortex/spin-down.

9. Dilute the lysed embryo 1:10 in for a total volume of 20 μl:

- 2 μl of lysed embryo + 18 μl of water = 20 μl total volume

10. Use the 1:10 dilution for qPCR assay. Return the lysed embryo to the -20°C.

qPCR

(Real-Time thermal cycler CFX96, BioRad)

Things to note:

- BioRad = 96-well plate; rows = A-H; columns = 1-12
- 3 assays per plate = mtDNA, nucDNA, *Wolbachia* DNA
- Each embryo in duplicate per assay

- 10 μl reaction per well
-

1. Make three master mixes:
   1) mtDNA primers
   2) nucDNA primers
   3) *Wolbachia* primers

- mtDNA primers (ND5) – 374 bp amplicon
  CTTCGACTTCCAAGACGTTC
  CCTAAAGCTCATGTTGAAGCTC
- nucDNA primers (3R)K – 246 bp amplicon
  TCTGACCCACTCTCCACTTG
  TCGAATGATATACGAAGCGTTTAC
- *Wolbachia* primers (Wsp_qp_1L and Wsp_qp_1R) 350 bp amplicon
  CCATATGTTGGTATTGGTGCAG
  ATTCAACACGTGCAGTTTCATC

Master mix preparation table for a reaction volume of 10 μl:

	1 reaction	3 reactions	12 (+2) reactions	36 (+3) reactions
l PCR-grade water	3.1	9.3	43.4	120.9
iQ SYBR-Green super mix (2x)	5	15	70	195
10 μM forward primer	0.2	0.6	2.8	7.8
10 μM reverse primer	0.2	0.6	2.8	7.8

2. After adding 8.5 μl of master mix to wells, add 1.5 μl of 1:10 embryo lysate per well.

- MAKE SURE THAT THE SEAL IS ON SECURELY ON THE PLATE!! (Evaporation of reaction will give false results)
- Use Figure 1 for set-up.

3. Thermal cycling parameters for program "Embryo test"

a) 95°C for 5 min
b) 95°C for 30 sec
c) 56°C for 30 sec
d) 72°C for 45 sec; data acquisition at this step
   set steps b-d for 40 cycles
e) 72°C for 10 min

4. Set program “Embryo test” in BioRad CFX96

Fill in well information with the corresponding data of embryo number and line as well as DNA type as target (follow example in Figure AI1)

5. Run the program (Ask for help if doing it for the first time)

	1	2	3	4	5	6	7	8	9	10	11	12
A	SU05-1	SU05-1	SU05-1	SU05-2	SU05-2	SU05-2	SU05-3	SU05-3	SU05-3	SU05-4	SU05-4	SU05-4
B	SU05-1	SU05-1	SU05-1	SU05-2	SU05-2	SU05-2	SU05-3	SU05-3	SU05-3	SU05-4	SU05-4	SU05-4
C	SU05-5	SU05-5	SU05-5	SU11-1	SU11-1	SU11-1	SU11-2	SU11-2	SU11-2	SU11-3	SU11-3	SU11-3
D	SU05-5	SU05-5	SU05-5	SU11-1	SU11-1	SU11-1	SU11-2	SU11-2	SU11-2	SU11-3	SU11-3	SU11-3
E	SU11-4	SU11-4	SU11-4	SU11-5	SU11-5	SU11-5	SU18-1	SU18-1	SU18-1	SU18-2	SU18-2	SU18-2
F	SU11-4	SU11-4	SU11-4	SU11-5	SU11-5	SU11-5	SU18-1	SU18-1	SU18-1	SU18-2	SU18-2	SU18-2
G	SU18-3	SU18-3	SU18-3	SU18-4	SU18-4	SU18-4	SU18-5	SU18-5	SU18-5	-	-	-
H	SU18-3	SU18-3	SU18-3	SU18-4	SU18-4	SU18-4	SU18-5	SU18-5	SU18-5	-	-	-

Color code

MIT

NUC

WOL

Figure AI1. qPCR 96-well plate example. Note the embryo names containing the line and embryo number

Checking embryo quality

5. Once the program has finished, save results.

6. Use lab Mac to fill in Excel sheet “Embryo qPCR results”. (See Figure AI2). Make sure you have the following information from each of the duplicate runs per embryo per target DNA:.

- Mean Cq $_{\text{embryo/DNA type}}$ = average Cq of individual embryo lysate for locus
- Cq sd = (standard deviation) of individual embryo lysate for locus
- Nuc/Mito = (mean Cq nuc)/(mean Cq mito) of individual embryo lysate
- Nuc/Wol = (mean Cq nuc)/(mean Cq wol) of individual embryo lysate

7. Embryo lysates with Cq value ratios (Nuc/Mito ≤ 2.5) and (Nuc/Wol < 1.2, ideally less than 1.0) will be further prepared for sequencing.

Important considerations for embryo selection include:

- Cq values of Mitochondrial and Nuclear DNA should be lower than 35. Best eggs show Cq values below cycle 30. As can be seen in Figure AI2 water is used to determine the cycles in which there is primer dimer signal (this occurs around cycle 40)
- *Wolbachia* may not be present in the tested embryo; therefore may not be detected at all. As long as Mitochondrial and Nuclear DNA are detected and their ratios match the expected values, the embryo can be used for sequencing.

KEEP the original embryo lysates. Throw-out original embryo lysates for those did not meet quality criteria; these are failures.

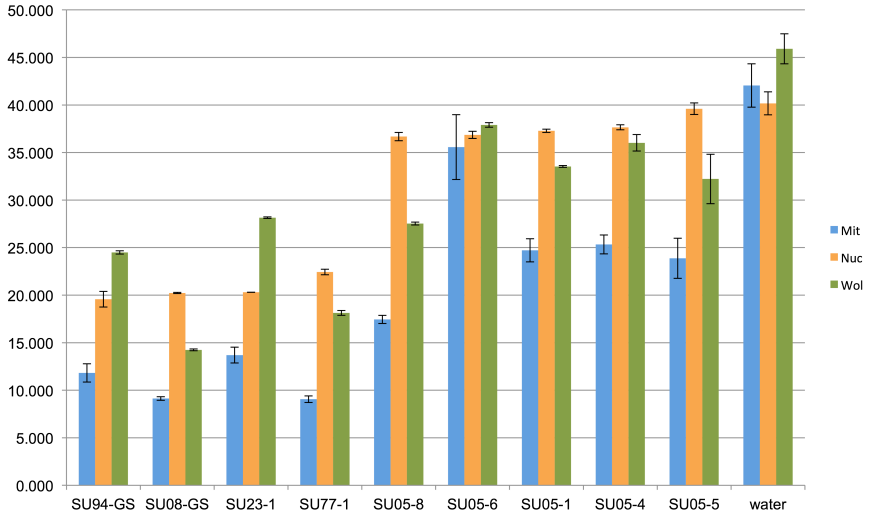


Figure AI2. Haploid embryo quality assessment.

Profiles of DNA content of selected embryos, SU94-GS, SU08-GS and SU5-1 met successfully met quality standards and also yielded high quality sequence data (see Table 5). Embryos SU23-1 and SU77-1 passed quality control but did not yield high quality genome sequences (Table 5). All other SU5 embryos were characterized by a low content of nuclear DNA.

Whole Genome Amplification (WGA) of Embryo

(Qiagen REPLI-g reagents)

1. Thaw REPLI-g Midi DNA Polymerase on ice. Thaw all other components at room temp, vortex, and spin.

- If precipitate in reaction buffer after thawing, vortex for 10 sec.

2. Prepare a WGA master mix (mm), vortex and spin. Make an additional 10% if more than one reaction.

- 32 μl WGA mm per embryo
    - 8 μl nuclease-free water
    - 23.2 μl REPLI-g Midi Reaction Buffer
    - 0.8 μl REPLI-g Midi DNA Polymerase

3. Add 32 μl master mix to 8 μl of denatured embryo lysate (2 μl of the original 10 μl was used for 1:10 dilution for the qPCR embryo assay).

4. Incubate at 30°C o/n.

5. Inactivate REPLI-g Midi DNA Polymerase by incubating at 65°C for 3 min.

Evaluation of WGA

(1% agarose gel)

1. Pour 1% agarose gel. Use a comb with wells for WGA samples, 1 Kb ladder, and 3 lambda DNA samples (125 ng, 250 ng, and 500 ng).

2. Use 8-well strip tubes to prepare WGA samples.

- 1:10 dilution of WGA reaction; 20ul total volume (2 μl sample + 18 μl water)
- Add 4 μl loading dye.
- Vortex and spin down.
- Load 6 μl in well.

3. Prepare known concentrations of lambda DNA using 100ng/ul working stock.

- 125ng = 1.25 μl of stock
- 250ng = 2.5 μl of stock
- 500ng = 5.0 μl of stock

4. Load gel.

- Note:
    - 6 μl of diluted WGA embryo per well
    - Prior to loading the three lambda DNAs, add 1ul of loading dye to each.

LITERATURE CITED

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics 139: 1067-1076.

Andersen, P. R., M. Domanski, M. S. Kristiansen, H. Storvall, E. Ntini *et al.*, 2013 The human cap-binding complex is functionally connected to the nuclear RNA exosome. Nature Structural & Molecular Biology 20: 1367-1376.

Anderson, A. R., A. A. Hoffmann and S. W. Mckechnie, 2005 Response to selection for rapid chill-coma recovery in *Drosophila melanogaster*: physiology and life-history traits. Genetics Research 85: 15-22.

Atwood, K. C., L. K. Schneider and F. J. Ryan, 1951 Periodic selection in *Escherichia coli*. Proceedings of the National Academy of Sciences 37: 146-155.

Axelsson, E., A. Ratnakumar, M.-L. Arendt, K. Maqbool, M. T. Webster *et al.*, 2013 The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495: 360-364.

Ayrinhac, A., V. Debat, P. Gibert, A. G. Kister, H. Legout *et al.*, 2004 Cold adaptation in geographical populations of *Drosophila melanogaster*: phenotypic plasticity is more important than genetic variability. Functional Ecology 18: 700-706.

Ayroles, J. F., M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman *et al.*, 2009 Systems genetics of complex traits in *Drosophila melanogaster*. Nature Genetics 41: 299-307.

Bächli, G., C. R. Vilela, S. A. Escher and A. Saura, 2005 *The Drosophilidae (Diptera) of Fennoscandia*, Leiden

Barrett, R. D. H., and H. E. Hoekstra, 2011 Molecular spandrels: tests of adaptation at the genetic level. Nature reviews Genetics 12: 767-780.

Barton, N. H., 2000 Genetic hitchhiking. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 355: 1553-1562.

Barton , N. H., and P. D. Keightley, 2002 Understanding quantitative genetic variation. Nature reviews Genetics 3: 11-21.

Barton , N. H., and M. Turelli, 1989 Evolutionary quantitative genetics: how little do we know? Annual Review of Genetics 23: 337-370.

Bateson, W., 1909 *Mendel's principles of heredity*. Dover Publications, Cambridge.

Beaumont, M. A., 2005 Adaptation and speciation: what can Fst tell us? Trends in Ecology & Evolution 20: 435-440.

Beaumont , M. A., and D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology 13: 969-980.

Beaumont, M. A., and R. A. Nichols, 1996 Evaluating Loci for Use in the Genetic Analysis of Population Structure. Proceedings of the Royal Society of London. Series B: Biological Sciences 263: 1619-1626.

Beisswanger , S., and W. Stephan, 2008 Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in *Drosophila*. Proceedings of the National Academy of Sciences 105: 5447-5452.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57: 289-300.

Bierne, N., and A. Eyre-Walker, 2004 The Genomic Rate of Adaptive Amino Acid Substitution in *Drosophila*. Molecular Biology and Evolution 21: 1350-1360.

Bierne, N., D. Roze and J. J. Welch, 2013 Pervasive selection or is it…? why are FST outliers sometimes so frequent? Molecular Ecology 22: 2061-2064.

Callahan, B., R. A. Neher, D. Bachtrog, P. Andolfatto and B. I. Shraiman, 2011 Correlated evolution of nearby residues in Drosophilid proteins. PLoS Genetics 7: e1001315.

Campo, D., K. Lehmann, C. Fjeldsted, T. Souaiaia, J. Kao *et al.*, 2013 Whole-genome sequencing of two North American *Drosophila melanogaster* populations reveals genetic differentiation and positive selection. Molecular Ecology 22: 5084-5097.

Cano, J. M., C. Matsuba, H. Mäkinen and J. Merilä, 2006 The utility of QTL-Linked markers to detect selective sweeps in natural populations — a case study of the EDA gene and a linked marker in threespine stickleback. Molecular Ecology 15: 4613-4621.

Capy, P., E. Pla and J. R. David, 1993 Phenotypic and genetic variability of morphometrical traits in natural populations of *Drosophila melanogaster* and *D. simulans*. I. Geographic variations. Gentics Selection Evolution 25: 517-536.

Carroll, S. B., 2005 Evolution at two levels: on genes and form. PLoS Biology 3: e245.

Chevin, L.-M., and F. Hospital, 2008 Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics 180: 1645-1660.

Clowers, K. J., R. F. Lyman, T. F. C. Mackay and T. J. Morgan, 2010 Genetic variation in *senescence marker protein-30* is associated with natural variation in cold tolerance in *Drosophila*. Genetics Research 92: 103-113.

Colinet, H., S. F. Lee and A. A. Hoffmann, 2009 Temporal expression of heat shock genes during cold stress and recovery from chill coma in adult *Drosophila melanogaster*. Journal of the Federation of European Biochemical Societies 277: 174-185.

Coop, G., D. Witonsky, A. Di Rienzo and J. K. Pritchard, 2010 Using Environmental Correlations to Identify Loci Underlying Local Adaptation. Genetics 185: 1411-1423.

Cordell, H. J., 2002 Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics 11: 2463-2468.

David, J., and P. Capy, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends in Genetics 4: 106-111.

Davis, A. W., J. Roote, T. Morley, K. Sawamura, S. Herrmann *et al.*, 1996 Rescue of hybrid sterility in crosses between *D. melanogaster* and *D. simulans*. Nature 380: 157-159.

De Luca, M., N. V. Roshina, G. L. Geiger-Thornsberry, R. F. Lyman, E. G. Pasyukova *et al.*, 2003 *Dopa decarboxylase* (*Ddc*) affects variation in *Drosophila* longevity. Nature Genetics 34: 429-433.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Molecular Biology and Evolution 15: 1788-1790.

Dietrich, M. R., and R. a. J. Skipper, 2012 A shifting terrain: a brief history of the adaptive landscape, pp. 3 - 14 in *The Adaptive Landscape in Evolutionary Biology*, edited by E. Svensson and R. Calsbeek. Oxford University Press, Oxford.

Duchen, P., 2013 Modeling the demographic history of *Drosophila melanogaster* using Approximate Bayesian Computation and Next Generation Sequencing data, pp. 134. Ludwig Maximilians Universität, Munich.

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

Enattah, N. S., A. Trudeau, V. Pimenoff, L. Maiuri, S. Auricchio *et al.*, 2007 Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. The American Journal of Human Genetics 81: 615-625.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa and M. Foll, 2013 Robust Demographic Inference from Genomic and SNP Data. PLoS Genetics 9: e1003905.

Excoffier, L., and M. Foll, 2011 fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27: 1332-1334.

Fabian, D. K., M. Kapun, V. Nolte, R. Kofler, P. S. Schmidt *et al.*, 2012 Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. Molecular Ecology 21: 4748-4769.

Falconer, D. S., and T. F. C. Mackay, 1996 Introduction to Quantitative Genetics, pp.

Fallis, L., J. Fanara and T. Morgan, 2011 Genetic variation in heat-stress tolerance among South American *Drosophila* populations. Genetica 139: 1331-1337.

Fallis, L. C., 2012 The evolution and genetics of thermal traits In *Drosophila melanogaster*, pp. 125. Kansas State University, Manhattan.

Fanara, J. J., K. O. Robinson, S. M. Rollmann, R. R. H. Anholt and T. F. C. Mackay, 2002 Vanaso is a candidate quantitative trait gene for *Drosophila* olfactory behavior. Genetics 162: 1321-1328.

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive darwinian selection. Genetics 155: 1405-1413.

Fiston-Lavier, A.-S., N. D. Singh, M. Lipatov and D. A. Petrov, 2010 *Drosophila melanogaster* recombination rate calculator. Gene 463: 18-20.

Foll, M., and O. Gaggiotti, 2008 A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180: 977-993.

François, O., M. G. B. Blum, M. Jakobsson and N. A. Rosenberg, 2008 Demographic history of European populations of Arabidopsis thaliana. PLoS Genetics 4: e1000075.

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

Gaggiotti, O., and M. Foll, 2010 Quantifying population structure using the F-model. Molecular Ecology Resources 10: 821-830.

Geber, M. A., 2011 Ecological and evolutionary limits to species geographic ranges. The American Naturalist 178: S1-S5.

Gibert, P., B. Moreteau, G. Pétavy, D. Karan and J. R. David, 2001 Chill-coma tolerance, a major climatic adaptation among *Drosophila* species. Evolution 55: 1063-1068.

Glaser-Schmitt, A., A. Catalán and J. Parsch, 2013 Adaptive divergence of a transcriptional enhancer between populations of *Drosophila melanogaster*. Philosophical Transactions of the Royal Society B: Biological Sciences 368.

Glinka, S., D. De Lorenzo and W. Stephan, 2006 Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster*. Molecular Biology and Evolution 23: 1869-1878.

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

Gompel, N., B. Prud'homme, P. J. Wittkopp, V. A. Kassner and S. B. Carroll, 2005 Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. Nature 433: 481-487.

Goto, S. G., T. Yoshida, K. Beppu and M. T. Kimura, 1999 Evolution of overwintering strategies in Eurasian species of the *Drosophila* obscura species group. Biological

Journal of the Linnean Society 68: 429-441.
Gould, S. J., 2002 *The structure of evolutionary theory*. Hardvard University Press, Cambridge.
Gould, S. J., and R. C. Lewontin, 1979 The spandrels of San Marco and the panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London. Series B, Biological Sciences 205: 581-598.
Gouy, M., S. Guindon and O. Gascuel, 2010 SeaView Version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27: 221-224.
Grant, P. R., and B. R. Grant, 2002 Unpredictable Evolution in a 30-Year Study of Darwin's Finches. Science 296: 707-711.
Graveley, B. R., G. May, A. N. Brooks, J. W. Carlson, L. Cherbas *et al.*, 2011 The *D. melanogaster* transcriptome: modENCODE RNA-Seq data for differing treatment conditions. P. C. T. Flybase.
Günther, T., and G. Coop, 2013 Robust Identification of Local Adaptation from Allele Frequencies. Genetics 195: 205-220.
Gurganus, M. C., S. V. Nuzhdin, J. W. Leips and T. F. C. Mackay, 1999 High-resolution mapping of quantitative trait loci for sternopleural bristle number in *Drosophila melanogaster*. Genetics 152: 1585-1604.
Hale, L. R., and R. S. Singh, 1991 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. IV. Mitochondrial DNA variation and the role of history vs. selection in the genetic structure of geographic populations. Genetics 129: 103-117.
Hancock, A. M., G. Alkorta-Aranburu, D. B. Witonsky and A. Di Rienzo, 2010 Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philosophical Transactions of the Royal Society B: Biological Sciences 365: 2459-2468.
Hancock, A. M., D. B. Witonsky, G. Alkorta-Aranburu, C. M. Beall, A. Gebremedhin *et al.*, 2011 Adaptations to climate-mediated selective pressures in humans. PLoS Genetics 7: e1001375.
Harbison, S. T., A. H. Yamamoto, J. J. Fanara, K. K. Norga and T. F. C. Mackay, 2004 Quantitative Trait Loci Affecting Starvation Resistance in *Drosophila melanogaster*. Genetics 166: 1807-1823.
Hellemans, J., G. Mortier, A. De Paepe, F. Speleman and J. Vandesompele, 2007 qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biology 8: R19.
Hermisson, J., and P. S. Pennings, 2005 Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. Genetics 169: 2335-2352.
Herold, N., C. L. Will, E. Wolf, B. Kastner, H. Urlaub *et al.*, 2009 Conservation of the protein composition and electron microscopy structure of *Drosophila melanogaster* and human spliceosomal complexes. Molecular and Cellular Biology 29: 281-301.
Hertz, G. Z., and G. D. Stormo, 1999 Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15: 563-577.
Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38: 226-231.
Hoffmann, A. A., A. Anderson and R. Hallas, 2002 Opposing clines for high and low temperature resistance in *Drosophila melanogaster*. Ecology Letters 5: 614-618.
Hoffmann, A. A., and M. W. Blows, 1994 Species borders: ecological and evolutionary perspectives. Trends in Ecology & Evolution 9: 223-227.

Hoffmann, A. A., R. Hallas, C. Sinclair and P. Mitrovski, 2001 Levels of variation in stress resistance in *Drosophila* among strains, local populations, and geographic regions: patterns for desiccation, starvation, cold resistance, and associated traits. Evolution 55: 1621-1630.

Hoffmann, A. A., and P. A. Parsons, 1989 An integrated approach to environmental stress tolerance and life-history variation: desiccation tolerance in *Drosophila*. Biological Journal of the Linnean Society 37: 117-136.

Hoffmann, A. A., M. Scott, L. Partridge and R. Hallas, 2003a Overwintering in *Drosophila melanogaster*: outdoor field cage experiments on clinal and laboratory selected populations help to elucidate traits under selection. Journal of Evolutionary Biology 16: 614-623.

Hoffmann, A. A., J. Shirriffs and M. Scott, 2005 Relative importance of plastic vs genetic factors in adaptive differentiation: geographical variation for stress resistance in *Drosophila melanogaster* from Eastern Australia. Functional Ecology 19: 222-227

Hoffmann, A. A., J. G. Sorensen and V. Loeschcke, 2003b Adaptation of *Drosophila* to temperature extremes: bringing together quantitative and molecular approaches. Journal of Thermal Biology 28: 175-216.

Hoffmann, A. A., and A. R. Weeks, 2007 Climatic selection on genes and traits after a 100 year-old invasion: a critical look at the temperate-tropical clines in *Drosophila melanogaster* from eastern Australia. Genetica 129: 133-147.

Huang, W., S. Richards, M. A. Carbone, D. Zhu, R. R. H. Anholt *et al.*, 2012 Epistasis dominates the genetic architecture of *Drosophila* quantitative traits. Proceedings of the National Academy of Sciences 109: 15553-15559.

Hübner, S., E. Rashkovetsky, Y. B. Kim, J. H. Oh, K. Michalak *et al.*, 2013 Genome differentiation of *Drosophila melanogaster* from a microclimate contrast in Evolution Canyon, Israel. Proceedings of the National Academy of Sciences.

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

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics 116: 153-159.

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

Innan, H., and W. Stephan, 2001 Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. Genetics 159: 389-399.

Izquierdo, J. I., 1991 How does *Drosophila melanogaster* overwinter? Entomologia Experimentalis et Applicata 59: 51–58.

Jablonski, N. G., and G. Chaplin, 2012 Human skin pigmentation, migration and disease susceptibility. Philosophical Transactions of the Royal Society B: Biological Sciences 367: 785-792.

Jensen, J. D., V. L. Bauer Dumont, A. B. Ashmore, A. Gutierrez and C. F. Aquadro, 2007 Patterns of sequence variability and divergence at the diminutive gene region of *Drosophila melanogaster*: complex patterns suggest an ancestral selective sweep. Genetics 177: 1071-1085.

Jensen, J. D., Y. Kim, V. B. Dumont, C. F. Aquadro and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401-1410.

Jordan, K. W., T. J. Morgan and T. F. C. Mackay, 2006 Quantitative trait loci for locomotor behavior in *Drosophila melanogaster*. Genetics 174: 271-284.

Jukes, T. H., 2000 The neutral theory of molecular evolution. Genetics 154: 956-958.
Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics 123: 887-899.
Kawecki, T. J., 2008 Adaptation to marginal habitats. Annual Review of Ecology, Evolution, and Systematics 39: 321-342.
Keightley, P. D., 1998 Genetic basis of response to 50 generations of selection on body weight in inbred mice. Genetics 148: 1931-1939.
Keightley, P. D., and G. Bulfield, 1993 Detection of quantitative trait loci from frequency changes of marker alleles under selection. Genetics Research 62: 195-203.
Keller, A., 2007 *Drosophila melanogaster*'s history as a human commensal. Current Biology 17: R77-R81.
Kelley, L. A., and M. J. E. Sternberg, 2009 Protein structure prediction on the Web: a case study using the Phyre server. Nature Protocols 4: 363-371.
Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics 146: 1197-1206.
Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513-1524.
Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765-777.
Kimura, M., 1985 The role of compensatory neutral mutations in molecular evolution. Journal of Genetics 64: 7-19.
Kimura, M. T., 1988 Adaptations to temperate climates and evolution of overwintering strategies in the *Drosophila*-melanogaster species group. Evolution 42: 1288-1297.
Kimura, M. T., 2004 Cold and heat tolerance of drosophilid flies with reference to their latitudinal distributions. Oecologia 140: 442-449.
King, M., and A. Wilson, 1975 Evolution at two levels in humans and chimpanzees. Science 188: 107-116.
Kirkpatrick, H., K. Johnson and A. Laughon, 2001 Repression of Dpp targets by binding of brinker to Mad Sites. Journal of Biological Chemistry 276: 18216-18222.
Kirkpatrick, M., and N. Barton, 1997 Evolution of a species' range. The American Naturalist 150: 1-23.
Kolaczkowski, B., A. D. Kern, A. K. Holloway and D. J. Begun, 2011 Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. Genetics 187: 245-260.
Koonin, E. V., 2009 Darwinian evolution in the light of genomics. Nucleic Acids Research 37: 1011-1034.
Koštál, V., D. Renault, A. Mehrabianová and J. Bastl, 2007 Insect cold tolerance and repair of chill-injury at fluctuating thermal regimes: role of ion homeostasis. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology 147: 231-238.
Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304: 412-417.
Kreitman, M., 1996 The neutral theory is dead. Long live the neutral theory. BioEssays 18: 678-683.
Lachaise, D., M. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup pp. 159–225 in *Evolutionary biology*, edited by B. W. M.K. Hecht, And G. T. Prance. Plenum Press, New York.
Lamason, R. L., M.-a. P. K. Mohideen, J. R. Mest, A. C. Wong, H. L. Norton *et al.*, 2005 SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and

humans. Science 310: 1782-1786.
Lande, R., 1983 The response to selection on major and minor mutations affecting a metrical trait. Heredity 50: 47-65.
Lander, E. S., and D. Botstein, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185-199.
Langley, C. H., M. Crepeau, C. Cardeno, R. Corbett-Detig and K. Stevens, 2011 Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. Genetics 188: 239-246.
Langley, C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider *et al.*, 2012 Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192: 533-598.
Laurent, S. J. Y., A. Werzner, L. Excoffier and W. Stephan, 2011 Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. Molecular Biology and Evolution 28: 2041-2051.
Laurie, C. C., S. D. Chasalow, J. R. Ledeaux, R. Mccarroll, D. Bush *et al.*, 2004 The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. Genetics 168: 2141-2155.
Lewis, P. O., 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50: 913-925.
Lewontin, R. C., 1997 Dobzhansky's Genetics and the Origin of Species: Is It Still Relevant? Genetics 147: 351-355.
Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-195.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
Li, H., J. Ruan and R. Durbin, 2008 Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18: 1851-1858.
Li, H., and W. Stephan, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genetics 2: e166.
Li, Y.-J., Y. Satta and N. Takahata, 1999 Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes & Genetic Systems 74: 117-127.
Linnen, C. R., Y.-P. Poh, B. K. Peterson, R. D. H. Barrett, J. G. Larson *et al.*, 2013 Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339: 1312-1316.
Lynch, M., and B. Walsh, 1998 *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland.
Mackay, T. F., and R. F. Lyman, 2005 *Drosophila* bristles and the nature of quantitative genetic variation. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 360: 1513-1527.
Mackay, T. F. C., 2001 Quantitative trait loci in *Drosophila*. Nature reviews Genetics 2: 11-20.
Mackay, T. F. C., 2014 Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nature reviews Genetics 15: 22-33.
Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles *et al.*, 2012 The *Drosophila melanogaster* genetic reference panel. Nature 482: 173-178.
Mackay, T. F. C., E. A. Stone and J. F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. Nature reviews Genetics 10: 565-577.
Macmillan , H. A., and B. J. Sinclair, 2011 Mechanisms underlying insect chill-coma. Journal of Insect Physiology 57: 12-20.

Macmillan, H. A., C. M. Williams, J. F. Staples and B. J. Sinclair, 2012 Reestablishment of ion homeostasis during chill-coma recovery in the cricket Gryllus pennsylvanicus. Proceedings of the National Academy of Sciences 109: 20750-20755.

Makalowski, W., and M. S. Boguski, 1998 Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. Proceedings of the National Academy of Sciences 95: 9407-9412.

Massouras, A., S. M. Waszak, M. Albarca-Aguilera, K. Hens, W. Holcombe *et al.*, 2012 Genomic variation and Its impact on gene expression in *Drosophila melanogaster*. PLoS Genetics 8: e1003055.

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

Messer, P. W., and D. A. Petrov, 2013 Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology & Evolution 28: 659-669.

Minami, M., N. Kinoshita, Y. Kamoshida, H. Tanimoto and T. Tabata, 1999 brinker is a target of Dpp in *Drosophila* that negatively regulates Dpp-dependent genes. Nature 398: 242-246.

Moore, J. A., 1983 Thomas Hunt Morgan, The Geneticist. American Zoologist 23: 855-865.

Moreno, E., K. Basler and G. Morata, 2002 Cells compete for Decapentaplegic survival factor to prevent apoptosis in *Drosophila* wing development. Nature 416: 755-759.

Negre, N., C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller *et al.*, 2011 A cis-regulatory map of the *Drosophila* genome. Nature 471: 527-531.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., and W. H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences 76: 5269-5273.

Nei, M., and T. Maruyama, 1975 Lewontin-Krakauer test for neutral genes. Genetics 80: 395.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Research 15: 1566-1575.

Nilsson, A. L., 1998 Deep flowers for long tongues. Trends in Ecology & Evolution 13: 259-260.

Nuzhdin, S. V., L. G. Harshman, M. Zhou and K. Harmon, 2007 Genome-enabled hitchhiking mapping identifies QTLs for stress resistance in natural *Drosophila*. Heredity 99: 313-321.

Nuzhdin, S. V., E. G. Pasyukova, C. L. Dilda, Z. B. Zeng and T. F. C. Mackay, 1997 Sex-specific quantitative trait loci affecting longevity in *Drosophila melanogaster*. Proceedings of the National Academy of Sciences 94: 9734-9739.

Nuzhdin, S. V., and T. L. Turner, 2013 Promises and limitations of hitchhiking mapping. Current Opinion in Genetics & Development 23: 694-699.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu *et al.*, 2012 Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS Genet 8: e1002685.

Orr, A., 1998 The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. Evolution 52: 935-949.

Orr, A., 2005 The genetic theory of adaptation: a brief history. Nature reviews Genetics 6: 119-127.

Pasyukova, E. G., N. V. Roshina and T. F. C. Mackay, 2004 Shuttle craft: a candidate quantitative trait gene for *Drosophila* lifespan. Aging Cell 3: 297-307.

Pasyukova, E. G., C. Vieira and T. F. C. Mackay, 2000 Deficiency mapping of quantitative tait loci affecting longevity in *Drosophila melanogaster*. Genetics 156: 1129-1146.

Pavlidis, P., S. Hutter and W. Stephan, 2008 A population genomic approach to map recent positive selection in model species. Molecular Ecology 17: 3585-3598.

Pavlidis, P., J. D. Jensen and W. Stephan, 2010 Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics 185: 907-922.

Pavlidis, P., D. Metzler and W. Stephan, 2012 Selective sweeps in multilocus models of quantitative traits. Genetics 192: 225-239.

Pavlidis, P., D. Živković, A. Stamatakis and N. Alachiotis, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Molecular Biology and Evolution 30: 2224-2234.

Perl, D., and F. X. Schmid, 2002 Some like it hot: the molecular determinants of protein thermostability. European Journal of Chemical Biology 3: 39-44.

Pfaffelhuber, P., A. Lehnert and W. Stephan, 2008 Linkage Disequilibrium Under Genetic Hitchhiking in Finite Populations. Genetics 179: 527-537.

Phillips, P. C., 2008 Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. Nature reviews Genetics 9: 855-867.

Pool, J. E., 2009 Notes regarding the collection of African *Drosophila melanogaster*. *Drosophila* Information Service 92: 130-134.

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

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno *et al.*, 2012 Population genomics of Sub-Saharan *Drosophila melanogaster*: African diversity and Non-African admixture. PLoS Genetics 8: e1003080.

Pritchard, J. K., and A. Di Rienzo, 2010 Adaptation – not by sweeps alone. Nature reviews Genetics 11: 665-667.

Pritchard, J. K., J. K. Pickrell and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology 20: R208-R215.

Przeworski, M., 2002 The Signature of Positive Selection at Randomly Chosen Loci. Genetics 160: 1179-1189.

Przeworski, M., G. Coop, J. D. Wall and M. Nachman, 2005 The signature of positive selection on standing genetic variation. Evolution 59: 2312-2323.

Pyrowolakis, G., B. Hartmann, B. Müller, K. Basler and M. Affolter, 2004 A simple molecular complex mediates widespread BMP-induced repression during *Drosophila* development. Developmental Cell 7: 229-240.

Régniere, J., J. Powell, B. Bentz and V. Nealis, 2012 Effects of temperature on development, survival and reproduction of insects: Experimental design, data analysis and modeling. Journal of Insect Physiology 58: 634-647.

Régnière, J., R. St-Amant and P. Duval, 2012 Predicting insect distributions under climate change from physiological responses: spruce budworm as an example. Biological Invasions 14: 1571-1586.

Remold, S. K., and R. E. Lenski, 2004 Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. Nat Genet 36: 423-426.

Remolina, S. C., P. L. Chang, J. Leips, S. V. Nuzhdin and K. A. Hughes, 2012 Genomic basis of aging and life-history evolution in *Drosophila melanogaster*. Evolution 66: 3390-3403.

Riebler, A., L. Held and W. Stephan, 2008 Bayesian variable selection for detecting

adaptive genomic differences among populations. Genetics 178: 1817-1829.
Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. Science 273: 1516-1517.
Rockman, M. V., 2012 The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution 66: 1-17.
Rodríguez-Trelles, F., R. Tarrío and M. Santos, 2013 Genome-wide evolutionary response to a heat wave in *Drosophila*. Biology Letters 9.
Rubin, C.-J., H.-J. Megens, A. Martinez Barrio, K. Maqbool, S. Sayyab *et al.*, 2012 Strong signatures of selection in the domestic pig genome. Proceedings of the National Academy of Sciences 109: 19529-19536.
Rubin, C.-J., M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood *et al.*, 2010 Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464: 587-591.
Rubinstein, M., and F. S. J. De Souza, 2013 Evolution of transcriptional enhancers and animal diversity. Philosophical Transactions of the Royal Society B: Biological Sciences 368.
Saminadin-Peter, S. S., C. Kemkemer, P. Pavlidis and J. Parsch, 2012 Selective sweep of a cis-regulatory sequence in a non-African population of *Drosophila melanogaster*. Molecular Biology and Evolution 29: 1167-1174.
Schmidt, P. S., L. Matzkin, M. Ippolito and W. F. Eanes, 2005 Geographic variation in diapause incidence, life-history traits, and climatic adaptation in *Drosophila melanogaster*. Evolution 59: 1721-1732.
Schmidt, P. S., and A. B. Paaby, 2008 Reproductive diapause and life-history clines in North American populations of *Drosophila melanogaster*. Evolution 62: 1204-1215.
Scriver, C. R., 2008 Garrod's Croonian Lectures (1908) and the charter "Inborn Errors of Metabolism" : Albinism, alkaptonuria, cystinuria, and pentosuria at age 100 in 2008. Journal of Inherited Metabolic Disease 31: 580-598.
Service, P. M., 2004 How good are quantitative complementation tests? Science of Aging Knowledge Environment. 2004: pe13.
Sexton, J. P., P. J. Mcintyre, A. L. Angert and K. J. Rice, 2009 Evolution and Ecology of Species Range Limits. Annual Review of Ecology, Evolution, and Systematics 40: 415-436.
Shao, H., L. C. Burrage, D. S. Sinasac, A. E. Hill, S. R. Ernest *et al.*, 2008 Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. Proceedings of the National Academy of Sciences 105: 19910-19914.
Stephan, W., and H. Li, 2007 The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98: 65-68.
Stephan, W., T. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theoretical Population Biology 41: 237–254.
Stutervant, A., 1913 The linear arrangement of six sex- linked factors in *Drosophila*, as shown by their mode of association. Journal of Experimental Zoology 14: 43-59.
Suissa, Y., O. Ziv, T. Dinur, E. Arama and O. Gerlitz, 2011 The NAB-Brk Signal Bifurcates at JNK to Independently Induce Apoptosis and Compensatory Proliferation. Journal of Biological Chemistry 286: 15556-15564.
Sun, G., and P. Schliekelman, 2010 A Genetical Genomics Approach to Genome Scans Increases Power for QTL Mapping. Genetics 187: 939-953.
Svetec, N., 2009 Searching for genes involved in the adaptation of *Drosophila melanogaster* to the European climate, pp. 132. Ludwig Maximilians Universität, Munich.
Svetec, N., P. Pavlidis and W. Stephan, 2009 Recent strong positive selection on

Drosophila melanogaster HDAC6, a gene encoding a stress surveillance factor, as revealed by population genomic analysis. Molecular Biology and Evolution 26: 1549-1556.

Svetec, N., A. Werzner, R. Wilches, P. Pavlidis, J. M. Álvarez-Castro *et al.*, 2011 Identification of X-linked quantitative trait loci affecting cold tolerance in *Drosophila melanogaster* and fine mapping by selective sweep analysis. Molecular Ecology 20: 530-544.

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

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

Tauber, E., M. Zordan, F. Sandrelli, M. Pegoraro, N. Osterwalder *et al.*, 2007 Natural selection favors a newly derived timeless allele in *Drosophila melanogaster*. Science 316: 1895-1898.

Telonis-Scott, M., R. Hallas, S. W. Mckechnie, C. W. Wee and A. A. Hoffmann, 2009 Selection for cold resistance alters gene transcript levels in *Drosophila melanogaster*. Journal of Insect Physiology 55: 549-555.

Thornton, K. R., and J. D. Jensen, 2007 Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175: 737-750.

Thornton, K. R., J. D. Jensen, C. Becquet and P. Andolfatto, 2007 Progress and prospects in mapping recent selection in the genome. Heredity 98: 340-348.

Turner, T. L., A. D. Stewart, A. T. Fields, W. R. Rice and A. M. Tarone, 2011 Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. PLoS Genetics 7: e1001336.

Uniprot-Consortium, 2014 Activities at the Universal Protein Resource (UniProt). Nucleic Acids Research 42: D191-D198.

Valenzuela, R. K., S. N. Forbes, P. Keim and P. M. Service, 2004 Quantitative trait loci affecting life span in replicated populations of *Drosophila melanogaster*. II. Response to selection. Genetics 168: 313-324.

Verhoeven, K. J. F., G. Casella and L. M. Mcintyre, 2010 Epistasis: obstacle or advantage for mapping complex traits? PLoS one 5: e12264.

Von Tschermak-Seysenegg, E., 1951 The rediscovery of Gregor Mendel's Work: an Historical Retrospect. Journal of Heredity 42: 163-171.

Vucetich, J., and T. Waite, 2003 Spatial patterns of demography and genetic processes across the species' range: Null hypotheses for landscape conservation genetics. Conservation Genetics 4: 639-645.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7: 256-276.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Asssociates Inc., Sunderland.

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

Werzner, A., 2011 Local adaptation in *Drosophila melanogaster* - Molecular and morphological aspects, pp. 132. Ludwig Maximilians Universität, Munich.

Willi, Y., J. Van Buskirk and A. A. Hoffmann, 2006 Limits to the adaptive potential of small populations, pp. 433-458 in *Annual Review of Ecology Evolution and Systematics*. Annual Reviews, Palo Alto.

Wilson, R. H., T. J. Morgan and T. F. C. Mackay, 2006 High-resolution mapping of quantitative trait loci affecting increased life span in *Drosophila melanogaster*. Genetics 173: 1455-1463.

Wright, S., 1931 Evolution in mendelian populations. Genetics 16: 97-159.

Yao, L.-C., S. Phin, J. Cho, C. Rushlow, K. Arora *et al.*, 2008 Multiple modular promoter elements drive graded brinker expression in response to the Dpp morphogen gradient. Development 135: 2183-2192.

Yasuda, G. K., G. Schubiger and B. T. Wakimoto, 1995 Genetic characterization of *ms(3)K81*, a paternal effect gene of *Drosophila melanogaster*. Genetics 140: 219-229.

Yoo, B. H., F. W. Nicholas and K. A. Rathie, 1980 Long-term selection for a quantitative character in large replicate populations of *Drosophila melanogaster*. Theoretical and Applied Genetics 57: 113-117.

Zhou, S., T. G. Campbell, E. A. Stone, T. F. C. Mackay and R. R. H. Anholt, 2012 Phenotypic plasticity of the *Drosophila* transcriptome. PLoS Genetics 8: e1002593.

Zichner, T., D. A. Garfield, T. Rausch, A. M. Stotz, E. Canna *et al.*, 2013 Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. Genome Research 23: 568-579.

Ziv, O., Y. Suissa, H. Neuman, T. Dinur, P. Geuking *et al.*, 2009 The co-regulator dNAB interacts with Brinker to eliminate cells with reduced Dpp signaling. Development 136: 1137-1145.

Živković, D., and T. Wiehe, 2008 Second-order moments of segregating sites under variable population size. Genetics 180: 341-357.

ACKNOWLEDGEMENTS

I would like to thank Professor W. Stephan for his excellent academic mentoring throughout these years of PhD. I want to highlight his help and support while writing the grant proposal for the Volkswagen Foundation, which allowed me to develop my research in Munich. I acknowledge his always-constructive exchange of ideas as projects progressed as well as his invaluable contribution, for instance, with manuscript preparations.

I thank the members of the Evolutionary Biology group in Munich who contributed in several ways to the development of my thesis project or by creating a friendly work atmosphere. I particularly want to address my close collaborators, P. Duchén for providing support with C codes tailored to the needs of my analyses and datasets, his patience and guidance with my own codes, as well for his opportune advice on my manuscripts. I express my gratitude to S. Laurent for his readiness to help and enriching exchange of ideas. I thank S. Voigt for support with all aspects regarding gene expression analyses and enriching discussions. I also would like to thank A. Steincke for her pivotal contribution to the successful completion of my project with her fly and lab work, likewise to H. Lainer and S. Lange for their excellent technical assistance. I acknowledge Professor J. Parsch and R. Morrison for having read parts of this manuscript and suggested style changes.

I want to make a special mention of gratitude to the following persons and institutions:

- To Frau I. Kroiss for her friendly and efficient help with bureaucratic issues.
- To the Volkswagen Foundation for making this work possible (through research grant I/848313).
- To the EES master and PhD community in Munich from 2007 to 2013 with whom I have had the opportunity to enjoy great times and build friendships.
- To my close friends and colleagues, especially A. Catalán, P. Duchén, M. Wittmann with whom our common interest for science extended to art, books, wine, food and evolved to the feeling of a family away from home.

Although addressed lastly, my first words of gratitude are for my family. My parents, H. and T. (de) Wilches, and my sisters C. and L. Wilches, whose love and care, in spite of the distance, has meant a standing source of support through this challenging step in my scientific and professional formation.

CURRICULUM VITAE

Ricardo Wilches

ricardoinbaveria@gmail.com

Current professional affiliation

LMU Biozentrum
Department of Evolutionary Biology
Großhadernerstraße 2
82152 Planegg-Martinsried

Email: wilches@bio.lmu.de

Education

- Doctor in Philosophy. Evolutionary Genetics, Ludwig Maximilians Universität. Munich. Germany.
- Master of Science in Biology, Munich graduate school in Evolution, Ecology and Systematics Ludwig Maximilians Universitatät. Munich. Germany. 2010.
- Bachelor degree in Biology (with honors). Pontificia Universidad Javeriana. Bogotá D.C. Colombia. 2004.

Fellowships, Grants and Awards

- Doctoral research grant. Volkswagen foundation. Germany. 2010 - 2014.
- Undergraduate travel award. Society for Molecular Biology and Evolution. SMBE Meeting. Barcelona. Spain. 2008.
- Scholarship for Master studies. German Academic Exchange Service (DAAD). Germany. 2007-2009.
- "Joven Inverstigador" Research training fellowship. Administrative Department of Science, Technology and Innovation (COLCIENCIAS). Colombia. 2005-2006

Publications

- Svetec N, Werzner A, Wilches R, et al. 2011. Identification of X-linked quantitative trait loci affecting cold tolerance in *Drosophila melanogaster* and fine mapping by selective sweep analysis. Molecular Ecology 20: 530–544.
- Mendoza E, Hernandez C, Wilches R, Varela L, Villareal J, Barrera L, Villanueva D. 2010. Genotype frequencies of *C/T -13910* and *G/A -22018* polymorphism in a Colombian Caribbean population do not correspond with lactase persistence prevalence reported in the region. Colombia Medica. 41:290-294. (Article only in Spanish).
- Morrison WR, Lohr JN, Duchen P, Wilches R, Trujillo D, Mair M. and Renner, S.S. 2009. The impact of taxonomic change on conservation: Does it kill, can it save, or is it just irrelevant? Biological Conservation 142: 3201-3206.
- Wilches R, Vega H, Echeverri O, Barrera LA. 2006. Colombian haplotypes of the Gaucher disease-causing *N370S* mutation may originate from a possible common ancestral haplotype. Biomédica . 26:433-441. (Article only in Spanish).

Research experience

- Department of Evolutionary Biology. LMU. Munich. Biozentrum Martinsried. Germany. 2010-present. PhD project title: Evolution of genes related to temperature adaptation in Drosophila melanogaster as revealed by QTL and population genetics analyses. Doctoral research guided by Professor Doctor Wolfgang Stephan.
- Max Planck Institute for Neurobiology. Martinsried. Germany. 2009-2010 Research helping student (German: Hilfswissenschaftler). Supervisor: Ilona Grunwald-Kadow, PhD.
- Institute for inborn errors of metabolism (IEIM). Faculty of Sciences. Pontificia Universidad Javeriana. Bogota D.C. Colombia. 2004-2006. Research assistant. Supervisor: Luis. A. Barrera, PhD.

Participation in (selected) events

- EES Conference. 8-9 October 2013 Biozentrum LMU. Martinsried, Germany (Talk)
- XIV Congress European society for evolutionary biology. 19-24 August 2013. Lisbon. Portugal (Poster presentation)
- Workshop in Evolutionary quantitative genetics. 5-10 August 2012. National Evolutionary Synthesis Center (NESCent) Durham (NC). USA.
- 4th International Conference on Quantitative Genetics. 17-22 June 2012. Edinburgh. UK. (Poster presentation)
- XII Congress European society for evolutionary biology. 24-29 August 2009. Turin. Italy (Poster presentation)
- Society for Molecular biology and evolution Annual meeting. 5-8 June 2008. Barcelona. Spain. (Poster presentation)

Language skills

- English: advanced command of spoken and written communication. TOEFL-computer based (2012) over all score 245/300
- Spanish: native speaker language command.
- German: intermediate understanding and use of the language (listening and reading skills) Spoken communication level B2 (CEFR Standards).
- Italian: intermediate understanding and use of the language (listening, reading, speaking skills)