# Signatures of polygenic adaptation in European *Drosophila melanogaster*

Dissertation
an der Fakultät für Biologie
der Ludwig-Maximilians-Universität
München

vorgelegt von

Vedran Bozicevic

aus Zagreb, Kroatien

München, Februar 2017

Dekan: Prof. Dr. Heinrich Leonhardt

1. Gutachter: Prof. Dr. Wolfgang Stephan

2. Gutachter: Prof. Dr. Wolfgang Enard

Tag der Abgabe: 20. Februar 2017

Tag der mündlichen Prüfung: 12. Juni 2017

## Erklärung:

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

## Eidesstattliche Erklärung:

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

München, den 20.02.2017

Vedran Bozicevic

## Contribution and Acknowledgements

In this thesis, I present the results of the doctoral research I conducted between October 2012 and June 2016 in the area of evolutionary biology. This research comprises predominantly population genetic analyses of the fruit fly *Drosophila melanogaster*, as well as the associated bioinformatics and statistical analyses, and was done in collaboration with several other scientists.

The project was initiated, designed, and coordinated by Andreas Wollstein, Wolfgang Stephan, and myself. Andreas Wollstein and myself carried out the bioinformatics and analyzed the genetic data. Stephan Hutter contributed bioinformatics resources and tools. Andreas Wollstein and myself performed the statistical analyses. Stefan Laurent and Pablo Duchén provided bioinformatic support in the early phases of the project.

## Original Publications

Božičević V, Hutter S, Stephan W, Wollstein A (2016): Population genetic evidence for cold adaptation in European *Drosophila melanogaster* populations. *Molecular Ecology*, **25**, 1175-91.

Croze M, Wollstein A, Božičević V, Živković D, Stephan W, Hutter S (2017). A genome-wide scan for genes under balancing selection in *Drosophila melanogaster*. *BMC Evolutionary Biology*, **17**, 15.

# Table of contents

# Summary

Research on this topic was motivated by recent advances in two empirically different approaches to the study of adaptation, population genetics and quantitative genetics. New sequencing methods now allow population geneticists to identify with great precision footprints of selective events at single loci, while recent quantitative genetics studies have provided high-powered genome-wide maps of causal alleles associated with phenotypic variation. A combination of insights from both fields has allowed researchers to ask whether loci detected by genome-wide association studies are enriched in signals of adaptation (i.e. polygenic adaptation). However, most research in polygenic adaptation so far has been carried out on human data, and generally only to identify individual loci with unusual allele frequency patterns.

The first part of my thesis (Results - Section 3.1) aims to fill this void by analyzing the recently available NGS and GWAS data of *Drosophila melanogaster* to uncover population genetic signatures of polygenic adaptation. We focused on three phenotypic traits that are known to be adaptively important: chill-coma recovery time, resistance to starvation stress, and startle response. My aim was to test the hypothesis that European populations of *Drosophila melanogaster*, an originally subtropical species, have adapted to the colder climate by correlated shifts in allele frequencies at SNPs associated with these three traits.
We show that SNPs associated to CCRT show overall higher levels of population differentiation, as estimated by pairwise $F_{ST}$ between Europe and Africa, and higher correlations with environmental variables, reported as Bayes factors by *Bayenv2*. We assess the mean pairwise $F_{ST}$ and mean Bayes factors over the associated sets by comparing them with sets of equal size randomly sampled from the genomic background. Furthermore, we assess the single outlier SNPs by comparison with simulated data from a likely demographic model.

In the second part of Section 3.1, we move from the signatures of adaptation on the level of SNPs to the signatures on the level of genes. We first review what is known about candidate genes involved in cold tolerance, resistance to starvation and locomotory responses. We then functionally classify all SNPs within those genes that show high pairwise $F_{ST}$ (in at least one population pair) or high Bayes factors (for correlation with at

least one environmental variable). Both high $F_{ST}$ and high Bayes factors we characterize as measured by the empirical *P*-values derived from neutral coalescent simulations. We show that, for instance, many genes previously related to cold tolerance contain a large number of highly differentiated intronic SNPs, and moreover that particular genes contain highly differentiated SNPs in multiple functional classes (intronic, 3' and 5' UTR, synonymous and nonsynonymous coding SNPs).

Next, we moved from previously characterized genes for cold tolerance and related traits to all the genes in the genome. We performed Gene Ontology and KEGG / Reactome pathway enrichment analyses of most highly significant candidate genes, as defined by their highest Bayes factor SNPs for each environmental variable. To reduce noise, we clustered the enriched categories that contained many overlapping genes, and assessed the most highly significant categories from each cluster. Thereafter we took the candidate genes from a study of latitudinal adaptation that examined a cline along the eastern coast of the US, and performed the same GO and pathway enrichment and clustering analyses. We then assessed the significance of the overlap of enriched categories from both studies, showing that there is significant overlap with the North American cline, and that there is even an overlap between the clusters of enriched categories, i.e. the most significant GO term within a cluster tends to be the same term in both studies. Altogether, this indicates that similar selective pressures may have shaped the allele frequency distributions in Europe and North America. Finally, we examine candidate genes that overlap between the North American cline, and extreme BFs for latitude and altitude in this study. We show that these 8 genes are functionally related to at least two other genes and propose a novel adaptive gene network.

Finally, we assessed the enrichment of GO terms and pathways for signals of adaptation from the perspective of all of their SNPs, not just the most highly significant candidate SNPs. We were most interested to find out if some of the most enriched gene sets (GO, pathway) remain enriched between the two types of enrichment analyses. We showed the importance of enrichments in pathways related to circadian rhythms, which seem to tie together all of our observations of local adaptive signatures in other traits. Finally, we aimed to showcase the advantages and the importance of using a set of different approaches for detecting selection.

# Abbreviations

*Adh* - Alcohol dehydrogenase

*Bayenv* - Bayesian environmental correlation

BF - Bayes factor

CCRT - chill-coma recovery time

*Ddc* - Dopa decarboxylase

*Catsup* - Catecholamines up

DGRP - *Drosophila* genetic reference panel

DNA - deoxyribonucleic acid

FR - France (Lyon) population

GO - gene ontology

GWAS - genome wide association study

KEGG - Kyoto encyclopedia of genes and genomes

NL - Netherlands (Leiden) population

RG - Rwanda (Gikongoro) population

RSS - resistance to starvation stress

*SimRel* - functional similarity relationship of GO terms

SNP - single nucleotide polymorphism

SR - startle response

UTR - untranslated region

ZI - Zambia (Siavonga) population

# List of figures

**Figure 3.13** Reactome pathways related to Transmembrane transport of small molecules and significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

**Figure 3.14** Reactome pathways related to Immune System and significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

**Figure 3.15** Reactome pathways related to Developmental Biology and significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

**Figure 3.16** Reactome pathways related to Neuronal System and to Hemostasis significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

**Figure S1** Proportions of genes supported by SNPs with strong evidence (*ln*(BF)>5 or $P < 0.0029$) for correlation with latitude (top panel) and altitude (bottom panel) that overlap with candidate genes from North America as quantified by $F_{ST}$ (Fabian *et al.* 2012).

**Figure S2** Proportions of genes supported by SNPs with strong evidence (*ln*(BF)>5 or $P < 0.0029$) for correlation with coldest month minimum temperature (top panel) and hottest month minimum temperature (bottom panel) that overlap with candidate genes from North America as quantified by $F_{ST}$ (Fabian *et al.* 2012)..

**Figure S3** Proportions of genes supported by SNPs with strong evidence (*ln*(BF)>5 or $P < 0.0029$) for correlation with yearly minimum temperatures (top panel) and yearly maximum temperatures (bottom panel) that overlap with candidate genes from North America as quantified by $F_{ST}$ (Fabian *et al.* 2012).

**Figure S4** Reactome pathways related to DNA repair significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

**Figure S5** Reactome pathways related to Cell Cycle, DNA Replication, and Programmed Cell Death significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

**Figure S6** Reactome pathways related to Organelle biogenesis and maintenance, Cellular responses to stress, and Extracellular matrix organization significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

**Figure S7** Reactome pathways related to Metabolism of proteins, Vesicle-mediated transport, and Cell-Cell communication significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

# List of tables

# CHAPTER 1: GENERAL INTRODUCTION

*Drosophila melanogaster* - an illustration (own work)

## 1.1 Population genetics and the new evolutionary synthesis

In the last decade of the 19th century and the first three decades of the 20th century, several important findings paved the way to modern genetics. In 1889, Hugo de Vries postulated “pangenes” based on Darwin’s Pangenesis theory (Stamhuis *et al.* 1999). In 1893, August Weismann developed germ plasm theory, demonstrating that inheritance is mediated only by gametes (Winther 2001). In 1900, Erich von Tschermak, a grandson of Eduard Fenzl, one of Mendel’s Viennese professors, rediscovered Mendel’s work on heredity (Harwood 2000). Around the same time, Mendel’s laws of heredity were

independently discovered by William Jasper Spillman, an American agronomist, as well as Hugo de Vries, a Dutch botanist who introduced the term “mutation” to biology, and Carl Erich Correns, a botanist from Munich who had to force de Vries to acknowledge Mendel’s primacy in the discovery (Lenay 2000). A few years later, in 1905, William Bateson was the first to use the term “genetics”, and did a lot to popularize Mendel’s work. He later also co-discovered genetic linkage, together with Reginald Punnett. Although he was to a great degree influenced by Darwin, Bateson did not believe that evolution was a gradual process mediated by natural selection. Rather, he was an adherent of saltationism, a view that held that evolution was a consequence of rather large and sudden changes from one generation to the next. Notable saltationists included de Vries, Punnett, Thomas Hunt Morgan, Charles Davenport, and Wilhelm Johannsen, a Danish plant physiologist who coined the terms “genotype” and “phenotype”, as well as the term “gene” in 1909. They all held that Mendelism and mutation were the most important evolutionary mechanisms. Opposed to them were, notably, Walter Weldon, Francis Galton, and Karl Pearson, the latter of which in 1911 founded the first University statistics department. They were gradualists, hugely influenced by Darwin (Galton was Darwin’s first cousin), and started using statistical modeling on various problems in biology, establishing the discipline of biometry. The debate between Mendelians and biometricians would finally be resolved by Fisher, Wright, and especially Haldane, who showed in “The Causes of Evolution” that Mendelian genetics works hand in hand with natural selection as the primary force of evolution. This new statistical framework finally united the fields of evolution and genetics.

The first seeds of the modern study of molecular adaptation could be traced to the precursor of the neo-Darwinian synthesis, in particular the work of Fisher, Haldane, and Wright (Figure 1.1). Their critical contributions explained continuous variation in terms of many discrete genetic loci (Fisher 1930), how selection operates in real-world examples of adaptation (Haldane 1932), and how drift and selection interact to push organisms towards their adaptive optima (Wright 1932). This work was the foundation of the new field of population genetics, making it possible for geneticists of the following decades to bridge the gap that existed prior to the 1930s between experimentalists (notably, Thomas Hunt Morgan) and naturalists (most notably, Francis Galton and Charles Darwin).

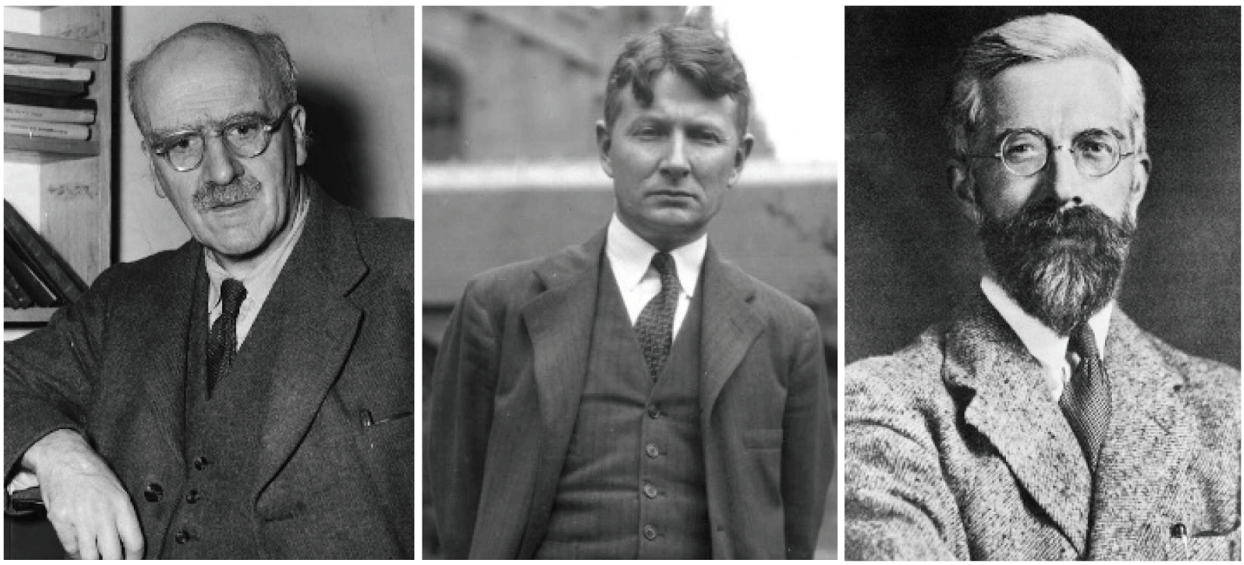


**Figure 1.1** Modern synthesists (from left to right): John Burdon Sanderson Haldane (1892-1964), Sewall Green Wright (1889-1988), and Sir Ronald Aylmer Fisher (1890-1962). From Whitfield (2008).

The theoretical framework laid out by Fisher, Haldane and Wright was not as impactful at first, largely because it had been formulated in highly abstract mathematical language few biologists could understand, it lacked empirical support, and did not address issues such as speciation and the role of geographic factors that were important for field naturalists (Hey *et al.* 2005). Nevertheless, by showing that genetic mechanisms were responsible for the geographic effects once attributable to Lamarckism, this research allowed Darwinism to be broadly accepted as the undisputed paradigm of biology (Bowler 1989). This integration of the approaches of experimentalists and naturalists was perhaps more important than the founding of population genetics itself in the creation of the new evolutionary synthesis (Greene 1981; Mayr & Provine 1981; Bowler 1989; Eldredge 1995).

## 1.2 Classical vs. balance hypothesis and empirical challenges in population genetics

Ernst Mayr (1959) criticized Fisher and Haldane's "beanbag" approach to genetics, with their emphasis on assigning fixed adaptive values on individual genes, and argued in favor of Wright's approach, which stressed the importance of interactions between genes, especially in small inbreeding populations. A set of Mendelian genetic factors with small effects on a phenotype could therefore generate small phenotypic differences to produce a continuous

range of variation. This line of thinking was important to get biometricians to accept Mendelian inheritance and a possibility that natural selection could act on such a wide range of genetic variation. However, laboratory geneticists such as Thomas Hunt Morgan and his student Hermann J. Muller continued to hold to the more refined “classical hypothesis” of population genetics (Muller 1949). This hypothesis held that genetic variability in most species is low, in contrast to the so-called “balance hypothesis”, which held that most loci are polymorphic and individuals typically heterozygous. The subsequent developments in molecular biology techniques that allowed for estimates of allele frequencies in natural populations showed that a fund of variability in each character already exists in any natural population, with many alleles circulating within the population in low frequency even if they confer no advantage whatsoever. When environmental conditions change, then natural selection has a ready supply of standing variation to act upon, and moreover, various forms of balancing selection actively maintain genetic variation even in the absence of environmental change (Bowler 1989).

The debate between the proponents of the classical and balanced views was a consequence of a central challenge for empirical population genetics: how to measure genetic variability to better understand selection, drift, and other forces (Avise 2004). Proponents of the classical view held that genetic variation in most species was low and that most individuals were homozygous at the majority of their loci for the same wild-type allele. Conversely, the balance view, championed by Dobzhansky, was that genetic variation is so high that most loci are polymorphic and no allele could properly be called “wild-type”, and that most individuals were heterozygous at close to 100% of their loci. Central to the disagreement was the concept of genetic load, a notion that variation on a population level would produce a heavy burden of reduced fitness thanks to many slightly deleterious alleles whose effect was too small to be eliminated from the population by purifying selection. Based on this, Muller predicted that only about 0.1% of all loci would prove to be heterozygous in most individuals. Accordingly, natural selection is mostly purifying selection in the classical view, and adaptive evolution due mostly to rare advantageous mutations that rapidly sweep to fixation. Since variability is so low, the predominant force in evolution would be *de novo* mutation, and recombination would be comparatively insignificant. In the balanced view, the predominant force would be different forms of balancing selection, producing polymorphisms through temporally or spatially varying

selection, heterosis, or frequency-dependent selection. Working on such high numbers of polymorphisms, recombination would then be far more important, and able to produce from generation to generation individuals with a large variability in fitness. Therefore, according to the balanced hypothesis, variation in a large population was not a curse, but rather a blessing, allowing it to rapidly respond to new and unpredictable environmental challenges.

## 1.3 *Drosophila* and the study of natural populations

The classical-balance controversy was related to other major questions in evolutionary biology, notably the debate between Fisher and Wright on the shifting balance theory, the dominance vs. overdominance hypothesis, and to a degree the later controversy surrounding the neutral theory of molecular evolution (Kimura 1983). A key development that came out of the modern synthesis was the introduction of *Drosophila* studies on evolution in natural populations (Crow 2008).

The first population genetics study of natural populations sought to measure the parameters that appeared in Wright's theory using data from Dobzhansky's studies of *Drosophila pseudoobscura* (Lewontin *et al.* 1981), paper number V). As a naturalist, Wright's model of adaptive landscapes appealed to Dobzhansky and he arguably did more than anyone else to popularize his views (Crow 2008). Similarly, Simpson popularized Wright's work among paleontologists.

For Wright, individual contributions of single genes to the phenotype were not as important as their interactions. A higher overall fitness coming from a favorable combination of alleles would be represented as a local peak on Wright's adaptive landscape. Since the totality of the interactions between these genes was beneficial, whereas the individual components were not, then the main question was how could natural selection drive the evolution to a higher adaptive peak, while crossing the less fit valleys. Wright's solution was encompassed in his shifting balance theory of population structure, which stated that the most beneficial state for the evolution of a population would be a metapopulation consisting of many smaller partially isolated subpopulations. From time to time, it would happen that one of the subpopulations would drift into a favorable combination of alleles purely by chance, at which point this subpopulation would

quickly grow, export a large number of migrants, and eventually the entire metapopulation would get fixed for the particular allelic combination. However, in order for the shifting balance theory to work, the metapopulation has to exist within relatively strict margins of subpopulation sizes, migration rates, and selection intensities, which has been shown to keep the overall fitness of the population low (Crow *et al.* 1990).

In contrast, Fisher put more emphasis on any individual allele change that might contribute to increasing the fitness of the population and refused to believe it would be possible for any single state of the population to exist where no further increase in fitness was possible. Because the environmental variables continually change, a large population would be more advantageous for evolution than a small subpopulation, because a large population holds more variability that natural selection could draw from, and because a large population is less susceptible to the effects of random genetic drift. Fisher importantly showed that selection was acting on the additive component of genetic variance, arguing that dominance variance and epistatic variance were comparatively unimportant. Kimura (1965) showed that a population under directional selection would tend to produce levels of linkage disequilibrium that would effectively cancel out the epistatic variance. Shifting balance is still a subject of debate today (Peck *et al.* 1998; Wade & Goodnight 1998; Coyne *et al.* 2000; Whitlock & Phillips 2000; Blum 2002; Johnson 2008; Chouteau & Angers 2012; Nahum *et al.* 2015). Modern population genetics has built on the foundations laid out by Fisher, Haldane and Wright in the 1920s, and perhaps in more than any other way by merging the theoretical predictions with empirical data. This empirical data was lacking at the beginning of what Huxley (1942) called “the modern synthesis”, which slowed the initial acceptance of the theory in the 1930s. Modern advances in technology have produced enormous quantities of molecular data, greatly strengthening the previously completely abstract models (Nordborg & Innan 2002). The field first started to take hold with the publication of Dobzhansky’s “Genetics and the origin of species” in 1937, which popularized Wright’s adaptive landscape paradigm among field biologists and naturalists. Another important development was the adoption of Dobzhansky and Wright of empirical data from natural populations of fruit flies (Lewontin *et al.* 1981), but a molecular evolution really took off in the mid 1960s with the introduction of protein polymorphism data, also in fruit flies (Lewontin & Hubby 1966), which sparked the

selectionist-neutralist controversy. A major problem prior to this era was how much variability there was in nature and how to measure it.

## 1.4 The importance of technology for population genetics

The theory of population genetics developed by Fisher, Haldane, Wright, Kolmogorov, Lewontin, Hamilton, Maynard Smith, Kimura, Moran, Kempthorne, Cockerham and others reached its golden years in the mid-1960s, seeking rules to generalize about the different contributions to adaptation of genetic drift, migration, mutation, and natural selection based on the position of the population in parameter space with a wide range of parameters (Felsenstein 2000; Kuhner *et al.* 2000). This was a necessary undertaking because, until the seminal paper of Lewontin and Hubby (1966), population geneticists lacked any sequence data. The development of computers that started to be used for genetic simulations in the 1950s (Barricelli 1954; Fraser 1957), combined with the arrival of electrophoretic and other techniques to estimate multilocus gene frequencies in the 1960s, gradually transformed the mostly theory-driven population genetics into a predominantly data analysis-driven evolutionary genetics (Felsenstein 2000).

As molecular biology matured in the 1950s, starting to uncover the physical and chemical nature of genes, the widening application of molecular techniques fundamentally changed the nature of the questions asked by evolutionary geneticists (Crow 2008). By the mid-1980s, almost a thousand different species were studied by electrophoresis (Nevo *et al.* 1984), showing larger than expected amounts of variability, but not being able to resolve the controversy of the classical vs. balanced hypotheses. What happened instead was that the questions surrounding genetic variability were largely dropped in favor of inquiries surrounding the relative contributions of drift and selection to evolutionary change (Crow 2008). Kimura (1968) and Kimura and Ohta (1969) related the intraspecies thinking of evolutionary geneticists and the interspecies thinking of molecular evolutionists, effectively unifying the insights coming from electrophoretic studies and the studies of sequence evolution (Felsenstein 2000). The neutral theory of molecular evolution acted as a catalyst for more data to flow into the field of evolutionary genetics, eventually precipitating the rise of evolutionary genomics. With the publication of the pioneering study by Martin Kreitman (1983) on the polymorphism at the *Adh* locus in populations of *Drosophila melanogaster* along a latitudinal cline, evolutionary genetics made a full circle that was initially started all

the way back in the 1930s by Dobzhansky and Wright. The stage was set for the further advancements in sequencing technologies to finally allow *bona fide* population genomics approaches in various organisms to identify single nucleotide polymorphisms (SNPs).

## 1.5 Local adaptation in natural populations

In his seminal book "Adaptation and Natural Selection" (1966), George C. Williams established what was later to be called a "gene-centered" view of evolution, and would play an important role in the group or multi-level selection debate. Equally importantly, however, this was a pioneering work outlining systematic and rigorous criteria for the analysis of local adaptation. It pointed out that divergent selection on local populations that differ in habitat would result in genotypes from the local habitat exhibiting higher relative fitness in their own habitat than genotypes from other habitats, regardless of the consequences of the traits that those genotypes underlie in other habitats (Williams 1966; Kawecki & Ebert 2004).

For example, directional selection for cold tolerance on local populations would result in individuals from a temperate habitat exhibiting, on average, higher relative fitness in European habitats than individuals coming from a tropical habitat. Even if a species is distributed in a continuous range along a gradient of environmental variables, local populations sampled from different points of the gradient may still show patterns of local adaptation. In addition, the classical study of adaptation - an improvement in phenotype that increases its relative fitness - involves the consideration of the historical role of natural selection in this process. Consequently, this methodology has to compare the derived populations to the ancestral ones, as well as their intermediate, presumably less well adapted ancestors. An alternative is offered by the local adaptation approach, which compares local populations that have evolved under disparate environmental conditions, and which - if detected - offers strong evidence of recent selection for specific environmental variables (Kawecki & Ebert 2004).

Another controversy with important implications for the detection of local adaptation arose in the early 1990s centered on the problem of explaining the interesting observation that genomic regions which have reduced levels of genetic variation (lower than what we would expect under neutrality) also experience lower rates of recombination. Charlesworth

and coworkers (1993) suggested that the reason is a form of strong negative selection, termed background selection, that removes recurrent deleterious alleles in a process balanced out by mutation. This has a similar effect on neutral variants linked to the selected site as the competing selective sweep model, in that both lead to a local reduction of genetic variation. Before the advent of modern sequencing technology, a lot of efforts were centered on trying to determine the relative importance of background selection versus selective sweeps (Stephan 2010). Regardless, both models were in conflict with Kimura's neutral theory, as they predicted that neutral traits could be affected by selection acting on linked non-neutral variants. This has important ramifications for the study local adaptation. The fact that local adaptation increases $F_{ST}$ even at neutral variants far away from the focal locus, and that background selection increases $F_{ST}$ as a consequence of reducing intra-population diversity means that those effects can be distinguishable from one another, as well as from balancing selection, which will leave diversity peaks in the region of the focal polymorphism within, as opposed to between subpopulations (Charlesworth *et al.* 1997). These theoretical considerations, together with the availability of cheaper and higher quality DNA sequence data on a genome-wide level, have led to the development of many new methods for the detection of natural selection (Vitti *et al.* 2013).

## 1.6 Adaptation in *Drosophila melanogaster*

*Drosophila melanogaster* is an ecological generalist and opportunist, and being a human commensal, it possesses an extraordinary colonizing ability that made it a cosmopolitan species (Lachaise *et al.* 1988; Lachaise & Silvain 2004; Keller 2007). Historical biogeographic and systematic studies suggested that the species originated in the tropics of sub-Saharan Africa, extending its range outside Africa with vegetational-climatic changes in late Pleistocene, and colonizing Europe in early Holocene, after the last glaciation (David & Capy 1988). These studies were subsequently corroborated by large-scale microsatellite, multilocus SNP, and later whole-genome DNA sequencing approaches (Glinka *et al.* 2003; Ometto *et al.* 2005; Pool *et al.* 2012), clearly supporting the hypothesis that *Drosophila melanogaster* originated in Africa as a tropical species and then colonized the rest of the world only relatively recently (Stephan & Li 2007).

A clear idea of the biogeographic and demographic history of *Drosophila melanogaster* is important to be able to make inferences about adaptation to spatially varying

environments (Adrion *et al.* 2015). Connecting the genetic variation underlying locally adapted populations with their phenotypic and fitness variation, for instance by sampling individuals from populations along clines, has been a successful approach (Endler 1977; Keller *et al.* 2013; Adrion *et al.* 2015; Porcelli *et al.* 2016). In the case of a cosmopolitan species such as *Drosophila melanogaster*, the expectation is that populations will be connected by significant gene flow, and that their optimum fitness will shift gradually with the environmental gradient. Such continuous-environment clines might still get sharp under abrupt changes in environmental variables. Clines of causative genetic variants should closely follow the environmental cues, while clines in selectively neutral variants should not (Adrion *et al.* 2015). Since the genetic model that underpins continuous environmental clines is likely to have a quantitative genetic architecture, an interesting empirical question is if and how all of the variants of these quantitative traits will follow the environmental gradient (Adrion *et al.* 2015). In order to answer this question, a key challenge we need to tackle is how to identify true causal variants of polygenic adaptation.

## 1.7 Detecting adaptive footprints of polygenic adaptation

Approaches to detect adaptive footprints have commonly used genome-wide scans of large numbers of molecular markers (such as SNPs) to screen for $F_{ST}$ outliers. Because frequency differences between populations could be caused by genetic drift, local $F_{ST}$ values must then be compared to a genomic background distribution. This approach relies on the fact that changes in allele frequency caused by natural selection should be local, whereas those caused by drift should encompass the entire genome. In addition, in cases where we can observe a clear phenotype that differs substantially between individuals from different populations, we can test the frequency shifts across multiple populations for correlations with environmental variables (Coop *et al.* 2010).

Environmental variables such as temperature, as well as variables that correlate with temperature, are of crucial importance for the ability of organisms, especially ectotherms, to survive and reproduce. For this reason, it is particularly interesting to examine how ecologically important traits respond on a genetic level to environments with different abiotic conditions (Stinchcombe & Hoekstra 2008). Due to increasingly lower costs of sequencing, it is now feasible to generate datasets of a sufficient number of fully sequenced individuals from different environments, such that estimates can be made

about unusual patterns of variation at individual SNPs that suggest adaptation to specific environmental variables. Such population genomics scans can be supplemented by data from quantitative genetics studies (Mackay *et al.* 2012), and candidate genes can further be examined in the context of the biochemical processes and pathways they are involved in.

An additional level of complexity is added by the fact that ecologically important traits can be polygenic, and genomic scans for footprints of local adaptation that focus on areas of reduced variability (i.e. selective sweeps) are in most cases not powerful enough to detect polygenic adaptation (Stephan 2015). In polygenic adaptation, loci with moderate to large effects are expected to be rare. Instead, what we expect to happen with polygenic adaptation are slight shifts in frequency at many loci with small individual effects on the phenotype (Pritchard & Di Rienzo 2010). The amount of change will depend on the distance of the trait to phenotypic optimum, and the effects of individual loci, but it is possible that small effects might on aggregate significantly influence the fitness of the phenotype as a whole. Taking all of this into consideration, it is not surprising that to detect such subtle shifts in allele frequencies over many sites would be quite cumbersome, especially because the same effects could be a consequence of genetic drift. However, observing such frequency changes in functionally relevant alleles from individuals fine-tuned to their respective environments can give us valuable insight into the causes and consequences of adaptation.

A common approach for biological interpretation of the results of various genome scans, which can sometimes be very long lists of candidate genes, is to test gene ontology terms for an overrepresentation of these genes. A simple gene ontology enrichment analysis might however lead to spurious conclusions allowing relatively easily for an ostensibly biologically meaningful, but still *ad hoc* interpretation (Pavlidis *et al.* 2012). In other words, it is possible for methodologically flawed computational and statistical approaches to identify high rates of false positive candidate genes, and then to construct a biologically sensible explanation of their results, which in turn validates the approach a posteriori (Pavlidis *et al.* 2012). For this reason, it is important to augment such analyses with additional sources of information, e.g. by cross-validation of significant GO terms using candidate genes from other similar studies. Additionally, recent studies on human population differentiation showed that in the analysis of signals of adaptation, classical GO enrichment approaches should be supplemented with a gene-set enrichment analysis that includes all SNPs related

to the GO term (Daub *et al.* 2013, 2015; Amorim *et al.* 2015; Dopazo *et al.* 2016). This is especially important in the context of ecologically relevant traits, which are mostly polygenic. As mentioned before, it is likely that such traits would adapt to local conditions by small shifts in allele frequencies across many loci. For this reason, a classical GO analysis, which by design focuses only on the most differentiated loci, would not be well-suited for detecting footprints of polygenic adaptation.

## 1.8 Cold tolerance traits in *Drosophila melanogaster*

### 1.8.1 Chill-coma recovery time

Almost every species needs to be able to adapt to thermal conditions (Colinet *et al.* 2010a). Studying insect temperature tolerance has given us important insights into thermal adaptation (Angiletta 2009). The so-called chill-coma recovery time (CCRT) is a metric widely used to quantify the tolerance to low, but non-lethal temperatures in ectothermic organisms. Most species of insects, including *Drosophila melanogaster*, are not able to use ice nucleating agents or antifreeze proteins to cope with the cellular damage caused by the formation of ice particles in their cells. Instead, they had to evolve a different mechanism to be able to survive periods of cold stress in temperate environments. At 0°C, as a result of the disruption in nerve and muscle excitability, fruit flies fall into a narcosis state called chill coma. This is in almost all cases a reversible state, and for this reason it is an important mechanism for winter survival (Angiletta 2009). Despite the wide use of CCRT in assessing cold tolerance (Macmillan & Sinclair 2011), the physiological details of how chill coma comes about are not well understood. However, it is likely driven by temperature effects on cell structures that maintain ion and water balance: the lipid membrane, its ATPases, and gated ion channels (Macmillan & Sinclair 2011). Briefly, at a certain low temperature point, a disruption occurs in the neuromuscular system as a consequence of direct effects of low temperature on ion pumps or ion channels and/or indirect effects on all membrane-bound proteins through changes of membrane fluidity.

Behavior, ecology, and fitness of all ectotherms is profoundly influenced not only by extremely low or high temperatures, but also by temperatures that merely approach those thermal limits (Angilletta 2009; Huey 2010). One key issue when considering insect adaptation to low temperature is the question of whether there is a correlation between

latitude and the optimal, as well as critical temperature limits. *Drosophila melanogaster*, like many other insects, is of tropical origin, but has invaded colder habitats following the last ice age. The large daily and seasonal variations in temperature that ectotherms like *Drosophila* encounter in the course of such invasions inevitably reduce their overall physiological performance to the extent where they may become locally extinct (Huey 2010). There are two main ways to answer those challenges: (1) increase the capacity for acclimation (by changes in behavior, biochemistry, and/or physiology), and (2) adapt genetically to reduce thermal sensitivity.

Analyses of the thermal sensitivity of Darwinian fitness, defined as the intrinsic rate of population growth (r), showed that in various species of insects, the optimal temperature that maximizes r decreases with absolute latitude (Huey & Berrigan 2001; Frazier *et al.* 2006; Deutsch *et al.* 2008; Huey 2010). Critical maximums and minimums also decline with latitude, and the decline is much more dramatic for minimums, which reflects the fact that yearly minimum temperatures decrease with latitude more sharply than maximum temperatures. However, to accurately gauge the thermal sensitivity of the fitness of most species is in a majority of cases infeasible. This is because constructing such a fitness curve requires a measurement of lifetime reproductive success in clones of individuals raised under controlled environments with varying temperature (Angilletta 2009). Additionally, due to the confounding effects of multiple biotic and abiotic factors that cannot be controlled in the laboratory, we cannot even be sure that those fitness curves accurately represent the state of nature. But most importantly, in almost all cases we are not so much interested in how fitness evolves, rather we want to know about the evolution of particular phenotypes that contribute to it (Angilletta 2009). In *Drosophila melanogaster*, chill coma recovery has been shown to relate negatively to latitude (Hoffmann *et al.* 2002, 2003; Ayrinhac *et al.* 2004), and that its tolerance to heat has evolved in parallel with tolerance to cold (Bubliy & Loeschcke 2005). Also interestingly, it has been found that diet supplementation with the amino-acid proline (Koštál *et al.* 2012) or artificial expression of antifreeze proteins of other insect species (Nicodemus *et al.* 2006) can transform *Drosophila melanogaster* into a partially freeze-tolerant organism.

An examination of selection pressures insects like *Drosophila* - originally tropical species - had to endure to survive in colder environments, is one of the key interests of evolutionary physiology. A macrophysiological approach to the study of insect adaptation

to low temperatures (Chown *et al.* 2004) shows that both optimal and critical temperatures correlate inversely with latitude (Angilletta 2009), indicating that high-latitude species have adapted evolutionarily to low temperatures (Huey 2010). Such insects, including *Drosophila melanogaster*, encounter multiple physiological challenges: seasonal bouts of low ambient temperature, large variance in daily temperatures, shorter growing seasons, and disruptions in photoperiodic timing, all of which reduce overall physiological performance and restrict population growth (Janzen 1967; Ragland & Kingsolver 2008; Huey 2010).

To adapt to colder temperatures, insects can either increase their ability to acclimate, or adaptively change their thermal sensitivity. In the first case, they may alter their behavioral patterns to avoid the cold, adjust morphological and physiological plasticity to tolerate temperature shifts, and compensate their response to photoperiodic cues in order to properly time reproduction, diapause and other seasonal activities. Alternatively, given enough time, adaptive shifts can occur in the genetic variation underlying the physiological or biochemical performance in the face of cold stress. Additionally, the ability for and the extent of a plasticity response to a cold environment both contain a genetic component and should thus be under selection in natural populations (Gabriel 2005; Overgaard *et al.* 2010).

### 1.8.2 Resistance to starvation stress

Suboptimal quantity and quality of food is a ubiquitous source of stress for insects, and even herbivorous species have to periodically cope with the lack of essential nutrients, which is why, like traits related to other forms of environmental stress, resistance to starvation stress (RSS) is a trait expected to be under strong natural selection (Hoffmann *et al.* 1991; Randall *et al.* 2002; Rion & Kawecki 2007).

RSS is normally quantified as the time until death under acute starvation, but with sufficient water provided, and can vary anywhere between 20 and 200+ hours (Harbison *et al.* 2005; Rion & Kawecki 2007). Females survive longer than males, and individuals kept under calorically restrictive diets longer than those on rich media. Experimentally selected flies also survive longer, normally after several generations of breeding the top 10%-50% of adults that survived total food deprivation (Bubliy & Loeschcke 2005). The same study showed that flies selected for increased cold shock resistance became more resistant to starvation (Bubliy & Loeschcke 2005). However, the picture seems to be more complex. For

example, (Hoffmann *et al.* 2005) found a negative correlation between RSS and cold stress measured by mortality in female flies selected for either trait, (Broughton *et al.* 2005) found flies whose insulin-producing cells have been ablated to put up more lipid reserves and be more resistant to starvation, but also less to cold and heat, and (Ayroles *et al.* 2009) found a tendency among flies with shorter CCRT (i.e. more resistant to cold) to have higher competitive fitness, but lower lifespan and RSS. Additionally, genetic variation of isofemale lines derived from populations sampled along a latitudinal and altitudinal gradient in South America showed an opposing latitudinal cline for RSS (Goenaga *et al.* 2013).

### 1.8.3 Startle response

Startle response (SR) is a type of vigorous reaction activity in *Drosophila melanogaster* induced by a mechanical disturbance, and measured immediately following it (Meehan & Wilson 1987). Locomotor activity encompasses various traits and behaviours that can last from a few seconds to a few days, and that differ with regards to speed, directionality, and amount of activity (Jordan *et al.* 2006). Startle response and other locomotory behaviors play an important role as the components of fitness, because they encompass a wide range of activities related to finding food, finding mates, defending one's territory, and escaping from predators and other dangers (Gilchrist *et al.* 1997; Gibert *et al.* 2001a; Jordan *et al.* 2006).

SR has been used extensively in quantitative genetics of *Drosophila* as a phenotype to study interactions between the nervous system, genes, and behavior (Jordan *et al.* 2006; Yamamoto *et al.* 2008, 2009; Swarup *et al.* 2012; Ober *et al.* 2012; Huang *et al.* 2012a). *Drosophila melanogaster* is a particularly good model to study the genetic architecture of behavioral traits such as SR, thanks to the possibility to combine approaches from genetics (can easily control for genetic background, as well as environment), and approaches from neuroanatomy (Yamamoto *et al.* 2008). In terms of adaptation to changing environmental variables, behavioral traits that make it possible for a fly to escape harmful conditions will obviously be useful for survival. Since the capacity for locomotion might have fitness trade-offs with both CCRT and RSS (Gibert *et al.* 2001a; Carbone *et al.* 2006; Adrion *et al.* 2015), it would not be unreasonable to expect that locomotory traits might locally adapt to colder temperate environments, particularly in an ectothermic species such as *Drosophila melanogaster*. For example, walking speed has been compared between flies from tropical

and temperate populations, reared at different temperatures, and also measured at different ambient temperatures and different ages (Gibert *et al.* 2001a). Flies from a natural population in France showed an inverse relationship between walking speed and developmental temperature, and those raised at a lower temperature performed better than flies from the Congo. Congoan flies raised at an intermediate temperature were even better, but this effect fell with age, to the point where speed actually increased with age for flies raised and maintained at low temperature (Gibert *et al.* 2001a). Altogether, this suggests that evolutionary responses to selection for locomotor performance might be difficult to predict.

Further understanding of SR has come from quantitative geneticists, who have attributed the variation in locomotion to multiple interacting quantitative trait loci (QTL) with environmentally sensitive effects (Jordan *et al.* 2006). The QTL approach to identify regions of the genome responsible for the variation in this trait measures a less precise, and more general behavioral response compared to the approach of behavioral geneticists (Baker *et al.* 2001). The reason is that quantitative geneticists had to optimize their assays for measuring startle response to be more rapid and high-throughput (Jordan *et al.* 2006). In addition, different kinds of startle behaviors are frequently mentioned in literature pertaining to circadian rhythms, in fact they are some of the most commonly used metrics for determining clock parameters in various species, including *Drosophila (Chiu et al. 2010)*.

QTLs for SR have been fine mapped to multiple candidate genes, the most interesting of which is *Dopa decarboxylase* (*Ddc*) (Jordan *et al.* 2006). This gene codes for an enzyme that catalyzes the synthesis of dopamine from L-dopa, as well as the final step in the synthesis of serotonin. Neurotransmitters such as dopamine affect locomotor activity of *Drosophila* in various ways (Connolly 1966, 1967; Partridge *et al.* 1987). For instance, walking performance is correlated with male mating success (Partridge *et al.* 1987) and, as mentioned, is affected by temperature (Crill *et al.* 1996; Gibert *et al.* 2001a).

Another gene, *Catecholamines up* (*Catsup*), has been studied for its effect on locomotion and other traits, and the results pointed to the importance of assessing the influence of variants with small effect sizes on the associated quantitative traits (Carbone *et al.* 2006). The protein product of *Catsup* regulates tyrosine hydroxylase, reducing the synthesis of the neurotransmitter dopamine, which reduces locomotory behaviors, including SR. It is a

highly pleiotropic gene, whose polymorphisms are associated with locomotor behavior, but also RSS, longevity, and the number of abdominal and sternopleural sensory bristles (Carbone *et al.* 2006). Interestingly, recombination between the causal polymorphisms inside the gene is so large that it allows them to evolve almost independently (Carbone *et al.* 2006). Furthermore, at least six of these sites affect RSS with individually small effects, but the two most extreme haplotypes have very different mean survival times (approx. 100 h and 38 h), indicating that the overall effect of these polymorphisms is significant. Yet only two of the six observed associations would have been detected with a common design of only genotyping common variants (Carbone *et al.* 2006).

The *Catsup* gene is an example of why, in order to properly understand the effects of candidate loci on complex traits, one has to take into account the individual polymorphisms at all of the sites related to the trait. Understanding the complexity of the relationships between SNPs, genes, gene sets, cold tolerance phenotypes, the environment, and fitness requires a multifaceted approach that includes different levels of organization. This was the main motivation behind this work.

## 1.9 The aim of this study

The principal aim of this study was to uncover the genome-wide basis of adaptation to cold in European populations of *Drosophila melanogaster* (Figure 1.2), with a particular interest in polygenic adaptation of cold tolerance traits. Chill-coma recovery time is one of the most commonly researched resistance traits, but the physiological mechanisms, genetic control, and genetic variability of CCRT are not well known. In this work, we have aimed at better connecting genotypic to phenotypic variability to study the evolution of cold tolerance. *Drosophila melanogaster* is particularly well suited for studies of cold adaptation and has well-described clinal distributions along latitudinal and altitudinal gradients (Gibert *et al.* 2001b; Hoffmann *et al.* 2002, 2003; Hoffmann & Weeks 2007).

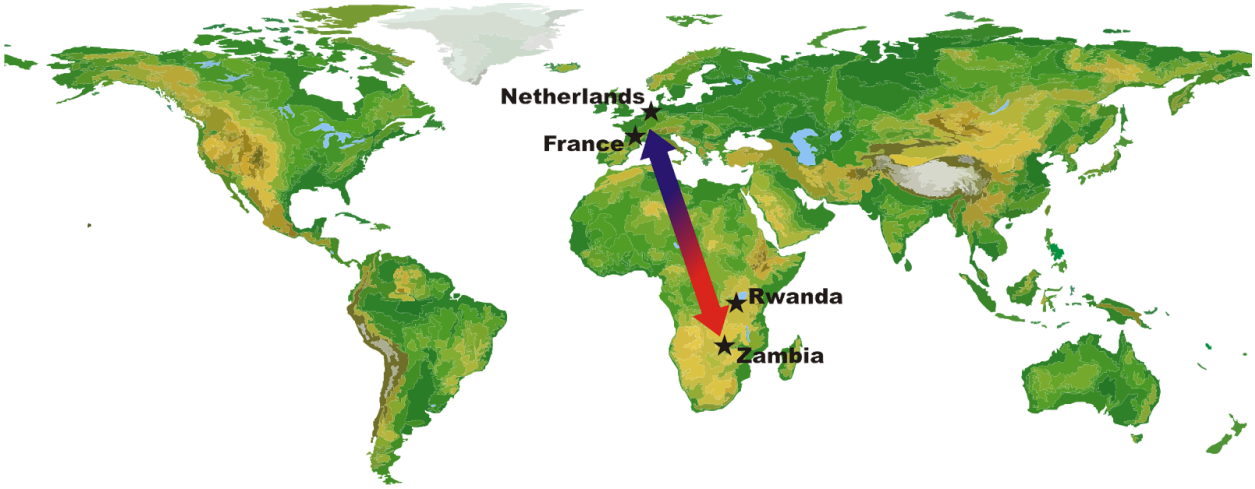



**Figure 1.2** Populations of *Drosophila melanogaster* used in this study.

Previous studies have characterized many genes involved in temperature tolerance, starvation stress, and mechanical disturbance, using a multitude of techniques: physiology (Lee *et al.* 2009), *P*-element insertions (Alic *et al.* 2014), RNA interference (Fedotov *et al.* 2014), transposon insertions (Yamamoto *et al.* 2008), quantitative trait locus mapping (Wilches *et al.* 2014), mutant complementation tests (Fallis 2012), genome-wide association studies (Organisti *et al.* 2015), exon expression (Telonis-Scott *et al.* 2013), expression knockdown (Colinet *et al.* 2010b), microarrays (Gates *et al.* 2011), behavioral assays (Meyer *et al.* 2014), proteomics (Colinet *et al.* 2013), transcriptomics (Vermeulen *et al.* 2013), targeted cell knockout (Zhang *et al.* 2013), selective sweep mapping (Svetec *et al.* 2011),

latitudinal variation (Fabian *et al.* 2012), and numerous others. Our approach used high quality genome-wide SNP data from multiple populations (Pool *et al.* 2012), and GWAS of CCRT, RSS, and SR (Huang *et al.* 2014) to carry out a genome-wide scan for polygenic adaptation.

### 1.9.1 Adaptation on gene level

Firstly, cold adaptation is examined focusing on genetic variation between tropical and temperate populations of *Drosophila melanogaster*, pertaining to three phenotypes known to be adaptive: CCRT, RSS, and SR. The aim was to accumulate evidence for annotated genes that may have experienced local adaptation in these phenotypes.

In all analyses, genetic variation was investigated by means of two test statistics. Firstly, pairwise $F_{ST}$ was estimated between African and European populations of *Drosophila melanogaster*. Such genome-wide scans of large numbers of molecular markers, in our case SNPs, have previously been widely used to screen for $F_{ST}$ outliers (Luikart *et al.* 2003; Storz 2005; Stinchcombe & Hoekstra 2008; Holsinger & Weir 2009; Lotterhos & Whitlock 2014). Effects of random genetic drift on allele frequency differences between populations must be taken into account in genome scans for selection, because factors such as population structure, bottlenecks, migration, and population expansions can create patterns that mimic selection on the genetic level (Teshima *et al.* 2006; Excoffier *et al.* 2009; Elhaik 2012; Vitti *et al.* 2013). Such effects can be disentangled from selection by, for example, comparing local $F_{ST}$ to the genomic background distribution (Holsinger & Weir 2009), by simulation of locus-specific, population-specific, and combined effects on $F_{ST}$ (Riebler *et al.* 2008), or by directly estimating the probability of selection for each locus (Foll & Gaggiotti 2008; Villemereuil & Gaggiotti 2015). Additional caution is necessary to guard against reduced statistical power and false positives due to background selection (Stephan 2010; Huber *et al.* 2016), intrinsic genetic incompatibilities (Bierne *et al.* 2011), genetic architecture of quantitative traits (Le Corre & Kremer 2012; Gagnaire & Gaggiotti 2016), and reduced recombination in the vicinity of centromeres (Roesti *et al.* 2012; Hemmer & Blumenstiel 2016).

In the second method used here, correlations between allele frequencies and different environmental variables were quantified by means of *Bayenv2 (Coop et al. 2010; Günther & Coop 2013)*. With an assumption that there is a phenotype that differs between certain

populations, and especially in the case of a complex phenotype, this is a powerful approach to test for selection because it tests for frequency shifts across all populations at the same time, and accounts for possible effects of neutrality (Berg & Coop 2014). In addition to environmental variables, different phenotypic distributions can also be used in such analyses (Mendizabal *et al.* 2012; Villemereuil *et al.* 2014).

The focus in the first part of the results is on three phenotypes that are known to be adaptive and related to cold adaptation: chill-coma recovery time (CCRT), resistance to starvation stress (RSS), and startle response (SR) . Chill-coma is a reversible narcosis state that helps fruit flies to cope with low, non-lethal temperatures (Gibert *et al.* 2001b). Measuring the time it takes for an individual to wake up from this state is a commonly used metric for assessing cold tolerance (David *et al.* 1998; Anderson *et al.* 2005; Rako & Hoffmann 2006; Macmillan & Sinclair 2011). Because it is reversible, it has become an important mechanism for survival in temperate climates (Bale 1993). Similarly, RSS and SR have important direct fitness consequences (Rion & Kawecki 2007; Kenny *et al.* 2008; Schwasinger-Schmidt *et al.* 2012).

### 1.9.2 Adaptation on gene network level

Secondly, we assess the enrichment of gene ontology terms and pathways for signals of adaptation using a gene-set enrichment approach. Rather than focusing only on outlier SNPs to define candidate genes in a classical overrepresentation analysis, we instead considered GO terms and pathways as gene sets defined by all of their underlying SNPs. We were most interested to find out if sets of SNPs defined by gene sets (GO terms and pathways) are enriched in SNPs with Bayes factors for environmental variables, which might suggest local adaptation to European environments.

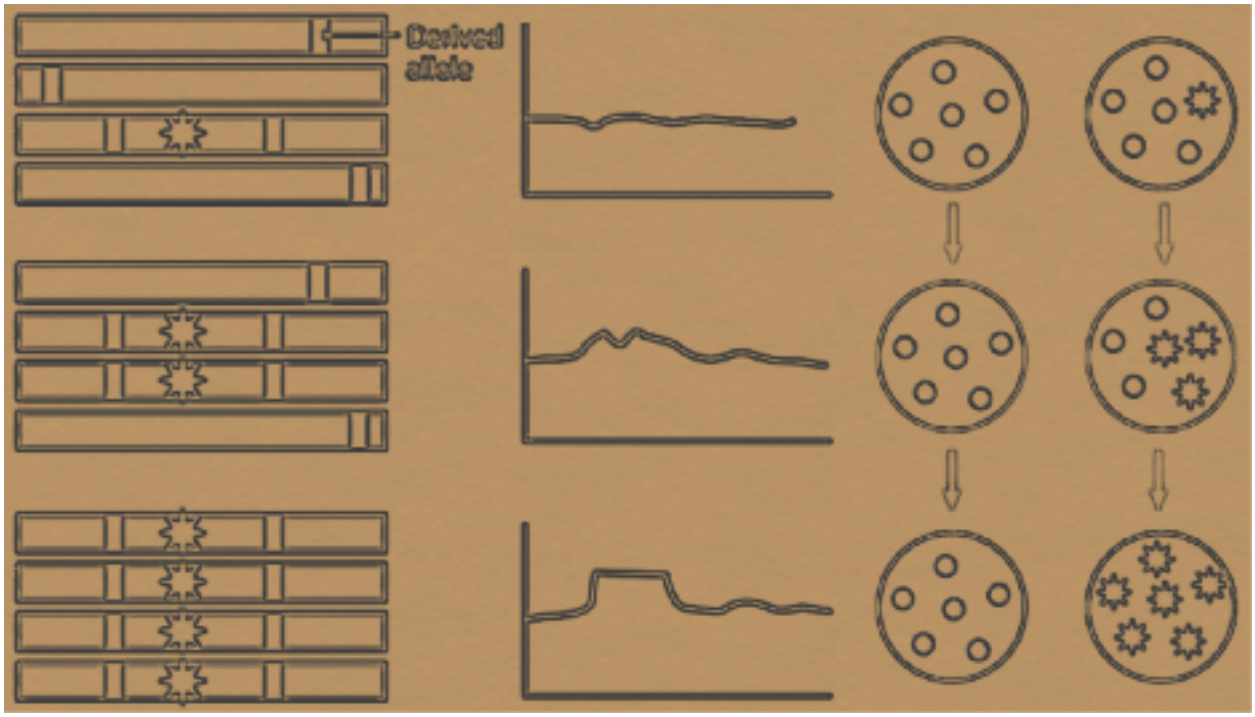
Recent studies on human data (Daub *et al.* 2013, 2015) indicate that gene-set enrichment for all of their genes produces very different results from the classical GO enrichment of only the top-scoring genes (candidate gene enrichment). For instance, Daub *et al.* (2013) showed that out of the 100 most significant genes in their study, only 14 genes were found related to their most significantly enriched candidate gene sets (retrieved from NCBI BioSystems database), and that a GO enrichment analysis of these 100 genes

revealed no significant biological processes. Using insights from these studies, we have carried out a similar gene-set enrichment analysis in *Drosophila melanogaster*.

Effectively, we treated each gene set, defined as either a gene ontology term or a pathway, as a "phenotype" whose selection footprint is the additive contribution of Bayes factors (as calculated by *Bayenv2*) of all of its respective SNPs. In this way, gene sets (effectively, SNP sets) that might not have been significant in the classical gene ontology enrichment analysis might still show an enrichment in candidate SNPs, as the contributions of all SNPs, and not just SNPs that map to top candidate genes, are taken into account.

This is particularly important in the context of possible polygenic adaptation (Pritchard & Di Rienzo 2010; Berg & Coop 2014; Stephan 2015). Namely, if adaptation proceeds by small allele frequency shifts spread out across many loci, then the additive contribution of all of these loci to the trait have to be taken into account. This might especially be the case for traits, such as cold tolerance, which are probably related to a number of biochemical pathways that direct almost every organ system in the body. Various studies have shown that complex phenotypes are able to adapt to local conditions to a considerable extent by covariance in small allele frequency changes (Barton & Bengtsson 1986; Latta 1998; Le Corre & Kremer 2003, 2012; Yeaman 2015). To characterize the pathways by means of which these changes might occur was the major aim of this work. Finally, we wanted to ascertain the extent of possible polygenic adaptation across the genome, and compare the results between different approaches to gene ontology and pathway enrichment.

# CHAPTER 2: MATERIALS AND METHODS




Selective sweeps - an illustration (adapted from Vitti *et al.* 2013)

## 2.1 Adaptation on gene level

Firstly, we investigated adaptation to cold on the level of individual genes by estimating the amount of population differentiation at SNPs across the genome, as well as the correlation of allele frequencies across the tested populations with environmental variables. We started from individual SNPs associated with CCRT, RSS, and SR, and broadened our analysis to genes, and then gene ontology terms and metabolic pathways. On the whole, we sought to clarify the intricacies of a multitude of selective pressures that we suspected might work in concert on many genes across the genome.

### 2.1.1 Data analysed

Genome-wide SNP data were generated for two African and two European populations of *Drosophila melanogaster*. Data for Gikongoro, Rwanda (RG, 27 lines), Siavonga, Zambia (ZI, 27 lines), and Lyon, France (FR, 8 lines) were taken from the Drosophila Genome Nexus (http://johnpool.net/genomes.html; (Lack *et al.* 2015)). For these populations, female flies were collected between 2008 and 2010 and iso-female lines were established. Whole-genome sequences for individual lines were created by next-generation sequencing of haploid embryos (Langley *et al.* 2011) and assembling of sequencing reads into full genomes by mapping them to the *D. melanogaster* reference sequence (Lack *et al.* 2015). Some of the African lines were found to harbor stretches of European admixture (Pool *et al.* 2012; Lack *et al.* 2015). These genomic regions were masked in the analysis here. Additionally, a population from Leiden, the Netherlands (NL, 11 lines) was analyzed, for which females were collected in 1999 (Bubliy & Loeschcke 2000) and kept in the lab as lines of iso-females. These fly lines were subjected to fifteen generations of full-sib mating, followed by the generation of whole-genome sequences by next-generation sequencing of adult flies, and genome assembly as detailed above (Voigt *et al.* 2015). Across all the populations and chromosomes, 9,995,420 polymorphic sites were observed. Genotypes with a PHRED score of less than 31 (≈0.08% chance of calling a base incorrectly) were considered as missing data (Ewing & Green 1998). Sites that were segregating for more than two alleles were excluded, as well as sites that contained more than 10% missing data across the populations considered. For *Bayenv2* analyses, sites that were not polymorphic across the tested populations were further excluded, as well as singletons (over all populations). After quality filtering, 3,663,890 polymorphic sites were finally retained for the autosomes and 867,049 for the X chromosome for pairwise $F_{ST}$ analyses, and 313,972 polymorphic sites for the autosomes and 39,304 for the X chromosome for *Bayenv2* analyses (Table 2.1).

From the filtered genomic background, subsets of SNPs were defined (Table 2.1) that were previously associated with the following quantitative traits: (i) time to recover from chill coma (CCRT), (ii) resistance to starvation stress (RSS), and (iii) startle-induced locomotor response (SR) (Huang *et al.* 2014). These SNPs, which were originally ascertained from a North American population, were defined by using their corresponding positions in our

European and African populations of interest. Annotation data that was used was as recorded in Flybase release 5 (Pierre *et al.* 2014).

**Table 2.1** Size of the trait-associated SNP datasets before and after filtering.

| Trait | Association study | Total SNPs associated | Associated SNPs after filtering ($F_{ST}$) | Associated SNPs after filtering (*Bayenv2*) |
|---|---|---|---|---|
| Chill Coma Recovery Time (**CCRT**) | Huang *et al.* 2014 (*DGRP Freeze 2.0*) | 119 | 59 | 14 |
| Resistance to Starvation Stress (**RSS**) | Huang *et al.* 2014 (*DGRP Freeze 2.0*) | 132 | 64 | 14 |
| Startle Response (**SR**) | Huang *et al.* 2014 (*DGRP Freeze 2.0*) | 78 | 51 | 12 |
| Total genomic background after filtering | | | 3 663 890 (autosomes) + 867 049 (X chromosome) | 313 972 (autosomes) + 39 304 (X chromosome) |

### 2.1.2 Statistical analyses

The amount of population differentiation was estimated between pairs of populations on SNPs across the genome using the Weir and Cockerham estimator of $F_{ST}$ (Weir & Cockerham 1984), which is known to be unbiased with respect to sample size (Willing *et al.* 2012). Furthermore, it is well known that the distribution of $F_{ST}$ can be skewed, for example by sub-selection of SNPs from GWA studies (Clark *et al.* 2005; Elhaik 2012). Demographic events, such as population bottlenecks, affect the allele frequencies patterns that are also reflected in $F_{ST}$ estimates. Therefore, to derive unconfounded *P*-values from the observed $F_{ST}$ distribution, a resampling approach was used here to assess the null distribution (i.e. the genomic background). For every trait-associated set of SNPs (3 traits × 4 population pairs = 12 SNP sets), mean $F_{ST}$ was calculated over each of 12 × 10000 SNP sets of random background SNPs. The quantile of mean $F_{ST}$ of trait-associated SNPs in the empirical distribution of 10000 equally sized sets represented its empirical *P*-value that needs no further adjustment for false discovery rate (Noble 2009). Testing for different phenotypes in different pairs of populations represents independent statistical trials.

To identify adaptive loci as a response to known environmental changes, we used *Bayenv2* (Coop *et al.* 2010; Günther & Coop 2013). This method takes into account covariance of allele frequencies across tested populations, which arises as a consequence of demographic history and spatial expansion. *Bayenv2* performs well for accurate identification of loci under spatially varying selection (Villemereuil *et al.* 2014; Lotterhos & Whitlock 2014). In *Bayenv2*, a covariance matrix between all pairs of populations from putatively neutral sites is used as a null model to test for relationships of the population frequencies of a given site to an environmental variable and the SNP-specific allele distribution. Here the covariance matrix was estimated from 5000 randomly sampled SNPs that are in linkage equilibrium and used the mean from ten matrices as the null model. To test the convergence of *Bayenv2*, several independent Markov Chain Monte Carlo runs were used with a maximum chain length of 10000 iterations. Convergence was observed after about 5000 iterations (Figure 2.1).

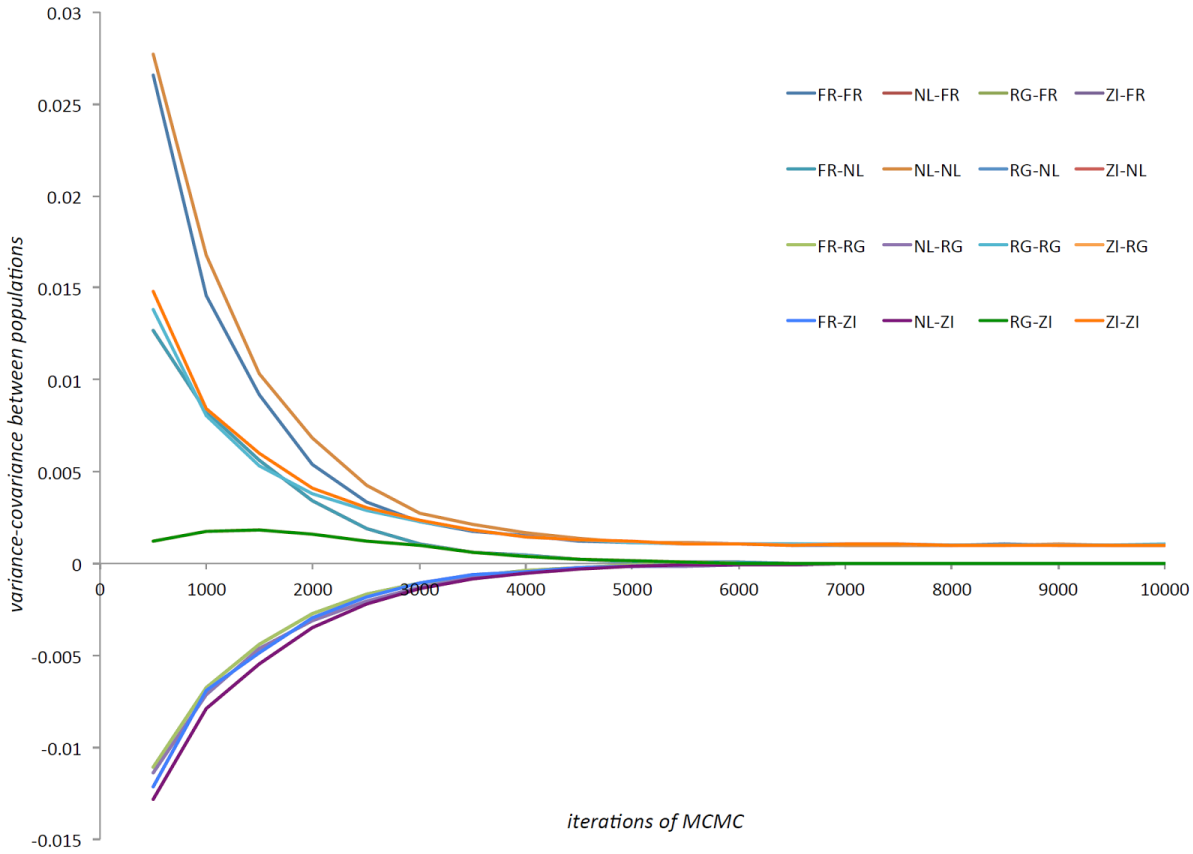



**Figure 2.1** Convergence of a *Bayenv2* covariance matrix estimation with a maximum chain length of 10000 iterations. Convergence is reached at about 5000 iterations.

However, these chains might converge to different solutions. To be most stringent, the median results from 10 independent runs were used (Blair *et al.* 2014). Subsequently, correlations between each single SNP and six environmental variables were tested for: (1) geographical latitude, (2) height above mean sea level, and (3 - 6) four temperature measures (average daily minimum of the coldest and warmest month, and average daily minimum and maximum throughout the year) (Table 2.2). The results for environmental variables are given as Bayes factors (BFs). A higher BF gives higher support to the model where the environmental variable has a significant effect on allele frequency distribution over an alternative model with no effect (Coop *et al.* 2010). Similar as above, what is finally reported is the median BF out of ten independent runs of each SNP, which has been shown to improve the proportion of false positives (Blair *et al.* 2014; Lotterhos & Whitlock 2014). With BF values, a resampling approach analogous to the one applied on $F_{ST}$ was used. Sets of SNPs of the same size were sampled randomly from the genomic background, and then the null distribution was assessed for BF of CCRT-, RSS-, and SR-associated SNPs.

**Table 2.2** The populations and environmental variables used in the analysis.

| Population | Latitude (degrees) | Altitude (m) | $T_{min}$ (°C) of the coldest month* | $T_{min}$ (°C) of the hottest month* | $T_{min}$ (°C) yearly average* | $T_{max}$ (°C) yearly average* |
|---|---|---|---|---|---|---|
| The Netherlands, Leiden (**NL**) | 52°09'29 N | -2 | 0.2 | 12.5 | 6.1 | 13.4 |
| France, Lyon (**FR**) | 45°45'00 N | 198 | 0.1 | 15.6 | 7.5 | 16.3 |
| Rwanda, Gikongoro (**RG**) | 02°27'50 S | 1796 | 13.9 | 14.9 | 14.4 | 24.8 |
| Zambia, Siavonga (**ZI**) | 16°32'17 S | 481 | 10 | 17.9 | 14.9 | 26.4 |

*Climate data were taken from World Weather Information Service – World Meteorological Organization (worldweather.wmo.int). Climatological information is based on monthly averages for the 30-year period 1961-1990.

To get an estimate of the extent to which random genetic drift has shaped the SNP allele frequencies, the parameters of a likely demographic model (Figure 2.2) were estimated by means of Approximate Bayesian Computation (Beaumont 2010; Bertorelle *et al.* 2010;

Csilléry *et al.* 2010; Laurent *et al.* 2011; Duchen *et al.* 2013). Coalescent simulations were performed with Hudson's ms (Hudson 2002). The correlation of the simulated and the observed summary statistics were high for both autosomes ($R^2$ > 0.96) and the X chromosome ($R^2$ > 0.97) (Table S1). Note that these estimates might not represent the true demographic history very well. More simulations and proper estimates would be required (Beaumont 2010). However, the combination of parameters produced by our simulations sufficiently represents the effects of genetic drift on $F_{ST}$ and *Bayenv2* values due to the demographic history.

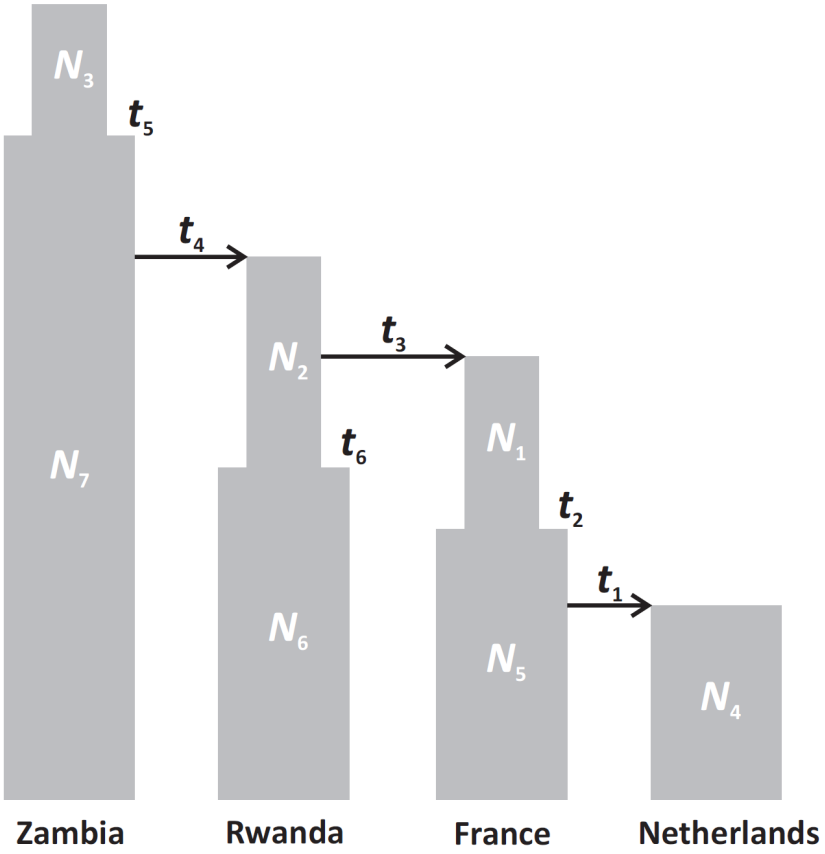



**Figure 2.2** Demographic model of the populations used in this study. The ancestral African population (Zambia) expands from its ancestral ($N_3$) to its present size ($N_7$) at time $t_5$, followed by a founder event of size $N_2$ that forms the African population (Rwanda) at time $t_4$. Rwanda undergoes a population size change at $t_6$ to $N_6$. The European populations diverge at time $t_3$ and size $N_1$, followed by a population size change at time $t_2$ to $N_5$ before Netherlands diverges from France at time $t_1$ and size $N_4$.

The empirical *P*-values generated from coalescent simulations were then assessed for all genome-wide calculated $F_{ST}$ and *Bayenv2* values by the following formula: *P*-val = (number of

simulated values > observed value) / (number of simulations). Note that these single SNP empirical *P*-values are distinct from empirical *P*-values obtained over a set of SNPs using the previously described resampling method. These *P*-values were additionally adjusted for false discovery rate according to Benjamini and Hochberg (1995).

**Table 2.3** Posterior distribution of model parameter provided in Figure 2.2 as estimated from autosomal and X chromosomal data. Times are provided in generations. Best simulations denote the parameter that produces simulations closest to the observed data. Std denotes the standard deviation as estimated from the 500 closest simulations. The correlation ($R^2$) depicts the correlation coefficient between summary statistics as derived from best simulations and observed data.

| | **Autosomes** | | **ChrX** | |
|---|---|---|---|---|
| **Parameter** | **Best simulation** | **Std** | **Best simulation** | **Std** |
| $N_0$ | 488598 | 296064 | 59493 | 969331 |
| $NC_1$ (EUR ancestr.) | 212220 | 2034838 | 517095 | 5444096 |
| $NC_2$ (RG ancestr.) | 3558864 | 1860214 | 326435 | 5528576 |
| $NC_3$ (ZI ancestr.) | 57087 | 163398 | 186785 | 1266591 |
| $N_4$ (NL) | 161734 | 6105889 | 223226 | 25147724 |
| $N_5$ (FR) | 2944039 | 7793518 | 72834 | 23919068 |
| $N_6$ (RG) | 2543659 | 6900832 | 280567 | 21739806 |
| $N_7$ (ZI) | 4815156 | 7141087 | 389559 | 22326748 |
| $T_1$ (FR->NL) | 12768 | 78218 | 215 | 33150 |
| $T_2$ (SC FR+NL) | 15652 | 245388 | 831 | 44665 |
| $T_3$ (ZI->RG) | 252899 | 351793 | 1362 | 67352 |
| $T_4$ (ANC->ZI) | 891861 | 498750 | 10557 | 114296 |
| $T_5$ (SC ANC) | 1360416 | 678186 | 13453 | 423855 |
| $T_6$ (SC RG) | 134331 | 202658 | 1352 | 47361 |
| **Correlation ($R^2$)** | 0.978 | | 0.960 | |

Using our demographic estimations, we further assessed the power in disentangling adaptive from neutrally evolving SNPs in our framework of four populations. For this analysis, we focused on the example of the CCRT trait, comprising 90 associated SNPs with information about phenotypic effects available as average difference of major and minor

alleles (Huang *et al.* 2014). Effect sizes were standardized in units of standard deviations of the CCRT distribution over male and female lines. From the set of associated SNPs, we further excluded sites with the lower effects from pairs in high LD ($R^2 > 0.8$), resulting in 74 SNPs. We used the forward equations of allele frequencies at independent sites that are given in de Vladar and Barton (2014) (Equation 6). The evolution of a number of independent polymorphic sites that contribute to a polygenic trait is defined forward in time dependent on the following parameters: (i) the selection coefficient (S = 0.1), which was chosen to be reasonably high as expected for a quantitative trait (de Vladar & Barton 2014), (ii) the population optimum ($z_0$), as expected from the mean latitude, and (iii) the initial frequency of the polymorphic sites as provided from our neutral simulations.

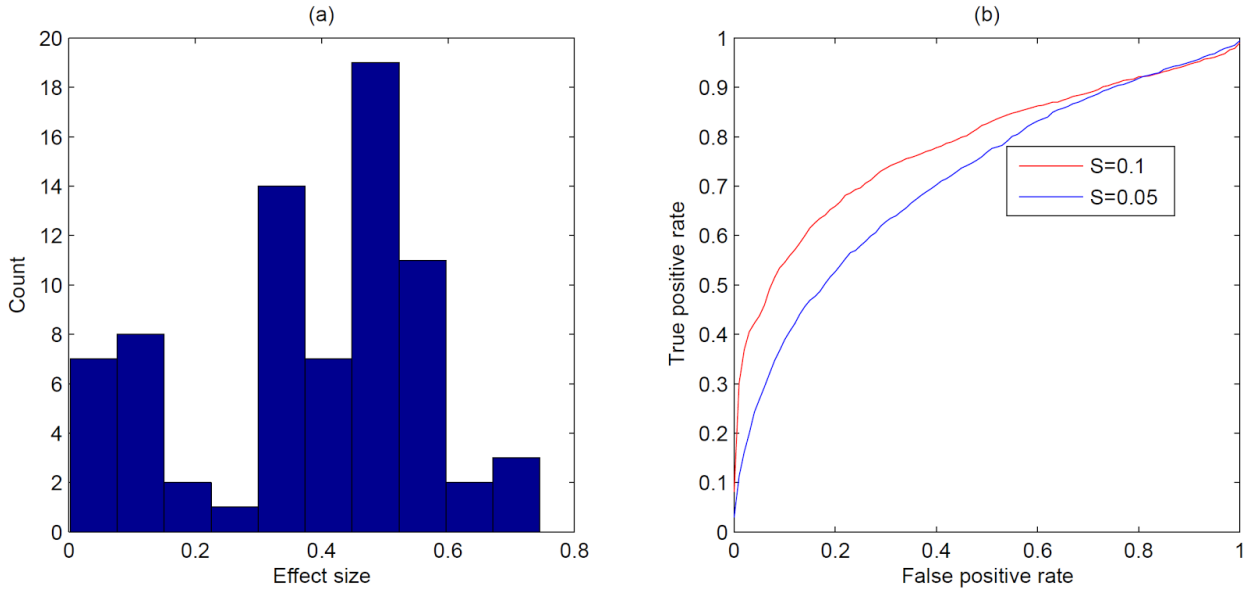



**Figure 2.3** Results from power analysis in recovering adaptive from neutral SNP with the setup of our study in the case example of CCRT. See materials and methods part for details. Panel (a) depicts the used distribution of effect sizes as quantified from CCRT phenotype. Effect sizes are provided in units of standard deviation of the respective phenotype. Panel (b) depicts the expected power in disentangling neutral from adaptive evolving SNPs associated with the CCRT phenotype for different selective pressure (S=0.1, S=0.05). Accepting 5% false positives 43% true positives would be maintained assuming a selection coefficient of S=0.1.

Additionally, we used the same mutation rate ($\mu$ = 1.13e-09) and estimated population sizes from our neutral simulations ($N_{FR}$ = 1.8263e+05, $N_{NL}$ = 9.1975e+05, $N_{RG}$ = 1.2728e+06, $N_{ZI}$ = 1.3258e+06). We then simulated the effect of genetic drift after each generation in the forward equation via multinomial resampling for all independent sites. We let each

simulation run independently for 200 generations with and without selection in all four populations. Most trajectories converged to equilibrium much earlier than 200 generations. We generated 100 replicates. From the resulting set, we applied *Bayenv2* as described above to the neutrally as well as to the adaptively evolving sites. We assessed the power by describing the proportion of correctly identified adaptive sites with a BF above the threshold obtained from the distribution of neutrally evolving sites (see Figure 2.3). From the simulations of the putative adaptive history of CCRT, we obtained a power of about 43% at a (neutral) acceptance threshold of 0.05. Thus, we believe that we are able to achieve sufficient power to detect selection on variants with intermediate to large effects that change rapidly in frequency due to environmental fluctuations (Pritchard *et al.* 2010; de Vladar & Barton 2014).

For gene and SNP annotation, we used the perl script from Ensembl's Variant Effect Predictor (VEP) tool (McLaren *et al.* 2010). For synonymous SNPs, possible codon usage bias was determined using data from the Codon Usage Database (www.kazusa.or.jp/codon) that was originally compiled from NCBI-GenBank (www.ncbi.nlm.nih.gov/genbank). Associated SNPs were checked for overlap of their respective genes with temperature tolerance genes from literature. Independently, we reported the overlap of genes and high $F_{ST}$ SNPs for CCRT, RSS, and SR.

Ontology and pathway analyses were performed using the Cytoscape plugins ClueGO (http://apps.cytoscape.org/apps/cluego) and CluePedia (Bindea *et al.* 2009, 2013) (http://apps.cytoscape.org/apps/cluepedia). We used Cytoscape version 2.1.6 (Shannon *et al.* 2003). Cohen's Kappa score (Cohen 1968) of 0.7 was used as a threshold for the proportion of genes shared between enriched ontology and pathway terms to link the terms into GO networks (Bindea *et al.* 2009) and networks of KEGG (Kanehisa & Goto 2000) and Reactome (Croft *et al.* 2011) metabolic pathways. With ClueGO and CluePedia enriched terms were integrated into networks. Enrichment and depletion of single terms were calculated using a two-sided hypergeometric test. The FDR correction was applied (Benjamini & Hochberg 1995), retaining the terms enriched with a FDR-corrected *P*-value of less than 0.01 that contained at least three candidate genes, or when the candidate genes represented at least 4% of the total number of genes related to the term.

## 2.2 Adaptation on gene network level

Secondly, we were in particular focused on polygenic adaptation signatures by means of enrichments of gene sets. We used the likelihood of selection at individual SNPs (Bayes factors), estimated using environmental correlations in Section 3.1, to assess the enrichment in signals of adaptation of thousands of gene sets representing GO terms and Reactome pathways. By treating each gene set effectively as a miniature phenotype that reflects the individual contributions of all associated SNPs, we came to radically different results compared to classical GO and pathway enrichment that relies only on outliers. Overall, our aim was to emphasize the need of future studies to utilize a range of approaches, especially when dealing with polygenic local adaptation.

### 2.2.1 Gene sets

We retrieved the connections between *Drosophila melanogaster* NCBI / Entrez gene IDs and the corresponding biochemical pathways using NCBI's Frequency weighted Links (FLink) resource (http://www.ncbi.nlm.nih.gov/Structure/flink/flink.cgi). NCBI's BioSystems database aggregates genetic pathways data from different databases (KEGG, Reactome, BioCyc, and WikiPathways) and relates them to their respective genes. We found a total of 1659 fruit fly biosystems on NCBI (http://www.ncbi.nlm.nih.gov/biosystems?term=%22Drosophila%20melanogaster%22%5BOrganism%5D). Next, we matched the NCBI gene IDs with FlyBase gene IDs using the biological Database network application (https://biodbnet-abcc.ncifcrf.gov/db/db2db.php). To map SNPs to genes, we retrieved the gene coordinates using FlyBase's Batch Download function (http://flybase.org/static_pages/downloads/ID.html). For each SNP, we noted BFs, as calculated by *Bayenv2*, and built the distributions for each gene set (biosystem or GO category) of their respective BFs.

### 2.2.2 Bayes factor set enrichment approach

To find out if a given gene set had a higher likelihood of selection, as given by the cumulative contribution of all of its SNPs' BFs, we used the following resampling approach. For each target gene set, we randomly sampled from the genomic background SNPs that map to stretches of DNA of the lengths that corresponded to the lengths of each gene in

the target gene set. We sampled 10 000 such sets of SNPs for each target gene set (100 000 for latitude because of the extremely small *P*-values of some gene sets), and then counted how many of them had larger or smaller BF medians compared to the target set, resulting in an empirical *P*-value. As discussed previously, such a *P*-value accounts for multiple testing error. Additionally, sampling of gene lengths for each gene set accounts for potential gene length bias, which might spuriously assign greater likelihood of selection to gene sets containing higher numbers of SNPs. We repeated this analysis for each of the six environmental variables (latitude, altitude, and four measures of temperature). For further analyses, focus was mainly on the first environmental variable (latitude), hence a particular gene set was considered "enriched" if its empirical *P*-value for latitude was less than 0.05.

### 2.2.3 Relating networks of enriched gene sets

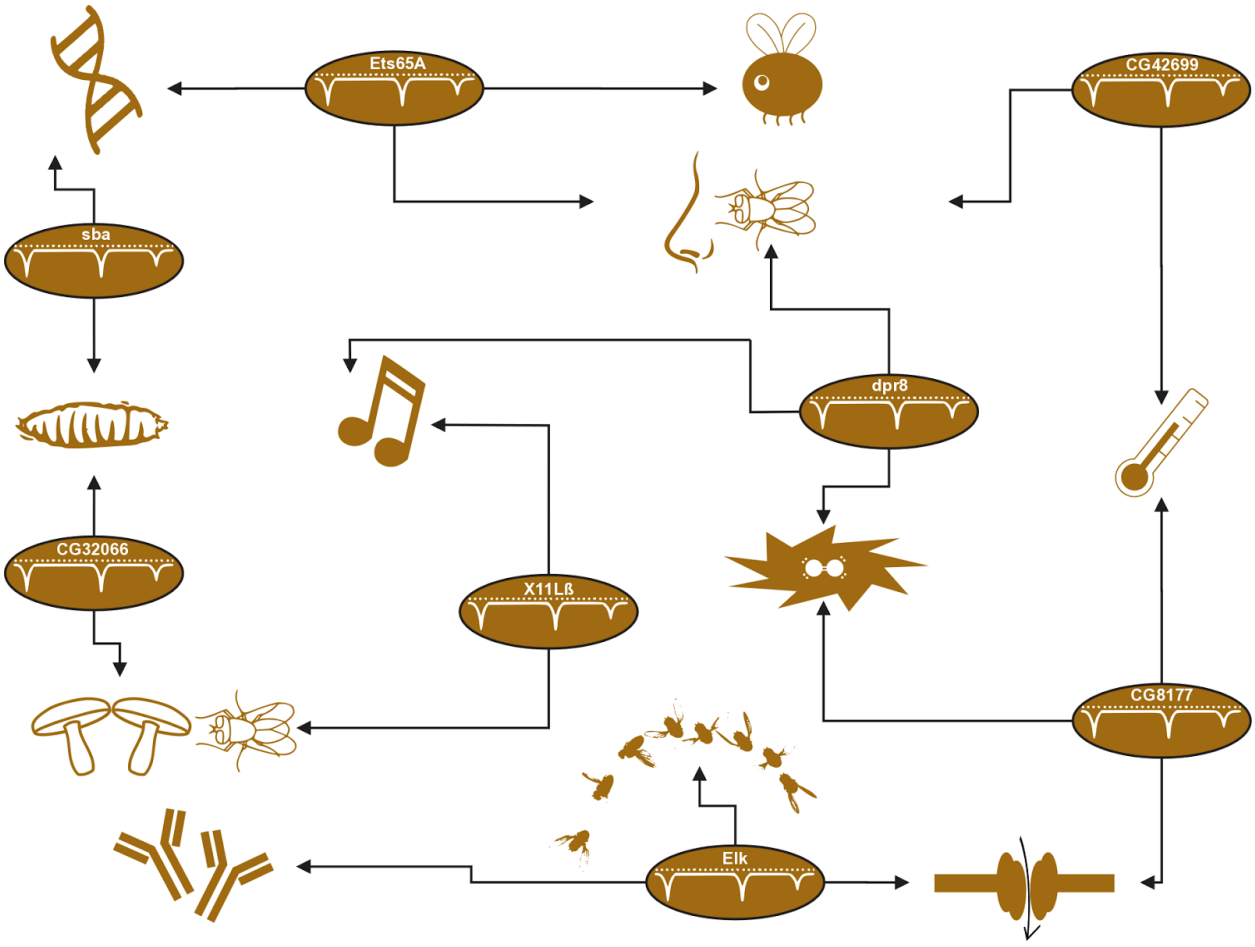
To understand how gene sets enriched for signals of adaptation relate to each other, we retrieved a list of genes associated with each gene set enriched at $P < 0.05$ for latitude. For each gene, we checked the relevant literature references on flybase.org that relate the gene to the particular gene set. Our first aim was simply to relate enriched gene sets to one another by noting the genes that they share in common, thus creating a large adaptive network. In particular, we wanted to know which gene sets may have multiple overlapping genes, because that might give us an insight into what the broader "phenotype" is that natural selection might be operating on.

Our second aim was to note any connection in the relevant reference that the gene might have with phenotypes other than the one broadly described by the gene ontology or pathway definition, but that still might be relevant for local adaptation to European environments. This is important for elucidating how local selection operates, because it allows us to gain an understanding how seemingly disparate biological processes might be related in a network of possible adaptive significance.

To identify the most important Gene Ontology sets, we clustered the significantly enriched ($P < 0.01$) sets using the SimRel algorithm (Schlicker *et al.* 2006) implemented in the software ReviGO (http://revigo.irb.hr/) (Supek *et al.* 2011). ReviGO is a tool that takes as input a list of GO terms ascertained by the user, and attempts to cluster those GO terms based on their functional (semantic) similarity. The SimRel algorithm calculates the functional similarity between gene ontology terms by assessing the number of ancestral

terms they share in the ontology. It creates a matrix of the GO terms' pairwise semantic similarities and then clusters them in a two-dimensional plot (see Figures 3.1 and 3.3) by reducing the dimensionality of this matrix by means of multidimensional scaling (Supek *et al.* 2011). We used the default cut-off of 0.7 for the SimRel similarity score, which means that all GO terms with a score higher than 0.7 were assigned to a cluster. Each cluster was named after the GO term in the cluster with the highest uniqueness. Uniqueness is calculated as (1 − (average semantic similarity of a term to all other terms)) (Supek *et al.* 2011). Finally, we examined the resulting clusters using literature in the context of possible polygenic adaptation to European environments.

# CHAPTER 3: RESULTS




Adaptive gene network - an illustration adapted from Fig. 3.2

## 3.1 Adaptation on gene level

### 3.1.1 Genetic differentiation of trait-associated SNPs

We first quantified the amount of differentiation for SNPs associated with phenotypic traits that are known to differ between temperate and tropical regions (Da Lage *et al.* 1990; Gibert *et al.* 2001a; b; Hoffmann *et al.* 2001; Kennington *et al.* 2001; De *et al.* 2013). As a

measure of differentiation, we estimated pairwise $F_{ST}$ of SNP sets that were previously associated with CCRT, RSS, and SR (Huang *et al.* 2012b, 2014) in four pairs of populations (see Table 2.1 for SNP numbers, Table 3.1 for results).

**Table 3.1** Mean population differentiation ($F_{ST}$) over associated SNP sets (Huang *et al.* 2014). Empirical *P*-value ($P_{empirical}$) was estimated by random resampling approach (see methods).

| Trait | Number of associated SNPs | Population comparison | Target mean $F_{ST}$ | Genome mean $F_{ST}$ | $P_{empirical}$ | significance |
|---|---|---|---|---|---|---|
| **CCRT** | 59 | RG-NL | 0.133 | 0.081 | 0.031 | * |
| | | RG-FR | 0.147 | 0.099 | 0.0654 | ∘ |
| | | ZI-NL | 0.125 | 0.085 | 0.0708 | ∘ |
| | | ZI-FR | 0.129 | 0.087 | 0.0765 | ∘ |
| **RSS** | 64 | RG-NL | 0.134 | 0.081 | 0.0259 | * |
| | | RG-FR | 0.122 | 0.099 | 0.1963 | ns |
| | | ZI-NL | 0.11 | 0.085 | 0.1613 | ns |
| | | ZI-FR | 0.091 | 0.087 | 0.4163 | ns |
| **SR** | 51 | RG-NL | 0.083 | 0.081 | 0.4278 | ns |
| | | RG-FR | 0.081 | 0.099 | 0.6915 | ns |
| | | ZI-NL | 0.096 | 0.085 | 0.314 | ns |
| | | ZI-FR | 0.089 | 0.087 | 0.4336 | ns |

A *P*-value has been obtained from a resampling approach as described in Materials and Methods. We find a significant enrichment of elevated $F_{ST}$ values (see Table 3.1) comparing the Netherlands (NL) and Rwanda (RG) samples (mean target $F_{ST}$ = 0.133; $P_{empirical}$ = 0.031) in the CCRT dataset (Huang *et al.* 2014), and NL and RG (mean target $F_{ST}$ = 0.134; $P_{empirical}$ = 0.0259) in the RSS dataset (Huang *et al.* 2014). The $F_{ST}$ differences between NL and RG were not mirrored in the FR vs. ZI for RSS. However, for CCRT nearly all pairs of populations show a significant ($P_{empirical}$ < 0.05), or nearly significant ($P_{empirical}$ < 0.08) enrichment of $F_{ST}$ values. This may suggest that the differentiation between Africa and Europe at trait-associated SNPs cannot simply be explained by their demographic history and that adaptive forces need to be invoked. At the same time, the marginal significance displayed in some pairs of

populations for CCRT, and no significant enrichment of $F_{ST}$ for SR (Table 3.1) requires more evidence than solely $F_{ST}$ before considering the possibility of selection. This evidence is presented in the following sections.

### 3.1.2 Environmental correlation with trait-associated SNPs

Next we used BFs instead of $F_{ST}$ and repeated the resampling approach, to test for an enrichment of high BFs in the associated SNPs. We tested fewer SNPs using *Bayenv2* compared to $F_{ST}$, because fewer SNPs were polymorphic across all the populations tested. Nevertheless, the *Bayenv2* results mostly follow qualitatively the results of the empirical $F_{ST}$ observations in that CCRT-associated SNPs were the most likely among the three traits to have BFs higher than the genomic background (Table 3.2). In the case of CCRT, we detected enrichments for altitude ($P_{empirical}$ = 0.015) and for coldest month minimum ($P_{empirical}$ = 0.001). For correlations with yearly minimum temperature, the empirical *P*-value of BFs associated with CCRT was still marginally significant ($P_{empirical}$ = 0.066), while latitude and both yearly maximum temperature and the hottest month minimum showed no significant correlation. Additionally, the SNPs associated with RSS and SR generally do not show a significantly higher correlation with environmental variables (higher BFs) than the genomic background (Table 3.2).

The magnitude of BFs conveys information on the likelihood of a site being under selection (e.g. a BF > 1 means that selection is more likely than neutrality). BFs are a much more stringent measure of the likelihood of selection than pairwise $F_{ST}$. For example, when we imposed a cutoff of *ln*(BF) > 1 (positive evidence as suggested by Kass and Raftery (1995) or $P < 0.0063$ from our simulations), only one CCRT-associated SNP (chr2R_18586714) was still significantly correlated with environmental variables ($P < 0.0039$). For instance, chr3L_6723212 was not significant with this cutoff (BF between 0.46 for env2 (altitude) and 2.46 for env5 ($T_{min}$, yearly average)), even though its $F_{ST}$ ranges from 0.44 (FR-RG) to 0.73 (NL-ZI). Note that a cutoff of *ln*(BF) > 1 corresponds to $P < 0.0063$ according to our neutral simulations.

**Table 3.2** Mean Bayes factors over associated SNPs sets calculated with *Bayenv2*. Empirical *P*-values were obtained by random resampling from genomic background (see methods). Asterisks indicate significance threshold ($P < 0.05$ (*); $P < 0.01$ (**); $P < 0.001$ (***)).

| Quantitative trait / association study | Environmental variable | Mean Bayes factor | $P_{empirical}$ | significance |
|---|---|---|---|---|
| **CCRT** ( $N = 14$ ) | Latitude | 16.8739 | 0.129 | ns |
| | Altitude | 1532.67 | 0.015 | * |
| | $T_{min}$ of the coldest month | 8601.49 | 0.001 | *** |
| | $T_{min}$ of the hottest month | 3373.85 | 0.755 | ns |
| | $T_{min}$ yearly average | 17.1501 | 0.066 | ° |
| | $T_{max}$ yearly average | 15.3386 | 0.155 | ns |
| **RSS** ( $N = 14$ ) | Latitude | 0.24811 | 0.300 | ns |
| | Altitude | 0.22003 | 0.168 | ns |
| | $T_{min}$ of the coldest month | 0.22871 | 0.430 | ns |
| | $T_{min}$ of the hottest month | 0.21495 | 0.763 | ns |
| | $T_{min}$ yearly average | 0.23986 | 0.475 | ns |
| | $T_{max}$ yearly average | 0.24788 | 0.262 | ns |
| **SR** ( $N = 12$ ) | Latitude | 0.21202 | 0.378 | ns |
| | Altitude | 0.21746 | 0.170 | ns |
| | $T_{min}$ of the coldest month | 0.20214 | 0.830 | ns |
| | $T_{min}$ of the hottest month | 0.19755 | 0.515 | ns |
| | $T_{min}$ yearly average | 0.21252 | 0.494 | ns |
| | $T_{max}$ yearly average | 0.20995 | 0.379 | ns |

### 3.1.3 Many genes related to cold tolerance are enriched for SNPs with high BF and $F_{ST}$ values

We retrieved a list of genes that are known to be related to cold or heat tolerance from the literature (see Table S1, Table S3 for description). To mitigate possible false positives, we aimed to include in this list candidates from a range of studies that have employed a variety of different techniques, including QTL mapping, physiology, gene knockdowns, *P*-element insertions, RNA interference, mutant complementation tests, and various gene

expression approaches (see Table S1 for references for each gene and the techniques used in each study). We quantified the number of SNPs with significant $F_{ST}$ outlier (Table S3) and BF outlier (Table S3 and S4) values within these genes. We found that 17 genes (*DnaJ-1*, *AnxB9*, *Lsp1beta*, *CG16700*, *psq*, *stan*, *lola*, *Oatp30B*, *E(spl)m7-HLH*, *cpo*, *whd*, *CG12054*, *Dyrk2*, *shep*, *chas*, *Ire1*, and *Octbeta3R*) contained SNPs with evidence in favor of selection (*ln*(BF)>1, (Kass & Raftery 1995) or $P < 0.0063$) for at least one environmental variable. Even more strikingly, 13 of these genes contained SNPs with strong evidence (*ln*(BF)>3, (Kass & Raftery 1995) or $P < 0.0043$), which means that a model including selection is about 20 times more likely than neutrality at multiple loci within these genes.

We used the VEP tool to retrieve functional annotations of the top 1% SNPs from the genome and for each gene we reported numbers of intron variants, 3' and 5'UTR variants, and synonymous and nonsynonymous coding variants. For nonsynonymous SNPs we also reported whether the alternate amino acid had a differently charged side chain, as this may lead to differences in the folding of the final protein product. We observed a general trend such that most SNPs with elevated $F_{ST}$ values are located in introns, followed by SNPs in untranslated (3' and 5') regions. This is consistent with previously observed patterns of selective constraints in noncoding DNA (Halligan *et al.* 2004; Halligan & Keightley 2006). Many genes show frequency changes suggestive of selection in multiple classes. We found 20 genes showing synonymous changes that might impact gene expression through codon usage bias. Overall, we found that from the 46 genes retrieved from literature (Table S1), 33 contained SNPs from the top 1% of the $F_{ST}$ distribution, for at least one pairwise population comparison (Table S3). We observed that in the heat-shock gene *Hsp26* there were two variants (chr3L_5743995 and chr3L_5743998) that code for codons with negative codon usage bias, which were both more common in Europe (with frequency 90.9% or more) than in Africa (frequency 52.4% or less). This suggests that they might be under less selective constraint, or under positive selection in African populations (Table S3). Notably, 6 genes (*CG31738*, *CG12943*, *CG30379*, *lola*, *nclb*, and *chas*) contained nonsynonymous variants with very high $F_{ST}$, indicating selection for a different amino acid. This possibility is most apparent in *CG30379* and *lola* because the associated amino acid changes also change the charge of their respective side chains, which could have an even greater influence on the folding of the protein. The gene *lola* is particularly interesting, because it contains high $F_{ST}$

variants along its entire length, including coding and noncoding, as well as regulatory regions.

### 3.1.4 Inversion analysis

Major cosmopolitan inversions are known to have an effect on clinal variation (Hoffmann & Weeks 2007; Hoffmann & Rieseberg 2008). In order to exclude the possibility that the differentiation at our candidate genes was due to the effect of inversions, we compared the genomic coordinates of the breakpoints of the four major cosmopolitan inversions, *In(2L)t*, *In(2R)NS*, *In(3L)P*, and *In(3R)P (Ashburner & Lemeunier 1976)*, with the coordinates of our candidate genes, i.e. genes with *ln*(BF)>5, as reported by *Bayenv2* ($P < 0.0034$) for correlation with latitude and altitude. We obtained the cytogenetic-absolute coordinate mapping from http://flybase.org/static_pages/downloads/FB2014_06/map_conversion/genome-cyto-seq.txt.gz. We also checked for overlap with the candidate genes retrieved from literature. We found no genes in close proximity to any inversion breakpoint; for example, the candidate gene closest to any inversion was *sty*, a latitudinal candidate gene located approximately 115 kb upstream of a *In(2L)P* breakpoint (see inversion_analysis.xlsx on Dryad). Of the genes from the literature, *Hsp83* comes closest (about 32 kb upstream of *In(2L)P*). It is therefore unlikely that these inversions would have a noticeable effect on our inferences.

### 3.1.5 Clinal genes in Europe overlap with clinal genes in North America

We mapped the SNPs with extremely high BFs (*ln*(BF)>5, as reported by *Bayenv2*; $P < 0.0448$ from neutral simulations) for correlation with latitude and altitude to genes, and then reported the overlaps with genes that were detected in a North American cline (Fabian *et al.* 2012) (see Figure 3.1, Figures S1-S3, and Table 3.3).

**Table 3.3** Number of genes with *ln*(BF)>5 ($P < 0.0029$) for correlation with latitude that overlap with candidate genes of Fabian et al. (2012). Final column gives the significance of the overlap from a hypergeometric test.

| **Latitudinal candidate genes (Fabian *et al.* 2012)** | ***N* genes (candidate, Fabian *et al*. 2012)** | ***N* genes (total, Fabian *et al.* 2012)** | **Latitude *ln*(BF)>5 overlap (of $N_{total}$=378)** | $P_{hyper}$ |
|---|---|---|---|---|
| Florida-Maine | 2010 | 11314 | 78 | 0.1018 |
| Florida-Pennsylvania | 2051 | 11314 | 82 | 0.0541 |
| Pennsylvania-Maine | 720 | 11314 | 51 | $2.36 \cdot 10^{-7}$ |

| | | | | |
|---|---|---|---|---|
| Clinal genes (all) | 1973 | 11314 | 72 | 0.2191 |
| Clinal genes (significant) | 140 | 11314 | 11 | 0.0072 |
| Across all populations | 3169 | 11314 | 131 | 0.0024 |

We found a total of 8 candidate genes across both latitude and altitude overlapping with the latitudinal selection candidates of the North American cline (Fabian *et al.* 2012): *Ets65A*, *Elk*, *sba*, *CG32066*, *dpr8*, *CG8177*, *X11Lbeta*, and *CG42699* (Figures 3.1 and 3.2). Four of these - *Ets65A*, *Elk*, *sba*, *CG32066* - are also clinal genes in North America (Fabian *et al.* 2012), so we were able to compare the direction of change in their candidate SNPs' allele frequencies and our own. In all four cases, the estimated frequency of the major allele in all candidate SNPs increases consistently from RG to FR to NL, while in ZI it is about the same as in RG.

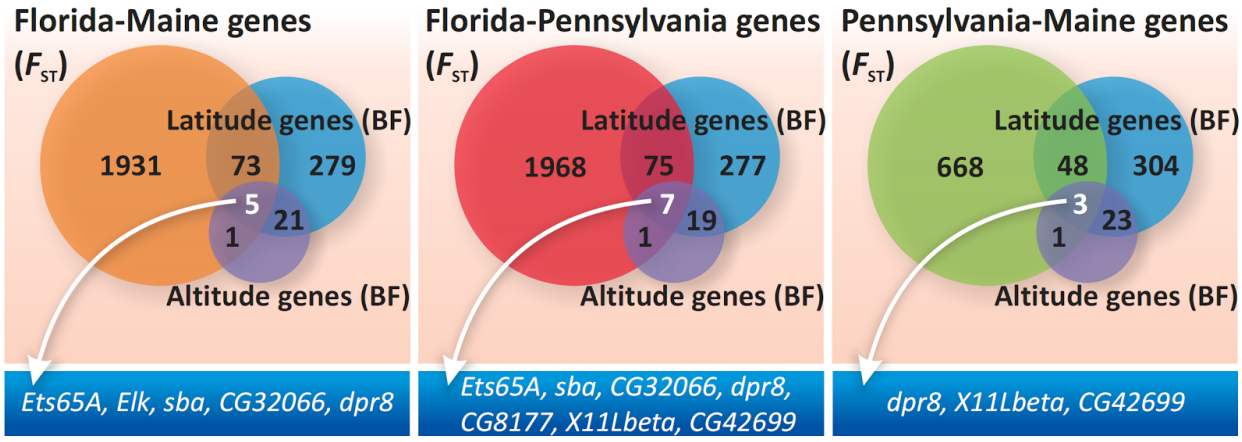



**Figure 3.1** Proportions of genes supported by SNPs with strong evidence (*ln*(BF)>5 or $P < 0.0029$) for correlation with latitude and altitude (*Bayenv2*) that overlap with candidate genes from North America(Fabian *et al.* 2012). The most interesting genes that overlap among all three sets are shown in the bottom panels. For overlaps between North America and other environmental variables, see also Figures S5 through S7.

A literature research on the function of these genes showed a loose network of similar overlapping functions with common biological themes (see Figure 3.2). Most notably, six of these genes (*Ets65A*, *Elk*, *CG32066*, *dpr8*, *X11Lbeta*, and *CG42699*) are related to various kinds of behavioral responses. Additionally, they have been found to influence temperature sensitivity (*CG8177* and *CG42699*), oxidative stress (*CG8177* and *dpr8*), courtship song (*dpr8* and *X11Lbeta*), and mushroom bodies (*X11Lbeta* and *CG32066*), structures in the fly brain essential for learning and memory. *Ets65A* is a transcription factor expressed in a nutrition-dependent manner in the adipose tissue (Baltzer *et al.* 2009), which may affect

lifespan (Ayroles *et al.* 2009; Durham *et al.* 2014), a life-history trait related to starvation resistance and cold resistance (Hoffmann *et al.* 2005; Ayroles *et al.* 2009). Intriguingly, 11 SNPs of *Ets65A* had been found to vary along a cline in North America, and those with the highest BFs, *chr3L_6112566* and *chr3L_6106869*, as well as 22 of the highest $F_{ST}$ variants, are intronic, suggesting an important adaptive role in gene regulation.

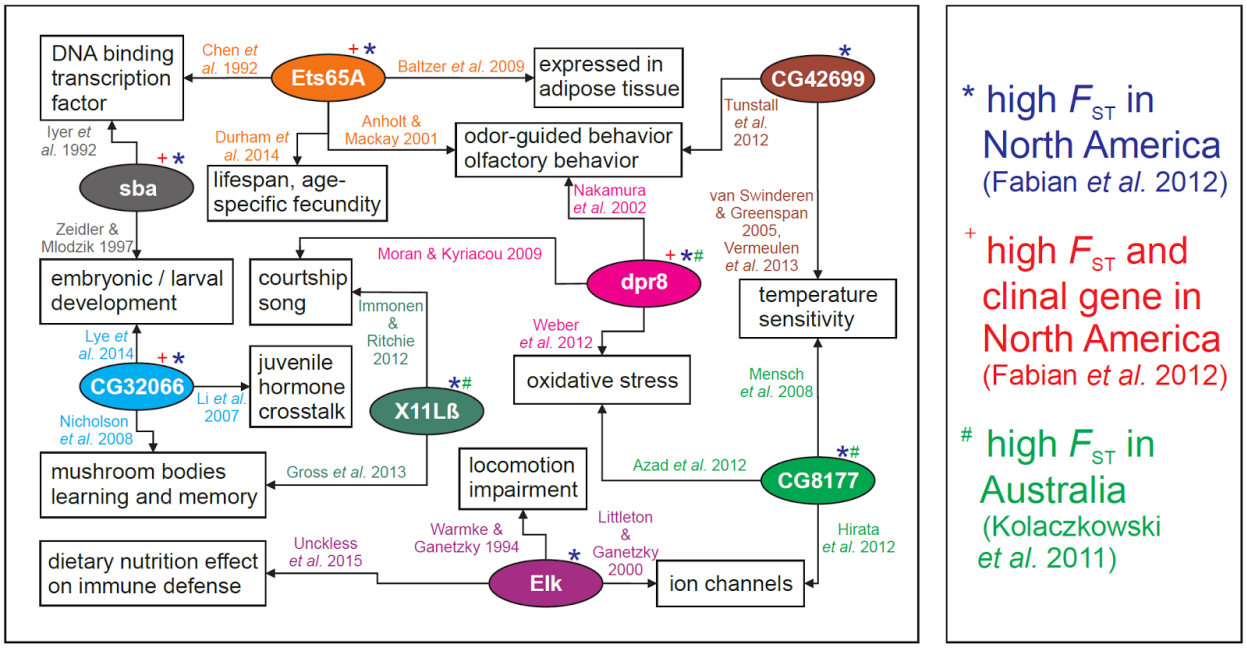



**Figure 3.2** Manually drawn network of candidate genes that overlap with previous studies (Australia (Kolaczkowski *et al.* 2011); North America (Fabian *et al.* 2012); see also Figures 3.1, and S1 to S3. Only the genes (coloured ellipses) of SNPs related to cold tolerance (*Bayenv2*, *ln*(BF)>5 or $P < 0.0029$) were considered. Open rectangles denote functional relevance from literature with lines exemplifying relationships between genes and functions, with the relevant references.

Even more interesting in this regard is the synonymous SNP *chr3L_6093812*, whose predominantly European allele, G ($f(G)_{FR}$=0.86, $f(G)_{NL}$=1, $f(G)_{RG}$=0.18, $f(G)_{ZI}$=0.33), generates a more frequently used codon (AAG, used 70% of the time). Another transcription factor that showed up in our results was *sba* (*six-banded*), an important developmental gene (Zeidler & Mlodzik 1997; Blanco *et al.* 2010; Iyer *et al.* 2013) that might be involved in changing the gut morphology in response to gut microbes (Sharon *et al.* 2010; Ridley *et al.* 2012; Newell & Douglas 2014; Broderick *et al.* 2014). Two additional genes from this analysis are also implicated in mating behavior: *dpr8* is involved in the production of male courtship song (Moran & Kyriacou 2009), while *X11Lbeta* is upregulated in females responding to male courtship song (Immonen & Ritchie 2012). Interestingly, the two genes have been found to

vary in both a North American (Fabian *et al.* 2012) and an Australian cline (Kolaczkowski *et al.* 2011). The third American-Australian cline candidate that also showed up in our analysis was *CG8177*. This is interesting in the context of its possible role in hypoxia tolerance (Azad *et al.* 2012), which it shares with *dpr8* (Weber *et al.* 2012), and its temperature-sensitive effects on developmental time (Mensch *et al.* 2008). Sensitivity to temperature seems to impact the function of another candidate gene, *CG42699*, both through its interaction with *Syx1A* (van Swinderen & Greenspan 2005), and its expression after exposure to cold shock (Vermeulen & Bijlsma 2004; Vermeulen *et al.* 2013). Likewise, *CG32066* might play a role in the crosstalk of the juvenile hormone and ecdysteroids (Li *et al.* 2007), which are known to impact reproductive diapause, an important overwintering mechanism (Mitrovski & Hoffmann 2001; Boulétreau-Merle & Fouillet 2002) that varies clinally (Schmidt *et al.* 2005). It is therefore less surprising that *CG32066* harbors clinal SNPs in North America (Fabian *et al.* 2012), and that it responds to non-optimal rearing temperatures (Chen *et al.* 2015). Our final candidate gene was the $Ca^{2+}$-gated $K^{+}$ channel *Elk*, whose mutations impair locomotion (Warmke & Ganetzky 1994; Littleton & Ganetzky 2000). Interestingly, it has been shown that locomotion may be correlated with temperature conditions (Crill *et al.* 1996; Gibert *et al.* 2001a), and that neurotransmitter release triggered by voltage-gated channels can depend on temperature (Chuang *et al.* 2004; Wu *et al.* 2005).

### 3.1.6 Overlap of enriched Gene Ontology terms with other studies

To gain a clearer understanding of the biological functions of our most significant clinal genes, we tested for significant enrichment of GO and KEGG/Reactome terms. Furthermore, we performed an equivalent enrichment analysis of the North American candidate genes (Fabian *et al.* 2012), and then assessed the overlap with our candidates. Using Cytoscape's ClueGO and CluePedia plugins, we integrated our candidate genes into GO networks and networks with KEGG/Reactome metabolic pathways. Table 3.4 shows the results of the ClueGO analysis for our latitudinal selection candidate genes (N = 378). ClueGO resulted in 111 significantly enriched GO terms (for the 20 most highly enriched categories, see Table S5). These categories are grouped into clusters using Cohen's Kappa statistics, based on their shared genes (Figure 3.3).

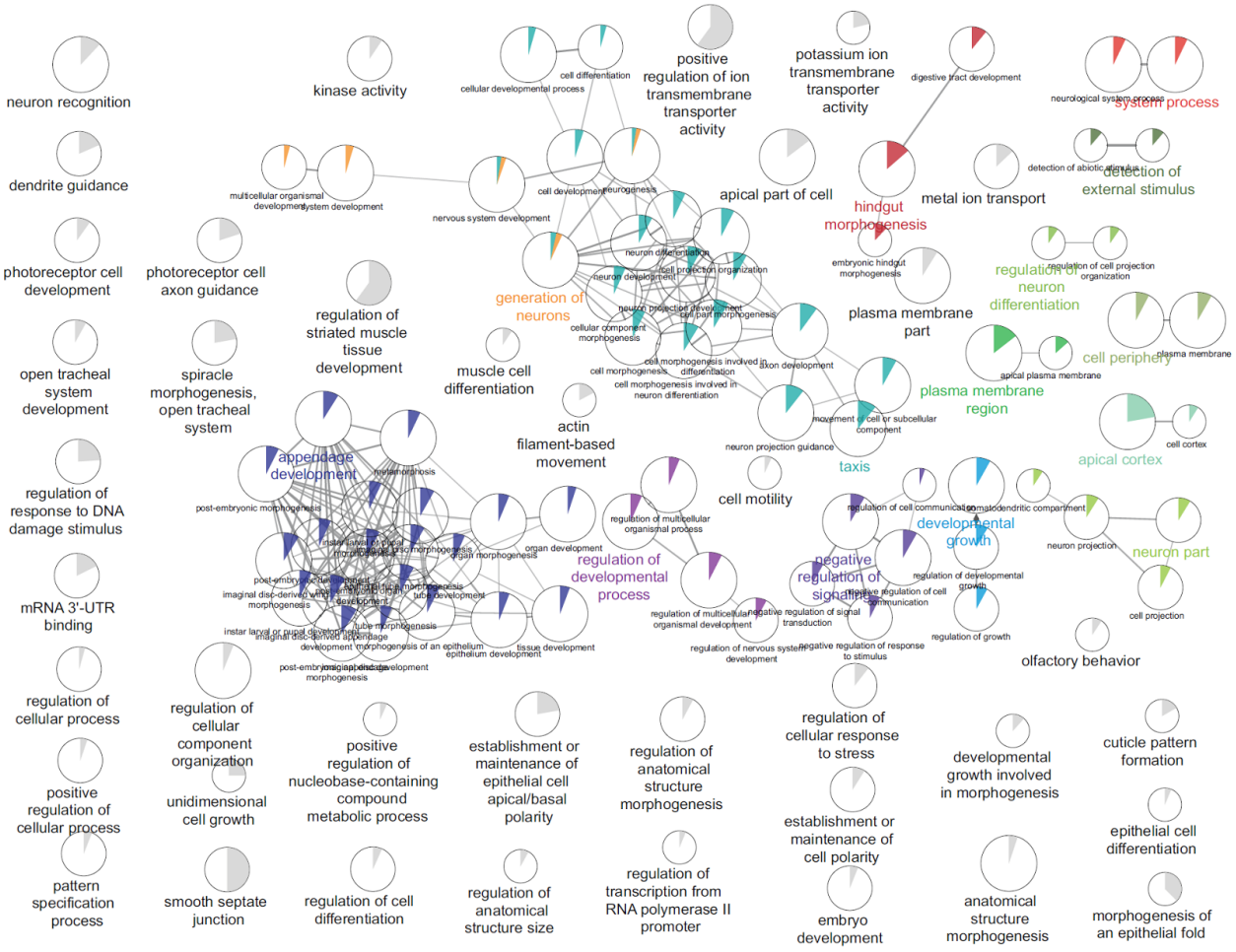



**Figure 3.3** Clusters of GO categories enriched with genes most correlated with latitude (*ln*(BF)>5 or $P < 0.0023$). A total of 111 GO categories were enriched, and 72 of them (designated by coloured pie charts) were grouped into a total of 14 clusters based on shared genes. Clusters were defined by Cohen's *kappa* > 0.7. Pie charts represent the percentage of clinal genes belonging to each enriched category. Coloured categories are those with the most significant *P*-value among the categories of each cluster.

Many terms that were grouped in these clusters could be related to various aspects of the nervous system, epithelium, wing, and tube development, and cover multiple developmental stages (Figure 3.3). Interestingly, the majority of these terms (12 out of 20 for appendage development, 10 out of 16 for taxis and 3 out of 5 for generation of neurons) were also significantly enriched in our ClueGO analysis of all three population pairs of the North American cline (Fabian *et al.* 2012). All the remaining terms from the

three clusters were enriched in at least one population pair of the North American cline (Table 3.4).

**Table 3.4** Number of GO terms (upper table) and KEGG and Reactome pathways (lower table) that overlap with (Fabian *et al.* 2012) in latitudinal differentiation. With altitudinal differentiation we could not find any significant overlap (data not shown).

| **Latitudinal candidate genes (Fabian *et al*. 2012)** | ***N* GO terms *P*<0.01 (Fabian *et al*. 2012)** | ***N* GO terms (total)** | **Latitude GO terms overlap (of $N_{total}$=111)** | $P_{hyper}$ |
|---|---|---|---|---|
| Florida-Maine | 101 | 4844 | 57 | $4.99 \cdot 10^{-73}$ |
| Florida-Pennsylvania | 131 | 4844 | 48 | $8.47 \cdot 10^{-49}$ |
| Pennsylvania-Maine | 79 | 4844 | 52 | $2.87 \cdot 10^{-71}$ |
| Clinal genes (all) | 152 | 4844 | 61 | $9.98 \cdot 10^{-67}$ |
| Clinal genes (significant) | 36 | 4844 | 6 | $1.41 \cdot 10^{-4}$ |

| **Latitudinal candidate genes (Fabian *et al*. 2012)** | ***N* pathway terms *P*<0.01 (Fabian *et al*. 2012)** | ***N* pathway terms (total)** | **Latitude pathway terms overlap (of $N_{total}$=75)** | $P_{hyper}$ |
|---|---|---|---|---|
| Florida-Maine | 117 | 3251 | 14 | $2.47 \cdot 10^{-7}$ |
| Florida-Pennsylvania | 54 | 3251 | 17 | $5.43 \cdot 10^{-16}$ |
| Pennsylvania-Maine | 45 | 3251 | 3 | 0.0840 |
| Clinal genes (all) | 115 | 3251 | 13 | $1.37 \cdot 10^{-6}$ |
| Clinal genes (significant) | 6 | 3251 | 0 | - |

The majority of the enriched KEGG and Reactome pathways were involved in signaling and various aspects of the nervous system (Figure 3.4). Similar to the GO terms, many of the enriched pathways were also enriched in our ClueGO analysis using candidate genes of Fabian et al. (2012). We therefore wanted to know if these overlaps could be generalized to the total set of enriched GO terms, as well as enriched KEGG and Reactome pathways for all the population pairs from North America.

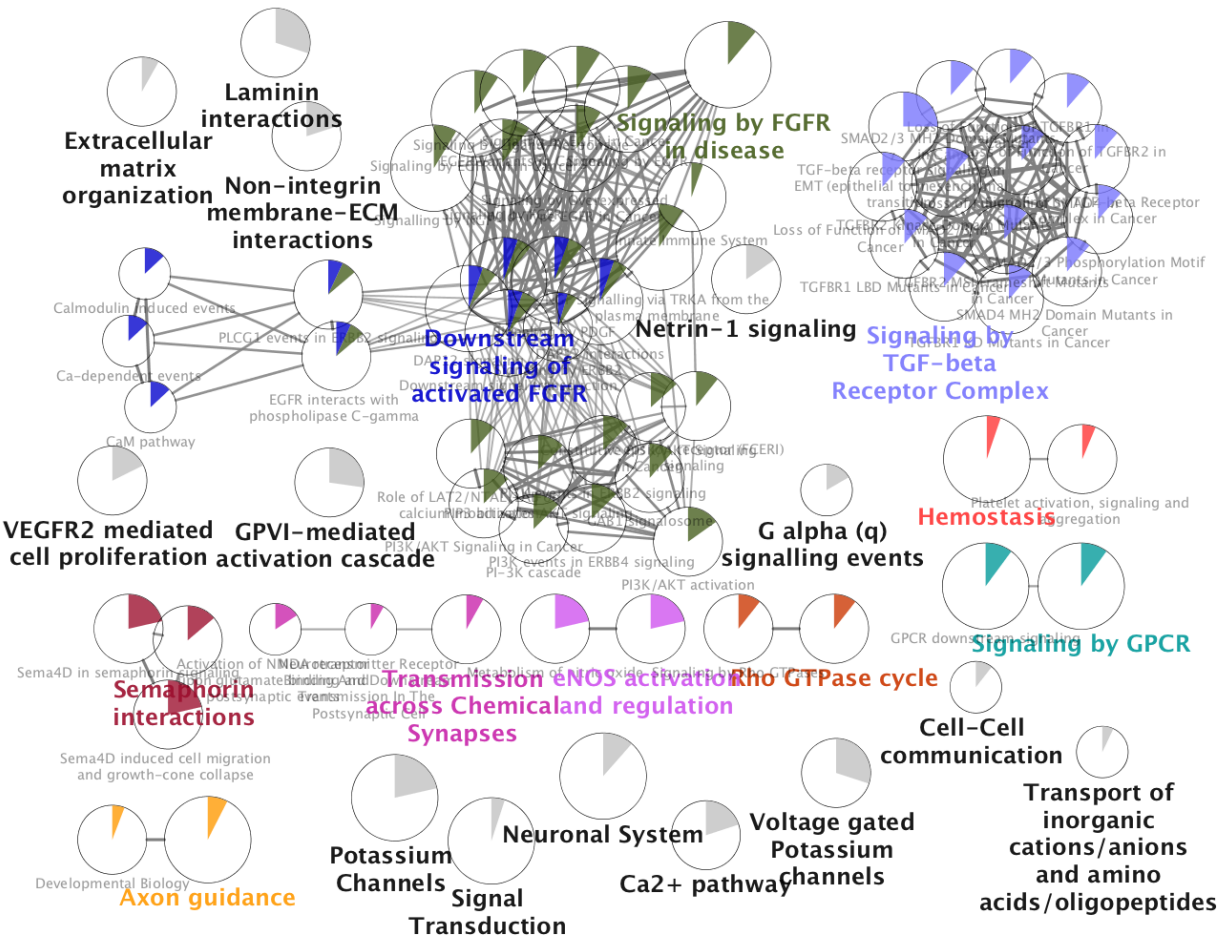



**Figure 3.4** Clusters of KEGG/Reactome pathways enriched with genes most correlated with latitude (*ln*(BF)>5 or $P < 0.0023$). A total of 75 KEGG and Reactome pathways were enriched, and 61 of them (designated by coloured pie charts) were grouped into a total of 10 clusters based on shared genes. Clusters were defined by Cohen's *kappa* > 0.7. Pie charts represent the percentage of clinal genes belonging to each enriched pathway. Coloured pathways are those with the most significant *P*-value among the categories of each cluster.

In the case of GO terms, Table 3.4 shows that the overlap in all three cases was not only substantial (between 43 and 51%), but also statistically significant (*P*-values between $10^{-16}$ and $10^{-5}$, hypergeometric test). Total clinal genes and the subset of significant clinal genes also resulted in significant overlaps between enriched GO terms (*P*-values of $9.98 \cdot 10^{-67}$ and $1.41 \cdot 10^{-4}$, hypergeometric test).

In the case of KEGG and Reactome pathways (Table 3.4, lower panel), the overlap was smaller percentage-wise, but still statistically significant in the case of overlaps with Florida-Maine and Florida-Pennsylvania (*P*-values of $2.47 \cdot 10^{-7}$ and $5.43 \cdot 10^{-16}$, hypergeometric test). While significant clinal genes showed no overlap with our data, likely due to the small overall number of enriched terms (only six), the overlap with total clinal genes was found to be statistically significant (at *P*-value of $1.37 \cdot 10^{-6}$, hypergeometric test).

Finally, we even found some overlap between the Kappa-defined clusters, which were defined according to the most significant GO term in the cluster. Table 3.4 shows the numbers of overlapping GO terms and pathways, as well as Kappa-defined clusters (all at *P*-value FDR cutoff of 0.01).

## 3.2 Adaptation on gene network level

In the second part of the results, we are focusing on the characterization of gene ontology terms and Reactome pathways that show evidence of local adaptation to environmental variables via small shifts in allele frequencies (and hence, Bayes factors obtained by means of *Bayenv2*) across all associated SNPs (i.e. polygenic adaptation). The distribution of BFs of these SNPs is generally expected to be higher (i.e. have a higher median BF) compared to sets of SNPs of equal size sampled from the genomic background.

Out of a total of 6497 GO terms, 340 were significant at latitude $P_{\text{empirical}} < 0.01$. Table 3.5 shows the results of clustering of GO terms with latitude $P_{\text{empirical}} < 0.01$ that contain at least 5 annotated genes by means of semantic similarity implemented in the software *ReviGO*. We found a total of five clusters whose head terms were significant at latitude $P_{\text{empirical}} < 0.0001$: (1) regulation of circadian sleep/wake cycle, sleep, (2) nephrocyte filtration, (3) mitochondrion organization, (4) tissue development, and (5) regulation of myoblast fusion. The remaining 6 GO terms in Table 3.5 all clustered under the head term *regulation of circadian sleep/wake cycle, sleep*.

Figures 3.5 and 3.7 show scatter plots of the results of ReviGO clustering, for terms with the latitude $P_{\text{empirical}} < 0.01$ and latitude $P_{\text{empirical}} < 0.001$, respectively. Regulation of myoblast fusion and tissue development are very close in terms of semantic similarity, as are nephrocyte filtration and regulation of circadian sleep/wake cycle, sleep. Figures 3.6 and 3.8 show similar plots of enriched clusters, where nodes represent clusters, and edge thicknesses represent similarity between them.

**Table 3.5** Results of *ReviGO* clustering of the most significant GO terms, with a focus on latitude. All of the head terms' latitude $P_{empirical}$ ($P_{env1}$ in the table) was less than 0.0001, and all of the terms contained at least five annotated genes. Italicized terms are clustered under the head term due to high semantic similarity. Head terms have the highest uniqueness among all the terms in the cluster. Empirical *P*-values for all six environmental variables are given in columns 5 through 10. A *P*-value of zero in the table denotes cases where resampling 10 000 (or in the case of latitude, i.e. $P_{env1}$, 100 000) random sets of SNPs from the genome was insufficient to determine the exact *P*-value (the true value is small, but positive). For details about environmental variables, see Section 3.1 and Table 2.2.

| GO term ID | description | uniq. | $N_{genes}$ | $P_{env1}$ | $P_{env2}$ | $P_{env3}$ | $P_{env4}$ | $P_{env5}$ | $P_{env6}$ |
|---|---|---|---|---|---|---|---|---|---|
| GO:0097206 | **nephrocyte filtration** | 0.841 | 7 | 0 | 0.0005 | 0.0001 | 0.0098 | 0 | 0 |
| GO:1901739 | **regulation of myoblast fusion** | 0.736 | 9 | 0.00002 | 0.0005 | 0.0073 | 0.4554 | 0 | 0 |
| GO:0045187 | **regulation of circadian sleep/wake cycle, sleep** | 0.599 | 15 | 0 | 0.0145 | 0 | 0.0048 | 0 | 0 |
| GO:0007622 | *rhythmic behavior* | 0.772 | 13 | 0.00024 | 0.0062 | 0.0019 | 0.1134 | 0 | 0.0001 |
| GO:0007623 | *circadian rhythm* | 0.876 | 41 | 0.00001 | 0.0006 | 0 | 0.1041 | 0 | 0 |
| GO:0048042 | *regulation of post-mating oviposition* | 0.659 | 12 | 0.00177 | 0.0027 | 0.0002 | 0.0108 | 0.0023 | 0.0015 |
| GO:2000252 | *negative regulation of feeding behavior* | 0.725 | 6 | 0.00201 | 0.0436 | 0.0559 | 0.0114 | 0.0069 | 0.0071 |
| GO:0009649 | *entrainment of circadian clock* | 0.768 | 12 | 0.00578 | 0 | 0.015 | 0.1301 | 0.0077 | 0.0198 |
| GO:0045475 | *locomotor rhythm* | 0.727 | 65 | 0.00093 | 0.002 | 0 | 0.0408 | 0 | 0 |
| GO:0007005 | **mitochondrion organization** | 0.883 | 37 | 0.00006 | 0.0134 | 0.0001 | 0.2539 | 0.0003 | 0.0001 |
| GO:0009888 | **tissue development** | 0.848 | 21 | 0.00002 | 0.038 | 0.0042 | 0.0175 | 0 | 0 |

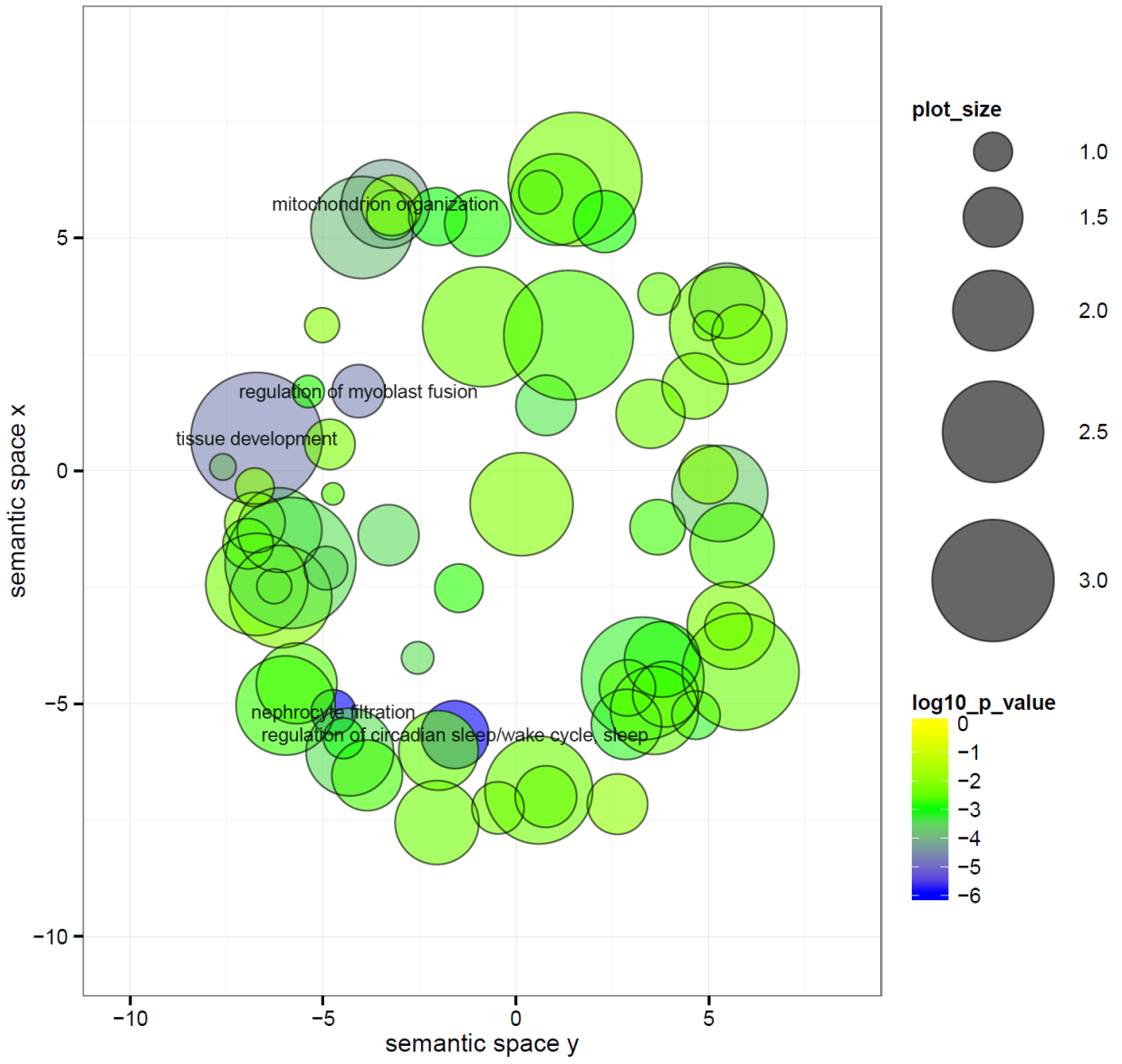



**Figure 3.5** A scatter plot of the results of ReviGO clustering of the most significant GO terms (latitude $P_{empirical}$ in all cases < 0.01, and $N_{genes} \geq 5$). X and Y axes denote dimensions in the semantic space of similarity between the GO terms, as determined by the SimRel algorithm (see Materials and Methods). GO terms / clusters closer to one another on the scatter plot have more common ancestors in the gene ontology, i.e. they have a high similarity score. Sizes of the circles denote relative numbers of associated genes. Colors from blue to green to yellow denote the $log_{10}$ of the empirical *P*-value for latitude. Labeled terms are the most highly significant ones.

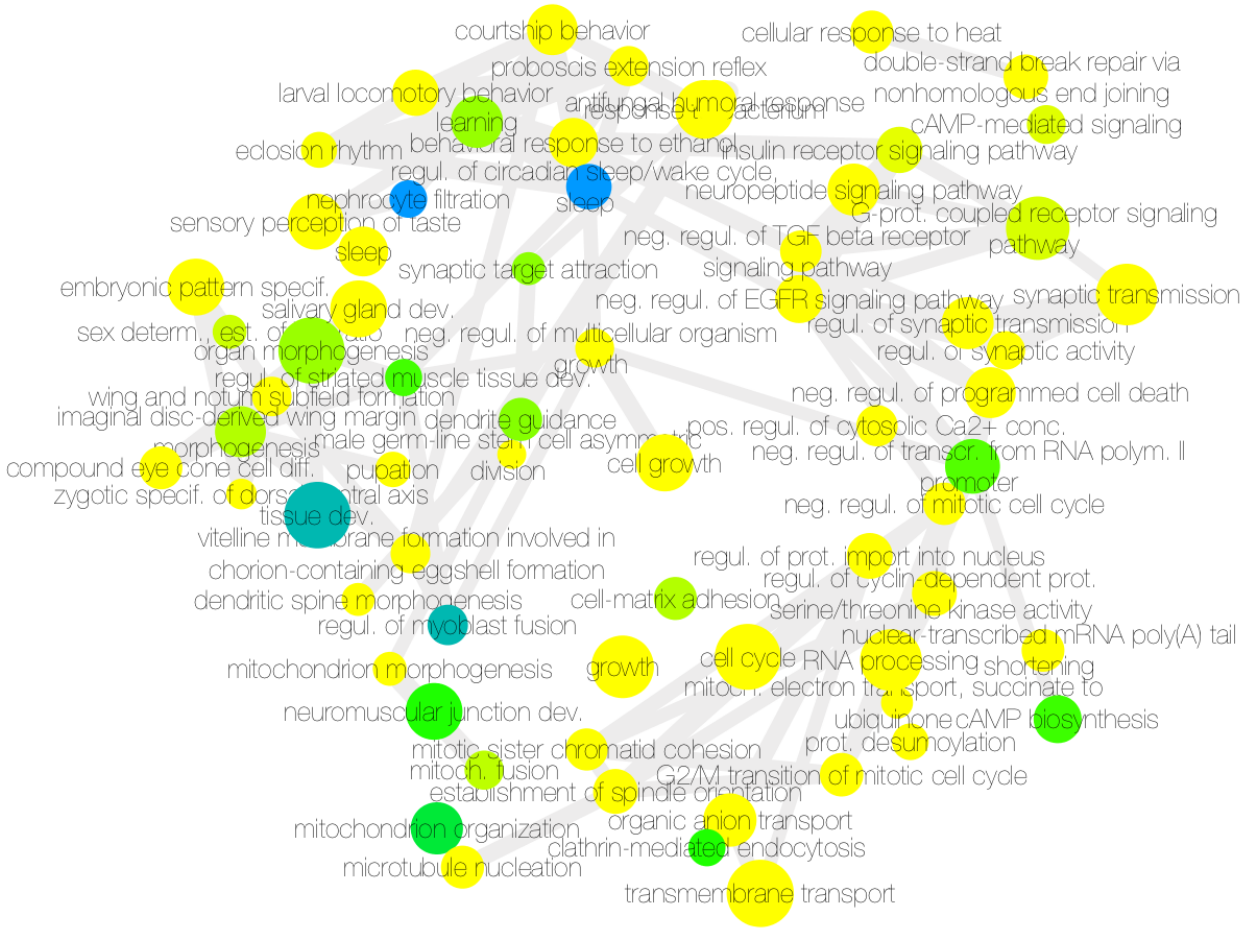



**Figure 3.6** A plot of all highly enriched (latitude $P_{empirical} < 0.01$) GO terms / clusters with $\geq 5$ genes, showing their semantic similarities, as calculated by the SimRel algorithm, in the form of edges of different thickness. Greater thickness denotes higher similarity between GO terms / clusters. Size of the circles denotes the relative numbers of their genes, while color (from blue to green to yellow) denotes statistical significance (latitude). Blue terms / clusters are the most highly significant.

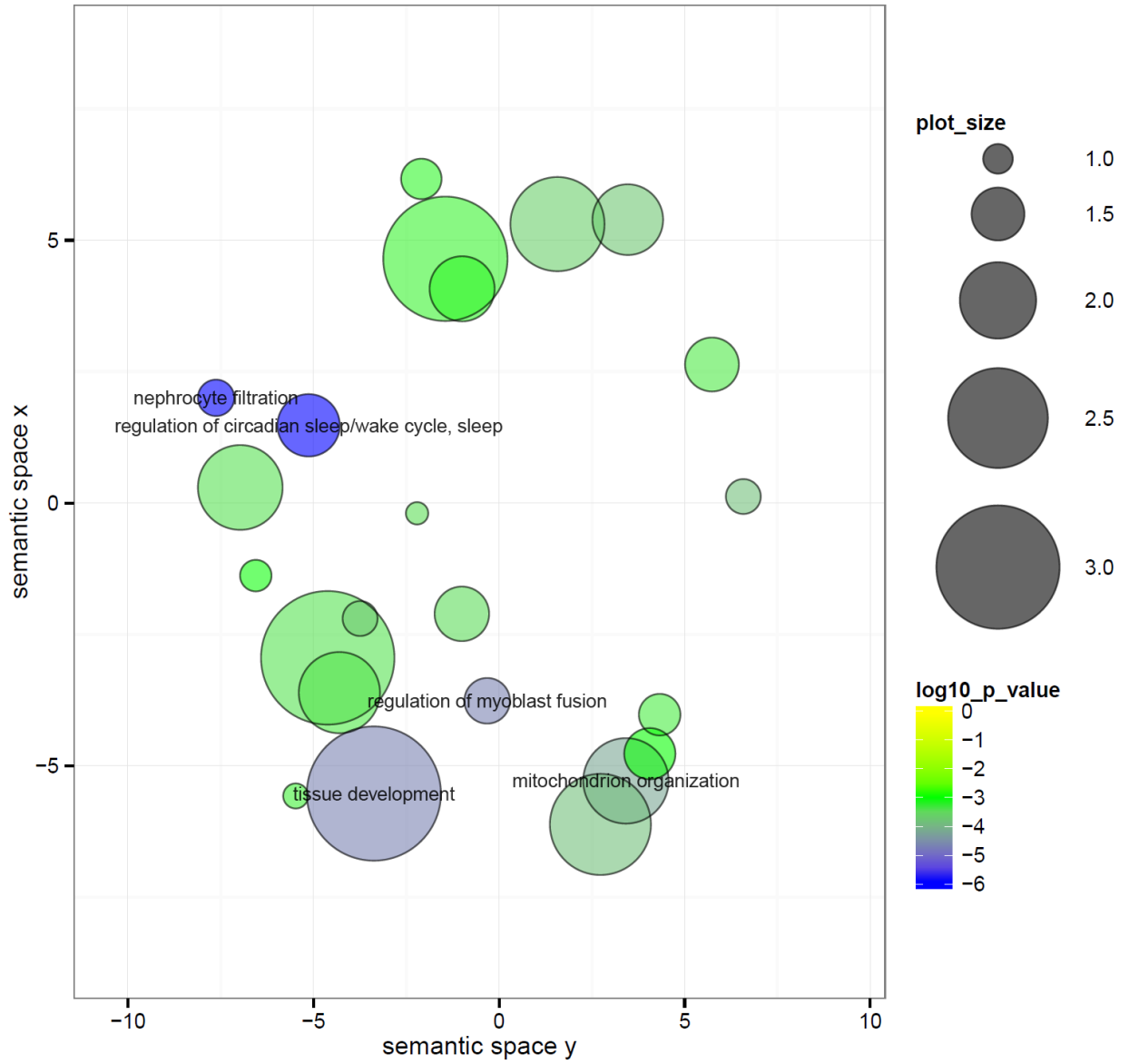



**Figure 3.7** A scatter plot of the results of ReviGO clustering of the most significant GO terms, but with more stringent conditions compared to Figure 3.5 (latitude $P_{empirical}$ in all cases < 0.001, and $N_{genes} \geq 5$). X and Y axes denote dimensions in the semantic space of similarity between the GO terms, as determined by the SimRel algorithm (see Materials and Methods). GO terms / clusters closer to one another on the scatter plot have more common ancestors in the gene ontology, i.e. they have a high similarity score. Sizes of the circles denote relative numbers of associated genes. Colors from blue to green to yellow denote the $log_{10}$ of the empirical *P*-value for latitude. Labeled terms are the most highly significant ones.

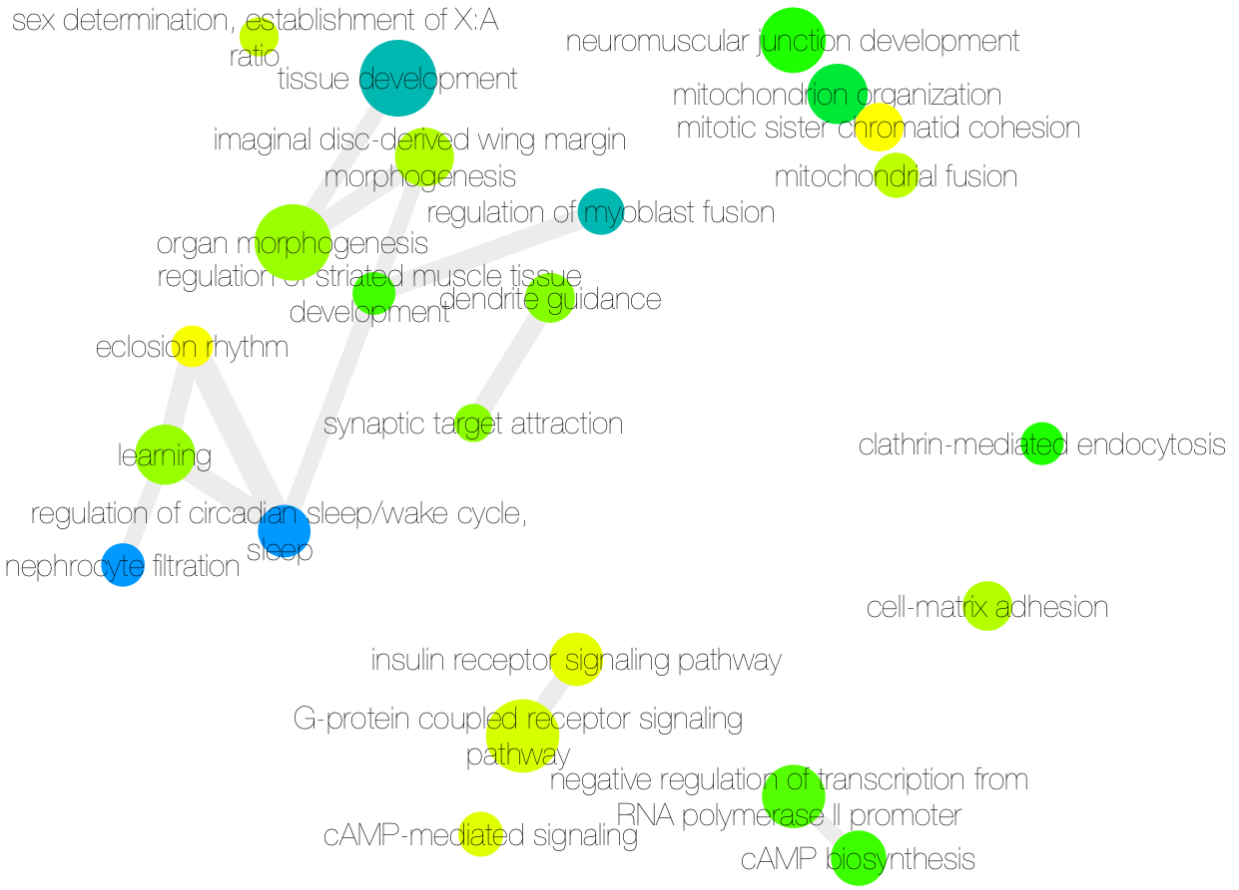



**Figure 3.8** A plot of all very highly enriched (latitude $P_{empirical}$ < 0.001) GO terms / clusters with ≥ 5 genes, showing their semantic similarities, as calculated by the SimRel algorithm, in the form of edges of different thickness. Greater thickness denotes higher similarity between GO terms / clusters. Size of the circles denotes the relative numbers of their genes, while color (from blue to green to yellow) denotes statistical significance (latitude). Blue terms / clusters are the most highly significant.

### 3.2.1 Top enriched GO terms / clusters

At the 0.7 cutoff for semantic similarity based on their ancestral GO terms in the ontology (default settings for the SimRel algorithm), the largest cluster produced by the SimRel algorithm was headed by the term GO:0045187, regulation of sleep/wake cycle, sleep (latitude $P_{empirical}$<0.00001). The remaining GO terms were, in order of decreasing significance: *circadian rhythm*, *rhythmic behavior*, *locomotor rhythm*, *regulation of post-mating oviposition*, *negative regulation of feeding behavior*, and *entrainment of circadian clock* (Table 3.5). If we relax the similarity threshold for SimRel from 0.7 to 0.5, and the cutoff for significance from 0.001 to 0.01, then the former GO terms are also joined by the following:

*learning*, *locomotor rhythm*, *eclosion rhythm*, *larval locomotory behavior*, *courtship behavior*, *memory*, *olfactory learning*, *circadian temperature homeostasis*, *retina homeostasis*, and *operant conditioning*. In either case, the GO terms inside the cluster are not only similar when it comes to shared ancestry in the gene ontology graph, but they also share a number of genes. Most notably, 13 genes are related to at least three of the circadian rhythm gene sets: discs overgrown (*dco*), Clock (*Clk*), cycle (*cyc*), period (*per*), timeless (*tim*), Cyclic-AMP response element binding protein B (*CrebB*), cryptochrome (*cry*), Casein kinase II β subunit (*CkIIβ*), Protein kinase, cAMP-dependent, regulatory subunit type 2 (*Pka-R2*), Protein kinase, cAMP-dependent, catalytic subunit 1 (*Pka-C1*), Pigment-dispersing factor receptor (*Pdfr*), shaggy (*sgg*), and vrille (*vri*). Additionally, five genes were shared between two of the circadian rhythm gene sets: Shaker (*Sh*), (*Fmr1*), (*Atx-2*), glass (*gl*), and takeout (*to*).

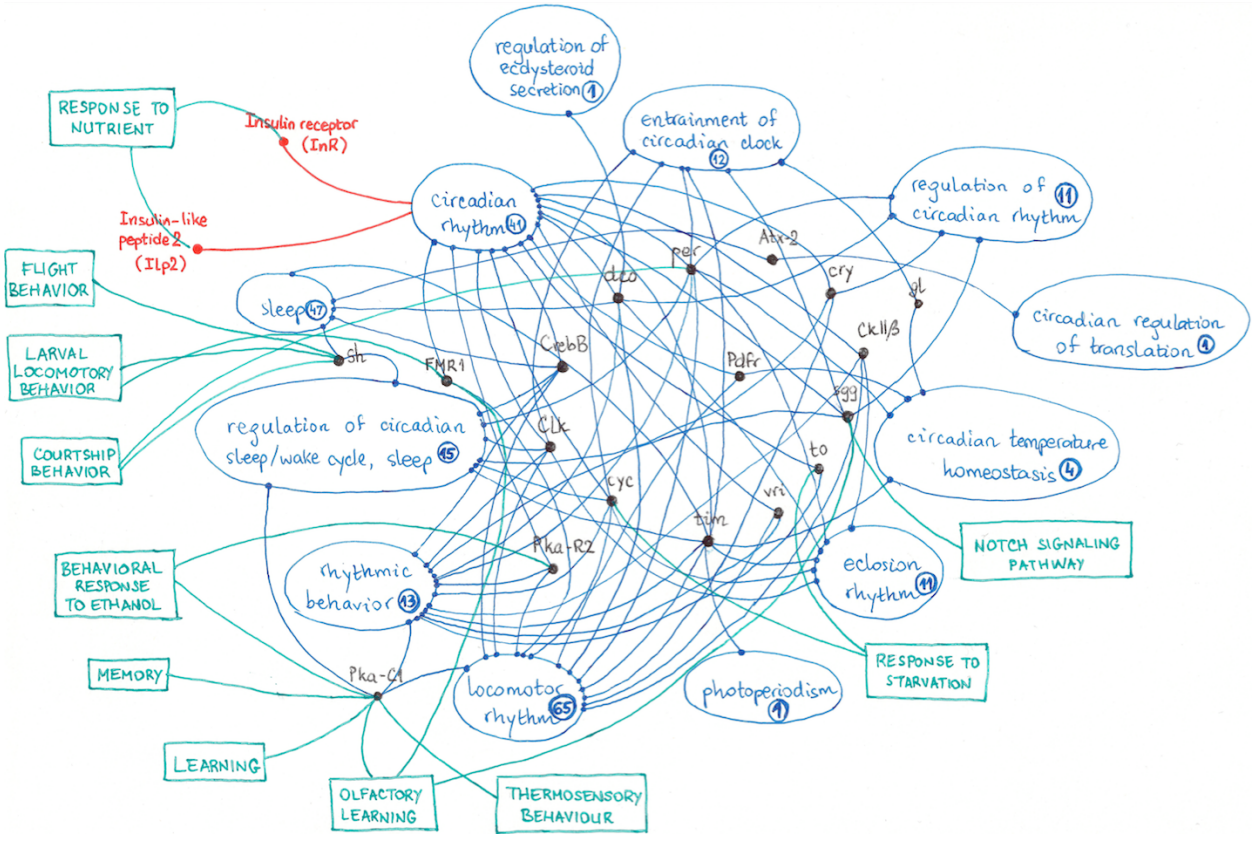



**Figure 3.9** Manually drawn Circadian rhythms cluster. Blue ovals represent the enriched ($P_{latitude}$ < 0.05) GO terms belonging to the cluster, black dots represent shared genes, and blue lines their connections to the cluster's GO terms. Green rectangles represent enriched GO terms from other clusters, and green lines their connections to the cluster's shared genes.

The remaining four GO terms from Table 3.5 - nephrocyte filtration, mitochondrion organization, tissue development, and regulation of myoblast fusion - were all the only GO term in their respective cluster, due to their high uniqueness scores, as calculated by SimRel at the 0.7 cutoff. Figure 3.8 shows that nephrocyte filtration is close to the circadian rhythm term in the semantic space, and that tissue development and myoblast fusion are also quite closely related. Mitochondrion organization is not closely related to any other cluster of significant GOs, even among the larger group of GO terms defined at a more relaxed cutoff of empirical *P*-value for latitude at 0.01 (Figures 3.5 and 3.6).

## 3.2.2 Enrichment of Reactome pathways

### 3.2.2.1 Signal Transduction and Metabolism

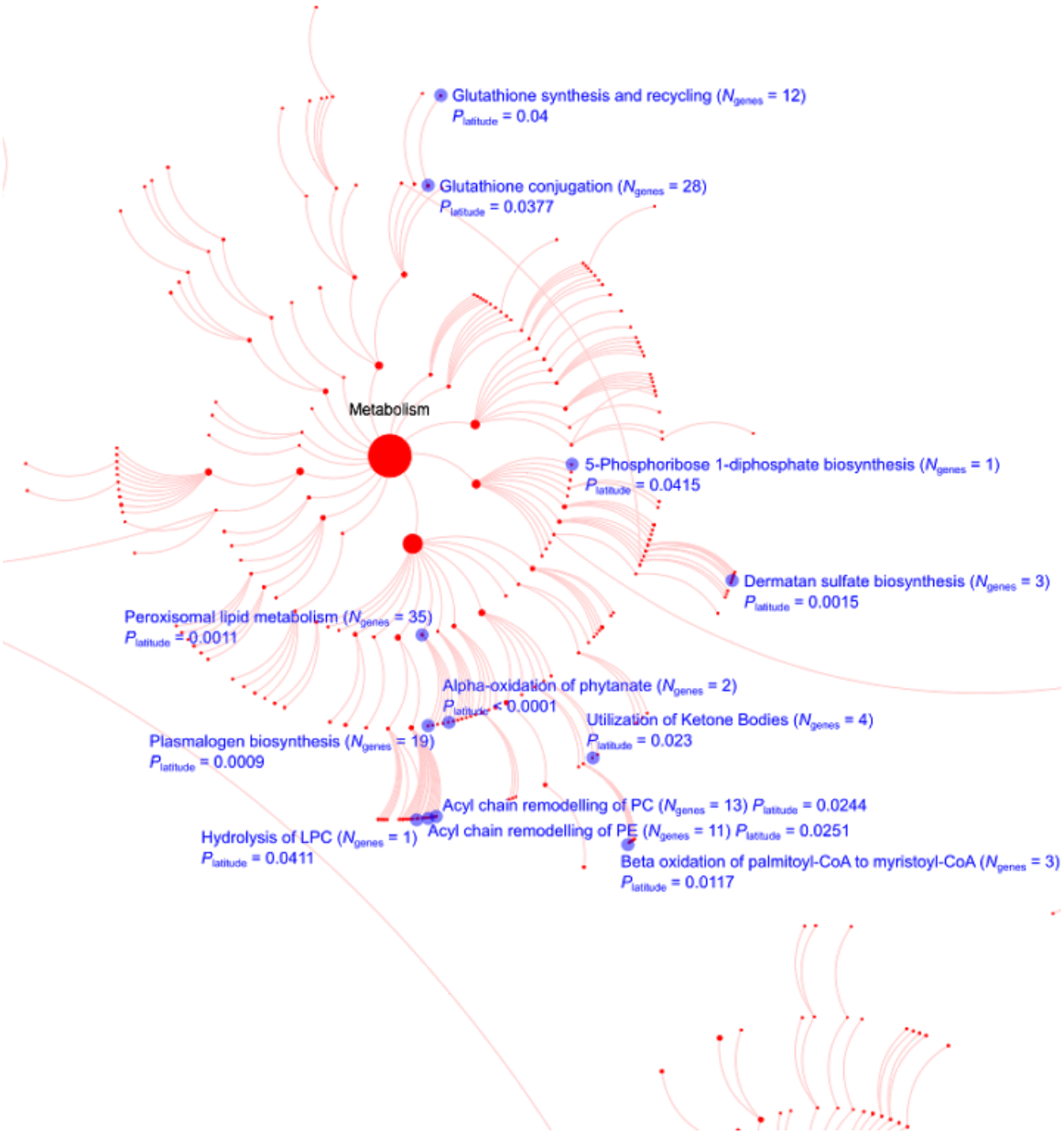



**Figure 3.10** Reactome pathways related to Metabolism and significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

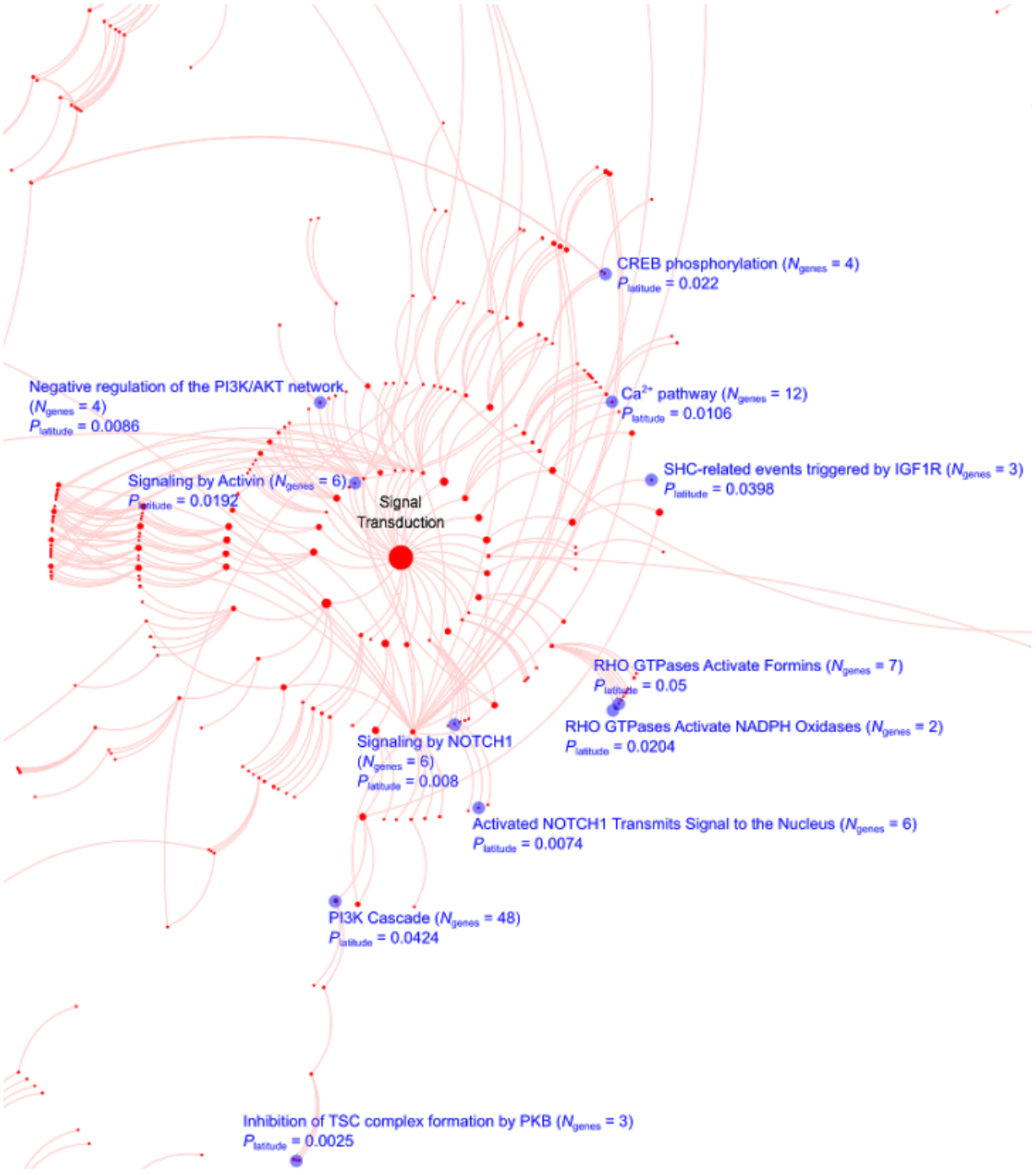



**Figure 3.11** Reactome pathways related to Signal Transduction and significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

For latitude, we found 11 significant (empirical $P_{latitude} < 0.05$) Reactome pathways involved in *Signal Transduction* (Figure 3.11) and 12 significant Reactome pathways involved in *Metabolism* (Figure 3.10). Various signaling cascades showed signals of selection: *Notch* signaling pathway, signaling by activin (a member of the TGF beta superfamily of ligands), signaling by insulin receptor and by Type 1 insulin-like growth factor 1 receptor (IGF1R), signaling by Phosphoinositide 3-kinase (PI3K) / Protein kinase B (AKT) / mechanistic target of rapamycin (mTOR), signaling by the Rho family of GTPases, signaling by integration 1 / Wingless (Wnt), and Nerve growth factor (NGF) signaling (Fig. 1). Additionally, selection signals were present in the $Ca^{2+}$ pathway and in the cAMP response element-binding protein (CREB) phosphorylation pathway. Six of these pathways are annotated with 5 or more genes, and out of these, *Activated NOTCH1 Transmits Signal to the Nucleus* (biosystem 1329452) had the highest significance ($P_{latitude} = 0.0074$). Its 6 genes contain a total of 195 SNPs in our analysis, the most significant of which had a BF of ≈ 9.35. Since the cut-off for our candidate genes in Section 3.1 was a BF of at least $e^5$ (≈ 148) (Kass & Raftery 1995), none of the genes survived, and the pathway therefore had no chance of being discovered by the classical enrichment approach. However, taking into account the distribution of BFs over all SNPs, it achieved significance under our SNP-set-enrichment approach.

Among the 12 metabolism-related pathways showing signals of selection, 8 were involved in various roles in lipid and lipoprotein metabolism, 2 in biological oxidations, and 2 in the metabolism of carbohydrates. Six are annotated with 5 or more genes, the most significant being *Plasmalogen biosynthesis* (biosystem 1328972), with $P_{latitude} = 0.0009$, 19 genes, and 348 SNPs, 3 of which surpassed the candidate SNP cut-off in our analysis (BF > $e^5$). However, since these 3 SNPs only map to 3 genes (*CG5065*, *CG8303*, and *wat*), they were not enough to push the pathway past the significance threshold in the classical pathway enrichment we performed in Section 3.1. However, in our analysis here, 5 out of 6 environmental variables had empirical *P*-values < 0.05.

### 3.2.2.2 Gene Expression and Transmembrane Transport of Small Molecules

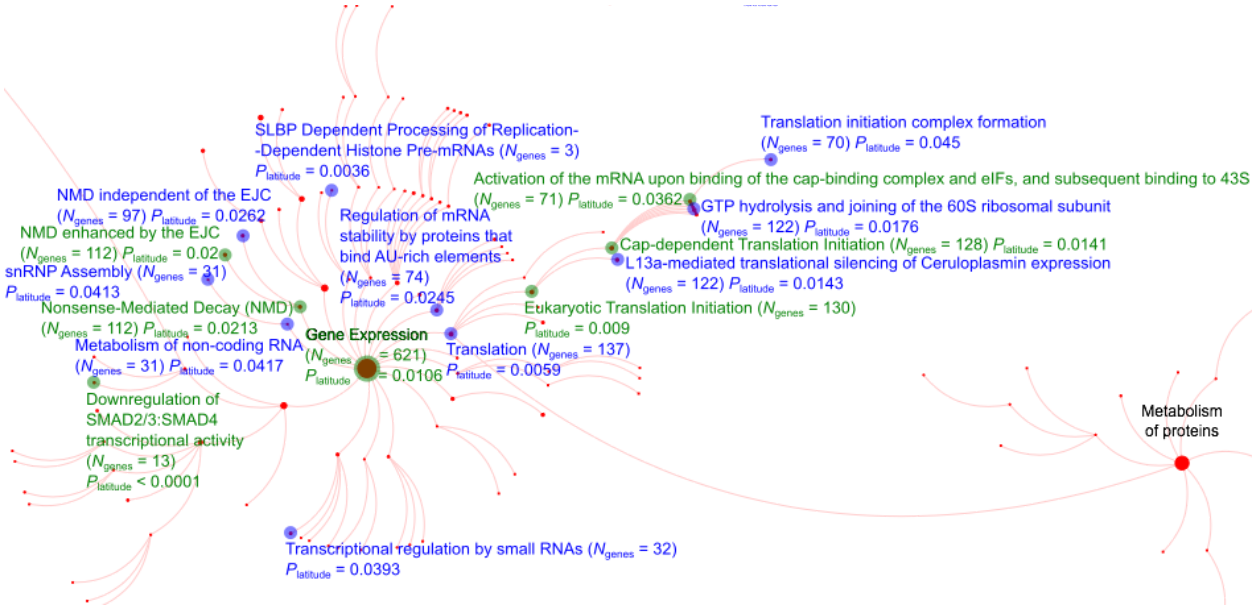



**Figure 3.12** Reactome pathways related to Gene Expression and significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical}$ < 0.05).

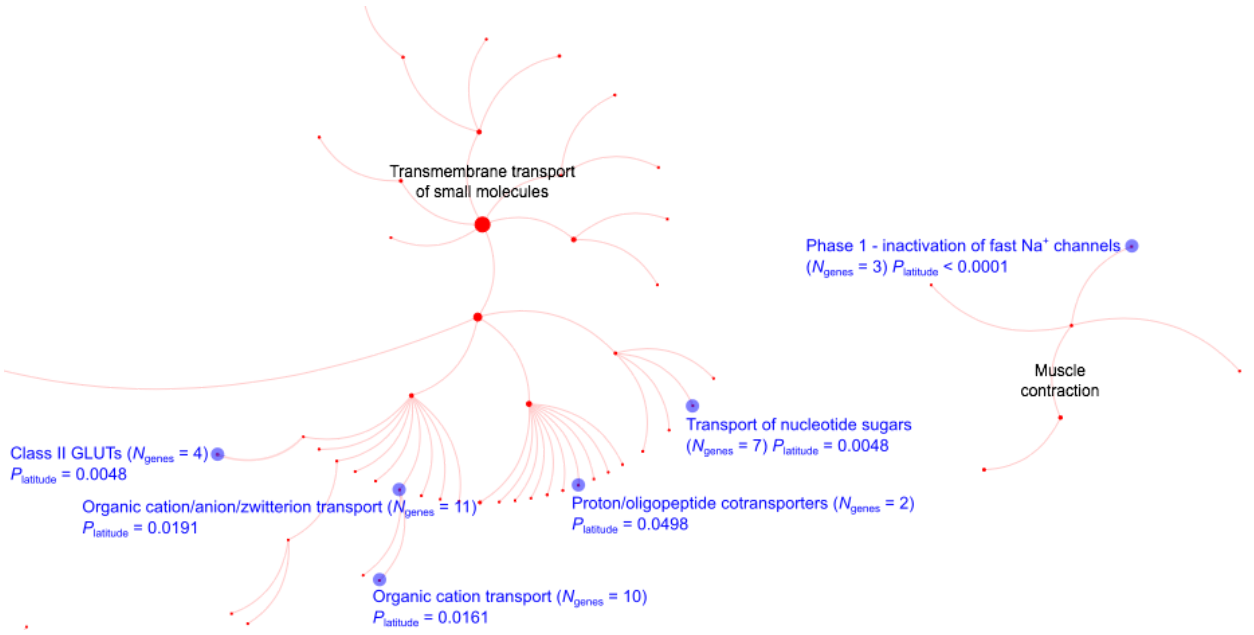



**Figure 3.13** Reactome pathways related to Transmembrane transport of small molecules and significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical}$ < 0.05).

We found a total of 16 pathways enriched in signals of selection related to gene expression (Figure 3.12) (17 including *Gene Expression* as a gene set), 5 pathways related to the transmembrane transport of small molecules (Figure 3.13), and one pathway related to muscle contraction (Figure 3.13). Translation (total $N_{genes}$ = 137) was highly enriched for latitude ($P_{empirical}$ = 0.0059), as well as 6 of its child pathways. Three additional 1st order pathways were enriched: metabolism of non-coding RNA ($P_{empirical}$ = 0.0417), regulation of mRNA stability by proteins that bind AU-rich elements ($P_{empirical}$ = 0.0245), and nonsense-mediated decay (NMD) ($P_{empirical}$ = 0.0213). *Gene Expression* (total $N_{genes}$ = 621) as a gene set was also enriched overall with highly significant SNPs for latitude ($P_{empirical}$ = 0.0106), as well as altitude ($P_{empirical}$ = 0.0454), coldest month temperature ($P_{empirical}$ = 0.0083), and the yearly minimum ($P_{empirical}$ = 0.0136) and maximum temperatures ($P_{empirical}$ = 0.0407). For latitude, *Downregulation of SMAD2/3:SMAD4 transcriptional activity* (biosystem 1329447) was by far the most enriched pathway, to the point where 10000 genomic background samplings were not enough to ascertain the exact empirical *P*-value (i.e. $P_{empirical} < 0.0001$). It was also significant for coldest month ($P_{empirical}$ = 0.0128), yearly minimum ($P_{empirical}$ = 0.001), and yearly maximum temperatures ($P_{empirical}$ = 0.0015). Its 13 genes contained 524 SNPs, four of which had BFs > $e^5$. Three of those four SNPs defined the candidate gene *Snoo* (FBgn0085450) in the Section 3.1 pathway enrichment analysis, while one defined the candidate gene *Smr* (FBgn0265523). Even though they were very strong outliers (BFs for latitude between 187.74 and 925.7), these two genes alone were again not sufficient for this pathway to be detected as significant in the classical enrichment we performed in Section 3.1.

The *Transmembrane transport of small molecules* cluster is notably represented by *Transport of nucleotide sugars* (biosystem 1328781), whose 7 genes and 55 SNPs show a signal of polygenic adaptation for all six environmental variables, with $P_{empirical}$ of 0.0048, 0.0409, 0.0051, 0.0151, 0.003, and 0.0036, for latitude, altitude, and the four measures of temperature, respectively. This is despite the fact that none of the individual SNPs achieved particularly large BFs for environmental correlation.

### 3.2.2.3 Developmental Biology and the Immune System

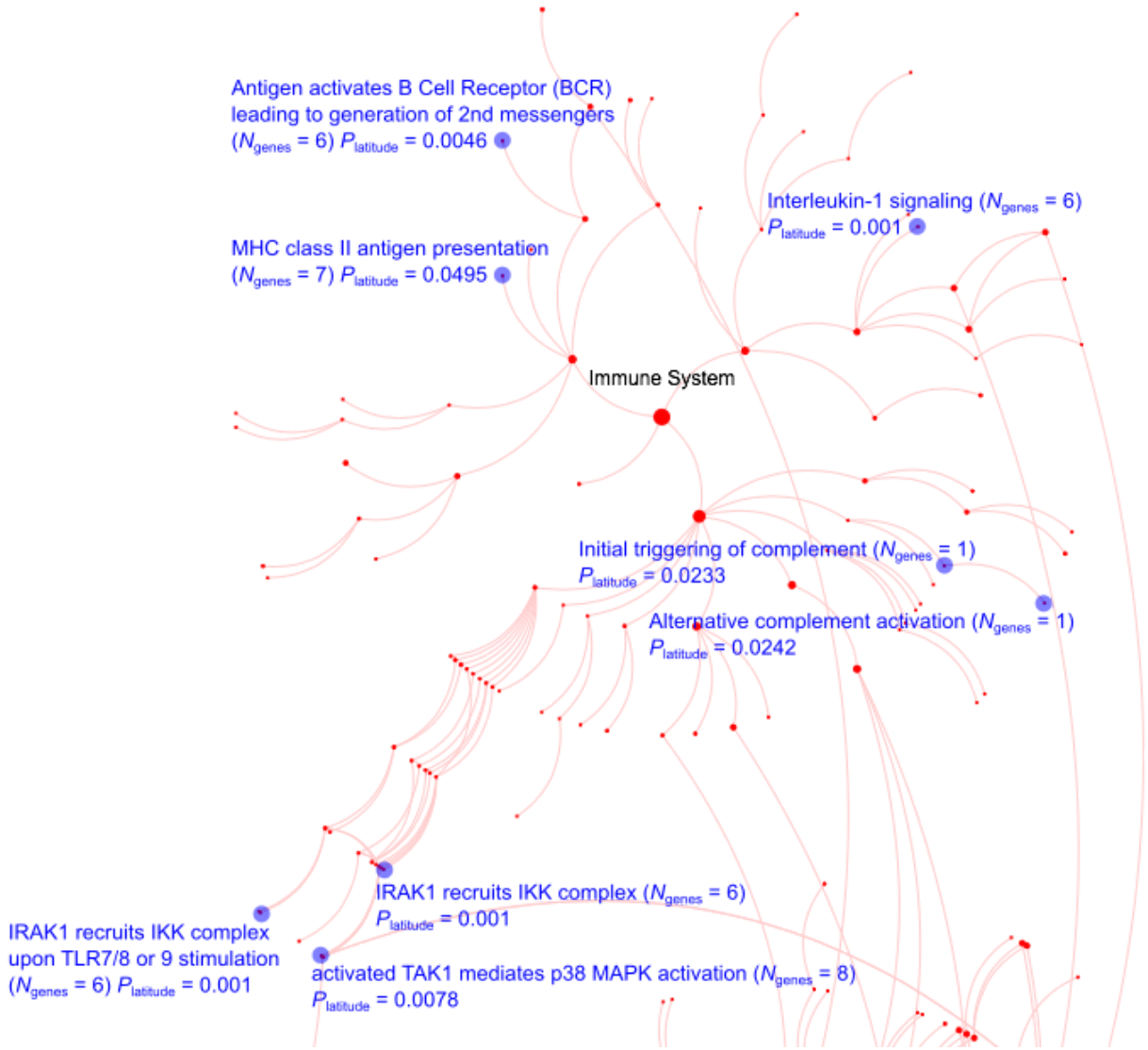



**Figure 3.14** Reactome pathways related to Immune System and significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

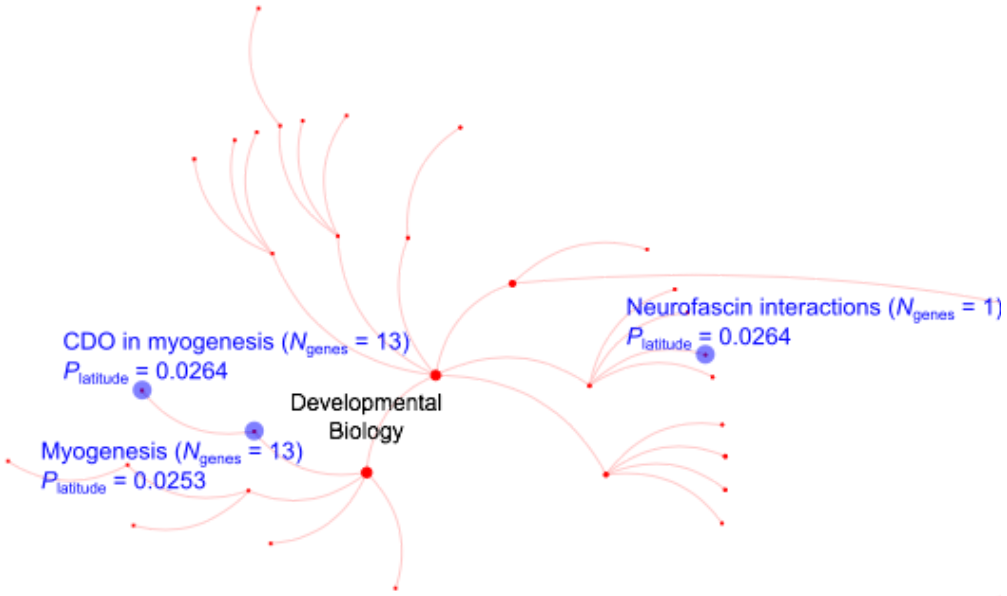



**Figure 3.15** Reactome pathways related to Developmental Biology and significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

Three pathways related to developmental biology (Figure 3.15) and 8 pathways related to the immune system (Figure 3.14) were enriched in signals of natural selection. Out of the developmental terms, two were related to myogenesis (*Myogenesis*: latitude $P_{empirical}$ = 0.0253, and *CDO in myogenesis*: latitude $P_{empirical}$ = 0.0264), and one to axon guidance (*Neurofascin interactions*: latitude $P_{empirical}$ = 0.0264). The most significant pathway of the *Developmental Biology* cluster was *Myogenesis* (biosystem 1328897), with 420 SNPs scattered over its 13 annotated genes, none of which individually show the strongest evidence of selection (BFs of the top two SNPs were 22.1 and 105.3 for latitude). However, overall the fact that this pathway affects the development of muscles, does fit into the context of many other genes and pathways that we found show signals of adaptation for locomotory performance (see Discussion).

Five out of eight enriched immune system pathways were related to the innate immune system, three of which were child terms of *Toll*-like receptor cascades, and two of which were related to the complement cascade. Of the three remaining immune system pathways, two were related to the adaptive immune system, and one to cytokine signaling.

Furthermore, five of the eight significant immune system pathways showed very strong signals of polygenic adaptation for latitude ($P_{empirical}$ < 0.001), five out of eight were significant for altitude ($P_{empirical}$ between 0.0175 and 0.0334), and six out of eight for coldest month temperature ($P_{empirical}$ between 0.0098 and 0.0156). The most significant pathway, *IRAK1 recruits IKK complex* (biosystem 1329245), showed strong evidence of polygenic adaptation for latitude, altitude, and 3 out of 4 temperature variables.

### 3.2.2.4 Neuronal System and Hemostasis

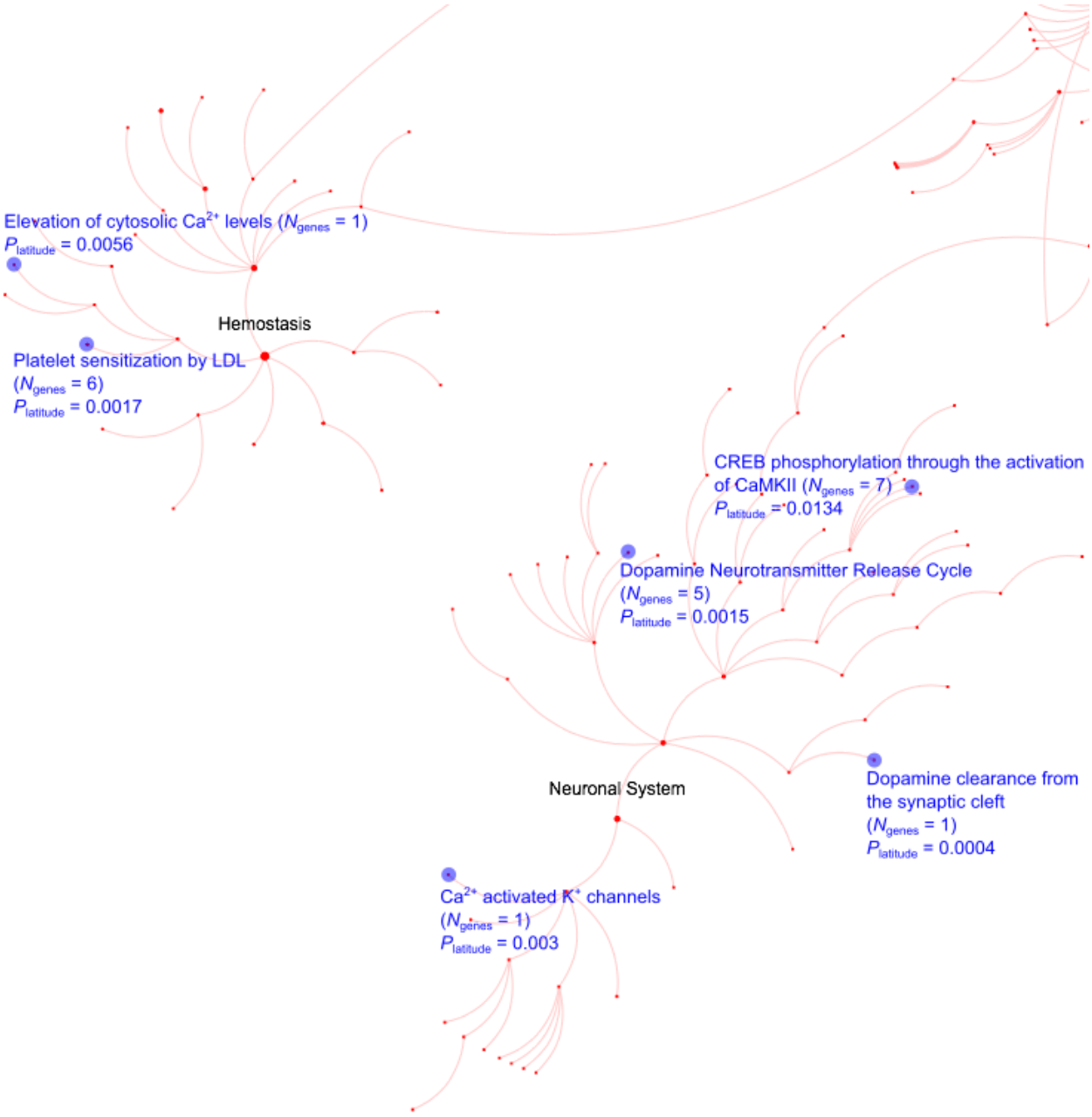



**Figure 3.16** Reactome pathways related to Neuronal System and to Hemostasis significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

Two hemostasis-related and four neuronal system-related Reactome pathways were significantly enriched for signals of adaptation to latitude ($P_{empirical} < 0.05$) (Figure 3.16). Both

hemostasis pathways were related to platelet homeostasis, while three out of four neuronal system pathways were related to transmission across chemical synapses, and the remaining one to a function in potassium channels.

From the *Neuronal System* cluster, *Dopamine Neurotransmitter Release Cycle* (biosystem 1328620) displayed strong signals of adaptation to latitude ($P_{empirical}$ = 0.0015), the strongest among pathways with 5 or more genes from the cluster.

#### 3.2.2.5 Other enriched Reactome pathways

The remaining enriched Reactome pathways are mostly related to housekeeping functions such as DNA replication and repair, cell cycle, and programmed cell death, and it is therefore difficult to speculate in what way their underlying genes might be adapting to local conditions.

*DNA Repair* as a whole was enriched with signals of selection (latitude $P_{empirical}$ = 0.0142), as were 13 of its child Reactome pathways. Eight enriched pathways were involved in DNA double-strand break repair, and their parent pathway *DNA Double-Strand Break Repair* was also enriched as a whole (latitude $P_{empirical}$ = 0.0017). Additionally, two enriched pathways were involved in mismatch repair, and three in global genome nucleotide excision repair.

A total of 11 Reactome pathways enriched for latitudinal selection were related to the cell cycle, one was involved in the regulation of DNA replication, and one in programmed cell death. The cell-cycle-related pathways with selection signals were involved in: (1) the mitotic phase (two pathways), (2) regulation of the mitotic phase cycle (four pathways), (3) the synthesis phase (one pathway), and (4) the mitotic G2-G2/M phase (four pathways). Additionally, *Cell Cycle, Mitotic* was enriched as a whole (latitude $P_{empirical}$ = 0.0365).

Among the remaining pathways, we found enrichments for three related to cellular responses to stress, three related to extracellular matrix formation, and two to organelle biogenesis and maintenance. Two of the three stress responses were related to oxygen, and one to senescence due to telomere shortening. Similarly, two of the ECM pathways were elastic fibre-related, and the remaining one laminin-related. Both organellar pathways with signals of selection were related to ciliary function.

Metabolism of proteins showed five Reactome networks with signals of selection: three related to a post-translational modification process called sumoylation, and two to the trimming of N-linked glycans. Four Reactome pathways showing signals of selection were vesicle-transport-related, all four with functions in membrane trafficking through various types of clathrin-coated vesicles. Two additional enriched pathways were *Cell-Cell communication* (latitude $P_{empirical}$ = 0.0158) and its 1st order child pathway *Nephrin interactions* (latitude $P_{empirical}$ = 0.0012) (Figure 3.13).

# CHAPTER 4: GENERAL DISCUSSION

Discussion - an illustration (*Creative Commons Zero*)

We divided the results into two main sections, the first one dealing with adaptation on the gene level (Section 3.1), and the second one with adaptation on the gene network level (Section 3.2). In Section 3.1, we used an approach incorporating multiple lines of evidence for adaptation to cold. To do this, we examined (1) SNPs related to cold tolerance from GWA studies, (2) candidate genes from literature, and (3) genes with SNPs that show strong evidence of being adaptive. We investigated the potential effects of linkage to QTLs known to affect cold tolerance, as well as cosmopolitan inversions. Furthermore, we looked for overlaps with candidate genes from other studies, and added information from GO enrichments of those candidate genes, as well as our own. In the first part of the Discussion, we discuss the results of these approaches, particularly in the context of selective forces and fitness tradeoffs, and propose an adaptive network of genes in the core of the complex cold tolerance phenotype.

In Section 3.2, we tested for enrichment of GO terms and Reactome pathways with an approach that aims to capture a signal of polygenic adaptation. We did this by taking into account Bayes factors for correlation with environmental variables of all the individual SNPs that can be mapped to GO terms and Reactome pathways. The approach was similar to the one we used to establish significance of $F_{ST}$ over SNPs that mapped to our three quantitative traits from GWAS. We clustered the most significant GO terms and examined the connections between the significant Reactome terms in the various domains of the Reactome network graph. Finally, we discuss these results in the context of local adaptation to European environments. For instance, the transcription factor *Myocyte enhancer factor 2* (*Mef2*) is involved in adult circadian locomotor behaviors (Blanchard *et al.* 2010), a gene set defined in our analysis by the gene ontology category GO:0045475 (locomotor rhythm), highly enriched with signals of selection for all six environmental variables. But it also operates as a critical transcriptional switch in the adult fat body between the lipid and glycogen metabolism and the innate immune response (Clark *et al.* 2013). In this example, we might note the connection because we know, for instance, that metabolic levels are related to ambient temperature (Berrigan & Partridge 1997), that circadian behaviors show clinal patterns of selection in European flies (Tauber *et al.* 2007), that the immune response often shows patterns of strong selection (Schlenke & Begun 2003; Lazzaro 2008), and finally, that nutrition has an important contribution to fly immunity (Unckless *et al.* 2015).

## 4.1 Evidence for adaptation to cold from genome-wide association studies

We started our analysis by quantifying the amount of population differentiation at SNPs associated with CCRT, RSS, and SR. Each of the three traits has an important adaptive role. It is possible that the observed difference in average $F_{ST}$ between the three traits is due to different fitness trade-offs among the traits. We know from previous studies that RSS is sensitive to varying environmental conditions, especially low temperature, and that it has a strong effect on fitness (Boulétreau-Merle & Fouillet 2002; Hoffmann 2010; Goenaga *et al.* 2010, 2012, 2013). In natural populations, RSS depends on pre-adult resource acquisition, which varies with developmental temperature (Chippindale *et al.* 1998). Both artificial selection experiments (Hoffmann *et al.* 2005) and quantitative genetics studies of these traits (Ayroles *et al.* 2009) have demonstrated a trade-off between cold and starvation

resistance. Lines with higher cold tolerance (i.e. shorter CCRT) mate more rapidly and have high competitive fitness, which comes at the expense of surviving starvation stress, while lines with higher RSS tend to have longer life spans at the cost of fitness (Ayroles *et al.* 2009). Additionally, lines able to postpone egg-laying in the autumn have a longer life expectancy and greater RSS, which is favored under low temperatures (Boulétreau-Merle & Fouillet 2002) and has implications for surviving cold conditions when food is scarce (Hoffmann *et al.* 2005; Goenaga *et al.* 2012). Another factor important for the RSS component of fitness is food availability during the cold period of the year. Interestingly, because fruit flies feed on decomposing fruit and on the bacteria and yeasts that grow on it (Markow & O'Grady 2008), a colder environment might reduce the rate of fruit decay, shortening the period of time when food is scarce, and thus neutralizing the need for increased RSS in lower temperatures (Goenaga *et al.* 2013). As for SR, there might be a similar trade-off between the fitness advantages of good locomotor performance (e.g. rapid mating, defense of territory, escape from predators) and the pressure to perform better in cold conditions. Indeed, both developmental and adult temperatures have been shown to affect locomotor performance (Crill *et al.* 1996; Gibert *et al.* 2001a). In short, the $F_{ST}$ and BFs of SNPs associated with phenotypic traits are consistent with the conclusion that European climate might be at the same time reducing both RSS and SR in favor of greater cold resistance, and incurring an additional fitness cost on RSS due to food being naturally preserved by colder temperatures.

## 4.2 Evidence from candidate genes from literature

The second line of evidence for adaptation comes from our study of candidate genes previously described in the literature (Tables S1-S2). Both $F_{ST}$ (Table S3) and BFs (Table S4) showed that genes related to tolerance to cold or heat, disturbance, and starvation stress generally contain SNPs whose allele frequency patterns are better explained by selection. We found no literature candidate genes in the neighborhood of the breakpoints of the most common cosmopolitan inversions (see inversion_analysis.xlsx on Dryad). This is particularly striking in (1) heat-shock genes *Hsp26* and *Hsp68*, whose synonymous variants suggest that they might be under fitness cost in European populations due to preferred codon usage; (2) the gene *lola*, which contained highly differentiated variants at both 3' and 5' regions, introns (including the intron variant chr2R_6394221 with strong evidence

(*ln*(BF)>3; $P < 0.03$) for 4 environmental variables), as well as synonymous and nonsynonymous sites; and (3) another 6 genes with highly differentiated nonsynonymous variants (*CG18140*, *CG31738*, *CG12943*, *CG30379*, *nclb*, and *chas*). While in the case of some genes not all population comparisons showed an elevated $F_{ST}$, the results are nevertheless indicative. Furthermore, our aim here was not to show that genes are disproportionately associated with CCRT, rather than the other two traits. Indeed, previous research has shown that there are fitness trade-offs between CCRT and starvation resistance, and likewise that locomotion is related to metabolism rate and dependent on the ambient temperature. We therefore expect that selection might be acting on all three traits in a concerted manner. Another interesting question is whether there is any overlap between genes related to different traits. Although the SNPs associated with the three traits in recent GWAS studies (Mackay *et al.* 2012; Huang *et al.* 2012b, 2014) showed no overlap, literature suggests that several genes involved in lipid regulation can vary along latitudinal clines with significantly different temperature regimes. For instance, *CG12054* codes for a transcription factor that is a target of *Forkhead box O*, itself a transcription factor that directly influences lifespan by regulating lipid metabolism (Alic *et al.* 2014). Furthermore, *Dyrk2*, *hdc*, *shep* and *chas* function in larval fat storage (Reis *et al.* 2010), *Ire1* is involved in the regulation of energy metabolism (Pile *et al.* 2003), and *Octbeta3R* mediates appetite for energy-rich foods (Zhang *et al.* 2013). Strikingly, all 7 of these genes have been related the North American latitudinal cline (Fabian *et al.* 2012), and most recently, all except for *Octbeta3R* have been found to respond to a range of different suboptimal rearing temperatures (Chen *et al.* 2015). The gene *shep* is particularly interesting because of functions that might make it important for both RSS and SR. As already mentioned, it plays a role in fat storage (Reis *et al.* 2010), but it was originally described in a forward genetic screen for gravitaxis (Armstrong *et al.* 2006), as its mutants impair the ability of the fly to perceive gravity. Many SNPs that map to *shep* were $F_{ST}$ outliers (top 1%, Table S3), but more interestingly, seven SNPs also had BFs with positive evidence of selection (*ln*(BF)>1 or nominal $P < 0.0014$) and two with strong evidence (*ln*(BF)>3 or $P < 0.03$) for at least one environmental variable (Tables S3 and S4). Six of these seven SNPs are intronic, but *chr3L_5154623* might be particularly adaptively important, as it maps to the 3'UTR region of the gene. The preponderance of candidate SNPs in the region of *shep* suggests its importance for adaptation related to both RSS and SR, and the fact that its SNPs also vary

clinally in North America (Fabian *et al.* 2012) and that it changes expression depending on rearing temperatures (Chen *et al.* 2015) also suggests its importance for CCRT.

Several previous QTL studies have attempted to localize regions of the genome responsible for heat and cold resistance (Morgan & Mackay 2006; Norry *et al.* 2007, 2008; Svetec *et al.* 2011; Wilches *et al.* 2014). To find out if genes from these QTLs show signatures of selection, we compared the candidate genes from these studies with our candidates significantly correlated with latitude and altitude (*ln*(BF)>5, as reported by *Bayenv2*, $P < 0.0273$). We found that four genes, *dlg1* (Norry *et al.* 2008), *CG1677* (Wilches *et al.* 2014), *rdgA* (Svetec *et al.* 2011), and *DnaJ-1* (Morgan & Mackay 2006), were also highly correlated with latitude in our study. None of them were also correlated with altitude. However, *rdgA* was also found to be a candidate gene in the North American cline (Fabian *et al.* 2012). Taken together, the evidence from both high $F_{ST}$ and from significant BFs confirms adaptive significance of several genes previously known to be important for CCRT, RSS or SR, and in some cases even all three traits.

## 4.3 Evidence from genome-wide top candidate genes

The final line of evidence for cold adaptation comes from our genome-wide analysis (Figure 3.1) of genes that contained variants with particularly strong evidence of environmental selection (*ln*(BF)>5, (Kass & Raftery 1995) or $P < 0.0273$), which means that the model including selection is more than $e^5$ (or ≈148) times more likely than neutrality. Since latitude and altitude are variables that may account for more conditions than just temperature (e.g. length of day, amount of insolation, seasonality, amount of oxygen, pressure), and because temperature variables were correlated to latitude, we decided to focus the analysis on genes with strong evidence of selection particularly for these two variables. An intersection of genes with overwhelming evidence for both altitudinal and latitudinal selection (Figure 3.1) would thus control for many potential confounding factors. As an additional level of control for false positives, we particularly closely examined genes from the overlap of latitude and altitude that have also been proposed from a North American cline (Fabian *et al.* 2012). We discovered that 8 genes (*Ets65A*, *Elk*, *sba*, *CG32066*, *dpr8*, *CG8177*, *X11Lbeta*, and *CG42699*) conformed to all of these strict conditions. Surprisingly, we found that these genes could be organized into a network, where each gene was functionally related to up to

three of its neighbors (Figure 3.2). Four of these eight genes are also clinal genes in North America (*Ets65A*, *Elk*, *sba*, *CG32066*), so we were able to compare the direction of change in allele frequencies between their candidate SNPs with available data from Fabian *et al.* (2012). In all four cases, the estimated frequency of the major allele increases consistently from RG to FR to NL, while in ZI it is about the same as in RG. The various functions of these genes suggest that adaptation to the more temperate local conditions in Europe has been a complex process involving many factors important for fitness, from direct tolerance to cold and oxidative stress, to developmental time, nutrient storage, locomotion, mating behavior, and even learning and memory. This may suggest the presence of epistatic fitness interactions.

Finally, we examined the GO and pathway enrichment and tested for significance of overlap with equivalent enrichment analyses that we performed using the candidate genes from North America (Fabian *et al.* 2012). GO analyses could increase the amount of information about adaptation in certain pathways that might be revealed by the joint effect of genes that individually might contribute only slightly to the trait. Thus, GO analyses might also complement our knowledge by indicating genes that did not pass our stringent significance threshold. Our strict criteria for ascertaining genes resulted in only 27 genes that correlated with altitude, so that we could not find enrichment of these genes. However, the results of overlaps with latitude were quite surprising. We found significant overlaps not only with all population pairs from North America, but even with only the genes with the steepest frequency change with latitude, termed “significantly clinal” genes by Fabian *et al.* (2012). To account for possible gene length bias, we performed GO enrichment using the software *Gowinda* with parameters corresponding to those used by Fabian *et al.* (2012), but allowing genes with multiple independent SNPs to be scored more than once for different GO terms. We recovered 72 significantly enriched terms for latitudinal SNP candidates (at $P_{FDR}<0.05$). 8 of these terms overlapped with terms enriched in our ClueGO analysis (Fisher’s exact test: $P<0.00021$).

Overall, the significance levels of overlaps of enriched terms were even more pronounced than those of genes. Moreover, even the clusters of enriched terms produced by the functional grouping in ClueGO showed overlap with the clusters we got from North American enrichment analyses. The largest clusters from terms enriched for our latitude candidate genes (appendage development, taxis, and generation of neurons) are in line

with selection possibly acting on the fly appendages (Bergmann's rule) and perhaps also influencing locomotion, behavior, memory, and learning. Taken together, the GO and pathway analyses showed that we could replicate the significance of gene overlaps with the candidates of Fabian *et al.* (2012), even when taking into account the broader functional roles of the candidate genes. Perhaps even more importantly, the particularly enriched terms and clusters make sense in the wider context of adaptation to colder European environments.

## 4.4 Evidence for adaptation on gene network level

### 4.4.1 Gene Ontology terms

We started by comparing the overall numbers of significantly enriched GO terms and pathways with the results of a classical GO / pathway enrichment performed in the first part of the results (Section 3.1). We were particularly interested in any overlaps, and in explaining the differences between the two approaches. We aimed to show that the approach used here is more appropriate for assessing adaptation signatures in general, and polygenic adaptation in particular. We do this by showing many examples of pathways enriched with overall higher median BF SNPs, but with an insufficient number of extreme outlier SNPs (and hence outlier genes), which makes them impossible to detect using the classical approach. Conversely, most GO terms and pathways enriched using the classical approach are ascertained mainly because of only a handful of outlier SNPs with extremely large BFs. Finally, we organize the enriched pathways into clusters and we look at literature that might explain the observed patterns in the context of local adaptation to European environments.

For latitude, the 47 top enriched GO terms (latitude $P_{empirical} < 0.0001$) contained a total of 212 unique genes between them, but only 22 of those genes were classified as candidate genes in our classical GO enrichment design (Section 3.1). The majority of the genes in significant gene sets contained no extreme outlier SNPs, as defined in Section 3.1 (defined with a cut-off of $ln(BF)>5$, or $BF>e^5$). In total, only eight GO terms were found significant in both analyses: tissue development (GO:0009888, latitude $P_{empirical} = 0.00002$), regulation of striated muscle tissue development (GO:0016202, latitude $P_{empirical} = 0.00018$), organ morphogenesis (GO:0009887, $P_{empirical} = 0.00040$), dendrite guidance (GO:0070983, latitude

$P_{empirical}$ = 0.00033), tube development (GO:0035295, latitude $P_{empirical}$ = 0.01566), cell morphogenesis involved in neuron differentiation (GO:0048667, latitude $P_{empirical}$ = 0.03276), smooth septate junction (GO:0005920, latitude $P_{empirical}$ = 0.04210), and cell projection (GO:0042995, latitude $P_{empirical}$ = 0.04739). In most cases however, candidate genes were too few within any of the GO terms to push them over the significance threshold of a classical GO enrichment.

The molecular biology of circadian rhythms and clock genes has been studied extensively (Blau & Young 1999; Martinek *et al.* 2001; Glossop *et al.* 2003; Sehgal 2004; Bae & Edery 2006; Allada & Chung 2010; Frenkel & Ceriani 2011; Vanin *et al.* 2012; McClung 2013; Dusik *et al.* 2014; Kunst *et al.* 2015; Flourakis *et al.* 2015). In *Drosophila melanogaster*, they are related to many traits that affect survival and reproduction: locomotion (Chiu *et al.* 2010; Lear *et al.* 2013; Vaccaro *et al.* 2016), courtship (Konopka *et al.* 1996; Dockendorff *et al.* 2002; De *et al.* 2013; Medina *et al.* 2015), neuronal plasticity (Petsakou *et al.* 2015), learning and behavior (Frenkel & Ceriani 2011), nitrogen homeostasis (Jeyaraj *et al.* 2012), apoptosis (Means *et al.* 2015), feeding behavior and immunity (Sarov-Blat *et al.* 2000; Stone *et al.* 2012; Allen *et al.* 2016), sleep (Liu *et al.* 2014; Kunst *et al.* 2015), sleep triggered by immune response (Kuo *et al.* 2010; Bollinger *et al.* 2010), as well as seasonal adaptation, preference, and entrainment of the clock in response to cold temperatures (Glaser & Stanewsky 2005; Chen *et al.* 2006; Kaneko *et al.* 2012; Lee & Montell 2013; Goda *et al.* 2014).

Circadian clocks have been shown to be adaptively important (Sandrelli *et al.* 2007; Tauber *et al.* 2007; Yerushalmi & Green 2009; Vaze & Sharma 2013), and most interestingly, clines have been found in clock genes that allow for fine tuning of the clock to local environmental conditions (Kyriacou *et al.* 2008). In light of all of these studies, our finding here that so many GO terms related to circadian rhythms are enriched in signals of polygenic adaptation, especially to latitude, is perhaps the most interesting finding in this study. Adaptation that happens to clock genes and circadian rhythms might easily lead to adaptive changes in chill-coma recovery time, startle response, and resistance to starvation stress. It is also in line with our adaptive network of genes from Section 3.1 (see Figure 3.2), affecting many of the traits those highly significant outliers are involved in, such as courtship, locomotion, learning and memory, temperature sensitivity, olfactory behavior, nutrition, and immunity.

Nephrocytes are essential for the disposal of nitrogen waste in *Drosophila* (Weavers *et al.* 2009; Cagan 2011), and their dysfunctions are related to excess of dietary sugars via the OGT-Polycomb-Knot-Sns pathway (Na *et al.* 2015). Also, Figures 3.1, 3.2, 3.3, and 3.4 show that *Nephrocyte filtration* is close in functional similarity to *Regulation of circadian sleep/wake cycle, sleep*. This is interesting in light of studies linking nitrogen homeostasis with circadian rhythms (Jeyaraj *et al.* 2012).

Given the ubiquitousness and importance of mitochondrial function, it is not surprising that changes in their function and organization affect starvation, locomotion, heat stress, and other traits that we have shown in Section 3.1 to be locally adapted in *Drosophila melanogaster*. For example, starved larvae have been shown to have lower mitochondrial activity in the fat body, and reduced expression of oxidative phosphorylation and glutamine metabolism genes (Baltzer *et al.* 2009). Mitochondrial defects are well known to be related to neuromuscular degeneration (López Del Amo *et al.* 2015), and the inner mitochondrial membrane contains heat shock proteins that are upregulated under stress in the whole embryo (Baena-López *et al.* 2008).

Figures 3.5 through 3.8 show that these two GO terms are functionally similar, which is not surprising given they are both developmental terms. Genes such as *eve*, *lbe*, and *slou* control the size of individual muscles by regulating the number of fusion events via actin dynamics and cell adhesion (Bataillé *et al.* 2010). Development of muscles is obviously related to locomotion, but other traits might also be related, for example via correlated responses to selection (Partridge & Fowler 1992; Partridge *et al.* 1999). We know, for example, that fitness trade-offs exist between speed of development and adult weight (Nunney 1996), and that body size in turn is related to temperature (Partridge *et al.* 1994).

### 4.4.2 Reactome pathways

The role of Notch signaling in the regulation of development is well known (Struhl & Adachi 1998). It is evolutionarily highly conserved (Kidd *et al.* 2015). Interestingly, more recent work has shown that Notch signaling also plays a role in energy metabolism through the control of the tricarboxylic acid cycle and glycolysis (Slaninova *et al.* 2016), that it is involved in late-stage skeletal myogenesis (Bi *et al.* 2016), and that it is required in adult flies for the

plasticity of the olfactory receptor neurons (Kidd *et al.* 2015). It is conceivable that these processes might all play a role in adaptation to more temperate environments.

Plasmalogens are ether phospholipids with a complicated evolutionary history, found in taxa from bacteria to mammals (Goldfine 2010). They are one of the most abundant lipids in the *Drosophila* head (Chintapalli *et al.* 2013), particularly the brain (Carvalho *et al.* 2012), and more abundant in males than females, even though males are smaller and have less lipid content overall (Carvalho *et al.* 2012). This might be interesting in the context of a possible adaptation in mating behavior. *Courtship behavior* was an enriched GO term in our SNP-set-enrichment analysis here ($P_{\text{latitude}}$ = 0.0042). Moreover, courtship was one of the functions associated with the cold adaptive network we characterized in Section 3.1 (see Figures 3.1 and 3.2). Interestingly, plasmalogens have been found to be abundant in the testes of vertebrates (Reisse *et al.* 2001), though it is not known whether this might also be the case in *Drosophila* (Carvalho *et al.* 2012).

SMAD proteins modulate transcription by taking signals from transforming growth factor beta ligands and activating downstream genes related to activities such as proliferation and differentiation (Sekelsky *et al.* 1995). They are important for the proper functioning of the neuromuscular junction (Higashi-Kovtun *et al.* 2010) and therefore probably also for proper locomotion. We have previously shown (see Tables S1-S3) that many genes related to locomotion show signals of selection. Moreover, even though it only has 3 annotated genes, *Phase 1 - inactivation of fast* $Na^+$ *channels* (biosystem 1328858) also showed strong signals of polygenic adaptation over its 339 SNPs. This is in line with our finding in Section 3.1 (see Figure 3.2) that the gene *Elk*, which is involved in locomotion impairments and the function of ion channels, is one of the most significant clinal genes.

More generally, the fact that we find so many gene expression pathways with signals of polygenic adaptation for latitude, and that *Gene Expression* (biosystem 1329531) was enriched overall (latitude $P_{\text{empirical}}$ = 0.0106, altitude $P_{\text{empirical}}$= 0.0454, coldest month temperature $P_{\text{empirical}}$ = 0.0083, coldest and warmest yearly temperatures $P_{\text{empirical}}$ = 0.0136 and 0.0407), suggests that adaptation by means of changes in gene expression may have played an important role in European local adaptation.

Nucleotide sugars are involved in development (Liu *et al.* 2010), and dietary sugars in general have been shown to affect cold tolerance in *Drosophila* by inducing system-wide

metabolic alteration (Colinet *et al.* 2012). Furthermore, three pathways related to developmental biology (Figure 3.15) and 8 pathways related to the immune system (Figure 3.14) were enriched in signals of natural selection. *IRAK1* (interleukin-1 receptor-associated kinase 1) plays an important role in the *Toll* pathway (Ferrandon *et al.* 2007). The immune system of *Drosophila* has been investigated extensively (Tzou *et al.* 2002; Hoffmann 2003), particularly for signatures of selection (Hoffmann & Reichhart 2002; Schlenke & Begun 2003; Lazzaro 2008; Obbard *et al.* 2009). More recently, a possibility of balancing selection on immune genes has been getting increasing attention (Unckless *et al.* 2016; Croze *et al.* 2016). Immunity is also closely related to circadian-regulated behavior (Allen *et al.* 2016), sleep (Bollinger *et al.* 2010; Lenz *et al.* 2015), metabolism (Clark *et al.* 2013), nutrition (Unckless *et al.* 2015), gut structure (Broderick *et al.* 2014), and even courtship (Immonen & Ritchie 2012). Overall, it is clear that immunity and its associated biochemical pathways have an important effect on fitness. In Section 3.1, we found that the gene *Elk*, which is related to dietary nutrition effect on the immune response (Unckless *et al.* 2015), was one of the most significant genes in our analysis (see also the above section discussing the *Gene Expression* cluster). In addition to adaptive signals on individual genes, the evidence from this study also suggests a strong role of small shifts in allele frequencies over many genes, i.e. polygenic adaptation.

Moreover, 2 hemostasis-related and 4 neuronal system-related Reactome pathways were significantly enriched (Figure 3.16). Both hemostasis pathways were related to platelet homeostasis. Platelets play an important role in healing wounds and immunity. Seen in the context of strong evidence of adaptation for many immune system-related pathways, it is less surprising that we also found hemostasis- related pathways with many SNPs correlating strongly with many environmental variables. For instance, *Elevation of cytosolic Ca2+ levels* (biosystem 1328678) shows high correlation with all six environmental variables ($P_{\text{empirical}}$ between 0.0017 and 0.016). Additionally, there is evidence that calcium is related to temperature sensitivity (Chuang *et al.* 2004; Wu *et al.* 2005).

As for the *Neuronal System* cluster, *Dopamine Neurotransmitter Release Cycle* had a strong signal of adaptation (latitude $P_{\text{empirical}}$ = 0.0015). Dopamine is important for locomotory behaviors in *Drosophila* (Meehan & Wilson 1987; Carbone *et al.* 2006; Jordan *et al.* 2006), and in the light of our findings from Section 3.1, dopamine-related pathways are

expected to have a strong effect on fitness. Thus a possibility of polygenic adaptation is in line with already presented evidence.

Finally, we found several enriched Reactome pathways that were mostly related to housekeeping functions, e.g. DNA replication and repair, cell cycle, and programmed cell death. It is hard to tell in what way the genes related to these functions might be adapting to local conditions. Interestingly, it has been shown, for instance, that apoptosis is related to circadian rhythms (Means *et al.* 2015), and that the efficiency of DNA repair depends on temperature (Lupu *et al.* 2004).

# CHAPTER 5: CONCLUSION

In the first part of this work, we have detected footprints of polygenic adaptation in *Drosophila melanogaster* to temperature-related traits. Our results suggest that these traits may have responded to selection in a concerted manner, most likely as a result of complex fitness tradeoffs. Additionally, we found that SNPs under strong environmental selection support genes that significantly overlap with clinal candidates from other continents. The overlaps are even more significant if assessed from common biological pathways and gene ontology terms enriched with candidate genes between studies. Lastly, we proposed a network of genes with the strongest evidence of selection, which suggests that adaptation to new environments in Europe involved a strong direct response to cold, but also changes in development, mating behavior, oxidative stress, locomotion, reproductive diapause, learning, and memory. Functional studies of these genes in the context of cold tolerance are needed to confirm these findings. Also, future studies of local adaptation to cold should take into account the intricacies of different selective pressures that may be operating on many genes across the genome simultaneously.

In the second part, we have shown evidence for polygenic local adaptation in gene sets defined as gene ontology terms and Reactome pathways. Perhaps the most interesting result was the enrichment of many GO terms that are functionally related to circadian rhythms. None of the circadian rhythm GO terms were found in our classical GO analysis (Section 3.1), which is less surprising given the evidence that clock genes might be evolving by small changes to fit local conditions (Kyriacou *et al.* 2008). Clearly, taking into account the little fitness contributions from many loci to investigating adaptation of traits such as circadian rhythms has advantages over approaches concentrating exclusively on candidate or outlier genes. Our results presented here show that studying polygenic local adaptation of traits as complex as cold tolerance is not straightforward, and highlight the need of future studies to take a multi-pronged approach in search of evidence of adaptation.

# APPENDIX

# SUPPLEMENTARY FIGURES AND TABLES

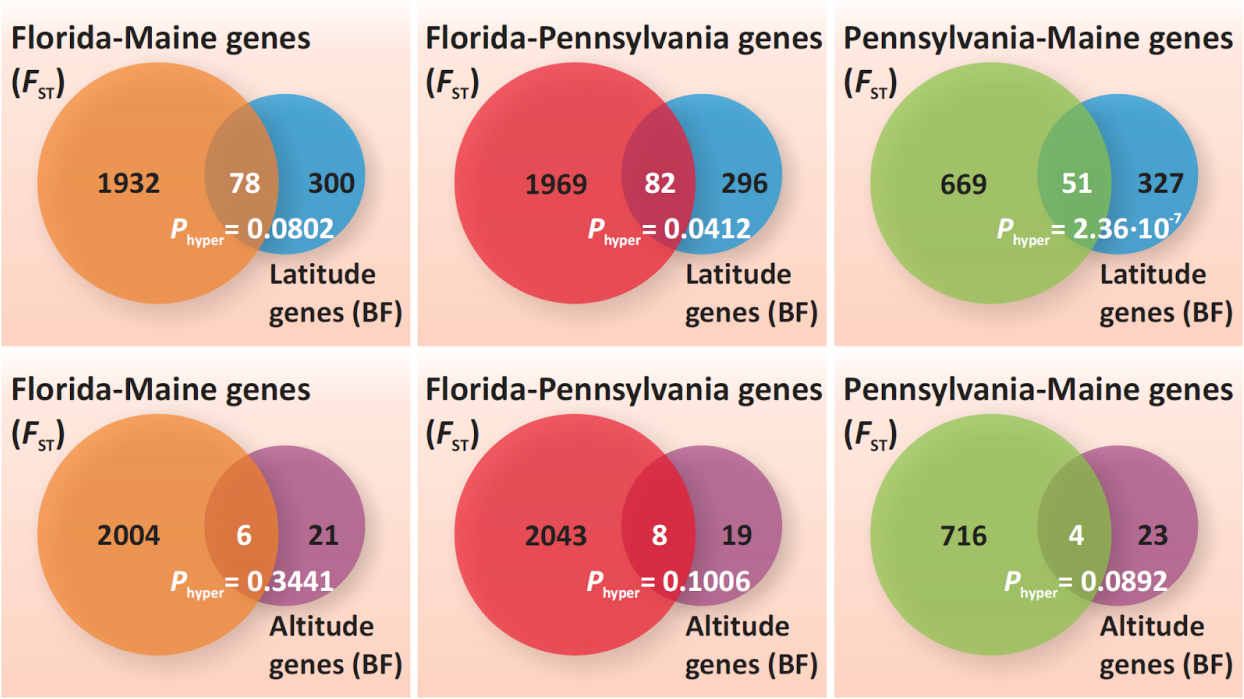



**Figure S1** Proportions of genes supported by SNPs with strong evidence (*ln*(BF)>5 or $P < 0.0029$) for correlation with latitude (top panel) and altitude (bottom panel) that overlap with candidate genes from North America as quantified by $F_{ST(\text{Fabian } et\ al.\ 2012)}$. The significances of overlaps were assessed using Fisher's exact tests and are shown under each overlap.

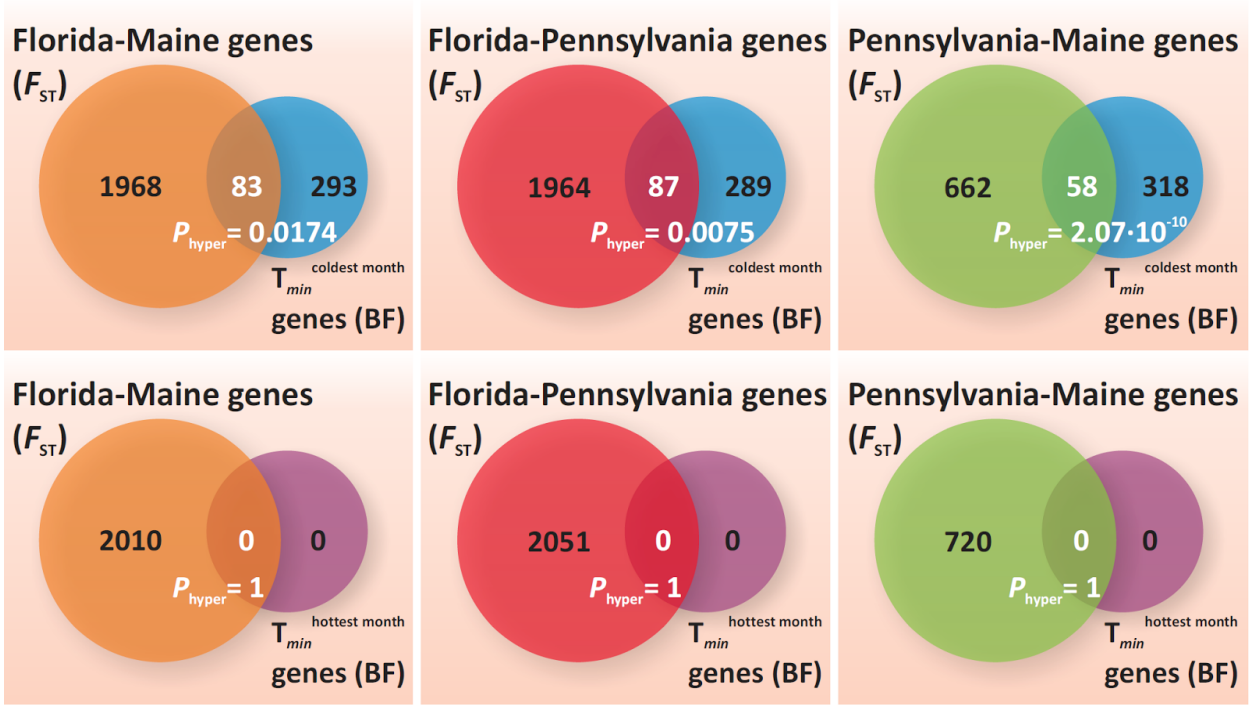



**Figure S2** Proportions of genes supported by SNPs with strong evidence (*ln*(BF)>5 or $P < 0.0029$) for correlation with coldest month minimum temperature (top panel) and hottest month minimum temperature (bottom panel) that overlap with candidate genes from North America as quantified by $F_{ST}$ (Fabian *et al.* 2012). The significances of overlaps were assessed using Fisher's exact tests and are shown under each overlap.

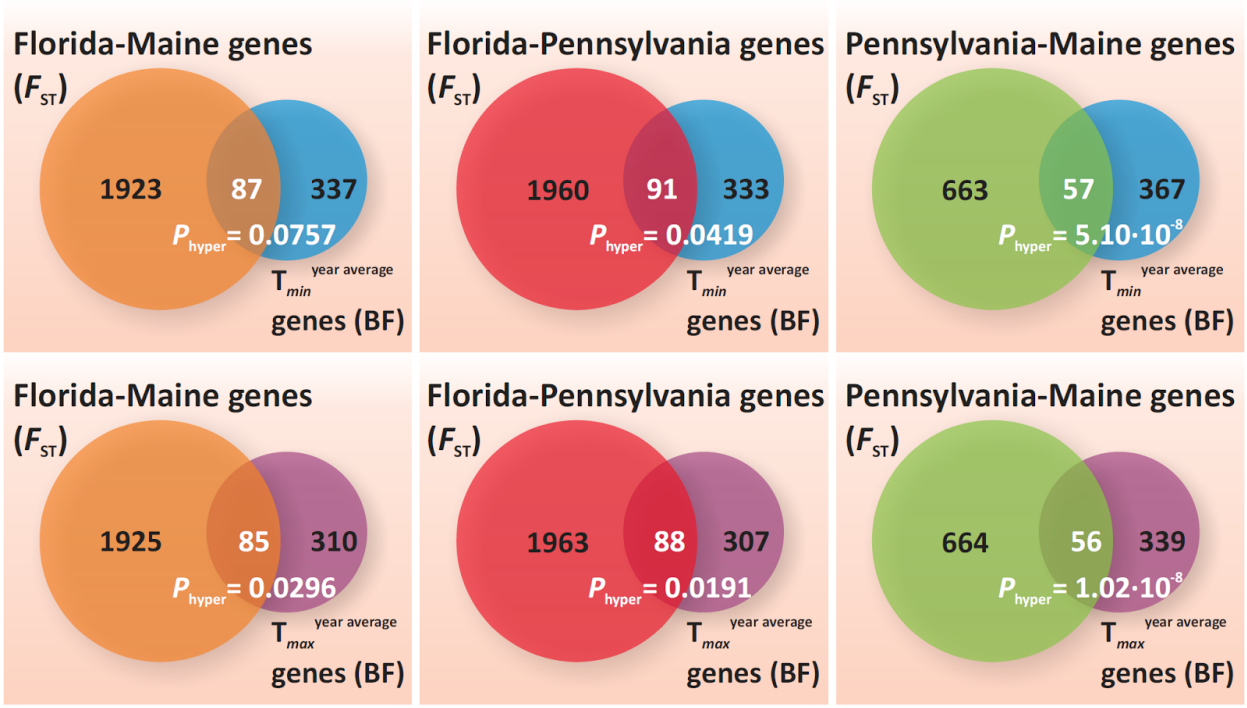



**Figure S3** Proportions of genes supported by SNPs with strong evidence (*ln*(BF)>5 or $P < 0.0029$) for correlation with yearly minimum temperatures (top panel) and yearly maximum temperatures (bottom panel) that overlap with candidate genes from North America as quantified by $F_{ST}$ (Fabian *et al.* 2012). The significances of overlaps were assessed using Fisher's exact tests and are shown under each overlap.


**Figure S4** Reactome pathways related to DNA repair significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

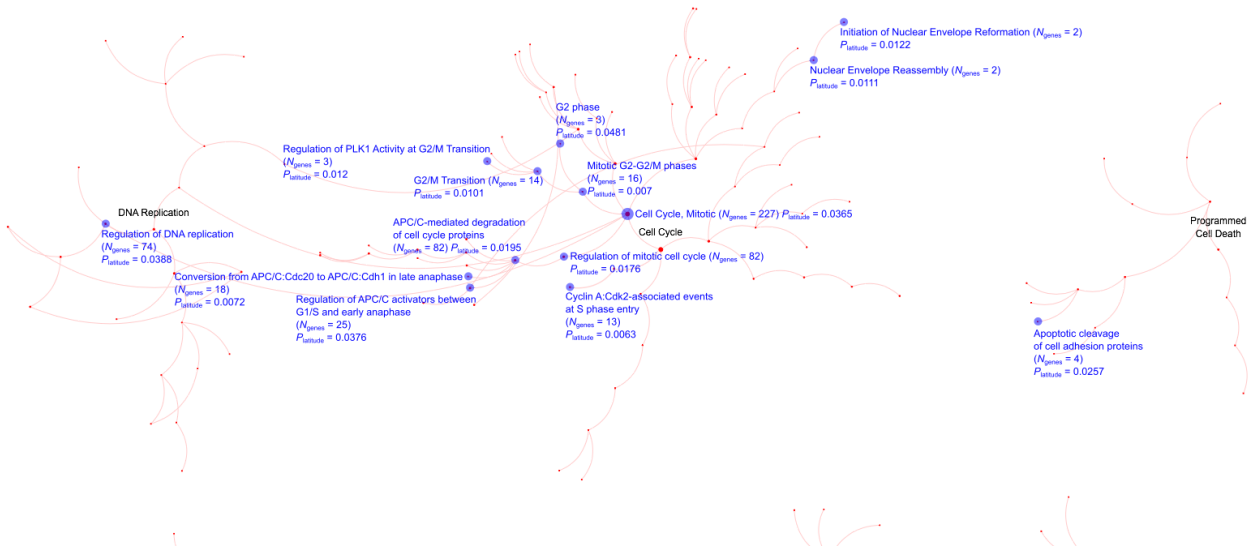



**Figure S5** Reactome pathways related to Cell Cycle, DNA Replication, and Programmed Cell Death significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

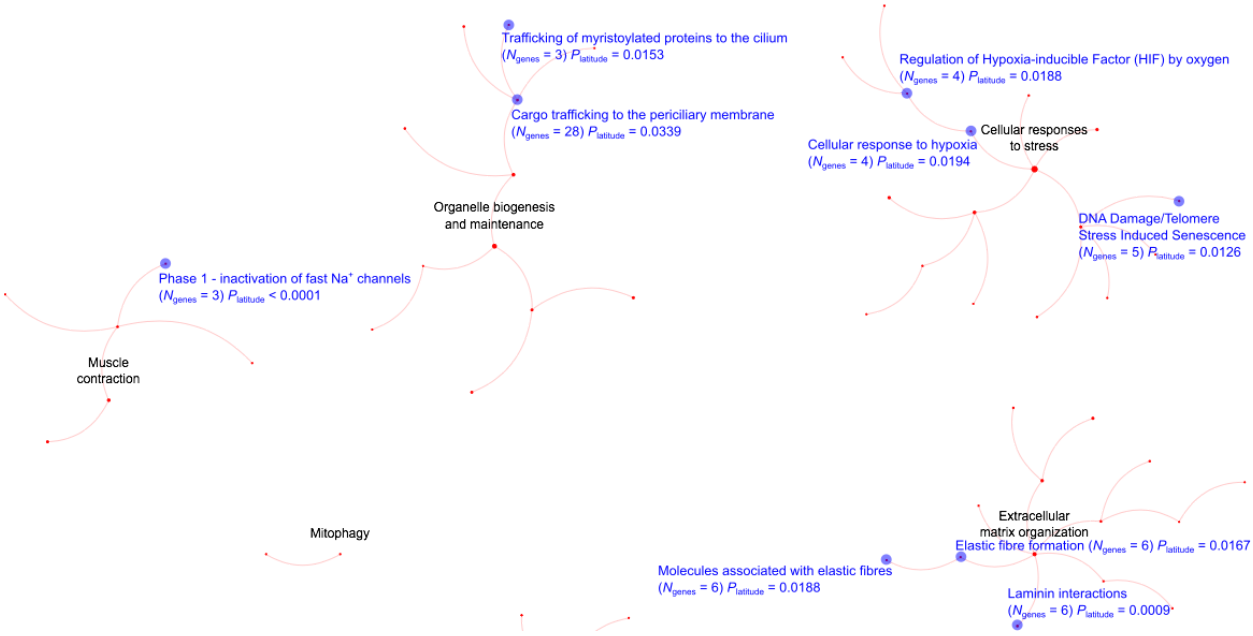



**Figure S6** Reactome pathways related to Organelle biogenesis and maintenance, Cellular responses to stress, and Extracellular matrix organization significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

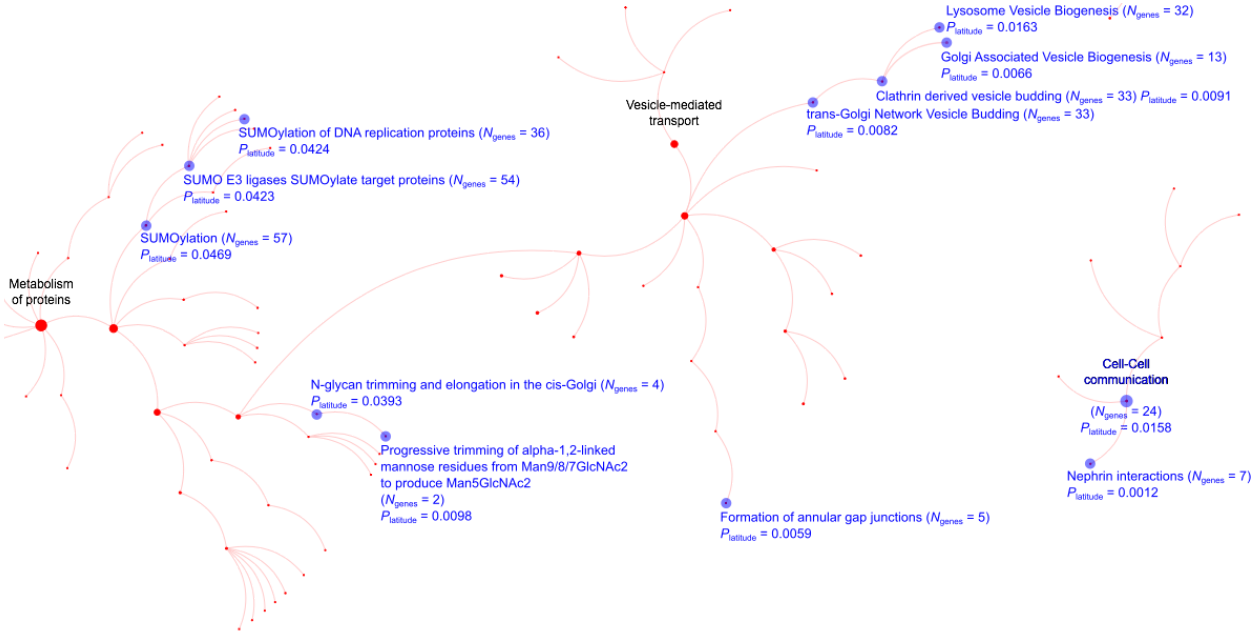



**Figure S7** Reactome pathways related to Metabolism of proteins, Vesicle-mediated transport, and Cell-Cell communication significantly enriched in signals of polygenic adaptation for latitude ($P_{empirical} < 0.05$).

**Table S1** Genes previously characterized for tolerance to cold or heat, disturbance, or starvation stress.

| gene ID | references | ascertained by | Flybase 5 coordinates |
|---|---|---|---|
| **genes for tolerance to cold or heat** | | | |
| *brinker/brk* | (Wilches et al. 2014) | P-element insertion, gene expression, QTL mapping | X:7,201,972..7,205,165 |
| *Frost/Fst* | (Sinclair et al. 2007; Colinet et al. 2010; Bing et al. 2012) | functional study; gene transcription; gene expression | 3R:5,470,700..5,471,876 |
| *Hsp70Aa* | (Overgaard et al. 2005; Sinclair et al. 2007; Colinet et al. 2010) | physiological study; gene transcription; expression | 3R:7,779,885..7,782,266 |
| *Hsp22* | (Colinet et al. 2010; Colinet et al. 2013; Vermeulen et al. 2013) | expression knockdown; proteomics; transcriptomics | 3L:9,366,031..9,368,064 |
| *Hsp23* | (Colinet et al. 2010; Telonis-Scott et al. 2013) | expression knockdown; exon expression analysis | 3L:9,374,982..9,375,865 |
| *Hsp26* | (Qin et al. 2005; Telonis-Scott et al. 2013) | transcript abundance after cold shock; exon expression | 3L:9,369,518..9,370,527 |
| *Hsp27* | (Carreira et al. 2013; Telonis-Scott et al. 2013) | exon expression; rearing temperature, P-element inserts | 3L:9,377,163..9,378,794 |
| *DnaJ-1* | (Telonis-Scott et al. 2013; Colinet et al. 2013) | exon expression; induced expression of stress genes | 3L:5,743,129..5,745,289 |
| *Hsp68* | (Telonis-Scott et al. 2013; Colinet et al. 2013) | exon expression; induced expression of stress genes | 3R:19,880,802..19,883,032 |
| *Hsp83/Hsp90* | (Telonis-Scott et al. 2013; Colinet et al. 2013; Goda et al. 2014) | exon expression; induced expression of stress genes | 3L:3,192,969..3,197,059 |
| *AnnIX/Anxb9* | (Telonis-Scott et al. 2009; Vermeulen et al. 2013) | gene expression after cold shock; transcriptomics | 3R:16,890,963..16,896,639 |
| *CG18140* | (Telonis-Scott et al. 2009) | gene expression after cold shock | 2L:22,892,306..22,918,647 |
| *CG31738* | (Telonis-Scott et al. 2009) | gene expression after cold shock | 2L:16,545,016..16,663,218 |
| *CG18180* | (Harbison et al. 2005; Telonis-Scott et al. 2009; Vermeulen et al. 2013) | QTL; gene expression after cold shock; transcriptomics | 3L:9,634,113..9,635,013 |
| *CG15353* | (Harbison et al. 2005; Telonis-Scott et al. 2009) | QTL; gene expression after cold shock | 2L:2,006,763..2,007,193 |
| *Jon99Ci* | (Harbison et al. 2005; Telonis-Scott et al. 2009; Vermeulen et al. 2013) | QTL; gene expression after cold shock; transcriptomics | 3R:25,750,948..25,751,911 |
| *Lsp1beta* | (Harbison et al. 2005; Telonis-Scott et al. 2009) | QTL; gene expression after cold shock | 2L:898,500..901,316 |
| *Jon25Bii* | (Harbison et al. 2005; Telonis-Scott et al. 2009; Vermeulen et al. 2013) | QTL; gene expression after cold shock; transcriptomics | 2L:4,952,243..4,953,136 |
| *CG16700* | (Ayroles et al. 2009; Svetec et al. 2011) | QTL, systems genetics; selective sweep mapping | X:16,985,887..16,992,521 |
| *smp-30* | (Qin et al. 2005; Clowers et al. 2010; Vermeulen et al. 2013) | transcript abundance after cold shock; transcriptomics | 3R:10,571,623..10,576,968 |
| *Hsromega* | (Collinge et al. 2008; Cockerell et al. 2014) | heat knockdown, protein synthesis, latitudinal variation | 3R:17,121,849..17,143,558 |
| *CG8791* | (Fallis 2012; Vermeulen et al. 2013) | mutant complementation tests; transcriptomics | 2R:3,814,158..3,817,427 |
| *psq* | (Winbush et al. 2012; Vermeulen et al. 2013) | gene expression; transcriptomics | 2R:6,445,393..6,504,785 |
| *stan* | (Toshima et al. 2014; Organisti et al. 2015) | GWAS; rescue, domain deletion assays | 2R:6,560,850..6,608,643 |
| *Lsm10* | (Fallis 2012; Vermeulen et al. 2013) | mutant complementation tests; transcriptomics | 2R:6,708,603..6,709,145 |
| *Taf5* | (Fallis 2012; Xie et al. 2014) | mutant complementation tests; RNAi and knockdown | 2R:6,764,817..6,767,178 |
| *CG30016* | (Fallis 2012; Vermeulen et al. 2013) | mutant complementation tests; transcriptomics | 2R:6,762,397..6,762,897 |
| *Pex6* | (Vermeulen et al. 2013) | transcriptomic analysis | 2R:6,767,479..6,770,707 |
| *stv* | (Telonis-Scott et al. 2013; Colinet et al. 2013; Vermeulen et al. 2013) | exon expression; induced expression; transcriptomics | 3L:13,470,641..13,476,615 |
| **genes affecting response to disturbance** | | | |
| *CG12943* | (Romero-Calderón et al. 2007; Vermeulen et al. 2013) | P-element insertion / deletion; transcriptomics | 2R:6,791,075..6,793,068 |
| *CG30379* | (Lee et al. 2009) | physiological study | 2R:3,827,461..3,829,375 |
| *clumsy* | (Lee et al. 2009) | physiological study | 2L:21,206,493..21,211,818 |
| *lola* | (Yamamoto et al. 2008; Gates et al. 2011; Fedotov et al. 2014) | transposon insertions; microarray; P-element / RNAi | 2R:6,369,399..6,430,796 |
| *Oatp30B* | (Meyer et al. 2014) | mutagenesis, immunostaining, behavioral assays | 2L:9,521,210..9,540,060 |
| *CG31619* | (Meyer et al. 2014) | mutagenesis, immunostaining, behavioral assays | 2L:21,684,106..21,729,051 |
| *E(spl)m7-HLH* | (Yamamoto et al. 2008) | transposon insertions, mutagenesis | 3R:26,037,038..26,038,019 |
| **genes affecting resistance to starvation stress** | | | |
| *cpo* | (Schmidt et al. 2008; Fabian et al. 2012; Cogni et al. 2014) | QTL mapping; latitudinal cline; diapause cline | 3R:13,745,554..13,844,614 |
| *whd* | (Wang et al. 2011; Vermeulen et al. 2013; Gingras et al. 2014) | mutagenesis; transcriptomics; knockdown | 2R:6,356,978..6,366,072 |
| *nclb* | (Casper et al. 2011; Toshima et al. 2014) | mutagenesis; GWAS | 2R:6,763,253..6,764,801 |
| *CG12054* | (Alic et al. 2014; Chen et al. 2015) | P-element insertion, microarray/RNA expression, ChIP | 3R:31,220,702..31,231,259 |
| *Dyrk2* | (Reis et al. 2010; Chen et al. 2015) | mutagenesis; induced expression | 2L:14,184,478..14,234,126 |
| *hdc* | (Reis et al. 2010; Chen et al. 2015) | mutagenesis; induced expression | 3R:30,277,932..30,372,382 |
| *shep* | (Reis et al. 2010; Chen et al. 2015) | mutagenesis; induced expression | 3L:5,155,821..5,277,944 |
| *chas* | (Reis et al. 2010; Chen et al. 2015) | mutagenesis; induced expression | X:17,672,458..17,698,730 |
| *Ire1* | (Pile et al. 2003; Vermeulen et al. 2013; Chen et al. 2015) | microarrays, mutagenesis; transcriptomics; expression | 3R:19,853,908..19,861,291 |
| *Octbeta3R* | (Zhang et al. 2013) | induced expression, targeted cell knockout | 3R:12,511,570..12,548,229 |

**Table S2** Names and descriptions of literature genes from Table S1. Retrieved from Flybase v6 (flybase.org).

| gene ID | Flybase ID | gene name | gene function |
|---|---|---|---|
| | | **genes affecting cold tolerance** | |
| *brk* | FBgn0024250 | brinker | a potential transcription factor that negatively regulates *decapentaplegic* target genes |
| *Fst* | FBgn0037724 | Frost | upregulated in response to cold exposure |
| *Hsp70Aa* | FBgn0013275 | Heat-shock-protein-70Aa | chaperone required for border cells migration |
| *Hsp22* | FBgn0001223 | Heat shock protein 22 | - |
| *Hsp23* | FBgn0001224 | Heat shock protein 23 | - |
| *Hsp26* | FBgn0001225 | Heat shock protein 26 | - |
| *Hsp27* | FBgn0001226 | Heat shock protein 27 | - |
| *DnaJ-1* | FBgn0263106 | DnaJ-like-1 | - |
| *Hsp68* | FBgn0001230 | Heat shock protein 68 | - |
| *Hsp83* | FBgn0001233 | Heat shock protein 83 | Molecular chaperone that promotes the maturation, structural maintenance and proper regulation of specific target proteins involved for instance in cell cycle control and signal transduction. Undergoes a functional cycle that is linked to its ATPase activity. This cycle probably induces conformational changes in the client proteins, thereby causing their activation. Interacts dynamically with various co-chaperones that modulate its substrate recognition, ATPase cycle and chaperone function. Together with Hop and piwi, mediates canalization, also known as developmental robustness, likely via epigenetic silencing of existing genetic variants and suppression of transposon-induced new genetic variation. Required for piRNA biogenesis by facilitating loading of piRNAs into PIWI proteins. {ECO:0000269\|PubMed:21186352, ECO:0000269\|PubMed:22902557}. |
| *AnxB9* | FBgn0000083 | Annexin B9 | - |
| *CG18140* | FBgn0250907 | Chitinase 3 | - |
| *CG31738* | FBgn0259735 | - | - |
| *CG18180* | FBgn0036024 | - | - |
| *CG15353* | FBgn0040718 | - | - |
| *CG15353* | FBgn0040718 | - | - |
| *Jon99Ci* | FBgn0003358 | Jonah 99Ci | Its major function may be to aid in digestion. {ECO:0000269\|PubMed:2469005}. |
| *Lsp1beta* | FBgn0002563 | Larval serum protein 1 beta | Larval storage protein (LSP) which may serve as a store of amino acids for synthesis of adult proteins. {ECO:0000250}. |
| *Jon25Bii* | FBgn0031654 | Jonah 25Bii | - |
| *CG16700* | FBgn0030816 | - | - |
| *smp-30* | FBgn0038257 | Senescence marker protein-30 | - |
| *Hsromega* | FBgn0001234 | Heat shock RNA omega | - |
| *CG8791* | FBgn0033234 | Major Facilitator Superfamily Transporter 12 | - |
| *psq* | FBgn0263102 | pipsqueak | - |
| *stan* | FBgn0024836 | starry night | Involved in the fz signaling pathway that controls wing tissue polarity. Also mediates homophilic cell adhesion. May play a role in initiating prehair morphogenesis. May play a critical role in tissue polarity and in formation of normal dendrite fields. {ECO:0000269\|PubMed:10490098, ECO:0000269\|PubMed:10556066}. |
| *Lsm10* | FBgn0033554 | Lsm10 | - |
| *Taf5* | FBgn0010356 | TBP-associated factor 5 | TFIID is a multimeric protein complex that plays a central role in mediating promoter responses to various activators and repressors. May play a role in helping to anchor Taf4 within the TFIID complex. May be involved in transducing signals from various transcriptional regulators to the RNA polymerase II transcription machinery. {ECO:0000269\|PubMed:8247000}. |
| *CG30016* | FBgn0050016 | - | - |
| *Pex6* | FBgn0033564 | Peroxin 6 | - |

| | | | |
|---|---|---|---|
| *stv* | FBgn0086708 | starvin | probably acts as a co-chaperone modulating the activity of Hsp70 chaperone machinery during recovery from cold stress |
| | | **genes affecting response to disturbance** | |
| *CG12943* | FBgn0033572 | polyphemus | - |
| *CG30379* | FBgn0050379 | - | - |
| *clumsy* | FBgn0026255 | clumsy | - |
| *lola* | FBgn0005630 | longitudinals lacking | Putative transcription factor required for axon growth and guidance in the central and peripheral nervous systems. Repels CNS axons away from the midline by promoting the expression of the midline repellent sli and its receptor robo. {ECO:0000269\|PubMed:11880341, ECO:0000269\|PubMed:8050351}. |
| *Oatp30B* | FBgn0032123 | Organic anion transporting polypeptide 30B | - |
| *CG31619* | FBgn0051619 | nolo / no long nerve cord | - |
| *E(spl)m7-HLH* | FBgn0002633 | Enhancer of split m7, helix-loop-helix | Participates in the control of cell fate choice by uncommitted neuroectodermal cells in the embryo. Transcriptional repressor. Binds DNA on N-box motifs: 5'-CACNAG-3'. |
| | | **genes affecting resistance to starvation stress** | |
| *cpo* | FBgn0263995 | couch potato | May play a role in the development or function of the peripheral nervous system by regulating the processing of nervous system-specific transcripts. {ECO:0000269\|PubMed:1427076}. |
| *whd* | FBgn0261862 | withered | - |
| *nclb* | FBgn0263510 | no child left behind | - |
| *CG12054* | FBgn0039831 | - | - |
| *Dyrk2* | FBgn0016930 | Dual-specificity tyrosine phosphorylation-regulated kinase 2 | In vitro; can phosphorylate exogenous substrates on Ser and Thr residues. May have a physiological role in development being involved in cellular growth and differentiation. {ECO:0000269\|PubMed:12786602}. |
| *hdc* | FBgn0010113 | headcase | Required for imaginal cell differentiation, may be involved in hormonal responsiveness during metamorphosis. Involved in an inhibitory signaling mechanism to determine the number of cells that will form unicellular sprouts in the trachea. Regulated by transcription factor esg. The longer hdc protein is completely functional and the shorter protein carries some function. {ECO:0000269\|PubMed:8575315, ECO:0000269\|PubMed:9531534}. |
| *shep* | FBgn0052423 | alan shepard | Has a role in the perception of gravity. {ECO:0000269\|PubMed:16594976}. |
| *chas* | FBgn0263258 | chascon | - |
| *Ire1* | FBgn0261984 | Inositol-requiring enzyme-1 | - |
| *Octbeta3R* | FBgn0250910 | Octopamine beta3 receptor | Receptor for octopamine. Octopamine (OA) is a neurotransmitter, neurohormone, and neuromodulator in invertebrates. The activity of this receptor is mediated by G proteins which activate adenylyl cyclase (By similarity). {ECO:0000250\|UniProtKB:Q9VCZ3}. |

**Table S3** $F_{ST}$-values ($P < 0.05$; before slash) and BF values (*ln*(BF)>1 behind slash) significant in at least one population pair or environmental variable respectively, for genes previously known to be involved for temperature tolerance as listed in Table S1 and Table S2.

| gene ID | 3'UTR variants | 5'UTR variants | Intron variants | Synonymous variants that **decrease** expression of major allele in Europe by codon usage bias | Synonymous variants that **increase** expression of major allele in Europe by codon usage bias | Nonsynonymous variants with change in side-chain charge | Nonsynonymous variants without change in side-chain charge |
|---|---|---|---|---|---|---|---|
| *brk* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Fst* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp70Aa* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp22* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp23* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp26* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp27* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *DnaJ-1* | 2 / 1 | - / - | - / - | - / - | 2 / - | - / - | - / - |
| *Hsp68* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp83* | - / - | - / - | 1 / - | - / - | - / - | - / - | - / - |
| *AnxB9* | - / - | 1 / - | 6 / - | 1 / - | - / - | - / - | - / - |
| *CG18140* | - / - | 1 / - | 1 / - | - / - | - / - | - / - | - / - |
| *CG31738* | - / - | 4 / - | 118 / 2 | 5 / - | - / - | - / - | 4 / - |
| *CG18180* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG15353* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Jon99Ci* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Lsp1beta* | - / - | - / - | 1 / - | 2 / - | - / - | - / - | - / - |
| *Jon25Bii* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG16700* | 3 / - | 1 / - | 10 / - | 4 / - | 3 / - | - / - | - / - |
| *smp-30* | - / - | - / - | - / - | - / - | 1 / - | - / - | - / - |
| *Hsromega* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG8791* | - / - | 2 / - | 2 / - | - / - | - / - | - / - | - / - |
| *psq* | 1 / - | 2 / - | 44 / - | 2 / - | 1 / - | - / - | - / - |
| *stan* | 1 / - | 2 / - | 24 / 22 | 1 / - | 2 / - | - / - | - / - |
| *Lsm10* | - / - | - / - | - / - | 1 / - | - / - | - / - | - / - |
| *Taf5* | 2 / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG30016* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| *Pex6* | - / - | - / - | - / - | 1 / - | - / - | - / - | - / - |
| *stv* | - / - | 3 / - | 6 / - | - / - | - / - | - / - | - / - |
| | | | | | | | |
| *CG12943* | - / - | - / - | - / - | - / - | - / - | - / - | 1 / - |
| *CG30379* | - / - | - / - | 1 / - | - / - | - / - | 1 / - | 2 / - |
| *clumsy* | - / - | - / - | - / - | 1 / - | - / - | - / - | - / - |
| *lola* | 8 / - | 3 / - | 52 / 1 | 7 / - | 4 / - | 2 / - | 2 / - |
| *Oatp30B* | 1 / - | - / - | 8 / - | 1 / - | - / - | - / - | - / - |
| *CG31619* | - / - | - / - | 7 | - / - | - / - | - / - | - / - |
| *E(spl)m7-HLH* | - / - | - / - | - / - | - / - | - / - | - / - | - / - |
| | | | | | | | |
| *cpo* | 3 / - | 1 / - | 103 / 7 | 1 / - | 1 / - | - / - | - / - |
| *whd* | - / - | 1 / - | 3 / - | 2 / - | - / - | - / - | - / - |
| *nclb* | - / - | - / - | - / - | - / - | - / - | - / - | 1 / - |
| *CG12054* | 1 / - | 1 / - | 3 / - | - / - | - / - | - / - | - / - |
| *Dyrk2* | 1 / - | 1 / - | 24 / 1 | 1 / - | - / - | - / - | - / - |
| *hdc* | 3 / - | - / - | 68 / 5 | 1 / - | - / - | - / - | - / - |
| *shep* | 1 / - | 3 / - | 109 / 5 | - / - | - / - | - / - | - / - |
| *chas* | 2 / - | 3 / - | 19 / - | 2 / - | 1 / - | - / - | 1 / - |
| *Ire1* | - / - | - / - | 3 / - | - / - | - / - | - / - | - / - |
| *Octbeta3R* | - / - | 1 / - | 29 / 3 | 1 / - | - / - | - / - | - / - |

**Table S4** BF values (*ln*(BF)>1 or $P < 0.0063$; before slash; *ln*(BF)>3 or $P < 0.0043$; behind slash) for genes from literature and various environmental variables (env1 = latitude, env2 = altitude, env3 = coldest month minimum, env4 = hottest month minimum, env5 = yearly minimum, env6 = yearly maximum).

| gene ID | chromosome | start position | end position | env1 | env2 | env3 | env4 | env5 | env6 |
|---|---|---|---|---|---|---|---|---|---|
| *brk* | chrX | 7201972 | 7205165 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Fst* | chr3R | 5470700 | 5471876 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp70Aa* | chr3R | 7779885 | 7782266 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp22* | chr3L | 9366031 | 9368064 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp23* | chr3L | 9374982 | 9375865 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp26* | chr3L | 9369518 | 9370527 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp27* | chr3L | 9377163 | 9378794 | - / - | - / - | - / - | - / - | - / - | - / - |
| *DnaJ-1* | chr3L | 5743129 | 5745289 | 2 / 2 | 2 / 1 | 2 / 2 | 2 / - | 2 / 2 | 2 / 2 |
| *Hsp68* | chr3R | 19880802 | 19883032 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsp83* | chr3L | 3192969 | 3197059 | - / - | - / - | - / - | - / - | - / - | - / - |
| *AnxB9* | chr3R | 16890963 | 16896639 | 1 / - | - / - | 1 / - | - / - | 1 / - | 1 / - |
| *CG18180* | chr3L | 9634113 | 9635013 | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG15353* | chr2L | 2006763 | 2007193 | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG15353* | chr2L | 2006763 | 2007193 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Jon99Ci* | chr3R | 25750948 | 25751911 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Lsp1beta* | chr2L | 898500 | 901316 | 1 / - | - / - | - / - | - / - | 1 / - | 1 / - |
| *Jon25Bii* | chr2L | 4952243 | 4953136 | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG16700* | chrX | 16985887 | 16992521 | 1 / 1 | 1 / - | 1 / 1 | - / - | 1 / 1 | 1 / 1 |
| *smp-30* | chr3R | 10571623 | 10576968 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Hsromega* | chr3R | 17121849 | 17143558 | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG8791* | chr2R | 3814158 | 3817427 | - / - | - / - | - / - | - / - | - / - | - / - |
| *psq* | chr2R | 6445393 | 6504785 | 2 / 1 | - / - | 2 / - | 1 / - | 2 / 1 | 2 / 1 |
| *stan* | chr2R | 6560850 | 6608643 | 5 / - | 1 / - | 4 / - | - / - | 5 / - | 5 / - |
| *Lsm10* | chr2R | 6708603 | 6709145 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Taf5* | chr2R | 6764817 | 6767178 | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG30016* | chr2R | 6762397 | 6762897 | - / - | - / - | - / - | - / - | - / - | - / - |
| *Pex6* | chr2R | 6767479 | 6770707 | - / - | - / - | - / - | - / - | - / - | - / - |
| *stv* | chr3L | 13470641 | 13476615 | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG12943* | chr2R | 6791075 | 6793068 | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG30379* | chr2R | 3827461 | 3829375 | - / - | - / - | - / - | - / - | - / - | - / - |
| *clumsy* | chr2L | 21206493 | 21211818 | - / - | - / - | - / - | - / - | - / - | - / - |
| *lola* | chr2R | 6369399 | 6430796 | 1 / - | - / - | 1 / - | - / - | 1 / - | 1 / - |
| *Oatp30B* | chr2L | 9521210 | 9540060 | 4 / 2 | 1 / - | 4 / 1 | 1 / - | 4 / 2 | 4 / 2 |
| *CG31619* | chr3R | 21684106 | 21729051 | - / - | - / - | - / - | - / - | - / - | - / - |
| *E(spl)m7-HLH* | chr2L | 21862760 | 21863741 | 1 / 1 | - / - | 1 / 1 | - / - | 1 / 1 | 1 / 1 |
| *cpo* | chr3R | 13745554 | 13844614 | 8 / 2 | 4 / 1 | 7 / 3 | 1 / - | 8 / 2 | 8 / 2 |
| *whd* | chr2R | 6356978 | 6366072 | 2 / - | - / - | - / - | 1 / - | 1 / - | 1 / - |
| *nclb* | chr2R | 6763253 | 6764801 | - / - | - / - | - / - | - / - | - / - | - / - |
| *CG12054* | chr3R | 27046424 | 27056981 | 1 / 1 | - / - | 1 / - | - / - | 1 / 1 | 1 / 1 |
| *Dyrk2* | chr2L | 14184478 | 14234126 | 3 / 2 | 2 / - | 3 / 2 | 1 / - | 3 / 2 | 2 / 2 |
| *hdc* | chr2R | 6125493 | 6133582 | - / - | - / - | - / - | - / - | - / - | - / - |
| *shep* | chr3L | 5148921 | 5271044 | 4 / 2 | 4 / - | 5 / 3 | - / - | 7 / 2 | 6 / 2 |
| *chas* | chrX | 17566491 | 17592763 | 2 / 2 | 1 / - | 1 / 1 | 1 / - | 2 / 1 | 2 / 1 |
| *Ire1* | chr3R | 15679630 | 15687013 | 1 / 1 | 1 / 1 | 1 / 1 | - / - | 1 / 1 | 1 / 1 |
| *Octbeta3R* | chr3R | 8337292 | 8373951 | 3 / 3 | 2 / 1 | 3 / 3 | 2 / 1 | 3 / 3 | 3 / 3 |

**Table S5** Top GO terms enriched for genes significantly correlated with latitude (*ln*(BF)>5 or $P < 0.0029$). The third column shows overlap with an equivalent GO enrichment analysis we performed on North American candidate genes (Fabian *et al.* 2012).

| GOID | Term | % overlap N. America | Clusters under | N genes | % genes of term | Term *P*-value | Term *P*-val. corrected |
|---|---|---|---|---|---|---|---|
| GO:0042330 | taxis | 75 | taxis | 38 | 10,73 | 2,01E-11 | 7,84E-09 |
| GO:0048666 | neuron development | 100 | taxis | 56 | 7,87 | 8,41E-11 | 1,64E-08 |
| GO:0048736 | appendage development | 50 | appendage development | 42 | 8,94 | 5,67E-10 | 3,68E-08 |
| GO:0031175 | neuron projection development | 100 | taxis | 48 | 8,19 | 5,26E-10 | 4,09E-08 |
| GO:0035120 | post-embryonic appendage morphogenesis | 75 | appendage development | 41 | 8,99 | 7,86E-10 | 4,37E-08 |
| GO:0035239 | tube morphogenesis | 100 | appendage development | 49 | 8,10 | 4,91E-10 | 4,78E-08 |
| GO:0048667 | cell morphogenesis involved in neuron differentiation | 100 | taxis | 42 | 8,73 | 1,16E-09 | 5,65E-08 |
| GO:0000904 | cell morphogenesis involved in differentiation | 100 | taxis | 45 | 8,57 | 4,77E-10 | 6,19E-08 |
| GO:0060562 | epithelial tube morphogenesis | 100 | appendage development | 46 | 8,13 | 1,71E-09 | 7,38E-08 |
| GO:0002009 | morphogenesis of an epithelium | 75 | appendage development | 52 | 7,45 | 3,61E-09 | 8,78E-08 |
| GO:0006928 | movement of cell or subcellular component | 100 | taxis | 45 | 8,01 | 4,30E-09 | 8,81E-08 |
| GO:0032990 | cell part morphogenesis | 75 | taxis | 48 | 7,77 | 3,88E-09 | 8,87E-08 |
| GO:0048737 | imaginal disc-derived appendage development | 75 | appendage development | 40 | 8,70 | 3,53E-09 | 9,16E-08 |
| GO:0048699 | generation of neurons | 100 | generation of neurons | 61 | 6,80 | 4,24E-09 | 9,16E-08 |
| GO:0071944 | cell periphery | 100 | cell periphery | 49 | 7,68 | 3,39E-09 | 9,43E-08 |
| GO:0030182 | neuron differentiation | 100 | taxis | 58 | 7,02 | 2,45E-09 | 9,54E-08 |
| GO:0030030 | cell projection organization | 100 | taxis | 52 | 7,49 | 3,29E-09 | 9,85E-08 |
| GO:0061564 | axon development | 75 | taxis | 33 | 10,03 | 2,97E-09 | 1,05E-07 |
| GO:0005886 | plasma membrane | 100 | cell periphery | 44 | 8,19 | 3,25E-09 | 1,05E-07 |
| GO:0000902 | cell morphogenesis | 100 | taxis | 54 | 7,16 | 6,97E-09 | 1,36E-07 |

# Bibliography


Adrion JR, Hahn MW, Cooper BS (2015) Revisiting classic clines in *Drosophila melanogaster* in the age of genomics. *Trends in Genetics,* **31**, 434-444.

Alic N, Giannakou ME, Papatheodorou I *et al.* (2014) Interplay of dFOXO and two ETS-family transcription factors determines lifespan in *Drosophila melanogaster*. *PLoS Genetics*, **10**, e1004619.

Allada R, Chung BY (2010) Circadian organization of behavior and physiology in *Drosophila*. *Annual Review of Physiology*, **72**, 605-624.

Allen VW, O'Connor RM, Ulgherait M *et al.* (2016) period-Regulated Feeding Behavior and TOR Signaling Modulate Survival of Infection. *Current Biology*, **26**, 1383.

Amorim CEG, Daub JT, Salzano FM, Foll M, Excoffier L (2015) Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PLoS one*, **10**, e0121557.

Anderson AR, Hoffmann AA, McKechnie SW (2005) Response to selection for rapid chill-coma recovery in *Drosophila melanogaster*: physiology and life-history traits. *Genetics Research*, **85**, 15–22.

Angilletta MJ (2009) *Thermal adaptation: a theoretical and empirical synthesis*. Oxford University Press.

Armstrong JD, Texada MJ, Munjaal R, Baker DA, Beckingham KM (2006) Gravitaxis in *Drosophila melanogaster*: a forward genetic screen. *Genes, Brain, and Behavior*, **5**, 222–239.

Ashburner M, Lemeunier F (1976) Relationships within the melanogaster species subgroup of the genus *Drosophila* (Sophophora). I. Inversion polymorphisms in *Drosophila melanogaster* and *Drosophila simulans*. *Proceedings of the Royal Society of London B: Biological Sciences*, **193**, 137–157.

Avise JC (2004) *Molecular markers, natural history and evolution*. 2nd Edition. Sinauer Associates, Sunderland, Massachusetts.

Ayrinhac A, Debat V, Gibert P *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 JF, Carbone MA, Stone EA *et al.* (2009) Systems genetics of complex traits in *Drosophila melanogaster*. *Nature Genetics*, **41**, 299–307.

Azad P, Zhou D, Zarndt R, Haddad GG (2012) Identification of genes underlying hypoxia tolerance in *Drosophila* by a P-element screen. *G3: Genes, Genomes, Genetics*, **2**, 1169–1178.

Bae K, Edery I (2006) Regulating a circadian clock's period, phase and amplitude by phosphorylation: insights from *Drosophila*. *Journal of biochemistry*, **140**, 609–617.

Baena-López LA, Alonso J, Rodriguez J, Santarén JF (2008) The expression of heat shock protein HSP60A reveals a dynamic mitochondrial pattern in *Drosophila melanogaster* embryos. *Journal of proteome research*, **7**, 2780–2788.

Baker BS, Taylor BJ, Hall JC (2001) Are complex behaviors specified by dedicated regulatory genes? Reasoning from *Drosophila*. *Cell*, **105**, 13–24.

Bale JS (1993) Classes of Insect Cold Hardiness. *Functional Ecology*, **7**, 751–753.

Baltzer C, Tiefenböck SK, Marti M, Frei C (2009) Nutrition controls mitochondrial biogenesis in the *Drosophila* adipose tissue through Delg and cyclin D/Cdk4. *PLoS one*, **4**, e6935.

Barricelli NA (1954) Esempi numerici di processi di evoluzione. *Methodos*, **6**, 45–68.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridising populations. *Heredity*, **57**, 357–376.

Bataillé L, Delon I, Da Ponte JP, Brown NH, Jagla K (2010) Downstream of identity genes: muscle-type-specific regulation of the fusion process. *Developmental cell*, **19**, 317–328.

Beaumont MA (2010) Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 379–406.

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

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. *PLoS Genetics*, **10**, e1004412.

Berrigan D, Partridge L (1997) Influence of temperature and activity on the metabolic rate of adult Drosophila melanogaster. *Comparative biochemistry and physiology. Part A, Physiology*, **118**, 1301–1307.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, **19**, 2609–2625.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

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

Bindea G, Mlecnik B, Hackl H *et al.* (2009) ClueGO: a Cytoscape plugin to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091–1093.

Bi P, Yue F, Sato Y *et al.* (2016) Stage-specific effects of Notch activation during skeletal myogenesis. *eLife*, **5**.

Blair LM, Granka JM, Feldman MW (2014) On the stability of the Bayenv method in assessing human SNP-environment associations. *Human Genomics*, **8**, 1.

Blanchard FJ, Collins B, Cyran SA *et al.* (2010) The transcription factor Mef2 is required for normal circadian behavior in *Drosophila*. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, **30**, 5855–5865.

Blanco E, Ruiz-Romero M, Beltran S *et al.* (2010) Gene expression following induction of regeneration in Drosophila wing imaginal discs. Expression profile of regenerating wing discs. *BMC Developmental Biology*, **10**, 94.

Blau J, Young MW (1999) Cycling vrille expression is required for a functional *Drosophila* clock. *Cell*, **99**, 661–671.

Blum MJ (2002) Rapid movement of a *Heliconius* hybrid zone: evidence for phase III of Wright's shifting balance theory? *Evolution*, **56**, 1992–1998.

Bollinger T, Bollinger A, Oster H, Solbach W (2010) Sleep, immunity, and circadian clocks: a mechanistic model. *Gerontology*, **56**, 574–580.

Boulétreau-Merle J, Fouillet P (2002) How to overwinter and be a founder: egg-retention phenotypes and mating status in *Drosophila melanogaster*. *Evolutionary Ecology*, **16**, 309–332.

Bowler PJ (1989) *Evolution: the history of an idea*. University of California Press.

Broderick NA, Buchon N, Lemaitre B (2014) Microbiota-Induced Changes in *Drosophila melanogaster* Host Gene Expression and Gut Morphology. *mBio*, **5**, e01117–14.

Broughton SJ, Piper MDW, Ikeya T *et al.* (2005) Longer lifespan, altered metabolism, and stress resistance in *Drosophila* from ablation of cells making insulin-like ligands. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 3105–3110.

Bubliy OA, Loeschcke V (2000) High stressful temperature and genetic variation of five quantitative traits in *Drosophila melanogaster*. *Genetica*, **110**, 79–85.

Bubliy OA, Loeschcke V (2005) Correlated responses to selection for stress resistance and longevity in a laboratory population of *Drosophila melanogaster*. *Journal of evolutionary biology*, **18**, 789–803.

Cagan RL (2011) The Drosophila nephrocyte. *Current opinion in nephrology and hypertension*, **20**, 409–415.

Carbone MA, Jordan KW, Lyman RF *et al.* (2006) Phenotypic variation and natural selection at catsup, a pleiotropic quantitative trait gene in *Drosophila*. *Current Biology*, **16**, 912–919.

Carvalho M, Sampaio JL, Palm W *et al.* (2012) Effects of diet and development on the Drosophila lipidome. *Molecular systems biology*, **8**, 600.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical research*, **70**, 155–174.

Chen W-F, Majercak J, Edery I (2006) Clock-gated photic stimulation of timeless expression at cold temperatures and seasonal adaptation in *Drosophila*. *Journal of biological rhythms*, **21**, 256–271.

Chen J, Nolte V, Schlötterer C (2015) Temperature Stress Mediates Decanalization and Dominance of Gene Expression in *Drosophila melanogaster*. *PLoS Genetics*, **11**, e1004883.

Chintapalli VR, Al Bratty M, Korzekwa D, Watson DG, Dow JAT (2013) Mapping an atlas of tissue-specific *Drosophila melanogaster* metabolomes by high resolution mass spectrometry. *PLoS one*, **8**, e78066.

Chippindale AK, Gibbs AG, Sheik M *et al.* (1998) Resource acquisition and the evolution of stress resistance in *Drosophila melanogaster*. *Evolution*, **52**, 1342–1352.

Chiu JC, Low KH, Pike DH, Yildirim E, Edery I (2010) Assaying locomotor activity to study circadian rhythms and sleep parameters in *Drosophila*. *Journal of visualized experiments: JoVE*, **43**, e2157-e2157.

Chouteau M, Angers B (2012) Wright's shifting balance theory and the diversification of aposematic signals. *PLoS one*, **7**, e34028.

Chown SL, Gaston KJ, Robinson D (2004) Macrophysiology: large-scale patterns in physiological traits and their ecological implications. *Functional Ecology*, **18**, 159–167.

Chuang H-H, Neuhausser WM, Julius D (2004) The super-cooling agent icilin reveals a mechanism of coincidence detection by a temperature-sensitive TRP channel. *Neuron*, **43**, 859–869.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, **15**, 1496–1502.

Clark RI, Tan SWS, Péan CB *et al.* (2013) MEF2 is an in vivo immune-metabolic switch. *Cell*, **155**, 435–447.

Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213–220.

Colinet H, Larvor V, Bical R, Renault D (2012) Dietary sugars affect cold tolerance of *Drosophila melanogaster*. *Metabolomics: Official journal of the Metabolomic Society*, **9**, 608–622.

Colinet H, Lee SF, Hoffmann A (2010a) Functional characterization of the Frost gene in *Drosophila melanogaster*: importance for recovery from chill coma. *PLoS one*, **5**, e10925.

Colinet H, Lee SF, Hoffmann A (2010b) Knocking down expression of Hsp22 and Hsp23 by RNA interference affects recovery from chill coma in *Drosophila melanogaster*. *Journal of Experimental Biology*, **213**, 4146–4150.

Colinet H, Overgaard J, Com E, Sørensen JG (2013) Proteomic profiling of thermal acclimation in *Drosophila melanogaster*. *Insect biochemistry and molecular biology*, **43**, 352–365.

Connolly K (1966) Locomotor activity in *Drosophila*. II. Selection for active and inactive strains. *Animal behaviour*, **14**, 444–449.

Connolly K (1967) Locomotor activity in *Drosophila* III. A distinction between activity and reactivity. *Animal behaviour*, **15**, 149–152.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Coyne JA, Barton NH, Turelli M (2000) Is Wright's shifting balance process important in evolution? *Evolution*, **54**, 306–317.

Crill WD, Huey RB, Gilchrist GW (1996) Within- and Between-Generation Effects of Temperature on the Morphology and Physiology of *Drosophila melanogaster*. *Evolution*, **50**, 1205–1218.

Croft D, O'Kelly G, Wu G *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, **39**, D691–7.

Crow JF (2008) Mid-century controversies in population genetics. *Annual review of genetics*, **42**, 1–16.

Crow JF, Engels WR, Denniston C (1990) Phase Three of Wright's Shifting-Balance Theory. *Evolution*, **44**, 233–247.

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

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, **25**, 410–418.

Da Lage JL, Capy P, David JR (1990) Starvation and desiccation tolerance in *Drosophila melanogaster*: differences between European, North African and Afrotropical populations. *Genetics, Selection, Evolution: GSE*, **22**, 381–391.

Daub JT, Dupanloup I, Robinson-Rechavi M, Excoffier L (2015) Inference of Evolutionary Forces Acting on Human Biological Pathways. *Genome biology and evolution*, **7**, 1546–1558.

Daub JT, Hofer T, Cutivet E *et al.* (2013) Evidence for polygenic adaptation to pathogens in the human genome. *Molecular biology and evolution*, **30**, 1544–1558.

David JR, Capy P (1988) Genetic variation of *Drosophila melanogaster* natural populations. *Trends in Genetics*, **4**, 106–111.

David JR, Gibert P, Pla E *et al.* (1998) Cold stress tolerance in *Drosophila*: analysis of chill coma recovery in *D. melanogaster*. *Journal of Thermal Biology*, **23**, 291–299.

Deutsch CA, Tewksbury JJ, Huey RB *et al.* (2008) Impacts of climate warming on terrestrial ectotherms across latitude. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 6668–6672.

De J, Varma V, Saha S, Sheeba V, Sharma VK (2013) Significance of activity peaks in fruit flies, *Drosophila melanogaster*, under seminatural conditions. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 8984–8989.

Dockendorff TC, Su HS, McBride SMJ *et al.* (2002) *Drosophila* lacking *dfmr1* activity show defects in circadian output and fail to maintain courtship interest. *Neuron*, **34**, 973–984.

Dopazo J, Amadoz A, Bleda M *et al.* (2016) 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. *Molecular biology and evolution*, **33**, 1205–1218.

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

Durham MF, Magwire MM, Stone EA, Leips J (2014) Genome-wide analysis in *Drosophila* reveals age-specific effects of SNPs on fitness traits. *Nature Communications*, **5**, 4338.

Dusik V, Senthilan PR, Mentzel B *et al.* (2014) The MAP kinase p38 is part of *Drosophila melanogaster*'s circadian clock. *PLoS Genetics*, **10**, e1004565.

Eldredge N (1995) *Reinventing Darwin: the Great Debate at the High Table of Evolutionary Theory*. Wiley, New York.

Elhaik E (2012) Empirical distributions of F(ST) from large-scale human polymorphism data. *PLoS ONE*, **7**, e49837.

Endler JA (1977) Geographic variation, speciation, and clines. *Monographs in population biology*, **10**, 1–246.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186–194.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Fabian DK, Kapun M, Nolte V, Kofler R (2012) Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology*, **21**, 4748-4769.

Fallis LC (2012) The evolution and genetics of thermal traits in *Drosophila melanogaster*. Kansas State University.

Fedotov SA, Bragina JV, Besedina NG *et al.* (2014) The effect of neurospecific knockdown of candidate genes for locomotor behavior and sound production in *Drosophila melanogaster*. *Fly*, **8**, 176–187.

Felsenstein J (2000) *From population genetics to evolutionary genetics.* In: *A View Through the Trees of Evolutionary Genetics: From Molecules to Morphology*, pp. 609–627. New York: Cambridge University Press.

Ferrandon D, Imler J-L, Hetru C, Hoffmann JA (2007) The *Drosophila* systemic immune response: sensing and signalling during bacterial and fungal infections. *Nature Reviews Immunology*, **7**, 862–874.

Fisher RA (1930) *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press.

Flourakis M, Kula-Eversole E, Hutchison AL *et al.* (2015) A Conserved Bicycle Model for Circadian Clock Control of Membrane Excitability. *Cell*, **162**, 836–848.

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

Fraser A (1957) Simulation of genetic systems by automatic digital computers. I. Introduction. *Australian journal of biological sciences*, **10**, 484–491.

Frazier M r., Huey R b., Berrigan D, Associate Editor and Editor: Jonathan B. Losos (2006) Thermodynamics Constrains the Evolution of Insect Population Growth Rates: "Warmer Is Better." *The American naturalist*, **168**, 512–520.

Frenkel L, Ceriani MF (2011) Circadian plasticity: from structure to behavior. *International review of neurobiology*, **99**, 107–138.

Gabriel W (2005) How stress selects for reversible phenotypic plasticity. *Journal of Evolutionary Biology*, **18**, 873–883.

Gagnaire P-A, Gaggiotti OE (2016) Detecting polygenic selection in marine populations by combining population genomics and quantitative genetics approaches. *Current Zoology*, **62**, 603-616.

Gates MA, Kannan R, Giniger E (2011) A genome-wide analysis reveals that the *Drosophila* transcription factor *Lola* promotes axon growth in part by suppressing expression of the actin nucleation factor *Spire*. *Neural development*, **6**, 37.

Gibert P, Huey RB, Gilchrist GW (2001a) Locomotor performance of *Drosophila melanogaster*: interactions among developmental and adult temperatures, age, and geography. *Evolution*, **55**, 205–209.

Gibert P, Moreteau B, Pétavy G, Karan D, David JR (2001b) Chill-coma tolerance, a major climatic adaptation among *Drosophila* species. *Evolution*, **55**, 1063–1068.

Gilchrist GW, Huey RB, Partridge L (1997) Thermal sensitivity of *Drosophila melanogaster*: evolutionary responses of adults and eggs to laboratory natural selection at different temperatures. *Physiological Zoology*, **70**, 403–414.

Glaser FT, Stanewsky R (2005) Temperature synchronization of the *Drosophila* circadian clock. *Current Biology*, **15**, 1352–1363.

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

Glossop NRJ, Houl JH, Zheng H *et al.* (2003) *VRILLE* feeds back to control circadian transcription of Clock in the *Drosophila* circadian oscillator. *Neuron*, **37**, 249–261.

Goda T, Sharp B, Wijnen H (2014) Temperature-dependent resetting of the molecular circadian oscillator in *Drosophila*. *Proceedings of the Royal Society B: Biological Sciences*, **281**.

Goenaga J, Fanara JJ, Hasson E (2010) A quantitative genetic study of starvation resistance at different geographic scales in natural populations of *Drosophila melanogaster*. *Genetics Research*, **92**, 253–259.

Goenaga J, Fanara JJ, Hasson E (2013) Latitudinal Variation in Starvation Resistance is Explained by Lipid Content in Natural Populations of *Drosophila melanogaster*. *Evolutionary Biology*, **40**, 601–612.

Goenaga J, Mensch J, Fanara JJ, Hasson E (2012) The effect of mating on starvation resistance in natural populations of *Drosophila melanogaster*. *Evolutionary Ecology*, **26**, 813–823.

Goldfine H (2010) The appearance, disappearance and reappearance of plasmalogens in evolution. *Progress in lipid research*, **49**, 493–498.

Greene JC (1981) *Science, Ideology, and World View Essays in the History of Evolutionary Ideas*. Monograph Collection (Matt - Pseudo).

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Haldane JBS (1932) The causes of evolution, 1932. *Princeton, NJ: Princeton University Press. Haldane The Causes of Evolution (1932)*.

Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD (2004) Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Research*, **14**, 273–279.

Halligan DL, Keightley PD (2006) Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Research*, **16**, 875–884.

Harbison ST, Chang S, Kamdar KP, Mackay TFC (2005) Quantitative genomics of starvation stress resistance in *Drosophila*. *Genome biology*, **6**, R36.

Harwood J (2000) The rediscovery of Mendelism in agricultural context: Erich von Tschermak as plant-breeder. *Comptes rendus de l'Académie des sciences. Serie III, Sciences de la vie*, **323**, 1061–1067.

Hemmer LW, Blumenstiel JP (2016) Holding it together: rapid evolution and positive selection in the synaptonemal complex of *Drosophila*. *BMC evolutionary biology*, **16**, 91.

Hey J, Fitch WM, Ayala FJ, Others (2005) *Systematics and the Origin of Species: On Ernst Mayr's 100th Anniversary*. National Academies Press.

Higashi-Kovtun ME, Mosca TJ, Dickman DK, Meinertzhagen IA, Schwarz TL (2010) Importin-11 Regulates Synaptic Phosphorylated Mothers Against Decapentaplegic, and Thereby Influences Synaptic Development and Function at the *Drosophila* Neuromuscular Junction. *Journal of Neuroscience*, **30**, 5253–5268.

Hoffmann JA (2003) The immune response of *Drosophila*. *Nature*, **426**, 33–38.

Hoffmann AA (2010) Physiological climatic limits in *Drosophila*: patterns and implications. *The Journal of Experimental Biology*, **213**, 870–880.

Hoffmann AA, Anderson A, Hallas R (2002) Opposing clines for high and low temperature resistance in *Drosophila melanogaster*. *Ecology Letters*, **5**, 614–618.

Hoffmann AA, Hallas R, Anderson AR, Telonis-Scott M (2005) Evidence for a robust sex-specific trade-off between cold resistance and starvation resistance in *Drosophila melanogaster*. *Journal of Evolutionary Biology,* **18**, 804–810.

Hoffmann AA, Hallas R, Sinclair C, Mitrovski P (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 AA, Parsons PA, Others (1991) *Evolutionary genetics and environmental stress*. Oxford University Press.

Hoffmann JA, Reichhart JM (2002) *Drosophila* innate immunity: an evolutionary perspective. *Nature Immunology*, **3**(2), 121-126.

Hoffmann AA, Rieseberg LH (2008) Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annual Review of Ecology, Evolution, and Systematics*, **39**, 21–42.

Hoffmann AA, Sørensen JG, Loeschcke V (2003) Adaptation of *Drosophila* to temperature extremes: bringing together quantitative and molecular approaches. *Journal of Thermal Biology*, **28**, 175–216.

Hoffmann AA, Weeks AR (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.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature Reviews Genetics*, **10**, 639–650.

Huang W, Massouras A, Inoue Y *et al.* (2014) Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Research*, **24**, 1193–1208.

Huang W, Richards S, Carbone MA *et al.* (2012) Epistasis dominates the genetic architecture of *Drosophila* quantitative traits. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 15553–15559.

Huber CD, DeGiorgio M, Hellmann I, Nielsen R (2016) Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology*, **25**, 142–156.

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

Huey RB (2010) Evolutionary physiology of insect thermal adaptation to cold environments. *Low temperature biology of insects*, 223–241.

Huey RB, Berrigan D (2001) Temperature, demography, and ectotherm fitness. *The American naturalist*, **158**, 204–210.

Huxley JS (1942) *Evolution, the Modern Synthesis*. London: Allen & Unwin

Immonen E, Ritchie MG (2012) The genomic response to courtship song stimulation in female *Drosophila melanogaster*. *Proceedings of the Royal Society B*, **279**, 1359–1365.

Iyer EPR, Iyer SC, Sullivan L *et al.* (2013) Functional genomic analyses of two morphologically distinct classes of *Drosophila* sensory neurons: post-mitotic roles of transcription factors in dendritic patterning. *PLoS one*, **8**, e72434.

Janzen DH (1967) Why Mountain Passes are Higher in the Tropics. *The American naturalist*, **101**, 233–249.

Jeyaraj D, Scheer FAJL, Ripperger JA *et al.* (2012) *Klf15* orchestrates circadian nitrogen homeostasis. *Cell metabolism*, **15**, 311–323.

Johnson N (2008) Sewall Wright and the development of shifting balance theory. *Nature Education*, **1**, 52.

Jordan KW, Morgan TJ, Mackay TFC (2006) Quantitative trait loci for locomotor behavior in *Drosophila melanogaster*. *Genetics*, **174**, 271–284.

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**, 27–30.

Kaneko H, Head LM, Ling J *et al.* (2012) Circadian rhythm of temperature preference and its neural control in *Drosophila*. *Current Biology*, **22**, 1851–1857.

Kass RE, Raftery AE (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters*, **7**, 1225–1241.

Keller A (2007) *Drosophila melanogaster*'s history as a human commensal. *Current Biology*, **17**, R77–81.

Keller I, Alexander JM, Holderegger R, Edwards PJ (2013) Widespread phenotypic and genetic divergence along altitudinal gradients in animals. *Journal of evolutionary biology*, **26**, 2527–2543.

Kennington WJ, Gilchrist AS, Goldstein DB, Partridge L (2001) The genetic bases of divergence in desiccation and starvation resistance among tropical and temperate populations of *Drosophila melanogaster*. *Heredity*, **87**, 363–372.

Kenny MC, Wilton A, Ballard JWO (2008) Seasonal trade-off between starvation resistance and cold resistance in temperate wild-caught *Drosophila simulans*. *Australian journal of entomology*, **47**, 20–23.

Kidd S, Struhl G, Lieber T (2015) Notch is required in adult *Drosophila* sensory neurons for morphological and functional plasticity of the olfactory circuit. *PLoS Genetics*, **11**, e1005244.

Kimura M (1965) A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proceedings of the National Academy of Sciences of the United States of America*, **54**, 731–736.

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

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kimura M, Ohta T (1969) The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics*, **63**, 701–709.

Kolaczkowski B, Kern AD, Holloway AK, Begun DJ (2011) Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. *Genetics*, **187**, 245–260.

Konopka RJ, Kyriacou CP, Hall JC (1996) Mosaic Analysis in the *Drosophila* Cns of Circadian and Courtship-Song Rhythms Affected by a Period Clock Mutation: Short Communication. *Journal of neurogenetics*, **11**, 117–139.

Koštál V, Šimek P, Zahradníčková H, Cimlová J, Štětina T (2012) Conversion of the chill susceptible fruit fly larva (*Drosophila melanogaster*) to a freeze tolerant organism. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 3270–3274.

Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature*, **304**, 412–417.

Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, **156**, 439–447.

Kunst M, Tso MCF, Ghosh DD, Herzog ED, Nitabach MN (2015) Rhythmic control of activity and sleep by class B1 GPCRs. *Critical reviews in biochemistry and molecular biology*, **50**, 18–30.

Kuo T-H, Pike DH, Beizaeipour Z, Williams JA (2010) Sleep triggered by an immune response in *Drosophila* is regulated by the circadian clock and requires the NFκB Relish. *BMC neuroscience*, **11**, 1.

Kyriacou CP, Peixoto AA, Sandrelli F, Costa R, Tauber E (2008) Clines in clock genes: fine-tuning circadian rhythms to the environment. *Trends in genetics: TIG*, **24**, 124–132.

Lachaise D, Cariou M-L, David JR *et al.* (1988) Historical Biogeography of the *Drosophila melanogaster* Species Subgroup. In: *Evolutionary Biology* Evolutionary Biology, pp. 159–225. Springer US.

Lachaise D, Silvain J-F (2004) How two Afrotropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster-D. simulans* palaeogeographic riddle. *Genetica*, **120**, 17–39.

Lack JB, Cardeno CM, Crepeau MW *et al.* (2015) The *Drosophila* Genome Nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics*, **199**, 1229–1241.

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

Latta RG (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *The American naturalist*, **151**, 283–292.

Laurent SJY, Werzner A, Excoffier L, Stephan W (2011) Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. *Molecular Biology and Evolution*, **28**, 2041–2051.

Lazzaro BP (2008) Natural selection on the *Drosophila* antimicrobial immune system. *Current opinion in microbiology*, **11**, 284–289.

Lear BC, Darrah EJ, Aldrich BT *et al.* (2013) UNC79 and UNC80, putative auxiliary subunits of the NARROW ABDOMEN ion channel, are indispensable for robust circadian locomotor rhythms in *Drosophila*. *PLoS one*, **8**, e78147.

Le Corre V, Kremer A (2003) Genetic Variability at Neutral Markers, Quantitative Trait Loci and Trait in a Subdivided Population Under Selection. *Genetics*, **164**, 1205–1219.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548–1566.

Lee J-Y, Bhatt D, Bhatt D, Chung W-Y, Cooper RL (2009) Furthering pharmacological and physiological assessment of the glutamatergic receptors at the *Drosophila* neuromuscular junction. *Comparative Biochemistry and Physiology. Toxicology & pharmacology: CBP*, **150**, 546–557.

Lee Y, Montell C (2013) Drosophila TRPA1 functions in temperature control of circadian rhythm in pacemaker neurons. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, **33**, 6716–6725.

Lenay C (2000) Hugo De Vries: from the theory of intracellular pangenesis to the rediscovery of Mendel. *Comptes rendus de l'Académie des sciences. Serie III, Sciences de la vie*, **323**, 1053–1060.

Lenz O, Xiong J, Nelson MD, Raizen DM, Williams JA (2015) FMRFamide signaling promotes stress-induced sleep in *Drosophila*. *Brain, behavior, and immunity*, **47**, 141–148.

Lewontin RC, Hubby JL (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, **54**, 595–609.

Lewontin RC, Provine JA, Wallace WB *et al.* (1981) *Dobzhansky's genetics of natural populations I-XLIII*.

Littleton JT, Ganetzky B (2000) Ion channels and synaptic organization: analysis of the *Drosophila* genome. *Neuron*, **26**, 35–43.

Liu S, Lamaze A, Liu Q *et al.* (2014) WIDE AWAKE mediates the circadian timing of sleep onset. *Neuron*, **82**, 151–166.

Liu L, Xu Y-X, Hirschberg CB (2010) The role of nucleotide sugar transporters in development of eukaryotes. *Seminars in cell & developmental biology*, **21**, 600–608.

Li Y, Zhang Z, Robinson GE, Palli SR (2007) Identification and characterization of a juvenile hormone response element and its binding proteins. *Journal of Biological Chemistry*, **282**, 37605–37617.

López Del Amo V, Seco-Cervera M, García-Giménez JL *et al.* (2015) Mitochondrial defects and neuromuscular degeneration caused by altered expression of *Drosophila* Gdap1: implications for the Charcot-Marie-Tooth neuropathy. *Human molecular genetics*, **24**, 21–36.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Lupu A, Pechkovskaya A, Rashkovetsky E, Nevo E, Korol A (2004) DNA repair efficiency and thermotolerance in *Drosophila melanogaster* from "Evolution Canyon." *Mutagenesis*, **19**, 383–390.

Mackay TFC, Richards S, Stone EA *et al.* (2012) The *Drosophila melanogaster* Genetic Reference Panel. *Nature*, **482**, 173–178.

Macmillan HA, Sinclair BJ (2011) Mechanisms underlying insect chill-coma. *Journal of Insect Physiology*, **57**, 12–20.

Markow TA, O'Grady P (2008) Reproductive ecology of *Drosophila*. *Functional Ecology*, **22**, 747–759.

Martinek S, Inonog S, Manoukian AS, Young MW (2001) A role for the segment polarity gene shaggy/GSK-3 in the *Drosophila* circadian clock. *Cell*, **105**, 769–779.

Mayr E (1959) Where are We. Genetics and Twentieth Century Darwinism. In: *Cold Spring Harbor Symposia on Quantitative Biology*, pp. 1–14.

Mayr E, Provine WB (1981) The Evolutionary Synthesis. *Bulletin of the American Academy of Arts and Sciences*, **34**, 17–32.

McClung CA (2013) How might circadian rhythms control mood? Let me count the ways. *Biological psychiatry*, **74**, 242–249.

McLaren W, Pritchard B, Rios D *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.

Means JC, Venkatesan A, Gerdes B *et al.* (2015) *Drosophila* spaghetti and doubletime link the circadian clock and light to caspases, apoptosis and tauopathy. *PLoS Genetics*, **11**, e1005171.

Medina I, Casal J, Fabre CCG (2015) Do circadian genes and ambient temperature affect substrate-borne signalling during *Drosophila* courtship? *Biology open*, **4**, 1549–1557.

Meehan MJ, Wilson R (1987) Locomotor activity in the Tyr-1 mutant of *Drosophila melanogaster*. *Behavior Genetics*, **17**, 503–512.

Mendizabal I, Marigorta UM, Lao O, Comas D (2012) Adaptive evolution of loci covarying with the human African Pygmy phenotype. *Human Genetics*, **131**, 1305–1317.

Mensch J, Lavagnino N, Carreira V *et al.* (2008) Identifying candidate genes affecting developmental time in *Drosophila melanogaster*: pervasive pleiotropy and gene-by-environment interaction. *BMC Developmental Biology*, **8**, 78.

Meyer S, Schmidt I, Klämbt C (2014) Glia ECM interactions are required to shape the *Drosophila* nervous system. *Mechanisms of development*, **133**, 105–116.

Mitrovski P, Hoffmann AA (2001) Postponed reproduction as an adaptation to winter conditions in *Drosophila melanogaster*: evidence for clinal variation under semi-natural conditions. *Proceedings of the Royal Society B: Biological Sciences*, **268**, 2163–2168.

Moran CN, Kyriacou CP (2009) Functional neurogenomics of the courtship song of male *Drosophila melanogaster*. *Cortex*, **45**, 18–34.

Morgan TJ, Mackay TFC (2006) Quantitative trait loci for thermotolerance phenotypes in *Drosophila melanogaster*. *Heredity*, **96**, 232–242.

Muller HJ (1949) The Darwinian and modern conceptions of natural selection. *Proceedings of the American Philosophical Society*, **93**, 459–470.

Nahum JR, Godfrey-Smith P, Harding BN *et al.* (2015) A tortoise-hare pattern seen in adapting structured and unstructured populations suggests a rugged fitness landscape in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 7530–7535.

Na J, Sweetwyne MT, Park ASD, Susztak K, Cagan RL (2015) Diet-Induced Podocyte Dysfunction in *Drosophila* and Mammals. *Cell reports*, **12**, 636–647.

Nevo E, Beiles A, Ben-Shlomo R (1984) The Evolutionary Significance of Genetic Diversity: Ecological, Demographic and Life History Correlates. In: *Evolutionary Dynamics of Genetic Diversity* Lecture Notes in Biomathematics., pp. 13–213. Springer Berlin Heidelberg.

Newell PD, Douglas AE (2014) Interspecies interactions determine the impact of the gut microbiota on nutrient allocation in *Drosophila melanogaster*. *Applied and Environmental Microbiology*, **80**, 788–796.

Nicodemus J, O'tousa JE, Duman JG (2006) Expression of a beetle, *Dendroides canadensis*, antifreeze protein in *Drosophila melanogaster*. *Journal of insect physiology*, **52**, 888–896.

Noble WS (2009) How does multiple testing correction work? *Nature Biotechnology*, **27**, 1135–1137.

Nordborg M, Innan H (2002) Molecular population genetics. *Current opinion in plant biology*, **5**, 69–73.

Norry FM, Gomez FH, Loeschcke V (2007) Knockdown resistance to heat stress and slow recovery from chill coma are genetically associated in a quantitative trait locus region of chromosome 2 in *Drosophila melanogaster*. *Molecular Ecology*, **16**, 3274–3284.

Norry FM, Scannapieco AC, Sambucetti P, Bertoli CI, Loeschcke V (2008) QTL for the thermotolerance effect of heat hardening, knockdown resistance to heat and chill-coma recovery in an intercontinental set of recombinant inbred lines of *Drosophila melanogaster*. *Molecular Ecology*, **17**, 4570–4581.

Nunney L (1996) The Response to Selection for Fast Larval Development in *Drosophila melanogaster* and its Effect on Adult Weight: An Example of a Fitness Trade-Off. *Evolution*, **50**, 1193–1204.

Obbard DJ, Welch JJ, Kim K-W, Jiggins FM (2009) Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genetics*, **5**, e1000698.

Ober U, Ayroles JF, Stone EA *et al.* (2012) Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1002685.

Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Molecular Biology and Evolution*, **22**, 2119–2130.

Organisti C, Hein I, Grunwald Kadow IC, Suzuki T (2015) Flamingo, a seven-pass transmembrane cadherin, cooperates with Netrin/Frazzled in *Drosophila* midline guidance. *Genes to Cells*, **20**, 50–67.

Overgaard J, Sørensen JG, Loeschcke V (2010) Genetic variability and evolution of cold-tolerance. In: *Low Temperature Biology of Insects*. Cambridge University Press.

Partridge L, Barrie B, Fowler K, French V (1994) Evolution and Development of Body Size and Cell Size in *Drosophila melanogaster* in Response to Temperature. *Evolution*, **48**, 1269–1276.

Partridge L, Fowler K (1992) Direct and Correlated Responses to Selection on Age at Reproduction in *Drosophila melanogaster*. *Evolution*, **46**, 76–91.

Partridge L, Hoffmann A, Jones JS (1987) Male size and mating success in *Drosophila melanogaster* and *D. pseudoobscura* under field conditions. *Animal Behaviour*, **35**, 468–476.

Partridge L, Prowse N, Pignatelli P (1999) Another set of responses and correlated responses to selection on age at reproduction in *Drosophila melanogaster*. *Proceedings of the Royal Society of London* B - *Biological Sciences*, **266**, 255–261.

Pavlidis P, Jensen JD, Stephan W, Stamatakis A (2012) A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular biology and evolution*, **29**, 3237–3248.

Peck SL, Ellner SP, Gould F (1998) A Spatially Explicit Stochastic Model Demonstrates the Feasibility of Wright's Shifting Balance Theory. *Evolution*, **52**, 1834–1839.

Petsakou A, Sapsis TP, Blau J (2015) Circadian Rhythms in Rho1 Activity Regulate Neuronal Plasticity and Network Hierarchy. *Cell*, **162**, 823–835.

Pierre SES, Ponting L, Stefancsik R, McQuilton P, the FlyBase Consortium (2014) FlyBase 102—advanced approaches to interrogating FlyBase. *Nucleic Acids Research*, **42**, D780–D788.

Pile LA, Spellman PT, Katzenberger RJ, Wassarman DA (2003) The SIN3 deacetylase complex represses genes encoding mitochondrial proteins: implications for the regulation of energy metabolism. *The Journal of Biological Chemistry*, **278**, 37840–37848.

Pool JE, Corbett-Detig RB, Sugino RP *et al.* (2012) Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture. *PLoS Genetics*, **8**, e1003080.

Porcelli D, Westram AM, Pascual M *et al.* (2016) Gene expression clines reveal local adaptation and associated trade-offs at a continental scale. *Scientific reports*, **6**, 32975.

Pritchard JK, Di Rienzo A (2010) Adaptation–not by sweeps alone. *Nature Reviews Genetics*, **11**, 665–667.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208–15.

Ragland GJ, Kingsolver JG (2008) Evolution of thermotolerance in seasonal environments: the effects of annual temperature variation and life-history timing in *Wyeomyia smithii*. *Evolution*, **62**, 1345–1357.

Rako L, Hoffmann AA (2006) Complexity of the cold acclimation response in *Drosophila melanogaster*. *Journal of Insect Physiology*, **52**, 94–104.

Randall D, Burggren WW, French K, Eckert R (2002) *Eckert animal physiology*. Macmillan.

Reisse S, Rothardt G, Völkl A, Beier K (2001) Peroxisomes and ether lipid biosynthesis in rat testis and epididymis. *Biology of reproduction*, **64**, 1689–1694.

Reis T, Van Gilst MR, Hariharan IK (2010) A buoyancy-based screen of *Drosophila* larvae for fat-storage mutants reveals a role for Sir2 in coupling fat storage to nutrient availability. *PLoS Genetics*, **6**, e1001206.

Ridley EV, Wong AC-N, Westmiller S, Douglas AE (2012) Impact of the resident microbiota on the nutritional phenotype of *Drosophila melanogaster*. *PLoS one*, **7**, e36765.

Riebler A, Held L, Stephan W (2008) Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics*, **178**, 1817–1829.

Rion S, Kawecki TJ (2007) Evolutionary biology of starvation resistance: what we have learned from *Drosophila*. *Journal of evolutionary biology*, **20**, 1655–1664.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake--stream stickleback population pairs. *Molecular ecology*, **21**, 2852–2862.

Sandrelli F, Tauber E, Pegoraro M *et al.* (2007) A molecular basis for natural selection at the timeless locus in *Drosophila melanogaster*. *Science*, **316**, 1898–1900.

Sarov-Blat L, So WV, Liu L, Rosbash M (2000) The *Drosophila* takeout gene is a novel molecular link between circadian rhythms and feeding behavior. *Cell*, **101**, 647–656.

Schlenke TA, Begun DJ (2003) Natural selection drives *Drosophila* immune system evolution. *Genetics*, **164**, 1471–1480.

Schlicker A, Domingues FS, Rahnenführer J, Lengauer T (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC bioinformatics*, **7**, 302.

Schmidt PS, Paaby AB, Heschel MS (2005) Genetic variance for diapause expression and associated life histories in *Drosophila melanogaster*. *Evolution*, **59**, 2616–2625.

Schwasinger-Schmidt TE, Kachman SD, Harshman LG (2012) Evolution of starvation resistance in *Drosophila melanogaster*: measurement of direct and correlated responses to artificial selection. *Journal of evolutionary biology*, **25**, 378–387.

Sehgal A (2004) *Molecular Biology of Circadian Rhythms*. John Wiley & Sons.

Sekelsky JJ, Newfeld SJ, Raftery LA, Chartoff EH, Gelbart WM (1995) Genetic characterization and cloning of mothers against dpp, a gene required for decapentaplegic function in *Drosophila melanogaster*. *Genetics*, **139**, 1347–1358.

Shannon P, Markiel A, Ozier O *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.

Sharon G, Segal D, Ringo JM *et al.* (2010) Commensal bacteria play a role in mating preference of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 20051–20056.

Slaninova V, Krafcikova M, Perez-Gomez R *et al.* (2016) Notch stimulates growth by direct regulation of genes involved in the control of glycolysis and the tricarboxylic acid cycle. *Open biology*, **6**, 150155.

Stamhuis IH, Meijer OG, Zevenhuizen EJ (1999) Hugo de Vries on heredity, 1889-1903. Statistics, Mendelian laws, pangenes, mutations. *Isis; an international review devoted to the history of science and its cultural influences*, **90**, 238–267.

Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 1245–1253.

Stephan W (2015) Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*, **25** (1), 79-88.

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

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Stone EF, Fulton BO, Ayres JS *et al.* (2012) The circadian clock protein timeless regulates phagocytosis of bacteria in *Drosophila*. *PLoS Pathogens*, **8**, e1002445.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*.

Struhl G, Adachi A (1998) Nuclear access and action of notch in vivo. *Cell*, **93**, 649–660.

Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS one*, **6**, e21800.

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.

Swarup S, Harbison ST, Hahn LE *et al.* (2012) Extensive epistasis for olfactory behaviour, sleep and waking activity in *Drosophila melanogaster*. *Genetics research*, **94**, 9–20.

van Swinderen B, Greenspan RJ (2005) Flexibility in a gene network affecting a simple behavior in *Drosophila melanogaster*. *Genetics*, **169**, 2151–2163.

Tauber E, Zordan M, Sandrelli F *et al.* (2007) Natural selection favors a newly derived timeless allele in *Drosophila melanogaster*. *Science*, **316**, 1895–1898.

Telonis-Scott M, van Heerwaarden B, Johnson TK, Hoffmann AA, Sgrò CM (2013) New levels of transcriptome complexity at upper thermal limits in wild *Drosophila* revealed by exon expression analysis. *Genetics*, **195**, 809–830.

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Research*, **16**, 702–712.

Tzou P, Meister M, Lemaitre B (2002) 27 Methods for studying infection and immunity in *Drosophila*. In: *Molecular Cellular Microbiology*, pp. 507–529. Elsevier.

Unckless RL, Howick VM, Lazzaro BP (2016) Convergent Balancing Selection on an Antimicrobial Peptide in *Drosophila*. *Current Biology*, **26**, 257–262.

Unckless RL, Rottschaefer SM, Lazzaro BP (2015) The complex contributions of genetics and nutrition to immunity in *Drosophila melanogaster*. *PLoS Genetics*, **11**, e1005030.

Vaccaro A, Birman S, Klarsfeld A (2016) Chronic jet lag impairs startle-induced locomotion in *Drosophila*. *Experimental gerontology*, **85**, 24–27.

Vanin S, Bhutani S, Montelli S *et al.* (2012) Unexpected features of *Drosophila* circadian behavioural rhythms under natural conditions. *Nature*, **484**, 371–375.

Vaze KM, Sharma VK (2013) On the adaptive significance of circadian clocks for their owners. *Chronobiology international*, **30**, 413–433.

Vermeulen CJ, Bijlsma R (2004) Characterization of conditionally expressed mutants affecting age-specific survival in inbred lines of *Drosophila melanogaster*: lethal conditions and temperature-sensitive periods. *Genetics*, **167**, 1241–1248.

Vermeulen CJ, Sørensen P, Kirilova Gagalova K, Loeschcke V (2013) Transcriptomic analysis of inbreeding depression in cold-sensitive *Drosophila melanogaster* shows upregulation of the immune response. *Journal of Evolutionary Biology*, **26**, 1890–1902.

Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.

Villemereuil P, Gaggiotti OE (2015) A new $F_{ST}$-based method to uncover local adaptation using environmental variables. *Methods in ecology and evolution / British Ecological Society*, **6**, 1248–1258.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual review of genetics*, **47**, 97–120.

de Vladar HP, Barton N (2014) Stability and response of polygenic traits to stabilizing selection and mutation. *Genetics*, **197**, 749–767.

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

Wade MJ, Goodnight CJ (1998) Perspective: The Theories of Fisher and Wright in the Context of Metapopulations: When Nature Does Many Small Experiments. *Evolution*, **52**, 1537–1553.

Warmke JW, Ganetzky B (1994) A family of potassium channel genes related to *eag* in *Drosophila* and mammals. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 3438–3442.

Weavers H, Prieto-Sánchez S, Grawe F *et al.* (2009) The insect nephrocyte is a podocyte-like cell with a filtration slit diaphragm. *Nature*, **457**, 322–326.

Weber AL, Khan GF, Magwire MM *et al.* (2012) Genome-wide association analysis of oxidative stress resistance in *Drosophila melanogaster*. *PLoS one*, **7**, e34745.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**, 1358–1370.

Whitfield J (2008) Biological theory: Postmodern evolution? *Nature*, **455**, 281–284.

Whitlock MC, Phillips PC (2000) The exquisite corpse: a shifting view of the shifting balance. *Trends in Ecology & Evolution*, **15**, 347–348.

Wilches R, Voigt S, Duchen P, Laurent S, Stephan W (2014) Fine-mapping and selective sweep analysis of QTL for cold tolerance in *Drosophila melanogaster*. *G3: Genes, Genomes, Genetics*, **4**, 1635–1645.

Williams GC (1966) Adaptation and Natural Selection. *Princeton University Press, Princeton, NJ*.

Willing EM, Dreyer C, van Oosterhout C (2012) Estimates of genetic differentiation measured by $F_{ST}$ do not necessarily require large sample sizes when using many SNP markers. *PLoS one*, **7**, e42649.

Winther RG (2001) August Weismann on germ-plasm variation. *Journal of the history of biology*, **34**, 517–555.

Wright S (1932) *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*. Proceedings of the Sixth International Congress of Genetics: 1, pp. 356-366.

Wu Y, Kawasaki F, Ordway RW (2005) Properties of short-term synaptic depression at larval neuromuscular synapses in wild-type and temperature-sensitive paralytic mutants of *Drosophila*. *Journal of Neurophysiology*, **93**, 2396–2405.

Yamamoto A, Anholt RRH, MacKay TFC (2009) Epistatic interactions attenuate mutations affecting startle behaviour in *Drosophila melanogaster*. *Genetics research*, **91**, 373–382.

Yamamoto A, Zwarts L, Callaerts P *et al.* (2008) Neurogenetic networks for startle-induced locomotion in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 12393–12398.

Yeaman S (2015) Local Adaptation by Alleles of Small Effect. *The American naturalist*, **186 Suppl 1**, S74–89.

Yerushalmi S, Green RM (2009) Evidence for the adaptive significance of circadian rhythms. *Ecology letters*, **12**, 970–981.

Zeidler MP, Mlodzik M (1997) six-banded, a novel *Drosophila* gene, is expressed in 6 segmental stripes during embryonic development and in the eye imaginal disc. *Biological Chemistry*, **378**, 1119–1124.

Zhang T, Branch A, Shen P (2013) Octopamine-mediated circuit mechanism underlying controlled appetite for palatable food in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 15431–15436.

CURRICULUM VITAE

# VEDRAN BOŽIČEVIĆ

Bioinformatic data analytics, insights & communications specialist

## Education

October 2012 – June 2017

Doctoral candidate, **LMU München**

- Population genomics, bioinformatics & statistics

October 2002 – June 2008

Undergraduate, Faculty of Science, **University of Zagreb**

- Ecology, evolutionary developmental biology

## Work experience

October 2012 – June 2016

**Biozentrum, LMU München**

Marie Curie Early Stage Researcher

- Gene network architecture and polygenic selection detection in fruit fly next-generation sequencing data

January 2015 – June 2015

**Era7 Bioinformatics, Granada, Spain**

- Research and development, *Bio4j* graph database

January 2012 – September 2012

**Biozentrum, LMU München**

Research Assistant

- Bioinformatic analysis of balancing selection

October 2008 – December 2011

**Ruđer Bošković Institute, Zagreb**

Researcher, Laboratory for Evolutionary Genetics

- Organ systems evolution, bioinformatics & statistics

June 2008 – September 2008

**Valamar Club Tamaris Hotel ****, Lanterna-Poreč**

- Restaurant hotel service, guest hospitality & support

## Skills and expertise

Key competencies

- Outstanding **written** and **verbal communication** skills, acquired during the course of 8+ years of research, workshops, conferences, meetings & volunteer work
- Keen mind for valuable **insights** derived from **complex data**, recognizing meaningful patterns & connections
- Extensive experience in **conceiving**, **organizing**, and **leadership** of projects involving various stakeholders
- Excellent **time management** of multiple concurrent projects, documentation and communication of work & research progress to the European Commission

## Summary of IT skills

| Level | | Experience (years) |
|---|---|---|
| Languages | | |
| ■■■■■■■□□□ | Shell scripting | 4,5 |
| ■■■■■■□□□□ | R (2.13, 2.14) | 4 |
| ■■■■■□□□□□ | Perl (5.10, 5.12) | 4 |
| ■■□□□□□□□□ | Python (2.7) | 0,5 |
| Databases | | |
| ■■□□□□□□□□ | MS Access (2000, 2003) | 0,5 |
| Operating systems | | |
| ■■■■■■■■□□ | Windows (9x, XP, 7, 10) | 20 |
| ■■■■■■□□□□ | Ubuntu (12.04, 14.04) | 4,5 |
| ■■■■■□□□□□ | Mac OS X (10.8 – 10.11) | 4 |
| Tools | | |
| ■■■■■■■■■□ | MS Office | 20 |
| ■■■■■■□□□□ | CorelDraw (X5) | 5 |
| ■■■■■□□□□□ | OpenOffice & LibreOffice | 4,5 |
| ■■■□□□□□□□ | Matlab (R2011a) | 1 |
| ■■■□□□□□□□ | Inkscape (0.91) | 1 |
| ■■□□□□□□□□ | Illustrator | 0,5 |

# Workshops & courses

April 2016 – LMU Munich
- Biobash & Biopython

April 2015 – Bayer AG Gent, Belgium
- Achievements conference
- European policies & career planning
- EU initiatives for entrepreneurship

October 2014 – Natural History Museum London
- Scientific communication, public speaking and presentation skills, and interacting with the press
- Ecological niche modelling

March 2014 – J. Fourier University Grenoble, France
- Bayesian inference & model-based analysis
- Mid-term meeting & progress assessment from European Commission representatives

October 2013 – Vetmeduni Vienna, Austria
- Coalescent theory and analysis of pop. structure
- Using genomic data for quantitative genetics

August 2013 – Era7 Bioinformatics Granada, Spain
- Cloud computing – Amazon Web Services
- Pipelines & configuration of virtual machines for Amazon cloud instances

June 2013 – Qiagen Aarhus, Denmark
- Next-generation sequencing data analysis using CLC Genomics Workbench, Perl software development

February 2013 – LMU Munich & JLU Gießen
- Multivariate statistics & QTL analysis in R
- Techniques in plant breeding and genomics

December 2012 – Queen Mary University London
- Collaboration, communication & negotiation strategies, personal development planning

November 2012 – LMU Munich
- Perl programming for biological data analysis

July 2012 – LMU Munich
- Conflict management, effective problem solving, communication, negotiation & mediation

# Languages

C2 English written & edited scientific publications

C1 Italian CILS certificate, Università di Siena

B2 German Institut für Berufliche Bildung, München

A2 Spanish elementary proficiency

Native Croatian / Bosnian / Serbian

# Volunteering & outreach

- Marie Curie Ambassador: scientific communication with the public (genomics, bioinformatics, evolution)
- ESR & Bioethics committees, Intercrossing network
- Student council, Life Science Munich (LSM) Graduate School, organization of social activities
- Tag der offenen Tür, LMU Munich, 2014 & 2015

# Publications

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

Croze M, Wollstein A, Božičević V, Živković D, Stephan W, Hutter S. *BMC Evolutionary Biology*. 2017 Jan 13;17(1):15.

Population genetic evidence for cold adaptation in European *Drosophila melanogaster* populations

Božičević V, Hutter S, Stephan W, Wollstein A. *Molecular Ecology*. 2016 Mar 1;25(5):1175-91.

Phylostratigraphic profiles reveal deep evolutionary history of the vertebrate head sensory systems

Šestak MS, Božičević V, Bakarić R, Dunjko V, Domazet-Lošo T. *Frontiers in Zoology*. 2013 Apr 12;10(1):18.