Geographic variation of the *for* gene reveals signatures of local adaptation in *Drosophila melanogaster*

Dylan J. Padilla Perez

School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA.

Corresponding author: dpadil10@asu.edu; **Phone:** +1 (480) 646-7769.

Article type: Research article.

Conflict of interest: The author has declared no competing interests.

Data accessibility: A fully reproducible workflow of the data analyses associated with this study, including R scripts and additional supporting material, is available in the following repositories: Github `https://dylan-padilla.github.io/for-evo/`, a Dryad link will be available upon acceptance .

Keywords: foraging mode, landscape genetics, polymorphism, heterogeneous environments, path length.

This manuscript includes:

Abstract: 219 words

Main text: $\approx$ 4,000 words

Figures: 6

References: 50

Abstract

In the early 1980s, the observation that *Drosophila melanogaster* larvae differed in their foraging behavior laid the foundation for the work that would later lead to the discovery of the *for* gene and its associated foraging phenotypes, rover and sitter. Since then, the molecular characterization of the *for* gene and our understanding of the mechanisms that maintain its allelic variants in the laboratory have progressed enormously. However, the significance and dynamics of such variation are yet to be investigated in nature. With the advent of next-generation sequencing, it is now possible to identify loci underlying adaptation of populations in response to environmental variation. Here, I present results of the first genotype-environment association analysis that quantifies variation in the *for* gene among samples of *D. melanogaster* structured across space and time. These samples consist of published genomes of adult flies collected worldwide, and at least twice per site of collection (during spring and fall). An analysis of genetic differentiation based on F_{st} and individual admixture estimates revealed an east-west gradient in population structure. The results indicate that different patterns of gene flow as expected under models of isolation by environment and isolation by distance are likely driving the genetic differentiation among populations. Overall, this study is essential for understanding the mechanisms underlying the evolution of foraging behavior in *D. melanogaster*.

Keywords: foraging mode, landscape genetics, polymorphism, heterogeneous environments, path length.

Introduction

Species distributed over broad geographic ranges face drastic variation in environmental conditions. Such species provide an opportunity to understand the mechanisms underlying evolutionary responses to environmental variation. *Drosophila melanogaster* represents a good example of a species with a widespread distribution range, being found on every continent and most islands (Markow and O'Grady, 2005). Since the early 1900s until the present days, this species has been a central model to understand the genetic basis of adaptation (Flatt, 2020). Identifying genes underlying adaptation of populations is a key issue in evolutionary biology (Tenaillon and Tiffin, 2008). For instance, the foraging gene (*for*) is best known for underlying different foraging strategies adopted by *D. melanogaster* in response to environmental variation (Anreiter and Sokolowski, 2019). Detailed characterization of this gene in laboratory experiments revealed the existence of two allelic variants: rover and sitter (Sokolowski, 1980). In the presence of nutritive yeast, larval rovers travel longer distances and eat less than sitters do. This difference is not evident in the presence of a non-nutritive agar medium (Kaun et al., 2007), suggesting a gene-by-environment interaction (Burns et al., 2012). Similarly, Pereira and Sokolowski (1993) showed that adult rovers walked significantly farther from the food source after eating than did sitters, but there were no differences in walking behavior in the absence of food. This behavioral variation should also be observed in the wild, because food availability varies temporally and spatially across the range of *D. melanogaster*. Because these behavioral differences between rovers and sitters arise from genetic variation in the *for* gene, one should expect the frequency of

alleles in this gene to vary over space and time.

Extensive geographic and temporal variation in various genetic markers has been identified in *D. melanogaster* (Hoffmann and Weeks, 2007; Kapun et al., 2020; Sezgin et al., 2004). In some cases, clinal genetic variants have been directly linked to phenotypic variants (Lee et al., 2013; Paaby et al., 2014, 2010; Schmidt et al., 2008). For example, associations between the *Adh* allozyme locus and development, as well as the *hsp70* gene and heat knockdown resistance have been discovered from a mid-latitude population along the east coast of Australia. There is also good evidence indicating that *D. melanogaster* evolved Bergmann's clines on more than one continent. When raised in a common environment, both South American and Australian flies from high latitudes developed faster and matured at larger sizes than their counterparts from low latitudes (James et al., 1997). Similarly, seasonal stresses at high latitudes, such as overwinter survival and ephemeral food resources, impose strong selection for increased body size and somatic maintenance, whereas warm climates with ample feeding opportunities favor reduced size and decreased maintenance (James et al., 1997; Paaby and Schmidt, 2009). Because the foraging behavior of organisms influences their ability to gather and assimilate food to fuel their maintenance requirements, one may expect the allele frequency of the *for* gene to differ in genotypes living at high latitudes and over stressful time (e.g., winter) relative to those living at lower latitudes and less stressful conditions.

Although recent studies have provided good insights on the maintenance of the *for* gene polymorphism in the laboratory, the significance and dynamics of this variation are yet to be investigated in nature. Studies based on experimental evolution suggest

that density-dependent selection during the larval stage of Drosophila, rather than the adult stage, can maintain genetic variation in foraging behavior. For instance, a high larval density favors the rover variant and low larval density favors the sitter variant (Sokolowski et al., 1997). Similarly, negative frequency-dependent selection promotes the increase in frequency of one variant when it is rare relative to the other variant (Fitzpatrick et al., 2007). In addition, Anreiter et al. (2017) showed evidence of an epigenetic regulator of the *for* gene (the *G9a* methyltransferase) responsible for rover-sitter differences in adult foraging behavior. Specifically, they showed that rovers bearing *G9a* alleles, had significantly greater foraging success than sitters, but this rover–sitter difference disappeared when they carried *G9a* null alleles. In sum, researchers have now detailed advances on the mechanisms that produce variation in the Drosophila *for* gene under controlled conditions. Yet, our understanding of the environmental drivers of the genetic variation that controls the foraging behavior of flies in nature remains unknown. A good approach to fill this gap is referred to as genotype-environment association analysis (Joost et al., 2007; Rellstab et al., 2015), an emerging field that integrates tools from landscape genetics and population genomics to identify the environmental factors that have shaped present-day genetic variation, and the gene variants that drive local adaptation.

Here, I present results of the first genotype-environment association analysis that quantifies variation in the *for* gene among samples of *D. melanogaster* structured across space and time. These samples consist of published genomes of adult flies collected worldwide, and at least twice per site of collection (during spring and fall). Thus, not only could I test whether the allele frequency of the *for* gene change geographically, but also in response to seasonality. I found a distinct genetic structure among the

samples collected in America and Europe. An analysis of genetic differentiation based on F_{st} estimates revealed two major genetic clusters; one with all the American samples grouped together, and the other one grouping the European samples. Similarly, there is some degree of genetic differentiation among samples within each continent, suggesting a total of 6 ancestral populations (k) with unique patterns of allelic variation in the *for* gene. The results indicate that different patterns of gene flow as expected under models of isolation by environment (IBE) and isolation by distance (IBD) are likely driving the genetic differentiation among populations of *D. melanogaster*. Overall, this study is essential for understanding the mechanisms underlying the evolution of foraging behavior in *D. melanogaster*.

Materials and Methods

Data source

The data used in this study represent a subset of a dataset assembled by the Drosophila Genome Nexus project (DGN), the European Drosophila Population Genomics Consortium (DrosEU), and the Real Time Evolution Consortium (DrosRTEC). This dataset constitutes the most comprehensive spatiotemporal sampling of *D. melanogaster* populations to date, including estimates of genome-wide allele frequencies for more than 270 populations. Collectively, the data are based on sequencing of more than 13,000 flies collected around the world over multiple years (Kapun et al., 2021). Importantly, the analyses of this study relied on pooled samples stored in the Genomic Data Structure (GDS) file named "dest.all.PoolSNP.001.50.10Nov2020.ann.gds", available at `https://dest.bio/data-files/SNP-tables`. Pooled samples from this file consist of

33-40 wild-caught males, which are more easily distinguishable morphologically from similar species than females are. Detailed information on quality control, mapping, SNP calling and dataset merging can be obtained from the DEST dataset mapping pipeline `https://dest.bio/sourcecode`.

I filtered the GDS object to restrict the focus on Single Nucleotide Polymorphism variants (SNPs) in the *for* gene. As annotated at `https://flybase.org/` (genome assembly version 6.32), the *for* gene is found on the left arm of chromosome 2 from base pairs 3,622,074-3,656,953. I further filtered the reads to get a subset of samples collected at least once during spring and winter. Because the filtering process drastically reduced the number of individuals (pools) in a few populations, I grouped together populations from the nearest neighbor localities to ensure a minimum sample size of 5. I also applied a mean effective coverage N_{eff} filter of 28 to avoid the influence of atypical observations driven by samples with low coverage (Nunez et al., 2022). The DEST dataset provides the coordinates of each collection site, enabling one to obtain data about the seasonality of temperature and precipitation. These climatic variables were derived from the monthly values of temperature and rainfall for global land areas at a resolution of 30 s (Fick and Hijmans, 2017). In addition, I extracted log-transformed values of net primary production (NPP)—the net amount of solar energy converted to plant organic matter through photosynthesis—measured in units of elemental carbon. This variable represents the primary source of trophic energy for the world's ecosystems (Imhoff et al., 2004).

Quality control

I removed pools and SNP loci with more than 20% missing genotypes. To do so, I used the *missingno* function from the "poppr" package of R v2.9.3 Kamvar et al. (2014). Subsequently, I tested for deviations from Hardy–Weinberg equilibrium (HWE) using the function *hw.test* from the "pegas" package of R v1.1 Paradis (2010). This function calculated p-values of an exact test with a Monte Carlo permutation of alleles (1,000 replicates). I considered loci to be out of HWE if they deviated significantly in more than 50% of populations. This decision was based on a "false discovery rate" correction of the p-values.

Genetic differentiation and population structure

To test for genetic differentiation among populations at the level of the *for* gene, I estimated pairwise values of F_{st} adopting the method suggested by Weir and Cockerham (1984). I used the function *gl.fst.pop* from the package "dartR" v2.7.2 (Gruber et al., 2018), running 1,000 bootstraps across all loci to generate 95% confidence intervals of the values of F_{st}. When the confidence intervals did not include zero, the F_{st} value in question was considered significantly different from zero. To explore patterns of isolation by environment (IBE) and isolation by distance (IBD), I first performed a Partial Mantel test on genetic distance (F_{st}) and primary production distance, holding geographic distance (coordinates) constant. I then performed the same test but holding the primary production distance matrix constant to account for the effect of geographic distance. To do this, I used the *partial.mantel* function from the R package "vegan" v2.6-4, which ran a Monte-Carlo test based on 1,000 permutations and assessed the significance of the correlation between the distance matrices. I also eval-

uated the relationship between these distances with a Mantel correlogram, a method that does not make the assumption of linearity. For this test, I ran the function *mgram* from the R package "ecodist" v2.0.7 (Goslee and Urban, 2007), setting the number of permutations to 1,000 and using the Sturge's rule to determine an appropriate number of classes.

I estimated population structure using two classical approaches: an admixture analysis and a principal component analysis (PCA). I performed the admixture analysis using a sparse negative matrix factorization method (SNMF), available in the R package "LEA" v3.2.0 (Frichot and François, 2015). Similar to Bayesian clustering programs such as CLUSTER, the function *snmf* in LEA estimates individual admixture coefficients from the genomic matrix. Assuming k ancestral populations, the function provides least-squares estimates of ancestry proportions and estimates an entropy criterion to evaluate the quality of fit of the statistical model through cross-validation. The entropy criterion enabled one to choose the number of ancestral populations that best explained the genomic data. Specifically, I used the function *snmf* in LEA to compute individual ancestry coefficients for a number of ancestral populations (k), ranging between $k = 1$ and $k = 10$. For each value of k, I ran the algorithm 50 times and calculated the cross-entropy for each run. In addition, I performed a PCA on the genomic and environmental data using the function *rda* from the R package "vegan" v2.6-4. This function computes the scores of a PCA and returns a scree plot for the eigenvalues of the sample covariance matrix. I identified the number of significant components for both the genomic and the environmental matrices based on a broken stick assessment. According to the broken-stick stopping rule, principal components should be retained as long as observed eigenvalues are higher than corresponding ran-

dom broken stick components. Because the number of ancestral populations is closely linked to the number of principal components that explain variation in the genomic data, this method provides a good complement to the admixture analysis, and expands one's ability to find the most likely value of k.

Detecting adaptation

I also implemented two approaches to detect selection signatures in the data: 1) a differentiation-based method based on the OutFLANK algorithm v0.2 (Whitlock and Lotterhos, 2015), and 2) a genotype-environment association method based on Latent Factor Mixed Models with ridge penalties (Jumentier et al., 2022). OutFLANK calculates a likelihood on a trimmed distribution of F_{st} values to infer the distribution of F_{st} for neutral markers. This distribution is then used to assign q-values to each locus to detect outliers that may be due to spatially heterogeneous selection. The LFMM method is a univariate test that builds a model for each SNP and each predictor variable. Thus, it ran $1,550$ SNPs $\times 1$ synthetic predictor $= 1,550$ separate LFMM tests. This method requires an estimate of the number of populations in the data (k). To determine the most likely value of k, I used both an admixture analysis and a PCA, as discussed earlier. Because these analyses found support for multiple k values, I ran LFMMs varying k accordingly. For each model fitted, I used the genomic inflation factor (GIF) to select the one that best accounted for population structure. I also applied a False Discovery Rate (FDR) of 0.05 by converting the p-values to q-values. The FDR is the expected proportion of false positives among the list of positive tests (Storey and Tibshirani, 2003). For example, an FDR threshold of 0.05 applied to a dataset with an ideal p-value distribution would produce 5% false positives (false

discoveries) among the set of positive tests. However, if the distribution of p-values deviates from the ideal distribution, this same FDR threshold of 0.05 would produce more or fewer false discoveries among the set of positive tests, depending on the skew of the distribution.

Results

In general, none of the loci nor populations were consistently out of Hardy-Weinberg equilibrium. I filtered out of the analyses only one sample that had more than 20% of missing genotypes and 62 loci that had more than 20% of missing data. The quality control procedure generated a new dataset of 1,550 loci (SNPs), and 3,100 alleles from 151 pooled samples distributed among 15 populations.

I found an overall F_{st} value of 0.032, but the pairwise differences in F_{st} differences ranged from 0 to 0.069. Interestingly, the results revealed two major clusters separating the samples collected in America from those collected in Europe (Figure 1). However, significant differentiation was also observed within each continent. The IBE and IBD analyses based on the Partial Mantel test showed a clear increase of genetic distance with NPP distance (Mantel statistic $r = 0.361$, $p - value < 0.001$) and geographic distance (Mantel statistic $r = 0.926$, $p - value < 0.001$). However, a smoothed local mean (read line), indicated that these relationships might not be linear (Figure 2a-b). Importantly, the Mantel correlograms supported the same pattern; populations are relatively similar at short distances, while this similarity decreases with distance (Figure 2c-d).

Figure 1. Heatmap depicting genetic differentiation among populations of *D. melanogaster* based on pairwise F_{st} values. In some cases, samples from adjacent localities were combined to avoid low sample sizes. Abbreviations are as follows: PA = Pennsylvania; MI-WI = Michigan and Wisconsin; NY-MA = New York and Massachusetts.

Figure 2. Patterns of isolation by distance and isolation by environment according to Partial Mantel tests and Correlograms. **(a-b)** Geographic and net primary production distances as a function of genetic distance. Colors represent estimated probability densities and the red line a smoothed local mean. **(c-d)** Mantel correlation at different distance classes as a function of geographic and net primary production distances. Filled dots indicate significant correlations.

Despite the observed genetic differentiation based of the F_{st} values, the OutFLANK algorithm revealed no evidence of selection. The distribution of SNPs showed no outliers that might have arisen due to spatially heterogeneous selection (Figure 3).

Figure 3. Manhattan plot of the distribution of F_{st} loci as a function of their heterozygosity. No F_{st} loci showed significant deviation from the genetic background.

In addition, the genetic admixture analysis suggested a pattern of genetic clusters that approximates that observed from the F_{st} values. The *snmf* run with $k = 9$ best accounted for population structure as it produced the lowest cross-entropy value (Figure 4a). Overall, I observed a distinct east-west gradient in population structure according to the admixture proportions (Figure 4b). This gradient seems to be stronger

among the samples collected in America than those collected in Europe, although the population structure of samples from Ukraine and Russia stands out (Figure 5).

Figure 4. Population structure analysis based on individual ancestry coefficients for a number of ancestral populations. **(a)** Cross-entropy values for each *snmf* run with k ranging between $k = 1$ and $k = 10$. The red arrow indicates the most likely value of k. **(b)** Admixture proportion across populations of *D. melanogaster*. Colors indicate genetic clusters.

Figure 5. Mean admixture proportions across populations of *D. melanogaster* surveyed in America and Europe. Colors in the pie charts represent genetic clusters.

In contrast to the *snmf* method, a PCA of the genomic data suggested that PC1 and PC2 explained more than the random broken stick components. This means that $k = 2$ best accounted for population structure among all samples based on a PCA assessment. Similarly, a PCA of the environmental variables showed that only the PC1 explained more than the broken stick components, with longitude, latitude, and NPP contributing to most of the variation. Accordingly, I selected the PC1 as a synthetic predictor variable to be used in the genotype-environment association analysis. Based on the genetic admixture analysis and the PCA, I extracted locus-specific z-scores and p-values from latent factor mixed models with k ranging between 2 and 9. The GIF and adjusted p-values indicated that a good choice for the number of latent factors was $k = 6$ (Figure 6a). Using $k = 6$, the default GIF correction, and an FDR threshold of

0.05, I detected 2 candidate SNPs under selection in response to the PC1 environmental predictor (Figure 6b).

Figure 6. Genotype-environment association test based on a latent factor mixed model with ridge penalty. **(a)** Distribution of adjusted p-values using the default genomic inflation factor. **(b)** Manhattan plot of loci (SNPs) potentially affected by the PC1 predictor variable. Loci highlighted in red were considered to be under selection according to a False Discovery Rate of 0.05.

Discussion

This study shows that populations of *D. melanogaster* are geographically differentiated based on SNP variants of the foraging gene (*for*). The significant correlation between genetic distance with both primary production distance and geographic distance indicates that genetic differentiation between populations is consistent with signatures of IBE and IBD. Accordingly, gene flow may be strong among populations that occur in close proximity and exploiting similar environments. Different patterns of gene flow are expected under IBE and IDB models (Sexton et al., 2014). One scenario predicts that with limited dispersal and no selection, genetic drift would cause populations to become more differentiated at greater distances, and that this process should be more pronounced as mean population size decreases (Wright, 1943). Thus, under strict IBD, distance predicts differentiation as a result of dispersal limitation and drift, irrespective of environmental differences. A second scenario predicts gene flow to be strong among similar environments to a greater extent than predicted under IBD. However, adaptation to local environments can disrupt patterns of IBD, and subsequently IBE

could arise if maladapted immigrants from different environments are selected against, causing strong barriers to gene flow. A good example of this scenario comes from plants growing on and near mine tailings, where adaptation to different soil types has occurred (Antonovics, 1968). McNeilly and Antonovics (1968) showed that reproductive barriers arose through differences in flower bud development as a result of local adaptation to different soil types. Interestingly, a pattern of isolation by geographical and environmental factors has been reported recently in *D. melanogaster* (Yue et al., 2021). Using microsatellite genotypes, the authors showed an effect of precipitation on patterns of genetic variation across a latitudinal gradient. Similarly, a significant correlation between pairwise F_{st} and geographic distance was discussed in the context of IBD in a study that examined putatively neutral SNPs, mitochondrial haplotypes, as well as inversion and transposable element insertion polymorphisms (Kapun et al., 2020).

Although IBE and IBD may arise from local adaptation and gene flow, the outlier analysis based on genetic differentiation (OutFLANK) showed no signatures of selection. This method is based on an idea that loci influenced by directional selection will show larger genetic differentiation than neutral loci, while the opposite is true for loci subject to balancing selection (Lewontin and Krakauer, 1973). Thus, loci that exhibit unusually high or low F_{st} are good candidates for being influenced by selection. OutFLANK, however, might lead to unreliable results as it is not accurate for the left tail of the distribution of F_{st}—loci subject to spatially uniform balancing selection. Furthermore, OutFLANK has not been formally tested in genomic analyses involving pool-seq data, such as those analyzed in this study (Whitlock and Lotterhos, 2015). The absence of selection as evident by the results of this method should then

be interpreted with caution. Still, the cost-efficient pool-seq data strategy is widely applied to different fields of genomics such as the genotype-environment association analyses (Fischer et al., 2013; Frachon et al., 2018). While SNP calling from pools has the difficulty of distinguishing rare variants from sequencing errors, this limitation is often overcome by establishing a minimum threshold allele frequency for the analysis (Guirao-Rico and González, 2021). In hopes of mitigating this issue, I filtered out the reads with an effective coverage (N_{eff}) lower than 28 in this study. I applied this filter based on a previous work that used some of the samples included in this analysis (Nunez et al., 2022). This study suggested that the use of samples with $N_{eff} < 28$ created outliers in a principal component analysis driven by N_{eff}, but samples with $N_{eff} > 28$ no longer generated atypical observations. As a result, I avoided having a high rate of false positive variants in the analyses. On the basis of its more accurate allele frequency estimation at a lower cost than sequencing of individuals, I highlight the versatility and efficacy of pool-seq data in facilitating a broad range of genomic analyses.

In contrast to OutFLANK, the outlier analysis based on the correlation of genetic variation with environmental data showed evidence of two loci under selection. The implementation of this method involved the use of a genotype–environment association analysis, which identifies loci that show strong correlations with one or more environmental variables, and those loci are interpreted as potentially under selection. Accordingly, I modeled the effect of the environment on genetic variation in the *for* gene using latent factor mixed models. This approach accounts for confounding effects, when the environment is correlated with population structure. Importantly, the results suggested that NPP is one of the main drivers of genetic variation in the for gene

among populations of *D. melanogaster*. NPP may be a good measure of food availability in nature as it constitutes the primary energy source for Earth's ecosystems and complex food webs by supplying food energy to the planet's heterotrophic organisms, such as *D. melanogaster* (Imhoff and Bounoua, 2006). If NPP varies spatially and temporally across the range of this species, genotypes better able to find food might prevail in environments with low NPP. Yet, organisms can adjust their foraging behavior to meet environmental changes in food availability (Schoener, 1971). For example, northern pikes and white-tailed deers may reduce food intake during winter, when food availability is low (Johnson, 1966; Silver et al., 1969). However, food shortages in nature may often favor increased mobility to find new food sources. Indeed, some evidence in *D. melanogaster* indicate that flies become highly active when deprived of food (Williams et al., 2004). In the context of the rover-sitter variants, a mechanism known as balancing selection could explain the maintenance of these variants if they are exposed to NPP heterogeneity in space and time. Although this mechanism seems plausible in Drosophila, a recent study suggested that strong balancing selection at the *for* gene beginning recently (i.e., following migration out of Africa) is unlikely (Turner et al., 2015).

Currently, a number of studies have demonstrated that the foraging behavior of *D. melanogaster* can be strongly influenced by the *for* gene. However, none of them have explored the environmental drivers of variation in this gene based on a genotype-environment association study. To my knowledge, this study provides the first evidence of variation in the *for* gene mediated by patterns of gene flow as expected under models of IBE and IBD. Our analyses also revealed the importance of using different approaches to detect potential signatures of selection. Although all approaches used in

this study suffer from either low power (differentiation-based methods) or high false-positive rates (genotype–environment association methods), a strategy based on the use of both types of methods seems to lead to reliable identification of outlier loci. Field studies could disentangle the contributions of selection and environmental structuring to patterns of gene flow. For example, conducting an experimental transplant study to understand how dispersal and habitat-specific selection interact to influence populations occupying heterogeneous environments. Although IBE and strong selection should promote local genetic adaptation, it is unclear how often this is the case, particularly in Drosophila.

Acknowledgements

I thank Drs Michael Angilletta and Joaquin Nunez for providing helpful comments on earlier versions of this study.

Data accessibility

A fully reproducible workflow of the data analyses associated with this study, including R scripts and additional supporting material, is available in the following repositories: Github `https://dylan-padilla.github.io/for-evo/`, a Dryad link will be available upon acceptance .

Conflict of interest

The author has declared no competing interests.

References

Anreiter, I., Kramer, J. M., and Sokolowski, M. B. (2017). Epigenetic mechanisms modulate differences in drosophila foraging behavior. *Proceedings of the National Academy of Sciences*, 114(47):12518–12523.

Anreiter, I. and Sokolowski, M. B. (2019). The foraging gene and its behavioral effects: Pleiotropy and plasticity. *Annual review of genetics*, 53:373–392.

Antonovics, J. (1968). Evolution in closely adjacent plant populations vi. manifold effects of gene flow. *Heredity*, 23(4):508–524.

Burns, J. G., Svetec, N., Rowe, L., Mery, F., Dolan, M. J., Boyce, W. T., and Sokolowski, M. B. (2012). Gene–environment interplay in drosophila melanogaster: Chronic food deprivation in early life affects adult exploratory and fitness traits. *Proceedings of the National Academy of Sciences*, 109(supplement_2):17239–17244.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12):4302–4315.

Fischer, M. C., Rellstab, C., Tedder, A., Zoller, S., Gugerli, F., Shimizu, K. K., Holderegger, R., and Widmer, A. (2013). Population genomic footprints of selection and associations with climate in natural populations of arabidopsis halleri from the alps. *Molecular ecology*, 22(22):5594–5607.

Fitzpatrick, M. J., Feder, E., Rowe, L., and Sokolowski, M. B. (2007). Maintaining a behaviour polymorphism by frequency-dependent selection on a single gene. *Nature*, 447(7141):210–212.

Flatt, T. (2020). Life-history evolution and the genetics of fitness components in drosophila melanogaster. *Genetics*, 214(1):3–48.

Frachon, L., Bartoli, C., Carrère, S., Bouchez, O., Chaubet, A., Gautier, M., Roby, D., and Roux, F. (2018). A genomic map of climate adaptation in arabidopsis thaliana at a micro-geographic scale. *Frontiers in plant science*, 9:967.

Frichot, E. and François, O. (2015). Lea: An r package for landscape and ecological association studies. *Methods in Ecology and Evolution*, 6(8):925–929.

Goslee, S. C. and Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, 22:1–19.

Gruber, B., Unmack, P. J., Berry, O. F., and Georges, A. (2018). dartr: An r package to facilitate analysis of snp data generated from reduced representation genome sequencing. *Molecular ecology resources*, 18(3):691–699.

Guirao-Rico, S. and González, J. (2021). Benchmarking the performance of pool-seq snp callers using simulated and real sequencing data. *Molecular Ecology Resources*, 21(4):1216–1229.

Hoffmann, A. A. and Weeks, A. R. (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.

Imhoff, M. L. and Bounoua, L. (2006). Exploring global patterns of net primary production carbon supply and demand using satellite observations and statistical data. *Journal of Geophysical Research: Atmospheres*, 111(D22).

Imhoff, M. L., Bounoua, L., Ricketts, T., Loucks, C., Harriss, R., and Lawrence, W. T. (2004). Global patterns in human consumption of net primary production. *Nature*, 429(6994):870–873.

James, A. C., Azevedo, R. B., and Partridge, L. (1997). Genetic and environmental responses to temperature of drosophila melanogaster from a latitudinal cline. *Genetics*, 146(3):881–890.

Johnson, L. (1966). Experimental determination of food consumption of pike, esox lucius, for growth and maintenance. *Journal of the Fisheries Board of Canada*, 23(10):1495–1505.

Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G., and Taberlet, P. (2007). A spatial analysis method (sam) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular ecology*, 16(18):3955–3969.

Jumentier, B., Caye, K., Heude, B., Lepeule, J., and François, O. (2022). Sparse latent factor regression models for genome-wide and epigenome-wide association studies. *Statistical Applications in Genetics and Molecular Biology*, 21(1).

Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014). Poppr: an r package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2:e281.

Kapun, M., Barrón, M. G., Staubach, F., Obbard, D. J., Wiberg, R. A. W., Vieira, J., Goubert, C., Rota-Stabelli, O., Kankare, M., Bogaerts-Márquez, M., et al. (2020). Genomic analysis of european drosophila melanogaster populations reveals longi-

tudinal structure, continent-wide selection, and previously unknown dna viruses. *Molecular Biology and Evolution*, 37(9):2661–2678.

Kapun, M., Nunez, J. C., Bogaerts-Márquez, M., Murga-Moreno, J., Paris, M., Outten, J., Coronado-Zamora, M., Tern, C., Rota-Stabelli, O., Guerreiro, M. P. G., et al. (2021). Drosophila evolution over space and time (dest): a new population genomics resource. *Molecular biology and evolution*, 38(12):5782–5805.

Kaun, K. R., Riedl, C. A., Chakaborty-Chatterjee, M., Belay, A. T., Douglas, S. J., Gibbs, A. G., and Sokolowski, M. B. (2007). Natural variation in food acquisition mediated via a drosophila cgmp-dependent protein kinase. *Journal of Experimental Biology*, 210(20):3547–3558.

Lee, S. F., Eyre-Walker, Y. C., Rane, R. V., Reuter, C., Vinti, G., Rako, L., Partridge, L., and Hoffmann, A. A. (2013). Polymorphism in the neurofibromin gene, nf1, is associated with antagonistic selection on wing size and development time in drosophila melanogaster. *Molecular ecology*, 22(10):2716–2725.

Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74(1):175–195.

Markow, T. A. and O'Grady, P. (2005). *Drosophila: a guide to species identification and use*. Elsevier.

McNeilly, T. and Antonovics, J. (1968). Evolution in closely adjacent plant populations iv. barriers to gene flow. *Heredity*, 23(2):205–218.

Nunez, J. C., Lenhart, B. A., Bangerter, A., Murray, C. S., Yu, Y., Nystrom, T. L.,

Tern, C., Erickson, P. A., and Bergland, A. O. (2022). A cosmopolitan inversion drives seasonal adaptation in overwintering drosophila. *bioRxiv*, pages 2022–12.

Paaby, A. B., Bergland, A. O., Behrman, E. L., and Schmidt, P. S. (2014). A highly pleiotropic amino acid polymorphism in the drosophila insulin receptor contributes to life-history adaptation. *Evolution*, 68(12):3395–3409.

Paaby, A. B., Blacket, M. J., Hoffmann, A. A., and Schmidt, P. S. (2010). Identification of a candidate adaptive polymorphism for drosophila life history by parallel independent clines on two continents. *Molecular ecology*, 19(4):760–774.

Paaby, A. B. and Schmidt, P. S. (2009). Dissecting the genetics of longevity in drosophila melanogaster. *Fly*, 3(1):29–38.

Paradis, E. (2010). pegas: an r package for population genetics with an integrated–modular approach. *Bioinformatics*, 26(3):419–420.

Pereira, H. S. and Sokolowski, M. B. (1993). Mutations in the larval foraging gene affect adult locomotory behavior after feeding in drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 90(11):5044–5046.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, 24(17):4348–4370.

Schmidt, P. S., Zhu, C.-T., Das, J., Batavia, M., Yang, L., and Eanes, W. F. (2008). An amino acid polymorphism in the couch potato gene forms the basis for climatic adaptation in drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 105(42):16207–16211.

Schoener, T. W. (1971). Theory of feeding strategies. *Annual review of ecology and systematics*, 2(1):369–404.

Sexton, J. P., Hangartner, S. B., and Hoffmann, A. A. (2014). Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution*, 68(1):1–15.

Sezgin, E., Duvernell, D. D., Matzkin, L. M., Duan, Y., Zhu, C.-T., Verrelli, B. C., and Eanes, W. F. (2004). Single-locus latitudinal clines and their relationship to temperate adaptation in metabolic genes and derived alleles in drosophila melanogaster. *Genetics*, 168(2):923–931.

Silver, H., Colovos, N., Holter, J., and Hayes, H. (1969). Fasting metabolism of white-tailed deer. *The Journal of Wildlife Management*, pages 490–498.

Sokolowski, M. B. (1980). Foraging strategies of drosophila melanogaster: a chromosomal analysis. *Behavior genetics*, 10:291–302.

Sokolowski, M. B., Pereira, H. S., and Hughes, K. (1997). Evolution of foraging behavior in drosophila by density-dependent selection. *Proceedings of the national academy of sciences*, 94(14):7373–7377.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Tenaillon, M. I. and Tiffin, P. L. (2008). The quest for adaptive evolution: a theoretical challenge in a maze of data. *Current opinion in plant biology*, 11(2):110–115.

Turner, T., Giauque, C. C., Schrider, D. R., and Kern, A. D. (2015). Behavioral

variation in drosophila melanogaster: no evidence for common alleles of large-effect at the foraging gene in a population from north carolina, usa. *USA. bioRxiv*, 4325.

Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *evolution*, pages 1358–1370.

Whitlock, M. C. and Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of f st. *The American Naturalist*, 186(S1):S24–S36.

Williams, A. E., Rose, M. R., and Bradley, T. J. (2004). The respiratory pattern in drosophila melanogaster selected for desiccation resistance is not associated with the observed evolution of decreased locomotory activity. *Physiological and Biochemical Zoology*, 77(1):10–17.

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114.

Yue, L., Cao, L.-J., Chen, J.-C., Gong, Y.-J., Lin, Y.-H., Hoffmann, A. A., and Wei, S.-J. (2021). Low levels of genetic differentiation with isolation by geography and environment in populations of drosophila melanogaster from across china. *Heredity*, 126(6):942–954.

Figures

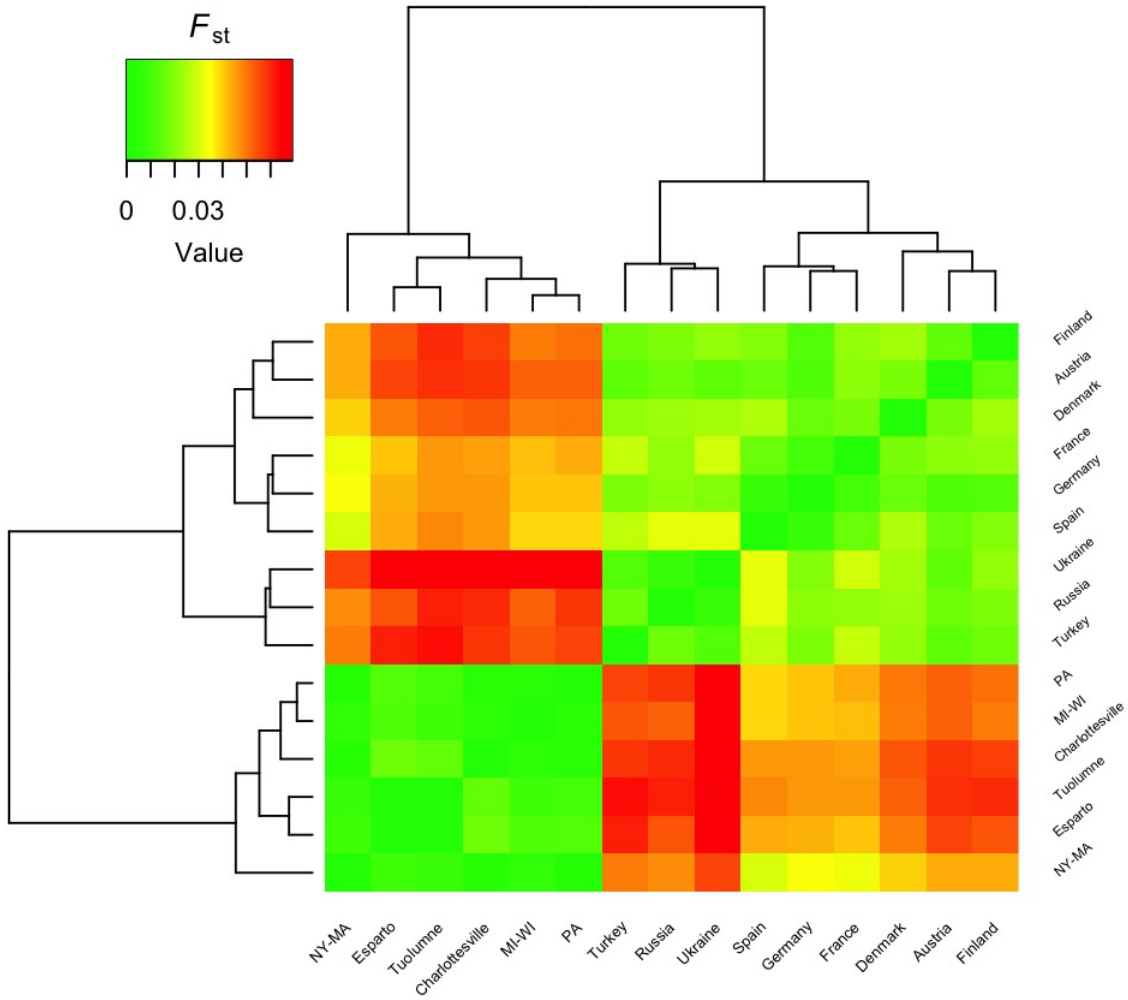

Figure 1: Heatmap depicting genetic differentiation among populations of *D. melanogaster* based on pairwise F_{st} values. In some cases, samples from adjacent localities were combined to avoid low sample sizes. Abbreviations are as follows: PA = Pennsylvania; MI-WI = Michigan and Wisconsin; NY-MA = New York and Massachusetts.

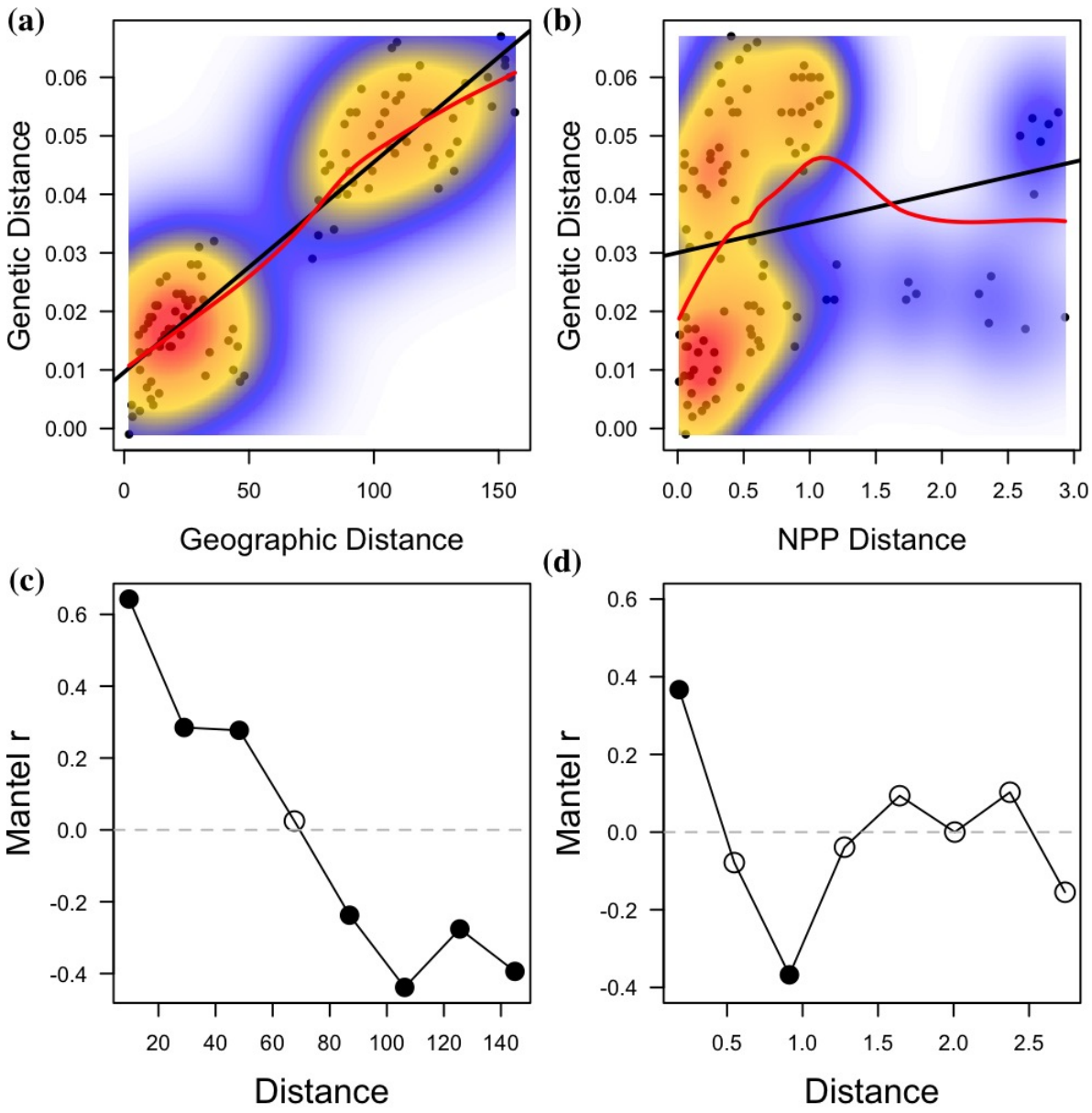

Figure 2: Patterns of isolation by distance and isolation by environment according to Partial Mantel tests and Correlograms. **(a-b)** Geographic and net primary production distances as a function of genetic distance. Colors represent estimated probability densities and the red line a smoothed local mean. **(c-d)** Mantel correlation at different distance classes as a function of geographic and net primary production distances. Filled dots indicate significant correlations.

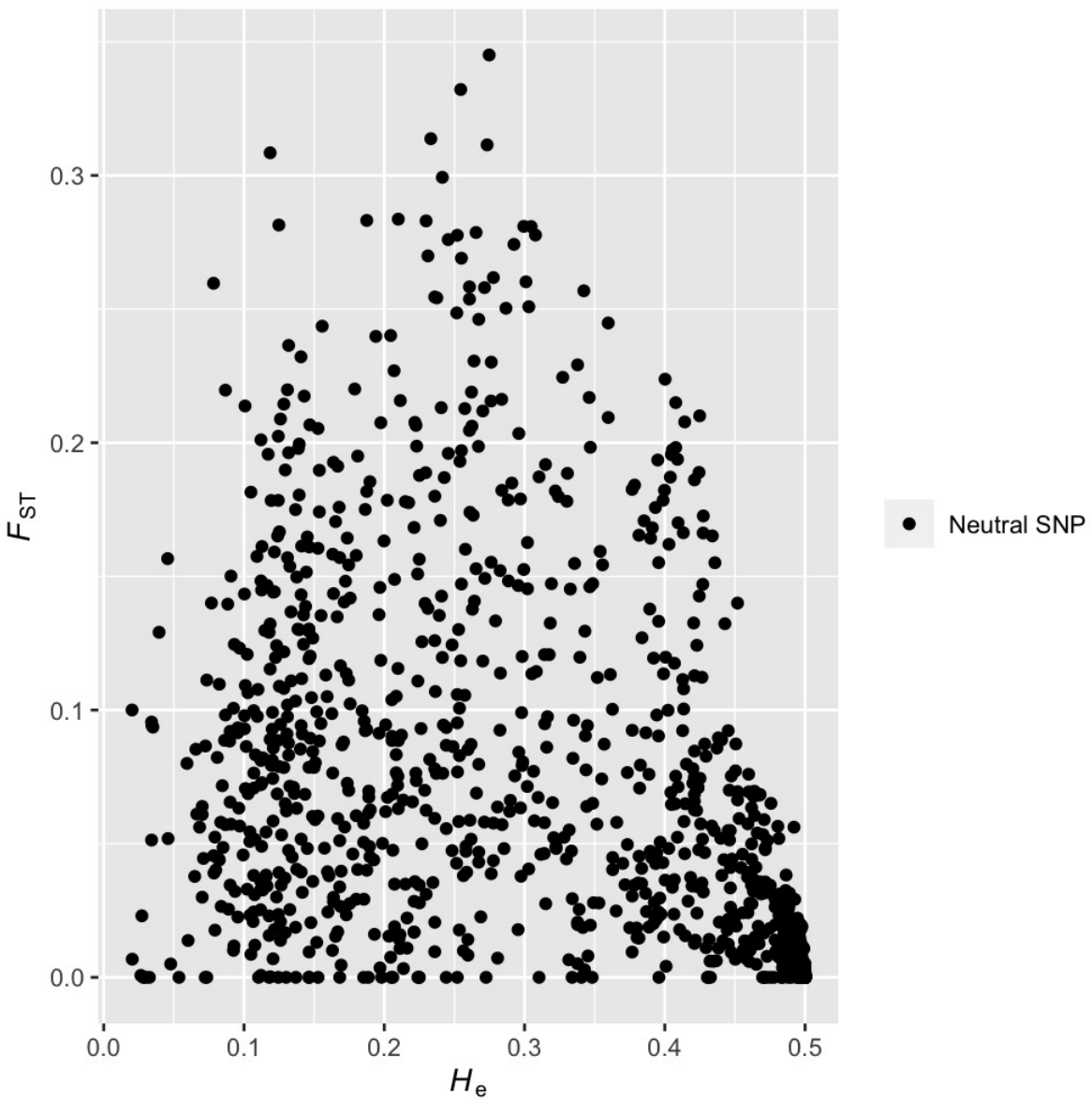

Figure 3: Manhattan plot of the distribution of F_{st} loci as a function of their heterozygosity. No F_{st} loci showed significant deviation from the genetic background.

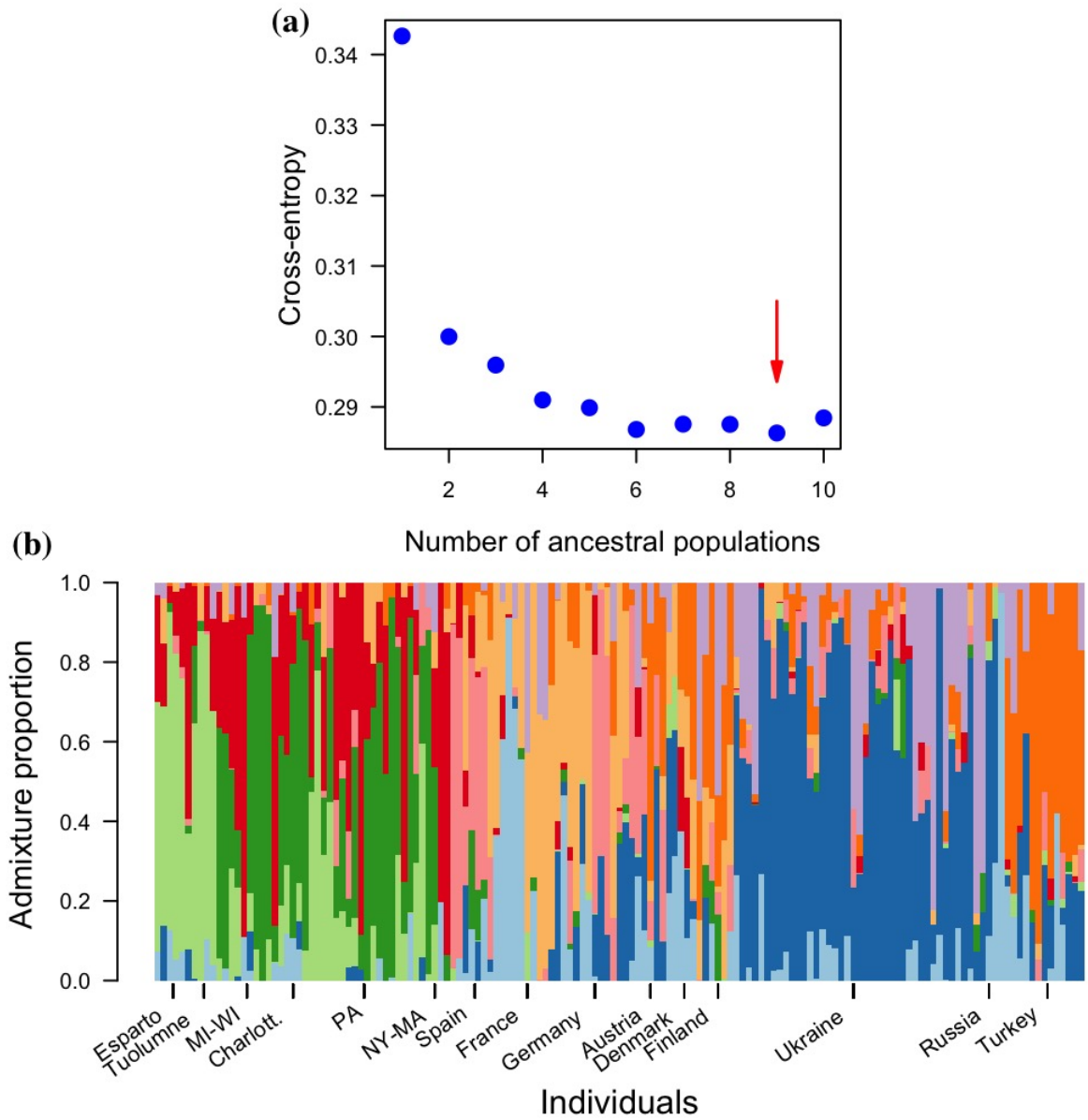

Figure 4: Population structure analysis based on individual ancestry coefficients for a number of ancestral populations. **(a)** Cross-entropy values for each *snmf* run with k ranging between $k = 1$ and $k = 10$. The red arrow indicates the most likely value of k. **(b)** Admixture proportion across populations of *D. melanogaster*. Colors indicate genetic clusters.

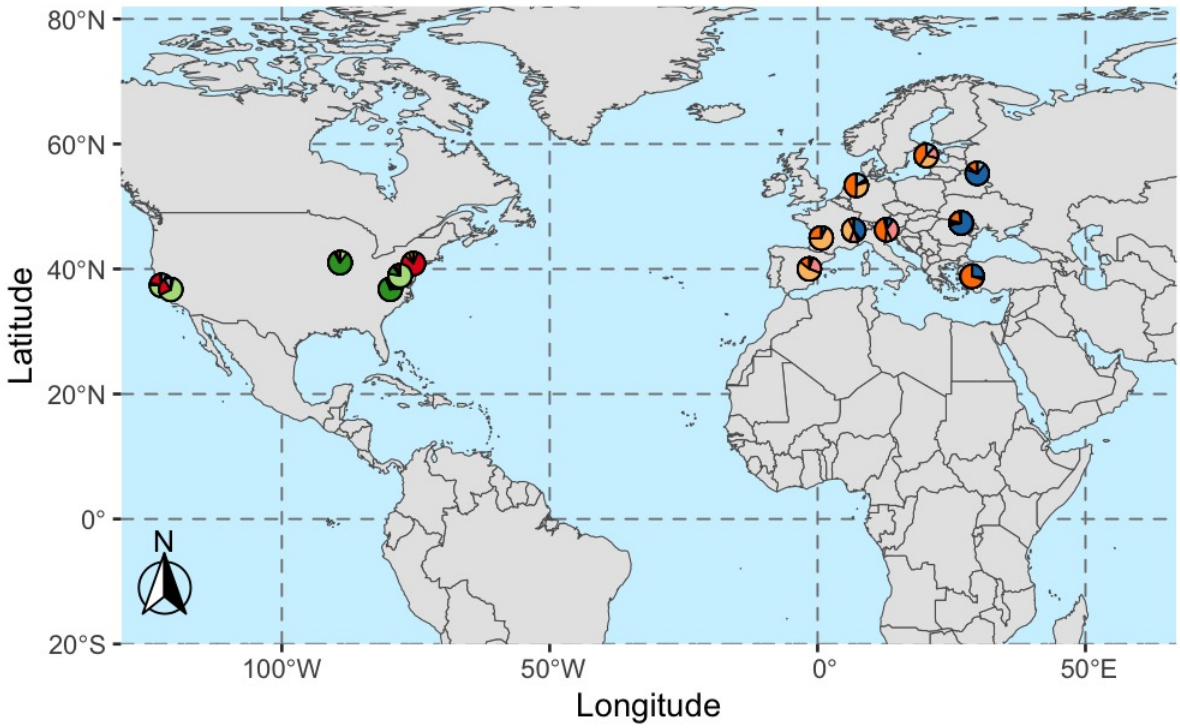

Figure 5: Mean admixture proportions across populations of *D. melanogaster* surveyed in America and Europe. Colors in the pie charts represent genetic clusters.

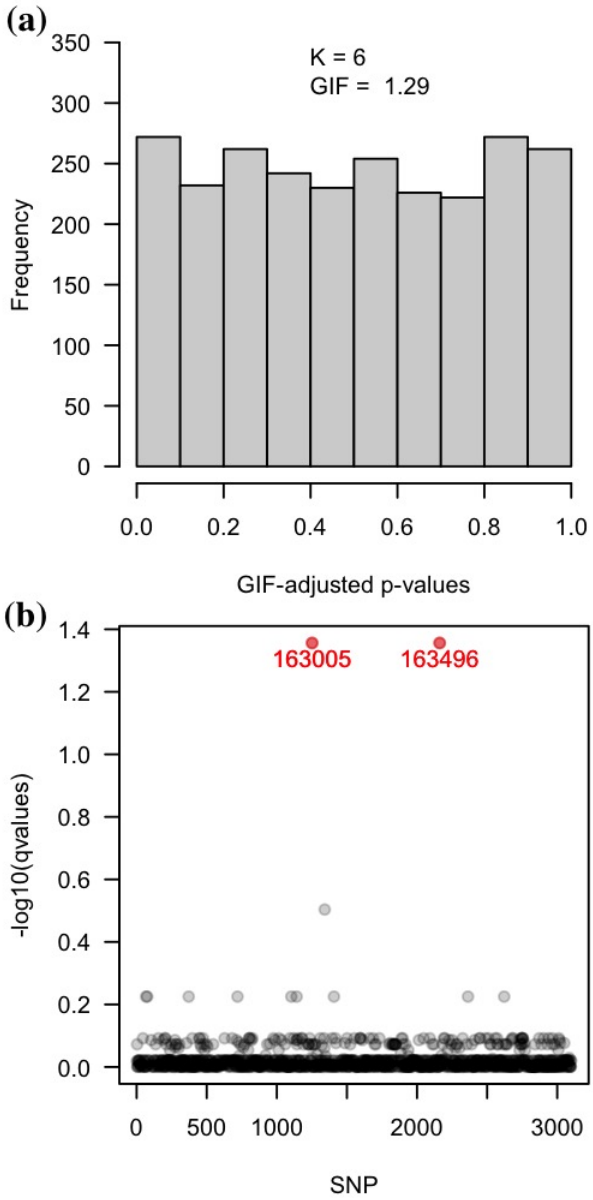

Figure 6: Genotype-environment association test based on a latent factor mixed model with ridge penalty. **(a)** Distribution of adjusted p-values using the default genomic inflation factor. **(b)** Manhattan plot of loci (SNPs) potentially affected by the PC1 predictor variable. Loci highlighted in red were considered to be under selection according to a False Discovery Rate of 0.05.