Genomic analysis of European Drosophila populations reveals longitudinal structure and continent-wide selection

Martin Kapun[1,2,3,*], Maite G. Barrón[1,4,*], Fabian Staubach[1,5,§], Jorge Vieira[1,6,7], Darren J. Obbard[1,8], Clément Goubert[1,9,10], Omar Rota-Stabelli[1,11], Maaria Kankare[1,12,§], Annabelle Haudry[1,9], R. Axel W. Wiberg[1,13,14], Lena Waidele[1,5], Iryna Kozeretska[1,15], Elena G. Pasyukova[1,16], Volker Loeschcke[1,17], Marta Pascual[1,18], Cristina P. Vieira[1,6,7], Svitlana Serga[1,15], Catherine Montchamp-Moreau[1,19], Jessica Abbott[1,20], Patricia Gibert[1,9], Damiano Porcelli[1,21], Nico Posnien[1,22], Sonja Grath[1,23], Élio Sucena[1,24,25], Alan O. Bergland[1,26,§], Maria Pilar Garcia Guerreiro[1,27], Banu Sebnem Onder[1,28], Eliza Argyridou[1,23], Lain Guio[1,4], Mads Fristrup Schou[1,17], Bart Deplancke[1,29], Cristina Vieira[1,9], Michael G. Ritchie[1,13], Bas J. Zwaan[1,30], Eran Tauber[1,31], Dorcas J. Orengo[1,18], Eva Puerma[1,18], Montserrat Aguadé[1,18], Paul S. Schmidt[1,32,§], John Parsch[1,23], Andrea J. Betancourt[1,33], Thomas Flatt[1,2,3,†,*,§], Josefa González[1,4,†,*,§]

[1] The European *Drosophila* Population Genomics Consortium (*DrosEU*)

[2] Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland

[3] Department of Biology, University of Fribourg, CH-1700 Fribourg, Switzerland

[4] Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona, Spain

[5] Department of Evolutionary Biology and Ecology, University of Freiburg, 79104 Freiburg, German

[6] Instituto de Biologia Molecular e Celular (IBMC) University of Porto, Porto, Portugal

[7] Instituto de Investigação e Inovação em Saúde (I3S), University of Porto, Porto, Portugal

[8] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom

[9] Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, University Lyon 1, Lyon, France

[10] Department of Molecular Biology and Genetics, 107 Biotechnology Building, Cornell University, Ithaca, New York 14853, USA

[11] Research and Innovation Centre, Fondazione Edmund Mach, San Michele all' Adige, Italy

[12] Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland

[13] Centre for Biological Diversity, School of Biology, University of St. Andrews, St Andrews, United Kingdom

[14] Evolutionary Biology, Zoological Institute, University of Basel, Basel, CH-4051, Switzerland

[15] General and Medical Genetics Department, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

[16] Laboratory of Genome Variation, Institute of Molecular Genetics of RAS, Moscow, Russia

[17] Department of Bioscience - Genetics, Ecology and Evolution, Aarhus University, Aarhus C, Denmark

[18] Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

[19] Laboratoire Evolution, Génomes, Comportement et Ecologie (EGCE) UMR 9191 CNRS - UMR247 IRD - Université Paris Sud - Université Paris Saclay. 91198 Gif sur Yvette Cedex, France.

[20] Department of Biology, Section for Evolutionary Ecology, Lund, Sweden

[21] Department of Animal and Plant Sciences, Sheffield, United Kingdom

[22]Universität Göttingen, Johann-Friedrich-Blumenbach-Institut für Zoologie und Anthropologie, Göttingen, Germany

[23] Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilians-Universität München, Planegg, Germany

[24] Instituto Gulbenkian de Ciência, Oeiras, Portugal

[25]Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal

[26] Department of Biology, University of Virginia, Charlottesville, VA, USA

[27] Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Barcelona, Spain

[28] Department of Biology, Faculty of Science, Hacettepe University, Ankara, Turkey

[29] Laboratory of Systems Biology and Genetics, EPFL-SV-IBI-UPDEPLA, CH-1015 Lausanne, Switzerland

[30] Laboratory of Genetics, Department of Plant Sciences, Wageningen University, Wageningen, Netherlands

[31] Department of Evolutionary and Environmental Biology and Institute of Evolution, University of Haifa, Haifa, Israel

[32] Department of Biology, University of Pennsylvania, Philadelphia, USA

[33] Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom

[†] **For correspondence**: thomas.flatt@unifr.ch, josefa.gonzalez@ibe.upf-csic.es

[*] These authors contributed equally to this work

[§] Members of the *Drosophila* Real Time Evolution (Dros-RTEC) Consortium

Competing interests: The authors declare that no competing interests exist.

Abstract

Genetic variation is the fuel of evolution. However, analyzing evolutionary dynamics in natural populations is challenging, sequencing of entire populations remains costly and comprehensive sampling logistically difficult. To tackle this issue and to define relevant spatial and temporal scales of variation, we have founded the European *Drosophila* Population Genomics Consortium (*DrosEU*). Here we present the first analysis of 48 *D. melanogaster* population samples collected across Europe in 2014. Our analysis uncovers novel patterns of variation at multiple levels: genome-wide neutral SNPs, mtDNA haplotypes, inversions, and TEs showing previously cryptic longitudinal population structure; signatures of selective sweeps shared among populations; presumably adaptive clines in inversions; and geographic variation in TEs. Additionally, we document highly variable microbiota and identify several new *Drosophila* viruses. Our study reveals novel aspects of the population biology of *D. melanogaster* and illustrates the power of extensive sampling and pooled sequencing of populations on a continent-wide scale.

Keywords: *Drosophila*, population genomics, demography, selection, clines, SNPs, structural variants, symbionts.

Introduction

Genetic variation is the raw material for evolutionary change. Understanding the processes that create and maintain variation in natural populations remains a fundamental goal in evolutionary biology. The identification of patterns of genetic variation within and among

taxa (Dobzhansky 1970; Lewontin 1974; Kreitman 1983; Kimura 1984; Hudson *et al.* 1987; McDonald & Kreitman 1991; e.g., Adrian & Comeron 2013) provides fundamental insights into the action of various evolutionary forces. Historically, due to technological constraints, studies of genetic variation were limited to single loci or small genomic regions and to static sampling of small numbers of individuals from natural populations. The development of population genomics has extended such analyses to patterns of variation on a genome-wide scale (e.g., Black *et al.* 2001; Jorde *et al.* 2001; Luikart *et al.* 2003; Begun *et al.* 2007; Sella *et al.* 2009; Charlesworth 2010; Casillas & Barbadilla 2017). This has resulted in fundamental advances in our understanding of historical and contemporaneous evolutionary dynamics in natural populations (e.g., Sella *et al.* 2009; Hohenlohe *et al.* 2010; Cheng *et al.* 2012; Fabian *et al.* 2012; Pool *et al.* 2012; Messer & Petrov 2013; Ellegren 2014; Harpur *et al.* 2014; Kapun *et al.* 2014; Bergland *et al.* 2014; Charlesworth 2015; Zanini *et al.* 2015; Kapun *et al.* 2016a; Casillas & Barbadilla 2017).

However, large-scale sampling and genome sequencing of entire populations remains largely prohibitive in terms of sequencing costs and labor-intensive sample collection, limiting the number of populations that can be analyzed. Evolution is a highly dynamic process across a variety of spatial scales in many taxa; thus, to generate a comprehensive context for population genomic analyses, it is essential to define the appropriate spatial scales of analysis, from meters to thousands of kilometers (Levins 1968; Endler 1977; Richardson *et al.* 2014). Furthermore, one-time sampling of natural populations provides only a static view of patterns of genetic variation. Allele frequency changes can be highly dynamic even across very short timescales (e.g., Umina *et al.* 2005; Bergland *et al.* 2014; Behrman *et al.* 2018), and theoretical work suggests that such temporal dynamics may be an important yet largely understudied mechanism by which genetic variation is maintained

126 (Wittmann *et al.* 2017). It is thus essential to define the relevant spatio-temporal scales for
127 sampling and population genomic analyses accordingly.
128
129 To generate a population genomic framework that can deliver appropriate high-resolution
130 sampling and to provide a unique resource to the research community, we formed the
131 European *Drosophila* Population Genomics Consortium (*DrosEU*; https://droseu.net). Our
132 primary objective is to utilize the strengths of this consortium to extensively sample and
133 sequence European populations of *Drosophila melanogaster* on a continent-wide scale
134 and across distinct timescales. In close cooperation with a complementary effort focused
135 on North American populations, the *Drosophila* Real Time Evolution Consortium (*Dros-*
136 *RTEC*; http://web.sas.upenn.edu/paul-schmidt-lab/dros-rtec/), our long-term goal is to
137 define the appropriate spatio-temporal scales at which populations should be sampled and
138 analyzed and to gain novel insights into the dynamics of genetic variation.
139
140 *D. melanogaster* offers several advantages for such a concerted sampling and analysis
141 effort: a relatively small genome, a broad geographic range, a multivoltine life history that
142 allows sampling across generations over short timescales, ease of sampling natural
143 populations using standardized techniques, an extensive research community and a well-
144 developed context for population genomic analysis (Powell 1997; Keller 2007; Hales *et al.*
145 2015). The species is native to sub-Saharan Africa and has subsequently expanded its
146 range into novel habitats in Europe over the last 10,000-15,000 years and in North
147 America and Australia in the last several hundred years (e.g., Lachaise *et al.* 1988; David
148 & Capy 1988; Keller 2007). On both the North American and Australian continents, the
149 prevalence of latitudinal clines in frequencies of alleles (e.g., Schmidt & Paaby 2008;
150 Turner *et al.* 2008; Kolaczkowski *et al.* 2011b; Fabian *et al.* 2012; Bergland *et al.* 2014;

151 Machado *et al.* 2016; Kapun *et al.* 2016a), structural variants such as chromosomal
152 inversions (Mettler *et al.* 1977; Voelker *et al.* 1978; Knibb *et al.* 1981; Knibb 1982; 1986;
153 Anderson *et al.* 1991; Rako *et al.* 2006; Kapun *et al.* 2014; Rane *et al.* 2015; Kapun *et al.*
154 2016a) and transposable elements (TEs) (Boussy *et al.* 1998; González *et al.* 2008; 2010),
155 as well as complex phenotypes (de Jong & Bochdanovits 2003; Schmidt & Paaby 2008;
156 Schmidt *et al.* 2008; Flatt *et al.* 2013; Adrion *et al.* 2015 and references therein; Kapun *et*
157 *al.* 2016b; Behrman *et al.* 2018) have been interpreted to result from local adaptation to
158 environmental factors that co-vary with latitude or as the legacy of an out-of-Africa
159 dispersal history. However, sampling across these latitudinal gradients has not been
160 replicated outside of a single transect on the east coasts of both continents. The observed
161 latitudinal clines on the east coasts of North America and Australia may have been
162 generated, at least in part, by demography and differential colonization histories of
163 populations at high and low latitudes (Bergland *et al.* 2016). In North America, for example,
164 temperate populations appear to be largely of European origin, whereas low-latitude
165 populations show evidence of greater admixture from ancestral African populations and
166 the Caribbean (Caracristi & Schlötterer 2003; Yukilevich & True 2008a; b; Duchen *et al.*
167 2013; Kao *et al.* 2015; Bergland *et al.* 2016). More intensive sampling and analysis of both
168 African as well as European populations is thus essential to disentangling the relative
169 importance of local adaptation versus colonization history and demography in generating
170 the clinal patterns that have been widely observed. While there has been a great deal of
171 progress in the analysis of ancestral African populations (e.g., Begun & Aquadro 1993;
172 Corbett-Detig & Hartl 2012; Pool *et al.* 2012; Fabian *et al.* 2015; Lack *et al.* 2015; 2016),
173 the European continent remains largely uncharacterized at the population genomic level
174 (Božičević *et al.* 2016; Pool *et al.* 2016; Mateo *et al.* 2018).
175

Here, we present the first analysis of the *DrosEU* pool-sequencing data from a set of 48 European population samples collected in 2014. Our main focus is on describing spatial variation across the European continent. A similar consortium has been organized mainly in the United States, the *Dros-RTEC* consortium. While the two consortia share the common goal of widespread and coordinate sampling, the *Dros-RTEC* consortium concentrates on seasonal dynamics in North American populations (Machado *et al.* 2018). We examine the 2014 *DrosEU* data at three levels: (1) patterns of variation at single-nucleotide polymorphisms (SNPs) in the nuclear (~5.5 x 10^6 SNPs) and mitochondrial (mtDNA) genomes; (2) variation in copy number of transposable elements (TEs); (3) cosmopolitan chromosomal inversions previously associated with climate adaptation; and (4) variation among populations in microbiota, including endosymbionts, bacteria, and viruses (Figure 1).

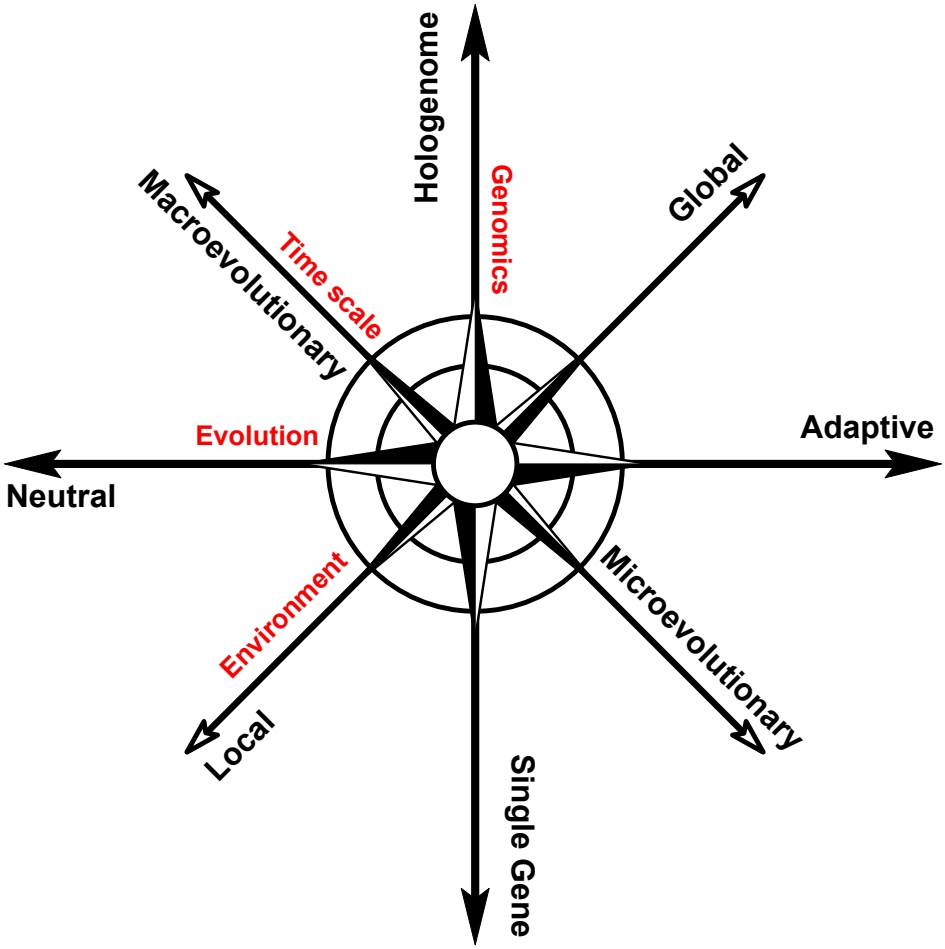


Figure 1. The conceptual framework of the *DrosEU* consortium. By intensive spatio-temporal sampling of natural populations of *Drosophila melanogaster*, the European *Drosophila* Population Genomics Consortium (*DrosEU*; http://droseu.net/), aims to uncover the factors that shape the evolutionary dynamics of this exemplary model organism. Each of the repeatedly and consistently sampled *DrosEU* populations is subject to evolutionary forces ("Evolution axis", from neutral to adaptive evolution) in interaction with the environment ("Environment axis", from local aspects to global patterns, including spatial factors). In addition, there are several dimensions along which the fly genomes can be studied, from single SNPs and genes to structural variants and co-evolving genomes ("Genomics axis"), both over short and long timescales ("Timescale axis").

We find that European populations of *D. melanogaster* exhibit novel patterns of variation at all levels investigated: neutral SNPs in the nuclear genome and mtDNA haplotypes that

reveal previously unknown longitudinal population structure; genomic regions consistent with selective sweeps that indicate selection on a continent-wide scale; new evidence for inversion clines in Europe; and spatio-temporal variation in TEs frequencies. We also identify four new DNA viruses and for the first time assemble the complete genome of a fifth. These novel features are revealed by the comprehensive magnitude of our coordinated sampling, thus demonstrating the utility of this approach.

Together with other large-scale genomic datasets for *D. melanogaster* (Casillas & Barbadilla 2017) our data provide a rich and powerful community resource for studies of molecular population genetics. Importantly, the *DrosEU* dataset represents the first comprehensive characterization of genetic variation in *D. melanogaster* on the European continent and might yield important insights into how this species has adapted to temperate climates after its migration out of Africa.

Results

As part of the *DrosEU* effort, we collected and sequenced 48 population samples of *D. melanogaster* from 32 geographical locations across Europe in 2014 (Table 1; Figure 2 and Figure 3A).

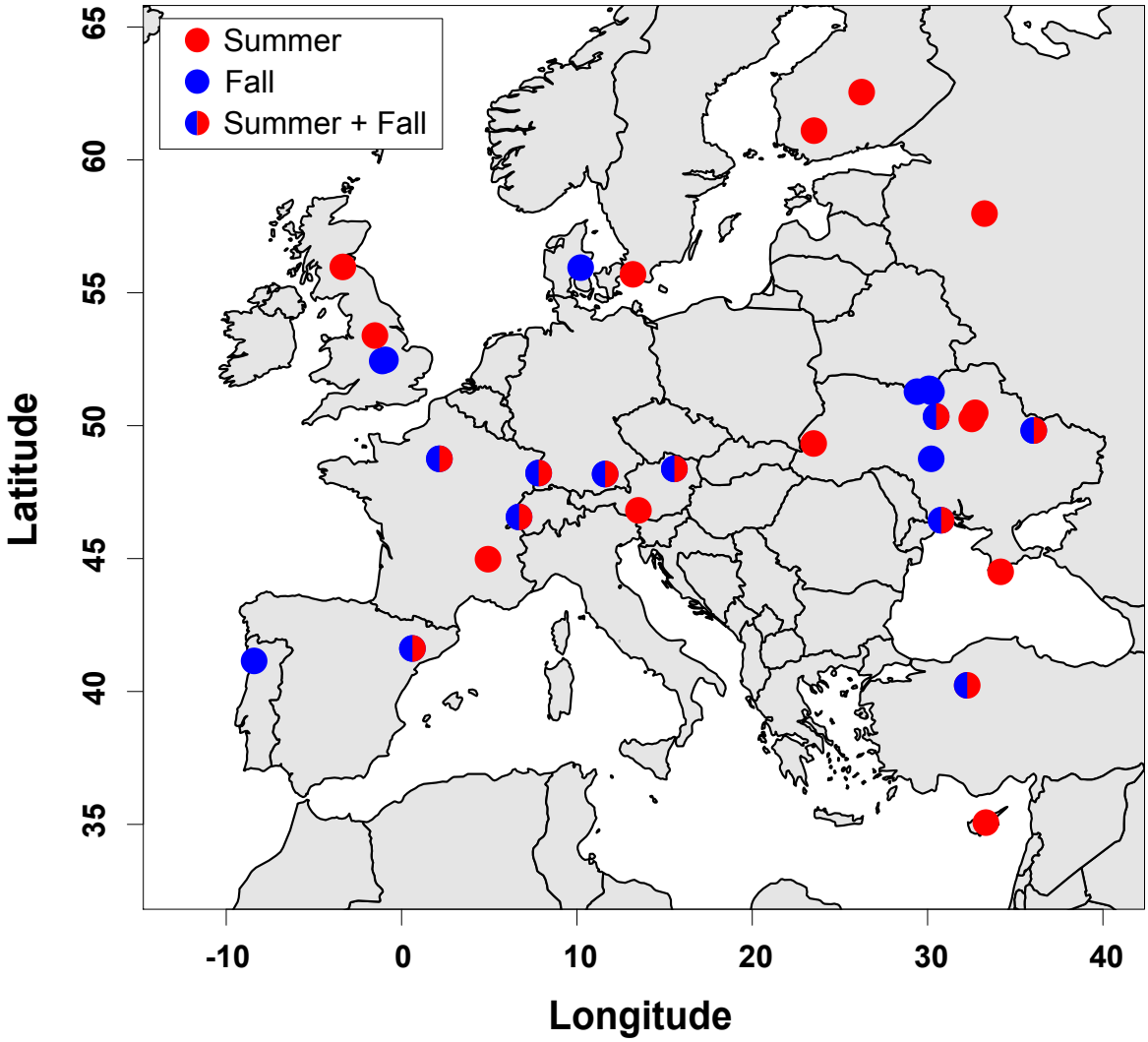


Figure 2. The geographic distribution of population samples. The map shows the geographic locations of all samples in the 2014 *DrosEU* dataset. The color of the circles indicates the sampling season for each location (see Table 1 and Supplemental Table 1). Note that some of the 12 Ukranian locations overlap in the map.

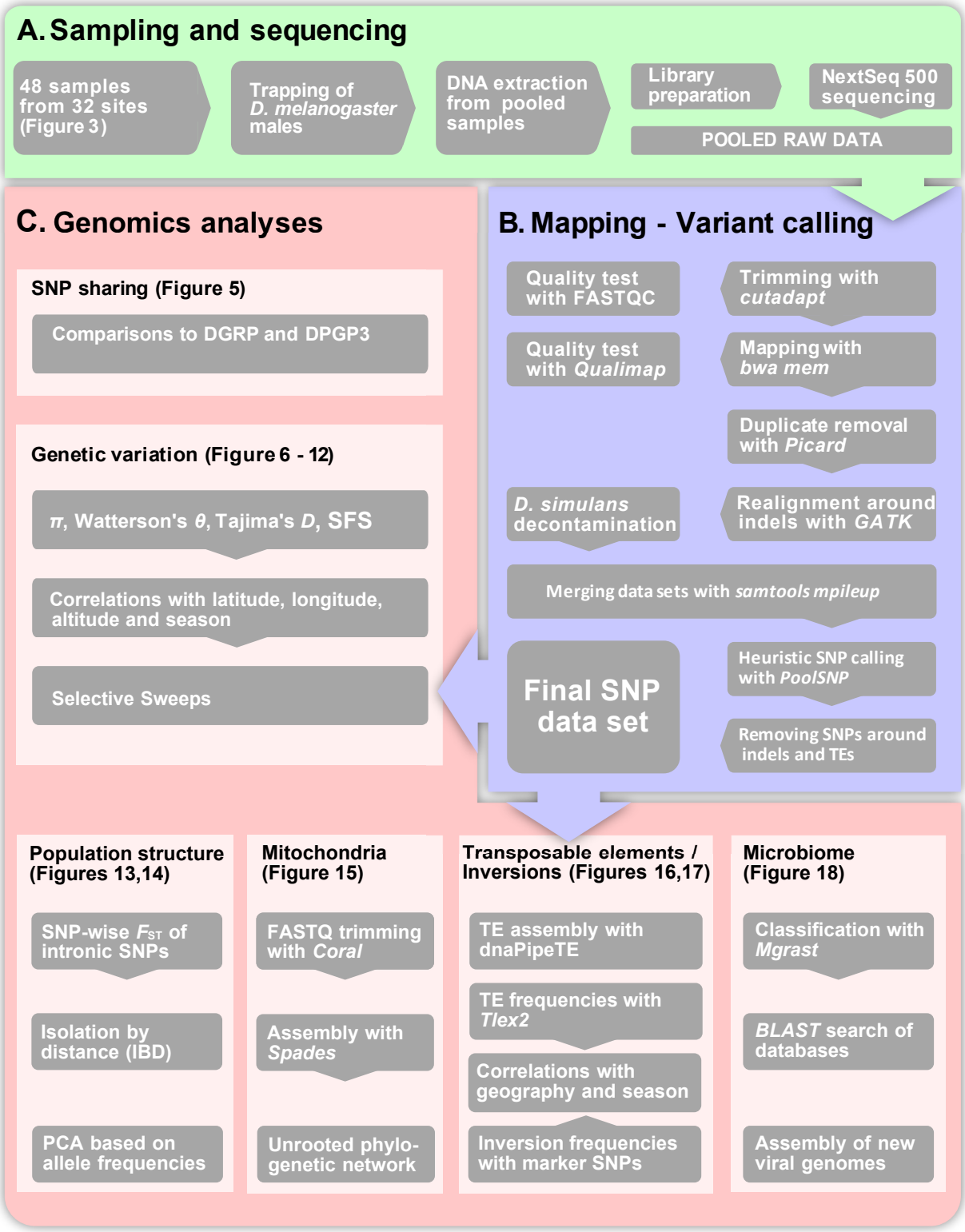


Figure 3. Sampling and data analysis pipeline. The schematic diagram shows the workflow of data collection and processing (A) followed by bioinformatic approaches used for quality assessment and read mapping (B) as well as the downstream analyses (C) conducted in this study (see Materials and Methods for further information).

While our analyses focus on spatial patterns, thirteen of the 32 locations were sampled repeatedly during the year (at least twice, once in summer and once in fall), allowing a first, crude analysis of seasonal changes in allele frequencies on a genome-wide level (Figure 2). For an extensive analysis of temporal (seasonal) patterns in mainly North American populations see the companion paper by Machado *et al*. (2018). All 48 samples were sequenced to high coverage, with a mean coverage per population of >50x (Table S1 and Figure 4).

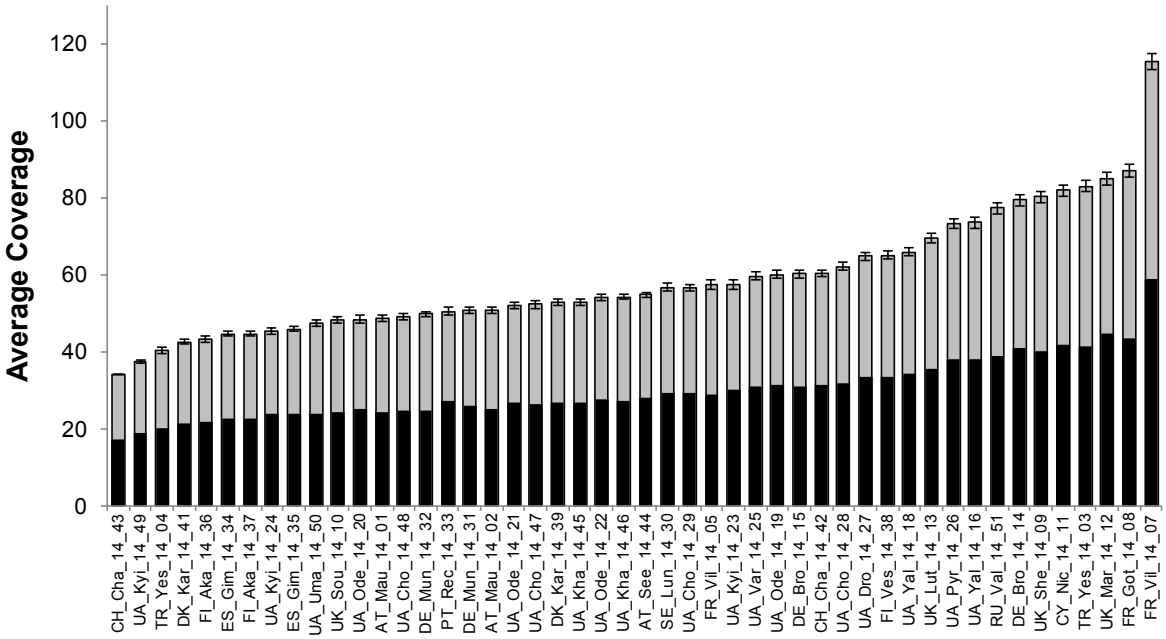


Figure 4. Chromosome-wide average coverages. Barplot showing the distribution of chromosome-wide coverages for *X*-chromosomes (black) and autosomes (grey). The error bars represent the standard deviation of coverages across all autosomal arms (*2L, 2R, 3L, 3R*).

Using this high-quality dataset, we performed the first comprehensive, continent-wide genomic analysis of European *D. melanogaster* populations (Figure 3). In addition to nuclear SNPs, we also investigated variation in mtDNA, TE insertions, chromosomal

244 inversion polymorphisms, and the *Drosophila*-associated microbiome (Figure 3).

245

246 **Most SNPs are widespread throughout Europe**

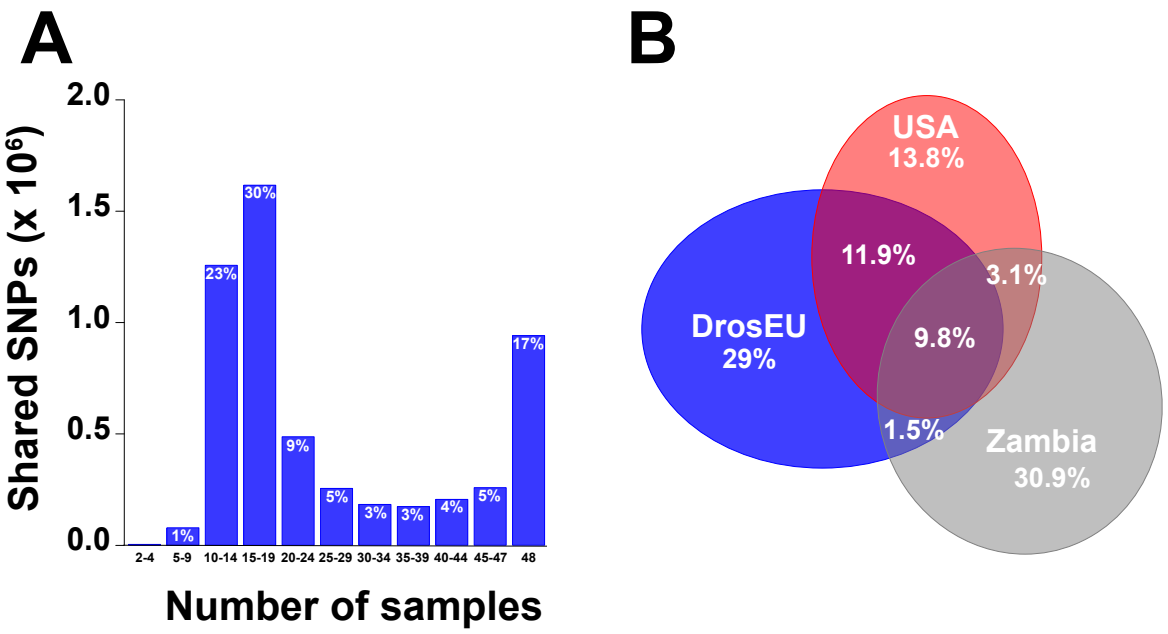
247 We identified a total of 5,558,241 “high confidence” SNPs with frequencies > 0.1% across

248 all 48 samples (Figure 3B, Table S1 and S2). Of these, 17% (941,080) were shared

249 among all samples, whereas 62% were polymorphic in fewer than 50% of the samples

250 (Figure 5A).

251



252

253 **Figure 5** with 2 supplements. **SNP sharing.** (A) Shared SNPs among *DrosEU* samples. Number and proportion of

254 SNPs in different samples, ranging from one specific sample to being shared among all 48 samples. (B) Shared SNPs

255 among three worldwide populations. Elliptic Venn diagram showing the number and proportion of SNPs overlapping

256 among the 5,361,256 biallelic SNPs in *DrosEU* (Europe), 3,953,804 biallelic SNPs in *DGRP* (North America) and

257 4,643,511 biallelic SNPs in Zambia (Africa) populations.

258

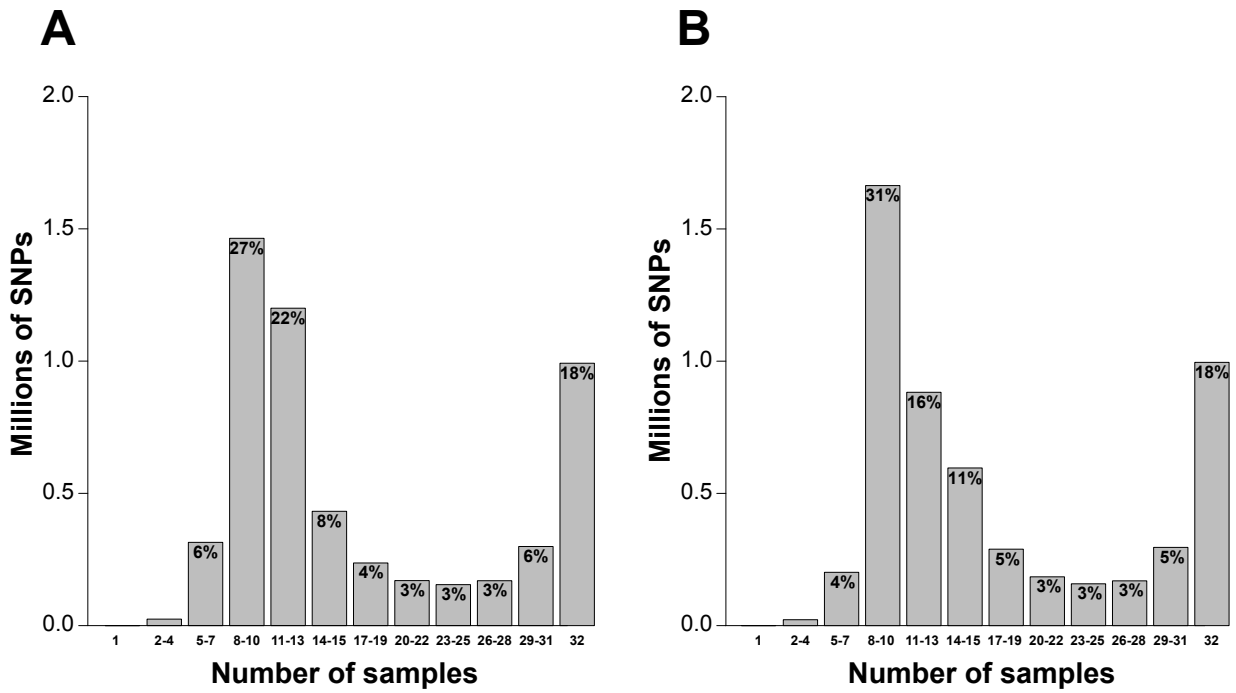


Figure 5 - figure supplement 1. SNP sharing among *DrosEU* samples. Number and proportion of SNPs in different 33 samples, each from one location. In the locations with more than one sample we took (A) the earliest sample or (B) the latest of the season.

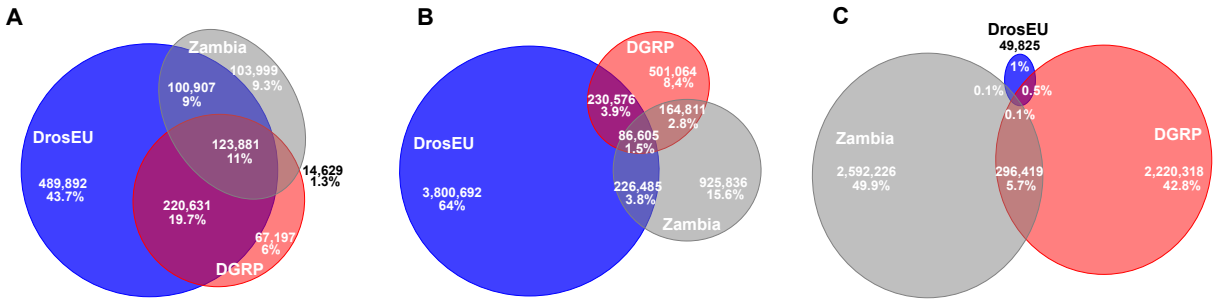


Figure 5 - figure supplement 2. SNP sharing with other populations by frequency bins. Elliptic Venn-diagram showing the proportion of SNPs shared and specific to each population, *DrosEU* (Europe), *DGRP* (North America) and Zambia (Africa) for different frequency bins: (A) frequency ≥ 0.5, (B) frequency ≥ 0.1 and < 0.5 and (C) frequency < 0.1. For proportions <1%, the number of SNPs is not depicted; *DrosEU*-Zambia: 4,088 SNPs, *DrosEU-DGRP*: 24,049 SNPs, *DrosEU-DGRP*-Zambia: 3,627 SNPs.

Due to our filtering scheme, SNPs that are private or nearly private to a sample will be recovered only if they are at a substantial frequency in that sample (~5%). In fact, only a small proportion of SNPs (1% = 3,645) was found in fewer than 10% of the samples, and only 0.004% (210) were specific to a single sample (Figure 5A). To avoid an excess contribution of SNPs from populations with multiple (seasonal) sampling, we repeated the analysis by considering only the earliest (Figure 5 - figure supplement 1A) or the latest (Figure 5 - figure supplement 1B) sample from populations with seasonal data. We observed similar patterns across the three analyses: (i) a very small number of sample-specific, private SNPs (210, 527 and 455, respectively), (ii) a majority of SNPs shared among 20% to 40% of the samples (53%, 52% and 52%, respectively), and (iii) a substantial proportion shared among all samples (17%, 20% and 19%, respectively; Figure 5A and Figure 5 - figure supplement 1). These results suggest that most SNPs are geographically widespread in Europe and that genetic differentiation among populations is moderate, consistent with high levels of gene flow across the European continent.

Derived European and North American populations share more SNPs with each other than they do with an ancestral African population

D. melanogaster originated in sub-Saharan Africa, migrated to Europe ~10,000-15,000 years ago, and subsequently colonized the rest of the world, including North America and Australia ~150 years ago (Lachaise *et al.* 1988; David & Capy 1988; Keller 2007). To search for genetic signatures of this shared history, we investigated the amount of allele sharing between African, European, and North American populations. We compared our SNP set to two published datasets, one from Zambia in sub-Saharan Africa (*DPGP3*; Lack *et al.* 2015) and one from North Carolina in North America (*DGRP*; Huang *et al.* 2014).

296 Populations from Zambia inhabit the ancestral geographical range of *D. melanogaster*
297 (Pool *et al.* 2012; Lack *et al.* 2015); North American populations are thought to be derived
298 from European populations, with some degree of admixture from African populations,
299 particularly in the southern United States and the Caribbean (Caracristi & Schlötterer
300 2003; Yukilevich & True 2008a; b; Yukilevich *et al.* 2010; Duchen *et al.* 2013; Kao *et al.*
301 2015; Pool 2015; Bergland *et al.* 2016). The population from North Carolina exhibits
302 primarily European ancestry, with ~15% admixture from Africa (Bergland *et al.* 2016).
303
304 Approximately 10% of the SNPs (~1 million) were shared among all three datasets (Figure
305 5B). Since the out-of-Africa range expansion and the subsequent colonization of the North
306 America continent by European (and to a lesser degree African) ancestors was likely
307 accompanied by founder effects, leading to a loss of African alleles, and adaptation to
308 temperate climates (Mettler *et al.* 1977), we predicted that a relatively high proportion of
309 SNPs would be shared between Europe and North America. As expected, the proportion
310 of shared SNPs was higher between Europe and North America (22%) than between
311 either Europe or North America and Zambia (11% and 13%, respectively; Figure 5B).
312 When we analyzed SNPs in variant frequency bins, the proportion of SNPs shared across
313 at least two continents increased from 26% to 41% for SNPs, with variant frequencies
314 larger than 50% (Figure 5 - figure supplement 2A). In contrast, only 6% of the SNPs at low
315 frequency (<10%; Figure 5 - figure supplement 2C) were shared. These results are
316 consistent with the loss of low-frequency variants during the colonization of the European
317 continent; they suggest that intermediate frequency alleles are more likely to be ancestral
318 and thus shared across broad geographic scales. Interestingly, as compared to Africa and
319 North America, we identified nearly 3 million private SNPs that are specific to Europe
320 (Figure 5B). Given that North American and Australian populations are – at least partly – of

European ancestry (see Lemeunier & Aulard 1992 for more details), future analysis of our data may be able to shed light on the demography and adaptation of these derived populations.

European and other derived populations exhibit similar amounts of genetic variation

Next, we estimated genome-wide levels of nucleotide diversity within the European population samples using population genetic summary statistics. Pairwise nucleotide diversity (π and Watterson's θ), corrected for pooling (Stalker 1976; Mettler *et al.* 1977; Voelker *et al.* 1978; Stalker 1980; Sezgin *et al.* 2004), ranged from 0.0047 to 0.0057 and from 0.0045 to 0.0064, respectively (Figure 6 and Figure 7), with our estimates being qualitatively similar to those from non-African *D. melanogaster* populations sequenced as individuals (Knibb *et al.* 1981; Knibb 1982; Anderson *et al.* 1987) or as pools (Inoue & Watanabe 1979; Inoue *et al.* 1984).

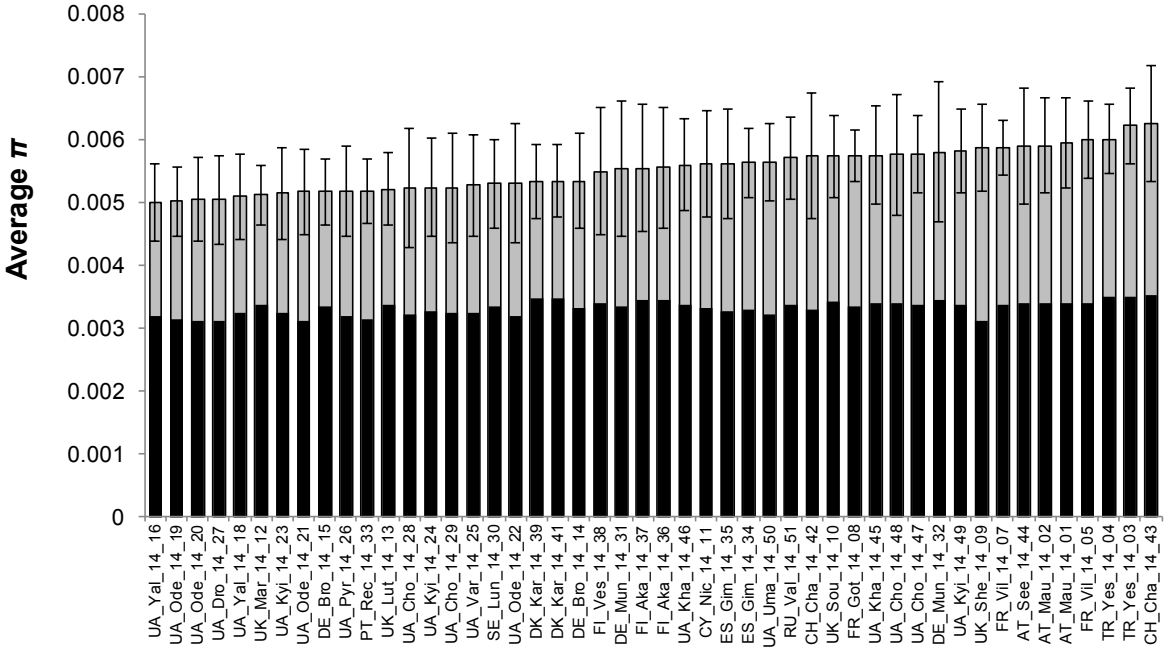
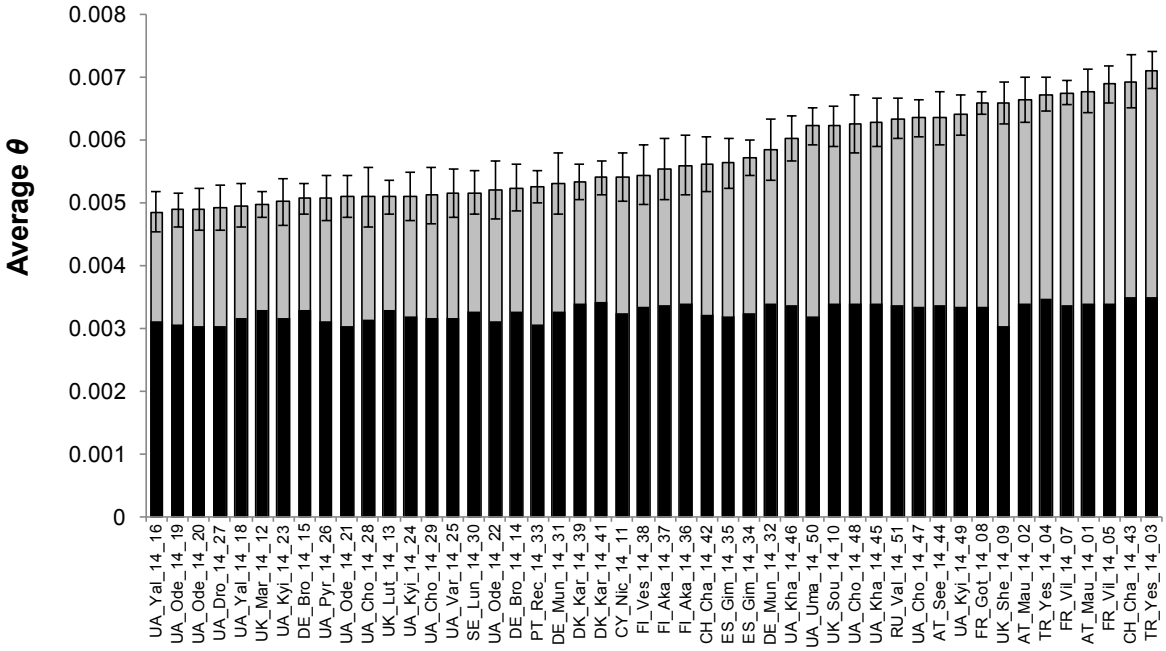


Figure 6. Chromosome-wide average Tajima's π. Barplot showing the distribution of chromosome-wide estimates of

337 Tajima's π for *X-chromosomes (*black) and autosomes (grey). The error bars represent the standard deviation of π
338 across all autosomal arms (*2L, 2R, 3L, 3R*).

339



340

341 **Figure 7. Chromosome-wide average Watterson's *θ*.** Barplot showing the distribution of chromosome-wide estimates
342 of Watterson's θ for *X-*chromosomes (black) and autosomes (grey). The error bars represent the standard deviation of θ
343 across all autosomal arms (*2L, 2R, 3L, 3R*).

344

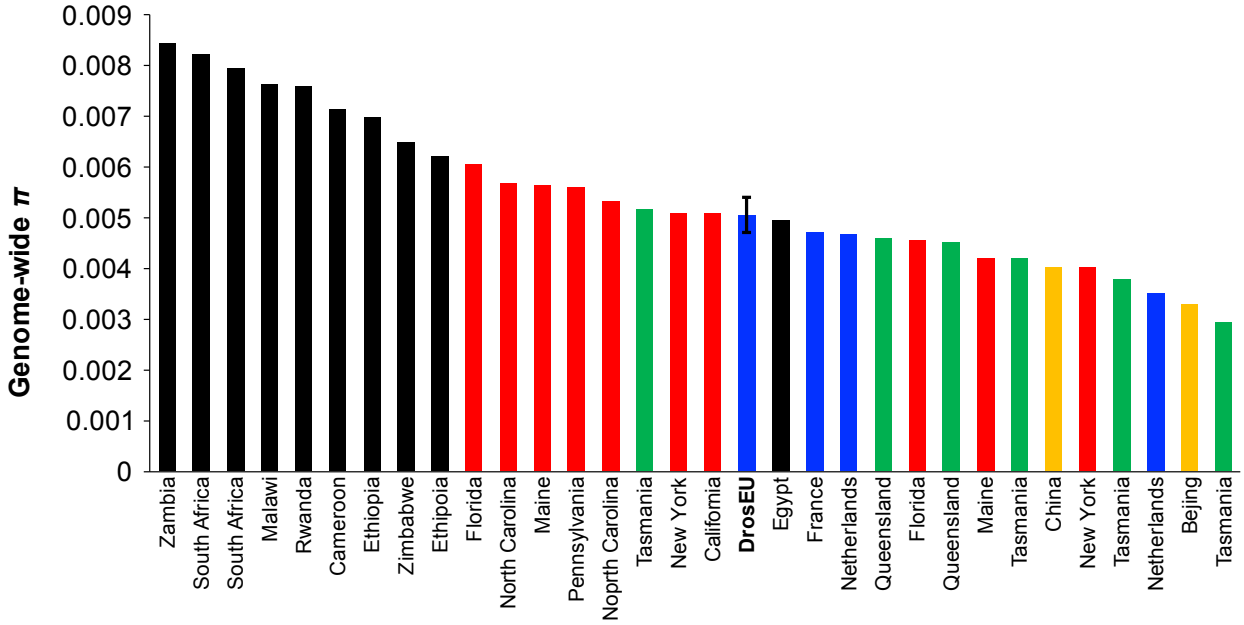


Figure 8. Genetic variation in worldwide samples. Barplot showing the distribution of genome-wide estimates of Tajima's π of the *DrosEU* and other genomic datasets (see Materials and Methods for more details) The error bars in the *DrosEU* dataset represent the standard deviation of π across all 48 population samples.

Estimates of π were slightly lower than, but in close agreement with, estimates of θ, leading to a slightly negative average of Tajima's *D* (Das & Singh 1990; 1991; Singh & Das 1992; Singh 2018). Due to our SNP calling approach (see Materials and Methods), we found a deficiency of alleles with frequencies ≤ 0.01, both in the sample-wise site frequency spectra (SFS) as well as in the combined SFS by SNP type, with the sample-wise SFS being skewed towards low frequency variants (Figure 9A).

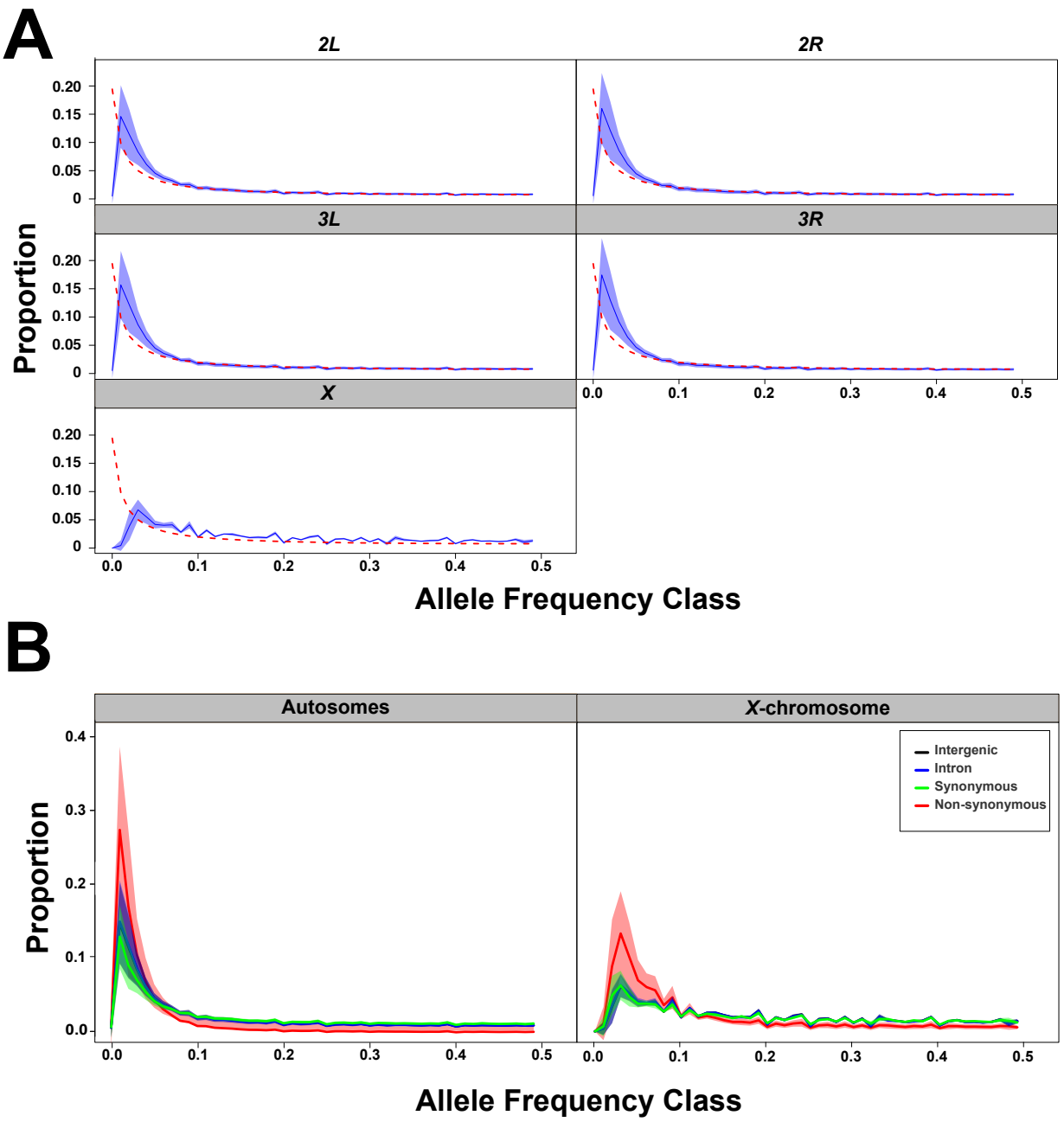


Figure 9. Site frequency spectra (SFS). (A) Chromosome-wise SFS: The five panels show the means (dark blue line) and standard deviations (light blue area) of the folded SFS across all 48 samples. The red dashed line indicates the expected SFS based on a θ of 10^{-4}. (B) SFS by SNP type: The two panels show the means (dark line) and standard deviations (light area) of the folded SFS for different SNP types (intergenic, intronic, non-synonymous and synonymous), combining SNPs across the autosomes (left) and the *X*-chromosome (right).

In addition, we observed an excess of low-frequency SNPs at non-synonymous sites as compared to other types of sites, which is consistent with purifying selection eliminating

366 deleterious non-synonymous mutations (Endler 1977).

367 Overall, we detected only minor differences in the amount of genetic variation among
368 populations. Specifically, genome-wide π ranged from 0.005 (Yalta, Ukraine) to 0.006
369 (Chalet à Gobet, Switzerland) for autosomes, and from 0.003 (Odesa, Ukraine) to 0.0035
370 (Chalet à Gobet, Switzerland) for the *X* chromosome (Table S1 and Figure 6). When
371 testing for associations between geographic variables and genome-wide average levels of
372 genetic variation, we found that both π and θ were strongly negatively correlated with
373 altitude, but neither was correlated with latitude or longitude (Table 2). There were no
374 correlations between the season in which the samples were collected and levels of
375 average genome-wide genetic variation as measured by π and θ (Table 2).

376

377 The *X* chromosome showed markedly lower genetic variation than the autosomes, with the
378 ratio of *X*-linked to autosomal variation (π_X/π_A) ranging from 0.53 to 0.66. These values
379 are well below the ratio of 0.75 (one-sample Wilcoxon rank test, $p < 0.001$) expected under
380 standard neutrality and equal sex ratio, but are consistent with previous findings for
381 European populations of *D. melanogaster* and can be attributed to either selection (Knibb
382 *et al.* 1981) or changes in population size (Knibb *et al.* 1981). This pattern is consistent
383 with previous estimates of relatively low *X*-linked diversity for European (Kapun *et al.*
384 2014) and other non-African populations (Kapun *et al.* 2016a). Interestingly, the ratio π_X/π_A
385 was significantly, albeit weakly, positively correlated with latitude (Spearman's $r = 0.315$, p
386 = 0.0289), with northern populations having slightly higher *X*/*A* ratios than southern
387 populations. This is at odds with the prediction of periodically bottlenecked populations
388 leading to a lower *X*/*A* ratio in the north and perhaps reflects more complex demographic
389 scenarios (Mettler *et al.* 1977; Voelker *et al.* 1978; Knibb *et al.* 1981; Knibb 1982; Das &
390 Singh 1991; Van 't Land *et al.* 2000; de Jong & Bochdanovits 2003; Anderson *et al.* 2005;

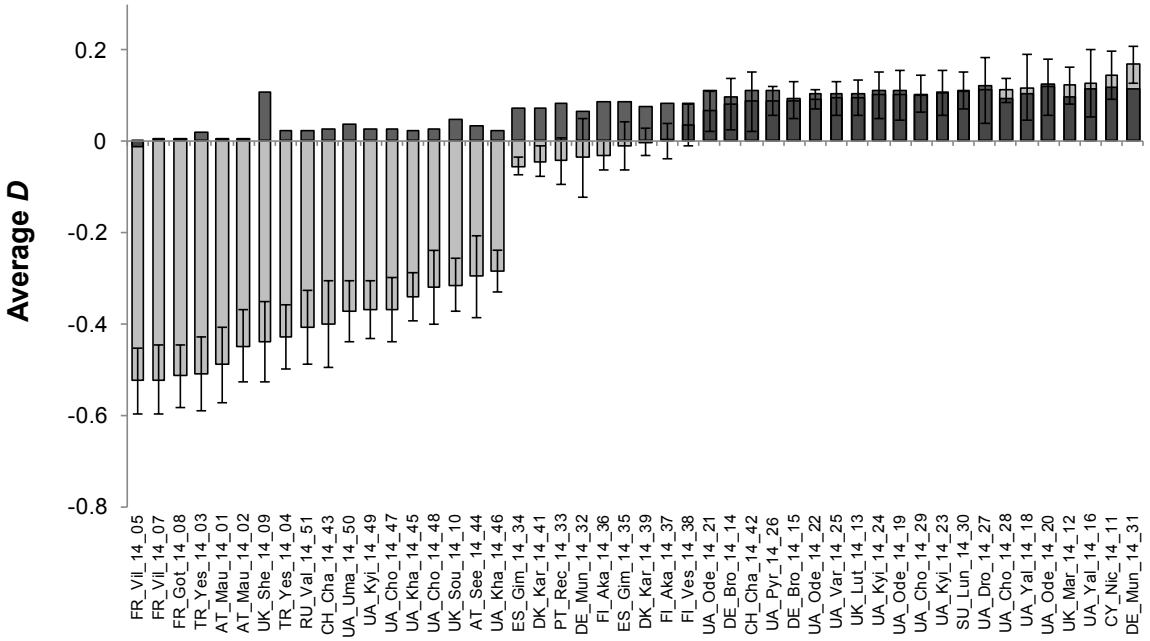
391 Umina *et al.* 2005; Rako *et al.* 2006; Kapun *et al.* 2014; 2016a).

392

393 In contrast to π and θ, we observed major differences in the genome-wide averages of

394 Tajima's *D* among samples (Figure 10).

395



396

397 **Figure 10. Chromosome-wide average Tajima's *D*.** Barplot showing the distribution of chromosome-wide estimates of

398 Tajima's *D* for *X*-chromosomes (black) and autosomes (grey). The error bars represent the standard deviation of *D*

399 across all autosomal arms (*2L, 2R, 3L, 3R*).

400

401 The chromosome-wide Tajima's *D* was negative in approximately half of all samples and

402 close to zero or slightly positive in the remaining samples, possibly due to heterogeneity in

403 the proportion of sequencing errors among the multiplexed sequencing runs. However,

404 models that included sequence run as a covariate did not explain more of the observed

405 variance than models without the covariate, suggesting that associations of π and θ with

406 geographic variables were not confounded by sequencing heterogeneity (see Supporting

Information; Table S4). Moreover, our results for π, θ and D are unlikely to be confounded by spatio-temporal autocorrelations: after accounting for similarity among spatial neighbors (Moran's $I \approx 0$, $p > 0.05$ for all tests), there were no significant residual autocorrelations among samples for these estimators.

Genetic variation was not distributed homogeneously across the genome. Both π and θ were markedly reduced close to centromeric and telomeric regions (Figure 11), which is in good agreement with previous studies reporting systematic reductions in genetic variation in regions with reduced recombination (Kennison 2008).

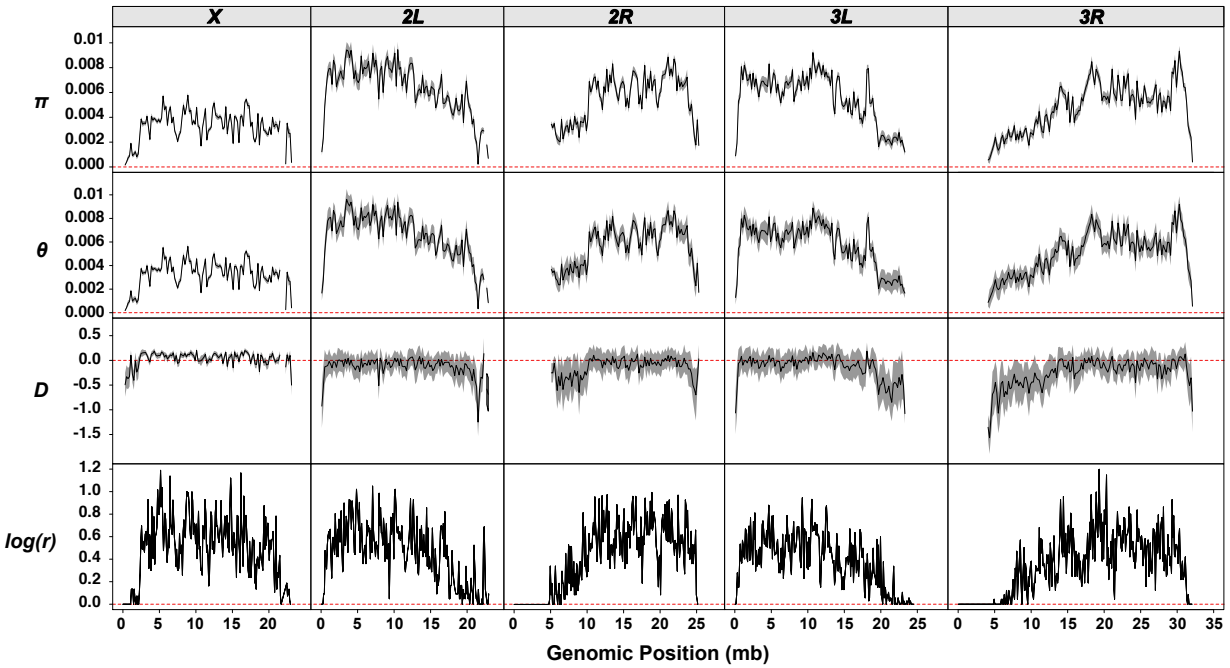
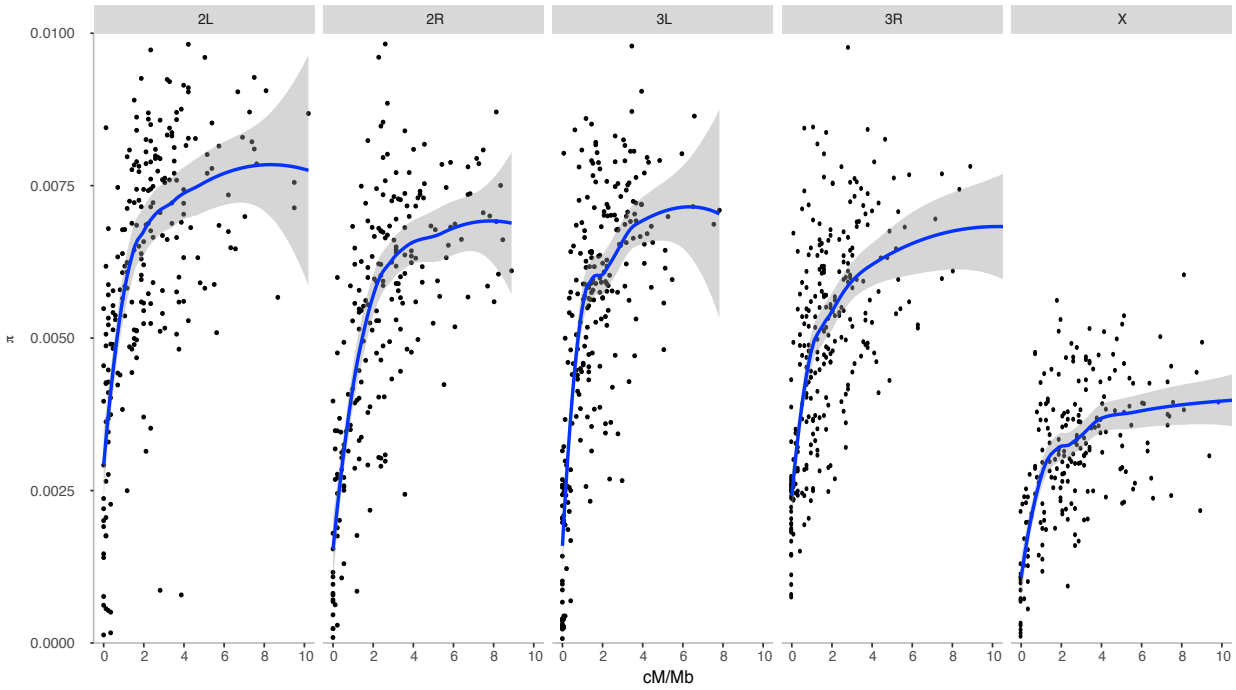


Figure 11 with 1 supplement**. Genome-wide estimates of genetic diversity and recombination rates**. The distribution of Tajima's π, Watterson's θ and Tajima's D (from top to bottom) in 200 kb non-overlapping windows plotted for each chromosomal arm separately. Bold black lines depict statistics which were averaged across all 48 samples and the upper and lower grey area show the corresponding standard deviations for each window. Red dashed lines highlight the vertical position of a zero value. The bottom row shows log-transformed recombination rates (r) in 100kb non-overlapping windows as obtained from Comeron et al. (2010).

424



425

426 **Figure 11 - figure supplement 1. Correlation between recombination and genetic diversity.** Smooth local
427 regression (LOESS) between Comeron et al. (2012) recombination rate in cM/Mb and the average of the 48 samples
428 genetic diversity (π) in 100kb non-overlapping windows by chromosome arm.

429

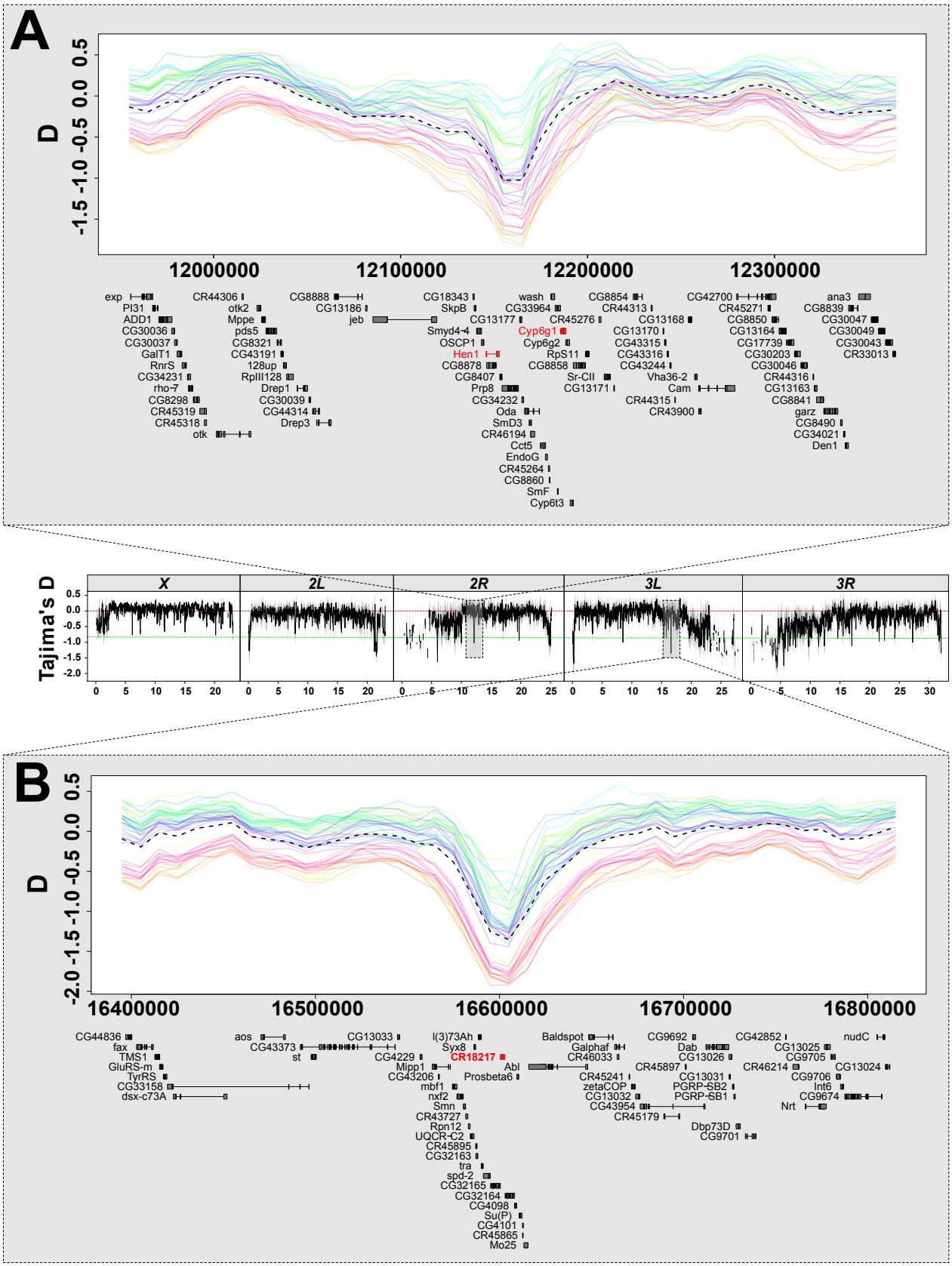
430 Consistent with this, we detected strong correlations with estimates of recombination rates
431 based on the data of Comeron *et al.* (2012) (linear regression, $p < 0.001$; not accounting
432 for autocorrelation), suggesting that the distribution of genome-wide genetic variation is
433 strongly influenced by the recombination landscape (Table S5). For autosomes, fine-scale
434 recombination rates explained 41-47% of the variation in π, whereas broad-scale
435 recombination rates (Roberts 1998; Pimpinelli *et al.* 2010) explained 50-56% of the
436 variation in diversity. We obtained similar results for *X*-chromosomes, with recombination
437 rates explaining 31-38% (Dobzhansky & Sturtevant 1938; Kunze-Mühl & Müller 1957;
438 Ashburner & Lemeunier 1976) or 24-33% (Wesley & Eanes 1994; Andolfatto *et al.* 1999;
439 Matzkin *et al.* 2005; Corbett-Detig *et al.* 2012) of the variation (Figure 11, Table S5, Figure

11 - figure supplement 1).

We also observed variation in Tajima's *D* with respect to genomic position (Figure 11). Notably, Tajima's *D* was markedly lower than the corresponding chromosome-wide average in the proximity of telomeric and centromeric regions on all chromosomal arms. These patterns possibly reflect purifying selection or selective sweeps close to heterochromatic regions (Navarro & Faria 2014; Kapun *et al.* 2014; 2016a), or might alternatively be a result of sequencing errors having a stronger effect on the SFS in low SNP density regions.

Localized reductions in Tajima's *D* are consistent with selective sweeps

We identified 144 genomic locations on the autosomes with non-zero recombination, reduced genetic variation, and a local reduction in Tajima's *D* (see Methods, Table S6), which jointly may be indicative of selective sweeps. Although we cannot rule out that these patterns are the result of non-selective demographic effects (e.g., bottlenecks), two observations suggest that at least some of these regions are affected by positive selection. First, bottlenecks are typically expected to cause genome-wide, non-localized reductions in Tajima's *D*. Second, several of the genomic regions in our data coincide with previously identified, well-supported selective sweeps in the proximity of *Hen1*, *Cyp6g1* (Andolfatto *et al.* 1999; Corbett-Detig & Hartl 2012), *wapl* (Andolfatto *et al.* 1999; Matzkin *et al.* 2005; Kennington *et al.* 2007; Corbett-Detig & Hartl 2012; Kennington & Hoffmann 2013; Kapun *et al.* 2014; 2016a), *HDAC6* (Begun 2015; Lavington & Kern 2017), and around the chimeric gene *CR18217* (Kirkpatrick 2010).



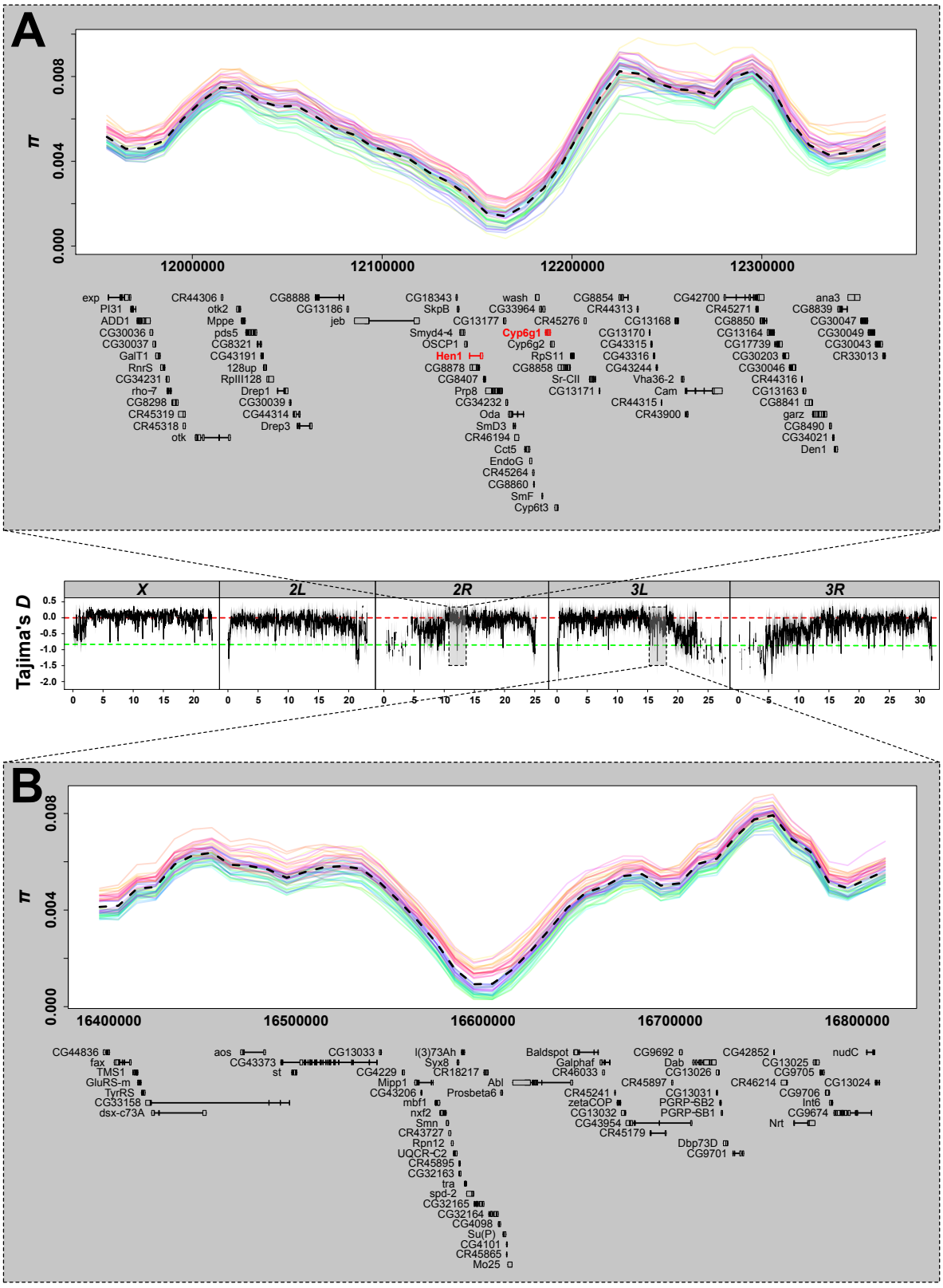
464

465 **Figure 12** with 3 supplements. **Signals of selective sweeps.** The central figure shows the distribution of Tajima's *D* in

466 50 kb sliding windows with 40 kb overlap. The red and green dashed lines show Tajima's *D* = 0 and Tajima's *D* = -1,

467 respectively. The top panel magnifies a genomic region on chromosomal arm *2R* that harbors well-known candidate loci
468 for pesticide resistance, *Cyp6g1* and *Hen1* (highlighted in red), where strong selection resulted in a selective sweep. This
469 sweep is characterized by an excess of low-frequency SNP variants, indicated by an overall negative Tajima's *D* in all
470 samples. The colored solid lines depict Tajima's *D* for each sample separately, whereas the black dashed line shows
471 Tajima's *D* averaged across all samples. (A legend for the color codes of the samples can be found in the Supporting
472 Information file in Figure 12 - figure supplement 3). The bottom figure shows a genomic region on *3L* which has not been
473 previously identified as a potential target of selection but shows Tajima's *D* patterns similar to the top figure. Notably,
474 both regions are also characterized by a strong reduction of genetic variation (Figure 12 - figure supplement 1).

475



476

477 **Figure 12 - figure supplement 1: Genetic variation in regions of putative selective sweeps.** This figure is equivalent

478 to Figure 12 in the main text but shows the distribution of genetic variation *(π)* in regions with depressed Tajima's *D*

around the well-studied *Cyp6g1* locus (A) and a previously unknown candidate region on *3L* (B). Similar to Tajima's *D, π* was calculated in 50 kb sliding windows with 40 kb overlap. See Table S6 for more examples. A legend for the color codes of the samples can be found in Figure 12 - figure supplement 3.

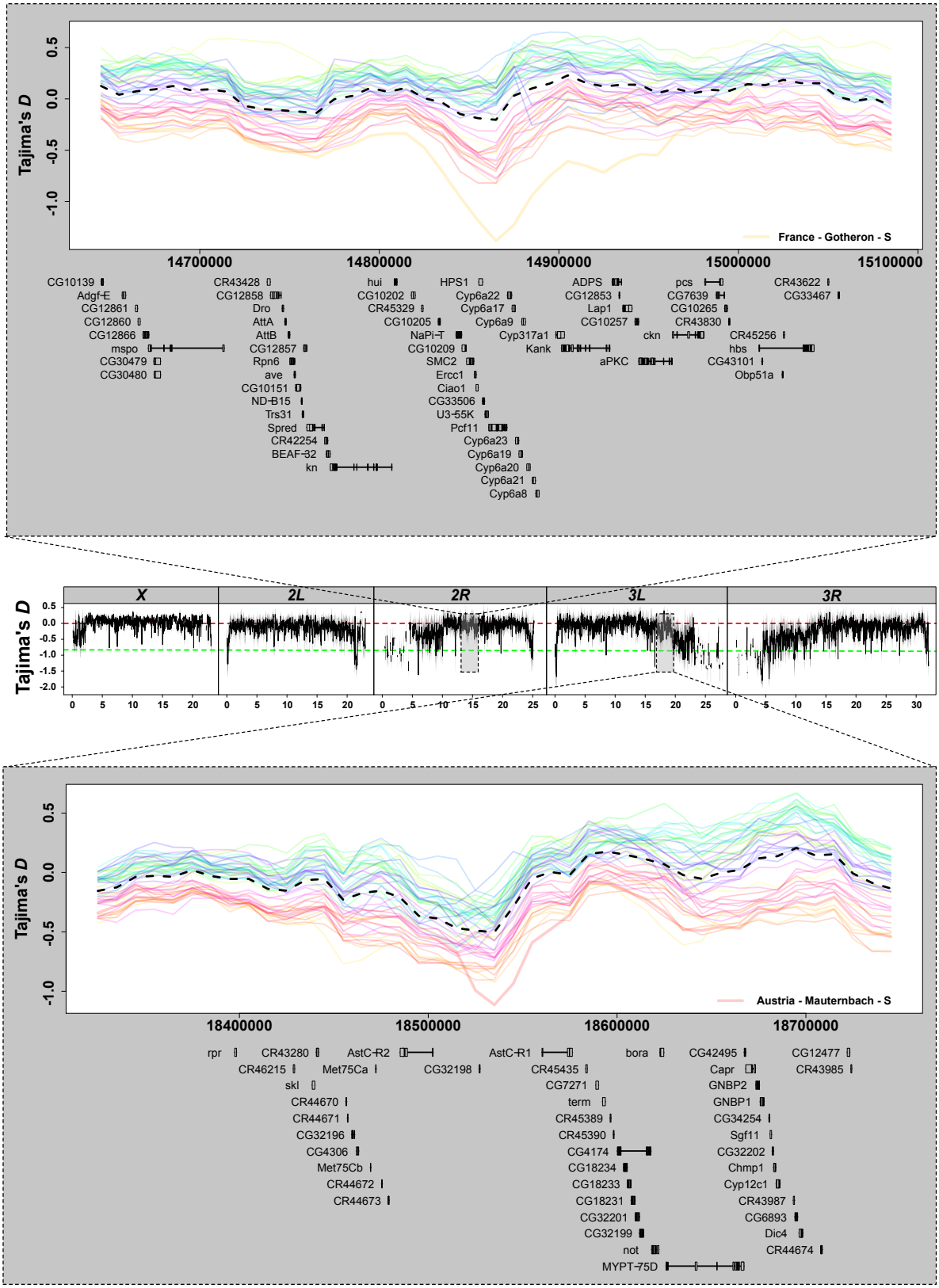


Figure 12 - figure supplement 2. Signals of selective sweeps in local populations. This figure is equivalent to Figure 12 in the main text but shows depressions in Tajima's *D* in single population samples. Such regions were identified by

enriching for samples with window-wise Tajima's *D* values < -0.9, which were smaller than 2 times the standard deviation of Tajima's *D* for all samples in the corresponding genomic region. The top and bottom panels show two such examples where samples with reduced Tajima's *D* are highlighted by thick lines and the corresponding sample names. See Table S6 for additional details. A legend for the color codes of the samples can be found in Figure 12 - figure supplement 3.

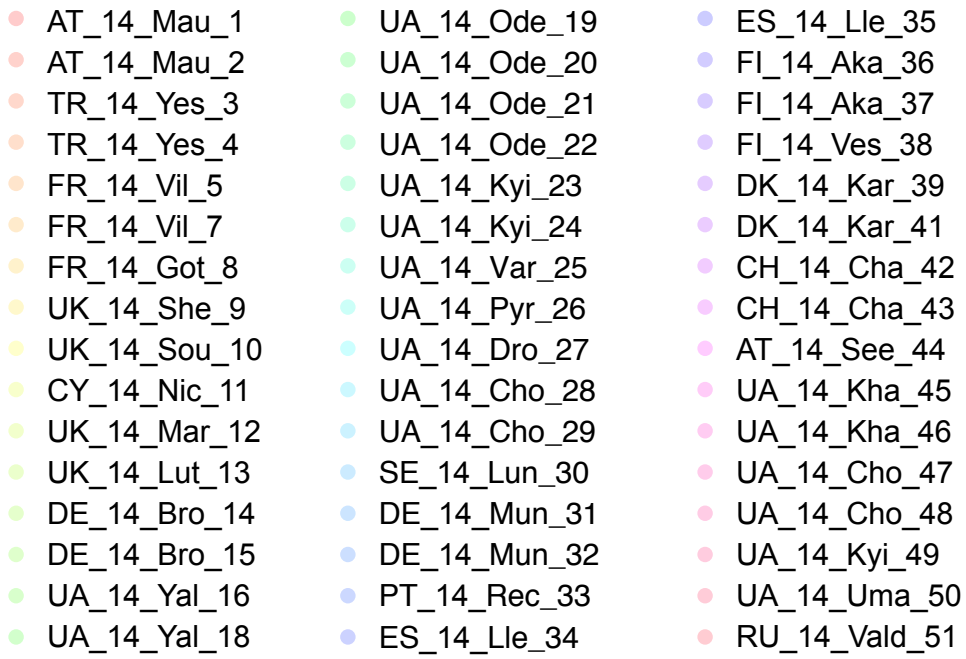


Figure 12 - figure supplement 3. *Legend for color code in Figure 12, Figure 12 - figure supplement 1 and Figure 12 - figure supplement 2.*

However, some regions, such as those around *wapl* or *HDAC6,* are characterized by low recombination rates (< 0.5 cM/Mb; Table S5), which can itself lead to reduced variation and Tajima's *D* (see also Nolte *et al.* 2013). Our screen also uncovered several regions that have not previously been described as harboring sweeps (Table S6). These represent promising candidate regions containing putative targets of positive selection. For several of these candidate regions, patterns of variation were highly similar across the majority of European samples, suggesting the existence of continent-wide selective sweeps that either predate the colonization of Europe (e.g., Beisswanger *et al.* 2006) or that have

swept across all European populations more recently. In contrast, some candidate regions were restricted to only a few populations and characterized by highly negative values of Tajima's *D*, i.e. deviating from the among-population average by more than two standard deviations, thus possibly hinting at cases of local, population-specific adaptation (Figure 12 - figure supplement 2 and Table S6 for examples).

European populations are strongly structured along an east-west gradient

We next investigated patterns of genetic differentiation due to demographic substructure. Overall, pairwise differentiation as measured by F_{ST} was relatively low, though markedly higher for *X*-chromosomes (0.043–0.076) than for autosomes (0.013–0.059; Student's *t*-test; $p < 0.001$; Figure 13), possibly reflecting differences in effective population size between the *X* chromosome and the autosomes (Hutter *et al.* 2007). One population, from Sheffield (UK), showed an unusually high amount of differentiation on the *X*-chromosome as compared to other populations (Figure 13).

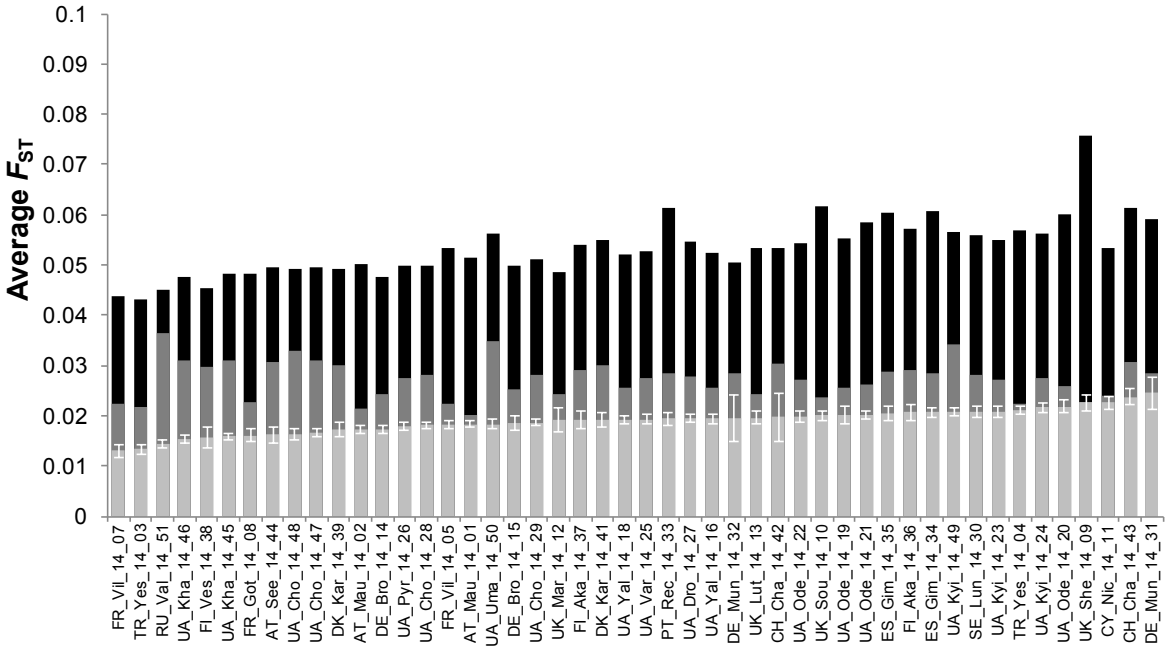


Figure 13. Chromosome-wide average F_{ST}. Bar-plot showing the distribution of chromosome-wide F_{ST} averaged across all possible pairwise comparisons for a given sample. The black bars show observed F_{ST} values and the dark grey bars expected F_{ST} values (based on autosomal values; see Materials and Methods) for the X chromosome. The light grey bars show autosomal F_{ST} values and error bars represent the standard deviation of average chromosome-wide F_{ST} across all autosomal arms (*2L, 2R, 3L, 3R*).

Despite these overall low levels of among-population differentiation, European populations showed some evidence of geographic substructure. To analyze this pattern in more detail, we focused on a set of SNPs located in short introns (< 60 bp), as these sites are relatively unaffected by selection (Haddrill *et al.* 2005; Singh *et al.* 2009; Parsch *et al.* 2010; Clemente & Vogl 2012; Lawrie *et al.* 2013). We analyzed the extent of isolation by distance (IBD) within Europe by correlating genetic and geographic distance and using pairwise F_{ST} between populations as a measure of genetic isolation. F_{ST} was overall low but significantly correlated with distance across the continent, indicating weak but significant IBD (Mantel test; $p < 0.001$; max. $F_{ST} \sim 0.05$; Figure 14A). We also examined those populations that were most and least separated by genetic differentiation, estimated

535 by pairwise F_{ST} (Figure 14B). In general, longitude had a stronger effect on isolation than
536 latitude, with populations showing the strongest differentiation separated along an east-
537 west, rather than a north-south, axis (Figure 14B). This pattern remained unchanged when
538 the number of populations sampled from Ukraine was reduced to avoid over-
539 representation (Figure 14 - figure supplement 1).
540

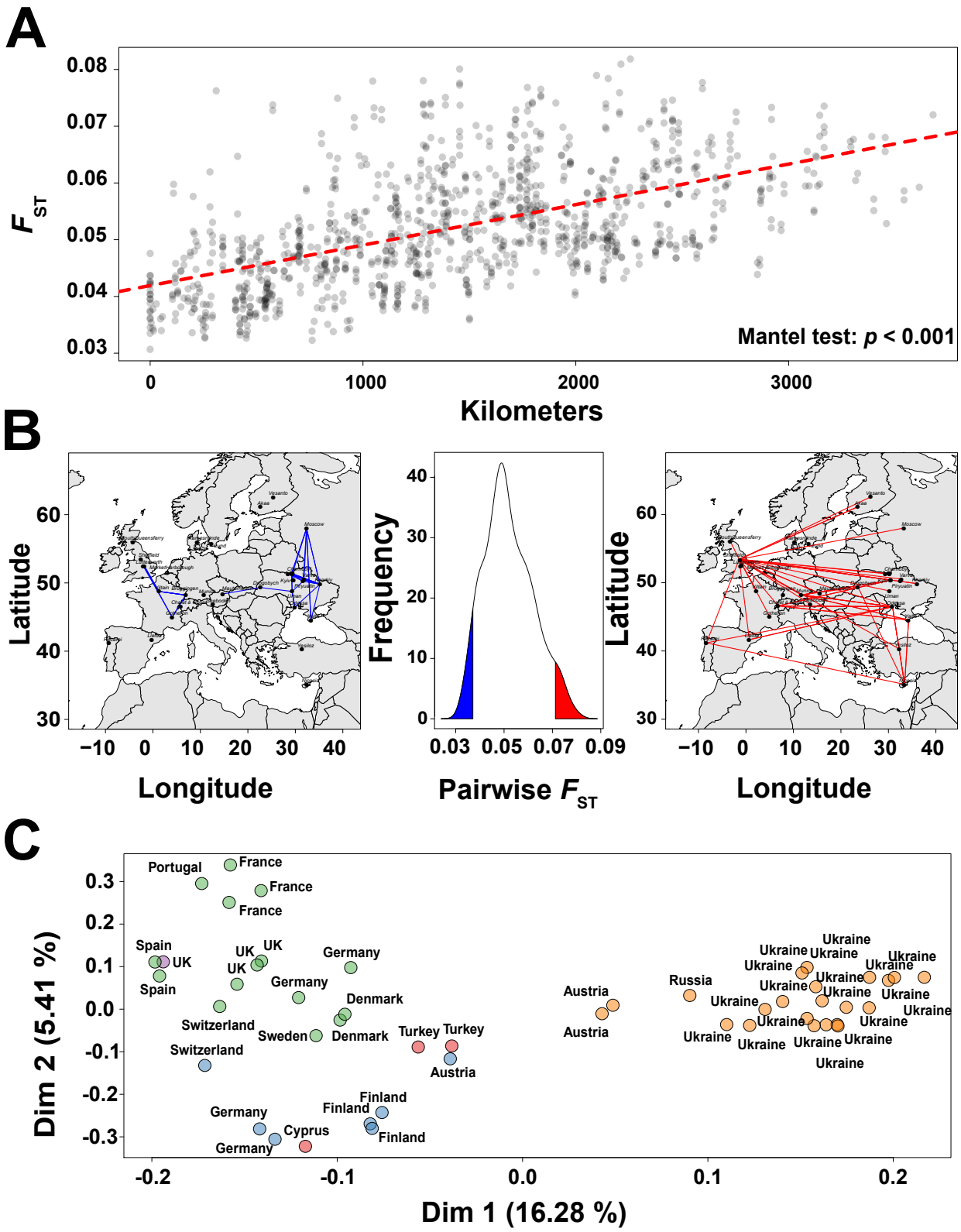


Figure 14 with 2 supplements. **Genetic differentiation among European populations**. (A) isolation by distance estimated by average genetic differentiation (F_{ST}) of 21,008 SNPs located in short introns (<60 bp) plotted against geographic distance. Mantel tests and linear regression (red dashed line and statistics in upper left box) indicate

significance. (B) Average neutral F_{ST} among populations. The center plot shows the distribution of average neutral F_{ST} values for all 1,128 pairwise combinations. Mean neutral F_{ST} values were calculated by averaging individual F_{ST} values from 20,008 genome-wide intronic SNPs for each pairwise comparison. The plots on the left and the right show population pairs in the lower (blue) and upper (red) 5% tails of the F_{ST} distribution. (C) Population structure of all *DrosEU* samples as determined by PCA of allele frequencies of 20,008 SNPs located in short introns (< 60 bp). The optimal number of five clusters was estimated by hierarchical model fitting using the first four principal components. Cluster assignment of each population, which was estimated by k-means clustering, is indicated by color.

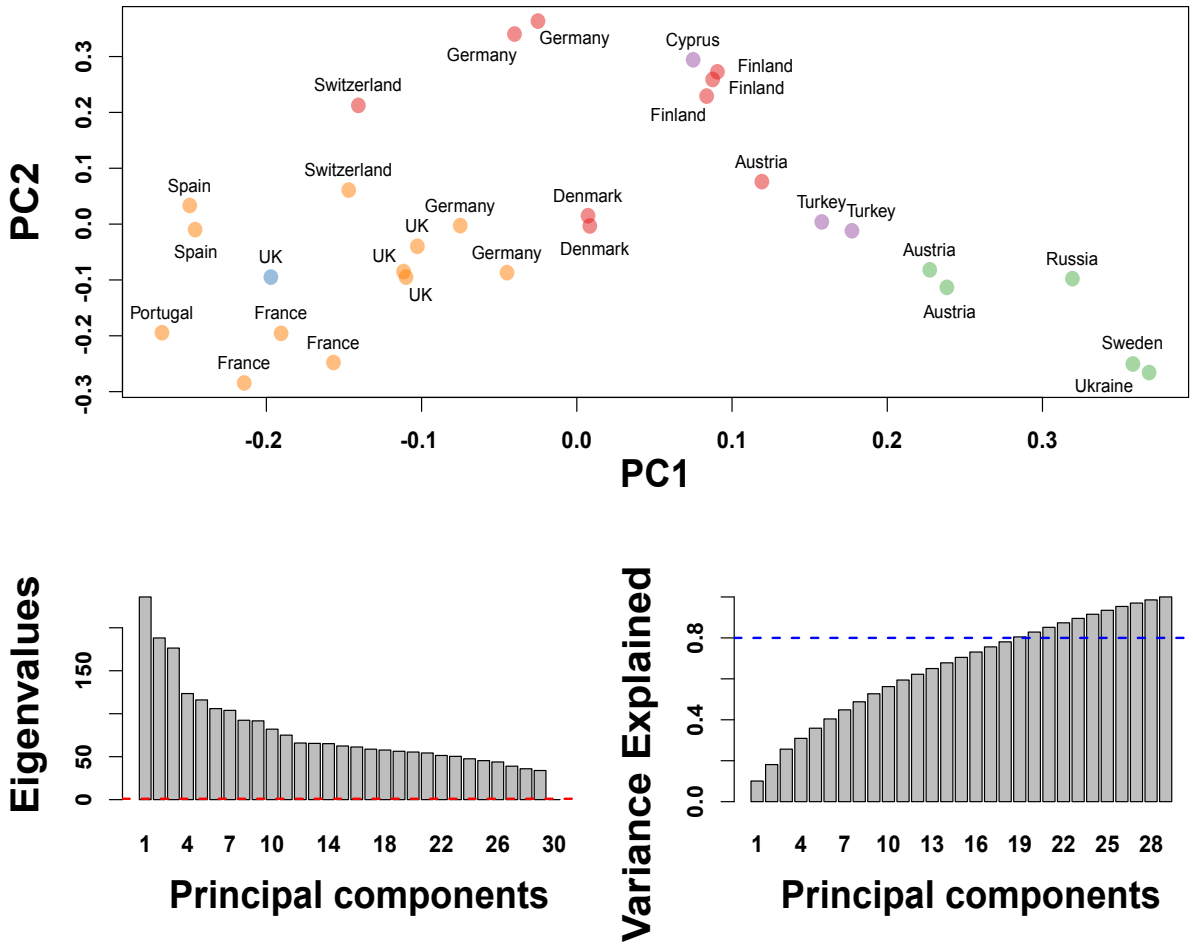
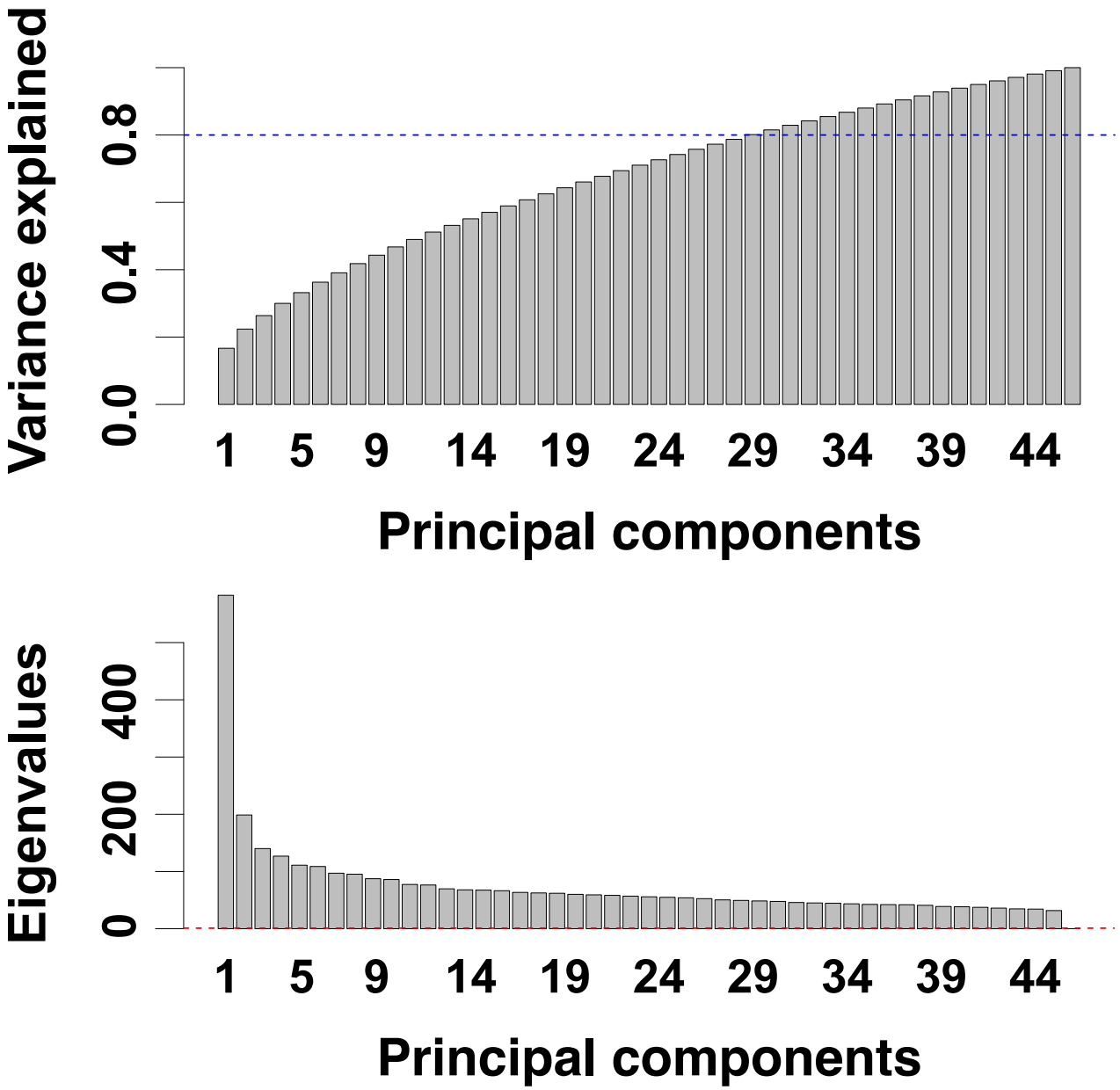


Figure 14 - figure supplement 1: Genetic differentiation among European populations. Similar to Figure 14, the top plot shows the population structure of all *DrosEU* samples as determined by PCA of allele frequencies of 20,008 SNPs located in short introns (< 60 bp) but only on a reduced dataset with one randomly drawn Ukrainian sample to test if the longitudinal pattern found in the full dataset was an artifact of the excessive sampling in the Ukraine. The bottom left plot shows the cumulative variance explained by each of the principal components (PC) and the bottom right barplot depicts

559 the Eigenvalues of each PC.

560



561

562 **Figure 14 - figure supplement 2:** PCA of neutrally evolving SNPs in the 48 DrosEU samples. The top plot shows the
563 cumulative variance explained by each of the principal components (PC) from a PCA based on the allele frequencies of
564 20,008 putatively neutrally evolving SNPs located in short (<60bp) introns distant from chromosomal inversions. The
565 bottom barplot depicts the Eigenvalues of each PC.

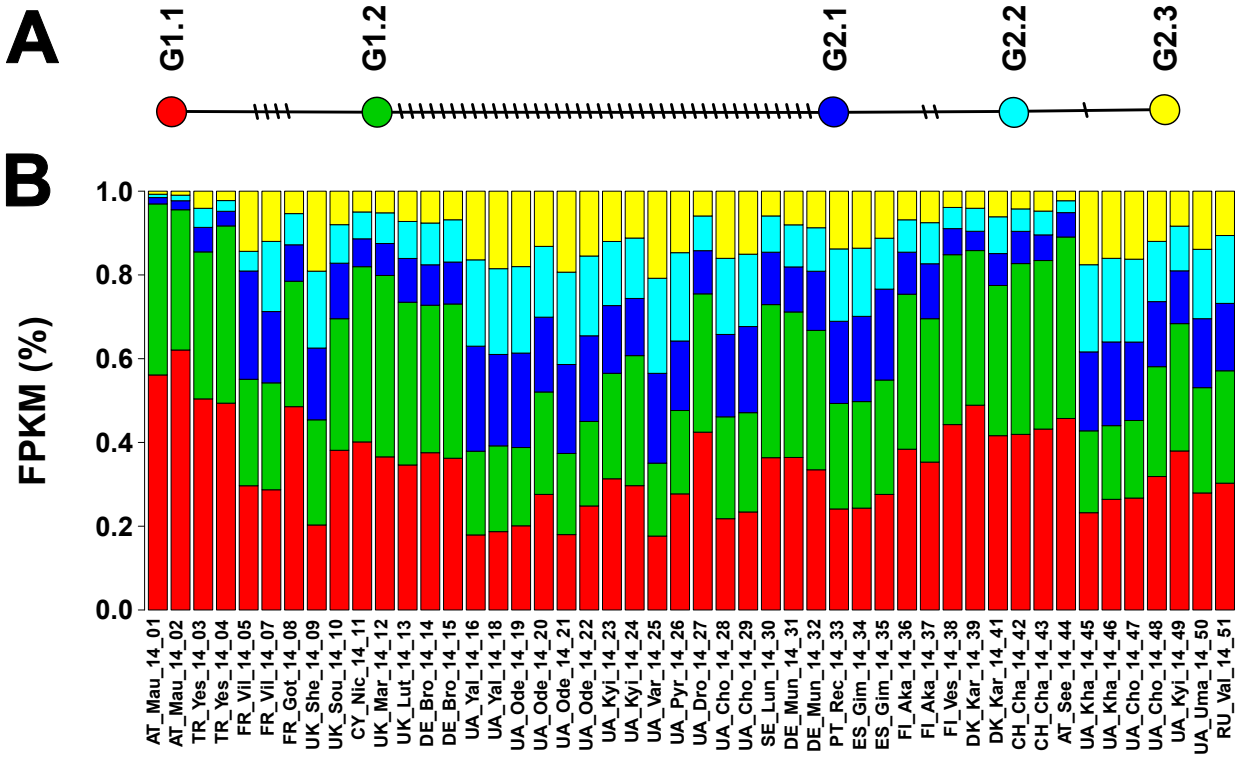
566

To further explore these patterns, we performed a principal component analysis (PCA) on the allele frequencies of SNPs in short introns. The first three principal components (PC) explained more than 25% of the total variance (PC1: 16.3%, PC2: 5.4%, PC3: 4.8%, eigenvalues = 599.2, 199.1, and 178.5 respectively; Figure 14C and Figure 14 - figure supplement 2). As expected, PC1 was strongly correlated with longitude. Despite significant signals of autocorrelation, as indicated by Moran's test on residuals from linear regressions with PC1, the association with longitude was not due to spatial autocorrelation, since a spatial error model also resulted in a significant association. PC2 was similarly, but to a lesser extent, correlated with longitude and also with altitude. PC3, by contrast, was not associated with any variable examined (Table 2). None of the major PC axes were correlated with season, indicating that there were no shared seasonal differences across samples in our dataset. Hierarchical model fitting based on the first three PC axes resulted in five distinct clusters (Figure 14C) that were oriented along the axis of PC1, supporting the notion of strong longitudinal differentiation among European populations. To the best of our knowledge, such a pronounced longitudinal signature of differentiation has not previously been reported in European *D. melanogaster*. Remarkably, this pattern is qualitatively similar to that observed for human populations (Cavalli-Sforza 1966; Xiao *et al.* 2004; Francalacci & Sanna 2008), perhaps consistent with co-migration of this commensal species.

Mitochondrial haplotypes also exhibit longitudinal population structure

Our finding that European populations are longitudinally structured is also supported by an analysis of mitochondrial haplotypes. We identified two main mitochondrial haplotypes in Europe, separated by at least 41 mutations (between G1.2 and G2.1; Figure 15A). Our findings are consistent with similar analyses of mitochondrial haplotypes from a North

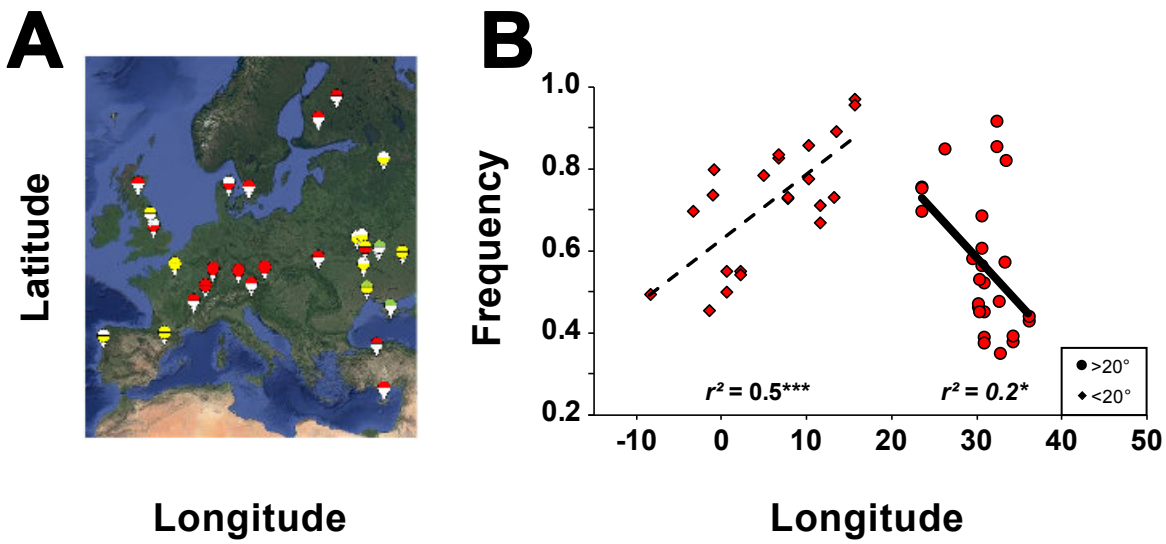
592 American *D. melanogaster* population (Cooper *et al.* 2015) as well as from worldwide
593 samples (Wolff *et al.* 2016), revealing varying degrees of differentiation among haplotypes,
594 ranging from only a few to hundreds of substitutions. The two G1 subtypes (G1.1 and
595 G1.2) are separated by only four mutations, and the three G2 subtypes are separated by a
596 maximum of four mutations (between G2.1 and G2.3). The estimated frequency of these
597 haplotypes varied greatly among populations (Figure 15B). Qualitatively, three types of
598 European populations can be distinguished based on these haplotypes, namely those with
599 (1) a high frequency (> 60%) of the G1 haplotypes, characteristic of central European
600 samples, (2) a low frequency (< 40%) of G1 haplotypes, a pattern common for Eastern
601 European populations in summer, and (3) a combined frequency of G1 haplotypes
602 between 40-60%, which is typical of samples from the Iberian Peninsula and from Eastern
603 Europe in fall (Figure 15 - figure supplement 1).
604



605
606 **Figure 15** with 1 supplement**. Mitochondrial haplotypes.** (A) TCS network showing the relationship of 5 common

607 mitochondrial haplotypes; (B) estimated frequency of each mitochondrial haplotype in 48 European samples.

608



609

610 **Figure 15 - figure supplement 1. Mitochondrial haplotypes.** (A) Graphical summary of the combined frequency of G1
611 haplotypes in Europe. Summer and Fall are represented at the top and bottom of the circles, respectively. White – no
612 information; green, yellow and red represent a combined frequency of G1 haplotypes lower than 40%, in between 40%
613 and 60% and higher than 60%, respectively. (B) Correlations between the combined frequency of G1 haplotypes and
614 longitude (red diamonds for populations below 20° and red circles for populations above 20°).

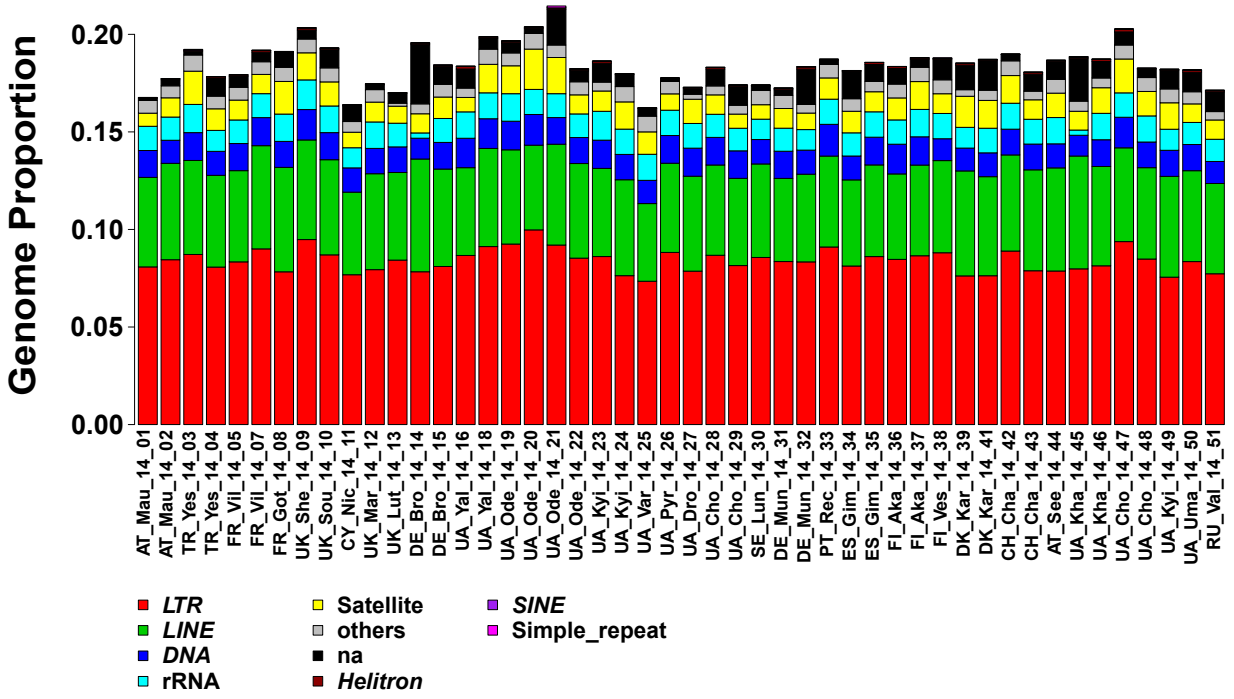
615

616 We observed a significant shift in the relative frequencies of the two haplotype classes
617 between summer and fall samples in only two of the nine possible comparisons among
618 haplotypes. While there was no correlation between latitude and the combined frequency
619 of G1 haplotypes, we found a weak but significant negative correlation between G1
620 haplotypes and longitude (r^2 = 0.10; p < 0.05), which is consistent with the longitudinal
621 east-west population structure observed for intronic SNPs. In a subsequent analysis, we
622 divided the dataset at 20º longitude into an eastern and a western subset since in northern
623 Europe 20º longitude corresponds to the division of two major climatic zones, namely C

(temperate) and D (cold), according to the Köppen-Geiger climate classification (Peel *et al.* 2007). When splitting the populations in a western (longitude < 20º E) and an eastern group (longitude > 20° E), we found a clear correlation between longitude and the combined frequency of G1 haplotypes, explaining as much as 50% of the variation in the western group (Figure 15 - figure supplement 1B). Similarly, in the eastern populations longitude and the combined frequency of G1 haplotypes were correlated, explaining approximately 20% of the variance (Figure 15 - figure supplement 1B). Thus, our data on mitochondrial haplotypes clearly confirm the existence of pronounced east-west population structure and differentiation in European *D. melanogaster*. While this might be due to climatic selection, as recently found for clinal mitochondrial haplotypes in Australia (Camus *et al.* 2017), we can presently not rule out an effect of demography.

The majority of TEs vary with longitude and altitude

To examine the population genetics of structural variants in our data, we first focused on transposable elements (TEs). The repetitive content of the 48 samples analyzed ranged from 16% to 21% with respect to nuclear genome size (Figure 16). The vast majority of detected repeats were TEs, mostly represented by long terminal repeats (LTR) and long interspersed nuclear elements (LINE; Class I), as well as a few DNA elements (Class II). LTR content best explained total TE content (LINE+LTR+DNA) (Pearson's $r = 0.87$, $p < 0.01$, *vs.* DNA $r = 0.58$, $p = 0.0117$, and LINE $r = 0.36$, $p < 0.01$ and Figure S16A).

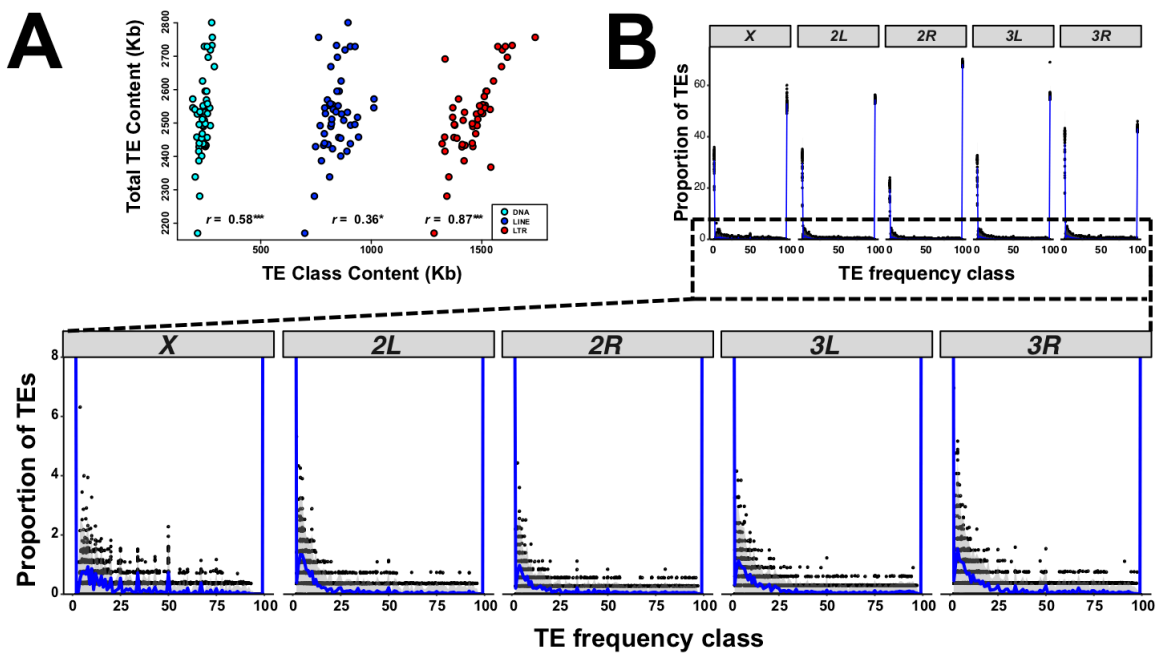


645

646 **Figure 16** with 1 supplement. **Transposable elements.** Relative abundances of repeats among samples. Proportion of

647 each repeat class was estimated from sampled reads with dnaPipeTE (2 samples per run, 0.1X coverage per sample).

648



649

650 **Figure 16 - figure supplement 1**. **Transposable Elements.** (A) shows the contribution of each of the main TE classes
651 to variation of the total TE content. Correlations (Pearson's correlation tests) between each of the three main TE classes
652 (LTR, LINE and DNA) and the total TE content of each pool (LTR+LINE+DNA) in Kb. (B) The site frequency spectrum of
653 TE frequencies per chromosome arm. Each dot represents the proportion of TEs in each bin per sample and a smoother
654 geometric line had been added to highlight the trend. Lower panel is a zoom in of the above panel.

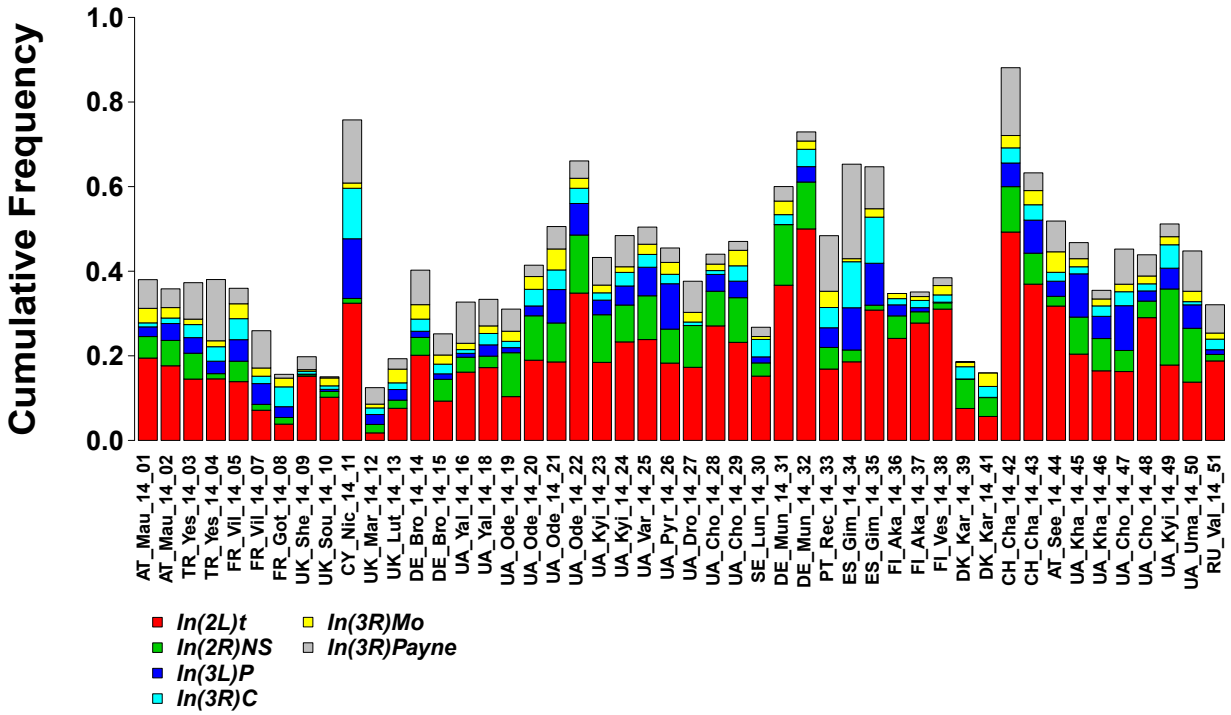
655

656 We next estimated population-wise frequencies of 1,630 TE insertions annotated in the *D.*
657 *melanogaster* reference genome *v.6.04* using *T-lex2* (Table S7, Fiston-Lavier *et al.* 2010).
658 On average, 56% of the TEs annotated in the reference genome were fixed in all samples.
659 The remaining polymorphic TEs usually segregated at low frequency in all samples (Figure
660 16 - figure supplement 1A), potentially due to the effect of purifying selection (González *et*
661 *al.* 2008; Petrov *et al.* 2011; Kofler *et al.* 2012; Cridland *et al.* 2013; Blumenstiel *et al.*
662 2014). However, we also observed 142 TE insertions present at intermediate (>10% and
663 <95%) frequencies (Figure 16 - figure supplement 1B), which might be consistent with
664 transposition-selection balance (Charlesworth *et al.* 1994).

In each of the 48 samples TE frequency and recombination rate were negatively correlated on a genome-wide level (Spearman rank sum test; $p < 0.01$), as previously reported (Bartolomé *et al.* 2002; Petrov *et al.* 2011; Kofler *et al.* 2012). This pattern still holds when only polymorphic TEs (population frequency <95%) are analyzed, although it becomes statistically non-significant for some chromosomes and populations (Table S8). In either case, the correlation is more negative when using broad-scale, rather than fine-scale, recombination rate estimates (Materials and methods, Tables S8B, S8D). This indicates that broad-scale recombination patterns may best capture long-term population recombination patterns.

We further tested whether the distribution of TE frequencies among samples could be explained by geographical or temporal variables. We focused on the 141 TE insertions that showed frequency variability among samples (interquartile range, (IQR) > 10; see Materials and Methods). Of these, 73 TEs showed significant associations with geographical or temporal variables after multiple testing correction (Table S9). Note that we used a conservative p-value threshold (< 0.001), and we did not find significant residual spatio-temporal autocorrelation among samples for any TE tested (Moran's I > 0.05 for all tests; Table S9). 16 out of 73 TEs were located in regions of very low recombination (0 cM/Mb for either of the two recombination measures used). Among the 57 significant TEs located in high recombination regions, we observed significant correlations of 13 TEs with longitude, 13 with altitude, 5 with latitude, and 3 with season (Table S9). In addition, the frequencies of the other 23 insertions were significantly correlated with more than one of the above-mentioned variables (Table S9). These significant TEs were scattered along the main five chromosome arms (Table S9). Among

690 the 57 significant TEs located in high recombination regions two TE families were enriched
691 (χ^2 p-values after Yate's correction < 0.05): the LTR *297* family with 11 copies, and the
692 DNA *pogo* family with 5 copies (Table S10). We also checked the genomic localization of
693 the 57 TEs. Most of them (42) were located inside genes: two in 5'UTR, four in 3'UTR, 18
694 in the first intron, and 18 TEs in subsequent introns. Additionally, 7 TEs are <1 kb from the
695 nearest gene, indicating that these might potentially affect the regulation of nearby genes
696 (Table S9). Interestingly, 14 of these 57 TEs coincide with previously identified candidate
697 adaptive TEs (Table S9), suggesting that our dataset might be enriched for adaptive
698 insertions. However, further analyses are needed to discard the effect of non-selective
699 forces on the patterns observed.
700
701 **Inversion polymorphisms in Europe exhibit latitudinal and longitudinal clines**
702 Chromosomal inversions are another class of important and common structural genomic
703 variants, often exhibiting frequency clines on multiple continents, some of which have been
704 shown to be adaptive (e.g. Knibb 1982; Umina *et al.* 2005; Kapun *et al.* 2014; 2016a).
705 However, little is known yet about the spatial distribution and clinality of inversions in
706 Europe. We used a panel of inversion-specific marker SNPs (Kapun *et al.* 2014) to
707 examine the presence and frequency of six cosmopolitan inversion polymorphisms
708 (*In(2L)t, In(2R)NS, In(3L)P, In(3R)C, In(3R)Mo, In(3R)Payne*) in the 48 samples. All
709 populations were polymorphic for one or more inversions (Figure 17). However, only
710 *In(2L)t* segregated at substantial frequencies in most populations (average frequency =
711 20.2%). All other inversions were either absent or occurred at low frequencies (average
712 frequencies: *In(2R)NS* = 6.2%, *In(3L)P* = 4%, *In(3R)C* = 3.1%, *In(3R)Mo* =2.2%,
713 *In(3R)Payne* = 5.7%).
714



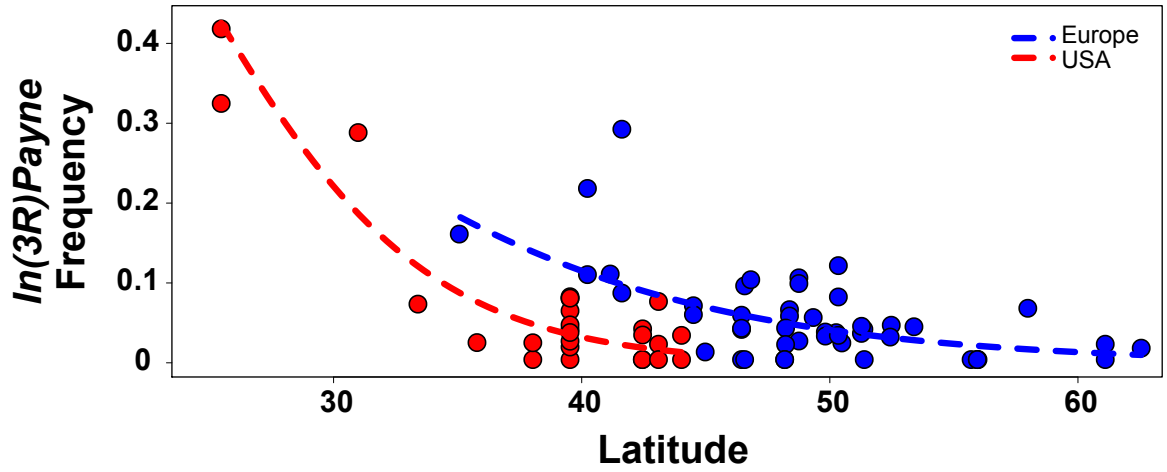
715

716 **Figure 17** with 1 supplement. **Distribution of inversion frequencies.** Cumulative bar plots showing the absolute

717 frequencies of six cosmopolitan inversions (*In(2L)t, In(2R)NS, In(3L)P, In(3R)C, In(3R)Mo, In(3R)Payne*) in all 48

718 population samples of the *DrosEU* dataset.

719



720

721 **Figure 17 - figure supplement 1. Clinal variation of the inversion *In(3R)Payne* across continents.** Parallel

722 frequency clines of *In(3R)Payne* along the latitudinal axis at the North American east coast (red) and in Europe (blue)
723 (see also Table S10).

724

725 Despite their overall low frequencies, several inversions exhibited clinal patterns across
726 space (Table 3). We observed significant latitudinal clines for *In(3L)P, In(3R)C* and
727 *In(3R)Payne*. Although they differed in overall frequencies, *In(3L)P* and *In(3R)Payne*
728 showed latitudinal clines in Europe that are qualitatively similar to the clines previously
729 observed along the North American and Australian east coasts (Figure S17 and Table
730 S11, Kapun *et al.* 2016a). For the first time, we also detected a longitudinal cline for *In(2L)t*
731 and *In(2R)NS*, with both inversions decreasing in frequency from east to west, a result that
732 is consistent with our finding of strong longitudinal among-population differentiation in
733 Europe. *In(2L)t* also increased in frequency with altitude (Table 3). Except for *In(3R)C,* we
734 did not find significant residual spatio-temporal autocorrelation among samples for any
735 inversion tested (Moran's $I \approx 0$, $p > 0.05$ for all tests; Table 3), suggesting that our analysis
736 was not confounded by spatial autocorrelation for most of the inversions. It will clearly be
737 interesting to examine the extent to which clines in inversions (and other genomic variants)
738 across Europe are shaped by selection and/or demography in future work.

739

740 **European *Drosophila* microbiomes contain trypanosomatids and novel viruses**

741 We were also interested in determining the abundance of microbiota associated with *D.*
742 *melanogaster* from the Pool-Seq data – these endosymbionts often have crucial functions
743 in affecting the life history, immunity, hormonal physiology, and metabolic homeostasis of
744 their fly hosts (e.g., Trinder *et al.* 2017; Martino *et al.* 2017). The taxonomic origin of a total
745 of 262 million non-*Drosophila* reads was inferred using MGRAST, which identifies and
746 counts short protein motifs ('features') within reads (Meyer *et al.* 2008). The largest fraction

of protein features was assigned to *Wolbachia* (on average 53.7%; Figure 18), a well-known endosymbiont of *Drosophila* (Werren *et al.* 2008). The relative abundance of *Wolbachia* protein features varied strongly between samples ranging from 8.8% in a sample from the UK to almost 100% in samples from Spain, Portugal, Turkey and Russia (Table 1). Similarly, *Wolbachia* loads varied 100x between samples if we use the ratio of *Wolbachia* protein features divided by the number of *Drosophila* sequences retrieved for that sample as a proxy for relative micro-organismal load (for a full table of micro-organismal loads standardized by *Drosophila* genome coverage see Table S12).

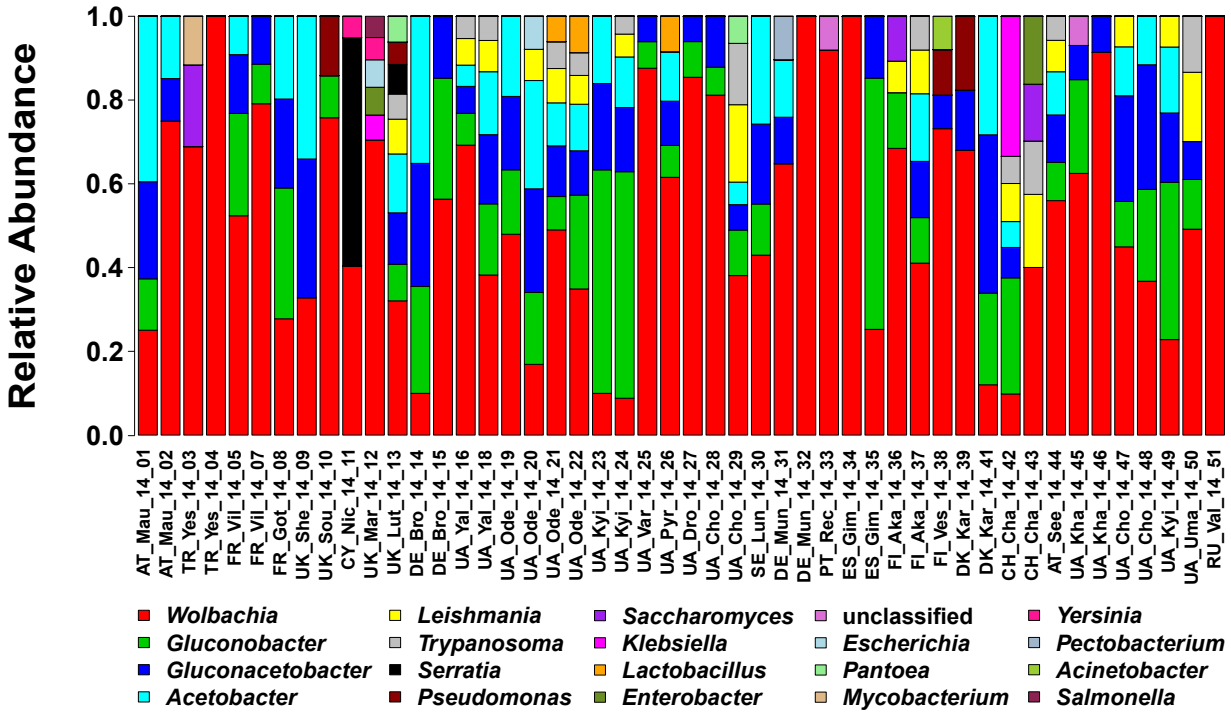


Figure 18: Microbiome. Relative abundance of *Drosophila*-associated microbes as assessed by MGRAST classified shotgun sequences. Microbes had to reach at least 3% relative abundance in one of the samples to be presented

Acetic acid bacteria of the genera *Gluconobacter*, *Gluconacetobacter*, and *Acetobacter* were the second largest group, with an average relative abundance of 34.4%.

762 Furthermore, we found evidence for the presence of several genera of Enterobacteria
763 (*Serratia, Yersinia, Klebsiella, Pantoea, Escherichia, Enterobacter, Salmonella*, and
764 *Pectobacterium*). *Serratia* occurs only at low frequencies or is absent from most of our
765 samples, but reaches a very high relative abundance in the Nicosia summer collection
766 (54.5%). This high relative abundance was accompanied by an 80x increase in *Serratia*
767 bacterial load. We detected several eukaryotic microorganisms, although they were less
768 abundant than the bacteria. The fraction of fungal protein features is larger than 3% in only
769 three of our samples from Finland, Austria and Turkey (Table 1). Interestingly, we detected
770 the presence of trypanosomatids in 16 of our samples, consistent with other recent
771 evidence that *Drosophila* can host these organisms (Wilfert *et al.* 2011; Chandler & James
772 2013; Hamilton *et al.* 2015).
773
774 Our data also allowed us to detect the presence of five different DNA viruses (Table S13).
775 These included approximately two million reads from *Kallithea* nudivirus (Webster *et al.*
776 2015), allowing us to assemble the complete *Kallithea* genome for the first time (>300-fold
777 coverage in the Ukrainian sample UA_Kha_14_46; Genbank accession KX130344). We
778 also identified around 1,000 reads from a novel nudivirus that is closely related to *Kallithea*
779 virus and to *Drosophila innubila* nudivirus (Unckless 2011) in sample DK_Kar_14_41 from
780 Karensminde, Denmark (Table 1). These sequences permitted us to identify a publicly
781 available dataset (SRR3939042: 27 male *D. melanogaster* from Esparto, California;
782 Machado *et al.* 2016) that contained sufficient reads to complete the genome (provisionally
783 named "*Esparto* Virus"; KY608910). We further identified two novel Densoviruses
784 (*Parvoviridae*), which we have provisionally named "*Viltain* virus", a relative of *Culex*
785 *pipiens* densovirus found at 94-fold coverage in sample FR_Vil_14_07 (Viltain; KX648535)
786 and "*Linvill Road* virus", a relative of *Dendrolimus punctatus* densovirus that was

represented by only 300 reads here, but which has previously been found to have a high coverage in dataset SRR2396966 from a North American sample of *D. simulans* (KX648536; Machado *et al.* 2016). In addition, we detected a novel member of the *Bidnaviridae* family,"*Vesanto* virus", a bidensovirus related to *Bombyx mori* densovirus 3 with approximately 900-fold coverage in sample FI_Ves_14_38 (Vesanto; KX648533 and KX648534), Using a detection threshold of >0.1% of the *Drosophila* genome copy number, the most commonly detected viruses were *Kallithea* virus (30/48 of the pools) and *Vesanto* virus (25/48), followed by *Linvill Road* virus (7/48) and *Viltain* virus (5/48), with *Esparto* virus being the rarest (2/48). In some samples, the viruses reached strikingly high titers: on 13 occasions the virus genome copy number in the pool exceeded the host genome copy number, reaching a maximum of nearly 20-fold in Vesanto.

This continent-wide analysis of the microbiota associated with fruit flies suggests that natural populations of European *D. melanogaster* differ greatly in the composition and relative abundance of microbes and viruses.

Discussion

In recent years, large-scale population resequencing projects have shed light on the biology of both model (Mackay *et al.* 2012; Langley *et al.* 2012; Consortium 2015; Lack *et al.* 2015; Alonso-Blanco *et al.* 2016; Lack *et al.* 2016) and non-model organisms (e.g., Hohenlohe *et al.* 2010; Wolf *et al.* 2010). Such massive datasets contribute greatly to our growing understanding of the processes that create and maintain genetic variation in natural populations. However, the relevant spatio-temporal scales for population genomic analyses remain largely unknown. Here we have applied, for the first time, a comprehensive sampling and sequencing strategy to European populations of *D.*

812 *melanogaster*, allowing us to uncover previously unknown aspects of this species'
813 population biology.
814
815 A main result from our analyses of SNPs located in short introns and presumably evolving
816 neutrally (Parsch *et al.* 2010) is that European *D. melanogaster* populations exhibit very
817 pronounced longitudinal differentiation, a pattern that – to the best of our knowledge – has
818 not been observed before for the European continent (for patterns of longitudinal
819 differentiation in Africa see e.g. Michalakis & Veuille 1996; Aulard *et al.* 2002; Fabian *et al.*
820 2015). Genetic differentiation was greatest between populations from eastern and western
821 Europe (Figure 14). The eastern populations included those from the Ukraine, Russia, and
822 Turkey, as well as one from eastern Austria, suggesting that there may be a region of
823 restricted gene flow in south-central Europe. However, populations from Finland and
824 Cyprus are more similar to western populations than to eastern populations, possibly as a
825 result of migration along shipping routes in the Baltic and Mediterranean seas. More data
826 from populations in the unsampled, intermediate regions are needed to better delineate
827 the geographic limits of the eastern and western population groups. Consistent with the
828 strong differentiation between eastern and western populations, our PCA analysis
829 revealed that longitude was the major factor associated with among-population
830 divergence, with no significant effect of latitude (Figure 14C; Table 2). Thus, the patterns
831 of neutral genetic differentiation in Europe contrast with those previously reported for North
832 America, where latitude impacts neutral differentiation (Machado *et al.* 2016; Kapun *et al.*
833 2016a). However, our present analysis does not exclude the existence of clinally varying
834 polymorphisms in European populations outside short introns: for example, we detected
835 latitudinal frequency clines both for TEs and inversion polymorphisms. A detailed analysis
836 of genome-wide patterns of clinal variation in the 2014 *DrosEU data* is beyond the scope

837 of this paper and currently under way.

838

839 The mitochondrial genome and several chromosomal inversions and TEs showed similar
840 patterns of differentiation as the rest of the genome, with the main axis of differentiation
841 being longitudinal. Uncovering the extent to which this pattern is driven by demography
842 and/or selection, and identifying the underlying environmental correlates (including any
843 potential role of co-migration with human populations), will be an important task for future
844 analyses. Due to the high density of samples and the large number of SNP markers
845 examined, our results reveal that European populations of *D. melanogaster* exhibit much
846 more differentiation and structure than previously thought (e.g., Baudry *et al.* 2004;
847 Dieringer *et al.* 2005; Schlötterer *et al.* 2006; Nunes *et al.* 2008; Mateo *et al.* 2018).

848

849 Within the eastern and western population groups there was a low – but detectable – level
850 of genetic differentiation among populations, including those that are geographically close
851 (Figure 14C). These population differences persisted over a timespan of at least 2–3
852 months, as there was less genetic differentiation between the summer and fall samples of
853 the 13 locations sampled at multiple time points than between neighboring populations
854 (Figure 14C). Thus, while the weak but significant signal of IBD suggests homogenizing
855 gene flow across geography, there is seasonally stable differentiation among populations.
856 The season in which samples were collected did not show a significant association with
857 genetic differentiation, except when considered in conjunction with longitude or altitude
858 (Table 2). However, the data analyzed here are from a single year only: demonstrating
859 recurrent shifts in SNP frequencies due to temporally varying selection will require analysis
860 of additional annual samples. For an extensive analysis of patterns of seasonal variation
861 across a broad geographic scale see Machado *et al.* (2018)

Our Pool-Seq data also allowed us to characterize geographic patterns in both inversions and TEs. In marked contrast to putatively neutral SNPs, the frequencies of several chromosomal inversions, including *In(3L)P, In(3R)C,* and *In(3R)Payne*, showed a significant correlation with latitude (Table 3). For *In(3L)P* and *In(3R)Payne*, the latitudinal clines were in qualitative agreement with parallel clines reported from North America and Australia, with the inversions decreasing in frequency as distance from the equator increases (Mettler *et al.* 1977; Knibb *et al.* 1981; Fabian *et al.* 2012; Kapun *et al.* 2014; Rane *et al.* 2015; Kapun *et al.* 2016a). This suggests that these inversions may contain genetic variants that are better adapted to warmer environments than to temperate climates. The overall frequencies of these inversions are, however, low in Europe (<5%), indicating that they might play only a minor role in local adaptation to European habitats. Some euchromatic TE insertions also showed geographic or seasonal patterns of variation (Table S7), indicating that they might play a role in local adaptation, particularly as many of them are located in regions where they could affect gene regulation. Importantly, several inversions and TEs also showed longitudinal frequency gradients, thus supporting the notion that European populations exhibit marked longitudinal differentiation.

We also examined signatures of selective sweeps in our dataset. We found 144 genomic regions that showed signatures of hard sweeps in regions of normal recombination (cM/Mb ≥ 0.5), and with reduced variation and negative Tajima's *D* (*D* ≤ -0.8) in all European populations (Figure 12, Table S6). Four of these regions were identified in previous studies as potential targets for positive selection.

The first region, at the center of chromosome arm *2R* (Figure 12A, Table S6), was

887 previously found to be strongly differentiated between African and North American
888 populations (Langley *et al.* 2012) and contains two genes, *Cyp6g1* and *Hen-1*, that are
889 associated with recent, strong selection. The cytochrome P450 gene *Cyp6g1* has been
890 linked to insecticide resistance (Daborn *et al.* 2002; Schmidt *et al.* 2010), shows evidence
891 for recent selection independently in both *D. melanogaster* and *D. simulans* (Schlenke &
892 Begun 2003; Catania *et al.* 2004), and is associated with a large differentiated region in
893 the Australian latitudinal cline (Kolaczkowski *et al.* 2011a). *Hen-1*, a methyltransferase
894 involved in maturation of small RNAs involved in virus and TE suppression, showed
895 marginally non-significant evidence for selective sweeps in North American and African
896 populations of *D. melanogaster* (Kolaczkowski *et al.* 2011b).
897
898 The second region previously implicated in a selective sweep is located on chromosome
899 arm *3L* (Figure 12B, Table S6) and centered around the chimeric gene *CR18217*, which
900 formed from the fusion of a gene encoding a DNA-repair enzyme (*CG4098*) and a
901 centriole gene (spd-2; Rogers & Hartl 2012). *CR18217* appears to be unique to *D.*
902 *melanogaster*, but – in spite of its recent origin – segregates at frequencies of around 90%
903 (Rogers & Hartl 2012), consistent with a recent strong sweep in this region of the genome.
904 This putative sweep region also spans *Prosbeta6*, which (like HDAC) encodes a gene
905 involved in proteolysis (Flybase v. FB2017_05; Gramates *et al.* 2017). *Prosbeta6* also
906 shows homology to genes involved in immune function (Lyne *et al.* 2007; Handu *et al.*
907 2015), which might explain why it has been a target of positive selection.
908
909 The third previously characterized sweep region, surrounding the *wapl* gene on the *X*
910 chromosome (Table S6), was identified as showing evidence of strong selective sweeps in
911 both African and European *D. melanogaster* populations (Beisswanger *et al.* 2006; Boitard

912 *et al.* 2012). The genic targets of selection in this region are unclear, but most likely are
913 *ph-p* in Europe and *ph-p* or *ph-d* in Africa (Beisswanger *et al.* 2006). These genes are
914 tandem duplicates involved in the Polycomb response pathway, which functions as an
915 epigenetic repressor of transcription (reviewed in Kassis *et al.* 2017).
916
917 The fourth previously observed sweep region, originally identified in African populations of
918 *D. melanogaster*, is also located on the *X* chromosome (Table S6), but 30 cM closer to the
919 telomere and thus not implicating the *wapl* region (Beisswanger *et al.* 2006; Boitard *et al.*
920 2012). Selection in this region has been attributed to the *HDAC6* gene (Svetec *et al.*
921 2009). *HDAC6*, although nominally a histone deacetylase, actually functions as a central
922 player in managing cytotoxic assaults, including in transport and degradation of misfolded
923 protein aggregates (reviewed in Matthias *et al.* 2008; Svetec *et al.* 2009).
924
925 Our data support the widespread occurrence of these previously identified sweeps in many
926 populations in Europe. Notably, practically all European populations examined showed
927 reduced variation and negative Tajima's *D* in these sweep regions. This is consistent with
928 the sweeps either pre-dating the colonization of Europe (e.g., Beisswanger *et al.* 2006) or
929 having swept across Europe more recently (also see Stephan 2010 for discussion). In
930 addition, we also uncovered several novel genomic regions with tentative evidence for
931 hard sweeps (Table S6) – these regions represent a valuable source for future analyses of
932 signals of adaptive evolution in European *Drosophila.*
933
934 Finally, we used our Pool-Seq data to identify microbes and viruses and to quantify their
935 presence in natural populations of *D. melanogaster* across the European continent.
936 *Wolbachia* was the most abundant bacterial genus associated with the flies, but its relative

937 abundance and load varied greatly among samples (Figure 18). The second most
938 abundant bacterial taxon was acetic acid bacteria (*Acetobacteraceae*), a group previously
939 found among the most abundant bacteria in natural *D. melanogaster* isolates (Chandler *et*
940 *al.* 2011; Staubach *et al.* 2013). Other microbes were highly variable abundance in relative
941 abundance. For example, *Serratia* abundance was low in most populations, but very high
942 in the Nicosia sample, which might reflect that there are individuals in the Nicosia sample
943 that carry a systemic *Serratia* infection generating high bacterial loads. Future sampling
944 may shed light on the temporal stability and/or population specificity of these patterns.
945 Contrary to expectation, we found relatively few yeast sequences. This is a bit surprising
946 because yeasts are commonly found on rotting fruit, the main food substrate of *D.*
947 *melanogaster*, and have been found in association with *Drosophila* before (Barata *et al.*
948 2012; Chandler *et al.* 2012). This suggests that, although yeasts can attract flies and play
949 a role in food choice (Becher *et al.* 2012; Buser *et al.* 2014), they might not be highly
950 prevalent in or on *D. melanogaster* bodies. While trypanosomatids have been reported in
951 association with *Drosophila* before (Wilfert *et al.* 2011; Chandler & James 2013; Hamilton
952 *et al.* 2015), our study provides the first systematic detection across a wide geographic
953 range in *D. melanogaster*. Despite being host to a wide diversity of RNA viruses (Huszar &
954 Imler 2008; Webster *et al.* 2015), only three DNA viruses have previously been reported in
955 association with *Drosophilidae*, and only one from *D. melanogaster* (Unckless 2011;
956 Webster *et al.* 2015; 2016). Here, we have discovered four new DNA viruses in *D.*
957 *melanogaster*. Although it is not possible to directly estimate viral prevalence from pooled
958 sequencing data, we found that the DNA viruses of *D. melanogaster* can be very
959 widespread, with *Kallithea* virus detectable at a low level in most populations.
960
961 A striking qualitative pattern in our microbiome data is the high level of variability among

962 populations in the composition and relative amounts of different microbiota and viruses.
963 Thus, an interesting open question is to what extent geographic differences in microbiota
964 might contribute to phenotypic differences and local adaptation among fly populations,
965 especially given that there might be tight and presumably local co-evolutionary interactions
966 between fly hosts and their endosymbionts (e.g., Haselkorn *et al.* 2009; Richardson *et al.*
967 2012; Staubach *et al.* 2013; Kriesner *et al.* 2016).
968
969 In conclusion, our study demonstrates that extensive sampling on a continent-wide scale
970 and pooled sequencing of natural populations can reveal new aspects of population
971 biology, even for a well-studied species such as *D. melanogaster*. Such extensive
972 population sampling is feasible due to the close cooperation and synergism within our
973 international consortium. Our efforts in Europe are paralleled in North America by the
974 *Drosophila* Real Time Evolution Consortium (*Dros-RTEC*), with whom we are currently
975 collaborating to compare population genomic data across continents. In future years, our
976 consortia will continue to sample and sequence European and North American *Drosophila*
977 populations in order to study these populations with increasing spatial and temporal
978 resolution and to provide an unprecedented resource for the *Drosophila* and population
979 genetics communities.
980

981 ## Materials and Methods

982 The 2014 *DrosEU* dataset analyzed here consists of 48 samples of *D. melanogaster*
983 collected from 32 geographical locations at different time-points across the European
984 continent, through a joint effort of 18 European research groups (see Figure 2, Table 1).
985 Field collections were performed with baited traps using a standardized protocol (see
986 Supplementary file for details). Up to 40 males from each collection were pooled, and DNA

extracted from each pool, using a standard phenol-chloroform based protocol. Each sample was processed in a single pool (Pool-Seq; Schlötterer *et al.* 2014), with each pool consisting of at least 33 wild-caught individuals. To exclude morphologically similar and co-occurring species, such as *D. simulans*, as potential contaminants from the samples, we only used wild-caught males and distinguished among species by examining genital morphology. Despite this precaution, we identified a low level of *D. simulans* contamination in our samples, and further steps were thus taken to exclude *D. simulans* sequences from our analysis (see below). The 2014 *DrosEU* dataset represents the most comprehensive spatio-temporal sampling of European *D. melanogaster* populations available to date (Table 1, Figure 3).

DNA extraction, library preparation and sequencing

DNA was extracted from pools of 33–40 males per sample after joint homogenization with bead beating and standard phenol/chloroform extraction. A detailed extraction protocol can be found in the Supporting Information file. In brief, 500 ng of DNA in a final volume of 55.5 µl were sheared with a Covaris instrument (Duty cycle 10, intensity 5, cycles/burst 200, time 30) for each sample separately. Library preparation was performed using NEBNext Ultra DNA Lib Prep-24 and NebNext Multiplex Oligos for Illumina-24 following the manufacturer's instructions. Each pool was sequenced as paired-end fragments on a Illumina NextSeq 500 sequencer at the Genomics Core Facility of Pompeu Fabra University (UPF; Barcelona, Spain). Samples were multiplexed in five batches of 10 samples each, except for one batch that contained only 8 samples (see Supplementary Table S1 for further information). Each multiplexed batch was sequenced on four lanes to obtain an approximate 50x raw coverage for each sample. Reads were sequenced to a length of 151 bp with a median insert size of 348 bp (ranging from 209 to 454 bp).

Mapping pipeline and variant calling

Prior to mapping, we trimmed and filtered raw FASTQ reads to remove low-quality bases (minimum base PHRED quality = 18; minimum sequence length = 75 bp) and sequencing adaptors using cutadapt (v. 1.8.3; Martin 2011). We only retained read pairs for which both reads fulfilled our quality criteria after trimming. FastQC analyses of trimmed and quality filtered reads showed overall high base-qualities (median ranging from 29 to 35 in all 48 samples) and indicated a loss of ~1.36% of all bases after trimming relative to the raw data. We used bwa mem (v. 0.7.15; Li 2013) with default parameters to map trimmed reads against a compound reference genome consisting of the genomes from *D. melanogaster* (v.6.12) and genomes from common commensals and pathogens, including *Saccharomyces cerevisiae* (GCF_000146045.2), *Wolbachia pipientis* (NC_002978.6), *Pseudomonas entomophila* (NC_008027.1), *Commensalibacter intestine* (NZ_AGFR00000000.1), *Acetobacter pomorum* (NZ_AEUP00000000.1), *Gluconobacter morbifer* (NZ_AGQV00000000.1), *Providencia burhodogranariea* (NZ_AKKL00000000.1), *Providencia alcalifaciens* (NZ_AKKM01000049.1), *Providencia rettgeri* (NZ_AJSB00000000.1), *Enterococcus faecalis* (NC_004668.1), *Lactobacillus brevis* (NC_008497.1), and *Lactobacillus plantarum* (NC_004567.2), to avoid paralogous mapping. We used Picard (v.1.109; http://picard.sourceforge.net) to remove duplicate reads and reads with a mapping quality below 20. In addition, we re-aligned sequences flanking insertions-deletions (indels) with GATK (v3.4-46; McKenna *et al.* 2010).

After mapping, Pool-Seq samples were tested for DNA contamination from *D. simulans*. To do this, we used a set of SNPs known to be divergent between *D. simulans* and *D. melanogaster* and assessed the frequencies of *D. simulans*-specific alleles following the

approach of Bastide *et al*. (2013). We combined the genomes of *D. melanogaster* (v.6.12) and *D. simulans* (Hu *et al.* 2013) and separated species-specific reads for samples with a contamination level > 1% via competitive mapping against the combined references using the pipeline described above. Custom software was used to remove reads uniquely mapping to *D. simulans*. In 9 samples, we identified contamination with *D. simulans*, ranging between 1.2 % and 8.7% (Table S1). After applying our decontamination pipeline, contamination levels dropped below 0.4 % in all 9 samples.

We used *Qualimap* (v. 2.2., Okonechnikov *et al.* 2016) to evaluate average mapping qualities per population and chromosome, which ranged from 58.3 to 58.8 (Table S1). We found heterogeneous sequencing depths among the 48 samples, ranging from 34x to 115x for autosomes and from 17x to 59x for *X*-chromosomes (Figure S1, Table S1). We then combined individual BAM files from all samples into a single *mpileup* file using samtools (v. 1.3; Li & Durbin 2009). Due to the large number of Pool-Seq datasets analyzed in parallel, we had to implement quality control criteria for all libraries jointly in order to call SNPs. To accomplish this, we implemented a novel custom SNP calling software to call SNPs with stringent heuristic parameters (PoolSNP; see Supplementary Information), available at Dryad (doi: https://doi.org/10.5061/dryad.rj1gn54). A site was considered polymorphic if (1) the minimum coverage from all samples was greater than 10x, (2) the maximum coverage from all samples was less than the 95th coverage percentile for a given chromosome and sample (to avoid paralogous regions duplicated in the sample but not in the reference), (3) the minimum read count for a given allele was greater than 20x across all samples pooled, and (4) the minimum read frequency of a given allele was greater than 0.001 across all samples pooled. The above threshold parameters were optimized based on simulated Pool-Seq data in order to maximize true

positives and minimize false positives (see Figure S18 and Supporting Information). Additionally, we excluded SNPs (1) for which more than 20% of all samples did not fulfill the above-mentioned coverage thresholds, (2) which were located within 5 bp of an indel with a minimum count larger than 10x in all samples pooled, and (3) which were located within known transposable elements (TE) based on the *D. melanogaster* TE library v.6.10. We further annotated our final set of SNPs with SNPeff (v.4.2; Cingolani *et al.* 2012) using the Ensembl genome annotation version BDGP6.82 (Figure 3).

Combined and population-specific site frequency spectra (SFS)

We quantified the amount of allelic variation with respect to different SNP classes. For this, we first combined the full dataset across all 48 samples and used the SNPeff annotation (see above) to classify the SNPs into four classes (intergenic, intronic, non-synonymous, and synonymous). For each class, we calculated the site frequency spectrum (SFS) based on minor allele frequencies for the *X*-chromosome and the autosomes, as well as for each sample and chromosomal arm separately, by counting alleles in 50 frequency bins of size 0.01.

Genetic variation in Europe

We characterized patterns of genetic variation among the 48 samples by estimating three standard population genetic parameters: π, Watterson's θ and Tajima's D (Watterson 1975; Nei 1987; Tajima 1989). We focused on SNPs located on the five major chromosomal arms (*X*, *2L*, *2R*, *3L*, *3R*) and calculated sample-wise π, θ and Tajima's D with corrections for Pool-Seq data (Kofler *et al.* 2011). Since PoPoolation, the most commonly used software for population genetics inference from Pool-Seq data, does not allow using predefined SNPs (which was desirable for our analyses), we implemented

corrected population genetic estimators described in Kofler *et al*. (2011) in Python (PoolGen; available at Dryad; doi: https://doi.org/10.5061/dryad.rj1gn54). Before calculating the estimators, we subsampled the data to an even coverage of 40x for the autosomes and 20x for the *X*-chromosome to control for the sensitivity to coverage variation of Watterson's θ and Tajima's *D* (Korneliussen *et al.* 2013). At sites with greater than 40x coverage, we randomly subsampled reads to 40x without replacement; at sites with below 40x coverage, we sampled reads 40 times with replacement. Using *R* (R Development Core Team 2009), we calculated sample-wise chromosome-wide averages for autosomes and *X* chromosomes separately and tested for correlations of π, θ and Tajima's *D* with latitude, longitude, altitude, and season using a linear regression model of the following form: $y_i = Lat + Lon + Alt + Season + \varepsilon_i$, where y_i is either π, θ and *D*. Here, latitude, longitude, and altitude are continuous predictors (Table 1), while 'season' is a categorical factor with two levels *S* ("summer") and *F* ("fall"), corresponding to collection dates before and after September 1st, respectively. We chose this arbitrary threshold for consistency with previous studies (Bergland *et al.* 2014; Kapun *et al.* 2016a). To further test for residual spatio-temporal autocorrelation among the samples (Kühn & Dormann 2012), we calculated Moran's *I* (Moran 1950) with the *R* package *spdep* (v.06-15., Bivand & Piras 2015). To do this, we used the residuals of the above-mentioned models, as well as matrices defining pairs of samples as neighbors weighted by geographical distances between the locations (samples within 10° latitude/longitude were considered neighbors). Whenever these tests revealed significant autocorrelation (indicating non-independence of the samples), we repeated the above-mentioned regressions using spatial error models as implemented in the *R* package *spdep*, which incorporate spatial effects through weighted error terms, as described above.

To test for confounding effects of variation in sequencing errors between runs, we extended the above-mentioned linear models including the run ID as a random factor using the *R* package *lme4* (v.1.1-14; see Supporting Information). Preliminary analyses showed that this model was not significantly better than simpler models, so we did not include sequencing run in the final analysis (see Supporting information and Table S4).

To investigate genome-wide patterns of variation, we averaged π, θ, and *D* in 200 kb non-overlapping windows for each sample and chromosomal arm separately and plotted the distributions in *R.* In addition, we calculated Tajima's *D* in 50 kb sliding windows with a step size of 10 kb to investigate fine-scale deviations from neutral expectations. We applied heuristic parameters to identify genomic regions harboring potential candidates for selective sweeps. To identify candidate regions with sweep patterns across most of the 48 samples, we searched for windows with log-transformed recombination rates ≥ 0.5, pairwise nucleotide diversity ($\pi \leq 0.004$), and average Tajima's *D* across all populations ≤ -0.8 (5% percentile). To identify potential selective sweeps restricted to a few population samples only, we searched for regions characterized as above but allowing one or more samples with Tajima's *D* being more than two standard deviations smaller than the window-wise average. To account for the effects of strong purifying selection in gene-rich genomic regions which can result in local negative Tajima's *D* (Tajima 1989) and thus confound the detection of selective sweeps, we repeated the analysis based on silent sites (4-fold degenerate sites, SNPs in short introns of ≤ 60 bp lengths and SNPs in intergenic regions in ≥ 2000 bp distance to the closest gene) only. Despite of the reduction in polymorphic sites available for this analysis, we found highly consistent sweep regions and therefore proceded with the full SNP datasets, which provided better resolution (results not shown).

For statistical analysis, the diversity statistics were log-transformed to normalize the data. We then tested for correlations between π and recombination rate using *R* in 100 kb non-overlapping windows and plotted these data using the *ggplot2* package (v.2.2.1., Wickham 2016). We used two different recombination rate measurements: (i) a fine-scale, high resolution genomic recombination rate map based on millions of SNPs in a small number of strains (Comeron *et al.* 2012), and (ii) the broad-scale Recombination Rate Calculator based on Marey maps generated by laboratory cross data fitting genetic and physical positions of 644 markers to a third-order polynomial curve for each chromosome arm (Fiston-Lavier *et al.* 2010). Both measurements were converted to version 6 of the *D. melanogaster* reference genome to match the genomic position of π estimates (see above).

SNP counts and overlap with other datasets

We used the panel of SNPs identified in the *DrosEU* dataset (available at Dryad; doi: https://doi.org/10.5061/dryad.rj1gn54) to describe the overlap in SNP calls with other published *D. melanogaster* population data: the *Drosophila* Population Genomics Project 3 (DPGP3) from Siavonga, Zambia (69 non-admixed lines; Lack *et al.* 2015; 2016) and the *Drosophila* Genetic Reference Panel (DGRP) from Raleigh, North Carolina, USA (205 inbred lines; Mackay *et al.* 2012; Huang *et al.* 2014). For these comparisons, we focused on biallelic SNPs on the 5 major chromosome arms. We used *bwa mem* for mapping and a custom pipeline for heuristic SNP calling (PoolSNP; Figure 3). To make the data from the 69 non-admixed lines from Zambia (Lack *et al.* 2015; 2016) comparable to our data, we reanalyzed these data using our pipeline for mapping and variant calling (Figure 3). The VCF file of the DGRP data was downloaded from http://dgrp2.gnets.ncsu.edu/ and

converted to coordinates according to the *D. melanogaster* reference genome v.6. We depicted the overlap of SNPs called in the three different populations using elliptic Venn diagrams with *eulerAPE* software (v3 3.0.0., Micallef & Rodgers 2014). While the *DrosEU* data were generated from sequencing pools of wild-caught individuals, both the DGRP and DPGP3 data are based on individual sequencing of inbreed lines and haploid individuals, respectively.

Genetic differentiation and population structure in European populations

To estimate genome-wide pairwise genetic differences, we used custom software to calculate SNP-wise F_{ST} using the approach of Weir and Cockerham (1984). We estimated SNP-wise F_{ST} for all possible pairwise combinations among samples. For each sample, we then averaged F_{ST} across all SNPs for all pairwise combinations that include this particular sample and finally ranked the 48 population samples by overall differentiation.

We inferred demographic patterns in European populations by focusing on 20,008 putatively neutrally evolving SNPs located in small introns (less than 60 bp length; Haddrill *et al.* 2005; Singh *et al.* 2009; Parsch *et al.* 2010; Clemente & Vogl 2012; Lawrie *et al.* 2013) that were at least 200 kb distant from the major chromosomal inversions (see below). To assess isolation by distance (IBD), we averaged F_{ST} values for each sample pair across all neutral markers and calculated geographic distances between samples using the haversine formula (Green & Smart 1985) which takes the spherical curvature of the planet into account. We tested for correlations between genetic differentiation and geographic distance using Mantel tests using the *R* package *ade4* (v.1.7-8., Dray & Dufour 2007) with 1,000,000 iterations. In addition, we plotted the 5% smallest and 5% largest F_{ST} values from all 1,128 pairwise comparisons among the 48 population samples onto a map

to visualize geographic patterns of genetic differentiation. From these putatively neutral SNPs, we used observed F_{ST} on the autosomes (F_{aut}) to calculate the expected F_{ST} on X chromosomes (F_X) as in Machado *et al*. (2016) using the equation

$$F_X = 1 - \left[\frac{9(z+1) * (1 - F_{aut})}{8(2z+1) - (1 - F_{aut}) * (7z - 1)}\right]$$

where z is the ratio of effective population sizes of males (N_m) and females (N_f), N_m/N_f (Ramachandran *et al.* 2004). For the purposes of this study we assume z = 1.

We further investigated genetic variation in our dataset by principal component analysis (PCA) based on allele frequencies of the neutral marker SNPs described above. We used the *R* package *LEA* (v. 1.2.0., Frichot *et al.* 2013) and performed PCA on unscaled allele frequencies as suggested by Menozzi *et al*. (1978) and Novembre and Stephens (2008). We focused on the first three principal components (PCs) and employed a model-based approach as implemented in the *R* package *mclust* (v. 5.2., Fraley & Raftery 2012) to identify the most likely number of clusters based on maximum likelihood and assigned population samples to clusters by *k*-means clustering in *R* (R Development Core Team 2009). Finally, we examined the first three PCs for correlations with latitude, longitude, altitude, and season using general linear models and tested for spatial autocorrelation as described above. A Bonferroni-corrected α threshold (α'= 0.05/3 = 0.017) was used to account for multiple testing.

Mitochondrial DNA

To obtain consensus mitochondrial sequences for each of the 48 European populations, reads from individual FASTQ files were aligned and minor variants replaced by the major

variant using *Coral* (Salmela & Schröder 2011). This way, ambiguities that might prevent the growth of contigs from reads during the assembly process can be eliminated. For each population, a genome assembly was obtained using *SPAdes* using standard parameters and *k*-mers of size 21, 33, 55, and 77 (Bankevich *et al.* 2012) and the corrected FASTQ files. Mitochondrial contigs were retrieved by blastn, using the *D. melanogaster* NC 024511 sequence as a query and each genome assembly as the database. To avoid nuclear mitochondrial DNA segments (numts), we ensured that only contigs with a much higher coverage than the average coverage of the genome were retrieved. When multiple contigs were available for the same region, the one with the highest coverage was selected. Possible contamination with *D. simulans* was assessed by looking for two or more consecutive sites that show the same variant as *D. simulans* and looking for alternative contigs for that region with similar coverage. As an additional quality control measure, we also examined the presence of pairs of sites showing four gametic types using *DNAsp* 6 (Rozas *et al.* 2017) – given that there is no recombination in mitochondrial DNA no such sites are expected. The very few sites presenting such features were rechecked by looking for alternative contigs for that region and were corrected if needed. The uncorrected raw reads for each population were mapped on top of the different consensus haplotypes using *Express* as implemented in *Trinity* (Grabherr *et al.* 2011). If most reads for a given population mapped to the consensus sequence derived for that population the consensus sequence was retained, otherwise it was discarded as a possible chimera between different mitochondrial haplotypes. The repetitive mitochondrial hypervariable region is difficult to assemble and was therefore not used; the mitochondrial region was thus analyzed as in Cooper *et al*. (2015). Mitochondrial genealogy was estimated using statistical parsimony (TCS network; Clement *et al.* 2000), as implemented in *PopArt* (http://popart.otago.ac.nz), and the surviving mitochondrial haplotypes.

Frequencies of the different mitochondrial haplotypes were estimated from FPKM values using the surviving mitochondrial haplotypes and expressed as implemented in *Trinity* (Grabherr *et al.* 2011).

Transposable elements

To quantify the transposable element (TE) abundance in each sample, we assembled and quantified the repeats from unassembled sequenced reads using *dnaPipeTE* (v.1.2. , Goubert *et al.* 2015). The vast majority of high-quality trimmed reads were longer than 135 bp. We thus discarded reads less than 135 bp before sampling. Reads matching mtDNA were filtered out by mapping to the *D. melanogaster* reference mitochondrial genome (NC_024511.2. 1) with *bowtie2* (v. 2.1.0., Langmead & Salzberg 2012). Prokaryotic sequences, including reads from symbiotic bacteria such as *Wolbachia*, were filtered out from the reads using the implementation of *blastx* (translated nucleic vs. protein database) *vs.* the non-redundant protein database (nr) using *DIAMOND* (v. 0.8.7., Buchfink *et al.* 2015). To quantify TE content, we subsampled a proportion of the raw reads (after filtering) corresponding to a genome coverage of 0.1X (assuming a genome size of 175 MB), and then assembled these reads with *Trinity* assembler (Grabherr *et al.* 2011). Due to the low coverage of the genome obtained with the subsampled reads, only repetitive DNA present in multiple copies should be fully assembled (Goubert *et al.* 2015). We repeated this process with three iterations per sample, as recommended by the program guidelines, to assess the repeatability of the estimates.

We further estimated frequencies of previously characterized TEs present in the reference genome with *T-lex2* (v. 2.2.2., Fiston-Lavier *et al.* 2015), using all annotated TEs (5,416 TEs) in version 6.04 of the *D. melanogaster* genome from flybase.org (Gramates *et al.*

2017). For 108 of these TEs, we used the corrected coordinates as described in Fiston-Lavier *et al*. (2015), based on the identification of target site duplications at the site of the insertion. We excluded TEs nested or flanked by other TEs (<100 bp on each side of the TE), and TEs which are part of segmental duplications, since *T-lex2* does not provide accurate frequency estimates in complex regions (Fiston-Lavier *et al.* 2015). We additionally excluded the INE-1 TE family, as this TE family is ancient, with thousands of insertions in the reference genome, which appear to be mostly fixed (2,234 TEs; Kapitonov & Jurka 2003).

After applying these filters, we were able to estimate frequencies of 1,630 TE insertions from 113 families from the three main orders, LTR, non-LTR, and DNA across all *DrosEU* samples. *T-lex2* contains three main modules: (i) the presence detection module, (ii) the absence detection module, and (iii) the combine module, which joins the results from the former two detection modules. In the presence module, *T-lex2* uses *Maq* (v. 0.7.1., Li *et al.* 2008) for the mapping of reads. As *Maq* only accepts reads 127 bp or shorter, we cut the trimmed reads following the general pipeline (Figure 3) and then used *Trimmomatic* (v. 0.35; Bolger *et al.* 2014) to cut trimmed reads longer than 100 bp into two equally sized fragments using CROP and HEADCROP parameters. Only the presence module was run with the cut reads.

To avoid inaccurate TE frequency estimates due to very low numbers of reads, we only considered frequency estimates based on at least 3 reads. Despite the stringency of *T-lex2* to select only high-quality reads, we additionally discarded frequency estimates supported by more than 90 reads, i.e. 3 times the average coverage of the sample with the lowest coverage (CH_Cha_14_43, Table 1), in order to avoid non-unique mapping reads.

This filtering allows to estimate TE frequencies for ~96% (92.9% to 97.8%) of the TEs in each population. For 85% of the TEs, we were able to estimate their frequencies in more than 44 out of 48 *DrosEU* samples.

We tested for correlations between TE insertion frequencies and recombination rates using Spearman's rank correlations as implemented in *R*. For SNPs, we used recombination rates from Comeron *et al*. (2012) and from the Recombination Rate Calculator (Fiston-Lavier *et al.* 2010) in non-overlapping 100 kb windows, and assigned to each TE insertion the recombination rate of the corresponding 100 kb genomic window. To test for spatio-temporal variation of TE insertions, we excluded TEs with an interquartile range (IQR) < 10. There were 141 TE insertions with variable population frequencies among the *DrosEU* samples. We tested the population frequencies of these insertions for correlations with latitude, longitude, altitude, and season using generalized linear models (ANCOVA) following the method used for SNPs but with a binomial error structure in *R*. We also tested for residual spatio-temporal autocorrelations, with Moran's *I* test (Moran 1950; Kühn & Dormann 2012). We used Bonferroni corrections to account for multiple testing (α'= 0.05/141 = 0.00035) and only considered Bonferroni-corrected p-values < 0.001 to be significant. TEs with a recombination rate that differed from 0 cM/Mb according to both used measures (see above) were considered as high recombination regions. To test TE family enrichment among the significant TEs we performed a χ^2 test and applied Yate's correction to account for the low number of some of the cells.

Inversion polymorphisms

Since Pool-Seq data precludes a direct assessment of the presence and frequencies of chromosomal inversions, we indirectly estimated inversion frequencies using a panel of

approximately 400 inversion-specific marker SNPs (Kapun *et al.* 2014) for six cosmopolitan inversions (*In(2L)t, In(2R)NS, In(3L)P, In(3R)C, In(3R)Mo, In(3R)Payne*). We averaged allele frequencies of these markers in each sample separately. To test for clinal variation in the frequencies of inversions, we tested for correlations with latitude, longitude, altitude and season using generalized linear models with a binomial error structure in *R* to account for the biallelic nature of karyotype frequencies. In addition, we tested for residual spatio-temporal autocorrelations as described above and Bonferroni-corrected the α threshold (α'= 0.05/7 = 0.007) to account for multiple testing.

Microbiome

Raw sequences were trimmed and quality filtered as described for the genomic data analysis. The remaining high quality sequences were mapped against the *D. melanogaster* genome (v.6.04) including mitochondria using bbmap (v. 35; Bushnell 2016) with standard settings. The unmapped sequences were submitted to the online classification tool, MGRAST (Meyer *et al.* 2008) for annotation. Taxonomy information was downloaded and analyzed in *R* (v. 3.2.3; R Development Core Team 2009) using the matR (v. 0.9; Braithwaite & Keegan) and RJSONIO (v. 1.3; Lang) packages. Metazoan sequence features were removed. For microbial load comparisons, the number of protein features identified by MGRAST for each taxon and sample was divided by the number of sequences that mapped to *D. melanogaster* chromosomes *X*, *Y, 2L*, *2R*, *3L*, *3R* and *4*.

We also surveyed the datasets for the presence of novel DNA viruses by performing *de novo* assembly of the non-fly reads using SPAdes 3.9.0 (Bankevich *et al.* 2012), and using conceptual translations to query virus proteins from Genbank using DIAMOND 'blastp' (Buchfink *et al.* 2015). In three cases (*Kallithea* virus, *Vesanto* virus, *Viltain* virus), reads

from a single sample pool were sufficient to assemble a (near) complete genome. In two other cases, fragmentary assemblies allowed us to identify additional publicly available datasets that contained sufficient reads to complete the genomes (*Linvill Road* virus, *Esparto* virus; completed using SRA datasets SRR2396966 and SRR3939042, respectively). Novel viruses were provisionally named based on the localities where they were first detected, and the corresponding novel genome sequences were submitted to Genbank (KX130344, KY608910, KY457233, KX648533-KX648536). To assess the relative amount of viral DNA, unmapped (non-fly) reads from each sample pool were mapped to repeat-masked *Drosophila* DNA virus genomes using bowtie2, and coverage normalized relative to virus genome length and the number of mapped *Drosophila* reads.

Acknowledgments

We are grateful to all members of the *DrosEU* and Dros-RTEC consortia and to Dmitri Petrov (Stanford University) for support and discussion. *DrosEU* is funded by a Special Topic Networks (STN) grant from the European Society for Evolutionary Biology (ESEB). Computational analyses were partially executed at the Vital-IT bioinformatics facility of the University of Lausanne (Switzerland) and at the computing facilities of the CC LBBE/PRABI in Lyon (France).

Additional information

Funding

Funder	Grant reference number	Author
University of Freiburg Research Innovation Fund 2014		Fabian Staubach

Academy of Finland	#268241	Maaria Kankare
Academy of Finland	#272927	Maaria Kankare
Russian Foundation of Basic Research	#15-54-46009 CT_a	Elena G. Pasyukova
Danish Natural Science Research Council	4002-00113	Volker Loeschcke
Ministerio de Economia y Competitividad	CTM2017-88080 (AEI/FEDER, UE)	Marta Pascual
CNRS	UMR 9191	Catherine Montchamp-Moreau
Vetenskapsrådet	2011-05679	Jessica Abbott
Vetenskapsrådet	2015-04680	Jessica Abbott
Emmy Noether Programme of the Deutsche Forschungsgemeinschaft, DFG	PO 1648/3-1	Nico Posnien
National Institute of Health (NIH)	R35GM119686	Alan O. Bergland
Ministerio de Economia y Competitividad	CGL2013-42432-P	Maria Pilar Garcia Guerreiro
Scientific and Technological Research Council of Turkey (TUBITAK)	#214Z238	Banu Sebnem Onder
ANR Exhyb	14-CE19-0016	Cristina Vieira
Network of Excellence LifeSpan	FP6 036894	Bas J. Zwaan
IDEAL	FP7/2007-2011/259679	Bas J. Zwaan
Israel Science Foundation	1737/17	Eran Tauber
National Institute of Health (NIH)	R01GM100366	Paul S. Schmidt

Deutsche Forschungsgemeinschaft	PA 903/8-1	John Parsch
Austrian Science Fund (FWF)	P27048	Andrea J. Betancourt
Biotechnology and Biological Sciences Research Council (BBSRC)	BB/P00685X/1	Andrea J. Betancourt
Swiss National Science Foundation (SNSF)	PP00P3_133641	Thomas Flatt
Swiss National Science Foundation (SNSF)	PP00P3_165836	Thomas Flatt
European Comission H2020-ERC-2014	CoG-647900	Josefa Gonzàlez
Ministerio de Economia y Competitividad	FEDER BFU2014-57779-P	Josefa Gonzàlez

Author contributions

Martin Kapun, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Supervision, Methodology, Investigation, Data curation, Project administration, Validation, Resources, Software; Maite G. Barrón, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Data curation, Project administration, Validation, Resources, Software; Fabian Staubach, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Supervision, Funding acquisition, Methodology, Investigation, Data curation, Validation, Resources, Software; Jorge Vieira, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Validation, Resources; Darren J. Obbard, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Validation,

Resources; Clément Goubert, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Investigation, Resources; Omar Rota-Stabelli, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Resources; Maaria Kankare, Writing-original draft preparation, Conceptualization, Writing-review & editing, Methodology, Investigation, Resources; Annabelle Haudry, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Investigation, Validation, Resources; R. Axel W. Wiberg, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Resources, Software; Lena Waidele, Svitlana Serga, Patricia Gibert, Damiano Porcelli, Sonja Grath, Eliza Argyridou, Lain Guio, Mads Fristrup Schou, Conceptualization, Writing-review & editing, Investigation, Resources; Iryna Kozeretska, Conceptualization, Writing-review & editing, Methodology, Investigation, Resources; Elena G. Pasyukova, Marta Pascual, Alan O. Bergland, Conceptualization, Writing-review & editing, Funding acquisition, Methodology, Investigation, Resources; Volker Loeschcke, Catherine Montchamp-Moreau, Jessica Abbott, Nico Posnien, Maria Pilar Garcia Guerreiro, Banu Sebnem Onder, Conceptualization, Writing-review & editing, Funding acquisition, Investigation, Resources; Cristina P. Vieira, Visualization, Formal analysis, Conceptualization, Writing-review & editing, Investigation, Resources; Élio Sucena, Conceptualization, Writing-review & editing, Methodology, Investigation, Project administration, Resources; Cristina Vieira, Michael G. Ritchie, Thomas Flatt, Josefa González, Writing-original draft preparation, Conceptualization, Writing-review & editing, Supervision, Funding acquisition, Methodology, Investigation, Project administration, Validation, Resources; Bart Deplancke, Conceptualization, Writing-review & editing, Funding acquisition, Investigation; Bas J. Zwaan, Visualization, Writing-original draft

preparation, Conceptualization, Writing-review & editing, Supervision, Funding acquisition, Methodology, Investigation, Project administration; Eran Tauber, Writing-original draft preparation, Conceptualization, Writing-review & editing, Funding acquisition, Methodology, Investigation, Resources; Dorcas J. Orengo, Eva Puerma, Conceptualization, Writing-review & editing, Investigation, Validation, Resources; Montserrat Aguadé, Writing-original draft preparation, Conceptualization, Writing-review & editing, Methodology, Investigation, Validation, Resources; Paul S. Schmidt, John Parsch, Writing-original draft preparation, Conceptualization, Writing-review & editing, Funding acquisition, Methodology, Investigation, Validation, Resources; Andrea J. Betancourt, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Supervision, Funding acquisition, Methodology, Investigation, Project administration, Validation, Resources

Author ORCIDs

Names	ORCID
Martin Kapun	0000-0002-3810-0504
Maite G. Barrón	0000-0001-6146-6259
Fabian Staubach	0000-0002-8097-2349
Jorge Vieira	0000-0001-7032-5220
Darren J. Obbard	0000-0001-5392-8142
Clément Goubert	0000-0001-8034-5559
Omar Rota-Stabelli	0000-0002-0030-7788
Maaria Kankare	0000-0003-1541-9050
Annabelle Haudry	0000-0001-6088-0909
R. Axel W. Wiberg	0000-0002-8074-8670

Lena Waidele	0000-0002-6323-6438
Iryna Kozeretska	0000-0002-6485-1408
Elena G. Pasyukova	0000-0002-6491-8561
Volker Loeschcke	0000-0003-1450-0754
Marta Pascual	0000-0002-6189-0612
Cristina P. Vieira	0000-0002-7139-2107
Svitlana Serga	0000-0003-1875-3185
Catherine Montchamp-Moreau	0000-0002-5044-9709
Jessica Abbott	0000-0002-8743-2089
Patricia Gibert	0000-0002-9461-6820
Damiano Porcelli	0000-0002-9019-5758
Nico Posnien	0000-0003-0700-5595
Sonja Grath	0000-0003-3621-736X
Élio Sucena	0000-0001-8810-870X
Alan O. Bergland	0000-0001-7145-7575
Maria Pilar Garcia Guerreiro	000-0001-9951-1879X
Banu Sebnem Onder	0000-0002-3003-248X
Eliza Argyridou	0000-0002-6890-4642
Lain Guio	0000-0002-5481-5200
Mads Fristrup Schou	0000-0001-5521-5269
Bart Deplancke	0000-0001-9935-843X
Cristina Vieira	0000-0003-3414-3993
Michael G. Ritchie	0000-0001-7913-8675
Bas J. Zwaan	0000-0002-8221-4998
Eran Tauber	0000-0003-4018-6535

Dorcas J. Orengo	0000-0001-7911-3224
Eva Puerma	0000-0001-7261-187X
Montserrat Aguadé	0000-0002-3884-7800
Paul S. Schmidt	0000-0002-8076-6705
John Parsch	0000-0001-9068-5549
Andrea J. Betancourt	0000-0001-9351-1413
Thomas Flatt	0000-0002-5990-1503
Josefa González	0000-0001-9824-027X

References

Adrian AB, Comeron JM (2013) The *Drosophila* early ovarian transcriptome provides insight to the molecular causes of recombination rate variation across genomes. *BMC Genomics*, **14**, 794.

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

Alonso-Blanco C, Andrade J, Becker C *et al.* (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. **166**, 481–491.

Anderson AR, Hoffmann AA, McKechnie SW, Umina PA, Weeks AR (2005) The latitudinal cline in the *In(3R)Payne* inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. *Molecular Ecology*, **14**, 851–858.

Anderson PR, Knibb WR, Oakeshott JG (1987) Observations on the extent and temporal stability of latitudinal clines for alcohol dehydrogenase allozymes and four chromosome inversions in *Drosophila melanogaster*. *Genetica*, **75**, 81–88.

1426 Anderson WW, Arnold J, Baldwin DG *et al.* (1991) Four decades of inversion
1427 polymorphism in *Drosophila pseudoobscura*. *Proceedings of the National Academy of*
1428 *Sciences of the United States of America*, **88**, 10367–10371.
1429 Andolfatto P, Wall JD, Kreitman M (1999) Unusual haplotype structure at the proximal
1430 breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster. Genetics*,
1431 **153**, 1297–1311.
1432 Ashburner M, Lemeunier F (1976) Relationships within the melanogaster Species
1433 Subgroup of the Genus Drosophila (Sophophora). I. Inversion Polymorphisms in
1434 Drosophila melanogaster and Drosophila simulans. *Proceedings of the Royal Society*
1435 *of London. Series B: Biological Sciences*, **193**, 137–157.
1436 Aulard S, David JR, Lemeunier F (2002) Chromosomal inversion polymorphism in
1437 Afrotropical populations of *Drosophila melanogaster*. *Genetic Research*, **79**, 49–63.
1438 Bankevich A, Nurk S, Antipov D *et al.* (2012) *SPAdes, a New Genome Assembly*
1439 *Algorithm and Its Applications to Single-cell Sequencing (7th Annual SFAF Meeting,*
1440 *2012).* Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY
1441 10801 USA.
1442 Barata A, Santos SC, Malfeito-Ferreira M, Loureiro V (2012) New insights into the
1443 ecological interaction between grape berry microorganisms and *Drosophila* flies during
1444 the development of sour rot. *Microbial ecology*, **64**, 416–430.
1445 Bartolomé C, Maside X, Charlesworth B (2002) On the Abundance and Distribution of
1446 Transposable Elements in the Genome of *Drosophila melanogaster*. *Molecular Biology*
1447 *and Evolution*, **19**, 926–937.
1448 Bastide H, Betancourt A, Nolte V *et al.* (2013) A genome-wide, fine-scale map of natural
1449 pigmentation variation in *Drosophila melanogaster*. (P Wittkopp, Ed,). *PLoS Genetics*,
1450 **9**, e1003534.

1451 Baudry E, Viginier B, Veuille M (2004) Non-African populations of *Drosophila*
1452 *melanogaster* have a unique origin. *Molecular Biology and Evolution*, **21**, 1482–1491.
1453 Becher PG, Flick G, Rozpędowska E *et al.* (2012) Yeast, not fruit volatiles mediate
1454 *Drosophila melanogaster* attraction, oviposition and development. *Functional Ecology*,
1455 **26**, 822–828.
1456 Begun DJ (2015) Parallel Gene Expression Differences between Low and High Latitude
1457 Populations of *Drosophila melanogaster* and *D. simulans.* (SV Nuzhdin, Ed,). *PLoS*
1458 *Genetics*, **11**, e1005184.
1459 Begun DJ, Aquadro CF (1993) African and North American populations of *Drosophila*
1460 *melanogaster* are very different at the DNA level. *Nature*, **365**, 548–550.
1461 Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population Genomics: Whole-Genome
1462 Analysis of Polymorphism and Divergence in *Drosophila simulans*. *PLoS Biology*, **5**,
1463 e310.
1464 Behrman EL, Howick VM, Kapun M *et al.* (2018) Rapid seasonal evolution in innate
1465 immunity of wild *Drosophila melanogaster*. *Proceedings of the Royal Society B:*
1466 *Biological Sciences*, **285**, 20172599.
1467 Beisswanger S, Stephan W, De Lorenzo D (2006) Evidence for a Selective Sweep in the
1468 *wapl* Region of *Drosophila melanogaster*. *Genetics*, **172**, 265–274.
1469 Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA (2014) Genomic Evidence
1470 of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in *Drosophila*.
1471 *PLoS Genetics*, **10**, e1004775.
1472 Bergland AO, Tobler R, González J, Schmidt P, Petrov D (2016) Secondary contact and
1473 local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila*
1474 melanogaster. *Molecular Ecology*, **25**, 1157–1174.

1475 Bivand R, Piras G (2015) Comparing Implementations of Estimation Methods for Spatial
1476 Econometrics. *Journal of Statistical Software*, **63**, 1–36.
1477 Black WC IV, Black WC IV, Baer CF, Antolin MF, DuTeau NM (2001) Population
1478 genomics: genome-wide sampling of insect populations. *Annual Review of*
1479 *Entomology*.
1480 Blumenstiel JP, Chen X, He M, Bergman CM (2014) An Age-of-Allele Test of Neutrality for
1481 Transposable Element Insertions. *Genetics*, **196**, 523–538.
1482 Boitard S, Schlötterer C, Nolte V, Pandey RV, Futschik A (2012) Detecting Selective
1483 Sweeps from Pooled Next-Generation Sequencing Samples. *Molecular Biology and*
1484 *Evolution*, **29**, 2177–2186.
1485 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina
1486 sequence data. *Bioinformatics*, **30**, 2114–2120.
1487 Boussy IA, Itoh M, Rand D, Woodruff RC (1998) Origin and decay of the P element-
1488 associated latitudinal cline in Australian *Drosophila melanogaster*. *Genetica*, **104**, 45–
1489 57.
1490 Božičević V, Hutter S, Stephan W, Wollstein A (2016) Population genetic evidence for cold
1491 adaptation in European *Drosophila melanogaster* populations. *Molecular Ecology*, **25**,
1492 1175–1191.
1493 Braithwaite DP, Keegan KP *matR: Metagenomics Analysis Tools for R*. https://CRAN.R-
1494 project.org/package=matR.
1495 Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using
1496 DIAMOND. *Nature Methods*, **12**, 59–60.
1497 Buser CC, Newcomb RD, Gaskett AC, Goddard MR (2014) Niche construction initiates the
1498 evolution of mutualistic interactions. *Ecology Letters*, **17**, 1257–1264.
1499 Bushnell B (2016) *BBMap short read aligner*. URL http://sourceforge. net/projects/bbmap.

1500 Camus MF, Wolff JN, Sgrò CM, Dowling DK (2017) Experimental Support That Natural
1501 Selection Has Shaped the Latitudinal Distribution of Mitochondrial Haplotypes in
1502 Australian *Drosophila melanogaster. Molecular Biology and Evolution*, **34**, 2600–2612.
1503 Caracristi G, Schlötterer C (2003) Genetic Differentiation Between American and
1504 European *Drosophila melanogaster* Populations Could Be Attributed to Admixture of
1505 African Alleles. *Molecular Biology and Evolution*, **20**, 792–799.
1506 Casillas S, Barbadilla A (2017) Molecular Population Genetics. *Genetics*, **205**, 1003–1035.
1507 Catania F, Kauer MO, Daborn PJ *et al.* (2004) World-wide survey of an Accord insertion
1508 and its association with DDT resistance in *Drosophila melanogaster. Molecular*
1509 *Ecology*, **13**, 2491–2504.
1510 Cavalli-Sforza LL (1966) Population Structure and Human Evolution. *Proceedings of the*
1511 *Royal Society of London. Series B: Biological Sciences*, **164**, 362–379.
1512 Chandler JA, James PM (2013) Discovery of trypanosomatid parasites in globally
1513 distributed *Drosophila* species. *PLoS ONE*, **8**, e61937.
1514 Chandler JA, Eisen JA, Kopp A (2012) Yeast communities of diverse *Drosophila* species:
1515 comparison of two symbiont groups in the same hosts. *Applied and Environmental*
1516 *Microbiology*, **78**, 7327–7336.
1517 Chandler JA, Lang JM, Bhatnagar S, Eisen JA, Kopp A (2011) Bacterial communities of
1518 diverse *Drosophila* species: ecological context of a host-microbe model system. *PLoS*
1519 *Genetics*, **7**, e1002272.
1520 Charlesworth B (2010) Molecular population genomics: a short history. *Genetical*
1521 *Research*, **92**, 397–411.
1522 Charlesworth B (2015) Causes of natural variation in fitness: evidence from studies of
1523 *Drosophila* populations. *Proceedings of the National Academy of Sciences of the*
1524 *United States of America*, **112**, 1662–1669.

1525 Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive
1526 DNA in eukaryotes. *Nature*, **371**, 215–220.
1527 Cheng C, White BJ, Kamdem C *et al.* (2012) Ecological genomics of *Anopheles gambiae*
1528 along a latitudinal cline: a population-resequencing approach. *Genetics*, **190**, 1417–
1529 1432.
1530 Cingolani P, Platts A, Wang LL *et al.* (2012) A program for annotating and predicting the
1531 effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila*
1532 *melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly (Austin)*, **6**, 80–92.
1533 Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene
1534 genealogies. *Molecular Ecology*, **9**, 1657–1659.
1535 Clemente F, Vogl C (2012) Unconstrained evolution in short introns? – An analysis of
1536 genome-wide polymorphism and divergence data from *Drosophila*. *Journal of*
1537 *Evolutionary Biology*, **25**, 1975–1990.
1538 Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in
1539 *Drosophila melanogaster*. (DA Petrov, Ed,). *PLoS Genetics*, **8**, e1002905.
1540 Consortium T (2015) A global reference for human genetic variation. **526**, 68–74.
1541 Cooper BS, Burrus CR, Ji C, Hahn MW, Montooth KL (2015) Similar Efficacies of
1542 Selection Shape Mitochondrial and Nuclear Genes in Both *Drosophila melanogaster*
1543 and *Homo sapiens*. *G3 (Bethesda, Md.)*, **5**, 2165–2176.
1544 Corbett-Detig RB, Hartl DL (2012) Population Genomics of Inversion Polymorphisms in
1545 *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1003056.
1546 Corbett-Detig RB, Cardeno C, Langley CH (2012) Sequence-based detection and
1547 breakpoint assembly of polymorphic inversions. *Genetics*, **192**, 131–137.

1548 Cridland JM, Macdonald SJ, Long AD, Thornton KR (2013) Abundance and distribution of
1549 transposable elements in two *Drosophila* QTL mapping resources. *Molecular Biology*
1550 *and Evolution*, **30**, 2311–2327.
1551 Daborn PJ, Yen JL, Bogwitz MR *et al.* (2002) A single p450 allele associated with
1552 insecticide resistance in *Drosophila*. *Science*, **297**, 2253–2256.
1553 Das A, Singh BN (1990) Chromosome inversions in Indian *Drosophila melanogaster*.
1554 *Genetica*, **81**, 85–88.
1555 Das A, Singh BN (1991) Genetic differentiation and inversion clines in Indian natural
1556 populations of Drosophila melanogaster. *Genome / National Research Council Canada*
1557 = *Génome / Conseil national de recherches Canada*, **34**, 618–625.
1558 David JR, Capy P (1988) Genetic variation of *Drosophila melanogaster* natural
1559 populations. *Trends in genetics : TIG*, **4**, 106–111.
1560 de Jong G, Bochdanovits Z (2003) Latitudinal clines in *Drosophila melanogaster*: body
1561 size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway.
1562 *Journal of Genetics*, **82**, 207–223.
1563 Dieringer D, Nolte V, Schlötterer C (2005) Population structure in African *Drosophila*
1564 *melanogaster* revealed by microsatellite analysis. *Molecular Ecology*, **14**, 563–573.
1565 Dobzhansky T (1970) *Genetics of the Evolutionary Process*. Columbia University Press.
1566 Dobzhansky T, Sturtevant AH (1938) Inversions in the Chromosomes of Drosophila
1567 Pseudoobscura. *Genetics*, **23**, 28–64.
1568 Dray S, Dufour A-B (2007) The ade4 Package: Implementing the Duality Diagram for
1569 Ecologists. *Journal of Statistical Software*, **22**.
1570 Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S (2013) Demographic inference
1571 reveals African and European admixture in the North American *Drosophila*
1572 *melanogaster* population. *Genetics*, **193**, 291–301.

1573 Ellegren H (2014) Genome sequencing and population genomics in non-model organisms.
1574 *Trends in Ecology & Evolution*, **29**, 51–63.
1575 Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University
1576 Press.
1577 Fabian DK, Kapun M, Nolte V *et al.* (2012) Genome-wide patterns of latitudinal
1578 differentiation among populations of *Drosophila melanogaster* from North America.
1579 *Molecular Ecology*, **21**, 4748–4769.
1580 Fabian DK, Lack JB, Mathur V *et al.* (2015) Spatially varying selection shapes life history
1581 clines among populations of *Drosophila melanogaster* from sub-Saharan Africa.
1582 *Journal of Evolutionary Biology*, **28**, 826–840.
1583 Fiston-Lavier A-S, Barrón MG, Petrov DA, González J (2015) T-lex2: genotyping,
1584 frequency estimation and re-annotation of transposable elements using single or
1585 pooled next-generation sequencing data. *Nucleic Acids Research*, **43**, e22–e22.
1586 Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA (2010) *Drosophila melanogaster*
1587 recombination rate calculator. *Gene*, **463**, 18–20.
1588 Flatt T, Amdam GV, Kirkwood TBL, Omholt SW (2013) Life-history evolution and the
1589 polyphenic regulation of somatic maintenance and survival. *The quarterly review of*
1590 *biology*, **88**, 185–218.
1591 Fraley C, Raftery AE (2012) *mclust Version 4 for R: Normal Mixture Modeling for Model-*
1592 *Based Clustering, Classification, and Density Estimation*. Seattle.
1593 Francalacci P, Sanna D (2008) History and geography of human Y-chromosome in
1594 Europe: a SNP perspective. *Journal of anthropological sciences*, **86**, 59–89.
1595 Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between
1596 loci and environmental gradients using latent factor mixed models. *Molecular Biology*
1597 *and Evolution*, **30**, 1687–1699.

González J, Karasov TL, Messer PW, Petrov DA (2010) Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in *Drosophila* (HS Malik, Ed,). *PLoS Genetics*, **6**, e1000905.

González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA (2008) High Rate of Recent Transposable Element–Induced Adaptation in *Drosophila melanogaster*. *PLoS Biology*, **6**, e251.

Goubert C, Modolo L, Vieira C *et al.* (2015) De Novo Assembly and Annotation of the Asian Tiger Mosquito (*Aedes albopictus*) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (*Aedes aegypti*). *Genome Biology and Evolution*, **7**, 1192–1205.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Gramates LS, Marygold SJ, Santos GD *et al.* (2017) FlyBase at 25: looking to the future. *Nucleic Acids Research*, **45**, D663–D671.

Green RM, Smart WM (1985) *Textbook on Spherical Astronomy*. Cambridge University.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P (2005) Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biology*, **6**, R67.

Hales KG, Korey CA, Larracuente AM, Roberts DM (2015) Genetics on the Fly: A Primer on the Drosophila Model System. *Genetics*, **201**, 815–842.

Hamilton PT, Votýpka J, Dostálová A *et al.* (2015) Infection Dynamics and Immune Response in a Newly Described *Drosophila*-Trypanosomatid Association. *mBio*, **6**, e01356–15.

Handu M, Kaduskar B, Ravindranathan R *et al.* (2015) SUMO-Enriched Proteome for *Drosophila* Innate Immune Response. *G3 (Bethesda, Md.)*, **5**, 2137–2154.

1623 Harpur BA, Kent CF, Molodtsova D *et al.* (2014) Population genomics of the honey bee
1624 reveals strong signatures of positive selection on worker traits. *Proceedings of the*
1625 *National Academy of Sciences of the United States of America*, **111**, 2614–2619.
1626 Haselkorn TS, Markow TA, Moran NA (2009) Multiple introductions of the *Spiroplasma*
1627 bacterial endosymbiont into *Drosophila*. *Molecular Ecology*, **18**, 1294–1305.
1628 Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel
1629 Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genetics*, **6**,
1630 e1000862.
1631 Hu TT, Eisen MB, Thornton KR, Andolfatto P (2013) A second-generation assembly of the
1632 *Drosophila simulans* genome provides new insights into patterns of lineage-specific
1633 divergence. *Genome Research*, **23**, 89–98.
1634 Huang W, Massouras A, Inoue Y *et al.* (2014) Natural variation in genome architecture
1635 among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome*
1636 *Research*, **24**, 1193–1208.
1637 Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on
1638 nucleotide data. *Genetics*, **116**, 153–159.
1639 Huszar T, Imler J-L (2008) *Drosophila* viruses and the study of antiviral host-defense.
1640 *Advances in virus research*, **72**, 227–265.
1641 Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W (2007) Distinctly Different Sex
1642 Ratios in African and European Populations of *Drosophila melanogaster* Inferred From
1643 Chromosomewide Single Nucleotide Polymorphism Data. *Genetics*, **177**, 469–480.
1644 Inoue Y, Watanabe TK (1979) Inversion polymorphisms in Japanese natural populations of
1645 *Drosophila melanogaster*. *Japanese Journal of Genetics*, **54**, 69–82.
1646 Inoue Y, Watanabe T, Watanabe TK (1984) Evolutionary Change of the Chromosomal
1647 Polymorphism in Drosophila melanogaster Populations. *Evolution*, **38**, 753.

1648 Jorde LB, Watkins WS, Bamshad MJ (2001) Population genomics: a bridge from
1649 evolutionary history to genetic medicine. *Human Molecular Genetics*, **10**, 2199–2207.
1650 Kao JY, Zubair A, Salomon MP, Nuzhdin SV, Campo D (2015) Population genomic
1651 analysis uncovers African and European admixture in *Drosophila melanogaster*
1652 populations from the south-eastern United States and Caribbean Islands. *Molecular*
1653 *Ecology*, **24**, 1499–1509.
1654 Kapitonov VV, Jurka J (2003) Molecular Paleontology of Transposable Elements in the
1655 *Drosophila melanogaster* Genome. *Proceedings of the National Academy of Sciences*
1656 *of the United States of America*, **100**, 6569–6574.
1657 Kapun M, Fabian DK, Goudet J, Flatt T (2016a) Genomic Evidence for Adaptive Inversion
1658 Clines in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **33**, 1317–1336.
1659 Kapun M, Schmidt C, Durmaz E, Schmidt PS, Flatt T (2016b) Parallel effects of the
1660 inversion *In(3R)Payne* on body size across the North American and Australian clines in
1661 *Drosophila melanogaster. Journal of Evolutionary Biology*, **29**, 1059–1072.
1662 Kapun M, van Schalkwyk H, McAllister B, Flatt T, Schlötterer C (2014) Inference of
1663 chromosomal inversion dynamics from Pool-Seq data in natural and laboratory
1664 populations of *Drosophila melanogaster. Molecular Ecology*, **23**, 1813–1827.
1665 Kassis JA, Kennison JA, Tamkun JW (2017) Polycomb and Trithorax Group Genes in
1666 *Drosophila*. *Genetics*, **206**, 1699–1725.
1667 Keller A (2007) *Drosophila melanogaster*'s history as a human commensal. *Current*
1668 *Biology*, **17**, R77–R81.
1669 Kennington WJ, Hoffmann AA (2013) Patterns of genetic variation across inversions:
1670 geographic variation in the *In(2L)t* inversion in populations of *Drosophila melanogaster*
1671 from eastern Australia. *BMC evolutionary biology*, **13**, 100.

1672 Kennington WJ, Hoffmann AA, Partridge L (2007) Mapping Regions Within Cosmopolitan
1673 Inversion *In(3R)Payne* Associated With Natural Variation in Body Size in *Drosophila*
1674 *melanogaster*. *Genetics*, **177**, 549–556.
1675 Kennison J (2008) Dissection of Larval Salivary Glands and Polytene Chromosome
1676 Preparation. *Cold Spring Harbor Protocols*, **2008**, pdb.prot4708–pdb.prot4708.
1677 Kimura M (1984) *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
1678 Kirkpatrick M (2010) How and why chromosome inversions evolve. *PLoS Biology*, **8**,
1679 e1000501.
1680 Knibb WR (1982) Chromosome inversion polymorphisms in *Drosophila melanogaster* II.
1681 Geographic clines and climatic associations in Australasia, North America and Asia.
1682 *Genetica*, **58**, 213–221.
1683 Knibb WR (1986) Temporal variation of *Drosophila melanogaster Adh* allele frequencies,
1684 inversion freqencies, and population sizes. *Genetica*, **71**, 175–190.
1685 Knibb WR, Oakeshott JG, Gibson JB (1981) Chromosome Inversion Polymorphisms in
1686 *Drosophila melanogaster*. I. Latitudinal Clines and Associations between Inversions in
1687 Australasian Populations. *Genetics*, **98**, 833–847.
1688 Kofler R, Betancourt AJ, Schlötterer C (2012) Sequencing of pooled DNA samples (Pool-
1689 Seq) uncovers complex dynamics of transposable element insertions in *Drosophila*
1690 *melanogaster. PLoS Genetics*, **8**, e1002487.
1691 Kofler R, Orozco-terWengel P, De Maio N *et al.* (2011) PoPoolation: A Toolbox for
1692 Population Genetic Analysis of Next Generation Sequencing Data from Pooled
1693 Individuals. *PLoS ONE*, **6**, e15925.
1694 Kolaczkowski B, Hupalo DN, Kern AD (2011a) Recurrent adaptation in RNA interference
1695 genes across the *Drosophila* phylogeny. *Molecular Biology and Evolution*, **28**, 1033–
1696 1042.

1697 Kolaczkowski B, Kern AD, Holloway AK, Begun DJ (2011b) Genomic Differentiation
1698 Between Temperate and Tropical Australian Populations of *Drosophila melanogaster*.
1699 *Genetics*, **187**, 245–260.
1700 Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R (2013) Calculation of Tajima's *D* and
1701 other neutrality test statistics from low depth next-generation sequencing data. *BMC*
1702 *Bioinformatics*, **14**, 289.
1703 Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of
1704 *Drosophila melanogaster*. *Nature*, **304**, 412–417.
1705 Kriesner P, Conner WR, Weeks AR, Turelli M, Hoffmann AA (2016) Persistence of a
1706 *Wolbachia* infection frequency cline in *Drosophila melanogaster* and the possible role
1707 of reproductive dormancy. *Evolution*, **70**, 979–997.
1708 Kunze-Mühl E, Müller E (1957) Weitere Untersuchungen über die chromosomale Struktur
1709 und die naturlichen Strukturtypen von *Drosophila subobscur*a coll. *Chromosoma*, **9**,
1710 559–570.
1711 Kühn I, Dormann CF (2012) Less than eight (and a half) misconceptions of spatial
1712 analysis. *Journal of Biogeography*, **39**, 995–998.
1713 Lachaise D, Cariou M-L, David JR *et al.* (1988) Historical Biogeography of the *Drosophila*
1714 *melanogaster* Species Subgroup. In: *Evolutionary Biology*, pp. 159–225. Springer,
1715 Boston, MA, Boston, MA.
1716 Lack JB, Cardeno CM, Crepeau MW *et al.* (2015) The *Drosophila* genome nexus: a
1717 population genomic resource of 623 *Drosophila melanogaster* genomes, including 197
1718 from a single ancestral range population. *Genetics*, **199**, 1229–1241.
1719 Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE (2016) A Thousand Fly
1720 Genomes: An Expanded *Drosophila* Genome Nexus. *Molecular Biology and Evolution*,
1721 **33**, msw195–3313.

1722 Lang DT *RJSONIO: Serialize R objects to JSON, JavaScript Object Notation*.
1723 https://CRAN.R-project.org/package=RJSONIO.
1724 Langley CH, Stevens K, Cardeno C *et al.* (2012) Genomic variation in natural populations
1725 of *Drosophila melanogaster*. *Genetics*, **192**, 533–598.
1726 Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature*
1727 *Methods*, **9**, 357–359.
1728 Lavington E, Kern AD (2017) The Effect of Common Inversion
1729 PolymorphismsIn(2L)tandIn(3R)Moon Patterns of Transcriptional Variation
1730 inDrosophila melanogaster. *G3 (Bethesda, Md.)*, **7**, 3659–3668.
1731 Lawrie DS, Messer PW, Hershberg R, Petrov DA (2013) Strong Purifying Selection at
1732 Synonymous Sites in *D. melanogaster*. *arXiv.org*, **q-bio.PE**, e1003527.
1733 Lemeunier F, Aulard S (1992) Inversion polymorphism in *Drosophila melanogaster*. In:
1734 *Drosophila Inversion Polymorphism* (eds Krimbas CB, Powell JR), p. 576. CRC Press.
1735 Levins R (1968) *Evolution in Changing Environments*. Princeton University Press.
1736 Lewontin RC (1974) *The Genetic Basis of Evolutionary Change*. Columbia University
1737 Press.
1738 Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-
1739 MEM. *arXiv.org*.
1740 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler
1741 transform. *Bioinformatics*, **25**, 1754–1760.
1742 Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants
1743 using mapping quality scores. *Genome Research*, **18**, 1851–1858.
1744 Luikart G, England PR, Tallmon D, Jordan S, Tableret P (2003) The power and promise of
1745 population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**,
1746 981–994.

1747 Lyne R, Smith R, Rutherford K *et al.* (2007) FlyMine: an integrated database for
1748 *Drosophila* and *Anopheles* genomics. *Genome Biology*, **8**, R129.
1749 Machado HE, Bergland AO, O'Brien KR *et al.* (2016) Comparative population genomics of
1750 latitudinal variation in *Drosophila simulans* and *Drosophila melanogaster*. *Molecular*
1751 *Ecology*, **25**, 723–740.
1752 Machado H, Bergland AO, Taylor R *et al.* (2018) Broad geographic sampling reveals
1753 predictable and pervasive seasonal adaptation in *Drosophila*. *bioRxiv*.
1754 https://doi.org/10.1101/337543
1755 Mackay TFC, Richards S, Stone EA *et al.* (2012) The *Drosophila melanogaster* Genetic
1756 Reference Panel. *Nature*, **482**, 173–178.
1757 Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing
1758 reads. *EMBnet.journal*, **17**, 10–12.
1759 Martino ME, Ma D, Leulier F (2017) Microbial influence on *Drosophila* biology. *Current*
1760 *Opinion in Microbiology*, **38**, 165–170.
1761 Mateo L, Rech GE, González J (2018) Genome-wide patterns of local adaptation in
1762 *Drosophila melanogaster*: adding intra European variability to the map. *bioRxiv*.
1763 Matthias P, Yoshida M, Khochbin S (2008) HDAC6 a new cellular stress surveillance
1764 factor. *Cell Cycle*, **7**, 7–10.
1765 Matzkin LM, Merritt TJS, Zhu C-T, Eanes WF (2005) The structure and population
1766 genetics of the breakpoints associated with the cosmopolitan chromosomal inversion
1767 *In(3R)Payne* in *Drosophila melanogaster. Genetics*, **170**, 1143–1152.
1768 McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in
1769 *Drosophila*. *Nature*, **351**, 652–654.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. *Science*, **201**, 786–792.

Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. **28**, 659–669.

Mettler LE, Voelker RA, Mukai T (1977) Inversion Clines in Populations of *Drosophila melanogaster. Genetics*, **87**, 169–176.

Meyer F, Paarmann D, D'Souza M *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Micallef L, Rodgers P (2014) euler*APE*: drawing area-proportional 3-Venn diagrams using ellipses. *PLoS ONE*, **9**, e101717.

Michalakis Y, Veuille M (1996) Length variation of CAG/CAA trinucleotide repeats in natural populations of *Drosophila melanogaster* and its relation to the recombination rate. *Genetics*, **143**, 1713–1725.

Moran PAP (1950) Notes on Continuous Stochastic Phenomena. *Biometrika*, **37**, 17.

Navarro A, Faria R (2014) Pool and conquer: new tricks for (c)old problems. *Molecular Ecology*, **23**, 1653–1655.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press.

Nolte V, Pandey RV, Kofler R, Schlötterer C (2013) Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. *Genome Research*, **23**, 99–110.

1794 Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial
1795 population genetic variation. *Nature Genetics*, **40**, 646–649.
1796 Nunes MDS, Neumeier H, Schlötterer C (2008) Contrasting patterns of natural variation in
1797 global *Drosophila melanogaster* populations. *Molecular Ecology*, **17**, 4470–4479.
1798 Okonechnikov K, Conesa A, García-Alcalde F (2016) Qualimap 2: advanced multi-sample
1799 quality control for high-throughput sequencing data. *Bioinformatics*, **32**, 292–294.
1800 Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P (2010) On the utility
1801 of short intron sequences as a reference for the detection of positive and negative
1802 selection in *Drosophila*. *Molecular Biology and Evolution*, **27**, 1226–1234.
1803 Peel MC, Finlayson BL, McMahon TA (2007) Updated world map of the Köppen-Geiger
1804 climate classification. *Hydrology and Earth System Sciences*, **11**, 1633–1644.
1805 Petrov DA, Fiston-Lavier AS, Lipatov M, Lenkov K, González J (2011) Population
1806 Genomics of Transposable Elements in *Drosophila melanogaster*. *Molecular Biology*
1807 *and Evolution*, **28**, 1633–1644.
1808 Pimpinelli S, Bonaccorsi S, Fanti L, Gatti M (2010) Preparation and Orcein Staining of
1809 Mitotic Chromosomes from Drosophila Larval Brain. *Cold Spring Harbor Protocols*,
1810 **2010**, pdb.prot5389–pdb.prot5389.
1811 Pool JE (2015) The Mosaic Ancestry of the *Drosophila* Genetic Reference Panel and the
1812 *D. melanogaster* Reference Genome Reveals a Network of Epistatic Fitness
1813 Interactions. *Molecular Biology and Evolution*, **32**, 3236–3251.
1814 Pool JE, Braun DT, Lack JB (2016) Parallel Evolution of Cold Tolerance Within *Drosophila*
1815 *melanogaster*. *Molecular Biology and Evolution*, msw232.
1816 Pool JE, Corbett-Detig RB, Sugino RP *et al.* (2012) Population Genomics of Sub-Saharan
1817 *Drosophila melanogaster*: African Diversity and Non-African Admixture. *PLoS*
1818 *Genetics*, **8**, e1003080.

1819 Powell JR (1997) *Progress and Prospects in Evolutionary Biology: The Drosophila Model*.
1820 Oxford University Press.
1821 R Development Core Team (2009) R: A Language and Environment for Statistical
1822 Computing. *R-project.org*.
1823 Rako L, Anderson AR, Sgrò CM, Stocker AJ, Hoffmann AA (2006) The association
1824 between inversion *In(3R)Payne* and clinally varying traits in *Drosophila melanogaster.*
1825 *Genetica*, **128**, 373–384.
1826 Ramachandran S, Rosenberg NA, Zhivotovsky LA, Feldman MW (2004) Robustness of
1827 the inference of human population structure: a comparison of X-chromosomal and
1828 autosomal microsatellites. *Human genomics*, **1**, 87–97.
1829 Rane RV, Rako L, Kapun M, LEE SF (2015) Genomic evidence for role of inversion *3RP*
1830 of *Drosophila melanogaster* in facilitating climate change adaptation. *Molecular*
1831 *Ecology*, **24**, 2423–2432.
1832 Richardson JL, Urban MC, Bolnick DI, Skelly DK (2014) Microgeographic adaptation and
1833 the spatial scale of evolution. *Trends in Ecology & Evolution*, **29**, 165–176.
1834 Richardson MF, Weinert LA, Welch JJ *et al.* (2012) Population Genomics of the *Wolbachia*
1835 Endosymbiont in *Drosophila melanogaster* (A Kopp, Ed,). *PLoS Genetics*, **8**,
1836 e1003129.
1837 Roberts DB (Ed.) (1998) *Drosophila: A Practical Approach The Practical Approach*.
1838 Practical Approach Series.
1839 Rogers RL, Hartl DL (2012) Chimeric genes as a source of rapid evolution in *Drosophila*
1840 *melanogaster*. *Molecular Biology and Evolution*, **29**, 517–529.
1841 Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC *et al.* (2017) DnaSP 6: DNA Sequence
1842 Polymorphism Analysis of Large Datasets. *Molecular Biology and Evolution*, **34**, 3299–
1843 3302.

1844 Salmela L, Schröder J (2011) Correcting errors in short reads by multiple alignments.
1845 *Bioinformatics*, **27**, 1455–1461.
1846 Schlenke TA, Begun DJ (2003) Natural selection drives *Drosophila* immune system
1847 evolution., **164**, 1471–1480.
1848 Schlötterer C, Neumeier H, Sousa C, Nolte V (2006) Highly structured Asian *Drosophila*
1849 *melanogaster* populations: a new tool for hitchhiking mapping? *Genetics*, **172**, 287–
1850 292.
1851 Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals - mining
1852 genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**,
1853 749–763.
1854 Schmidt JM, Good RT, Appleton B *et al.* (2010) Copy number variation and transposable
1855 elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genetics*, **6**,
1856 e1000998.
1857 Schmidt PS, Paaby AB (2008) Reproductive Diapause and Life-History Clines in North
1858 American Populations of *Drosophila melanogaster*. *Evolution*, **62**, 1204–1215.
1859 Schmidt PS, Zhu CT, Das J *et al.* (2008) An amino acid polymorphism in the *couch potato*
1860 gene forms the basis for climatic adaptation in *Drosophila melanogaster*. *Proceedings*
1861 *of the National Academy of Sciences of the United States of America*, **105**, 16207–
1862 16211.
1863 Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive Natural Selection in the
1864 *Drosophila* Genome? *PLoS Genetics*, **5**, e1000495.
1865 Sezgin E, Duvernell DD, Matzkin LM *et al.* (2004) Single-locus latitudinal clines and their
1866 relationship to temperate adaptation in metabolic genes and derived alleles in
1867 *Drosophila melanogaster. Genetics*, **168**, 923–931.

1868 Singh AK (2018) Chromosomal variability due to inversions in indian populations of
1869 *Drosophila*. *Indian J Sci Res*.
1870 Singh BN, Das A (1992) Further evidence for latitudinal inversion clines in natural
1871 populations of *Drosophila melanogaster* from India. *Journal of Heredity*, **83**, 227–230.
1872 Singh ND, Arndt PF, Clark AG, Aquadro CF (2009) Strong evidence for lineage and
1873 sequence specificity of substitution rates and patterns in *Drosophila*. *Molecular Biology*
1874 *and Evolution*, **26**, 1591–1605.
1875 Stalker HD (1976) Chromosome studies in wild populations of *D. melanogaster. Genetics*,
1876 **82**, 323–347.
1877 Stalker HD (1980) Chromosome Studies in Wild Populations of *Drosophila melanogaster*.
1878 II. Relationship of Inversion Frequencies to Latitude, Season, Wing-Loading and Flight
1879 Activity. *Genetics*, **95**, 211–223.
1880 Staubach F, Baines JF, Künzel S, Bik EM, Petrov DA (2013) Host species and
1881 environmental effects on bacterial communities associated with *Drosophila* in the
1882 laboratory and in the natural environment. *PLoS ONE*, **8**, e70749.
1883 Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and
1884 its implications. *Philosophical Transactions Of The Royal Society Of London Series B-*
1885 *Biological Sciences*, **365**, 1245–1253.
1886 Svetec N, Pavlidis P, Stephan W (2009) Recent strong positive selection on *Drosophila*
1887 *melanogaster HDAC6*, a gene encoding a stress surveillance factor, as revealed by
1888 population genomic analysis. *Molecular Biology and Evolution*, **26**, 1549–1556.
1889 Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA
1890 polymorphism. *Genetics*, **123**, 585–595.
1891 Trinder M, Daisley BA, Dube JS, Reid G (2017) *Drosophila melanogaster* as a High-
1892 Throughput Model for Host-Microbiota Interactions. *Frontiers in microbiology*, **8**, 751.

1893 Turner TL, Levine MT, Eckert ML, Begun DJ (2008) Genomic analysis of adaptive
1894 differentiation in *Drosophila melanogaster*. *Genetics*, **179**, 455–473.
1895 Umina PA, Weeks AR, Kearney MR, McKechnie SW, Hoffmann AA (2005) A rapid shift in
1896 a classic clinal pattern in *Drosophila* reflecting climate change. *Science*, **308**, 691–693.
1897 Unckless RL (2011) A DNA virus of *Drosophila*. *PLoS ONE*, **6**, e26564.
1898 Van 't Land J, Van Putten WF, Villarroel H, Kamping A, van Delden W (2000) Latitudinal
1899 variation for two enzyme loci and an inversion polymorphism in Drosophila
1900 melanogaster from Central and South America. *Evolution*, **54**, 201–209.
1901 Voelker RA, Cockerham CC, Johnson FM *et al.* (1978) INVERSIONS FAIL TO ACCOUNT
1902 FOR ALLOZYME CLINES. *Genetics*, **88**, 515–527.
1903 Watterson GA (1975) On the number of segregating sites in genetical models without
1904 recombination. *Theoretical Population Biology*, **7**, 256–276.
1905 Webster CL, Longdon B, Lewis SH, Obbard DJ (2016) Twenty-Five New Viruses
1906 Associated with the Drosophilidae (Diptera). *Evolutionary bioinformatics online*, **12**,
1907 13–25.
1908 Webster CL, Waldron FM, Robertson S *et al.* (2015) The Discovery, Distribution, and
1909 Evolution of Viruses Associated with *Drosophila melanogaster. PLoS Biology*, **13**,
1910 e1002210.
1911 Weir BS, Cockerham CC (1984) Estimating *F*-Statistics for the Analysis of Population
1912 Structure. *Evolution*, **38**, 1358–1370.
1913 Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate
1914 biology. *Nature Reviews Microbiology*, **6**, 741–751.
1915 Wesley CS, Eanes WF (1994) Isolation and analysis of the breakpoint sequences of
1916 chromosome inversion *In(3L)Payne* in *Drosophila* melanogaster. *Proceedings of the*
1917 *National Academy of Sciences of the United States of America*, **91**, 3132–3136.

1918 Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer.

1919 Wilfert L, Longdon B, Ferreira AGA, Bayer F, Jiggins FM (2011) Trypanosomatids are
1920 common and diverse parasites of *Drosophila*. *Parasitology*, **138**, 858–865.

1921 Wittmann MJ, Bergland AO, Feldman MW, Schmidt PS, Petrov DA (2017) Seasonally
1922 fluctuating selection can maintain polymorphism at many loci via segregation lift.
1923 *Proceedings of the National Academy of Sciences of the United States of America*,
1924 **114**, E9932–E9941.

1925 Wolf JBW, Bayer T, Haubold B *et al.* (2010) Nucleotide divergence vs. gene expression
1926 differentiation: comparative transcriptome sequencing in natural isolates from the
1927 carrion crow and its hybrid zone with the hooded crow. *Molecular Ecology*, **19**, 162–
1928 175.

1929 Wolff JN, Camus MF, Clancy DJ, Dowling DK (2016) Complete mitochondrial genome
1930 sequences of thirteen globally sourced strains of fruit fly (*Drosophila melanogaster*)
1931 form a powerful model for mitochondrial research. *Mitochondrial DNA. Part A, DNA*
1932 *mapping, sequencing, and analysis*, **27**, 4672–4674.

1933 Xiao F-X, Yotova V, Zietkiewicz E *et al.* (2004) Human *X*-chromosomal lineages in Europe
1934 reveal Middle Eastern and Asiatic contacts. *European Journal of Human Genetics*, **12**,
1935 301–311.

1936 Yukilevich R, True JR (2008a) Incipient sexual isolation among cosmopolitan *Drosophila*
1937 *melanogaster* populations. *Evolution*, **62**, 2112–2121.

1938 Yukilevich R, True JR (2008b) African morphology, behavior and phermones underlie
1939 incipient sexual isolation between us and Caribbean Drosophila melanogaster.
1940 *Evolution*, **62**, 2807–2828.

1941 Yukilevich R, Turner TL, Aoki F, Nuzhdin SV, True JR (2010) Patterns and processes of
1942 genome-wide divergence between North American and African *Drosophila*
1943 *melanogaster*. *Genetics*, **186**, 219–239.
1944 Zanini F, Brodin J, Thebo L *et al.* (2015) Population genomics of intrapatient HIV-1
1945 evolution. (AK Chakraborty, Ed,). *eLife*, **4**, e11282.
1946

Tables

Table 1. Sample information for all populations in the *DrosEU* dataset. The table shows the origin, collection data and season and sample size (number of chromosomes: *n*) of the 48 samples in the *DrosEU* dataset. Additional information can be found in the supporting information in Table S1.

ID	Country	Location	Coll. Date	Number ID	Lat (°)	Lon (°)	Alt (m)	Season	n	Coll. name
AT_Mau_14_01	Austria	Mauternbach	2014-07-20	1	48.38	15.56	572	S	80	Andrea J. Betancourt
AT_Mau_14_02	Austria	Mauternbach	2014-10-19	2	48.38	15.56	572	F	80	Andrea J. Betancourt
TR_Yes_14_03	Turkey	Yesiloz	2014-08-31	3	40.23	32.26	680	S	80	Banu Sebnem Onder
TR_Yes_14_04	Turkey	Yesiloz	2014-10-23	4	40.23	32.26	680	F	80	Banu Sebnem Onder
FR_Vil_14_05	France	Viltain	2014-08-18	5	48.75	2.16	153	S	80	Catherine Montchamp-Moreau
FR_Vil_14_07	France	Viltain	2014-10-27	7	48.75	2.16	153	F	80	Catherine Montchamp-Moreau
FR_Got_14_08	France	Gotheron	2014-07-08	8	44.98	4.93	181	S	80	Cristina Vieira
UK_She_14_09	United Kingdom	Sheffield	2014-08-25	9	53.39	-1.52	100	S	80	Damiano Porcelli
UK_Sou_14_10	United Kingdom	South Queensferry	2014-07-14	10	55.97	-3.35	19	S	80	Darren Obbard
CY_Nic_14_11	Cyprus	Nicosia	2014-08-10	11	35.07	33.32	263	S	80	Eliza Argyridou
UK_Mar_14_12	United Kingdom	Market Harborough	2014-10-20	12	52.48	-0.92	80	F	80	Eran Tauber
UK_Lut_14_13	United Kingdom	Lutterworth	2014-10-20	13	52.43	-1.10	126	F	80	Eran Tauber

DE_Bro_14_14	Germany	Broggingen	2014-06-26	14	48.22	7.82	173	S	80	Fabian Staubach
DE_Bro_14_15	Germany	Broggingen	2014-10-15	15	48.22	7.82	173	F	80	Fabian Staubach
UA_Yal_14_16	Ukraine	Yalta	2014-06-20	16	44.50	34.17	72	S	80	Iryna Kozeretska
UA_Yal_14_18	Ukraine	Yalta	2014-08-27	18	44.50	34.17	72	S	80	Iryna Kozeretska
UA_Ode_14_19	Ukraine	Odesa	2014-07-03	19	46.44	30.77	54	S	80	Iryna Kozeretska
UA_Ode_14_20	Ukraine	Odesa	2014-07-22	20	46.44	30.77	54	S	80	Iryna Kozeretska
UA_Ode_14_21	Ukraine	Odesa	2014-08-29	21	46.44	30.77	54	S	80	Iryna Kozeretska
UA_Ode_14_22	Ukraine	Odesa	2014-10-10	22	46.44	30.77	54	F	80	Iryna Kozeretska
UA_Kyi_14_23	Ukraine	Kyiv	2014-08-09	23	50.34	30.49	179	S	80	Iryna Kozeretska
UA_Kyi_14_24	Ukraine	Kyiv	2014-09-08	24	50.34	30.49	179	F	80	Iryna Kozeretska
UA_Var_14_25	Ukraine	Varva	2014-08-18	25	50.48	32.71	125	S	80	Oleksandra Protsenko
UA_Pyr_14_26	Ukraine	Pyriatyn	2014-08-20	26	50.25	32.52	114	S	80	Oleksandra Protsenko
UA_Dro_14_27	Ukraine	Drogobych	2014-08-24	27	49.33	23.50	275	S	80	Iryna Kozeretska
UA_Cho_14_28	Ukraine	Chornobyl	2014-09-13	28	51.37	30.14	121	F	80	Iryna Kozeretska
UA_Cho_14_29	Ukraine	Chornobyl Yaniv	2014-09-13	29	51.39	30.07	121	F	80	Iryna Kozeretska
SE_Lun_14_30	Sweden	Lund	2014-07-31	30	55.69	13.20	51	S	80	Jessica Abbott
DE_Mun_14_31	Germany	Munich	2014-06-19	31	48.18	11.61	520	S	80	John Parsch
DE_Mun_14_32	Germany	Munich	2014-09-03	32	48.18	11.61	520	F	80	John Parsch
PT_Rec_14_33	Portugal	Recarei	2014-09-26	33	41.15	-8.41	175	F	80	Jorge Vieira

ES_Gim_14_34	Spain	Gimenells (Lleida)	2014-10-20	34	41.62	0.62	173	F	80	Lain Guio
ES_Gim_14_35	Spain	Gimenells (Lleida)	2014-08-13	35	41.62	0.62	173	S	80	Lain Guio
FI_Aka_14_36	Finland	Akaa	2014-07-25	36	61.10	23.52	88	S	80	Maaria Kankare
FI_Aka_14_37	Finland	Akaa	2014-08-27	37	61.10	23.52	88	S	80	Maaria Kankare
FI_Ves_14_38	Finland	Vesanto	2014-07-26	38	62.55	26.24	121	S	66	Maaria Kankare
DK_Kar_14_39	Denmark	Karensminde	2014-09-01	39	55.95	10.21	15	F	80	Mads Fristrup Schou
DK_Kar_14_41	Denmark	Karensminde	2014-11-25	41	55.95	10.21	15	F	80	Mads Fristrup Schou
CH_Cha_14_42	Switzerland	Chalet à Gobet	2014-07-24	42	46.57	6.70	872	S	80	Martin Kapun
CH_Cha_14_43	Switzerland	Chalet à Gobet	2014-10-05	43	46.57	6.70	872	F	80	Martin Kapun
AT_See_14_44	Austria	Seeboden	2014-08-17	44	46.81	13.51	591	S	80	Martin Kapun
UA_Kha_14_45	Ukraine	Kharkiv	2014-07-26	45	49.82	36.05	141	S	80	Svitlana Serga
UA_Kha_14_46	Ukraine	Kharkiv	2014-09-14	46	49.82	36.05	141	F	80	Svitlana Serga
UA_Cho_14_47	Ukraine	Chornobyl Applegarden	2014-09-13	47	51.27	30.22	121	F	80	Svitlana Serga
UA_Cho_14_48	Ukraine	Chornobyl Polisske	2014-09-13	48	51.28	29.39	121	F	70	Svitlana Serga
UA_Kyi_14_49	Ukraine	Kyiv	2014-10-11	49	50.34	30.49	179	F	80	Svitlana Serga
UA_Uma_14_50	Ukraine	Uman	2014-10-01	50	48.75	30.21	214	F	80	Svitlana Serga
RU_Val_14_51	Russia	Valday	2014-08-17	51	57.98	33.24	217	S	80	Elena Pasyukova

Table 2. **Clinality of genetic variation and population structure.** Effects of geographic variables and/or seasonality on genome-wide average levels of diversity (π, θ and Tajima's D; top rows) and on the first three axes of a PCA based on allele frequencies at neutrally evolving sites (bottom rows). The values represent F-ratios from general linear models. Bold type indicates F-ratios that are significant after Bonferroni correction (adjusted α'=0.0055). Asterisks in parentheses indicate significance when accounting for spatial autocorrelation by spatial error models. These models were only calculated when Moran's I test, as shown in the last column, was significant. *p < 0.05; **p < 0.01; ***p < 0.001.

Factor	***Latitude***	***Longitude***	***Altitude***	***Season***	***Moran's I***
$\pi_{(X)}$	4.11*	1.62	**15.23*** **	1.65	0.86
$\pi_{(Aut)}$	0.91	2.54	**27.18*** **	0.16	-0.86
$\theta_{(X)}$	2.65	1.31	**15.54*** **	2.22	0.24
$\theta_{(Aut)}$	0.48	1.44	**13.66*** **	0.37	-1.13
$D_{(X)}$	0.02	0.38	5.93*	3.26	-2.08
$D_{(Aut)}$	0.09	0.76	5.33*	0.71	-1.45
PC1	0.06	**120.72*****(*****)**	5.35*(*)	2.53	4.15***
PC2	3.5	**10.22****	**15.21*** **	1.97	-1.96

	PC3	0.14	0.11	0.01	1.29	0.22
1961						
1962						

Table 3. **Clinality and/or seasonality of chromosomal inversions.** The values represent *F*-ratios from generalized linear models with a binomial error structure to account for frequency data. Bold type indicates deviance values that were significant after Bonferroni correction (adjusted *α'*=0.0071). Stars in parentheses indicate significance when accounting for spatial autocorrelation by spatial error models. These models were only calculated when Moran's *I* test, as shown in the last column, was significant. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Factor	***Latitude***	***Longitude***	***Altitude***	***Season***	***Moran's I***
In(2L)t	2.2	**10.09****	**43.94*****	0.89	-0.92
In(2R)NS	0.25	**14.43*****	2.88	2.43	1.25
In(3L)P	**21.78*****	2.82	0.62	3.6	-1.61
In(3R)C	**18.5***(***)**	0.75	1.42	0.04	2.79**
In(3R)Mo	0.3	0.09	0.35	0.03	-0.9
In(3R)Payne	**43.47*****	0.66	1.69	1.55	-0.89