Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses

Martin Kapun,1,†,‡ Maite G. Barrón,1,4 Fabian Stauber,1,5 Darren J. Obbard,1,6 R. Axel W. Wiberg,1,7,8 Jorge Vieira,1,9,10 Clément Goubert,1,11,12 Omar Rota-Stabelli,1,13 Maaria Kankare,1,14 María Bogaerts-Márquez,1,4 Annabelle Haudry,1,11 Lena Waidele,1,5 Iryna Kozeretska,1,15,16 Elena G. Pasyukova,1,17 Volker Loeschcke,1,18 Marta Pascual,1,19,20 Clément Goubert,1,11,12 Omar Rota-Stabelli,1,13 Maaria Kankare,‡,1,14 María Bogaerts-Márquez,1,4 Annabelle Haudry,1,11 Lena Waidele,1,5 Iryna Kozeretska,1,15,16 Elena G. Pasyukova,1,17 Volker Loeschcke,1,18 Marta Pascual,1,19,20 Cristina Vieira,1,9,10 Svitlana Serga,1,15 Catherine Montchamp-Moreau,1,121 Jessica Abbott,1,22 Patricia Gilbert,1,11 Damiano Porcelli,1,23 Nico Posnien,1,24 Alejandro Sánchez-Gracia,1,19,20 Sonja Grath,1,125 Elio Sucena,1,26,27 Alan O. Bergland,‡,1,28 Maria Pilar Garcia Guerreiro,1,29 Banu Sebnem Onder,1,30 Eliza Argyridou,1,25 Lain Guio,1,4 Mads Fristrup Schou,1,18,22 Bart Deplancke,1,31 Cristina Vieira,1,11 Michael G. Ritchie,1,7 Bas J. Zwaan,1,32 Eran Tauber,1,33,34 Dorcas J. Orenge,1,19,20 Eva Puerma,1,19,20 Montserrat Aguadé,1,19,20 Paul Schmidt,1,35 John Parsch,1,25 Andrea J. Betancourt,1,36 Thomas Flatt,1,‡‡ and Josefa González*†‡

1The European Drosophila Population Genomics Consortium (DrosEU)
2Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
3Department of Biology, University of Fribourg, Fribourg, Switzerland
4Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona, Spain
5Department of Evolutionary Biology and Ecology, University of Freiburg, Freiburg, Germany
6Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom
7Centre for Biological Diversity, School of Biology, University of St. Andrews, St Andrews, Scotland
8Department of Environmental Sciences, Zoological Institute, University of Basel, Basel, Switzerland
9Instituto de Biología Molecular e Celular (IBMC), University of Porto, Porto, Portugal
10Instituto de Investigación e Inovação em Saúde (I3S), University of Porto, Porto, Portugal
11Laboratoire de Biométrie et Biologie Évolutive UMR 5558, CNRS, Université Lyon 1, Université de Lyon, Villeurbanne, France
12Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY
13Research and Innovation Centre, Fondazione Edmund Mach, San Michele all’Adige, Italy
14Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland
15General and Medical Genetics Department, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
16State Institution National Antarctic Scientific Center of Ministry of Education and Science of Ukraine, Kyiv, Ukraine
17Laboratory of Genome Variation, Institute of Molecular Genetics of RAS, Moscow, Russia
18Department of Bioscience—Genetics, Ecology and Evolution, Aarhus University, Aarhus C, Denmark
19Departamento de Genética, Microbiologia e Estatística, Facultad de Biología, Universitat de Barcelona, Barcelona, Spain
20Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
21Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198, Gif-sur-Yvette, France
22Section for Evolutionary Ecology, Department of Biology, Lund University, Lund, Sweden
23Department of Animal and Plant Sciences, Sheffield, United Kingdom
24Johann-Friedrich-Blumenbach-Institut für Zoologie und Anthropologie, Universität Göttingen, Göttingen, Germany
25Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilians-Universität München, Planegg, Germany
26Instituto Gulbenkian de Ciência, Oeiras, Portugal
27Departamento de Biología Animal, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal
28Department of Biology, University of Virginia, Charlottesville, VA
29Department of Genética i Microbiologia, Universitat Autònoma de Barcelona, Barcelona, Spain
30Department of Biology, Faculty of Science, Hacettepe University, Ankara, Turkey
31Institute of Bio-engineering, School of Life Sciences, EPFL, Lausanne, Switzerland

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Genetic variation is the fuel of evolution, with standing genetic variation especially important for short-term evolution and local adaptation. To date, studies of spatiotemporal patterns of genetic variation in natural populations have been challenging, as comprehensive sampling is logistically difficult, and sequencing of entire populations costly. Here, we address these issues using a collaborative approach, sequencing 48 pooled population samples from 32 locations, and perform the first continent-wide genomic analysis of genetic variation in European *Drosophila melanogaster*. Our analyses uncover longitudinal population structure, provide evidence for continent-wide selective sweeps, identify candidate genes for local climate adaptation, and document clines in chromosomal inversion and transposable element frequencies. We also characterize variation among populations in the composition of the fly microbiome, and identify five new DNA viruses in our samples.

**Key words:** population genomics, adaptation, demography, selection, clines, SNPs, structural variants.

**Abstract**

Understanding processes that influence genetic variation in natural populations is fundamental to understanding the process of evolution (Dobzhansky 1970; Lewontin 1974; Kreitman 1983; Kimura 1984; Hudson et al. 1987; McDonald and Kreitman 1991; Adrian and Comeron 2013). Until recently, technological constraints have limited studies of natural genetic variation to small regions of the genome and small numbers of individuals. With the development of population genomics, we can now analyze patterns of genome-wide genetic variation for large numbers of individuals, with samples structured across space and time. As a result, we have new insight into the evolutionary dynamics of genetic variation in natural populations (Begun et al. 2007; Hohenlohe et al. 2010; Cheng et al. 2012; Pool et al. 2012; Harpur et al. 2014; Zanini et al. 2015). But, despite this technological progress, extensive large-scale sampling and genome sequencing of populations remains prohibitively expensive and too labor-intensive for most individual research groups.

Here, we present the first comprehensive, continent-wide genomic analysis of genetic variation of European *Drosophila melanogaster*, based on 48 pool-sequencing samples from 32 localities collected in 2014 (fig. 1) by the European *Drosophila* Population Genomics Consortium (*DroSeU*; https://droseu.net). *Drosophila melanogaster* offers several advantages for genomic studies of evolution in space and time. It boasts a relatively small genome, a broad geographic range, a multi-voltine life history which allows sampling across generations on short timescales, simple standard techniques for collecting wild samples, and a well-developed context for population genomic analysis (Powell 1997; Keller 2007; Hales et al. 2015). Importantly, this species is studied by an extensive international research community, with a long history of developing shared resources (Hales et al. 2015; Bilder and Irvine 2017; Haudry et al. 2020).

Our study complements and extends previous studies of genetic variation in *D. melanogaster*, both from its native range in Southeastern Africa and from its world-wide expansion as a human commensal. The expansion into Europe is thought to have occurred ~4,100–19,000 years ago and into North America and Australia in the last few centuries (David and Capy 1988; Lachaise et al. 1988; Li and Stephan 2006; Keller 2007; Kapopoulou, Kapun, et al. 2018; Arguello et al. 2019; Sprengelmeyer et al. 2020). The colonization of novel habitats and climate zones on multiple continents makes *D. melanogaster* especially useful for studying parallel local adaptation, with previous studies finding pervasive latitudinal clines in allele frequencies (Schmidt and Paaby 2008; Turner et al. 2008; Kolačzkowski et al. 2011; Fabian et al. 2012; Bergland et al. 2014; Kapun, Fabian, et al. 2016; Machado et al. 2016), structural variants such as chromosomal inversions (reviewed in Kapun and Flatt 2019), transposable elements (TEs) (Boussy et al. 1998; González et al. 2008, 2010), and complex phenotypes (de Jong and Bochdanovits 2003; Schmidt and Paaby 2008; Schmidt et al. 2008; Kapun, Schmidt, et al. 2016; Behrman et al. 2018), especially along...
A description of the basic patterns of genetic variation of these European D. melanogaster population samples, based on SNPs, is provided in Supplementary Material online (see supplementary results and table S1, Supplementary Material online). For each sample, we estimated genome-wide levels of \( \pi \), Watterson’s \( \theta \), and Tajima’s \( D \) (corrected for pooling; Futschik and Schlötterer 2010; Kofler et al. 2011). In brief, patterns of genetic variability and Tajima’s \( D \) were largely consistent with what has been previously observed on other continents (Fabian et al. 2012; Langley et al. 2012; Lack et al. 2015, 2016), and genetic diversity across the genome varies mainly with recombination rate (Langley et al. 2012). We also found little spatiotemporal variation among European populations in overall levels of sequence variability (table 2).

Below we focus on the identification of selective sweeps, previously unknown longitudinal population structure across the European continent, patterns of local adaptation and clines, and microbiota.

Several Genomic Regions Show Signatures of Continent-Wide Selective Sweeps

To identify genomic regions that have likely undergone selective sweeps in European populations of D. melanogaster, we used Pool-hmm (Boitard et al. 2013; see supplementary table S2A, Supplementary Material online), which identifies candidate sweep regions via distortions in the allele frequency spectrum. We ran Pool-hmm independently for each sample and identified several genomic regions that coincide with previously identified, well-supported sweeps in the proximity of Hen1 (Kolaczkowski et al. 2011), Cyp6g1 (Daborn et al. 2002), wapi (Beisswanger et al. 2006), and around the chimeric gene CR18217 (Rogers and Hartl 2012), among others (supplementary table S2B, Supplementary Material online). These regions also showed local reductions in \( \pi \) and Tajima’s \( D \), consistent with selective sweeps (fig. 2 and supplementary figs. S1 and S2, Supplementary Material online). The putative sweep regions that we identified in the European populations included 145 of the 232 genes previously identified using Pool-hmm in an Austrian population (Boitard et al. 2012; supplementary table S2C, Supplementary Material online). We also identified other regions which have not previously been described as targets of selective sweeps (supplementary table S2A, Supplementary Material online). Of the regions analyzed, 64 showed signatures of selection across all European populations (supplementary table S2D, Supplementary Material online). Of these, 52 were located in the 10% of regions with the lowest values of Tajima’s \( D \) (SuperExactTest; \( P < 0.001 \)). These may represent continent-wide sweeps that predate the colonization of Europe (Beisswanger et al. 2006) or which have recently swept across the majority of European populations (supplementary table S2D, Supplementary Material online). Indeed, 43 of the 64 genes (67%) that showed signatures of selection across all European populations were located in regions with reduced Tajima’s \( D \) (lowest 10%) in African populations, suggesting that selective sweeps in these genes might predate the out-of-Africa
Table 1. Sample Information for All Populations in the DrosEU Data Set.

<table>
<thead>
<tr>
<th>ID</th>
<th>Country</th>
<th>Location</th>
<th>Coll. Date</th>
<th>Number</th>
<th>Lat (°)</th>
<th>Lon (°)</th>
<th>Alt (m)</th>
<th>Season</th>
<th>n</th>
<th>Coll. Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT_Mau_14_01</td>
<td>Austria</td>
<td>Mauternbach</td>
<td>2014-07-20</td>
<td>1</td>
<td>48.38</td>
<td>15.56</td>
<td>572</td>
<td>S</td>
<td>80</td>
<td>Andrea J. Betancourt</td>
</tr>
<tr>
<td>AT_Mau_14_02</td>
<td>Austria</td>
<td>Mauternbach</td>
<td>2014-10-19</td>
<td>2</td>
<td>48.38</td>
<td>15.56</td>
<td>572</td>
<td>F</td>
<td>80</td>
<td>Andrea J. Betancourt</td>
</tr>
<tr>
<td>TR_Yes_14_03</td>
<td>Turkey</td>
<td>Yesiloz</td>
<td>2014-08-31</td>
<td>3</td>
<td>40.23</td>
<td>32.26</td>
<td>680</td>
<td>S</td>
<td>80</td>
<td>Banu Sebnem Onder</td>
</tr>
<tr>
<td>TR_Yes_14_04</td>
<td>Turkey</td>
<td>Yesiloz</td>
<td>2014-10-23</td>
<td>4</td>
<td>40.23</td>
<td>32.26</td>
<td>680</td>
<td>F</td>
<td>80</td>
<td>Banu Sebnem Onder</td>
</tr>
<tr>
<td>FR_Vil_14_05</td>
<td>France</td>
<td>Villain</td>
<td>2014-08-18</td>
<td>5</td>
<td>48.75</td>
<td>2.16</td>
<td>153</td>
<td>S</td>
<td>80</td>
<td>Catherine Montchamp-Moreau</td>
</tr>
<tr>
<td>FR_Vil_14_07</td>
<td>France</td>
<td>Villain</td>
<td>2014-10-27</td>
<td>7</td>
<td>48.75</td>
<td>2.16</td>
<td>153</td>
<td>F</td>
<td>80</td>
<td>Catherine Montchamp-Moreau</td>
</tr>
<tr>
<td>UK_She_14_09</td>
<td>United Kingdom</td>
<td>Sheffield</td>
<td>2014-08-25</td>
<td>9</td>
<td>53.39</td>
<td>-1.52</td>
<td>100</td>
<td>S</td>
<td>80</td>
<td>Cristina Vieira</td>
</tr>
<tr>
<td>UK_Sou_14_12</td>
<td>United Kingdom</td>
<td>Market Harborough</td>
<td>2014-10-20</td>
<td>12</td>
<td>52.43</td>
<td>-0.92</td>
<td>80</td>
<td>F</td>
<td>80</td>
<td>Eran Tauber</td>
</tr>
<tr>
<td>UK_Lut_14_13</td>
<td>United Kingdom</td>
<td>Lutterworth</td>
<td>2014-10-20</td>
<td>13</td>
<td>52.43</td>
<td>-1.10</td>
<td>126</td>
<td>F</td>
<td>80</td>
<td>Eran Tauber</td>
</tr>
</tbody>
</table>

NOTE.—Origin, collection date, season, and sample size (number of chromosomes: n) of the 48 samples in the DrosEU 2014 data set. Additional information can be found in supplementary table S1, Supplementary Material online.

Expansion (supplementary table S2D, Supplementary Material online).

We then asked if there was any indication of selective sweeps particular to a certain habitat. To this end, we classified the populations according to the Köppen–Geiger climate classification (Peel et al. 2007) and identified several putative sweeps exclusive to arid, temperate, and cold regions (supplementary table S2A, Supplementary Material online). To shed light on potential phenotypes affected by selective sweeps, we performed a gene ontology (GO) analysis. For temperate climates, this analysis showed enrichment for functions such as “response to stimulus,” “transport,” and “nervous system development.” For cold climates, it showed enrichment for “vitamin and cofactor metabolic processes” (supplementary table S2E, Supplementary Material online). There was no enrichment of any GO category for sweeps associated with arid regions.

Thus, we identified several new candidate selective sweeps in European populations of D. melanogaster, many of which...
European Populations Are Structured along an East-West Gradient

We next investigated whether patterns of genetic differentiation might be due to demographic substructuring. Overall, pairwise differentiation as measured by $F_{ST}$ was relatively low, particularly for the autosomes (autosomal $F_{ST}$ 0.013–0.059; X-chromosome $F_{ST}$ 0.043–0.076; Mann–Whitney U test; $P < 0.001$; supplementary table S1, Supplementary Material online). The X chromosome is expected to have higher $F_{ST}$ than the autosomes, given its relatively smaller effective population size (Mann–Whitney U test; $P < 0.001$; Hutter et al. 2007). One population, from Sheffield (UK), was unusually differentiated from the others (average pairwise $F_{ST} = 0.027$; SE = 0.00043 vs. $F_{ST} = 0.04$; SE = 0.00055 for comparisons without this population and with this population only; supplementary table S1, Supplementary Material online). Including this sample in the analysis could potentially lead to exaggerated patterns of geographic differentiation, as it is both highly differentiated and the furthest west. We therefore excluded it from the following analyses of geographic differentiation, as this approach is conservative. (For details, see Supplementary Material online; including or excluding this population did not qualitatively change our results and their interpretation.)

Despite low overall levels of among-population differentiation, we found that European populations exhibit clear evidence of geographic substructuring. For this analysis, we focused on SNPs located within short introns, with a length ≤60 bp and which most likely reflect neutral population structure (Haddrill et al. 2005; Singh et al. 2009; Parsch et al. 2010; Clemente and Vogl 2012; Lawrie et al. 2013). We restricted our analyses to polymorphisms in regions of high recombination ($r > 3$ cM/Mb; Comeron et al. 2012) and to SNPs at least 1 Mb away from the breakpoints of common inversions (excluding the inversion bodies themselves), resulting in 4,034 SNPs used for demographic analysis.

We found two signatures of geographic differentiation using these putatively neutral SNPs. First, we identified a weak but significant correlation between pairwise $F_{ST}$ and geographic distance, consistent with isolation by distance (IBD; Mantel test; $P < 0.001$; $R^2 = 0.12$, max. $F_{ST} \sim 0.045$; fig. 3A). Second, a principal components analysis (PCA) on allele frequencies showed that the three most important PC axes explain >25% of the total variance (PC1: 16.71%, PC2: 5.83%, PC3: 4.6%, eigenvalues = 159.8, 55.7, and 44, respectively; fig. 3B). The first axis, PC1, was strongly correlated with longitude ($F_{1,42} = 118.08, P < 0.001$; table 2). Again, this pattern is consistent with IBD, as the European continent extends further in longitude than latitude. We repeated the above PCA using SNPs in 4-fold degenerate sites, as these are also assumed to be relatively unaffected by selection (Akashi 1995; Halligan and Keightley 2006; supplementary fig. S3, Supplementary Material online), and found highly consistent results.

Because there was a significant spatial autocorrelation between samples (as indicated by Moran’s test on residuals from linear regressions with PC1; $P < 0.001$; table 2), we repeated the analysis with an explicit spatial error model; the association between PC1 and longitude remained significant. To a lesser extent PC2 was likewise correlated with longitude ($F_{1,42} = 7.15, P < 0.05$), but also with altitude ($F_{1,42} = 11.77, P < 0.01$) and latitude ($F_{1,42} = 4.69, P < 0.05$; table 2). Similar to PC2, PC3 was strongly correlated with altitude ($F_{1,42} = 19.91, P < 0.001$; table 2). We also examined these data for signatures of genetic differentiation between samples collected at different times of the year. For the data set as a whole, no major PC axes were correlated with season, indicating that there were no strong differences in allele frequencies shared between all our summer and fall samples ($P > 0.05$ for all analyses; table 2). For the ten locations sampled in both summer and fall, we performed separate PC analyses for summer and fall. Summer and fall values of PC1 (adjusted $R^2$: 0.98; $P < 0.001$), PC2 ($R^2$: 0.74; $P < 0.001$) and PC3 ($R^2$: 0.81; $P < 0.001$) were strongly correlated across seasons. This indicates a high degree of seasonal stability in local genetic variation.

Next, we attempted to determine if populations could be statistically classified into clusters of similar populations. Using hierarchical model fitting based on the first four PC axes from the PCA mentioned above, we found two distinct

---

**Table 2. Clinality of Genetic Variation and Population Structure.**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Altitude</th>
<th>Season</th>
<th>Moran’s $I$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_{(X)}$</td>
<td>4.11*</td>
<td>1.62</td>
<td>15.23***</td>
<td>1.65</td>
<td>0.86</td>
</tr>
<tr>
<td>$\pi_{(Aut)}$</td>
<td>0.91</td>
<td>2.54</td>
<td>27.18***</td>
<td>0.16</td>
<td>-0.86</td>
</tr>
<tr>
<td>$\theta_{(X)}$</td>
<td>2.65</td>
<td>1.31</td>
<td>15.54***</td>
<td>2.22</td>
<td>0.24</td>
</tr>
<tr>
<td>$\theta_{(Aut)}$</td>
<td>0.48</td>
<td>1.44</td>
<td>13.66***</td>
<td>0.37</td>
<td>-1.13</td>
</tr>
<tr>
<td>$D_{(X)}$</td>
<td>0.02</td>
<td>0.38</td>
<td>5.93*</td>
<td>3.26</td>
<td>-2.08</td>
</tr>
<tr>
<td>$D_{(Aut)}$</td>
<td>0.09</td>
<td>0.76</td>
<td>5.33*</td>
<td>0.71</td>
<td>-1.45</td>
</tr>
<tr>
<td>PC1</td>
<td>0.63</td>
<td>118.08***</td>
<td>3.64</td>
<td>0.75</td>
<td>4.2***</td>
</tr>
<tr>
<td>PC2</td>
<td>4.69*</td>
<td>7.15*</td>
<td>11.77**</td>
<td>1.68</td>
<td>-0.32</td>
</tr>
<tr>
<td>PC3</td>
<td>0.39</td>
<td>0.23</td>
<td>19.91***</td>
<td>0.28</td>
<td>1.38</td>
</tr>
</tbody>
</table>

*Note:* 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. Italic type indicates $F$ ratios that are significant after Bonferroni correction (adjusted $R^2 = 0.0055$). 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.

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. 

---


clusters (fig. 3B) separated along PC1, supporting the notion of strong longitudinal differentiation among European populations. Similarly, model-based spatial clustering also showed that populations were separated mainly by longitude (fig. 3C; using ConStruct, with \( K = 3 \) spatial layers chosen based on model selection procedure via cross-validation). We also inferred levels of admixture among populations from this analysis, based on the relationship between \( F_{ST} \) and migration rate (Wright 1949) and using recent estimates of \( N_e \) in European populations (\( N_e / C^2 \approx 3.1 \times 10^6 \); Duchen et al. 2013; for pairwise migration rates see supplementary table S3, Supplementary Material online). Within the Western European cluster and between the clusters, \( 4N_em \) was similar (\( 4N_em - WE = 43.76 \), \( 4N_em - between = 45.97 \)); in Eastern

---

**Fig. 2.** Candidate signals of selective sweeps in European populations. The central panel shows the distribution of Tajima’s \( D \) in 50-kb sliding windows with 40-kb overlap, with red and green dashed lines indicating Tajima’s \( D = 0 \) and \(-1\), respectively. The top panel shows a detail of a genomic region on chromosomal arm 2R in the vicinity of Cyp6g1 and Hen1 (highlighted in red), genes reportedly involved in pesticide resistance. This strong sweep signal is characterized by an excess of low-frequency SNP variants and overall negative Tajima’s \( D \) in all samples. Colored solid lines depict Tajima’s \( D \) for each sample (see supplementary fig. S2, Supplementary Material online, for color codes, Supplementary Material online); the black dashed line shows Tajima’s \( D \) averaged across all samples. The bottom panel shows a region on 3L previously identified as a potential target of selection, which shows a similar strong sweep signature. Notably, both regions show strongly reduced genetic variation (supplementary fig. S1, Supplementary Material online).
Europe, estimates of $4N_e m$ indicate significantly higher levels of admixture, despite the larger geographic range covered by these samples ($4N_e m = 74.17$; Mann-Whitney $U$ test; $P < 0.001$). This result suggests that longitudinal differentiation in Europe might be partly driven by high levels of genetic exchange in Eastern Europe, perhaps due to migration and recolonization after harsh winters in that region. However, these estimates of gene flow must be interpreted with caution, as unknown demographic events can confound estimates of migration rates from $F_{ST}$ (Whitlock and McCauley 1999).

In addition to restricted gene flow between geographic areas, local adaptation may explain population substructure, even at neutral sites, if nearby and closely related populations are responding to similar selective pressures. We investigated whether any of 19 climatic variables, obtained from the WorldClim database (Hijmans et al. 2005), were associated with the genetic structure in our samples. These climatic variables represent interpolated averages across 30 years of observation at the geographic coordinates corresponding to our sampling locations. As many of these variables are highly intercorrelated, we analyzed their joint effects on genetic variation, by using PCA to summarize the information they capture. The first three climatic PC axes capture >77% of the variance in the 19 climatic variables (Supplementary table S4, Supplementary Material online). PC1 explained 36% of the variance and was strongly correlated ($r > 0.75$ or $r < -0.75$) with climatic variables differentiating “hot and dry” from

---

**FIG. 3.** Genetic differentiation among European populations. (A) Average $F_{ST}$ among populations at putatively neutral sites. The center plot shows the distribution of $F_{ST}$ values for all 1,128 pairwise population comparisons, with the $F_{ST}$ values for each comparison obtained from the mean across all 4,034 SNPs used in the analysis. Plots on the left and the right show population pairs in the lower (blue) and upper (red) 5% tails of the $F_{ST}$ distribution. (B) PCA analysis of allele frequencies at the same SNPs reveals population substructuring in Europe. Hierarchical model fitting using the first four PCs showed that the populations fell into two clusters (indicated by red and blue), with cluster assignment of each population subsequently estimated by $k$-means clustering. (C) Admixture proportions for each population inferred by model-based clustering with ConStruct are highlighted as pie charts (left plot) or structure plots (center). The optimal number of three spatial layers ($K$) was inferred by cross-validation (right plot).
Local Climate Adaptation

Climatic Predictors Identify Genomic Signatures of Local Climate Adaptation

To further explore climatic patterns, and to identify signatures of local adaptation caused by climatic differences among populations independent of neutral demographic effects, we tested for associations of SNP alleles with climatic PC1 and PC2 using BayeScEnv (de Villeneuve and Gaggiotti 2015). The total number of SNPs tested and the number of "top SNPs" (q value < 0.05) are given in supplementary table SSA, Supplementary Material online. A large proportion of the top SNPs were intergenic (PC1: 33.5%; PC2: 32.2%) or intronic variants (PC1: 50.1%; PC2: 50.5%). Manhattan plots of q values for all SNPs are shown in figure 4. These figures show some distinct "peaks" of highly differentiated SNPs along with some broader regions of moderately differentiated SNPs (fig. 4). For example, the circadian rhythm gene timeout and the ecdysone signaling genes Eip74EF and Eip75B all lie near peaks associated with climatic PC1 ("hot/dry" vs. "cold/wet"; fig. 4, top panels). We note that the corresponding genes have been identified in previous studies of clinal (latitudinal) differentiation in North American D. melanogaster (Fabian et al. 2012; Machado et al. 2016). We found a significant overlap between genes associated with PC1 and PC2 (both of which are correlated with latitude) in our study and candidate gene sets from these previous studies of latitudinal
clines (SuperExactTest; P < 0.001; Fabian et al. 2012; Machado et al. 2016). For example, out of 1,974 latitudinally varying loci along the North American east coast identified by Fabian et al. (2012), we found 403 (20%) and 505 (26%) of them to also be associated with PC1 and PC2 in European populations, respectively (supplementary table S5 and C, Supplementary Material online). Moreover, the BayeScEnv analysis and Pool-hmm analysis together identify four regions with both climatic associations and evidence for continent-wide selective sweeps (supplementary table S5 and C, Supplementary Material online). Finally, four other BayeScEnv candidate genes were previously identified as targets of selection in African and North American populations based on significant McDonald–Kreitman tests (Langley et al. 2012; see supplementary table S5 and C, Supplementary Material online).

We next asked whether any insights into the targets of local selection could be gleaned from examining the functions of genes near the BayeScEnv peaks. We examined annotated features within 2 kb of significantly associated SNPs (PC1: 3,545 SNPs near 2,078 annotated features; PC2: 5,572 SNPs near 2,717 annotated features; supplementary table S5B and E, Supplementary Material online). First, we performed a GO term analysis with GOwinda (Kofler and Schlötterer 2012) to ask whether SNPs associated with climatic PCs are enriched for any gene functions. For PC1, we found no GO term enrichment. For PC2, we found enrichment for “cuticle development” and “UDP-glucosyltransferase activity.” Next, we performed functional annotation clustering with DAVID (v6.8; Huang et al. 2009), and identified 37 and 47 clusters with an enrichment score > 1.3 for PC1 and PC2, respectively (supplementary table SSD and E, Supplementary Material online, Huang et al. 2009). PC1 was enriched for categories such as “sex differentiation” and “response to nicotine,” whereas PC2 was enriched for functional categories such as “response to nicotine,” “integral component of membrane,” and “sensory perception of chemical stimulus” (supplementary table SSD and E, Supplementary Material online).

We also asked whether the SNPs identified by BayeScEnv show consistent signatures of local adaptation. Many associated genes (1,205) were also shared between PC1 and PC2. Some genes have indeed been previously implicated in climatic and clinal adaptation, such as the circadian rhythm genes *timeless*, *timeout*, and *clock*, the sexual differentiation gene *fruitless*, and the *couch potato* locus which underlies the latitudinal cline in reproductive dormancy in North America (Tauber et al. 2007; Schmidt et al. 2008; Fabian et al. 2012). Notably, these also include the major insulin signaling genes *insulin-like receptor (InR)* and *forkhead box subgroup O (foxo)*, which have strong genomic and experimental evidence implicating these loci in clinal, climatic adaptation along the North America east coast (Paaby et al. 2010, 2014; Fabian et al. 2012; Durmaz et al. 2019). Thus, European populations share multiple potential candidate targets of selection with North American populations (cf. Fabian et al. 2012; Machado et al. 2016; also see Božičević et al. 2016). We next turned to examining polymorphisms other than SNPs, that is, mitochondrial haplotypes as well as inversion and TE polymorphisms.

**Mitochondrial Haplotypes Also Exhibit Longitudinal Population Structure**

Mitochondrial haplotypes also showed evidence of longitudinal demographic structure in European population. We identified two main alternative mitochondrial haplotypes in Europe, G1 and G2, each with several subhaplotypes (G1.1 and G1.2 and G2.1, G2.2, and G2.3). The two subtypes, G1.2 and G2.1, are separated by 41 mutations (fig. 5A). The frequencies of the alternative G1 and G2 haplotypes varied among populations between 35.1% and 95.6% and between 4.4% and 64.9%, respectively (fig. 5B). Qualitatively, three types of European populations could be distinguished based on these haplotypes: 1) central European populations, with a high frequency (>60%) of G1 haplotypes, 2) Eastern European populations in summer, with a low frequency (<40%) of G1 haplotypes, and 3) Iberian and Eastern

**Fig. 5.** Mitochondrial haplotypes. (A) Network showing the relationship of five common mitochondrial haplotypes. (B) Estimated frequency of each mitochondrial haplotype in 48 European samples.
European populations in fall, with a frequency of G1 haplotypes between 40% and 60% (supplementary fig. S4, Supplementary Material online). Analyses of mitochondrial haplotypes from a North American population (Cooper et al. 2015) as well as from world-wide samples (Wolff et al. 2016) also revealed high levels of haplotype diversity.

Although there was no correlation between the frequency of G1 haplotypes and latitude, G1 haplotypes and longitude were weakly but significantly correlated ($r^2 = 0.10; P < 0.05$). We thus divided the data set into an eastern and a western subset along the 20° meridian, corresponding to the division of two major climatic zones, temperate (oceanic) versus cold (continental) (Peel et al. 2007). This split revealed a clear correlation ($r^2 = 0.5; P < 0.001$) between longitude and the frequency of G1 haplotypes, explaining as much as 50% of the variation in the western group (supplementary fig. S4B, Supplementary Material online). Similarly, in eastern populations, longitude and the frequency of G1 haplotypes were correlated ($r^2 = 0.2; P < 0.001$), explaining ~20% of the variance (supplementary fig. S4B, Supplementary Material online). Thus, these mitochondrial haplotypes appear to follow a similar east–west population structure as observed for the nuclear SNPs described above.

The Frequency of Polymorphic TEs Varies with Longitude and Altitude

To examine the population genomics of structural variants, we first focused on TEs. Similar to previous findings, the repetitive content of the 48 samples ranged from 16% to 21% of the nuclear genome size (Quesneville et al. 2005; fig. 6). The vast majority of detected repeats were TEs, mostly long-terminal repeat elements (LTRs; range 7.55–10.15%) and long interspersed nuclear elements (LINEs; range 4.18–5.52%), along with a few DNA elements (range 1.16–1.65%) (supplementary table S6, Supplementary Material online). LTRs have been previously described as being the most abundant TEs in the *D. melanogaster* genome (Kaminker et al. 2002; Bergman et al. 2006). Correspondingly, variation in the proportion of LTRs best explained variation in 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 supplementary fig. S5A, Supplementary Material online).

For each of the 1,630 euchromatic TE insertion sites annotated in the *D. melanogaster* reference genome v.6.04, we estimated the frequency at which a copy of the TE was present at that site using T-lex (Fiston-Lavier et al. 2015; see supplementary table S7, Supplementary Material online). On average, 56% were fixed in all samples. The remaining polymorphic TEs mostly segregated at low frequency in all samples (supplementary fig. S5B, Supplementary Material online), potentially due to purifying selection (González et al. 2008; Petrov et al. 2011; Kofler et al. 2012; Cridland et al. 2013; Blumenstiel et al. 2014). However, 246 were present at intermediate frequencies (>10% and <95%) and located in regions of nonzero recombination (Fiston-Lavier et al. 2010; Comeron et al. 2012; see supplementary table S7, Supplementary Material online). Although some of these insertions might be segregating neutrally at transposition–selection balance (Charlesworth et al. 1994; see supplementary fig. S5B, Supplementary Material online), they are likely enriched for candidate adaptive mutations (Rech et al. 2019).

In each of the 48 samples, TE frequency and recombination rate were negatively correlated genome-wide (Spearman rank sum test; $P < 0.01$), as has also been previously reported for...
D. melanogaster (Bartolomé et al. 2002; Petrov et al. 2011; Kofler et al. 2012). This remains true when fixed (population frequency ≥95%) TE insertions were excluded from the analysis, although it was not statistically significant for some chromosomes and populations (supplementary table S8, Supplementary Material online). In both cases, the correlation was stronger when broad-scale (Fiston-Lavier et al. 2010) rather than fine-scale (Comeron et al. 2012) recombination rate estimates were used, indicating that the former may best capture long-term population recombination patterns (see supplementary materials and methods and table S8, Supplementary Material online).

We next tested whether variation in TE frequencies among samples was associated with spatially or temporally varying factors. We focused on 111 TE insertions that segregated at intermediate frequencies, were located in nonzero recombination regions, and that showed an interquartile range (IQR) > 10 (see supplementary materials and methods, Supplementary Material online). Of these insertions, 57 were significantly associated with at least one variable of interest after multiple testing correction (supplementary table S9A, Supplementary Material online): 13 were significantly associated with longitude, 13 with latitude, five with latitude, three with season, and 23 insertions with more than one of these variables (supplementary table S9A, Supplementary Material online). These 57 TEs were mainly located inside genes (42 out of 57; Fisher’s exact test, P > 0.05; supplementary table S9A, Supplementary Material online).

The 57 TEs significantly associated with these environmental variables were enriched for two TE families: the LTR retrotransposons and populations (supplementary table S8, Supplementary Material online). Interestingly, 17 of the 57 TEs coincided with previously identified adaptive candidate TEs, suggesting that our data set might be enriched for adaptive insertions (SuperExactTest, P < 0.001), several of which exhibit spatial frequency clines that deviate from neutral expectation (SuperExactTest, P < 0.001, supplementary table S9A, Supplementary Material online; cf. Rech et al. 2019). Moreover, 18 of the 57 TEs also show significant correlations with either geographical or temporal variables in North American populations (SuperExactTest, P < 0.001, supplementary table S9A, Supplementary Material online; cf. Lerat et al. 2019).

Polymorphic chromosomal inversions, another class of structural variants besides TEs, are well-known to exhibit pronounced spatial (clinal) patterns in North American, Australian, and other populations, possibly due to spatially varying selection (reviewed in Kapun and Flatt 2019; also see Mettler et al. 1977; Knibb et al. 1981; Lemeunier and Aulard 1992; Hoffmann and Weeks 2007; Fabian et al. 2012; Kapun et al. 2014; Adirion et al. 2015; Rane et al. 2015; Kapun, Fabian, et al. 2016). However, in contrast to North America and Australia, inversion clines in Europe remain very poorly characterized (Lemeunier and Aulard 1992; Kapun and Flatt 2019). We therefore sought to examine the presence and frequency of six cosmopolitan inversions (In(2L)t, In(2R)NS, In(3L)P, In(3R)C, In(3R)Mo, In(3R)Payne) in our European samples, using a panel of highly diagnostic inversion-specific marker SNPs, identified through sequencing of cytologically determined karyotypes by Kapun et al. (2014) (also see Kapun, Fabian, et al. 2016). All 48 samples were polymorphic for one or more inversions (fig. 6). However, only In(2L)t segregated at substantial frequencies in most populations (average frequency = 20.2%); all other inversions were either absent or rare (average frequencies: In(2R)NS = 6.2%, In(3L)P = 4%, In(3R)C = 3.1%, In(3R)Mo = 2.2%, In(3R)Payne = 5.7%) (cf. Kapun, Fabian, et al. 2016; Kapun and Flatt 2019).

Despite their overall low frequencies, several inversions showed pronounced clinality, in qualitative agreement with findings from other continents (Lemeunier and Aulard 1992; Kapun and Flatt 2019). For the analyses below, we tested for potentially confounding effects of significant residual spatial autocorrelation among samples; all of these tests were negative, except for In(3R)C (Moran’s I ≈ 0, P > 0.05 for all tests; table 3).

We observed significant latitudinal clines in Europe for In(3L)P, In(3R)C, and In(3R)Payne (binomial generalized linear model: Inversion frequency ~ Latitude + Longitude + Altitude + Season; effect of Latitude: P < 0.001 for all; see table 3). Clines for In(3L)P and In(3R)Payne were similar between Europe and North America (with frequencies for both decreasing with latitude, P < 0.05; see supplementary table S10, Supplementary Material online). However, all inversions differed in their frequency at the same latitude between North America and Europe (P < 0.001 for the Latitude × Continent interaction; supplementary table S10, Supplementary Material online).

Inversions Exhibit Latitudinal and Longitudinal Clines in Europe

Polymorphic chromosomal inversions, another class of structural variants besides TEs, are well-known to exhibit pronounced spatial (clinal) patterns in North American, Australian, and other populations, possibly due to spatially varying selection (reviewed in Kapun and Flatt 2019; also see Mettler et al. 1977; Knibb et al. 1981; Lemeunier and Aulard 1992; Hoffmann and Weeks 2007; Fabian et al. 2012; Kapun et al. 2014; Adirion et al. 2015; Rane et al. 2015; Kapun, Fabian, et al. 2016). However, in contrast to North America and Australia, inversion clines in Europe remain very poorly characterized (Lemeunier and Aulard 1992; Kapun and Flatt 2019). We therefore sought to examine the presence and frequency of six cosmopolitan inversions (In(2L)t, In(2R)NS, In(3L)P, In(3R)C, In(3R)Mo, In(3R)Payne) in our European samples, using a panel of highly diagnostic inversion-specific marker SNPs, identified through sequencing of cytologically determined karyotypes by Kapun et al. (2014) (also see Kapun, Fabian, et al. 2016). All 48 samples were polymorphic for one or more inversions (fig. 6). However, only In(2L)t segregated at substantial frequencies in most populations (average frequency = 20.2%); all other inversions were either absent or rare (average frequencies: In(2R)NS = 6.2%, In(3L)P = 4%, In(3R)C = 3.1%, In(3R)Mo = 2.2%, In(3R)Payne = 5.7%) (cf. Kapun, Fabian, et al. 2016; Kapun and Flatt 2019).

Despite their overall low frequencies, several inversions showed pronounced clinality, in qualitative agreement with findings from other continents (Lemeunier and Aulard 1992; Kapun and Flatt 2019). For the analyses below, we tested for potentially confounding effects of significant residual spatial autocorrelation among samples; all of these tests were negative, except for In(3R)C (Moran’s I ≈ 0, P > 0.05 for all tests; table 3).

We observed significant latitudinal clines in Europe for In(3L)P, In(3R)C, and In(3R)Payne (binomial generalized linear model: Inversion frequency ~ Latitude + Longitude + Altitude + Season; effect of Latitude: P < 0.001 for all; see table 3). Clines for In(3L)P and In(3R)Payne were similar between Europe and North America (with frequencies for both decreasing with latitude, P < 0.05; see supplementary table S10, Supplementary Material online). However, all inversions differed in their frequency at the same latitude between North America and Europe (P < 0.001 for the Latitude × Continent interaction; supplementary table S10, Supplementary Material online).

**Table 3. Clinality and/or Seasonality of Chromosomal Inversions.**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Altitude</th>
<th>Season</th>
<th>Moran’s I</th>
</tr>
</thead>
<tbody>
<tr>
<td>In(2L)t</td>
<td>2.2</td>
<td>10.09***</td>
<td>43.94***</td>
<td>0.89</td>
<td>−0.92</td>
</tr>
<tr>
<td>In(2R)NS</td>
<td>0.25</td>
<td>14.43***</td>
<td>2.88</td>
<td>2.43</td>
<td>1.25</td>
</tr>
<tr>
<td>In(3L)P</td>
<td>21.78***</td>
<td>2.82</td>
<td>0.62</td>
<td>3.6</td>
<td>−1.61</td>
</tr>
<tr>
<td>In(3R)C</td>
<td>18.5***(***)</td>
<td>0.75</td>
<td>1.42</td>
<td>0.04</td>
<td>2.79**</td>
</tr>
<tr>
<td>In(3R)Mo</td>
<td>0.3</td>
<td>0.09</td>
<td>0.35</td>
<td>0.03</td>
<td>−0.9</td>
</tr>
<tr>
<td>In(3R)Payne</td>
<td>43.47***</td>
<td>0.66</td>
<td>1.69</td>
<td>1.55</td>
<td>−0.89</td>
</tr>
</tbody>
</table>

Note.—The values represent F ratios from binomial generalized linear models to account for frequency data. Underlined italic type indicates deviance values that were significant after Bonferroni correction (adjusted α = 0.0007). 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 < 0.001.**
Latitudinal inversion clines previously observed along the North American and Australian east coasts (supplementary fig. S6 and table S10, Supplementary Material online; Kapun, Fabian, et al. 2016) have been attributed to spatially varying selection, especially in the case of In(3R)Payne (Anderson et al. 2005; Umina et al. 2005; Kennington et al. 2006; Rako et al. 2006; Kapun, Fabian, et al. 2016; Kapun, Schmidt, et al. 2016; Durmaz et al. 2018; Kapun and Flatt 2019). Similar to patterns in North America (Kapun, Fabian, et al. 2016), we observed that clinality of the three inversion polymorphisms was markedly stronger than for putatively neutral SNPs in short introns (see supplementary table S11, Supplementary Material online), suggesting that these polymorphisms are maintained nonneutral. Together, these findings suggest that latitudinal inversion clines in Europe are shaped by spatially varying selection, as they are in North America (Kapun, Fabian, et al. 2016; Kapun and Flatt 2019).

We also detected longitudinal clines for In(2L)t and In(2R)NS, with both polymorphisms decreasing in frequency from east to west (see table 3; P < 0.01; also cf. Kapun and Flatt 2019). Longitudinal clines for these two inversions have also been found in North America (cf. Kapun and Flatt 2019). One of these inversions, In(2L)t, also changed in frequency with altitude (table 3; P < 0.001). These longitudinal and altitudinal inversion clines did, however, not deviate from neutral expectation (supplementary table S11, Supplementary Material online).

European Drosophila Microbiomes Contain Entomophthora, Trypanosomatid, and Previously Unknown DNA Viruses

The microbiota can affect life-history traits, immunity, hormonal physiology, and metabolic homeostasis of their fly hosts (Martino et al. 2017; Trinder et al. 2017) and might thus reveal interesting patterns of local adaptation. We therefore examined the bacterial, fungal, protist, and viral microbiota sequence content of our samples. To do this, we characterized the taxonomic origin of the non-Drosophila reads in our data set using MGRAST, which identifies and counts short protein motifs (features) within reads (Meyer et al. 2008). We examined 262 million reads in total. Of these, most were assigned to Wolbachia (mean 53.7%; fig. 7 and supplementary table S1, Supplementary Material online), a well-known endosymbiont of Drosophila (Werren et al. 2008). The abundance of Wolbachia protein features relative to other microbial protein features (relative abundance) varied strongly between samples, ranging from 8.8% in a sample from Ukraine to almost 100% in samples from Spain, Portugal, Turkey, and Russia (supplementary table S12, Supplementary Material online). Similarly, Wolbachia loads varied 100-fold between samples, as estimated from the ratio of Wolbachia protein features to Drosophila protein features (supplementary table S12, Supplementary Material online). In contrast to a previous study (Kriesner et al. 2016), there was no evidence for clinality of Wolbachia loads (P = 0.13, longitude; P = 0.41, latitude; Kendall’s rank correlation). However, these authors measured infection frequencies whereas we measured Wolbachia loads in pooled samples. Because the frequency of infection does not necessarily correlate with microbial loads measured in pooled samples, we might not have been able to detect such a signal in our data.

Acetic acid bacteria of the genera Gluconobacter, Gluconacetobacter, and Acetobacter were the second largest group, with an average relative abundance of 34.4% among microbial protein features. Furthermore, we found evidence for the presence of several genera of Enterobacteria (Serratia, Yersinia, Klebsiella, Pantoea, Escherichia, Enterobacter, Salmonella, and Pectobacterium). Serratia occurs only at low frequencies or is absent from most of our samples, but reaches a very high relative abundance among microbial protein features in the Nicosia (Cyprus) summer collection (54.5%). This high relative abundance was accompanied by an 80× increase in Serratia bacterial load.

We also detected several eukaryotic microorganisms, although they were less abundant than the bacteria. We found trypanosomatids, previously reported to be associated with Drosophila in other studies (Wilfert et al. 2011; Chandler and
In our data set, Supplementary table S13, Supplementary Material online). This is somewhat surprising because yeasts are commonly found on rotting fruit, the main food substrate of D. melanogaster, and co-occur with flies (Barata et al. 2012; Chandler et al. 2012). This result suggests that, although yeasts can attract flies and play a role in food choice (Becher et al. 2012; Buser et al. 2014), they might not be highly prevalent in or on D. melanogaster bodies. One reason might be that they are actively digested and thus not part of the microbiome. We also found the fungal pathogen Entomophthora muscae in 14 samples, making up 0.18% of the reads (Elya et al. 2018).

Our data also allowed us to identify DNA viruses. Only one DNA virus has been previously described for D. melanogaster (Kallithea virus; Webster et al. 2015; Palmer et al. 2018) and only two additional ones from other Drosophilid species (D. innubila Nudivirus, Unckless 2011; Invertebrate Iridovirus 31 in D. obscura and D. immigrans, Webster et al. 2016). In our data set, ~2 million reads came from Kallithea nudivirus (Webster et al. 2015), allowing us to assemble the first complete Kallithea genome (>300-fold coverage in the Ukrainian sample UA_Kha_14_46; GenBank accession number KX130344).

We also found reads from five additional DNA viruses that were previously unknown (supplementary table S13, Supplementary Material online). First, around 1,000 reads come from a novel nudivirus closely related to both Kallithea virus and to D. innubila nudivirus (Unckless 2011) in sample DK_Kar_14_41 from Karensminde, Denmark (supplementary table S13, Supplementary Material online). As the reads from this virus were insufficient to assemble the genome, we identified a publicly available data set (SRR3939042: 27 male D. melanogaster from Esparto, California; Machado et al. 2016) with sufficient reads to complete the genome (provisionally named “Esparto Virus”; KY608910). Second, we also identified two novel Densoviruses (Parvoviridae). The first is a relative of Culex pipiens densovirus, provisionally named “Viltain virus,” found at 94-fold coverage in sample FR_Vil_14_07 (Viltain; KX648535). The second is “Linvill Road virus,” a relative of Dengroivirus punctatus densovirus, represented by only 300 reads here, but with high coverage in data set SRR2396966 from a North American sample of D. simulans, permitting assembly (KX648536; Machado et al. 2016). Third, we detected a novel member of the Bidnavirinae family, “Vesanto virus,” a bidenvirus related to Bombyx mori densovirus 3 with ~900-fold coverage in sample FI_Ves_14_38 (Vesanto; KX648533 and KX648534). Finally, in one sample (UA_Yal_14_16), we detected a substantial number of reads from an Entomopox-like virus, which we were unable to fully assemble (supplementary table S13, Supplementary Material online).

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 and the entomopox-like virus being the rarest (2/48 and 1/48, respectively). Because Wolbachia can protect Drosophila from viruses (Teixeira et al. 2008), we hypothesized that Wolbachia loads might correlate negatively with viral loads, but found no evidence of such a correlation (P = 0.83 Kallithea virus; P = 0.76 Esparto virus; P = 0.52 Viltain virus; P = 0.96 Vesanto 1 virus; P = 0.93 Vesanto 2 virus; P = 0.5 Linvill Road virus; Kendall’s rank correlation). Perhaps this is because the Kallithea virus, the most prevalent virus in our data set, is not expected to be affected by Wolbachia (Palmer et al. 2018). Similarly, Shi et al. (2018) found no link between Wolbachia and the prevalence or abundance of RNA viruses in data from individual flies.

The variation in bacterial microbiomes across space and time reported here is analyzed in more detail in Wang et al. (2020); this study suggests that some of this variation is structured geographically (cf. Walters et al. 2020). Thus, microbiome composition might contribute to phenotypic differences and local adaptation among populations (Haselkorn et al. 2009; Richardson et al. 2012; Staubs et al. 2013; Kriener et al. 2016; Wang and Staubs 2018).

Conclusions

Here, we have comprehensively sampled and sequenced European populations of D. melanogaster for the first time (fig. 1). We find that European D. melanogaster populations are longitudinally differentiated for putatively neutral SNPs, mitochondrial haplotypes as well as for inversion and TE insertion polymorphisms. Potentially adaptive polymorphisms also show this pattern, possibly driven by the transition from oceanic to continental climate along the longitudinal axis of Europe. We note that this longitudinal differentiation qualitatively resembles the one observed for human populations in Europe (Cavalli-Sforza 1966; Xiao et al. 2004; Francalacci and Sanna 2008; Novembre et al. 2008). Given that D. melanogaster is a human commensal (Keller 2007; Arguello et al. 2019), it is thus tempting to speculate that the demographic history of European populations might have been influenced by past human migration. Outside Europe, east-west structure has been influenced by past human migration. Outside Europe, east–west structure has been previously found in sub-Saharan Africa populations of D. melanogaster, with the split between eastern and western African populations having occurred ~70 ka (Michalakis and Veilleux 1996; Aulard et al. 2002; Kapopoulou, Pfeifer, et al. 2018), a period that coincides with a wave of human migration from eastern into western Africa (Nielsen et al. 2017). However, in contrast to the pronounced pattern observed in Europe, African east–west structure has not been found in Asia or the Americas. It is thus tempting to speculate that the demographic history of European populations might have been influenced by past human migration. Outside Europe, east–west structure has been previously found in sub-Saharan Africa populations of D. melanogaster, with the split between eastern and western African populations having occurred ~70 ka (Michalakis and Veilleux 1996; Aulard et al. 2002; Kapopoulou, Pfeifer, et al. 2018), a period that coincides with a wave of human migration from eastern into western Africa (Nielsen et al. 2017). However, in contrast to the pronounced pattern observed in Europe, African east–west structure has been previously found in sub-Saharan Africa populations of D. melanogaster, with the split between eastern and western African populations having occurred ~70 ka (Michalakis and Veilleux 1996; Aulard et al. 2002; Kapopoulou, Pfeifer, et al. 2018), a period that coincides with a wave of human migration from eastern into western Africa (Nielsen et al. 2017). However, in contrast to the pronounced pattern observed in Europe, African east–west structure has been previously found in sub-Saharan Africa populations of D. melanogaster, with the split between eastern and western African populations having occurred ~70 ka (Michalakis and Veilleux 1996; Aulard et al. 2002; Kapopoulou, Pfeifer, et al. 2018), a period that coincides with a wave of human migration from eastern into western Africa (Nielsen et al. 2017). However, in contrast to the pronounced pattern observed in Europe, African east–west structure has been previously found in sub-Saharan Africa populations of D. melanogaster, with the split between eastern and western African populations having occurred ~70 ka (Michalakis and Veilleux 1996; Aulard et al. 2002; Kapopoulou, Pfeifer, et al. 2018), a period that coincides with a wave of human migration from eastern into western Africa (Nielsen et al. 2017).
Europe are paralleled in North America by the Dros-RTEC consortium (Machado et al. 2019), with whom we are collaborating to compare population genomic data across continents. Together, we have sampled both continents annually since 2014; we aim to continue to sample and sequence European and North American Drosophila populations with increasing spatiotemporal resolution in future years. With these efforts, we hope to provide a rich community resource for biologists interested in molecular population genetics and adaptation genomics.

Materials and Methods

A detailed description of the Materials and Methods is provided in supplementary materials and methods, Supplementary Material online; here, we give a brief overview of the data set and the basic methods used. The 2014 DrosEU data set represents the most comprehensive spatiotemporal sampling of European D. melanogaster populations to date (fig. 1 and supplementary table S1, Supplementary Material online). It comprises 48 samples of D. melanogaster collected from 32 geographical locations across Europe at different time points in 2014 through a joint effort of 18 research groups. Collections were mostly performed with baited traps using a standardized protocol (see supplementary materials and methods, Supplementary Material online). From each collection, we pooled 33–40 wild-caught males. We used males as they are more easily distinguishable morphologically from similar species than females. Despite our precautions, we identified a low level of D. simulans contamination in our sequences; we computationally filtered these sequences from the data prior to further analysis (see Supplementary Material online). To sequence these samples, we extracted DNA and barcoded each sample, and sequenced the ~40 flies per sample as a pool (Pool-Seq: Schlötterer et al. 2014), as paired-end fragments on an Illumina NextSeq 500 sequencer at the Genomics Core Facility of Pompeu Fabra University. Samples were multiplexed in five batches of ten samples, except for one batch of eight samples (supplementary table S1, Supplementary Material online). Each multiplexed batch was sequenced on four lanes at ~50× raw coverage per sample. The read length was 151 bp, with a median insert size of 348 bp (range 209–454 bp). Our genomic data set is available under NCBI Bioproject accession PRJNA388788. Sequences were processed and mapped to the D. melanogaster reference genome (v.6.12) and reference sequences from common commensals and pathogens. Our bioinformatic pipeline is available at https://github.com/capoony/DrosEU_pipeline (last accessed May 22, 2020). To call SNPs, we developed custom software (PoolSNP; see supplementary materials and methods, Supplementary Material online; https://github.com/capoony/PoolSNP, last accessed May 22, 2020), using stringent heuristic parameters. In addition, we obtained genome sequences from African flies from the Drosophila Genome Nexus (DGN; https://www.johnpool.net/genomes.html, last accessed May 22, 2020; see supplementary table S14, Supplementary Material online, for SRA accession numbers). We used data from 14 individuals from Rwanda and 40 from Siavonga (Zambia). We mapped these data to the D. melanogaster reference genome using the same pipeline as for our own data above, and built consensus sequences for each haploid sample by only considering alleles with >0.9 allele frequencies. We converted consensus sequences to VCF and used VCFtools (Danecek et al. 2011) for downstream analyses. Additional steps in the mapping and variant calling pipeline and further downstream analyses of the data are detailed in supplementary materials and methods, Supplementary Material online.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI). We thank three anonymous reviewers and the editors for their helpful comments on a previous version of our article. We are grateful to the 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), the computing facilities of the CC LBBE/PRABI in Lyon (France), the bwUniCluster of the state of Baden–Württemberg (bwHPC), and the University of St Andrews Bioinformatics Unit which is funded by a Wellcome Trust ISSF award (Grant No. 105621/Z/14/Z). We are also grateful to Simon Boitard and Oscar Gaggiotti for their helpful technical advice on Pool-hmm and BayeScEnv analyses, respectively. This study was supported by University of Freiburg Research Innovation Fund 2014, Deutsche Forschungsgemeinschaft (DFG) (Grant No. STA1154/4-1Project 408908608 to F.S.); Academy of Finland (Grant Nos. 268241 and 272927 to M.K.); Russian Foundation of Basic Research (Grant No. 15-54-46009 CT_a to E.G.P.); Danish Natural Science Research Council (Grant No. 4002-00113 to V.L.); Ministerio de Economia y Competitividad (Grant Nos. CTM2017-88080 [AEI/FEDER, UE] and CGL2013-42432-P to M.P. and M.P.G.G., respectively); Centre National de la Recherche Scientifique (Grant No. 14-CE19-0016 to C.V.); FP7, IDEAL (Grant No. FP7/2007-2011/259679 to B.J.Z.); Agence Nationale de la Recherche Exyb (Grant No. 14-CE19-0016 to C.V.); FP6, Network of Excellence LifeSpan (Grant No. FP6 036894 to B.S.O.); FP7, IDEAL (Grant No. FP7/2007-2011/259679 to B.J.Z.); Israel Science Foundation (Grant No. 1737/17 to E.T.); Deutsche Forschungsgemeinschaft (Grant No. PA 903/
References


8-1 to J.P.); Austrian Science Fund (FWF) (Grant Nos. P32275 and P27048 to M.K. and A.J.B., respectively); Biotechnology and Biological Sciences Research Council (BBSRC) (Grant No. BB/P00685X/1 to A.J.B.); Swiss National Science Foundation (SNSF) (Grant Nos. PP00P3_133641, PP00P3_165836, and 31003A_182262 to T.F.); European Commission (Grant No. H2020-ERC-2014CoG-647900 to J.G.); Secretaria d’Universitats i Recerca, Department Economia i Coneixement, Generalitat de Catalunya (Grant No. GRC 2017 SGR 880 to J.G.); Ministerio de Economía y Competitividad/FEDER (Grant No. BFU2014-57779-P to J.G.); and Ministerio de Ciencia e Innovación/AEI/FEDER, EU (Grant No. BFU2017-82937-P to J.G.)


Elya C, Lok TC, Spencer QE, Mcausland H, Martinez CC, Eisen MB. 2018. Robust manipulation of the behavior of Drosophila melanogaster by a fungal pathogen in the laboratory. eLife 7:e34414.


