bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genomic analysis of European *Drosophila* populations reveals major longitudinal structure, continent-wide selection, and unknown DNA viruses

Martin Kapun[1,2,3,4,†,*], Maite G. Barrón[1,5,*], Fabian Staubach[1,6,§], Jorge Vieira[1,7,8], Darren J. Obbard[1,9], Clément Goubert[1,10,11], Omar Rota-Stabelli[1,12], Maaria Kankare[1,13,§], María Bogaerts-Márquez[1,5], Annabelle Haudry[1,10], R. Axel W. Wiberg[1,14,15], Lena Waidele[1,6], Iryna Kozeretska[1,16,17,], Elena G. Pasyukova[1,18], Volker Loeschcke[1,19], Marta Pascual[1,20], Cristina P. Vieira[1,7,8], Svitlana Serga[1,16], Catherine Montchamp-Moreau[1,21], Jessica Abbott[1,22], Patricia Gibert[1,10], Damiano Porcelli[1,23], Nico Posnien[1,24], Alejandro Sánchez-Gracia[1,20], Sonja Grath[1,25], Élio 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[1,5], Mads Fristrup Schou[1,19,22], Bart Deplancke[1,31], Cristina Vieira[1,10], Michael G. Ritchie[1,14], Bas J. Zwaan[1,32], Eran Tauber[1,33], Dorcas J. Orengo[1,20], Eva Puerma[1,20], Montserrat Aguadé[1,20], Paul S. Schmidt[1,34,§], John Parsch[1,25], Andrea J. Betancourt[1,35], Thomas Flatt[1,2,3,†,*,§], Josefa González[1,5,†,*,§]

[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] Current affiliations: Department of Evolutionary Biology and Environmental Sciences, University of Zürich, CH-8057 Zürich, Switzerland; Department of Cell and Developmental Biology, Medical University of Vienna, AT-1090 Vienna, Austria

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

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

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

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

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

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

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

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

[17] State Institution National Antarctic Center of Ministry of Education and Science of Ukraine, 16 Taras Shevchenko Blvd., 01601, Kyiv, Ukraine

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

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

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

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

[26] Instituto Gulbenkian de Ciência, Oeiras, Portugal

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

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

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

[†] **For correspondence**: martin.kapun@uzh.ch, 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 but analysing the spatio-temporal dynamics of genetic changes in natural populations is challenging, comprehensive sampling logistically difficult, and sequencing of entire populations costly. Here we address these issues by performing the first continent-wide genomic analysis of genetic variation in European *Drosophila melanogaster*, based on 48 pool-sequencing samples from 32 populations. Our analyses uncover a novel pattern of major longitudinal population structure; establish previously unknown clines in inversions and transposable elements across Europe; and provide evidence for non-local, continent-wide selective sweeps that are shared among the majority of populations. We also find pronounced variation among populations in the composition of the fly microbiome and identify five new DNA viruses adding to a single example known so far for this species. Our study has important implications for the evolution and demography of *D. melanogaster*, an ancestrally African species that first colonized Europe before becoming cosmopolitan.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Keywords: *P*opulation genomics, demography, selection, clines, SNPs, structural variants, symbionts, viruses.

Introduction

Studying the processes that create and maintain 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 & Kreitman 1991; Adrian & Comeron 2013). Until recently, technological constraints have limited studies of natural genetic variation to small genomic regions and small numbers of individuals. With the development of population genomics, we can now analyse patterns of genetic variation on a genome-wide scale for large numbers of individuals, with sampling structured across space and time. As a result, we have gained fundamental new insights into evolutionary dynamics of genetic variation in natural populations (e.g., Hohenlohe *et al.* 2010; Cheng *et al.* 2012; Begun *et al.* 2007; Pool *et al.* 2012; Harpur *et al.* 2014; Zanini *et al.* 2015). Despite this recent technological progress, extensive large-scale sampling and genome sequencing of populations remains prohibitively expensive in terms of cost and labor for any individual research group.

Here, we present the first comprehensive, continent-wide genomic analysis of genetic variation in European *Drosophila melanogaster*, based on 48 pool-sequencing samples from 32 populations collected in 2014 (Figure 1) by the European *Drosophila* Population Genomics Consortium (*DrosEU*; https://droseu.net). *D. melanogaster* offers several advantages for studying the relevant spatio-temporal scales of evolution: a relatively small genome, a broad geographic range, a multivoltine life history that allows sampling across

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

generations over short timescales, ease of sampling natural populations using standardized techniques, and a well-developed context for population genomic analysis (e.g., Powell 1997; Keller 2007; Hales *et al.* 2015). Importantly, this species is studied by an extensive research community, with a long history of developing shared resources (Larracuente & Roberts 2015; Bilder & Irvine 2017).

The current study complements and extends previous studies of genetic variation in *D. melanogaster*, both from its native range in sub-Saharan Africa and from its world-wide expansion as a human commensal into Europe 10–20,000 years ago and into North America and Australia in the last few centuries (e.g., Lachaise *et al.* 1988; David & Capy 1988; Li & Stephan 2006; Keller 2007; Sprengelmeyer et al 2018; Arguello *et al.* 2019; also cf. Kapopoulou *et al.* 2018a). The colonization of novel habitats and climate zones on multiple continents makes *D. melanogaster* especially powerful for studying parallel local adaptation. Previous studies of genomic variation have uncovered latitudinal clines in allele frequencies (e.g., Schmidt & Paaby 2008; Turner *et al.* 2008; Kolaczkowski *et al.* 2011b; Fabian *et al.* 2012; Bergland *et al.* 2014; Machado *et al.* 2016; Kapun *et al.* 2016a), structural variants such as chromosomal inversions (reviewed in Kapun & Flatt 2019),) transposable elements (TEs) (Boussy *et al.* 1998; González *et al.* 2008; 2010), and complex phenotypes (de Jong & Bochdanovits 2003; Schmidt & Paaby 2008; Schmidt *et al.* 2008; Kapun *et al.* 2016b; Behrman *et al.* 2018). Thus far, sampling across these latitudinal gradients has been restricted to single transects on the east coasts of Australia and North America; in addition to parallel local adaptation, clines on these continents may be due to admixture between cohorts of flies with different colonization histories (Caracristi & Schlötterer 2003; Yukilevich & True 2008a; b; Duchen *et al.* 2013; Kao *et al.* 2015; Bergland *et al.* 2016).

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In contrast, the population genomics of *D. melanogaster* on the European continent remains largely uncharacterized (Božičević *et al.* 2016; Pool *et al.* 2016; Mateo *et al.* 2018). Because Eurasia was the first continent colonized by *D. melanogaster* as they migrated out of Africa, we sought to understand how this species has adapted to new habitats and climate zones in Europe, where it has been established the longest (Lachaise *et al.* 1988; David & Capy 1988). We analyse our data at three levels: (1) variation at single-nucleotide polymorphisms (SNPs) in nuclear and mitochondrial (mtDNA) genomes (~5.5 x 10^6 SNPs in total); (2) structural variation, including TE insertions and chromosomal inversions; and (3) variation in the microbiota associated with flies, including bacteria, fungi, protists, and viruses.

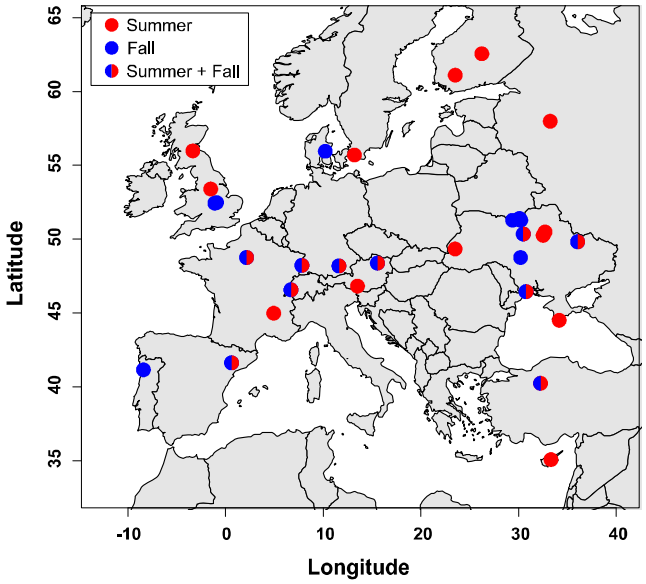

Figure 1. The geographic distribution of population samples. Locations of all samples in the 2014 *DrosEU* data set. The color of the circles indicates the sampling season for each location: ten of the 32 locations were sampled at least twice, once in summer and once in fall (see Table 1 and Supplemental Table 1). Note that some of the 12 Ukrainian locations overlap in the map.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

As part of the *DrosEU* consortium, we collected 48 population samples of *D. melanogaster* from 32 geographical locations across Europe in 2014 (Table 1; Figure 1). We performed pooled sequencing (Pool-Seq) of all 48 samples, with an average autosomal coverage ≥50x (Table S1). Of the 32 locations, 10 were sampled at least once in summer and once in fall (Figure 1), allowing a preliminary analysis of seasonal change in allele frequencies on a genome-wide scale.

European and other derived populations exhibit similar amounts of genetic variation

For each sample, we estimated genome-wide levels of nucleotide diversity (π and Watterson's θ, corrected for pooling; Futschik 2010; Kofler *et al.* 2011). We find that most European populations have similar levels of genetic variation (Table S1). Moreover, our estimates of pairwise nucleotide diversity are similar to those from derived non-African (North American and Australian) populations, whether sequenced as individuals or as pools (Figure 2 and Table S2). Thus, although European populations are considerably older than North American and Australian populations, they exhibit similar levels of DNA sequence variability.

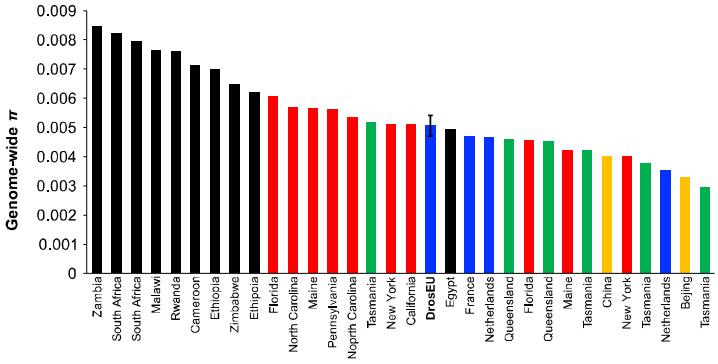

Figure 2. Genetic variation in worldwide samples. Bar plot showing the distribution of genome-wide estimates of Tajima's π of the *DrosEU* and other genomic datasets (also see Table S2 and Materials and Methods) The error bar in the *DrosEU* dataset represent the standard deviation of π across all 48 population samples.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We next tested for associations between geographic variables and genome-wide average levels of genetic variation. We found that neither π nor θ was correlated with latitude or longitude, but both strongly decreased with altitude (Table 2). This contrasts with previous studies of flies collected from a broader range of altitudes, which found increased genetic diversity in high-elevation populations (Lian *et al.* 2018). Finally, we tested for a correlation between genome-wide variation and the season of collection, finding no relationship (Table 2). Together, these results suggest that there is little spatio-temporal variation among European populations in overall levels of sequence variability.

For all populations, the ratio of *X*-linked to autosomal variation (π_X/π_A) was well below the value of 0.75 expected under neutrality with equal sex ratios (ranging from 0.53 to 0.66, one-sample Wilcoxon rank test, $p < 0.001$). These estimates are broadly consistent with those from previous studies of European and other non-African populations (e.g. Andolfatto 2001; Kauer *et al.* 2002; Hutter *et al.* 2007; Betancourt *et al.* 2004; Mackay *et al.* 2012; Langley *et al.* 2012). Surprisingly, the π_X/π_A ratio increased significantly, significantly, albeit weakly, with latitude (Spearman's $\rho = 0.315$, $p = 0.0289$). This observation is at odds with a the predictions of a simple model of periodic bottlenecks leading to a lower X/A ratio in northern populations (Hutter *et al.* 2007; Pool & Nielsen 2007), but might be consistent with stronger selection or more male-biased sex-ratios in the south as compared to the north (Charlesworth 2001; Hutter *et al*. 2007).

Genetic variation was heterogeneous across the genome, as has been previously reported (Begun & Aquadro 1992; Mackay *et al.* 2012; Langley *et al.* 2012; Huang *et al.* 2014). Both π and θ were markedly reduced close to centromeric and telomeric regions (Figure 3), and strongly positively correlated with recombination rate (linear regression against fine-scale

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recombination rate estimates from Comeron *et al.* (2012), $p < 0.001$; not accounting for autocorrelation; Table S3). Recombination rate explained 41–47% and 31–38% of the variation in π, for the autosomes and X chromosome, respectively. Using broad-scale recombination rate estimates (Fiston-Lavier *et al.* 2010) yielded a qualitatively similar, but slightly stronger correlation in autosomes and weaker in the X chromosome (Figure 3, Table S3, Figure 3 - figure supplement 1).

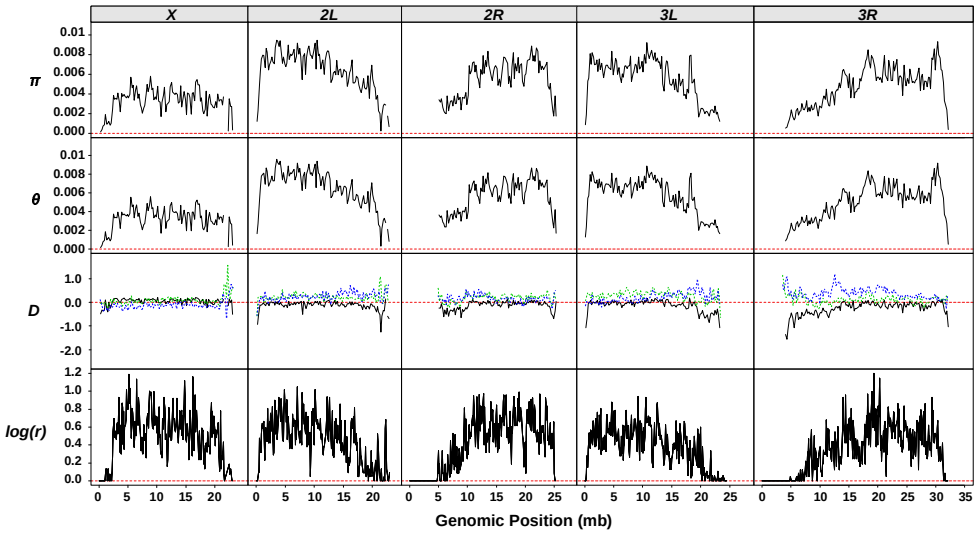

Figure 3 with 1 supplement**. Genome-wide estimates of genetic diversity and recombination rates**. The distribution of Tajima's π, Watterson's θ and Tajima's D (from top to bottom) in 200 kb non-overlapping windows plotted for each chromosomal arm separately. The dashed blue and green lines show estimates for 14 individuals from Rwanda and Zambia, respectively. Bold black lines depict statistics, that were averaged across all 48 samples and the upper and lower grey areas 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 100 kb non-overlapping windows as obtained from Comeron et al. (2010).

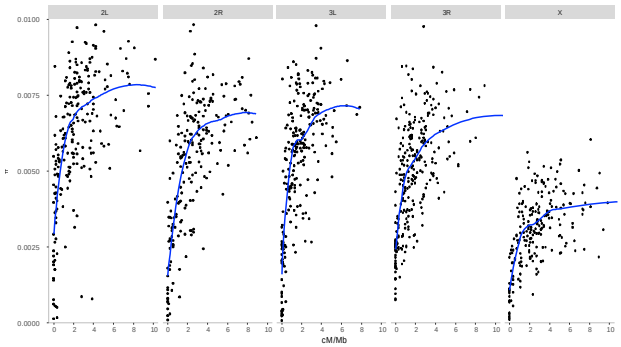

Figure 3 - figure supplement 1. Correlation between recombination and genetic diversity. Smooth local regression (LOESS) between recombination rate in cM/Mb (Comeron et al. 2012) and the average of the 48 samples' genetic diversity (π) in 100 kb non-overlapping windows by chromosome arm.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In contrast to π and θ, the European populations showed major differences in mean Tajima's *D* (Table S1). Tajima's *D* measures deviations from neutral expectations in allele frequencies, which can be due either to selection or complex demography, with negative *D* indicating an excess of low-frequency variants (Tajima 1983). Approximately half of the European samples have negative *D*, It is possible that this result is artefactual, caused by heterogeneity in the proportion of sequencing errors among multiplexed sequencing runs. However, this is unlikely, because including sequence run as a covariate in the statistical model did not improve its fit (Supplementary File 2; Table S4). In all of these analyses, we controlled for confounding effects of spatio-temporal autocorrelations between samples by accounting for similarity among spatial neighbors (Moran's $I \approx 0$, $p > 0.05$ for all tests). When comparing *D* in European samples with ancestral African populations from Zambia and Rwanda, the values were generally lower in the European populations, possibly due to the recent range and population size expansion (Figure 3 and Table S5). Similar to genetic diversity, *D* was also heterogeneous across the genome. Tajima's *D* was broadly reduced in the vicinity of telomeric and centromeric regions, possibly reflecting extended purifying selection or selective sweeps close to heterochromatic regions, and due to reduced recombination.

Several genomic regions show signatures of continent-wide selective sweeps

Genomic regions that show localized reductions in Tajima's *D* are attractive candidates for having undergone recent selective sweeps. To identify such genomic regions, we used *Pool-hmm* (Boitard *et al.* 2013; Table S6A), which – like Tajima's *D* – identifies candidate sweep regions *via* distortions in the allele frequency spectrum. Several genomic regions identified in this way coincide with previously identified, well-supported sweeps in the proximity of *Hen1* (Kolaczkowski *et al.* 2011b), *Cyp6g1* (Daborn *et al.* 2002), *wapl*

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Beisswanger *et al.* 2006), and around the chimeric gene *CR18217* (Rogers & Hartl 2012), among others (Table S6B). These regions also showed local reductions in Tajima's *D* and genetic variation, again consistent with selection (Figure 4 and Figure 4-figure supplement 1 and 2). The putative sweep regions included 145 of the 232 genes previously identified using *Pool-hmm* in an Austrian population (Boitard et al 2012; Table S6C). Other regions identified have not previously been described as harboring sweeps; these represent potential novel targets of positive selection deserving of further investigation (Table S6A). Overall, we identified 64 genes that showed signatures of selection across all European populations analysed (Table S6D); thirty-five of them were located in regions with low Tajima's *D*. This pattern suggests the existence of continent-wide sweeps that either predate the colonization of Europe (e.g., Beisswanger *et al.* 2006), or that have swept across the majority of European populations more recently (Table S6D). Finally, we classified the populations according to the Köppen-Geiger climate classification (Peel *et al.* 2007) and identified several candidate sweeps exclusive to arid, temperate or cold regions; Table S6A). For temperate climates, candidate sweep regions were enriched for functions such as 'response to stimulus', 'transport', and 'nervous system development'; for cold climates, they were enriched for 'vitamin and co-factor metabolic processes' (Table S6E). In contrast, we did not find any significant GO enrichment for arid candidate sweep regions. In summary, this dataset represents a rich genomic resource for future in-depth studies of selective sweeps and adaptation to different climates in *Drosophila*.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

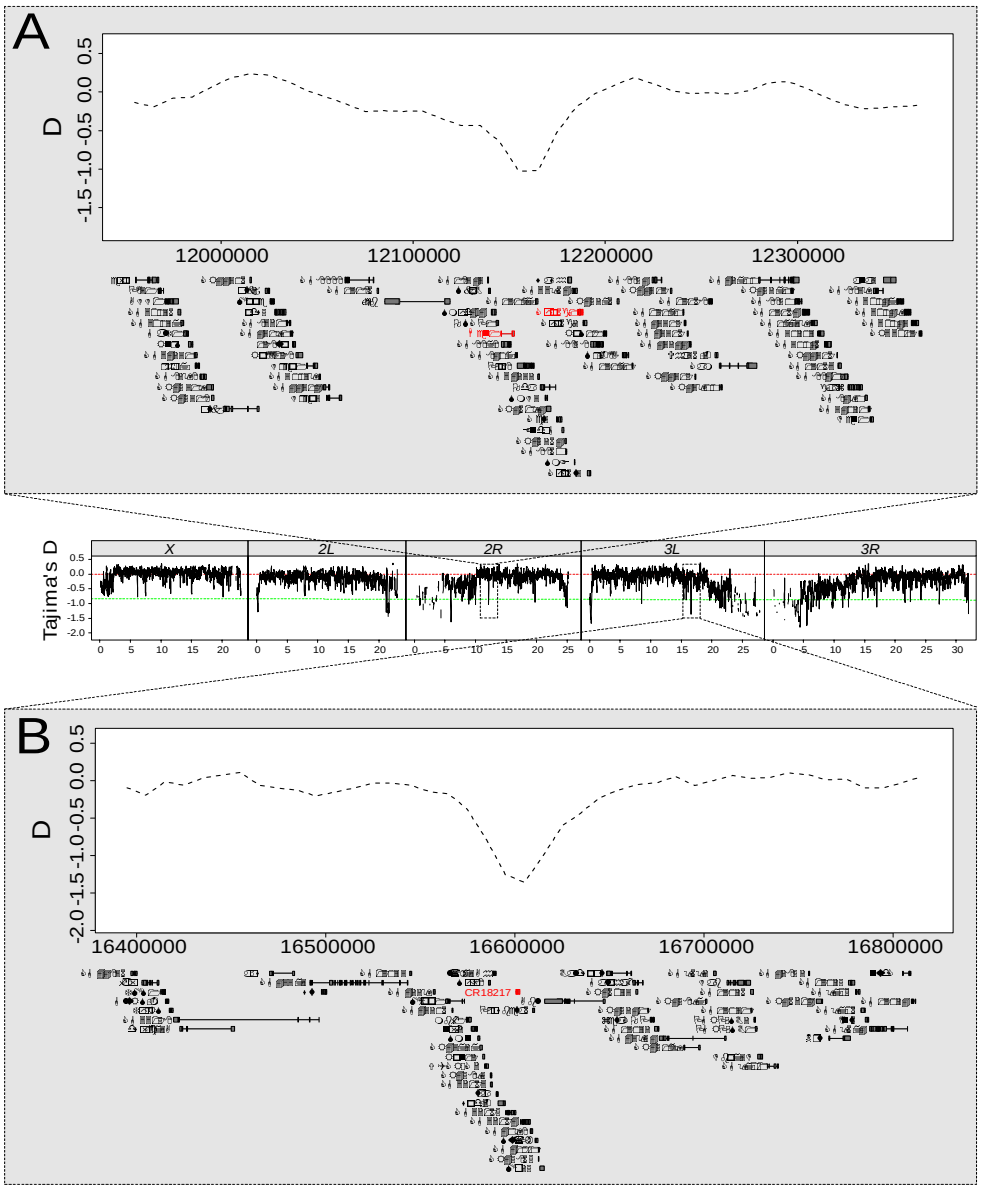

Figure 4 with 2 supplements. **Signals of selective sweeps.** The central panel shows the distribution of Tajima's *D* in 50 kb sliding windows with 40 kb overlap, with red and green dashed lines indicating Tajima's *D* = 0 and -1, respectively. The top panel shows a detail of a genomic region on chromosomal 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. Coloured solid lines depict Tajima's *D* for each sample (see SI Figure 4 for color codes); 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 (Figure 4 - figure supplement 1).

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

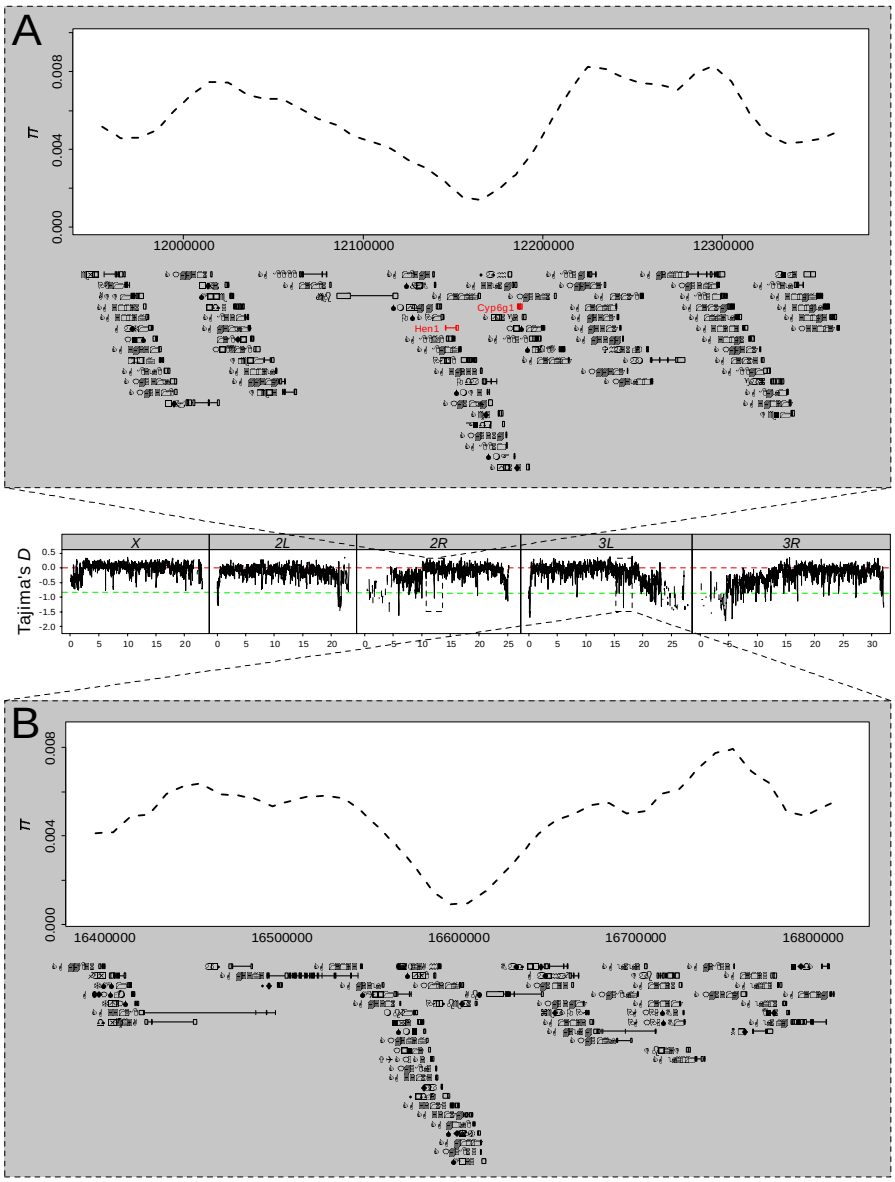

Figure 4 - figure supplement 1: Genetic variation in regions of putative selective sweeps. This figure is equivalent to Figure 4 in the main text but shows the distribution of genetic variation *(π)* in regions with depressed Tajima's *D* around the well-studied *Cyp6g1* locus (A) and around a previously known 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 4 - figure supplement 2.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AT_14_Mau_1	UA_14_Ode_19	ES_14_Lle_35
AT_14_Mau_2	UA_14_Ode_20	FI_14_Aka_36
TR_14_Yes_3	UA_14_Ode_21	FI_14_Aka_37
TR_14_Yes_4	UA_14_Ode_22	FI_14_Ves_38
FR_14_Vil_5	UA_14_Kyi_23	DK_14_Kar_39
FR_14_Vil_7	UA_14_Kyi_24	DK_14_Kar_41
FR_14_Got_8	UA_14_Var_25	CH_14_Cha_42
UK_14_She_9	UA_14_Pyr_26	CH_14_Cha_43
UK_14_Sou_10	UA_14_Dro_27	AT_14_See_44
CY_14_Nic_11	UA_14_Cho_28	UA_14_Kha_45
UK_14_Mar_12	UA_14_Cho_29	UA_14_Kha_46
UK_14_Lut_13	SE_14_Lun_30	UA_14_Cho_47
DE_14_Bro_14	DE_14_Mun_31	UA_14_Cho_48
DE_14_Bro_15	DE_14_Mun_32	UA_14_Kyi_49
UA_14_Yal_16	PT_14_Rec_33	UA_14_Uma_50
UA_14_Yal_18	ES_14_Lle_34	RU_14_Vald_51

Figure 4 - figure supplement 2. Legend for color code in Figure 4, Figure 4 - figure supplement 1.

European populations are 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, particularly for the autosomes (autosomal F_{ST} 0.013–0.059; *X*-chromosome F_{ST}: 0.043–0.076; Mann-Whitney-U test; $p < 0.001$; Table S1). The slightly elevated F_{ST} for the *X* chromosome is expected given its smaller effective population size (Hutter *et al.* 2007). One population, from Sheffield (UK), was unusually differentiated from the others (Table S1) and was excluded from analyses of neutral genetic differentiation. Despite overall low levels of among-population differentiation, European populations showed evidence of geographic substructure. To analyse this pattern in detail, we focused on SNPs most likely to reflect neutral population structure, those at 4-fold degenerate sites, in regions outside those showing signatures of selective sweeps, in regions of high recombination (r > 3cM/Mb; Comeron et al. 2011) and at least 1 Mb away from the breakpoints of common inversions.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The final filtered data set consisted of 8,727 SNPs. Within Europe, we found a weak but significant pattern of isolation by distance (IBD). That is, pairwise F_{ST}, though low overall, increased significantly with geographic distance (Mantel test; $p < 0.001$; r=0.65, max. F_{ST} ~ 0.05; Figure 5A and Figure 5A – figure supplement 1A).

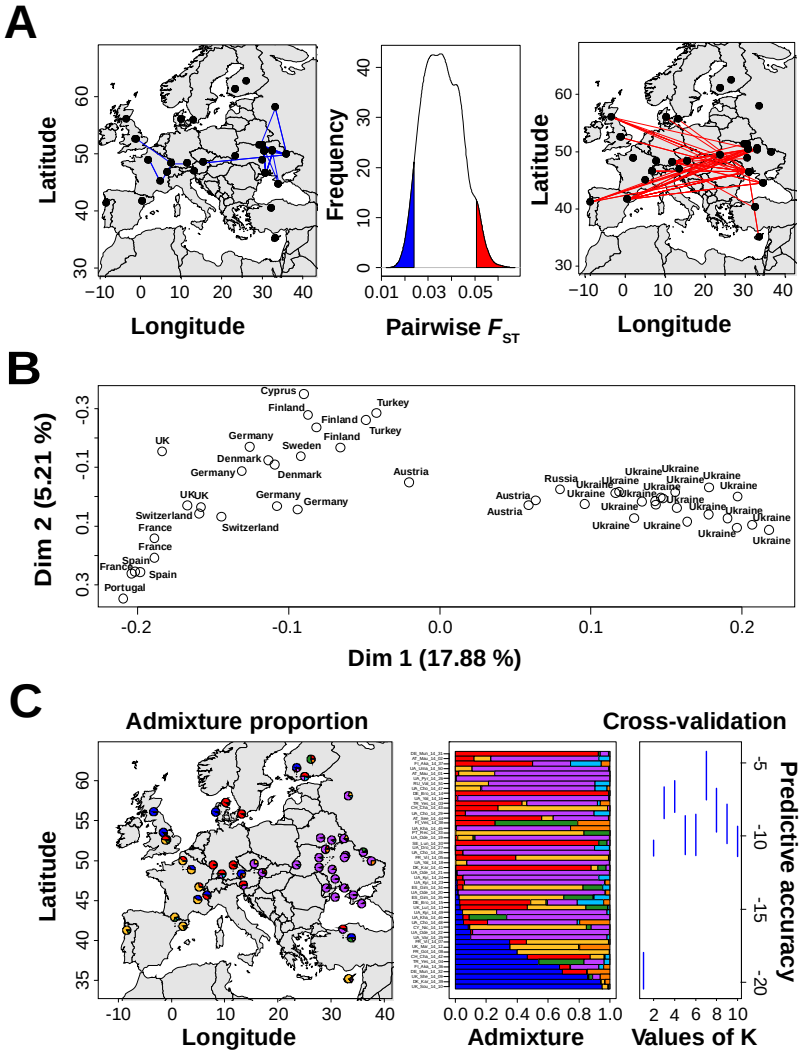

Figure 5 with 1 supplement**. Genetic differentiation among European populations**. (A) Average F_{ST} among populations at putatively neutral sites. The centre plot shows the distribution of F_{ST} values for all 1,128 pairwise population comparisons, with the F_{ST} values for each comparison obtained from the mean across all 8,727 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 three clusters (indicated by colour), 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 (centre). The optimal number of 7 spatial layers (K) was inferred by cross-validation (right plot).

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

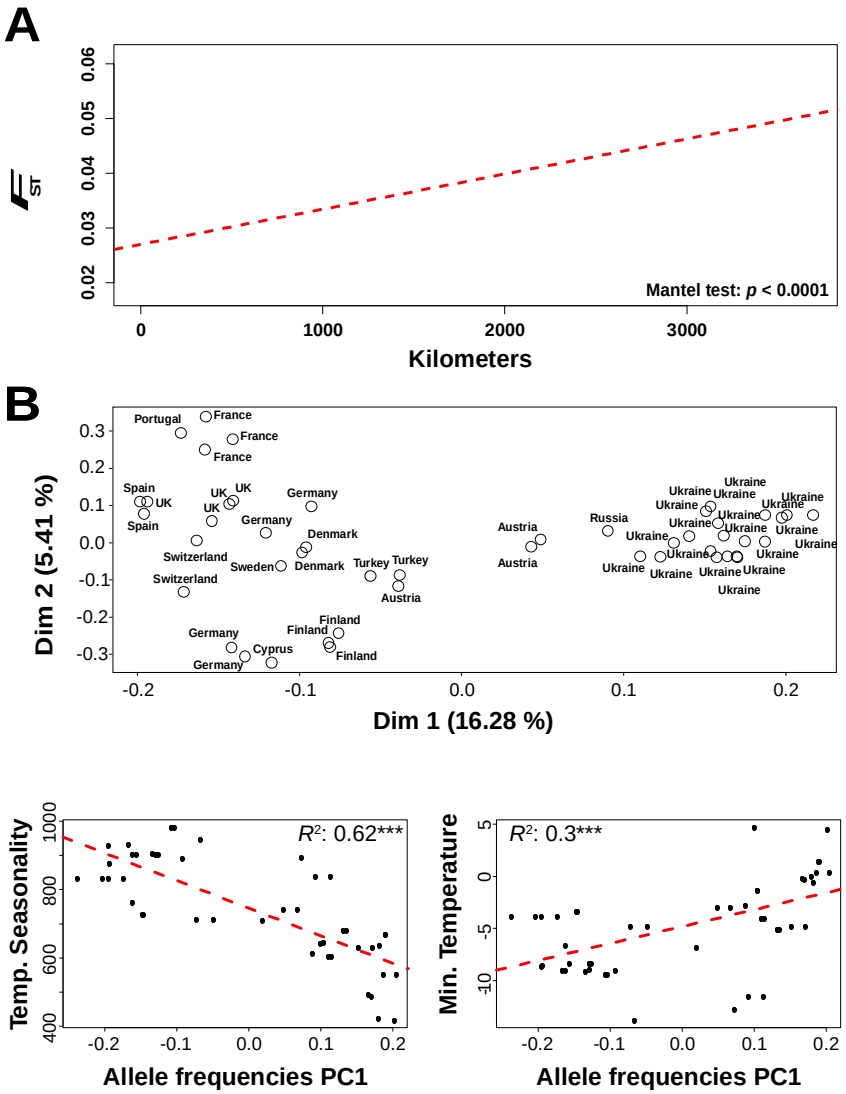

Figure 5 - figure supplement 1: Genetic differentiation among European populations. (A) Average F_{ST} for 8,727 putatively neutral SNPs is significantly negatively correlated with geographic distance (red dashed line shows the linear regression) (B) PCA-based inference of population structure similar to Figure 5B in the main text, but based on 20,008 SNPs located in short introns (<60bp). (C) We tested the top 5 PC for significant associations with 8 climatic variables obtained from the WorldClim database; the two significant regressions, between PC1 and Temperature seasonality (WorldClim Biovar 4; left) and between PC1 and minimum temperature of the coldest month (WorldClim Biovar 6; right) are shown.

We investigated population substructure using principal components analysis (PCA) on allele frequencies from the same set of SNPs at 4-fold degenerate sites. The first three PC axes explained >25% of the total variance (PC1: 17.88%, PC2: 5.2%, PC3: 4.7%, eigenvalues = 410, 101, and 92, respectively), with PC1 strongly correlated with longitude

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and to a lesser extent with altitude (Table 2). This longitudinal stratification is expected under a simple model of IBD, as the continent extends further in longitude than latitude. As there was significant spatial autocorrelation between samples (as indicated by Moran's test on residuals from linear regressions with PC1), we repeated the analysis with an explicit spatial error model; the association between PC1 and longitude remained significant. Like PC1, PC2 is correlated with longitude and altitude. PC3, by contrast, is not associated with any variable examined (Table 2). No major PC axes were correlated with season, indicating that there were no shared seasonal differences across samples in our data. However, based on linear regressions comparing summer and fall values of PC1 (adjusted R^2: 0.98; p-value < 0.001), PC2 (R^2: 0.79; p-value < 0.001) and PC3 (R^2: 0.93; p-value < 0.001), we found very strong associations of genetic variation across seasons in the 10 locations that were sampled in summer and fall. This indicates a high degree of spatio-temporal stability in the levels of genetic variation.

Hierarchical model fitting based on the first three PC axes resulted in three distinct clusters (Figure 5B) separated along PC1, supporting the notion of strong longitudinal differentiation among European populations. Importantly, these results remain qualitatively unchanged when restricting the analysis to SNPs located in short introns (< 60 bp), which are also assumed to be relatively unaffected by selection (Figure 5 – figure supplement 1B; Haddrill *et al.* 2005; Singh *et al.* 2009; Parsch *et al.* 2010; Clemente & Vogl 2012; Lawrie *et al.* 2013).

Model-based spatial clustering showed qualitatively similar results, with populations separated mainly by longitude (Figure 5C; using ConStruct, with K=7 spatial layers chosen based on model selection procedure *via* cross-validation). We could also infer levels of

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

admixture among populations from this analysis; population samples from eastern and northwestern Europe showed low levels of admixture, while those from central Europe appeared locally well-mixed (Figure 5C).

In addition to restricted gene flow between geographic areas, local adaptation may explain population substructuring, even at neutral sites, if closely related populations tend to respond to similar selective pressures. We thus probed whether this spatial substructuring is associated with any of nineteen climatic variables, obtained from the WorldClim database (Hijmans *et al.* 2005). These climatic variables represent averages interpolated averages across more than 50 years of observation at the geographic coordinates corresponding to our sampling locations. Only two variables are significant after Bonferroni correction (adjusted $\alpha = 0.0026$): between PC1 and 'temperature seasonality' (BioVar 4; Hijmans et al. 2005; $R^2 = 0.62$, $P<0.001$; Figure 5 – figure supplement 1C) and between PC1 and 'minimum temperature of the coldest month' ($R^2 = 0.3$, $P<0.001$; Figure 5 – figure supplement 1C). This suggests that the pronounced longitudinal differentiation along the European continent could at least partly be driven by the transition from oceanic to continental climate, leading to gradual changes in temperature seasonality and the severity of winter conditions which might impact demography, especially local survival. To the best of our knowledge, such strongly pronounced longitudinal structure and differentiation on a continent-wide scale has not yet been reported for *D. melanogaster.*

Mitochondrial haplotypes also exhibit longitudinal population structure

Our finding that European populations show strong longitudinal structure is also supported by an analysis of mitochondrial haplotypes. We identified two main mitochondrial haplotypes in Europe, separated by 41 mutations (G1.2 and G2.1; Figure 6A), with highly

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variable frequencies among populations (Figure 6B). Qualitatively, three types of European populations can be distinguished based on these haplotypes: (1) central European populations with a high frequency (> 60%) of the G1 haplotypes, (2) Eastern European populations in summer, with a low frequency (< 40%) of G1 haplotypes, and (3) Iberian and Eastern European populations in fall, with a combined frequency of G1 haplotypes between 40-60% (Figure 6 - figure supplement 1A). These results are consistent with analyses of mitochondrial haplotypes from a North American population (Cooper *et al.* 2015) as well as from worldwide samples (Wolff *et al.* 2016), which revealed a high level of haplotype diversity.

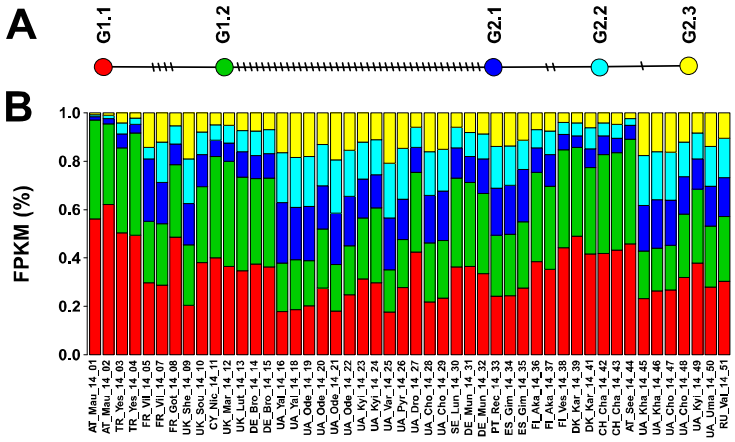

Figure 6 with 1 supplement**. Mitochondrial haplotypes.** (A) TCS network showing the relationship of 5 common mitochondrial haplotypes; (B) estimated frequency of each mitochondrial haplotype in 48 European samples.

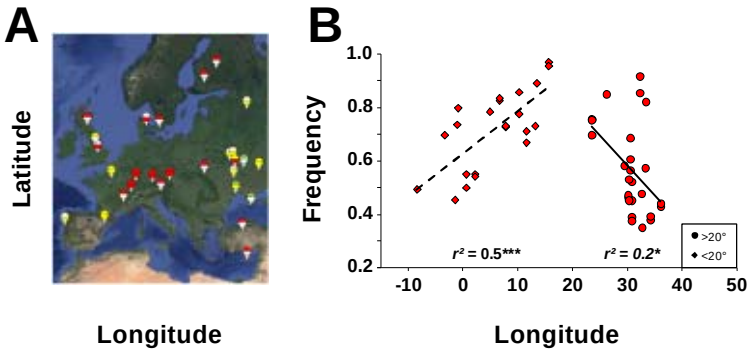

Figure 6 - figure supplement 1. Mitochondrial haplotypes. (A) Graphical summary of the combined frequency of G1

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

haplotypes in Europe. Summer and Fall are represented at the top and bottom of the circles, respectively. White – no information; green, yellow and red represent a combined frequency of G1 haplotypes lower than 40%, in between 40% and 60% and higher than 60%, respectively. (B) Correlations between the combined frequency of G1 haplotypes and longitude (red diamonds for western populations below 20° and red circles for eastern populations above 20°).

Mitochondrial haplotypes also showed shifts in the relative frequencies of the two haplotype classes between summer and fall, but only in 2 of 9 possible comparisons. While there was no correlation between latitude and the frequency of G1 haplotypes, we found a weak but significant negative correlation between G1 haplotypes and longitude ($r^2 = 0.10$; $p < 0.05$), consistent with the longitudinal east-west population structure observed for SNPs at 4-fold degenerate sites. In a subsequent analysis, we divided the dataset at 20º longitude into an eastern and a western subset because in northern Europe 20º longitude corresponds to the division of two major climatic zones, temperate and cold (Peel *et al.* 2007). This split revealed 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 6 - figure supplement 1B). Similarly, in eastern populations, longitude and the combined frequency of G1 haplotypes were correlated, explaining approximately 20% of the variance (Figure 6 - figure supplement 1B). Thus, these data on mitochondrial haplotypes clearly confirm the pronounced east-west structure and differentiation among European populations of *D. melanogaster*.

The frequency of polymorphic TEs varies with longitude and altitude

To examine the population genetics of structural variants, we first focused on transposable elements (TEs). The repetitive content of the 48 samples ranged from 16% to 21% with respect to nuclear genome size (Figure 7). The vast majority of detected repeats were TEs, mostly represented by long terminal repeats (LTR) and long interspersed nuclear elements (LINE), as well as a few DNA elements (Class II). LTRs best explained total TE

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 7- figure supplement 1A).

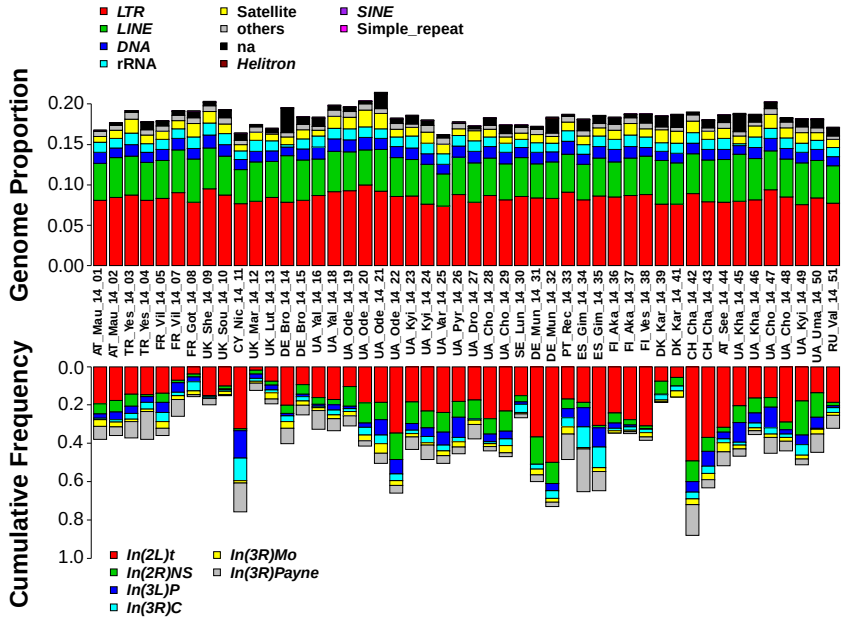

Figure 7 with 2 supplements. **Geographic patterns in structural variants**. The upper panel shows stacked bar plots with the relative abundances of TEs in all 48 population samples**.** The proportion of each repeat class was estimated from sampled reads with dnaPipeTE (2 samples per run, 0.1X coverage per sample). The lower panel shows stacked bar plots depicting absolute frequencies of six cosmopolitan inversions in all 48 population samples.

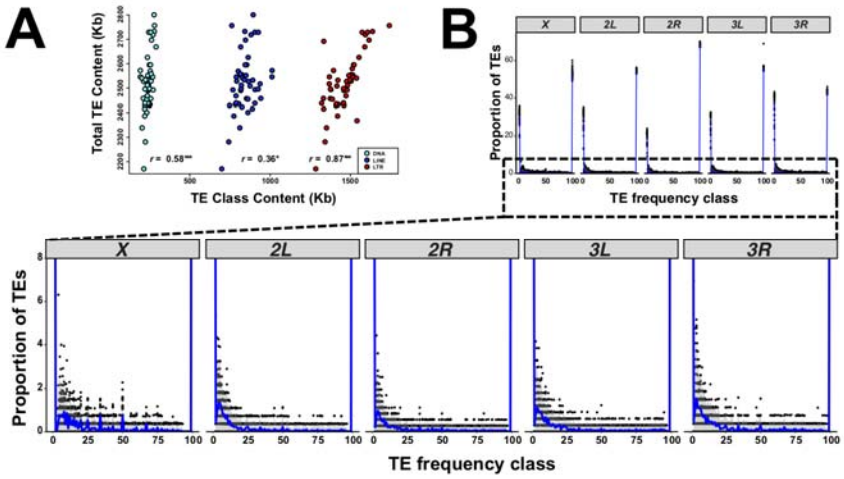

Figure 7- figure supplement 1. **Transposable Elements genome content and frequency distributions.** (A) Pearson's correlations between each main TE class (LTR, LINE and DNA) and the total TE content of each pool (LTR+LINE+DNA) in kb. (B) The site frequency spectrum of TE frequencies per chromosome arm. Each dot represents the proportion of TEs in each bin per sample and a smoother geometric line had been added to highlight the trend. Lower panel is a zoom in of the above panel.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We next estimated population frequencies of 1,630 TE insertions annotated in the *D. melanogaster* reference genome *v.6.04* using *T-lex2* (Table S7, Fiston-Lavier *et al.* 2015). On average, 56% of the TEs annotated in the reference genome were fixed in all samples. The majority of the remaining polymorphic TEs segregated at low frequency in all samples (Figure 7 - figure supplement 1A), potentially due to the effect of purifying selection (González *et al.* 2008; Petrov *et al.* 2011; Kofler *et al.* 2012; Cridland *et al.* 2013; Blumenstiel *et al.* 2014). However, we also observed 142 TE insertions present at intermediate (>10% and <95%) frequencies, which might be consistent with transposition-selection balance (Figure 7 - figure supplement 1B; 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 held when only polymorphic TEs (population frequency <95%) were analysed, although it was not statistically significant for some chromosomes and populations (Table S8). In either case, the correlation was more negative when using broad-scale (Fiston-Lavier *et al.* 2010), rather than fine-scale (Comeron et al 2012), recombination rate estimates, indicating that broad-scale recombination patterns may best capture long-term population recombination patterns (Materials and methods, Tables S8).

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) and were located in regions of non-zero recombination according to both fine-scale (Comeron et al. 2012), and broad-scale (Fiston-Lavier et al. (2010)

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

estimations. Of these, 57 TEs showed significant associations with geographical or temporal variables after multiple testing correction (Table S9). We found significant correlations of 13 TEs with longitude, 13 with altitude, five with latitude, and three 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. These TEs were scattered along the five main chromosome arms, with the majority located inside genes (42 out of 57; Table S9).

Two TE families were enriched in the 57 TE dataset: the LTR *297* family with 11 copies, and the DNA *pogo* family with five copies (χ^2-values after Yate's correction < 0.05; Table S10). Interestingly, 14 of these 57 TEs coincide with previously identified adaptive candidate TEs, suggesting that our dataset might be enriched for adaptive insertions, several of which seem to exhibit spatial frequency clines (Table S9; Rech *et al*. 2019).

Inversions exhibit latitudinal and longitudinal clines in Europe

Another class of structural variants, chromosomal inversions, show spatial patterns in North American and Australian populations, potentially due to selection (reviewed in Kapun & Flatt 2019). In contrast to North America and Australia, inversion clines in Europe are poorly characterized (Lemeunier & Aulard 1992). Here, we examined 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 inversion-specific marker SNPs (Kapun *et al.* 2014). All samples were polymorphic for one or more inversions (Figure 7). 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%,

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In(3R)Payne = 5.7%).

Despite their overall low frequencies, several inversions exhibited pronounced clinality (Table 3). In particular, we observed significant latitudinal clines for *In(3L)P, In(3R)C* and *In(3R)Payne*. Although they differed in overall frequencies, *In(3L)P* and *In(3R)Payne* showed latitudinal clines in Europe qualitatively similar to clines previously observed along the North American and Australian east coasts (Figure 7 - figure supplement 2 and Table S11, Kapun *et al.* 2016a), which, at least in the case of *In(3R)Payne*, are maintained by spatially varying selection (Kapun *et al*. 2016a,b; Durmaz *et al.* 2018; Anderson *et al.* 2005; Umina *et al.* 2005; Kennington *et al.* 2006; Rako *et al.* 2006).

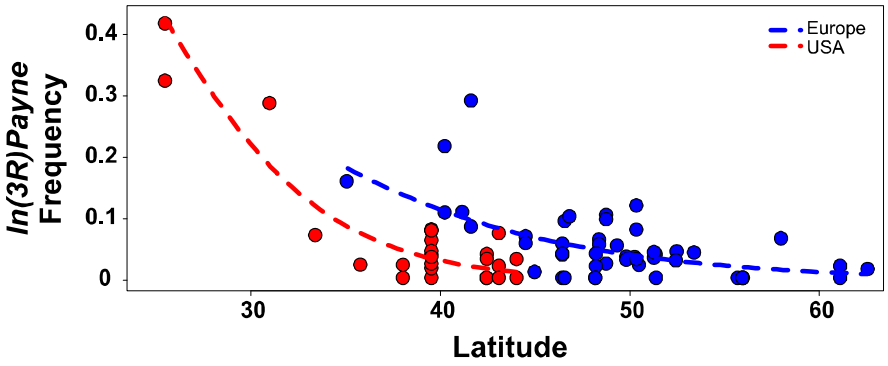

Figure 7 - figure supplement 2. Clinal variation of the inversion *In(3R)Payne* across continents. Parallel frequency clines of *In(3R)Payne* along the latitudinal axis at the North American east coast (red) and in Europe (blue) (see also Table S11).

We also detected – for the first time – longitudinal clines for *In(2L)t* and *In(2R)NS*, with both polymorphisms decreasing in frequency from east to west, a result consistent with the strong longitudinal population differentiation in Europe. *In(2L)t* also increased in frequency with altitude (Table 3). Except for *In(3R)C,* we did not find significant residual spatio-temporal autocorrelation among samples for any inversion tested (Moran's $I \approx 0$, $p > 0.05$

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for all tests; Table 3), suggesting that our analysis was not confounded by spatial autocorrelation for most of the inversions. Further studies are necessary to determine the extent to which these clines of inversion frequencies in Europe are shaped by selection.

European *Drosophila* microbiomes contain Entomophthora, trypanosomatids and unknown DNA viruses

We examined the bacterial, fungal, protist, and viral microbiota associated with *D. melanogaster* using the Pool-Seq data. The microbiota can affect life history traits, immunity, hormonal physiology, and metabolic homeostasis of their fly hosts (e.g., Trinder *et al.* 2017; Martino *et al.* 2017).

We characterised the taxonomic origin of the non-*Drosophila* reads in our dataset using MGRAST, which identifies and counts short protein motifs ('features') within reads (Meyer *et al.* 2008). We examined 262 million reads in total and of these most were assigned to *Wolbachia* (mean 53.7%; Figure 8), 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 the UK to almost 100% in samples from Spain, Portugal, Turkey and Russia (Table S12). Similarly, *Wolbachia* loads varied 100-fold between samples, as estimated from the ratio of *Wolbachia* protein features to *Drosophila* protein features (Table S12).

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

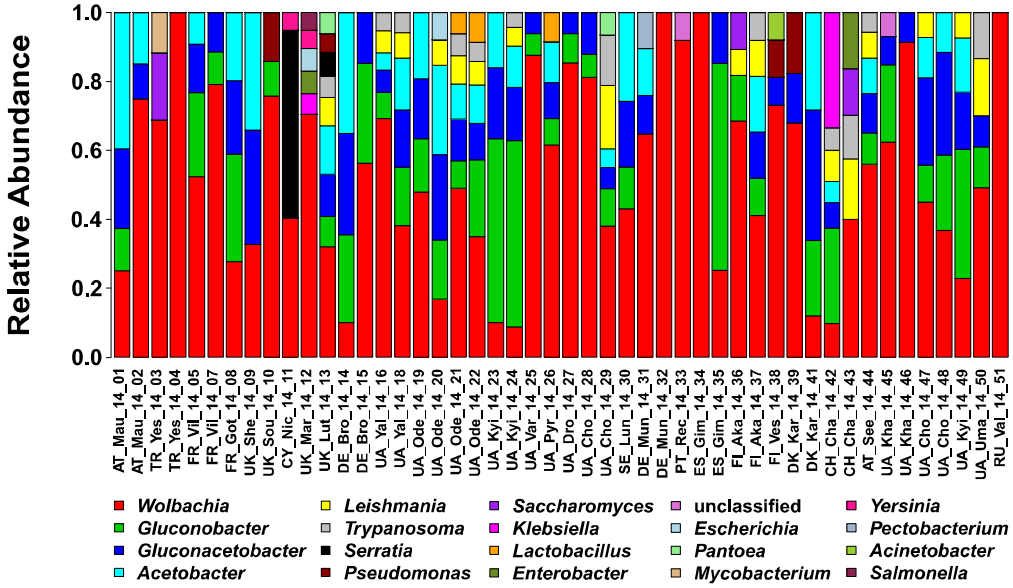

Figure 8: Microbiome. Relative abundance of *Drosophila*-associated microbes as assessed by MGRAST classified shotgun sequences. Microbes had to reach at least 3% relative abundance in one of the samples to be represented

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 80x increase in *Serratia* bacterial load.

We also detected several eukaryotic microorganisms, although they were less abundant than the bacteria. The fraction of fungal protein features, for example, is larger than 3% in only three samples (from Finland, Austria and Turkey; Table S12). Among the eukaryotic microbiota, we found trypanosomatids in 16 samples. Trypanosomatids have been

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

previously reported to be associated with *Drosophila* (Wilfert *et al.* 2011; Chandler & James 2013; Hamilton *et al.* 2015), and this appeared to have been confirmed in this first systematic survey across a wide geographic range in *D. melanogaster*. We also found the fungal pathogen *Entomophthora muscae* in 14 samples (Elya C et al. 2018). Somewhat surprisingly, we found few yeast sequences. Yeasts are commonly found on rotting fruit, the main food substrate of *D. melanogaster*, and have been found in association with *Drosophila* before (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 but are rather actively digested and thus not part of the microbiome.

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 others more from other Drosophilid species (*Drosophila innubila* Nudivirus [Unckless 2011], Invertebrate Iridovirus 31 in *D. obscura* and *D. immigrans* [Webster *et al.* 2016]).

Here, we found six different DNA viruses, five of which are new (Table S13). Approximately two 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 KX130344). We also identified around 1,000 reads from a novel nudivirus closely related to both *Kallithea* virus and to *Drosophila innubila* nudivirus (Unckless 2011) in sample DK_Kar_14_41 from Karensminde, Denmark (Table S13). As the reads from this virus in our data set were insufficient to assemble the genome, we identified a publicly available dataset (SRR3939042: 27 male *D. melanogaster* from Esparto, California; Machado *et al.* 2016)

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with sufficient reads to complete the genome (provisionally named "*Esparto* Virus"; KY608910).

We further 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 *Dendrolimus punctatus* densovirus, represented by only 300 reads here, but with 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). 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 (Table S13). 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).

Discussion

In recent years, large-scale population re-sequencing projects have produced major insights into the biology of both model (Mackay *et al.* 2012; Langley *et al.* 2012; Auton et al. 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). In particular, 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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

scales for population genomic analyses remain largely unknown (e.g., Guirao-Rico and González 2019). Here we have applied – for the first time – a continent-wide sampling and sequencing strategy to European populations of *D. melanogaster* (Figure 1), allowing us to uncover previously unknown aspects of this species' population biology and evolutionary genetics. This is particularly important because the population genomics of this species in Europe has been poorly characterized to date.

We find that European *D. melanogaster* populations exhibit pronounced longitudinal differentiation. We observed this pattern for a genome-wide set of SNPs at 4-fold degenerate sites, which presumably evolve neutrally (Figure 5), as well as for mitochondrial haplotypes, inversions and TEs which might be subject to spatially varying selection (Figure 6 and 7). Longitudinal differentiation might be due to the transition from oceanic to continental climate along the longitudinal axis (Figure 5-Figure 5 supplement 1). While spatial differences in climatic conditions likely play a major role in driving this pattern, we note that it is remarkably similar to that observed for human populations (e.g., Cavalli-Sforza 1966; Xiao et al. 2004; Francalacci & Sanna 2008; Novembre *et al.* 2008). Indeed, 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 kya ago (Michalakis & Veuille 1996; Aulard *et al.* 2002; Kapopoulou *et al.* 2018b), a period that – interestingly – 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 is relatively weak, explaining only ~2.7% of variation, and is due to an inversion whose frequency varies longitudinally. In contrast, our demographic analyses are based on SNPs located in >1 Mb distance from the breakpoints of the most common inversions. This makes it very unlikely that the strong

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

longitudinal pattern we have observed is driven by inversions.

Spatial patterns of differentiation were stronger for longitude than for latitude. In contrast, differentiation in North America has mainly been observed across latitude, for both neutral and adaptive polymorphisms (e.g., Machado *et al.* 2016; Kapun *et al.* 2016a; reviewed in Adrion *et al.* 2015). Although our present analysis showed that putatively neutral SNPs were primarily differentiated along longitude, latitudinal clines may still exist for adaptive polymorphisms. In fact, we detected latitudinal frequency clines for both inversions and TEs (Table 3 and Table S9). For the inversions *In(3L)P* and *In(3R)Payne*, the observed 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; Leumeunier & Aulard 1992; Fabian *et al.* 2012; Kapun *et al.* 2014; Rane *et al.* 2015; Kapun *et al.* 2016a). This pattern is widely thought to be a result of climate adaptation, with the inversions containing variants that make them better adapted to tropical or subtropical than to temperate, more seasonal climates (e.g., Kapun *et al*. 2016a). Several euchromatic TE insertions also showed geographic (or seasonal) patterns of variation (Table S9), indicating that they might play a role in local adaptation, particularly since many of them are located in regions where they might affect gene regulation. Further, 17 of them also show significant correlations with either geographical or temporal variables in North American populations (Lerat *et al.* 2019). Additionally, several inversions and TEs also exhibited longitudinal gradients.

We also examined signatures of selective sweeps in our data. Several of the identified regions have previously been reported as potential targets of positive selection (Figure 4,

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S6B and SC). However, most of these sweeps were originally identified by analysing a small number of populations (*e.g.* Kolaczkowski *et al.* 2011b; Daborn *et al.* 2002; Rogers & Hartl 2012). Here, we identified 64 genes (including *wapl*, *CR18217*, and *mgl*) which showed clear signatures of selection and which were widespread across Europe, thus strengthening the case for their adaptive significance. In addition, we found several regions with evidence of hard sweeps, some of them showing evidence of local climatic adaptation (Table S6); these candidate regions represent a valuable resource for future analyses of adaptation in European *Drosophila.*

Finally, our continent-wide analysis of the microbiota suggests that natural populations of European *D. melanogaster* vary greatly in the composition and abundance of microbes and viruses over space and time. Recent work suggests that at least parts of this variation in microbiomes follows geographic patterns (Walters et al 2018, Wang et al 2019) and contribute to phenotypic differences and local adaptation among populations, especially given that there might be tight and presumably local co-evolutionary interactions between fly hosts and their endosymbionts (e.g., Haselkorn *et al.* 2009; Richardson *et al.* 2012; Staubach *et al.* 2013; Kriesner *et al.* 2016; Wang and Staubach 2018). Most notably, we discovered five new DNA viruses of *D. melanogaster*. Despite this species being host to a wide diversity of RNA viruses, we now have found that the DNA viruses of *D. melanogaster* are also widespread, for instance with *Kallithea* virus detected in most populations.

Our study demonstrates that sampling on a continent-wide scale and pooled sequencing of a large number of natural populations can reveal fundamental and novel aspects of population biology, even for a well-studied model species such as *D. melanogaster*. Our

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

extensive sampling was feasible only due to synergistic collaboration among many research groups. Our efforts in Europe are paralleled in North America by the *Dros-RTEC* consortium, 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 spatio-temporal 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

The 2014 *DrosEU* dataset represents the most comprehensive spatio-temporal sampling of European *D. melanogaster* populations to date (Table 1). 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 File 2). 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 below).

DNA extraction, library preparation and sequencing

We extracted DNA from each sample after homogenization with bead beating and standard phenol/chloroform extraction. A detailed extraction protocol can be found in the Supplementary File 2. In preparation for sequencing, 500 ng of DNA from each sample

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

was sheared with a *Covaris* instrument (Duty cycle 10, intensity 5, cycles/burst 200, time 30). Library preparation was performed using *NEBNext Ultra DNA Lib Prep-24* and *NebNext Multiplex Oligos* for Illumina-24 following the manufacturer's instructions. Each sample was sequenced as a pool (Pool-Seq; Schlötterer *et al.* 2014), as paired-end fragments on a *Illumina NextSeq 500* sequencer at the Genomics Core Facility of Pompeu Fabra University. Samples were multiplexed in 5 batches of 10 samples, except for one batch of 8 samples (Table S1). Each multiplexed batch was sequenced on 4 lanes at ~50x raw coverage per sample. The read length was 151 bp, with a median insert size of 348 bp (range 209-454 bp). The data are available from NCBI Bioproject PRJNA388788.

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 retained only 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 range 29-35), with ~1.36% of bases lost after trimming. We used *bwa mem* (v. 0.7.15; Li 2013) with default parameters to map the trimmed reads. To avoid paralogous mapping, we mapped to a compound reference, consisting of the genomes of *D. melanogaster* (v.6.12) and 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*

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

faecalis (NC_004668.1), *Lactobacillus brevis* (NC_008497.1), and *Lactobacillus plantarum* (NC_004567.2). 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 indels with GATK (v3.4-46; McKenna *et al.* 2010).

After mapping, we filtered reads due to *D. simulans* contamination, using the method of Bastide *et al*. (2013). To do this, we used fixed differences between *D. simulans* and *D. melanogaster* to identify reads from *D. simulans*. For the nine samples that had a contamination level > 1% (range 1.2 - 8.7%; Table S1), we used custom software to remove reads that mapped preferentially to the *D. simulans* genome (Hu *et al.* 2013) using competitive mapping to references from both species. After applying our decontamination pipeline, contamination levels dropped below 0.4 % for all nine 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). Sequencing depth ranged from 34x to 115x for autosomes and from 17x to 59x for *X*-chromosomes (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 samples, we implemented quality control criteria for all libraries jointly to call SNPs. To call SNPs, we developed custom software (*PoolSNP*; see Supplementary File 2; available at doi: https://doi.org/10.5061/dryad.rj1gn54) using stringent heuristic parameters: (1) minimum coverage 10x for each sample, (2) maximum coverage < 95th coverage percentile for a given chromosome and sample (to avoid paralogous regions duplicated in the sample but not in the reference), (3) for each allele, a minimum read count > 20x and a minimum read frequency > 0.001, across all samples pooled. These parameters were

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

optimized based on simulated Pool-Seq data to maximize true positives and minimize false positives (Supplementary File 2). We also excluded SNPs (1) for which more than 20% of all samples did not fulfil 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 TEs based on the *D. melanogaster* TE library v.6.10. We annotated our final set of SNPs with *SNPeff* (v.4.2; Cingolani *et al.* 2012) using the Ensembl genome annotation version BDGP6.82.

Additional samples

We obtained genome sequences from African flies from the *Drosophila* Genome Nexus (DGN; http://www.johnpool.net/genomes.html; see Table S5 for SRA accession numbers). We used data from 14 individuals from Rwanda and 40 from Siavonga (Zambia). We mapped these data as described 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.

Genetic variation in Europe

We characterized patterns of genetic variation among the 48 samples for the five major chromosomal arms (*X*, *2L*, *2R*, *3L*, *3R*) by estimating π, Watterson's θ and Tajima's *D* (Watterson 1975; Nei 1987; Tajima 1989), using corrections for Pool-Seq data (Kofler *et al.* 2011). To perform these analyses for our set of SNPs, we re-implemented the methods of Kofler *et al*. (2011) in Python (PoolGen; doi: https://doi.org/10.5061/dryad.rj1gn54). To calculate unbiased window-wise estimates of parameters, we used an output file of our SNP calling pipeline (*PoolSNP*; doi: https://doi.org/10.5061/dryad.rj1gn54), which indicates for any given site in the reference, if it passed the filtering parameters used for SNP

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

calling. These data allow for the calculation of the effective window-size, which is the difference between the total window-size and the number of sites that did not pass the quality criteria. Using effective windows-sizes as the denominator for the calculation of window-wise averages yields unbiased average estimates. In contrast, dividing the summed statistics in a given window by the total window-size, which is common practice in most software tools, results in an underestimation of averaged parameters. Before calculating the estimators, we subsampled the data to an even coverage of 40x for autosomes and 20x for the *X*-chromosome, as Watterson's θ and Tajima's *D* are sensitive to coverage variation (Korneliussen *et al.* 2013). We calculated chromosome-wide averages of π, θ and Tajima's *D* for autosomes and *X* chromosomes using *R* (R Development Core Team 2009). We tested for correlations between these estimators and latitude, longitude, altitude, and season using a linear regression model: $y_i = Lat + Lon + Alt + Season + \varepsilon_i$, where y_i represents π, θ or *D*. We used *Lat*, *Lon* and *Alt* as continuous predictors (Table 1) and *Season* as a categorical factor with two levels, corresponding to collection dates before and after 1st September ('summer' and 'fall'), respectively, following Bergland *et al.* (2014) and Kapun *et al.* (2016a). To 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) for the residuals of the above models. For this analysis, we considered samples within 10° latitude / longitude to be neighbours, based on the pairwise geographical distances between collection locations. Whenever these tests revealed significant autocorrelations indicating non-independence, we repeated the above regressions using a spatial weights matrix based on nearest neighbours as described above to test for remaining spatial patterning in residuals as implemented in *spdep*. We also fitted models with run ID as a random factor using the *R* package *lme4* (v.1.1-14; see Supplementary File 2) to test for confounding

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

effects of variation in error rates among sequencing runs. As these models did not fit significantly better than simpler models, we excluded it from final analysis (see Supplementary File 2 and Table S3).

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, to investigate fine-scale deviations from neutral expectations, we also calculated Tajima's *D* in 50 kb sliding windows with a step size of 10 kb. We normalized diversity statistics using log-transformation and tested for correlations between π and recombination rate for 100 kb non-overlapping windows in *R* and plotted these data using the *ggplot2* (v.2.2.1., Wickham 2016). We used both fine-scale (Comeron *et al.* 2012) and broad-scale (Fiston-Lavier *et al.* 2010) estimates of recombination rate, after converting their coordinates to reference genome v 6.

To identify regions under selection, we used *Pool-hmm* to calculate the SFS (Site Frequency Spectrum) for each sample in the *pileup* format file with the following parameters *–prefix* (to assign a name to each sample), *-n* (number of chromosomes), *--only-spectrum* (for the SFS calculation), *--theta* 0.005 (default), and *-r* 100 (subsampling of 1/100 SNPs). We then split the *pileups* by chromosome and ran *Pool-hmm* with the following parameters: *--prefix*, *-n*, *-k* (per site transition probability between hidden states), *-s* (frequency spectrum file from previous step) and *-e sanger* (*Phred quality* = 33). For the 18 samples for which Tajima's *D* was very low, *Pool-hmm* identified the majority of the genome to be under selection; we thus removed those samples from our analysis. We used three different *k parameters* depending on the sample: $k=1e^{-10}$, $k=1e^{-30}$, and $k=1e^{-40}$ (Table S6A). For windows with significantly low Tajima's D in euchromatic regions, we

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

identified genes using *bedtools intersect* (v2.27.1) and the *D. melanogaster* v6.12 annotation file from Flybase (Thurmond et al 2019). For genes significant in all populations, we checked whether average Tajima's *D* was among the lowest 10% per chromosome. We tested for enrichment of involvement in particular biological processes using *DAVID* with default parameters (Huang et al 2009).

Genetic differentiation and population structure in European populations

To estimate genome-wide pairwise genetic differences, we used custom software to estimate SNP-wise F_{ST} using the approach of Weir and Cockerham (1984) for all pairwise combinations of samples. For each sample, we averaged pairwise F_{ST} between that sample and the other 47 samples and ranked the 48 population samples by overall differentiation.

We inferred demographic patterns by focusing on putatively neutrally evolving SNPs. For this, we used either 4-fold degenerate sites (defined using the genome sequences and the annotation features of the *D. melanogaster* reference genome version 6.12) or short introns (<60 bp; Haddrill *et al.* 2005; Singh *et al.* 2009; Parsch *et al.* 2010; Clemente & Vogl 2012; Lawrie *et al.* 2013). We also restricted our analyses to SNPs that were at least 1 Mb distant from major chromosomal inversions (see below) and those located in genomic regions with high recombination rates (r > 3cM/Mb; Comeron et al. 2012) to minimize the effects of linkage, which may confound analyses of neutral evolution. As the Sheffield (UK) population showed unusually high differentiation from other populations, we repeated the following analyses without the Sheffield sample. To assess isolation by distance (IBD), we averaged pairwise F_{ST} values across all neutral markers. We calculated geographic distance using the haversine formula (Green & Smart 1985), which takes the

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

spherical curvature of the planet into account. We tested for correlations between linearized genetic differentiation (Slatkin's distance: $F_{ST}/([1-F_{ST}])$ and $\log_{10}$-scaled geographic distance (Slatkin 1985) using Mantel tests implemented in *ade4* (v.1.7-8., Dray & Dufour 2007) with 1,000,000 iterations. In addition, we plotted the 5% smallest and 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.

We tested for population substructure using two different approaches. First, we performed principal component analysis (PCA) based on unscaled allele frequencies of the neutral marker SNPs, as suggested by Menozzi *et al*. (1978) and Novembre and Stephens (2008), using *LEA* (v. 1.2.0., Frichot *et al.* 2013). We focused on the first three principal components (PCs) and used *mclust* (v. 5.2., Fraley & Raftery 2012) to estimate the number of clusters *via* maximum likelihood and assigned population samples to clusters *via k*-means. In addition, we examined the first three PCs for correlations with latitude, longitude, altitude, and season using general linear models and tested for spatial autocorrelation as above. A Bonferroni-corrected α threshold ($\alpha' = 0.05/3 = 0.017$) was used to correct for multiple testing.

In a second, complementary approach, we inferred population delineation using model-based clustering as implemented in *ConStruct* (v.1.0.2; Bradburd *et al.* 2018). In contrast to most clustering-based methods, *ConStruct* incorporates continuous isolation by distance to avoid inflating estimates of the number of clusters and allows estimating admixture among populations. We ran spatial models with three MCMC chains per run and 10,000 iterations and compared the goodness of fit for models incorporating 1 to 10 spatial layers by cross-validation.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mitochondrial DNA

To obtain consensus mitochondrial sequences for each of the 48 European populations, we aligned reads from individual FASTQ files and replaced minor variants with the major variant using *Coral* (Salmela & Schröder 2011). This method prevents ambiguities from interfering with the assembly process. We assembled a genome for each population from the modified FASTQ files using *SPAdes* with standard parameters and *k*-mers of size 21, 33, 55, and 77 (Bankevich *et al.* 2012). 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 higher than 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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

region was thus analysed 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 transposable element (TE) abundance in each sample, we assembled and quantified repeats from unassembled sequenced reads using *dnaPipeTE* (v.1.2., Goubert *et al.* 2015). Only the left read of each pair were used. As the vast majority of high-quality trimmed reads were longer than 135 bp, we discarded reads shorter than this 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 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* (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). To assess the constancy of the estimates, we repeated this process with three iterations per sample, as recommended by the program guidelines.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We further estimated frequencies of 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 2,234 insertions in the reference genome, which appear to be mostly fixed (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. Because the mapper used by *T-lex2* to detect the presence of insertions (presence module) only accepts reads ≤127 bp, we trimmed reads longer than 100 bp into two equally sized fragments using *Trimmomatic* (v. 0.35; Bolger *et al.* 2014) with the CROP and HEADCROP parameters.

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 S1), in order to avoid non-uniquely 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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recombination rates from Comeron *et al*. (2012) and from Fiston-Lavier *et al.* (2010) in non-overlapping 100 kb windows and assigned to each TE insertion the recombination rate of the corresponding window.

To test for spatio-temporal variation of TE insertions, we excluded TEs with an interquartile range (IQR) < 10. We tested the population frequencies of the remaining 141 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 further tested if significant correlations with either of the predictor variables deviated from expectations under neutral evolution. To this end, we repeated the ANCOVA analyses on 8,727 presumably neutrally evolving 4-fold degenerate sites that we described previously in the demographic analyses. Based on *F*-ratios obtained from the ANCOVA models for each neutral SNP and predictor, we built empirical density functions and calculated empirical *p*-values for each TE by integrating over the area of the curve that is delineated by the *F*-value specific for the given TE and the maximum *F*-ratio in the neutral dataset.

We also tested for residual spatio-temporal autocorrelations in TE insertion frequencies, 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. 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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Bonferroni-corrected the α threshold (α'= 0.05/7 = 0.007) to account for multiple testing, accounted for residual spatio-temporal autocorrelations and tested if *F*-ratios of the ANCOVAs deviated from neutral expectations as explained above.

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 analysed 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*'

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(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), at the computing facilities of the CC LBBE/PRABI in Lyon (France) and at the bwUniCluster of the state of Baden-Württemberg (bwHPC).

Additional information

Funding

Funder	Grant reference number	Author

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

University of Freiburg Research Innovation Fund 2014. Deutsche Forschungsgemeinschaft	STA1154/4-1Project 408908608	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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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-2014CoG-647900	Josefa González
Secretaria d'Universitats i Recerca. Dept Economia i Coneixement. Generalitat de Catalunya	GRC 2017 SGR 880	Josefa González
Ministerio de Economia y Competitividad	FEDER BFU2014-57779-P	Josefa González

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,

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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; María Bogaerts-Márques, Alejandro Sánchez-Gracia, Formal analysis, Writing-review & editing, Investigation, Validation, 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;

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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
María Bogaerts-Márquez	0000-0001-9107-984X
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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Alejandro Sánchez-Gracia	0000-0003-4543-4577
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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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**, 1-12.

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*. *Cell,* **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.

Andolfatto P (2001) Contrasting Patterns of *X*-Linked and Autosomal Nucleotide Variation in *Drosophila melanogaster* and *Drosophila simulans*. *Molecular Biology and Evolution*, **18**, 279–290.

Arguello JR, Laurent S, Clark AG. 2019. Demographic History of the Human Commensal *Drosophila melanogaster*. *Genome Biology and Evolution* **11**:844–854.

Aulard S, David JR, Lemeunier F (2002) Chromosomal inversion polymorphism in Afrotropical populations of *Drosophila melanogaster*. *Genetic Research*, **79**, 49–63.

Auton A, Abecasis GR, Altshuler DM (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Bankevich A, Nurk S, Antipov D *et al.* (2012) *SPAdes, a New Genome Assembly Algorithm and Its Applications to Single-cell Sequencing (7th Annual SFAF Meeting, 2012).* Mary Ann Liebert Inc.

Barata A, Santos SC, Malfeito-Ferreira M, Loureiro V (2012) New insights into the ecological interaction between grape berry microorganisms and *Drosophila* flies during the development of sour rot. *Microbial Ecology*, **64**, 416–430.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bartolomé C, Maside X, Charlesworth B (2002) On the Abundance and Distribution of Transposable Elements in the Genome of *Drosophila melanogaster*. *Molecular Biology and Evolution*, **19**, 926–937.

Bastide H, Betancourt A, Nolte V *et al.* (2013) A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *PLoS Genetics*, **9**, e1003534.

Baudry E, Viginier B, Veuille M (2004) Non-African populations of *Drosophila melanogaster* have a unique origin. *Molecular Biology and Evolution*, **21**, 1482–1491.

Becher PG, Flick G, Rozpędowska E *et al.* (2012) Yeast, not fruit volatiles mediate *Drosophila melanogaster* attraction, oviposition and development. *Functional Ecology*, **26**, 822–828.

Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**, 519–520.

Begun DJ, Aquadro CF (1993) African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature*, **365**, 548–550.

Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in *Drosophila simulans*. *PLoS Biology*, **5**, e310.

Behrman EL, Howick VM, Kapun M *et al.* (2018) Rapid seasonal evolution in innate immunity of wild *Drosophila melanogaster*. *Proceedings of the Royal Society of London B*, **285**, 20172599.

Beisswanger S, Stephan W, De Lorenzo D (2006) Evidence for a Selective Sweep in the *wapl* Region of *Drosophila melanogaster*. *Genetics*, **172**, 265–274.

Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA (2014) Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in *Drosophila*. *PLoS Genetics*, **10**, e1004775.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bergland AO, Tobler R, González J, Schmidt P, Petrov D (2016) Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila* melanogaster. *Molecular Ecology*, **25**, 1157–1174.

Betancourt AJ, Kim Y, Orr HA (2004) A pseudohitchhiking model of X vs. autosomal diversity. *Genetics*, **168**, 2261–2269.

Betancourt AJ, Welch JJ, Charlesworth B (2009) Reduced effectiveness of selection caused by a lack of recombination. *Current Biology*, **19**, 655–660.

Bilder D, Irvine KD (2017) Taking Stock of the *Drosophila* Research Ecosystem. *Genetics* **206**, 1227–1236

Bivand R, Piras G (2015) Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, **63**, 1–36.

Black WC IV, Black WC IV, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology*, **46,** 441–469

Blumenstiel JP, Chen X, He M, Bergman CM (2014) An Age-of-Allele Test of Neutrality for Transposable Element Insertions. *Genetics*, **196**, 523–538.

Boitard S, Schlötterer C, Nolte V, Pandey RV, Futschik A (2012) Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. *Molecular Biology and Evolution*, **29**, 2177–2186.

Boitard S, Kofler R, Françoise P, Robelin D, Schlötterer C, Futschik A (2013) Pool-hmm: a Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. *Mol Ecol Resour,* **13**, 337–340.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Boussy IA, Itoh M, Rand D, Woodruff RC (1998) Origin and decay of the P element-associated latitudinal cline in Australian *Drosophila melanogaster*. *Genetica*, **104**, 45–57.

Božičević V, Hutter S, Stephan W, Wollstein A (2016) Population genetic evidence for cold adaptation in European *Drosophila melanogaster* populations. *Molecular Ecology*, **25**, 1175–1191.

Bradburd GS, Coop GM, Ralph PL (2018) Inferring Continuous and Discrete Population Genetic Structure Across Space. *Genetics* **210**, 33–52.

Braithwaite DP, Keegan KP *matR: Metagenomics Analysis Tools for R*. https://CRAN.R-project.org/package=matR.

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, **12**, 59–60.

Buser CC, Newcomb RD, Gaskett AC, Goddard MR (2014) Niche construction initiates the evolution of mutualistic interactions. *Ecology Letters*, **17**, 1257–1264.

Bushnell B (2016) *BBMap short read aligner*. URL http://sourceforge. net/projects/bbmap.

Caracristi G, Schlötterer C (2003) Genetic Differentiation Between American and European *Drosophila melanogaster* Populations Could Be Attributed to Admixture of African Alleles. *Molecular Biology and Evolution*, **20**, 792–799.

Casillas S, Barbadilla A (2017) Molecular Population Genetics. *Genetics*, **205**, 1003–1035.

Catania F, Kauer MO, Daborn PJ *et al.* (2004) World-wide survey of an Accord insertion and its association with DDT resistance in *Drosophila melanogaster. Molecular Ecology*, **13**, 2491–2504.

Cavalli-Sforza LL (1966) Population Structure and Human Evolution. *Proceedings of the Royal Society of London B*, **164**, 362–379.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Chandler JA, James PM (2013) Discovery of trypanosomatid parasites in globally distributed *Drosophila* species. *PLoS ONE*, **8**, e61937.

Chandler JA, Eisen JA, Kopp A (2012) Yeast communities of diverse *Drosophila* species: comparison of two symbiont groups in the same hosts. *Applied and Environmental Microbiology*, **78**, 7327–7336.

Chandler JA, Lang JM, Bhatnagar S, Eisen JA, Kopp A (2011) Bacterial communities of diverse *Drosophila* species: ecological context of a host-microbe model system. *PLoS Genetics*, **7**, e1002272.

Charlesworth B (2001) The effect of life-history and mode of inheritance on neutral genetic variability. *Genetical Research* **77**, 153–166.

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, **371**, 215–220.

Cheng C, White BJ, Kamdem C *et al.* (2012) Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics*, **190**, 1417–1432.

Cingolani P, Platts A, Wang LL *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly (Austin)*, **6**, 80–92.

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657–1659.

Clemente F, Vogl C (2012) Unconstrained evolution in short introns? – An analysis of genome-wide polymorphism and divergence data from *Drosophila*. *Journal of Evolutionary Biology*, **25**, 1975–1990.

Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1002905.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cooper BS, Burrus CR, Ji C, Hahn MW, Montooth KL (2015) Similar Efficacies of Selection Shape Mitochondrial and Nuclear Genes in Both *Drosophila melanogaster* and *Homo sapiens*. *G3*, **5**, 2165–2176.

Corbett-Detig RB, Hartl DL (2012) Population Genomics of Inversion Polymorphisms in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1003056.

Cridland JM, Macdonald SJ, Long AD, Thornton KR (2013) Abundance and distribution of transposable elements in two *Drosophila* QTL mapping resources. *Molecular Biology and Evolution*, **30**, 2311–2327.

Daborn PJ, Yen JL, Bogwitz MR *et al.* (2002) A single p450 allele associated with insecticide resistance in *Drosophila*. *Science*, **297**, 2253–2256.

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

de Jong G, Bochdanovits Z (2003) Latitudinal clines in *Drosophila melanogaster*: body size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway. *Journal of Genetics*, **82**, 207–223.

Dieringer D, Nolte V, Schlötterer C (2005) Population structure in African *Drosophila melanogaster* revealed by microsatellite analysis. *Molecular Ecology*, **14**, 563–573.

Dobzhansky T (1970) *Genetics of the Evolutionary Process*. Columbia University Press.

Dray S, Dufour A-B (2007) The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, **22**. 1–20

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Durmaz E, Benson C, Kapun M, Schmidt P, Flatt T (2018) An Inversion Supergene in *Drosophila* Underpins Latitudinal Clines in Survival Traits. *Journal of Evolutionary Biology*, in press.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

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

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

Fabian DK, Lack JB, Mathur V *et al.* (2015) Spatially varying selection shapes life history clines among populations of *Drosophila melanogaster* from sub-Saharan Africa. *Journal of Evolutionary Biology*, **28**, 826–840.

Fiston-Lavier A-S, Barrón MG, Petrov DA, González J (2015) T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Research*, **43**, e22–e22.

Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA (2010) *Drosophila melanogaster* recombination rate calculator. *Gene*, **463**, 18–20.

Fraley C, Raftery AE (2012) *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. https://cran.r-project.org/web/packages/mclust

Francalacci P, Sanna D (2008) History and geography of human Y-chromosome in Europe: a SNP perspective. *Journal of Anthropological Sciences*, **86**, 59–89.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Futschik A (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

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

Grenier JK, Arguello JR, Moreira MC *et al.* (2015) Global Diversity Lines-A Five-Continent Reference Panel of Sequenced *Drosophila melanogaster* Strains. *G3*, **5**, 593–603.

Guirao-Rico S, González J (2019) Evolutionary insights from large scale resequencing datasets in Drosophila melanogaster. *Current Opinion in Insect Science*, Insect genomics • Development and regulation **31**, 70–76.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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*, **5**, 2137–2154.

Harpur BA, Kent CF, Molodtsova D *et al.* (2014) Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 2614–2619.

Haselkorn TS, Markow TA, Moran NA (2009) Multiple introductions of the *Spiroplasma* bacterial endosymbiont into *Drosophila*. *Molecular Ecology*, **18**, 1294–1305.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genetics*, **6**, e1000862.

Hu TT, Eisen MB, Thornton KR, Andolfatto P (2013) A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Research*, **23**, 89–98.

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols,* **4**, 44–57.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W (2007) Distinctly Different Sex Ratios in African and European Populations of *Drosophila melanogaster* Inferred From Chromosomewide Single Nucleotide Polymorphism Data. *Genetics*, **177**, 469–480.

Jorde LB, Watkins WS, Bamshad MJ (2001) Population genomics: a bridge from evolutionary history to genetic medicine. *Human Molecular Genetics*, **10**, 2199–2207.

Kao JY, Zubair A, Salomon MP, Nuzhdin SV, Campo D (2015) Population genomic analysis uncovers African and European admixture in *Drosophila melanogaster* populations from the south-eastern United States and Caribbean Islands. *Molecular Ecology*, **24**, 1499–1509.

Kapopoulou A, Kapun M, Pavlidis P, *et al*. (2018a) Early split between African and European populations of *Drosophila melanogaster*. Preprint at *bioRxiv*, doi: https://doi.org/10.1101/340422

Kapopoulou A, Pfeifer S, Jensen J, Laurent S (2018b). The demographic history of African *Drosophila melanogaster*. Preprint at *bioRxiv*, doi:10.1101/340406

Kapitonov VV, Jurka J (2003) Molecular Paleontology of Transposable Elements in the *Drosophila melanogaster* Genome. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6569–6574.

Kapun M, Flatt T (2019) The adaptive significance of chromosomal inversion polymorphisms in *Drosophila melanogaster*. *Molecular Ecology*, **28**, 1263-1282

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Kapun M, Fabian DK, Goudet J, Flatt T (2016a) Genomic Evidence for Adaptive Inversion Clines in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **33**, 1317–1336.

Kapun M, Schmidt C, Durmaz E, Schmidt PS, Flatt T (2016b) Parallel effects of the inversion *In(3R)Payne* on body size across the North American and Australian clines in *Drosophila melanogaster. Journal of Evolutionary Biology*, **29**, 1059–1072.

Kapun M, van Schalkwyk H, McAllister B, Flatt T, Schlötterer C (2014) Inference of chromosomal inversion dynamics from Pool-Seq data in natural and laboratory populations of *Drosophila melanogaster. Molecular Ecology*, **23**, 1813–1827.

Kassis JA, Kennison JA, Tamkun JW (2017) Polycomb and Trithorax Group Genes in *Drosophila*. *Genetics*, **206**, 1699–1725.

Kauer M, Zangerl B, Dieringer D, Schlötterer C (2002) Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. *Genetics*, **160**, 247–256.

Keller A (2007) *Drosophila melanogaster*'s history as a human commensal. *Current Biology*, **17**, R77–R81.

Kennington JW, Partridge L, Hoffmann AA (2006) Patterns of Diversity and Linkage Disequilibrium Within the Cosmopolitan Inversion *In(3R)Payne* in *Drosophila melanogaster* Are Indicative of Coadaptation. *Genetics*, **172**, 1655 – 1663.

Kimura M (1984) *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Knibb WR, Oakeshott JG, Gibson JB (1981) Chromosome Inversion Polymorphisms in *Drosophila melanogaster*. I. Latitudinal Clines and Associations between Inversions in Australasian Populations. *Genetics*, **98**, 833–847.

Kofler R, Betancourt AJ, Schlötterer C (2012) Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster. PLoS Genetics*, **8**, e1002487.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Kofler R, Orozco-terWengel P, De Maio N *et al.* (2011) PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS ONE*, **6**, e15925.

Kolaczkowski B, Hupalo DN, Kern AD (2011a) Recurrent adaptation in RNA interference genes across the *Drosophila* phylogeny. *Molecular Biology and Evolution*, **28**, 1033–1042.

Kolaczkowski B, Kern AD, Holloway AK, Begun DJ (2011b) Genomic Differentiation Between Temperate and Tropical Australian Populations of *Drosophila melanogaster*. *Genetics*, **187**, 245–260.

Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R (2013) Calculation of Tajima's *D* and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, **14**, 289.

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

Kriesner P, Conner WR, Weeks AR, Turelli M, Hoffmann AA (2016) Persistence of a *Wolbachia* infection frequency cline in *Drosophila melanogaster* and the possible role of reproductive dormancy. *Evolution*, **70**, 979–997.

Kühn I, Dormann CF (2012) Less than eight (and a half) misconceptions of spatial analysis. *Journal of Biogeography*, **39**, 995–998.

Lachaise D, Cariou M-L, David JR *et al.* (1988) Historical Biogeography of the *Drosophila melanogaster* Species Subgroup. In Hecht MK, Wallace B, Prance GT (Eds.) *Evolutionary Biology* (pp. 159–225) Boston: Springer.

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE (2016) A Thousand Fly Genomes: An Expanded *Drosophila* Genome Nexus. *Molecular Biology and Evolution*, **33**, 3308–3313.

Lang DT (2014) *RJSONIO: Serialize R objects to JSON, JavaScript Object Notation*. https://CRAN.R-project.org/package=RJSONIO.

Langley CH, Stevens K, Cardeno C *et al.* (2012) Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics*, **192**, 533–598.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

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

Lawrie DS, Messer PW, Hershberg R, Petrov DA (2013) Strong Purifying Selection at Synonymous Sites in *D. melanogaster*. *PLoS Genetics,* **9**, e1003527.

Lerat E, Goubert C, Guirao-Rico S, Merenciano M, Dufour A-B, Vieira C, González J (2019) Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Molecular Ecology*, **28**,1506–1522.

Lemeunier F, Aulard S (1992). Inversion polymorphism in *Drosophila melanogaster.* In: Krimbas CB, & Powell JR (Eds.), *Drosophila Inversion Polymorphism* (pp. 339–405), New York: CRC Press.

Lewontin RC (1974) *The Genetic Basis of Evolutionary Change*. Columbia University Press.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv.org*, 1303.3997

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, **18**, 1851–1858.

Lian T, Li D, Tan X, Che T, Xu Z, Fan X, Wu N, Zhang L, Gaur U, Sun B, Yang M (2018) Genetic diversity and natural selection in wild fruit flies revealed by whole-genome resequencing. *Genomics,* **110**, 304–309.

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

Lyne R, Smith R, Rutherford K *et al.* (2007) FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. *Genome Biology*, **8**, R129.

Machado HE, Bergland AO, O'Brien KR *et al.* (2016) Comparative population genomics of latitudinal variation in *Drosophila simulans* and *Drosophila melanogaster*. *Molecular Ecology*, **25**, 723–740.

Machado H, Bergland AO, Taylor R *et al.* (2018) Broad geographic sampling reveals predictable and pervasive seasonal adaptation in *Drosophila*. Preprint at *bioRxiv*, doi: https://doi.org/10.1101/337543.

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

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.

Martino ME, Ma D, Leulier F (2017) Microbial influence on *Drosophila* biology. *Current Opinion in Microbiology*, **38**, 165–170.

Mateo L, Rech GE, González J (2018) Genome-wide patterns of local adaptation in *Drosophila melanogaster*: adding intra European variability to the map. Preprint at *bioRxiv*, doi: https://doi.org/10.1101/269332

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Matthias P, Yoshida M, Khochbin S (2008) HDAC6 a new cellular stress surveillance factor. *Cell Cycle*, **7**, 7–10.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, **351**, 652–654.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analysing 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. *Trends in Ecology & Evolution*, **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.

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

Nielsen R, Akey JM, Jakobsson M *et al.* (2017) Tracing the peopling of the world through genomics. *Nature*, **541**, 302-310.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

Novembre J, Johnson T, Bryc K, *et al.* (2008) Genes mirror geography within Europe. *Nature*, **456**, 98-101.

Nunes MDS, Neumeier H, Schlötterer C (2008) Contrasting patterns of natural variation in global *Drosophila melanogaster* populations. *Molecular Ecology*, **17**, 4470–4479.

Okonechnikov K, Conesa A, García-Alcalde F (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, **32**, 292–294.

Palmer WH, Medd NC, Beard PM, Obbard DJ (2018) Isolation of a natural DNA virus of Drosophila melanogaster, and characterisation of host resistance and immune responses. *PLOS Pathogens*, **14**, e1007050

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P (2010) On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Molecular Biology and Evolution*, **27**, 1226–1234.

Peel MC, Finlayson BL, McMahon TA (2007) Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, **11**, 1633–1644.

Petrov DA, Fiston-Lavier AS, Lipatov M, Lenkov K, González J (2011) Population Genomics of Transposable Elements in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **28**, 1633–1644.

Pool JE (2015) The Mosaic Ancestry of the *Drosophila* Genetic Reference Panel and the *D. melanogaster* Reference Genome Reveals a Network of Epistatic Fitness Interactions. *Molecular Biology and Evolution*, **32**, 3236–3251.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pool JE, Nielsen R (2007) Population size changes reshape genomic patterns of diversity. *Evolution*, **61**, 3001–3006.

Pool JE, Braun DT, Lack JB (2016) Parallel Evolution of Cold Tolerance Within *Drosophila melanogaster*. *Molecular Biology and Evolution*, **34,** 349–360.

Pool JE, Corbett-Detig RB, Sugino RP *et al.* (2012) Population Genomics of Sub-Saharan *Drosophila melanogaster*: African Diversity and Non-African Admixture. *PLoS Genetics*, **8**, e1003080.

Powell JR (1997) *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press.

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. *R-project.org*.

Rako L, Anderson AR, Sgrò CM, Stocker AJ, Hoffmann AA (2006) The association between inversion *In(3R)Payne* and clinally varying traits in *Drosophila melanogaster. Genetica*, **128**, 373–384.

Rane RV, Rako L, Kapun M, LEE SF (2015) Genomic evidence for role of inversion *3RP* of *Drosophila melanogaster* in facilitating climate change adaptation. *Molecular Ecology*, **24**, 2423–2432.

Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL, Horváth V, Fiston-Lavier A-S, Luyten I, Venkataram S, Quesneville H, Petrov DA, González J (2019) Stress response, behavior, and development are shaped by transposable element-induced mutations in *Drosophila*. *PLOS Genetics,* **15**, e1007900.

Reinhardt JA, Kolaczkowski B, Jones CD, Begun DJ, Kern AD (2014) Parallel Geographic Variation in *Drosophila melanogaster*. *Genetics*, **197**, 361–373.

Richardson MF, Weinert LA, Welch JJ *et al.* (2012) Population Genomics of the *Wolbachia* Endosymbiont in *Drosophila melanogaster. PLoS Genetics*, **8**, e1003129.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Rogers RL, Hartl DL (2012) Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **29**, 517–529.

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC *et al.* (2017) DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. *Molecular Biology and Evolution*, **34**, 3299–3302.

Salmela L, Schröder J (2011) Correcting errors in short reads by multiple alignments. *Bioinformatics*, **27**, 1455–1461.

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

Schlötterer C, Neumeier H, Sousa C, Nolte V (2006) Highly structured Asian *Drosophila melanogaster* populations: a new tool for hitchhiking mapping? *Genetics*, **172**, 287–292.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**, 749–763.

Schmidt JM, Good RT, Appleton B *et al.* (2010) Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genetics*, **6**, e1000998.

Schmidt PS, Paaby AB (2008) Reproductive Diapause and Life-History Clines in North American Populations of *Drosophila melanogaster*. *Evolution*, **62**, 1204–1215.

Schmidt PS, Zhu CT, Das J *et al.* (2008) An amino acid polymorphism in the *couch potato* gene forms the basis for climatic adaptation in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 16207–16211.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive Natural Selection in the *Drosophila* Genome? *PLoS Genetics*, **5**, e1000495.

Singh ND, Arndt PF, Clark AG, Aquadro CF (2009) Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila*. *Molecular Biology and Evolution*, **26**, 1591–1605.

Slatkin M. 1985. Gene Flow in Natural Populations. *Annual Review of Ecology and Systematics* **16**:393–430.

Staubach F, Baines JF, Künzel S, Bik EM, Petrov DA (2013) Host species and environmental effects on bacterial communities associated with *Drosophila* in the laboratory and in the natural environment. *PLoS ONE*, **8**, e70749.

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

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, Kaufman TC, Calvi BR, FlyBase Consortium (2019) FlyBase 2.0: the next generation. *Nucleic Acids Res,* **47**, D759–D765.

Trinder M, Daisley BA, Dube JS, Reid G (2017) *Drosophila melanogaster* as a High-Throughput Model for Host-Microbiota Interactions. *Frontiers in Microbiology*, **8**, 751.

Turner TL, Levine MT, Eckert ML, Begun DJ (2008) Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. *Genetics*, **179**, 455–473.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Umina PA, Weeks AR, Kearney MR, McKechnie SW, Hoffmann AA (2005) A rapid shift in a classic clinal pattern in *Drosophila* reflecting climate change. *Science*, **308**, 691–693.

Unckless RL (2011) A DNA virus of *Drosophila*. *PLoS ONE*, **6**, e26564.

Walters AM, Matthews MK, Hughes R, Malcolm Jaanna, Rudman S, Newell PD, Douglas AE, Schmidt PS, Chaston JM (2018) The microbiota influences the Drosophila melanogaster life history strategy. bioRxiv. 471540

Wang Y, Kapun M, Waidele L, Kuenzel S, Bergland AO, Staubach F (2019) Continent-wide structure of bacterial microbiomes of European Drosophila melanogaster suggests host-control. bioRxiv. 527531

Wang Y, Staubach F (2018); Individual variation of natural *D.melanogaster*-associated bacterial communities, *FEMS Microbiology Letters*, **365**, fny017

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Webster CL, Longdon B, Lewis SH, Obbard DJ (2016) Twenty-Five New Viruses Associated with the Drosophilidae (Diptera). *Evolutionary Bioinformatics Online*, **12**, 13–25.

Webster CL, Waldron FM, Robertson S *et al.* (2015) The Discovery, Distribution, and Evolution of Viruses Associated with *Drosophila melanogaster. PLoS Biology*, **13**, e1002210.

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

Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate biology. *Nature Reviews Microbiology*, **6**, 741–751.

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

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

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

Yukilevich R, Turner TL, Aoki F, Nuzhdin SV, True JR (2010) Patterns and processes of genome-wide divergence between North American and African *Drosophila melanogaster*. *Genetics*, **186**, 219–239.

Zanini F, Brodin J, Thebo L *et al.* (2015) Population genomics of intrapatient HIV-1 evolution. *eLife*, **4**, e11282.

Tables

Table 1. Sample information for all populations in the *DrosEU* dataset. 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 Table S1.

				Number						
ID	**Country**	**Location**	**Coll. Date**	**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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/313759; this version posted September 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Files

Supplementary File 1. Supplementary Tables.

This file contains the 13 supplementary tables mention in the text.

Supplementary File 2. Additional methods.

This file contains the additional methods mention in the text.