Genomic analysis of European *Drosophila melanogaster* populations on a dense spatial scale reveals longitudinal population structure and continent-wide selection

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[24] Instituto Gulbenkian de Ciência, Oeiras, Portugal

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

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

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

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

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

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

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

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

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

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

[*] These authors contributed equally to this work

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

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

Abstract

Genetic variation is the fuel of evolution. However, analyzing dynamics of evolutionary change in natural populations is challenging, genome sequencing of entire populations remains costly and comprehensive sample collection logistically challenging. To tackle this issue and to define relevant spatial and temporal scales of variation for a population genetic model system, the fruit fly *Drosophila melanogaster*, we have founded the European *Drosophila* Population Genomics Consortium (*DrosEU*). Our principal objective is to employ the strengths of this collaborative consortium to extensively sample and sequence natural populations on a continent-wide scale and across distinct timescales. Here we present the first analysis of the first *DrosEU* pool-sequencing dataset, consisting of 48 population samples collected across the European continent in 2014. The analysis of this comprehensive dataset uncovers novel patterns of variation at multiple levels: genome-wide neutral SNPs, mtDNA haplotypes, inversions and TEs that exhibit previously cryptic longitudinal population structure across the European continent; signatures of selective sweeps shared among the majority of European populations; presumably adaptive clines in inversions; and geographic variation in TEs. Additionally, we document highly variable microbiota among European fruit fly populations and identify several new *Drosophila* viruses. Our study reveals novel aspects of the population biology of *D. melanogaster* and illustrates the power of extensive sampling and pooled sequencing of natural populations on a continent-wide scale.

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

Introduction

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

However, large-scale sampling and genome sequencing of entire populations remains largely prohibitive in terms of sequencing costs and labor-intensive sample collection, limiting the number of populations that can be analyzed. Evolution is a highly dynamic process across a variety of spatial scales in many taxa; thus, to generate a comprehensive

context for population genomic analyses, it is essential to define the appropriate spatial scales of analysis, from meters to thousands of kilometers (Levins 1968; Endler 1977; Richardson *et al.* 2014). Furthermore, one-time sampling of natural populations provides only a static view of patterns of genetic variation. Allele frequency changes can be highly dynamic even across very short timescales (e.g., Umina *et al.* 2005; Bergland *et al.* 2014; Behrman *et al.* 2018), and theoretical work suggests that such temporal dynamics may be an important yet largely understudied mechanism by which genetic variation is maintained (Wittmann *et al.* 2017). It is thus essential to define the relevant spatio-temporal scales for sampling and population genomic analyses accordingly.

To generate a population genomic framework that can deliver appropriate high-resolution sampling and to provide a unique resource to the research community, we formed the European *Drosophila* Population Genomics Consortium (*DrosEU*; https://droseu.net). Our primary objective is to utilize the strengths of this consortium to extensively sample and sequence European populations of *Drosophila melanogaster* on a continent-wide scale and across distinct timescales

D. melanogaster offers several advantages for such an concerted sampling and analysis effort: a relatively small genome, a broad geographic range, a multivoltine life history that allows sampling across generations over short timescales, ease of sampling natural populations using standardized techniques, an extensive research community and a well-developed context for population genomic analysis (Powell 1997; Keller 2007; Hales *et al.* 2015). The species is native to sub-Saharan Africa and has subsequently expanded its range into novel habitats in Europe over the last 10,000-15,000 years and in North America and Australia in the last several hundred years (e.g., Lachaise *et al.* 1988; David

& Capy 1988; Keller 2007). On both the North American and Australian continents, the prevalence of latitudinal clines in frequencies of alleles (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 (Mettler *et al.* 1977; Voelker *et al.* 1978; Knibb *et al.* 1981; Knibb 1982; 1986; Anderson *et al.* 1991; Rako *et al.* 2006; Kapun *et al.* 2014; Rane *et al.* 2015; Kapun *et al.* 2016a) and 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; Flatt *et al.* 2013; Adrion *et al.* 2015 and references therein; Kapun *et al.* 2016b; Behrman *et al.* 2018) have been interpreted to result from local adaptation to environmental factors that co-vary with latitude or as the legacy of an out-of-Africa dispersal history. However, sampling across these latitudinal gradients has not been replicated outside of a single transect on the east coasts of both continents. The observed latitudinal clines on the east coasts of North America and Australia may have been generated, at least in part, by demography and differential colonization histories of populations at high and low latitudes (Bergland *et al.* 2016). In North America, for example, temperate populations appear to be largely of European origin, whereas low latitude populations show evidence of greater admixture from ancestral African populations and the Caribbean (Caracristi & Schlötterer 2003; Yukilevich & True 2008a; b; Duchen *et al.* 2013; Kao *et al.* 2015; Bergland *et al.* 2016). More intensive sampling and analysis of both African as well as European populations is essential to disentangling the relative importance of local adaptation versus colonization history and demography in generating the clinal patterns that have been widely observed. While there has been a great deal of progress in the analysis of ancestral African populations (e.g., Begun & Aquadro 1993; Corbett-Detig & Hartl 2012; Pool *et al.* 2012; Fabian *et al.* 2015; Lack *et al.* 2015; 2016),

176 Europe remains largely uncharacterized at the population genomic level (Božičević *et al.*
177 2016; Pool *et al.* 2016; Mateo *et al.* 2018).
178
179 Here, we present the first analysis of the *DrosEU* pool-sequencing data from a set of 48
180 European population samples collected in 2014. We examine the 2014 *DrosEU* data at
181 three levels: (1) patterns of variation at ~5.5 x 10^6 single-nucleotide polymorphisms
182 (SNPs) in the nuclear and mitochondrial (mtDNA) genomes; (2) variation in copy number
183 of transposable elements (TEs); (3) cosmopolitan chromosomal inversions previously
184 associated with climate adaptation; and (4) large amounts of variation among populations
185 in microbiota, including endosymbionts, bacteria, and viruses. We find that European
186 populations of *D. melanogaster* exhibit novel patterns of variation at all levels investigated:
187 neutral SNPs in the nuclear genome and mtDNA haplotypes that reveal previously
188 unknown longitudinal population structure; genomic regions consistent with selective
189 sweeps that indicate selection on a continent-wide scale; new evidence for inversion clines
190 in Europe; and spatio-temporal variation in TEs frequencies. We also identify four new
191 DNA viruses and for the first time assemble the complete genome of a fifth. These novel
192 features are revealed by the comprehensive magnitude of our coordinated sampling, thus
193 demonstrating the utility of this approach.
194
195 Together with other genomic datasets for *D. melanogaster* (e.g., DGRP, DPGP, DGN;
196 reviewed in Casillas & Barbadilla 2017) our data provide a rich and powerful community
197 resource for studies in molecular population genetics. Importantly, the *DrosEU* data
198 represent the first comprehensive characterization of genetic variation in *D. melanogaster*
199 on the European continent.
200

Results

As part of the *DrosEU* effort, we collected and sequenced 48 population samples of *D. melanogaster* from 32 geographical locations across Europe in 2014 (Table 1; Figure 2 and Figure 3A). While our analyses focus on spatial patterns, thirteen of the 32 locations were sampled repeatedly (at least twice) during the year, allowing a first, crude analysis of seasonal changes in allele frequencies between summer and fall on a genome-wide level (Figure 2). All 48 samples were sequenced to high coverage, with a mean coverage per population of >50x (Table S1 and Figure S1). Using this high-quality dataset, we performed the first comprehensive, continent-wide population genomic analysis of European *D. melanogaster* (Figure 2). In addition to nuclear SNPs, we also analyzed mtDNA, TE insertions, chromosomal inversion polymorphisms, and the *Drosophila*-associated microbiome (Figure 3).

Most SNPs are widespread throughout Europe

We identified a total of 5,558,241 “high confidence” SNPs with frequencies > 0.1% across all 48 samples (Figure 3B, Table S1 and S2). Of these, 17% (941,080) were shared among all samples, whereas 62% were polymorphic in fewer than 50% of the samples (Figure 4A). Due to our filtering scheme, SNPs that are private or nearly private to a sample will be recovered only if they are at a substantial frequency in that sample (~5%). In fact, only a small proportion of SNPs (1% = 3,645) was found in fewer than 10% of the samples, and only 0.004% (210) were specific to a single sample (Figure 4A). To avoid an excess contribution of SNPs from populations with multiple (seasonal) sampling, we repeated the analysis by considering only the earliest (Figure S2A) or latest (Figure S2B) sample from populations with seasonal data. We observed very similar patterns across the three analyses: (i) a very small number of sample-specific, private SNPs (210, 527 and

455, respectively), (ii) a majority of SNPs shared among 20% to 40% of the samples (53%, 52% and 52%, respectively), and (iii) a substantial proportion shared among all samples (17%, 20% and 19%, respectively; Figure 4A and Figure S2). These results suggest that most SNPs are geographically widespread in Europe and that genetic differentiation among populations is moderate, consistent with high levels of gene flow across the European continent.

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

D. melanogaster originated in sub-Saharan Africa, migrated to Europe ~10,000-15,000 years ago, and subsequently colonized the rest of the world, including North America and Australia ~150 years ago (Lachaise *et al.* 1988; David & Capy 1988; Keller 2007). To search for genetic signatures of this shared history, we investigated the amount of allele sharing between African, European, and North American populations. We compared our SNP set to two published datasets, one from Zambia in sub-Saharan Africa (DPGP3; Lack *et al.* 2015) and one from North Carolina in North America (DGRP; Huang *et al.* 2014). Populations from Zambia inhabit the ancestral geographical range of *D. melanogaster* (Pool *et al.* 2012; Lack *et al.* 2015); North American populations are thought to be derived from European populations, with some degree of admixture from African populations, particularly in the southern United States and the Caribbean (Caracristi & Schlötterer 2003; Yukilevich & True 2008a; b; Yukilevich *et al.* 2010; Duchen *et al.* 2013; Kao *et al.* 2015; Pool 2015; Bergland *et al.* 2016). The population from North Carolina exhibits primarily European ancestry, with approximately 15% admixture from Africa (Bergland *et al.* 2016).

Approximately 10% of the SNPs (~1 million) were shared among all three datasets (Figure 4B). Since the out-of-Africa range expansion and the subsequent colonization of the North America continent by European and - to a lesser degree - African ancestors was likely accompanied by founder effects, leading to a loss of African alleles, as well as adaptation to novel, temperate climates (de Jong & Bochdanovits 2003; Adrion *et al.* 2015) we predicted that a relatively high proportion of SNPs would be shared between Europe and North America. As expected, we found that the proportion of shared SNPs was higher between Europe and North America (22%) than between either Europe or North America and Zambia (11% and 13%, respectively; Figure 4B). When we analyzed SNPs in variant frequency bins, the proportion of SNPs shared across at least two continents increased from 26% to 41% for SNPs with variant frequencies larger than 50% (Figure S3A). In contrast, only 6% of the SNPs at low variant frequency (<10%; Figure S3C) were shared. These results are consistent with the loss of low-frequency variants during the colonization of new continents; they suggest that intermediate frequency alleles are more likely to be ancestral and thus shared across broad geographic scales. Interestingly, as compared to Africa and North America, we identified nearly 3 million private SNPs that are specific to Europe (Figure 4B). Given that North American and Australian populations are – at least partly – of European ancestry (e.g., Bergland *et al.* 2016), future analysis of our data may be able to shed light on the demography and adaptation of these derived populations.

European and other derived populations exhibit similar amounts of genetic variation

Next, we estimated genome-wide levels of nucleotide diversity within the European population samples using population genetic summary statistics. Pairwise nucleotide diversity (π and Watterson's θ), corrected for pooling (Futschik 2010; Kofler *et al.* 2011), ranged from 0.0047 to 0.0057 and from 0.0045 to 0.0064, respectively (Figure S4 and

Figure S5), with the estimates being qualitatively similar to those from non-African *D. melanogaster* populations sequenced as individuals (see Table S3 and Figure S6; Mackay *et al.* 2012; Langley *et al.* 2012; Huang *et al.* 2014; Grenier *et al.* 2015) or as pools (Kolaczkowski *et al.* 2011b; Fabian *et al.* 2012; Reinhardt *et al.* 2014). Estimates of π were slightly lower than, but in close agreement with, estimates of θ, leading to a slightly negative average of Tajima's *D* (Tajima 1989). Due to our SNP calling approach (see Materials and Methods), we found a deficiency of alleles with frequencies ≤ 0.01, both in the sample-wise site frequency spectra (SFS) and in the combined SFS by SNP type, with the sample-wise SFS being skewed towards low frequency variants (Figure S7A). In addition, we observed an excess of low-frequency SNPs at non-synonymous sites as compared to other types of site, which is consistent with purifying selection eliminating deleterious non-synonymous mutations (Figure 8B; Grenier *et al.* 2015).

Overall, we observed only minor differences in the amount of genetic variation among populations. Specifically, genome-wide π ranged from 0.005 (Yalta, Ukraine) to 0.006 (Chalet à Gobet, Switzerland) for autosomes, and from 0.003 (Odesa, Ukraine) to 0.0035 (Chalet à Gobet, Switzerland) for the *X* chromosome (Table S1 and Figure S4). When testing for associations between geographic variables and genome-wide average levels of genetic variation, we found that both π and θ were strongly negatively correlated with altitude, but neither was correlated with latitude or longitude (Table 2). No correlations were found between the season in which the samples were collected and levels of average genome-wide genetic variation as measured by π and θ (Table 2).

The *X* chromosome showed markedly lower genetic variation than the autosomes, with the ratio of *X*-linked to autosomal variation (π_X/π_A) ranging from 0.53 to 0.66. These values

are well below the ratio of 0.75 (one-sample Wilcoxon rank test, $p < 0.001$) that is expected under standard neutrality and equal sex ratios, but are consistent with previous findings for European populations of D. *melanogaster* and variously attributed to selection (e.g., Hutter *et al.* 2007), or changes in population size (Pool & Nielsen 2007). This pattern is consistent with previous estimates of relatively low *X*-linked diversity for European (Andolfatto 2001; Kauer *et al.* 2002; Hutter *et al.* 2007) and other non-African populations (also see Betancourt *et al.* 2004 and references therein; Mackay *et al.* 2012; Langley *et al.* 2012). Interestingly, the π_X/π_A was significantly, albeit weakly, positively correlated with latitude (Spearman's $r = 0.315$, $p = 0.0289$), with northern populations having slightly higher *X*/*A* ratios than southern populations, contrary to our naive prediction of periodically bottlenecked populations leading to a lower *X*/*A* ratio in the north, perhaps reflecting more complex demographic scenarios (Hutter *et al.* 2007; Pool & Nielsen 2007).

In contrast to π and θ, we observed major differences in the genome-wide averages of Tajima's *D* among samples (Figure S8). The chromosome-wide Tajima's *D* was negative in approximately half of all samples and close to zero or slightly positive in the remaining samples, possibly due to heterogeneity in the proportion of sequencing errors among the multiplexed sequencing runs. However, models that included sequence run as a covariate did not explain more of the observed variance than models without the covariate, suggesting that associations of π and θ with geographic variables were not confounded by sequencing heterogeneity (see Supporting Information; Table S4). Moreover, our results for π, θ and *D* are unlikely to be confounded by spatio-temporal autocorrelations: after accounting for similarity among spatial neighbors (Moran's $I \approx 0$, $p > 0.05$ for all tests), there were no significant residual autocorrelations among samples for these estimators.

Genetic variation was not distributed homogeneously across the genome. Both π and θ were markedly reduced close to centromeric and telomeric regions (Figure 5), which is in good agreement with previous studies reporting systematic reductions in genetic variation in regions with reduced recombination (Begun & Aquadro 1992; Mackay *et al.* 2012; Langley *et al.* 2012; Huang *et al.* 2014). Consistent with this, we detected strong correlations with estimates of recombination rates based on the data of Comeron *et al.* (2012; linear regression, $p < 0.001$; not accounting for autocorrelation), suggesting that the distribution of genome-wide genetic variation is strongly influenced by the recombination landscape (Table S5). For autosomes, fine-scale recombination rates explained 41-47% of the variation in π, whereas broad-scale recombination rates (based on the recombination rate calculator of Fiston-Lavier *et al.* 2010) explained 50-56% of the variation in diversity. We obtained similar results for *X*-chromosomes, with recombination rates explaining 31-38% (based on Comeron *et al.* 2012) or 24-33% (based on Fiston-Lavier *et al.* 2010) of the variation (Figure 5, Table S5, Figure S9).

We also observed variation in Tajima's *D* with respect to genomic position (Figure 5). Notably, Tajima's *D* was markedly lower than the corresponding chromosome-wide average in the proximity of telomeric and centromeric regions on all chromosomal arms. These patterns possibly reflect purifying selection or selective sweeps close to heterochromatic regions (but see Betancourt *et al.* 2009), or might alternatively be a result of sequencing errors having a stronger effect on the SFS in low SNP density regions.

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

We identified 144 genomic locations on the autosomes with non-zero recombination, reduced genetic variation, and a local reduction in Tajima's *D* (see Methods, Table S6),

which jointly may be indicative of selective sweeps. Although we cannot rule out that these patterns are the result of non-selective demographic effects (e.g., bottlenecks), two observations suggest that at least some of these regions are affected by positive selection. First, bottlenecks are typically expected to cause genome-wide, non-localized reductions in Tajima's *D*. Second, several of these genomic regions coincide with previously identified, well-supported selective sweeps in the proximity of *Hen1*, *Cyp6g1* (Figure 6A and Figure S10A; Daborn *et al.* 2002), *wapl* (Beisswanger *et al.* 2006), *HDAC6* (Svetec *et al.* 2009), or around the chimeric gene *CR18217* (Figure 6B, Figure S10B; Rogers & Hartl 2012). Note, however, that some regions, such as those around *wapl* or *HDAC6,* are characterized by low recombination rates (< 0.5 cM/Mb; Table S5), which can itself lead to reduced variation and Tajima's *D* (see also Nolte *et al.* 2013). Our screen also uncovered several regions that have not previously been described as harboring sweeps (Table S6). These represent promising candidate regions containing putative targets of positive selection. For several of these candidate sweep regions, patterns of variation were highly similar across the majority of European samples, suggesting continent-wide selective sweeps that either predate the colonization of Europe (e.g., Beisswanger *et al.* 2006) or that have swept across all European populations more recently. In contrast, some candidate sweep regions were restricted to only a few populations and characterized by highly negative values of Tajima's *D*, i.e. deviating from the among-population average by more than two standard deviations, thus possibly hinting at cases of local, population-specific adaptation (Figure S11 and Table S6 for examples).

European populations are strongly structured along an east-west gradient

We next investigated patterns of genetic differentiation due to demographic substructure. Overall, pairwise differentiation as measured by F_{ST} was relatively low, though markedly

376 higher for *X*-chromosomes (0.043–0.076) than for autosomes (0.013–0.059; Student's *t*-
377 test; $p < 0.001$; Figure S12), possibly reflecting differences in effective population size
378 between the *X* chromosome and the autosomes (Hutter *et al.* 2007). One population, from
379 Sheffield, UK, showed an unusually high amount of differentiation on the *X*-chromosome
380 as compared to other populations (Figure S12).
381
382 Despite these overall low levels of among-population differentiation, European populations
383 showed some evidence of geographic substructure. To analyze this structure in more
384 detail, we focused on a set of SNPs located in short introns (< 60 bp), as these sites are
385 relatively unaffected by selection (Haddrill *et al.* 2005; Singh *et al.* 2009; Parsch *et al.*
386 2010; Clemente & Vogl 2012; Lawrie *et al.* 2013). We analyzed the extent of isolation by
387 distance (IBD) within Europe by correlating genetic and geographic distance and using
388 pairwise F_{ST} between populations as a measure of genetic isolation. F_{ST} was overall low
389 but significantly correlated with distance across the continent, indicating weak but
390 significant IBD (Mantel test; $p < 0.001$; max. $F_{ST} \sim 0.05$; Figure 7A). We also investigated
391 the populations that were most and least separated by genetic differentiation, estimated by
392 pairwise F_{ST} (Figure 7B). In general, we found that longitude had a stronger effect on
393 isolation than latitude, with populations showing the strongest differentiation separated
394 along an east-west, rather than a north-south, axis (Figure 7B). This pattern remained
395 when the number of populations sampled from Ukraine was reduced to avoid over-
396 representation (Figure S13).
397
398 To further explore these patterns, we performed a principal component analysis (PCA) on
399 the allele frequencies of SNPs in short introns. The first three principal components (PC)
400 explained more than 25% of the total variance (PC1: 16.3%, PC2: 5.4%, PC3: 4.8%,

401 eigenvalues = 599.2, 199.1, and 178.5 respectively; Figure 7C and Figure S14). As
402 expected, PC1 was strongly correlated with longitude. Despite significant signals of
403 autocorrelation, as indicated by Moran's test on residuals from linear regressions with
404 PC1, the association with longitude was not due to spatial autocorrelation, since a spatial
405 error model also resulted in a significant association. PC2 was similarly, but to a lesser
406 extent, correlated with longitude and also with altitude. PC3, by contrast, was not
407 associated with any variable examined (Table 2). None of the major PC axes were
408 correlated with season, indicating that there were no shared seasonal differences across
409 samples in our dataset. Hierarchical model fitting based on the first three PC axes resulted
410 in five distinct clusters (Figure 7C) that were oriented along the axis of PC1, supporting the
411 notion of strong longitudinal differentiation among European populations. To the best of
412 our knowledge, such a pronounced longitudinal signature of differentiation has not
413 previously been reported in European *D. melanogaster*. Remarkably, this pattern is
414 qualitatively similar to those observed for human populations (Cavalli-Sforza 1966; Xiao *et*
415 *al.* 2004; Francalacci & Sanna 2008), perhaps consistent with co-migration of this
416 commensal species.
417
418 **Mitochondrial haplotypes also exhibit longitudinal population structure**
419 Our finding that European populations are longitudinally structured is also supported by an
420 analysis of mitochondrial haplotypes. We identified two main mitochondrial haplotypes in
421 Europe, separated by at least 41 mutations (between G1.2 and G2.1; Figure 8A). Our
422 findings are consistent with similar analyses of mitochondrial haplotypes from a North
423 American *D. melanogaster* population (Cooper *et al.* 2015) as well as from worldwide
424 samples (Wolff *et al.* 2016), revealing varying degrees of differentiation among haplotypes,
425 ranging from only a few to hundreds of substitutions. The two G1 subtypes (G1.1 and

426 G1.2) are separated by only four mutations, and the three G2 subtypes are separated by a
427 maximum of four mutations (between G2.1 and G2.3). The estimated frequency of these
428 haplotypes varied greatly among populations (Figure 8B). Qualitatively three types of
429 European populations can be distinguished based on these haplotypes, namely those with
430 (1) a high frequency (> 60%) of the G1 haplotypes, characteristic of central European
431 samples, (2) a low frequency (< 40%) of G1 haplotypes, a pattern common for Eastern
432 European populations in summer, and (3) a combined frequency of G1 haplotypes,
433 between 40-60%, which is typical of samples from the Iberian Peninsula and from Eastern
434 Europe in fall (Figure S15A).
435
436 We observed a significant shift in the relative frequencies of the two haplotype classes
437 between summer and fall samples in only two of the nine possible comparisons among
438 haplotypes. While there was no correlation between latitude and the combined frequency
439 of G1 haplotypes, we found a weak but significant negative correlation between G1
440 haplotypes and longitude ($r^2 = 0.10$; $p < 0.05$), which is consistent with the longitudinal
441 east-west population structure observed for intronic SNPs. In a subsequent analysis, we
442 divided the dataset at 20º longitude into an eastern and a western subset since in northern
443 Europe 20º longitude corresponds to the division of two major climatic zones, namely C
444 (temperate) and D (cold), according to the Köppen-Geiger climate classification (Peel *et al.*
445 2007). When splitting the populations in a western (longitude < 20º E) and an eastern
446 group (longitude > 20° E), we found a clear correlation between longitude and the
447 combined frequency of G1 haplotypes, explaining as much as 50% of the variation in the
448 western group (Figure S15B). Similarly, in the eastern populations we found a correlation
449 between longitude and the combined frequency of G1 haplotypes, which explains
450 approximately 20% of the variance (Figure S15B). Thus, the data on mitochondrial

haplotypes clearly confirm the existence of pronounced east-west population structure and differentiation in European *D. melanogaster*. While this might be due to climatic selection, as recently found for clinal mitochondrial haplotypes in Australia (Camus *et al.* 2017), we can presently not rule out an effect of demography.

The majority of TEs vary with longitude and altitude

To examine the population genetics of structural variants in our dataset, we first focused on TEs. The repetitive content of the 48 samples analyzed ranged from 16% to 21% with respect to nuclear genome size (Figure 9). The vast majority of detected repeats were transposable elements, mostly represented by long terminal repeats (LTR) and long interspersed nuclear elements (LINE; Class I), as well as a few DNA elements (Class II). LTR content best explained total TE content (LINE+LTR+DNA) (Pearson's $r = 0.87$, $p < 0.01$, *vs.* DNA $r = 0.58$, $p = 0.0117$, and LINE $r = 0.36$, $p < 0.01$ and Figure S16A).

We estimated population-wise 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.* 2010). On average, 56% of the TEs annotated in the reference genome were fixed in all samples. The remaining polymorphic TEs usually segregated at low frequency in all samples (Figure S16A), 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 (Figure S16B), which might be consistent with transposition-selection balance (Charlesworth *et al.* 1994).

In each of the 48 samples TE frequency and recombination rate were negatively correlated

476 on a genome-wide level (Spearman rank sum test; $p < 0.01$), as previously reported
477 (Bartolomé *et al.* 2002; Petrov *et al.* 2011; Kofler *et al.* 2012). Qualitatively, this pattern still
478 holds when only polymorphic TEs (population frequency <95%) are analyzed, although it
479 becomes statistically non-significant for some chromosomes and populations (Table S8).
480 In either case, the correlation is more negative when using broad-scale, rather than fine-
481 scale, recombination rate estimates (Materials and methods, Tables S8B, S8D). This
482 suggests that broad-scale recombination patterns may best capture long-term population
483 recombination patterns.
484
485 We further tested whether the distribution of TE frequencies among samples could be
486 explained by geographical or temporal variables. We focused on the 141 TE insertions that
487 showed frequency variability among samples (interquartile range, (IQR) > 10; see
488 Materials and Methods). Of these, 73 TEs showed significant associations with
489 geographical or temporal variables after multiple testing correction (Table S9). Note that
490 we used a conservative *p*-value threshold (< 0.001), and we did not find significant
491 residual spatio-temporal autocorrelation among samples for any TE tested (Moran's *I* >
492 0.05 for all tests; Table S9). Sixteen out of seventy-three TEs were located in regions of
493 very low recombination (0 cM/Mb for either of the two recombination measures used).
494 Among the 57 significant TEs located in high recombination regions, we observed
495 significant correlations of 13 TE's with longitude, of 13 with altitude, of five with latitude,
496 and of three with season (Table S9). In addition, the frequencies of the other 23 insertions
497 were significantly correlated with more than one of the above-mentioned variables (Table
498 S9). These significant TEs were scattered along the main five chromosome arms (Table
499 S9). Among the 57 significant TEs located in high recombination regions two TE families
500 were enriched (χ^2 p-values after Yate's correction < 0.05): the LTR *297* family with 11

copies, and the DNA *pogo* family with 5 copies (Table S10). We also checked the genomic localization of the 57 TEs. Most of them (42) were located inside genes: two in 5'UTR, four in 3'UTR, 18 in the first intron, and 18 TEs in subsequent introns. Additionally, 7 TEs are <1 kb from the nearest gene, suggesting that these could also affect the regulation of the nearby genes (Table S9). Interestingly, 14 of these 57 TEs coincide with previously identified candidate adaptive TEs (Table S9), suggesting that our dataset might be enriched for adaptive insertions. However, further analyses are needed to discard the effect of non-selective forces on the patterns observed.

Inversion polymorphisms in Europe exhibit latitudinal and longitudinal clines

Chromosomal inversions are another class of important and common structural genomic variants, often exhibiting frequency clines on multiple continents, some of which have been shown to be adaptive (e.g. Kapun *et al.* 2014; 2016a). However, little is known yet about the spatial distribution and clinality of inversions in Europe. We used a panel of inversion-specific marker SNPs (Kapun *et al.* 2014) to test for the presence and quantify the frequency of six cosmopolitan inversion polymorphisms (*In(2L)t, In(2R)NS, In(3L)P, In(3R)C, In(3R)Mo, In(3R)Payne*) in the 48 samples. All sampled populations were polymorphic for one or more inversions (Figure 10). However, only *In(2L)t* segregated at substantial frequencies in most populations (average frequency = 20.2%). All other inversions were either absent or occurred at low frequencies (average frequencies: *In(2R)NS* = 6.2%, *In(3L)P* = 4%, *In(3R)C* = 3.1%, *In(3R)Mo* =2.2%, *In(3R)Payne* = 5.7%).

Despite their overall low frequencies, several inversions exhibited clinal patterns across space (Table 3). 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 that are qualitatively similar to the clines previously observed along the North American and Australian east coasts (Figure S17 and Table S11, Kapun *et al.* 2016a). For the first time, we also detected a longitudinal cline for *In(2L)t* and *In(2R)NS*, with both inversions decreasing in frequency from East to West, a result that is consistent with our finding of strong longitudinal among-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 $I \approx 0$, $p > 0.05$ for all tests; Table 3), suggesting that our analysis was not confounded by spatial autocorrelation for most of the inversions. It will be interesting to examine in future work the extent to which clines in inversions (and other genomic variants) across Europe are shaped by selection and/or demography.

European *Drosophila* microbiomes contain trypanosomatids and novel viruses

Finally, we determined the abundance of microbiota associated with *D. melanogaster* from the Pool-Seq data – these endosymbionts often have crucial functions in affecting the life history, immunity, hormonal physiology, and metabolic homeostasis of their fly hosts (e.g., Trinder *et al.* 2017; Martino *et al.* 2017). The taxonomic origin of a total of 262 million non-*Drosophila* reads was inferred using MGRAST, which identifies and counts short protein motifs ('features') within reads (Meyer *et al.* 2008). The largest fraction of protein features was assigned to *Wolbachia* (on average 53.7%; Figure 11), a well-known endosymbiont of *Drosophila* (Werren *et al.* 2008). The relative abundance of *Wolbachia* protein features varied strongly between samples ranging from 8.8% in a sample from the UK to almost 100% in samples from Spain, Portugal, Turkey and Russia (Table 1). Similarly, *Wolbachia* loads varied 100x between samples if we use the ratio of *Wolbachia* protein features divided by the number of *Drosophila* sequences retrieved for that sample as a proxy for

551 relative micro-organismal load (for a full table of micro-organismal loads standardized by
552 *Drosophila* genome coverage see Table S12).
553
554 Acetic acid bacteria of the genera *Gluconobacter*, *Gluconacetobacter*, and *Acetobacter*
555 were the second largest group, with an average relative abundance of 34.4%.
556 Furthermore, we found evidence for the presence of several genera of Enterobacteria
557 (*Serratia, Yersinia, Klebsiella, Pantoea, Escherichia, Enterobacter, Salmonella*, and
558 *Pectobacterium*). *Serratia* occurs only at low frequencies or is absent from most of our
559 samples, but reaches a very high relative abundance in the Nicosia summer collection
560 (54.5%). This high relative abundance was accompanied by an 80x increase in *Serratia*
561 bacterial load. We detected several eukaryotic microorganisms, although they were less
562 abundant than the bacteria. The fraction of fungal protein features is larger than 3% in only
563 three of our samples from Finland, Austria and Turkey (Table 1). Interestingly, we detected
564 the presence of trypanosomatids in 16 of our samples, consistent with other recent
565 evidence that *Drosophila* can host these organisms (Wilfert *et al.* 2011; Chandler & James
566 2013; Hamilton *et al.* 2015).
567
568 Our data also allowed us to detect the presence of five different DNA viruses (Table S13).
569 These included approximately two million reads from *Kallithea* nudivirus (Webster *et al.*
570 2015), allowing us to assemble the complete *Kallithea* genome for the first time (>300-fold
571 coverage in the Ukrainian sample UA_Kha_14_46; Genbank accession KX130344). We
572 also identified around 1000 reads from a novel nudivirus that is closely related to *Kallithea*
573 virus and to *Drosophila innubila* nudivirus (Unckless 2011) in sample DK_Kar_14_41 from
574 Karensminde, Denmark (Table 1). These sequences permitted us to identify a publicly
575 available dataset (SRR3939042: 27 male D. *melanogaster* from Esparto, California;

576 Machado *et al.* 2016) that contained sufficient reads to complete the genome (provisionally
577 named "*Esparto* Virus"; KY608910). We further identified two novel Densoviruses
578 (Parvoviridae), which we have provisionally named "*Viltain* virus", a relative of *Culex*
579 *pipiens* densovirus found at 94-fold coverage in sample FR_Vil_14_07 (Viltain; KX648535)
580 and "*Linvill Road* virus", a relative of *Dendrolimus punctatus* densovirus that was
581 represented by only 300 reads here, but which has previously been found to have a high
582 coverage in dataset SRR2396966 from a North American sample of *D. simulans*
583 (KX648536; Machado *et al.* 2016). In addition, we detected a novel member of the
584 Bidnaviridae family,"*Vesanto* virus", a bidensovirus related to *Bombyx mori* densovirus 3
585 with approximately 900-fold coverage in sample FI_Ves_14_38 (Vesanto; KX648533 and
586 KX648534), Using a detection threshold of >0.1% of the *Drosophila* genome copy number,
587 the most commonly detected viruses were *Kallithea* virus (30/48 of the pools) and *Vesanto*
588 virus (25/48), followed by *Linvill Road* virus (7/48) and *Viltain* virus (5/48), with *Esparto*
589 virus being the rarest (2/48). In some samples, the viruses reached strikingly high titers: on
590 13 occasions the virus genome copy number in the pool exceeded the host genome copy
591 number, reaching a maximum of nearly 20-fold in Vesanto.
592
593 In summary, our continent-wide analysis of the microbiota associated with flies suggests
594 that natural populations of European *D. melanogaster* differ greatly in the composition and
595 relative abundance of microbes and viruses.
596

597 Discussion

598 In recent years, large-scale population resequencing projects have shed light on the
599 biology of both model (Mackay *et al.* 2012; Langley *et al.* 2012; Consortium 2015; Lack *et*
600 *al.* 2015; Alonso-Blanco *et al.* 2016; Lack *et al.* 2016) and non-model organisms (e.g.,

601 Hohenlohe *et al.* 2010; Wolf *et al.* 2010). Such massive datasets contribute significantly to
602 our growing understanding of the processes that create and maintain genetic variation in
603 natural populations and of adaptation. However, the relevant spatio-temporal scales for
604 population genomic analyses remain largely unknown. Here we have applied, for the first
605 time, a comprehensive sampling and sequencing strategy to European populations of
606 *Drosophila melanogaster*, allowing us to uncover previously unknown aspects of this
607 species' population biology.
608
609 A main result from our analyses of SNPs, located in short introns and presumably evolving
610 neutrally (Parsch *et al.* 2010), is that European *D. melanogaster* populations exhibit very
611 pronounced longitudinal differentiation, a pattern that – to the best of our knowledge – has
612 not been observed before for the European continent (for patterns of longitudinal
613 differentiation in Africa see e.g. Michalakis & Veuille 1996; Aulard *et al.* 2002; Fabian *et al.*
614 2015). Genetic differentiation was greatest between populations from eastern and western
615 Europe (Figure 7). The eastern populations included those from the Ukraine, Russia, and
616 Turkey, as well as one from eastern Austria, which suggests that there may be a region of
617 restricted gene flow in south-central Europe. However, populations from Finland and
618 Cyprus are more similar to western populations than to eastern populations, possibly as a
619 result of migration along shipping routes in the Baltic and Mediterranean seas. More data
620 from populations in the unsampled, intermediate regions are needed to better delineate
621 the geographic limits of the eastern and western population groups. Consistent with the
622 strong differentiation between eastern and western populations, our PCA analysis
623 revealed that longitude was the major factor associated with among-population
624 divergence, with no significant effect of latitude (Figure 7C; Table 2). Thus, the patterns of
625 neutral genetic differentiation in Europe contrast with those previously reported for North

626 America, where latitude impacts neutral differentiation (Machado *et al.* 2016; Kapun *et al.*
627 2016a). However, note that our analysis does not exclude the existence of clinally varying
628 polymorphisms in European populations outside short introns: for example, we detected
629 latitudinal frequency clines both for TEs and inversion polymorphisms (see below). A
630 detailed analysis of clinal variation in the *DrosEU data* is beyond the scope of this paper
631 and currently under way.
632
633 The mitochondrial genome and several chromosomal inversions and TEs (also see below)
634 showed similar patterns of differentiation as the rest of the genome, with the main axis of
635 differentiation being longitudinal. Uncovering the extent to which this pattern might be
636 driven by demography and/or selection, and the underlying environmental correlates
637 (including any potential role of co-migration with human populations), will be an important
638 task for future analyses. Due to the high density of samples and the large number of SNP
639 markers, our results show that European populations of *D. melanogaster* exhibit much
640 more differentiation and structure than previously thought (e.g., Baudry *et al.* 2004;
641 Dieringer *et al.* 2005; Schlötterer *et al.* 2006; Nunes *et al.* 2008; Mateo *et al.* 2018).
642
643 Within the eastern and western population groups there was a low – but detectable – level
644 of genetic differentiation among populations, including those that are geographically close
645 (Figure 7C). These population differences persisted over a timespan of at least 2–3
646 months, as there was less genetic differentiation between the summer and fall samples of
647 the 13 locations sampled at multiple time points than between neighboring populations
648 (Figure 7C). Thus, while the weak but significant signal of isolation by distance suggests
649 homogenizing gene flow across geography, there is seasonally stable differentiation
650 among populations. The season in which samples were collected did not show a

651 significant association with genetic differentiation, except when considered in conjunction
652 with longitude or altitude (Table 2). Note, however, that the data here are from a single
653 year only; demonstrating recurrent shifts in SNP frequencies due to temporally varying
654 selection will require analysis of additional annual samples.
655
656 Our Pool-Seq data also allowed us to characterize geographic patterns in both inversions
657 and TEs. In marked contrast to putatively neutral SNPs, the frequencies of several
658 chromosomal inversions, including *In(3L)P, In(3R)C,* and *In(3R)Payne*, showed a
659 significant correlation with latitude (Table 3). For *In(3L)P* and *In(3R)Payne*, the latitudinal
660 clines were in qualitative agreement with parallel clines reported in North America and
661 Australia, with the inversions decreasing in frequency as distance from the equator
662 increases (Mettler *et al.* 1977; Knibb *et al.* 1981; Fabian *et al.* 2012; Kapun *et al.* 2014;
663 Rane *et al.* 2015; Kapun *et al.* 2016a). This suggests that the inversions may contain
664 genetic variants that are better adapted to warmer environments than to temperate
665 climates. Note, however, that the overall frequencies of these inversions are low within
666 Europe (<5%), indicating that they might play only a minor role in local adaptation to
667 European habitats. Some euchromatic TE insertions also showed geographic or seasonal
668 patterns of variation (Table S7), suggesting that they have the potential to play a role in
669 local adaptation, particularly as many of them are located in regions where they might
670 affect gene regulation. Importantly, several inversions and TEs also showed longitudinal
671 frequency gradients, thus supporting the notion that European populations exhibit marked
672 longitudinal differentiation.
673
674 We also examined signatures of selective sweeps in our dataset. We found 144 genomic
675 regions that showed signatures of hard sweeps (i.e., in regions of normal recombination

676 (cM/Mb ≥ 0.5), but with reduced variation and negative Tajima's *D* (*D* ≤ -0.8)) in all
677 European populations (Figure 6, Table S6). Four of these regions were identified in
678 previous studies as potential targets for positive selection.
679
680 The first region, at the center of chromosome arm *2R* (Figure 6A), was previously found to
681 be strongly differentiated between African and North American populations (Langley *et al.*
682 2012) and contains two genes, *Cyp6g1* and *Hen-1*, that are associated with recent, strong
683 selection. The cytochrome P450 gene *Cyp6g1* that has been linked to insecticide
684 resistance (Daborn *et al.* 2002; Schmidt *et al.* 2010), shows evidence for recent selection
685 independently in both *D. melanogaster* and *D. simulans* (Schlenke & Begun 2003; Catania
686 *et al.* 2004), and is associated with a large differentiated region in the Australian latitudinal
687 cline (Kolaczkowski *et al.* 2011a). *Hen-1*, a methyltransferase involved in maturation of
688 small RNAs involved in virus and TE suppression, showed marginally non-significant
689 evidence for selective sweeps in North American and African populations of *D.*
690 *melanogaster* (Kolaczkowski *et al.* 2011b).
691
692 The second region previously implicated in a selective sweep is on chromosome arm *3L*
693 (Table S6) and is centered around the chimeric gene *CR18217*, which formed from the
694 fusion of a gene encoding a DNA-repair enzyme (*CG4098*) and a centriole gene (spd-2;
695 Rogers & Hartl 2012). *CR18217* appears to be unique to *D. melanogaster*, but, in spite of
696 its recent origin, segregates at frequencies of around 90% (Rogers & Hartl 2012),
697 consistent with a recent strong sweep in this region of the genome. This putative sweep
698 region also spans *Prosbeta6*, which (like HDAC) encodes a gene involved in proteolysis
699 (Flybase v. FB2017_05; Gramates *et al.* 2017). *Prosbeta6* also shows homology to genes
700 involved in immune function (Lyne *et al.* 2007; Handu *et al.* 2015), which might explain

701 why it has been a target of positive selection.

702

703 The third previously characterized sweep region, surrounding the *wapl* gene on the *X*
704 chromosome, was identified as showing evidence of a strong selective sweep (or sweeps)
705 in both African and European *D. melanogaster* populations (Beisswanger *et al.* 2006;
706 Boitard *et al.* 2012). The gene(s) targeted by selection is unclear, but is most likely *ph-p* in
707 Europe and *ph-p* or *ph-d* in Africa (Beisswanger *et al.* 2006). These genes are tandem
708 duplicates involved in the Polycomb response pathway, which functions as an epigenetic
709 repressor of transcription (reviewed in Kassis *et al.* 2017).

710

711 The fourth previously known putative sweep region, originally identified in African
712 populations of *D. melanogaster*, is also on the *X* chromosome, but 30 cM closer to the
713 telomere and thus not implicating the *wapl* sweep region (Beisswanger *et al.* 2006; Boitard
714 *et al.* 2012). Selection in this region has been attributed to the *HDAC6* gene (Svetec *et al.*
715 2009). *HDAC6*, although nominally a histone deacetylase, actually functions as a central
716 player in managing cytotoxic assaults, including in transport and degradation of misfolded
717 protein aggregates (reviewed in Matthias *et al.* 2008; Svetec *et al.* 2009).

718

719 Our data thus support the widespread occurrence of these previously identified sweeps in
720 many populations in Europe. Notably, practically all European populations showed
721 reduced variation and negative Tajima's *D* in the identified sweep regions. This is
722 consistent with the sweeps either pre-dating the colonization of Europe (e.g., Beisswanger
723 *et al.* 2006) or having swept across Europe more recently (also see Stephan 2010 for
724 discussion). In addition, we further uncovered several novel genomic regions showing
725 evidence for hard sweeps (Table S6), which represent a valuable source for future in-

726 depth analyses of signals for adaptive evolution in European *Drosophila.*

727

728 Finally, we used our Pool-Seq data to identify microbes and viruses and to quantify their
729 presence in natural populations of *D. melanogaster* across the European continent.
730 *Wolbachia* was the most abundant bacterial genus associated with the flies, but its relative
731 abundance and load varied greatly among samples. The second most abundant bacterial
732 taxon was acetic acid bacteria (Acetobacteraceae), a group previously found among the
733 most abundant bacteria in natural *D. melanogaster* isolates (Chandler *et al.* 2011;
734 Staubach *et al.* 2013). Other microbes were highly variable abundance in relative
735 abundance. For example, *Serratia* abundance was low in most populations, but very high
736 in the Nicosia sample, which might reflect that there are individuals in the Nicosia sample
737 that carry a systemic *Serratia* infection generating high bacterial loads. Future sampling
738 may shed light on the temporal stability and/or population specificity of these patterns.
739 Contrary to our expectation, we found relatively few yeast sequences. This is something of
740 a surprise, because yeasts are commonly found on rotting fruit, the main food substrate of
741 *D. melanogaster*, and have been found in association with *Drosophila* before (Barata *et al.*
742 2012; Chandler *et al.* 2012). This suggests that although yeasts can attract flies and play a
743 role in food choice (Becher *et al.* 2012; Buser *et al.* 2014), they are not highly prevalent in
744 or on *D. melanogaster* bodies. Although trypanosomatids have been reported in
745 association with *Drosophila* before (Wilfert *et al.* 2011; Chandler & James 2013; Hamilton
746 *et al.* 2015), our study provides the first systematic detection across a wide geographic
747 range in *D. melanogaster*.

748

749 Despite being host to a wide diversity of RNA viruses (Huszar & Imler 2008; Webster *et al.*
750 2015), only three DNA viruses have previously been reported in association with the

751 *Drosophilidae*, and only one from *D. melanogaster* (Unckless 2011; Webster *et al.* 2015;
752 2016). Here, we discovered four new DNA viruses. While it is not possible to directly
753 estimate viral prevalence from pooled sequencing data, we found that the DNA viruses of
754 *D. melanogaster* can be widespread, with *Kallithea* virus detectable at a low level in most
755 populations.
756
757 A striking qualitative pattern in our microbiome data is the high level of variability among
758 populations in the composition and relative amounts of different microbiota and viruses.
759 Thus, an interesting open question is to what extent geographic differences in microbiota
760 might contribute to phenotypic differences and local adaptation among fly populations,
761 especially given that there might be tight and presumably local co-evolutionary interactions
762 between fly hosts and their endosymbionts (e.g., Haselkorn *et al.* 2009; Richardson *et al.*
763 2012; Staubach *et al.* 2013; Kriesner *et al.* 2016).
764
765 In conclusion, our study demonstrates that extensive sampling on a continent-wide scale
766 and pooled sequencing of natural populations can reveal new aspects of population
767 biology, even for a well-studied species such as *D. melanogaster*. Such extensive
768 population sampling is feasible due to the close cooperation and synergism within our
769 international consortium. Our efforts in Europe are paralleled in North America by the
770 *Drosophila* Real Time Evolution Consortium (*Dros-RTEC*), with whom we are currently
771 collaborating to compare population genomic data across continents. In future years, our
772 consortia will continue to sample and sequence European and North American *Drosophila*
773 populations to study these populations with increasing spatial and temporal resolution,
774 providing an unprecedented resource for the *Drosophila* and population genetics
775 communities.

Materials and Methods

The *DrosEU* dataset analyzed here consists of 48 samples of *D. melanogaster* collected from 32 geographical locations at different time-points across the European continent, through a joint effort of 18 European research groups in 2014 (see Figure 3). Field collections were performed with baited traps using a standardized protocol (see Supplementary file for details). Up to 40 males from each collection were pooled, and DNA extracted from each, using a standard phenol-chloroform based protocol. Each sample was processed in a single pool (Pool-Seq; Schlötterer *et al.* 2014), with each pool consisting of at least 33 wild-caught individuals. To exclude morphologically similar and co-occurring species, such as *D. simulans*, as potential contaminants from the samples, we only used wild-caught males and distinguished among species by examining genital morphology. Despite this precaution, we identified a low level of *D. simulans* contamination in our samples, and further steps were thus taken to exclude *D. simulans* sequences from our analysis (see below). The *DrosEU* dataset represents the most comprehensive spatio-temporal sampling of European *D. melanogaster* populations available to date (Table 1, Figure 3).

DNA extraction, library preparation and sequencing

DNA was extracted from pools of 33–40 males per sample after joint homogenization with bead beating and standard phenol/chloroform extraction. A detailed extraction protocol can be found in the Supporting Information file. In brief, 500 ng of DNA in a final volume of 55.5 µl were sheared with a Covaris instrument (Duty cycle 10, intensity 5, cycles/burst 200, time 30) for each sample separately. Library preparation was performed using NEBNext Ultra DNA Lib Prep-24 and NebNext Multiplex Oligos for Illumina-24 following the

manufacturer's instructions. Each pool was sequenced as paired-end fragments on a Illumina NextSeq 500 sequencer at the Genomics Core Facility of Pompeu Fabra University (UPF; Barcelona, Spain). Samples were multiplexed in five batches of 10 samples each, except for one batch that contained only 8 samples (see Supplementary Table S1 for further information). Each multiplexed batch was sequenced on four lanes to obtain an approximate 50x raw coverage for each sample. Reads were sequenced to a length of 151 bp with a median insert size of 348 bp (ranging from 209 to 454 bp).

Mapping pipeline and variant calling

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

826 mapping. We used Picard (v.1.109; http://picard.sourceforge.net) to remove duplicate
827 reads and reads with a mapping quality below 20. In addition, we re-aligned sequences
828 flanking insertions-deletions (indels) with GATK (v3.4-46; McKenna *et al.* 2010).
829
830 After mapping, Pool-Seq samples were tested for DNA contamination from *D. simulans*.
831 To do this, we used a set of SNPs known to be divergent between *D. simulans* and *D.*
832 *melanogaster* and assessed the frequencies of *D. simulans*-specific alleles following the
833 approach of Bastide *et al*. (2013). We combined the genomes of *D. melanogaster* (v.6.12)
834 and *D. simulans* (Hu *et al.* 2013) and separated species-specific reads for samples with a
835 contamination level > 1% via competitive mapping against the combined references using
836 the pipeline described above. Custom software was used to remove reads uniquely
837 mapping to *D. simulans*. In 9 samples, we identified contamination with *D. simulans*,
838 ranging between 1.2 % and 8.7% (Table S1). After applying our decontamination pipeline,
839 contamination levels dropped below 0.4 % in all 9 samples.
840
841 We used *Qualimap* (v. 2.2., Okonechnikov *et al.* 2016) to evaluate average mapping
842 qualities per population and chromosome, which ranged from 58.3 to 58.8 (Table S1). We
843 found heterogeneous sequencing depths among the 48 samples, ranging from 34x to
844 115x for autosomes and from 17x to 59x for *X*-chromosomes (Figure S1, Table S1). We
845 then combined individual BAM files from all samples into a single *mpileup* file using
846 samtools (v. 1.3; Li & Durbin 2009). Due to the large number of Pool-Seq datasets
847 analyzed in parallel, we had to implement quality control criteria for all libraries jointly in
848 order to call SNPs. To accomplish this, we implemented a novel custom SNP calling
849 software to call SNPs with stringent heuristic parameters (PoolSNP; see Supplementary
850 Information), available at Dryad (doi: https://doi.org/10.5061/dryad.rj1gn54). A site was

851 considered polymorphic if (1) the minimum coverage from all samples was greater than
852 10x, (2) the maximum coverage from all samples was less than the 95th coverage
853 percentile for a given chromosome and sample (to avoid paralogous regions duplicated in
854 the sample but not in the reference), (3) the minimum read count for a given allele was
855 greater than 20x across all samples pooled, and (4) the minimum read frequency of a
856 given allele was greater than 0.001 across all samples pooled. The above threshold
857 parameters were optimized based on simulated Pool-Seq data in order to maximize true
858 positives and minimize false positives (see Figure S18 and Supporting Information).
859 Additionally, we excluded SNPs (1) for which more than 20% of all samples did not fulfill
860 the above-mentioned coverage thresholds, (2) which were located within 5 bp of an indel
861 with a minimum count larger than 10x in all samples pooled and (3) which were located
862 within known transposable elements (TE) based on the *D. melanogaster* TE library v.6.10.
863 We further annotated our final set of SNPs with SNPeff (v.4.2; Cingolani *et al.* 2012) using
864 the Ensembl genome annotation version BDGP6.82 (Figure 3).
865
866 **Combined and population-specific site frequency spectra (SFS)**
867 We quantified the amount of allelic variation with respect to different SNP classes. For this,
868 we first combined the full dataset across all 48 samples and used the SNPeff annotation
869 (see above) to classify the SNPs into four classes (intergenic, intronic, non-synonymous,
870 and synonymous). For each class, we calculated the site frequency spectrum (SFS) based
871 on minor allele frequencies for the *X*-chromosome and the autosomes, as well as for each
872 sample and chromosomal arm separately, by counting alleles in 50 frequency bins of size
873 0.01.
874
875 **Genetic variation in Europe**

We characterized patterns of genetic variation among the 48 samples by estimating three standard population genetic parameters: π, Watterson's θ and Tajima's D (Watterson 1975; Nei 1987; Tajima 1989). We focused on SNPs located on the five major chromosomal arms (*X*, *2L*, *2R*, *3L*, *3R*) and calculated sample-wise π, θ and Tajima's D with corrections for Pool-Seq data (Kofler *et al.* 2011). Since PoPoolation, the most commonly used software for population genetics inference from Pool-Seq data, does not allow using predefined SNPs (which was desirable for our analyses), we implemented corrected population genetic estimators described in Kofler *et al*. (2011) in Python (PoolGen; available at Dryad; doi: https://doi.org/10.5061/dryad.rj1gn54). Before calculating the estimators, we subsampled the data to an even coverage of 40x for the autosomes and 20x for the *X*-chromosome to control for the sensitivity to coverage variation of Watterson's θ and Tajima's D (Korneliussen *et al.* 2013). At sites with greater than 40x coverage, we randomly subsampled reads to 40x without replacement; at sites with below 40x coverage, we sampled reads 40 times with replacement. Using *R* (R Development Core Team 2009), we calculated sample-wise chromosome-wide averages for autosomes and *X* chromosomes separately and tested for correlations of π, θ and Tajima's D with latitude, longitude, altitude, and season using a linear regression model of the following form: $y_i = Lat + Lon + Alt + Season + \varepsilon_i$, where y_i is either π, θ and D. Here, latitude, longitude, and altitude are continuous predictors (Table 1), while 'season' is a categorical factor with two levels *S* ("summer") and *F* ("fall"), corresponding to collection dates before and after September 1st, respectively. We chose this arbitrary threshold for consistency with previous studies (Bergland *et al.* 2014; Kapun *et al.* 2016a). To further test for residual spatio-temporal autocorrelation among the samples (Kühn & Dormann 2012), we calculated Moran's *I* (Moran 1950) with the *R* package *spdep* (v.06-15., Bivand & Piras 2015). To do this, we used the residuals of the above-mentioned models, as well

as matrices defining pairs of samples as neighbors weighted by geographical distances between the locations (samples within 10° latitude/longitude were considered neighbors). Whenever these tests revealed significant autocorrelation (indicating non-independence of the samples), we repeated the above-mentioned regressions using spatial error models as implemented in the *R* package *spdep*, which incorporate spatial effects through weighted error terms, as described above.

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

To investigate genome-wide patterns of variation, we averaged π, θ, and D in 200 kb non-overlapping windows for each sample and chromosomal arm separately and plotted the distributions in *R.* In addition, we calculated Tajima's D in 50 kb sliding windows with a step size of 10 kb to investigate fine-scale deviations from neutral expectations. We applied heuristic parameters to identify genomic regions harboring potential candidates for selective sweeps. To identify candidate regions with sweep patterns across most of the 48 samples, we searched for windows with log-transformed recombination rates ≥ 0.5, pairwise nucleotide diversity ($\pi \leq 0.004$), and average Tajima's D across all populations ≤ -0.8 (5% percentile). To identify potential selective sweeps restricted to a few population samples only, we searched for regions characterized as above but allowing one or more samples with Tajima's D being more than two standard deviations smaller than the window-wise average. To account for the effects of strong purifying selection in gene-rich

926 genomic regions which can result in local negative Tajima's *D* (Tajima 1989) and thus
927 confound the detection of selective sweeps, we repeated the analysis based on silent sites
928 (4-fold degenerate sites, SNPs in short introns of ≤ 60 bp lengths and SNPs in intergenic
929 regions in ≥ 2000 bp distance to the closest gene) only. Despite of the reduction in
930 polymorphic sites available for this analysis, we found highly consistent sweep regions and
931 therefore proceeded with the full SNP datasets, which provided better resolution (results
932 not shown)
933
934 For statistical analysis, the diversity statistics were log-transformed to normalize the data.
935 We then tested for correlations between π and recombination rate using *R* in 100 kb non-
936 overlapping windows and plotted these data using the *ggplot2* package (v.2.2.1., Wickham
937 2016). We used two different recombination rate measurements: (i) a fine-scale, high
938 resolution genomic recombination rate map based on millions of SNPs in a small number
939 of strains (Comeron *et al.* 2012), and (ii) the broad-scale Recombination Rate Calculator
940 based on Marey maps generated by laboratory cross data fitting genetic and physical
941 positions of 644 markers to a third-order polynomial curve for each chromosome arm
942 (Fiston-Lavier *et al.* 2010). Both measurements were converted to version 6 of the *D.*
943 *melanogaster* reference genome to match the genomic position of π estimates (see
944 above).
945
946 **SNP counts and overlap with other datasets**
947 We used the panel of SNPs identified in the *DrosEU* dataset (available at Dryad; doi:
948 https://doi.org/10.5061/dryad.rj1gn54) to describe the overlap in SNP calls with other
949 published *D. melanogaster* population data: the *Drosophila* Population Genomics Project 3
950 (DPGP3) from Siavonga, Zambia (69 non-admixed lines; Lack *et al.* 2015; 2016) and the

951 *Drosophila* Genetic Reference Panel (DGRP) from Raleigh, North Carolina, USA (205
952 inbred lines; Mackay *et al.* 2012; Huang *et al.* 2014). For these comparisons, we focused
953 on biallelic SNPs on the 5 major chromosome arms. We used *bwa mem* for mapping and
954 a custom pipeline for heuristic SNP calling (PoolSNP; Figure 3). To make the data from
955 the 69 non-admixed lines from Zambia (Lack *et al.* 2015; 2016) comparable to our data,
956 we reanalyzed these data using our pipeline for mapping and variant calling (Figure 3).
957 The VCF file of the DGRP data was downloaded from http://dgrp2.gnets.ncsu.edu/ and
958 converted to coordinates according to the *D. melanogaster* reference genome v.6. We
959 depicted the overlap of SNPs called in the three different populations using elliptic Venn
960 diagrams with *eulerAPE* software (v3 3.0.0., Micallef & Rodgers 2014). While the *DrosEU*
961 data were generated from sequencing pools of wild-caught individuals, both the DGRP
962 and DPGP3 data are based on individual sequencing of inbreed lines and haploid
963 individuals, respectively.
964
965 **Genetic differentiation and population structure in European populations**
966 To estimate genome-wide pairwise genetic differences, we used custom software to
967 calculate SNP-wise F_{ST} using the approach of Weir and Cockerham (1984). We estimated
968 SNP-wise F_{ST} for all possible pairwise combinations among samples. For each sample, we
969 then averaged F_{ST} across all SNPs for all pairwise combinations that include this particular
970 sample and finally ranked the 48 population samples by overall differentiation.
971
972 We inferred demographic patterns in European populations by focusing on 21,008
973 putatively neutrally evolving SNPs located in small introns (less than 60 bp length; Haddrill
974 *et al.* 2005; Singh *et al.* 2009; Parsch *et al.* 2010; Clemente & Vogl 2012; Lawrie *et al.*
975 2013) that were at least 200 kb distant from the major chromosomal inversions (see

below). To assess isolation by distance (IBD), we averaged F_{ST} values for each sample pair across all neutral markers and calculated geographic distances between samples using the haversine formula (Green & Smart 1985) which takes the spherical curvature of the planet into account. We tested for correlations between genetic differentiation and geographic distance using Mantel tests using the *R* package *ade4* (v.1.7-8., Dray & Dufour 2007) with 1,000,000 iterations. In addition, we plotted the 5% smallest and 5% largest F_{ST} values from all 1,128 pairwise comparisons among the 48 population samples onto a map to visualize geographic patterns of genetic differentiation. From these putatively neutral SNPs, we used observed F_{ST} on the autosomes (F_{aut}) to calculate the expected F_{ST} on X chromosomes (F_X) as in Machado *et al*. (2016) using the equation

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

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

We further investigated genetic variation in our dataset by principal component analysis (PCA) based on allele frequencies of the neutral marker SNPs described above. We used the *R* package *LEA* (v. 1.2.0., Frichot *et al.* 2013) and performed PCA on unscaled allele frequencies as suggested by Menozzi *et al*. (1978) and Novembre and Stephens (2008). We focused on the first three principal components (PCs) and employed a model-based approach as implemented in the *R* package *mclust* (v. 5.2., Fraley & Raftery 2012) to identify the most likely number of clusters based on maximum likelihood and assigned population samples to clusters by *k*-means clustering in *R* (R Development Core Team 2009). Finally, we examined the first three PCs for correlations with latitude, longitude,

1001 altitude, and season using general linear models and tested for spatial autocorrelation as
1002 described above. A Bonferroni-corrected α threshold (α'= 0.05/3 = 0.017) was used to
1003 account for multiple testing.
1004
1005 **Mitochondria**
1006 To obtain consensus mitochondrial sequences for each of the 48 European populations,
1007 reads from individual FASTQ files were aligned and minor variants replaced by the major
1008 variant using *Coral* (Salmela & Schröder 2011). This way, ambiguities that might prevent
1009 the growth of contigs from reads during the assembly process can be eliminated. For each
1010 population, a genome assembly was obtained using *SPAdes* using standard parameters
1011 and *k*-mers of size 21, 33, 55, and 77 (Bankevich *et al.* 2012) and the corrected FASTQ
1012 files. Mitochondrial contigs were retrieved by blastn, using the *D. melanogaster* NC
1013 024511 sequence as a query and each genome assembly as the database. To avoid
1014 nuclear mitochondrial DNA segments (numts), we ensured that only contigs with a much
1015 higher coverage than the average coverage of the genome were retrieved. When multiple
1016 contigs were available for the same region, the one with the highest coverage was
1017 selected. Possible contamination with *D. simulans* was assessed by looking for two or
1018 more consecutive sites that show the same variant as *D. simulans* and looking for
1019 alternative contigs for that region with similar coverage. As an additional quality control
1020 measure, we also examined the presence of pairs of sites showing four gametic types
1021 using *DNAsp* 6 (Rozas *et al.* 2017) – given that there is no recombination in mitochondrial
1022 DNA no such sites are expected. The very few sites presenting such features were
1023 rechecked by looking for alternative contigs for that region and were corrected if needed.
1024 The uncorrected raw reads for each population were mapped on top of the different
1025 consensus haplotypes using *Express* as implemented in *Trinity* (Grabherr *et al.* 2011). If

most reads for a given population mapped to the consensus sequence derived for that population the consensus sequence was retained, otherwise it was discarded as a possible chimera between different mitochondrial haplotypes. The repetitive mitochondrial hypervariable region is difficult to assemble and was therefore not used; the mitochondrial region was thus analyzed as in Cooper *et al*. (2015). Mitochondrial genealogy was estimated using statistical parsimony (TCS network; Clement *et al.* 2000), as implemented in *PopArt* (http://popart.otago.ac.nz), and the surviving mitochondrial haplotypes. Frequencies of the different mitochondrial haplotypes were estimated from FPKM values using the surviving mitochondrial haplotypes and expressed as implemented in *Trinity* (Grabherr *et al.* 2011).

Transposable elements

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

DNA present in multiple copies should be fully assembled (Goubert *et al.* 2015). We repeated this process with three iterations per sample, as recommended by the program guidelines, to assess the repeatability of the estimates.

We further estimated frequencies of previously characterized TEs present in the reference genome with *T-lex2* (v. 2.2.2., Fiston-Lavier *et al.* 2015), using all annotated TEs (5,416 TEs) in version 6.04 of the *Drosophila melanogaster* genome from flybase.org (Gramates *et al.* 2017). For 108 of these TEs, we used the corrected coordinates as described in Fiston-Lavier *et al*. (2015), based on the identification of target site duplications at the site of the insertion. We excluded TEs nested or flanked by other TEs (<100 bp on each side of the TE), and TEs which are part of segmental duplications, since *T-lex2* does not provide accurate frequency estimates in complex regions (Fiston-Lavier *et al.* 2015). We additionally excluded the INE-1 TE family, as this TE family is ancient, with thousands of insertions in the reference genome, which appear to be mostly fixed (2,234 TEs; Kapitonov & Jurka 2003).

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

with the cut reads.

To avoid inaccurate TE frequency estimates due to very low numbers of reads, we only considered frequency estimates based on at least 3 reads. Despite the stringency of *T-lex2* to select only high-quality reads, we additionally discarded frequency estimates supported by more than 90 reads, i.e. 3 times the average coverage of the sample with the lowest coverage (CH_Cha_14_43, Table 1), in order to avoid non-unique mapping reads. This filtering allows to estimate TE frequencies for ~96% (92.9% to 97.8%) of the TEs in each population. For 85% of the TEs, we were able to estimate their frequencies in more than 44 out of 48 *DrosEU* samples.

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

measures (see above) were considered as high recombination regions. To test TE family enrichment among the significant TEs we performed a χ^2 test and applied Yate's correction to account for the low number of some of the cells.

Inversion polymorphisms

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

Microbiome

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

identified by MGRAST for each taxon and sample was divided by the number of sequences that mapped to *D. melanogaster* chromosomes *X*, *Y, 2L*, *2R*, *3L*, *3R* and *4*.

We also surveyed the datasets for the presence of novel DNA viruses by performing *de novo* assembly of the non-fly reads using SPAdes 3.9.0 (Bankevich *et al.* 2012), and using conceptual translations to query virus proteins from Genbank using DIAMOND 'blastp' (Buchfink *et al.* 2015). In three cases (*Kallithea* virus, *Vesanto* virus, *Viltain* virus), reads from a single sample pool were sufficient to assemble a (near) complete genome. In two other cases, fragmentary assemblies allowed us to identify additional publicly available datasets that contained sufficient reads to complete the genomes (*Linvill Road* virus, *Esparto* virus; completed using SRA datasets SRR2396966 and SRR3939042, respectively). Novel viruses were provisionally named based on the localities where they were first detected, and the corresponding novel genome sequences were submitted to Genbank (KX130344, KY608910, KY457233, KX648533-KX648536). To assess the relative amount of viral DNA, unmapped (non-fly) reads from each sample pool were mapped to repeat-masked *Drosophila* DNA virus genomes using bowtie2, and coverage normalized relative to virus genome length and the number of mapped *Drosophila* reads.

Acknowledgments

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

1151

1152 # Additional information

1153 ## Funding

Funder	Grant reference number	Author
University of Freiburg Research Innovation Fund 2014		Fabian Staubach
Academy of Finland	#268241	Maaria Kankare
Academy of Finland	#272927	Maaria Kankare
Russian Foundation of Basic Research	#15-54-46009 CT_a	Elena G. Pasyukova
Danish Natural Science Research Council	4002-00113	Volker Loeschcke
Ministerio de Economia y Competitividad	CTM2017-88080 (AEI/FEDER, UE)	Marta Pascual
CNRS	UMR 9191	Catherine Montchamp-Moreau
Vetenskapsrådet	2011-05679	Jessica Abbott
Vetenskapsrådet	2015-04680	Jessica Abbott
Emmy Noether Programme of the Deutsche Forschungsgemeinschaft, DFG	PO 1648/3-1	Nico Posnien
National Institute of Health (NIH)	R35GM119686	Alan O. Bergland
Ministerio de Economia y Competitividad	CGL2013-42432-P	Maria Pilar Garcia Guerreiro

Scientific and Technological Research Council of Turkey (TUBITAK)	#214Z238	Banu Sebnem Onder
ANR Exhyb	14-CE19-0016	Cristina Vieira
Network of Excellence LifeSpan	FP6 036894	Bas J. Zwaan
IDEAL	FP7/2007-2011/259679	Bas J. Zwaan
Israel Science Foundation	1737/17	Eran Tauber
National Institute of Health (NIH)	R01GM100366	Paul S. Schmidt
Deutsche Forschungsgemeinschaft	PA 903/8-1	John Parsch
Austrian Science Fund (FWF)	P27048	Andrea J. Betancourt
Biotechnology and Biological Sciences Research Council (BBSRC)	BB/P00685X/1	Andrea J. Betancourt
Swiss National Science Foundation (SNSF)	PP00P3_133641	Thomas Flatt
Swiss National Science Foundation (SNSF)	PP00P3_165836	Thomas Flatt
European Comission H2020-ERC-2014	CoG-647900	Josefa Gonzàlez
Ministerio de Economia y Competitividad	FEDER BFU2014-57779-P	Josefa Gonzàlez

Author contributions

Martin Kapun, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Supervision, Methodology, Investigation, Data curation, Project administration, Validation, Resources, Software; Maite G. Barrón, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Data curation, Project administration, Validation, Resources, Software; Fabian Staubach, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Supervision, Funding acquisition, Methodology, Investigation, Data curation, Validation, Resources, Software; Jorge Vieira, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Validation, Resources; Darren J. Obbard, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Validation, Resources; Clément Goubert, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Investigation, Resources; Omar Rota-Stabelli, Visualization, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Resources; Maaria Kankare, Writing-original draft preparation, Conceptualization, Writing-review & editing, Methodology, Investigation, Resources; Annabelle Haudry, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Investigation, Validation, Resources; R. Axel W. Wiberg, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Methodology, Investigation, Resources, Software; Lena Waidele, Svitlana Serga, Patricia Gibert, Damiano Porcelli, Sonja Grath, Eliza Argyridou, Lain Guio, Mads Fristrup Schou, Conceptualization, Writing-review & editing, Investigation, Resources; Iryna Kozeretska, Conceptualization, Writing-

review & editing, Methodology, Investigation, Resources; Elena G. Pasyukova, Marta Pascual, Alan O. Bergland, Conceptualization, Writing-review & editing, Funding acquisition, Methodology, Investigation, Resources; Volker Loeschcke, Catherine Montchamp-Moreau, Jessica Abbott, Nico Posnien, Maria Pilar Garcia Guerreiro, Banu Sebnem Onder, Conceptualization, Writing-review & editing, Funding acquisition, Investigation, Resources; Cristina P. Vieira, Visualization, Formal analysis, Conceptualization, Writing-review & editing, Investigation, Resources; Élio Sucena, Conceptualization, Writing-review & editing, Methodology, Investigation, Project administration, Resources; Cristina Vieira, Michael G. Ritchie, Thomas Flatt, Josefa González, Writing-original draft preparation, Conceptualization, Writing-review & editing, Supervision, Funding acquisition, Methodology, Investigation, Project administration, Validation, Resources; Bart Deplancke, Conceptualization, Writing-review & editing, Funding acquisition, Investigation; Bas J. Zwaan, Visualization, Writing-original draft preparation, Conceptualization, Writing-review & editing, Supervision, Funding acquisition, Methodology, Investigation, Project administration; Eran Tauber, Writing-original draft preparation, Conceptualization, Writing-review & editing, Funding acquisition, Methodology, Investigation, Resources; Dorcas J. Orengo, Eva Puerma, Conceptualization, Writing-review & editing, Investigation, Validation, Resources; Montserrat Aguadé, Writing-original draft preparation, Conceptualization, Writing-review & editing, Methodology, Investigation, Validation, Resources; Paul S. Schmidt, John Parsch, Writing-original draft preparation, Conceptualization, Writing-review & editing, Funding acquisition, Methodology, Investigation, Validation, Resources; Andrea J. Betancourt, Writing-original draft preparation, Formal analysis, Conceptualization, Writing-review & editing, Supervision, Funding acquisition, Methodology, Investigation, Project administration, Validation, Resources

1211

1212 **Author ORCIDs**

Names	ORCID
Martin Kapun	0000-0002-3810-0504
Maite G. Barrón	0000-0001-6146-6259
Fabian Staubach	0000-0002-8097-2349
Jorge Vieira	0000-0001-7032-5220
Darren J. Obbard	0000-0001-5392-8142
Clément Goubert	0000-0001-8034-5559
Omar Rota-Stabelli	0000-0002-0030-7788
Maaria Kankare	0000-0003-1541-9050
Annabelle Haudry	0000-0001-6088-0909
R. Axel W. Wiberg	0000-0002-8074-8670
Lena Waidele	0000-0002-6323-6438
Iryna Kozeretska	0000-0002-6485-1408
Elena G. Pasyukova	0000-0002-6491-8561
Volker Loeschcke	0000-0003-1450-0754
Marta Pascual	0000-0002-6189-0612
Cristina P. Vieira	0000-0002-7139-2107
Svitlana Serga	0000-0003-1875-3185
Catherine Montchamp-Moreau	0000-0002-5044-9709
Jessica Abbott	0000-0002-8743-2089
Patricia Gibert	0000-0002-9461-6820
Damiano Porcelli	0000-0002-9019-5758
Nico Posnien	0000-0003-0700-5595

Sonja Grath	0000-0003-3621-736X
Élio Sucena	0000-0001-8810-870X
Alan O. Bergland	0000-0001-7145-7575
Maria Pilar Garcia Guerreiro	000-0001-9951-1879X
Banu Sebnem Onder	0000-0002-3003-248X
Eliza Argyridou	0000-0002-6890-4642
Lain Guio	0000-0002-5481-5200
Mads Fristrup Schou	0000-0001-5521-5269
Bart Deplancke	0000-0001-9935-843X
Cristina Vieira	0000-0003-3414-3993
Michael G. Ritchie	0000-0001-7913-8675
Bas J. Zwaan	0000-0002-8221-4998
Eran Tauber	0000-0003-4018-6535
Dorcas J. Orengo	0000-0001-7911-3224
Eva Puerma	0000-0001-7261-187X
Montserrat Aguadé	0000-0002-3884-7800
Paul S. Schmidt	0000-0002-8076-6705
John Parsch	0000-0001-9068-5549
Andrea J. Betancourt	0000-0001-9351-1413
Thomas Flatt	0000-0002-5990-1503
Josefa González	0000-0001-9824-027X

1213

1214

1215

1216

References

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

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

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

Anderson WW, Arnold J, Baldwin DG *et al.* (1991) Four decades of inversion polymorphism in *Drosophila pseudoobscura*. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 10367–10371.

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.

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

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. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.

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.

1240 Bartolomé C, Maside X, Charlesworth B (2002) On the Abundance and Distribution of
1241 Transposable Elements in the Genome of *Drosophila melanogaster*. *Molecular Biology*
1242 *and Evolution*, **19**, 926–937.
1243 Bastide H, Betancourt A, Nolte V *et al.* (2013) A genome-wide, fine-scale map of natural
1244 pigmentation variation in *Drosophila melanogaster*. (P Wittkopp, Ed,). *PLoS Genetics*,
1245 **9**, e1003534.
1246 Baudry E, Viginier B, Veuille M (2004) Non-African populations of *Drosophila*
1247 *melanogaster* have a unique origin. *Molecular Biology and Evolution*, **21**, 1482–1491.
1248 Becher PG, Flick G, Rozpędowska E *et al.* (2012) Yeast, not fruit volatiles mediate
1249 *Drosophila melanogaster* attraction, oviposition and development. *Functional Ecology*,
1250 **26**, 822–828.
1251 Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate
1252 with recombination rates in *D. melanogaster*. *Nature*, **356**, 519–520.
1253 Begun DJ, Aquadro CF (1993) African and North American populations of Drosophila
1254 melanogaster are very different at the DNA level. *Nature*, **365**, 548–550.
1255 Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population Genomics: Whole-Genome
1256 Analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biology*, **5**,
1257 e310.
1258 Behrman EL, Howick VM, Kapun M *et al.* (2018) Rapid seasonal evolution in innate
1259 immunity of wild *Drosophila melanogaster*. *Proceedings of the Royal Society B:*
1260 *Biological Sciences*, **285**, 20172599.
1261 Beisswanger S, Stephan W, De Lorenzo D (2006) Evidence for a Selective Sweep in the
1262 *wapl* Region of *Drosophila melanogaster*. *Genetics*, **172**, 265–274.

1263 Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA (2014) Genomic Evidence
1264 of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in *Drosophila*.
1265 *PLoS Genetics*, **10**, e1004775.
1266 Bergland AO, Tobler R, González J, Schmidt P, Petrov D (2016) Secondary contact and
1267 local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila*
1268 melanogaster. *Molecular Ecology*, **25**, 1157–1174.
1269 Betancourt AJ, Kim Y, Orr HA (2004) A pseudohitchhiking model of X vs. autosomal
1270 diversity. *Genetics*, **168**, 2261–2269.
1271 Betancourt AJ, Welch JJ, Charlesworth B (2009) Reduced effectiveness of selection
1272 caused by a lack of recombination. *Current biology*, **19**, 655–660.
1273 Bivand R, Piras G (2015) Comparing Implementations of Estimation Methods for Spatial
1274 Econometrics. *Journal of Statistical Software*, **63**, 1–36.
1275 Black WC IV, Black WC IV, Baer CF, Antolin MF, DuTeau NM (2001) Population
1276 genomics: genome-wide sampling of insect populations. *Annual Review of*
1277 *Entomology*.
1278 Blumenstiel JP, Chen X, He M, Bergman CM (2014) An Age-of-Allele Test of Neutrality for
1279 Transposable Element Insertions. *Genetics*, **196**, 523–538.
1280 Boitard S, Schlötterer C, Nolte V, Pandey RV, Futschik A (2012) Detecting Selective
1281 Sweeps from Pooled Next-Generation Sequencing Samples. *Molecular Biology and*
1282 *Evolution*, **29**, 2177–2186.
1283 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina
1284 sequence data. *Bioinformatics*, **30**, 2114–2120.
1285 Boussy IA, Itoh M, Rand D, Woodruff RC (1998) Origin and decay of the P element-
1286 associated latitudinal cline in Australian *Drosophila melanogaster*. *Genetica*, **104**, 45–
1287 57.

1288 Božičević V, Hutter S, Stephan W, Wollstein A (2016) Population genetic evidence for cold
1289 adaptation in European *Drosophila melanogaster* populations. *Molecular Ecology*, **25**,
1290 1175–1191.
1291 Braithwaite DP, Keegan KP *matR: Metagenomics Analysis Tools for R*. https://CRAN.R-
1292 project.org/package=matR.
1293 Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using
1294 DIAMOND. *Nature Methods*, **12**, 59–60.
1295 Buser CC, Newcomb RD, Gaskett AC, Goddard MR (2014) Niche construction initiates the
1296 evolution of mutualistic interactions. *Ecology Letters*, **17**, 1257–1264.
1297 Bushnell B (2016) *BBMap short read aligner*. URL http://sourceforge. net/projects/bbmap.
1298 Camus MF, Wolff JN, Sgrò CM, Dowling DK (2017) Experimental Support That Natural
1299 Selection Has Shaped the Latitudinal Distribution of Mitochondrial Haplotypes in
1300 Australian *Drosophila melanogaster. Molecular Biology and Evolution*, **34**, 2600–2612.
1301 Caracristi G, Schlötterer C (2003) Genetic Differentiation Between American and
1302 European *Drosophila melanogaster* Populations Could Be Attributed to Admixture of
1303 African Alleles. *Molecular Biology and Evolution*, **20**, 792–799.
1304 Casillas S, Barbadilla A (2017) Molecular Population Genetics. *Genetics*, **205**, 1003–1035.
1305 Catania F, Kauer MO, Daborn PJ *et al.* (2004) World-wide survey of an Accord insertion
1306 and its association with DDT resistance in *Drosophila melanogaster. Molecular*
1307 *Ecology*, **13**, 2491–2504.
1308 Cavalli-Sforza LL (1966) Population Structure and Human Evolution. *Proceedings of the*
1309 *Royal Society of London. Series B: Biological Sciences*, **164**, 362–379.
1310 Chandler JA, James PM (2013) Discovery of trypanosomatid parasites in globally
1311 distributed *Drosophila* species. *PLoS ONE*, **8**, e61937.

1312 Chandler JA, Eisen JA, Kopp A (2012) Yeast communities of diverse *Drosophila* species:
1313 comparison of two symbiont groups in the same hosts. *Applied and Environmental*
1314 *Microbiology*, **78**, 7327–7336.
1315 Chandler JA, Lang JM, Bhatnagar S, Eisen JA, Kopp A (2011) Bacterial communities of
1316 diverse *Drosophila* species: ecological context of a host-microbe model system. *PLoS*
1317 *Genetics*, **7**, e1002272.
1318 Charlesworth B (2010) Molecular population genomics: a short history. *Genetical*
1319 *Research*, **92**, 397–411.
1320 Charlesworth B (2015) Causes of natural variation in fitness: evidence from studies of
1321 *Drosophila* populations. *Proceedings of the National Academy of Sciences of the*
1322 *United States of America*, **112**, 1662–1669.
1323 Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive
1324 DNA in eukaryotes. *Nature*, **371**, 215–220.
1325 Cheng C, White BJ, Kamdem C *et al.* (2012) Ecological genomics of *Anopheles gambiae*
1326 along a latitudinal cline: a population-resequencing approach. *Genetics*, **190**, 1417–
1327 1432.
1328 Cingolani P, Platts A, Wang LL *et al.* (2012) A program for annotating and predicting the
1329 effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila
1330 melanogaster strain w1118; *iso-2*; *iso-3*. *Fly (Austin)*, **6**, 80–92.
1331 Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene
1332 genealogies. *Molecular Ecology*, **9**, 1657–1659.
1333 Clemente F, Vogl C (2012) Unconstrained evolution in short introns? – An analysis of
1334 genome-wide polymorphism and divergence data from *Drosophila*. *Journal of*
1335 *Evolutionary Biology*, **25**, 1975–1990.

1336 Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in
1337 *Drosophila melanogaster*. (DA Petrov, Ed,). *PLoS Genetics*, **8**, e1002905.
1338 Consortium T (2015) A global reference for human genetic variation. **526**, 68–74.
1339 Cooper BS, Burrus CR, Ji C, Hahn MW, Montooth KL (2015) Similar Efficacies of
1340 Selection Shape Mitochondrial and Nuclear Genes in Both *Drosophila melanogaster*
1341 and *Homo sapiens*. *G3 (Bethesda, Md.)*, **5**, 2165–2176.
1342 Corbett-Detig RB, Hartl DL (2012) Population Genomics of Inversion Polymorphisms in
1343 *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1003056.
1344 Cridland JM, Macdonald SJ, Long AD, Thornton KR (2013) Abundance and distribution of
1345 transposable elements in two *Drosophila* QTL mapping resources. *Molecular Biology*
1346 *and Evolution*, **30**, 2311–2327.
1347 Daborn PJ, Yen JL, Bogwitz MR *et al.* (2002) A single p450 allele associated with
1348 insecticide resistance in *Drosophila*. *Science*, **297**, 2253–2256.
1349 David JR, Capy P (1988) Genetic variation of *Drosophila melanogaster* natural
1350 populations. *Trends in genetics : TIG*, **4**, 106–111.
1351 de Jong G, Bochdanovits Z (2003) Latitudinal clines in *Drosophila melanogaster*: body
1352 size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway.
1353 *Journal of Genetics*, **82**, 207–223.
1354 Dieringer D, Nolte V, Schlötterer C (2005) Population structure in African *Drosophila*
1355 *melanogaster* revealed by microsatellite analysis. *Molecular Ecology*, **14**, 563–573.
1356 Dobzhansky T (1970) *Genetics of the Evolutionary Process*. Columbia University Press.
1357 Dray S, Dufour A-B (2007) The ade4 Package: Implementing the Duality Diagram for
1358 Ecologists. *Journal of Statistical Software*, **22**.

1359 Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S (2013) Demographic inference
1360 reveals African and European admixture in the North American *Drosophila*
1361 *melanogaster* population. *Genetics*, **193**, 291–301.
1362 Ellegren H (2014) Genome sequencing and population genomics in non-model organisms.
1363 *Trends in Ecology & Evolution*, **29**, 51–63.
1364 Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University
1365 Press.
1366 Fabian DK, Kapun M, Nolte V *et al.* (2012) Genome-wide patterns of latitudinal
1367 differentiation among populations of *Drosophila melanogaster* from North America.
1368 *Molecular Ecology*, **21**, 4748–4769.
1369 Fabian DK, Lack JB, Mathur V *et al.* (2015) Spatially varying selection shapes life history
1370 clines among populations of *Drosophila melanogaster* from sub-Saharan Africa.
1371 *Journal of Evolutionary Biology*, **28**, 826–840.
1372 Fiston-Lavier A-S, Barrón MG, Petrov DA, González J (2015) T-lex2: genotyping,
1373 frequency estimation and re-annotation of transposable elements using single or
1374 pooled next-generation sequencing data. *Nucleic Acids Research*, **43**, e22–e22.
1375 Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA (2010) *Drosophila melanogaster*
1376 recombination rate calculator. *Gene*, **463**, 18–20.
1377 Flatt T, Amdam GV, Kirkwood TBL, Omholt SW (2013) Life-history evolution and the
1378 polyphenic regulation of somatic maintenance and survival. *The quarterly review of*
1379 *biology*, **88**, 185–218.
1380 Fraley C, Raftery AE (2012) *mclust Version 4 for R: Normal Mixture Modeling for Model-*
1381 *Based Clustering, Classification, and Density Estimation*. Seattle.
1382 Francalacci P, Sanna D (2008) History and geography of human Y-chromosome in
1383 Europe: a SNP perspective. *Journal of anthropological sciences*, **86**, 59–89.

1384 Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between
1385 loci and environmental gradients using latent factor mixed models. *Molecular Biology*
1386 *and Evolution*, **30**, 1687–1699.
1387 Futschik A (2010) The next generation of molecular markers from massively parallel
1388 sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.
1389 González J, Karasov TL, Messer PW, Petrov DA (2010) Genome-Wide Patterns of
1390 Adaptation to Temperate Environments Associated with Transposable Elements in
1391 *Drosophila* (HS Malik, Ed,). *PLoS Genetics*, **6**, e1000905.
1392 González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA (2008) High Rate of Recent
1393 Transposable Element–Induced Adaptation in *Drosophila melanogaster*. *PLoS Biology*,
1394 **6**, e251.
1395 Goubert C, Modolo L, Vieira C *et al.* (2015) De Novo Assembly and Annotation of the
1396 Asian Tiger Mosquito (*Aedes albopictus*) Repeatome with dnaPipeTE from Raw
1397 Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (*Aedes*
1398 *aegypti*). *Genome Biology and Evolution*, **7**, 1192–1205.
1399 Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from
1400 RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.
1401 Gramates LS, Marygold SJ, Santos GD *et al.* (2017) FlyBase at 25: looking to the future.
1402 *Nucleic Acids Research*, **45**, D663–D671.
1403 Green RM, Smart WM (1985) *Textbook on Spherical Astronomy*. Cambridge University.
1404 Grenier JK, Arguello JR, Moreira MC *et al.* (2015) Global Diversity Lines-A Five-Continent
1405 Reference Panel of Sequenced *Drosophila melanogaster* Strains. *G3 (Bethesda, Md.)*,
1406 **5**, 593–603.

1407 Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P (2005) Patterns of intron sequence
1408 evolution in *Drosophila* are dependent upon length and GC content. *Genome Biology*,
1409 **6**, R67.
1410 Hales KG, Korey CA, Larracuente AM, Roberts DM (2015) Genetics on the Fly: A Primer
1411 on the Drosophila Model System. *Genetics*, **201**, 815–842.
1412 Hamilton PT, Votýpka J, Dostálová A *et al.* (2015) Infection Dynamics and Immune
1413 Response in a Newly Described *Drosophila*-Trypanosomatid Association. *mBio*, **6**,
1414 e01356–15.
1415 Handu M, Kaduskar B, Ravindranathan R *et al.* (2015) SUMO-Enriched Proteome for
1416 *Drosophila* Innate Immune Response. *G3 (Bethesda, Md.)*, **5**, 2137–2154.
1417 Harpur BA, Kent CF, Molodtsova D *et al.* (2014) Population genomics of the honey bee
1418 reveals strong signatures of positive selection on worker traits. *Proceedings of the*
1419 *National Academy of Sciences of the United States of America*, **111**, 2614–2619.
1420 Haselkorn TS, Markow TA, Moran NA (2009) Multiple introductions of the *Spiroplasma*
1421 bacterial endosymbiont into *Drosophila*. *Molecular Ecology*, **18**, 1294–1305.
1422 Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel
1423 Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genetics*, **6**,
1424 e1000862.
1425 Hu TT, Eisen MB, Thornton KR, Andolfatto P (2013) A second-generation assembly of the
1426 *Drosophila simulans* genome provides new insights into patterns of lineage-specific
1427 divergence. *Genome Research*, **23**, 89–98.
1428 Huang W, Massouras A, Inoue Y *et al.* (2014) Natural variation in genome architecture
1429 among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome*
1430 *Research*, **24**, 1193–1208.

1431 Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on
1432 nucleotide data. *Genetics*, **116**, 153–159.
1433 Huszar T, Imler J-L (2008) *Drosophila* viruses and the study of antiviral host-defense.
1434 *Advances in virus research*, **72**, 227–265.
1435 Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W (2007) Distinctly Different Sex
1436 Ratios in African and European Populations of *Drosophila melanogaster* Inferred From
1437 Chromosomewide Single Nucleotide Polymorphism Data. *Genetics*, **177**, 469–480.
1438 Jorde LB, Watkins WS, Bamshad MJ (2001) Population genomics: a bridge from
1439 evolutionary history to genetic medicine. *Human Molecular Genetics*, **10**, 2199–2207.
1440 Kao JY, Zubair A, Salomon MP, Nuzhdin SV, Campo D (2015) Population genomic
1441 analysis uncovers African and European admixture in *Drosophila melanogaster*
1442 populations from the south-eastern United States and Caribbean Islands. *Molecular*
1443 *Ecology*, **24**, 1499–1509.
1444 Kapitonov VV, Jurka J (2003) Molecular Paleontology of Transposable Elements in the
1445 *Drosophila melanogaster* Genome. *Proceedings of the National Academy of Sciences*
1446 *of the United States of America*, **100**, 6569–6574.
1447 Kapun M, Fabian DK, Goudet J, Flatt T (2016a) Genomic Evidence for Adaptive Inversion
1448 Clines in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **33**, 1317–1336.
1449 Kapun M, Schmidt C, Durmaz E, Schmidt PS, Flatt T (2016b) Parallel effects of the
1450 inversion *In(3R)Payne* on body size across the North American and Australian clines in
1451 *Drosophila melanogaster. Journal of Evolutionary Biology*, **29**, 1059–1072.
1452 Kapun M, van Schalkwyk H, McAllister B, Flatt T, Schlötterer C (2014) Inference of
1453 chromosomal inversion dynamics from Pool-Seq data in natural and laboratory
1454 populations of *Drosophila melanogaster. Molecular Ecology*, **23**, 1813–1827.

1455 Kassis JA, Kennison JA, Tamkun JW (2017) Polycomb and Trithorax Group Genes in
1456 *Drosophila*. *Genetics*, **206**, 1699–1725.
1457 Kauer M, Zangerl B, Dieringer D, Schlötterer C (2002) Chromosomal patterns of
1458 microsatellite variability contrast sharply in African and non-African populations of
1459 *Drosophila melanogaster*. *Genetics*, **160**, 247–256.
1460 Keller A (2007) *Drosophila melanogaster*'s history as a human commensal. *Current*
1461 *Biology*, **17**, R77–R81.
1462 Kimura M (1984) *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
1463 Knibb WR (1982) Chromosome inversion polymorphisms in *Drosophila melanogaster* II.
1464 Geographic clines and climatic associations in Australasia, North America and Asia.
1465 *Genetica*, **58**, 213–221.
1466 Knibb WR (1986) Temporal variation of *Drosophila melanogaster Adh* allele frequencies,
1467 inversion freqencies, and population sizes. *Genetica*, **71**, 175–190.
1468 Knibb WR, Oakeshott JG, Gibson JB (1981) Chromosome Inversion Polymorphisms in
1469 *Drosophila melanogaster*. I. Latitudinal Clines and Associations between Inversions in
1470 Australasian Populations. *Genetics*, **98**, 833–847.
1471 Kofler R, Betancourt AJ, Schlötterer C (2012) Sequencing of pooled DNA samples (Pool-
1472 Seq) uncovers complex dynamics of transposable element insertions in *Drosophila*
1473 *melanogaster. PLoS Genetics*, **8**, e1002487.
1474 Kofler R, Orozco-terWengel P, De Maio N *et al.* (2011) PoPoolation: A Toolbox for
1475 Population Genetic Analysis of Next Generation Sequencing Data from Pooled
1476 Individuals. *PLoS ONE*, **6**, e15925.
1477 Kolaczkowski B, Hupalo DN, Kern AD (2011a) Recurrent adaptation in RNA interference
1478 genes across the *Drosophila* phylogeny. *Molecular Biology and Evolution*, **28**, 1033–
1479 1042.

1480 Kolaczkowski B, Kern AD, Holloway AK, Begun DJ (2011b) Genomic Differentiation
1481 Between Temperate and Tropical Australian Populations of *Drosophila melanogaster*.
1482 *Genetics*, **187**, 245–260.
1483 Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R (2013) Calculation of Tajima's *D* and
1484 other neutrality test statistics from low depth next-generation sequencing data. *BMC*
1485 *Bioinformatics*, **14**, 289.
1486 Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of
1487 Drosophila melanogaster. *Nature*, **304**, 412–417.
1488 Kriesner P, Conner WR, Weeks AR, Turelli M, Hoffmann AA (2016) Persistence of a
1489 *Wolbachia* infection frequency cline in *Drosophila melanogaster* and the possible role
1490 of reproductive dormancy. *Evolution*, **70**, 979–997.
1491 Kühn I, Dormann CF (2012) Less than eight (and a half) misconceptions of spatial
1492 analysis. *Journal of Biogeography*, **39**, 995–998.
1493 Lachaise D, Cariou M-L, David JR *et al.* (1988) Historical Biogeography of the *Drosophila*
1494 *melanogaster* Species Subgroup. In: *Evolutionary Biology*, pp. 159–225. Springer,
1495 Boston, MA, Boston, MA.
1496 Lack JB, Cardeno CM, Crepeau MW *et al.* (2015) The *Drosophila* genome nexus: a
1497 population genomic resource of 623 *Drosophila melanogaster* genomes, including 197
1498 from a single ancestral range population. *Genetics*, **199**, 1229–1241.
1499 Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE (2016) A Thousand Fly
1500 Genomes: An Expanded *Drosophila* Genome Nexus. *Molecular Biology and Evolution*,
1501 **33**, msw195–3313.
1502 Lang DT *RJSONIO: Serialize R objects to JSON, JavaScript Object Notation*.
1503 https://CRAN.R-project.org/package=RJSONIO.

1504 Langley CH, Stevens K, Cardeno C *et al.* (2012) Genomic variation in natural populations
1505 of *Drosophila melanogaster*. *Genetics*, **192**, 533–598.
1506 Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature*
1507 *Methods*, **9**, 357–359.
1508 Lawrie DS, Messer PW, Hershberg R, Petrov DA (2013) Strong Purifying Selection at
1509 Synonymous Sites in *D. melanogaster*. *arXiv.org*, **q-bio.PE**, e1003527.
1510 Levins R (1968) *Evolution in Changing Environments*. Princeton University Press.
1511 Lewontin RC (1974) *The Genetic Basis of Evolutionary Change*. Columbia University
1512 Press.
1513 Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-
1514 MEM. *arXiv.org*.
1515 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler
1516 transform. *Bioinformatics*, **25**, 1754–1760.
1517 Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants
1518 using mapping quality scores. *Genome Research*, **18**, 1851–1858.
1519 Luikart G, England PR, Tallmon D, Jordan S, Tableret P (2003) The power and promise of
1520 population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**,
1521 981–994.
1522 Lyne R, Smith R, Rutherford K *et al.* (2007) FlyMine: an integrated database for
1523 *Drosophila* and *Anopheles* genomics. *Genome Biology*, **8**, R129.
1524 Machado HE, Bergland AO, O'Brien KR *et al.* (2016) Comparative population genomics of
1525 latitudinal variation in *Drosophila simulans* and *Drosophila melanogaster*. *Molecular*
1526 *Ecology*, **25**, 723–740.
1527 Mackay TFC, Richards S, Stone EA *et al.* (2012) The *Drosophila melanogaster* Genetic
1528 Reference Panel. *Nature*, **482**, 173–178.

1529 Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing
1530 reads. *EMBnet.journal*, **17**, 10–12.
1531 Martino ME, Ma D, Leulier F (2017) Microbial influence on *Drosophila* biology. *Current*
1532 *Opinion in Microbiology*, **38**, 165–170.
1533 Mateo L, Rech GE, González J (2018) Genome-wide patterns of local adaptation in
1534 *Drosophila melanogaster*: adding intra European variability to the map. *bioRxiv*. DOI:
1535 10.1101/269332
1536 Matthias P, Yoshida M, Khochbin S (2008) HDAC6 a new cellular stress surveillance
1537 factor. *Cell Cycle*, **7**, 7–10.
1538 McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in
1539 Drosophila. *Nature*, **351**, 652–654.
1540 McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: A MapReduce
1541 framework for analyzing next-generation DNA sequencing data. *Genome Research*,
1542 **20**, 1297–1303.
1543 Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies
1544 in Europeans. *Science*, **201**, 786–792.
1545 Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective
1546 sweeps. **28**, 659–669.
1547 Mettler LE, Voelker RA, Mukai T (1977) Inversion Clines in Populations of *Drosophila*
1548 *melanogaster. Genetics*, **87**, 169–176.
1549 Meyer F, Paarmann D, D'Souza M *et al.* (2008) The metagenomics RAST server - a public
1550 resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC*
1551 *Bioinformatics*, **9**, 386.
1552 Micallef L, Rodgers P (2014) euler*APE*: drawing area-proportional 3-Venn diagrams using
1553 ellipses. *PLoS ONE*, **9**, e101717.

1554 Michalakis Y, Veuille M (1996) Length variation of CAG/CAA trinucleotide repeats in
1555 natural populations of *Drosophila melanogaster* and its relation to the recombination
1556 rate. *Genetics*, **143**, 1713–1725.
1557 Moran PAP (1950) Notes on Continuous Stochastic Phenomena. *Biometrika*, **37**, 17.
1558 Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press.
1559 Nolte V, Pandey RV, Kofler R, Schlötterer C (2013) Genome-wide patterns of natural
1560 variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila*
1561 *mauritiana*. *Genome Research*, **23**, 99–110.
1562 Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial
1563 population genetic variation. *Nature Genetics*, **40**, 646–649.
1564 Nunes MDS, Neumeier H, Schlötterer C (2008) Contrasting patterns of natural variation in
1565 global *Drosophila melanogaster* populations. *Molecular Ecology*, **17**, 4470–4479.
1566 Okonechnikov K, Conesa A, García-Alcalde F (2016) Qualimap 2: advanced multi-sample
1567 quality control for high-throughput sequencing data. *Bioinformatics*, **32**, 292–294.
1568 Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P (2010) On the utility
1569 of short intron sequences as a reference for the detection of positive and negative
1570 selection in *Drosophila*. *Molecular Biology and Evolution*, **27**, 1226–1234.
1571 Peel MC, Finlayson BL, McMahon TA (2007) Updated world map of the Köppen-Geiger
1572 climate classification. *Hydrology and Earth System Sciences*, **11**, 1633–1644.
1573 Petrov DA, Fiston-Lavier AS, Lipatov M, Lenkov K, González J (2011) Population
1574 Genomics of Transposable Elements in *Drosophila melanogaster*. *Molecular Biology*
1575 *and Evolution*, **28**, 1633–1644.
1576 Pool JE (2015) The Mosaic Ancestry of the *Drosophila* Genetic Reference Panel and the
1577 *D. melanogaster* Reference Genome Reveals a Network of Epistatic Fitness
1578 Interactions. *Molecular Biology and Evolution*, **32**, 3236–3251.

1579 Pool JE, Nielsen R (2007) Population size changes reshape genomic patterns of diversity.
1580 *Evolution*, **61**, 3001–3006.
1581 Pool JE, Braun DT, Lack JB (2016) Parallel Evolution of Cold Tolerance Within *Drosophila*
1582 *melanogaster*. *Molecular Biology and Evolution*, msw232.
1583 Pool JE, Corbett-Detig RB, Sugino RP *et al.* (2012) Population Genomics of Sub-Saharan
1584 *Drosophila melanogaster*: African Diversity and Non-African Admixture. *PLoS*
1585 *Genetics*, **8**, e1003080.
1586 Powell JR (1997) *Progress and Prospects in Evolutionary Biology: The Drosophila Model*.
1587 Oxford University Press.
1588 R Development Core Team (2009) R: A Language and Environment for Statistical
1589 Computing. *R-project.org*.
1590 Rako L, Anderson AR, Sgrò CM, Stocker AJ, Hoffmann AA (2006) The association
1591 between inversion *In(3R)Payne* and clinally varying traits in *Drosophila melanogaster.*
1592 *Genetica*, **128**, 373–384.
1593 Ramachandran S, Rosenberg NA, Zhivotovsky LA, Feldman MW (2004) Robustness of
1594 the inference of human population structure: a comparison of X-chromosomal and
1595 autosomal microsatellites. *Human genomics*, **1**, 87–97.
1596 Rane RV, Rako L, Kapun M, LEE SF (2015) Genomic evidence for role of inversion *3RP*
1597 of *Drosophila melanogaster* in facilitating climate change adaptation. *Molecular*
1598 *Ecology*, **24**, 2423–2432.
1599 Reinhardt JA, Kolaczkowski B, Jones CD, Begun DJ, Kern AD (2014) Parallel Geographic
1600 Variation in *Drosophila melanogaster*. *Genetics*, **197**, 361–373.
1601 Richardson JL, Urban MC, Bolnick DI, Skelly DK (2014) Microgeographic adaptation and
1602 the spatial scale of evolution. *Trends in Ecology & Evolution*, **29**, 165–176.

1603 Richardson MF, Weinert LA, Welch JJ *et al.* (2012) Population Genomics of the *Wolbachia*
1604 Endosymbiont in *Drosophila melanogaster* (A Kopp, Ed,). *PLoS Genetics*, **8**,
1605 e1003129.
1606 Rogers RL, Hartl DL (2012) Chimeric genes as a source of rapid evolution in *Drosophila*
1607 *melanogaster*. *Molecular Biology and Evolution*, **29**, 517–529.
1608 Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC *et al.* (2017) DnaSP 6: DNA Sequence
1609 Polymorphism Analysis of Large Datasets. *Molecular Biology and Evolution*, **34**, 3299–
1610 3302.
1611 Salmela L, Schröder J (2011) Correcting errors in short reads by multiple alignments.
1612 *Bioinformatics*, **27**, 1455–1461.
1613 Schlenke TA, Begun DJ (2003) Natural selection drives *Drosophila* immune system
1614 evolution., **164**, 1471–1480.
1615 Schlötterer C, Neumeier H, Sousa C, Nolte V (2006) Highly structured Asian *Drosophila*
1616 *melanogaster* populations: a new tool for hitchhiking mapping? *Genetics*, **172**, 287–
1617 292.
1618 Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals - mining
1619 genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**,
1620 749–763.
1621 Schmidt JM, Good RT, Appleton B *et al.* (2010) Copy number variation and transposable
1622 elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genetics*, **6**,
1623 e1000998.
1624 Schmidt PS, Paaby AB (2008) Reproductive Diapause and Life-History Clines in North
1625 American Populations of *Drosophila melanogaster*. *Evolution*, **62**, 1204–1215.
1626 Schmidt PS, Zhu CT, Das J *et al.* (2008) An amino acid polymorphism in the *couch potato*
1627 gene forms the basis for climatic adaptation in *Drosophila melanogaster*. *Proceedings*

1628 *of the National Academy of Sciences of the United States of America*, **105**, 16207–
1629 16211.
1630 Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive Natural Selection in the
1631 Drosophila Genome? *PLoS Genetics*, **5**, e1000495.
1632 Singh ND, Arndt PF, Clark AG, Aquadro CF (2009) Strong evidence for lineage and
1633 sequence specificity of substitution rates and patterns in *Drosophila*. *Molecular Biology*
1634 *and Evolution*, **26**, 1591–1605.
1635 Staubach F, Baines JF, Künzel S, Bik EM, Petrov DA (2013) Host species and
1636 environmental effects on bacterial communities associated with *Drosophila* in the
1637 laboratory and in the natural environment. *PLoS ONE*, **8**, e70749.
1638 Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and
1639 its implications. *Philosophical Transactions Of The Royal Society Of London Series B-*
1640 *Biological Sciences*, **365**, 1245–1253.
1641 Svetec N, Pavlidis P, Stephan W (2009) Recent strong positive selection on *Drosophila*
1642 *melanogaster HDAC6*, a gene encoding a stress surveillance factor, as revealed by
1643 population genomic analysis. *Molecular Biology and Evolution*, **26**, 1549–1556.
1644 Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA
1645 polymorphism. *Genetics*, **123**, 585–595.
1646 Trinder M, Daisley BA, Dube JS, Reid G (2017) *Drosophila melanogaster* as a High-
1647 Throughput Model for Host-Microbiota Interactions. *Frontiers in microbiology*, **8**, 751.
1648 Turner TL, Levine MT, Eckert ML, Begun DJ (2008) Genomic analysis of adaptive
1649 differentiation in *Drosophila melanogaster*. *Genetics*, **179**, 455–473.
1650 Umina PA, Weeks AR, Kearney MR, McKechnie SW, Hoffmann AA (2005) A rapid shift in
1651 a classic clinal pattern in *Drosophila* reflecting climate change. *Science*, **308**, 691–693.
1652 Unckless RL (2011) A DNA virus of *Drosophila*. *PLoS ONE*, **6**, e26564.

1653 Voelker RA, Cockerham CC, Johnson FM *et al.* (1978) Inversions fail to account for
1654 allozyme clines., **88**, 515–527.
1655 Watterson GA (1975) On the number of segregating sites in genetical models without
1656 recombination. *Theoretical Population Biology*, **7**, 256–276.
1657 Webster CL, Longdon B, Lewis SH, Obbard DJ (2016) Twenty-Five New Viruses
1658 Associated with the Drosophilidae (Diptera). *Evolutionary bioinformatics online*, **12**,
1659 13–25.
1660 Webster CL, Waldron FM, Robertson S *et al.* (2015) The Discovery, Distribution, and
1661 Evolution of Viruses Associated with *Drosophila melanogaster. PLoS Biology*, **13**,
1662 e1002210.
1663 Weir BS, Cockerham CC (1984) Estimating *F*-Statistics for the Analysis of Population
1664 Structure. *Evolution*, **38**, 1358–1370.
1665 Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate
1666 biology. *Nature Reviews Microbiology*, **6**, 741–751.
1667 Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer.
1668 Wilfert L, Longdon B, Ferreira AGA, Bayer F, Jiggins FM (2011) Trypanosomatids are
1669 common and diverse parasites of *Drosophila*. *Parasitology*, **138**, 858–865.
1670 Wittmann MJ, Bergland AO, Feldman MW, Schmidt PS, Petrov DA (2017) Seasonally
1671 fluctuating selection can maintain polymorphism at many loci via segregation lift.
1672 *Proceedings of the National Academy of Sciences of the United States of America*,
1673 **114**, E9932–E9941.
1674 Wolf JBW, Bayer T, Haubold B *et al.* (2010) Nucleotide divergence vs. gene expression
1675 differentiation: comparative transcriptome sequencing in natural isolates from the
1676 carrion crow and its hybrid zone with the hooded crow. *Molecular Ecology*, **19**, 162–
1677 175.

1678 Wolff JN, Camus MF, Clancy DJ, Dowling DK (2016) Complete mitochondrial genome
1679 sequences of thirteen globally sourced strains of fruit fly (*Drosophila melanogaster*)
1680 form a powerful model for mitochondrial research. *Mitochondrial DNA. Part A, DNA*
1681 *mapping, sequencing, and analysis*, **27**, 4672–4674.
1682 Xiao F-X, Yotova V, Zietkiewicz E *et al.* (2004) Human *X*-chromosomal lineages in Europe
1683 reveal Middle Eastern and Asiatic contacts. *European Journal of Human Genetics*, **12**,
1684 301–311.
1685 Yukilevich R, True JR (2008a) Incipient sexual isolation among cosmopolitan *Drosophila*
1686 *melanogaster* populations. *Evolution*, **62**, 2112–2121.
1687 Yukilevich R, True JR (2008b) African morphology, behavior and phermones underlie
1688 incipient sexual isolation between us and Caribbean Drosophila melanogaster.
1689 *Evolution*, **62**, 2807–2828.
1690 Yukilevich R, Turner TL, Aoki F, Nuzhdin SV, True JR (2010) Patterns and processes of
1691 genome-wide divergence between North American and African *Drosophila*
1692 *melanogaster*. *Genetics*, **186**, 219–239.
1693 Zanini F, Brodin J, Thebo L *et al.* (2015) Population genomics of intrapatient HIV-1
1694 evolution. (AK Chakraborty, Ed,). *eLife*, **4**, e11282.
1695

Figure legends

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

Figure 2. The geographic distribution of population samples. The map shows the geographic locations of all samples in the 2014 *DrosEU* dataset. The color of the circles indicates the sampling season for each location.

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

Figure 4. SNP sharing. (A) Shared SNPs among *DrosEU* samples. Number and proportion of SNPs in different samples, ranging from one specific sample to being shared

among all 48 samples. (B) Shared SNPs among three worldwide populations. Elliptic Venn diagram showing the number and proportion of SNPs overlapping among the 5,361,256 biallelic SNPs in *DrosEU* (Europe), 3,953,804 biallelic SNPs in DGRP (North America) and 4,643,511 biallelic SNPs in Zambia (Africa) populations.

Figure 5. Genome-wide estimates of genetic diversity and recombination rates. The distribution of Tajima's π, Watterson's θ and Tajima's *D* (from top to bottom) in 200 kb non-overlapping windows plotted for each chromosomal arm separately. Bold black lines depict statistics which were averaged across all 48 samples and the upper and lower grey area show the corresponding standard deviations for each window. Red dashed lines highlight the vertical position of a zero value. The bottom row shows log-transformed recombination rates (*r*) in 100kb non-overlapping windows as obtained from Comeron *et al.* (2010).

Figure 6. Signals of selective sweeps. The central figure shows the distribution of Tajima's *D* in 50 kb sliding windows with 40 kb overlap. The red and green dashed lines highlight Tajima's *D* = 0 and Tajima's *D* = -1 respectively. The top panel shows the magnification of a genomic region on chromosomal arm *2R* that harbors well-known candidate loci for pesticide resistance, *Cyp6g1* and *Hen1* (highlighted in red), where strong selection resulted in a selective sweep. This sweep is characterized by an excess of low-frequency SNP variants, indicated by an overall negative Tajima's *D* in all samples. The colored solid lines depict Tajima's *D* for each sample separately, whereas the black dashed line shows Tajima's *D* averaged across all samples. (A legend for the color codes of the samples can be found in the Supporting Information file in Figure S19). The bottom figure shows a genomic region on *3L* which has not been identified as a potential target of

selection but shows Tajima's *D* patterns similar to the top figure. Notably, both regions are also characterized by a strong reduction of genetic variation (Figure S10).

Figure 7. Genetic differentiation among European populations. (A) isolation by distance estimated by average genetic differentiation (F_{ST}) of 21,008 SNPs located in short introns (<60 bp) plotted against geographic distance. Mantel tests and linear regression (red dashed line and statistics in upper left box) indicate significance. (B) Average neutral F_{ST} among populations. The center plot shows the distribution of average neutral F_{ST} values for all 1,128 pairwise combinations. Mean neutral F_{ST} values were calculated by averaging individual F_{ST} values from 20,008 genome-wide intronic SNPs for each pairwise comparison. The plots on the left and the right show population pairs in the lower (blue) and upper (red) 5% tails of the F_{ST} distribution. (C) Population structure of all *DrosEU* samples as determined by PCA of allele frequencies of 20,008 SNPs located in short introns (< 60 bp). The optimal number of five clusters was estimated by hierarchical model fitting using the first four principal components. Cluster assignment of each population, which was estimated by *k*-means clustering, is indicated by color.

Figure 8. Mitochondria. (A) TCS network showing the relationship of 5 common mitochondrial haplotypes; (B) estimated frequency of each mitochondrial haplotype in 48 European populations;

Figure 9. Transposable elements. Relative abundances of repeats among pools**.** Proportion of each repeat class was estimated from sampled reads with dnaPipTE (2 samples per run, 0.1X coverage per sample).

Figure 10. Distribution of inversion frequencies. Cumulative bar plots showing the absolute frequencies of six cosmopolitan inversions (*In(2L)t, In(2R)NS, In(3L)P, In(3R)C, In(3R)Mo, In(3R)Payne)* in all 48 population samples of the *DrosEU* dataset.

Figure 11. Microbiome. Relative abundance of *Drosophila* associated microbes as assessed by MGRAST classified shotgun sequences. Microbes had to reach at least 3% relative abundance in one of the samples to be presented.

Tables

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

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

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

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

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

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

PC3	0.14	0.11	0.01	1.29	0.22

1792

1793

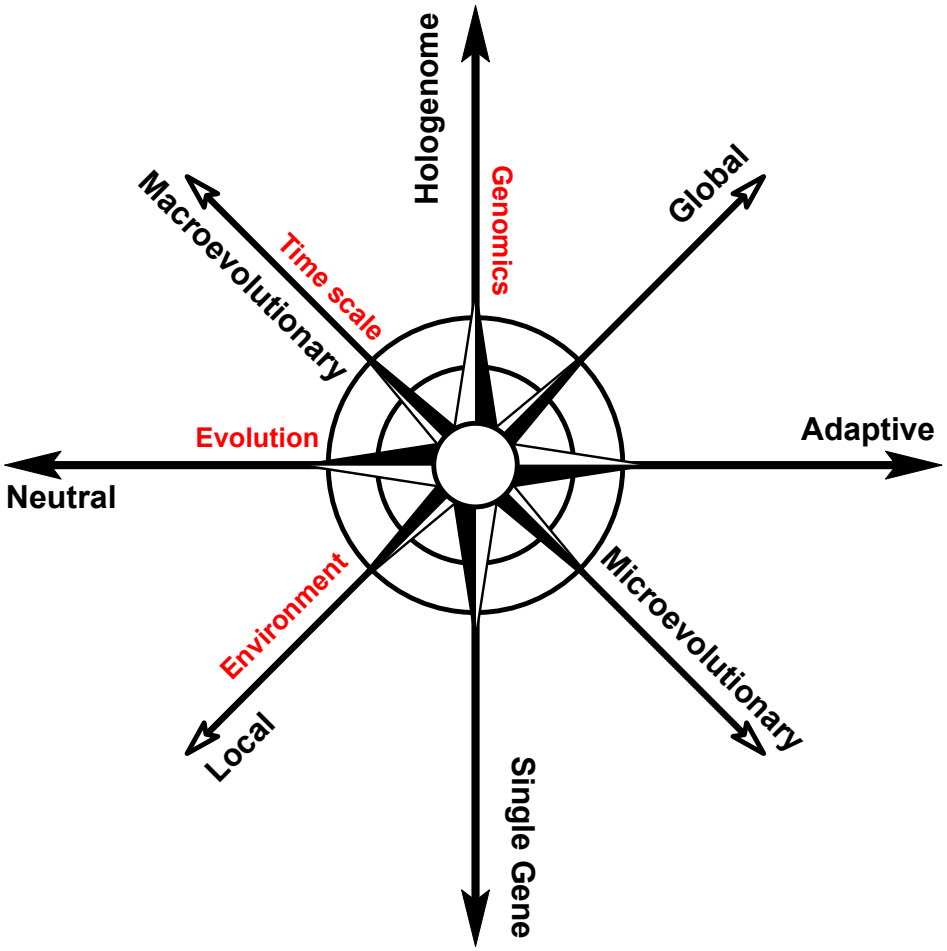
Table 3. **Clinality and/or seasonality of chromosomal inversions.** The values represent *F*-ratios from generalized linear models with a binomial error structure to account for frequency data. Bold type indicates deviance values that were significant after Bonferroni correction (adjusted *α'*=0.0071). 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

1801 **Figures**

1802

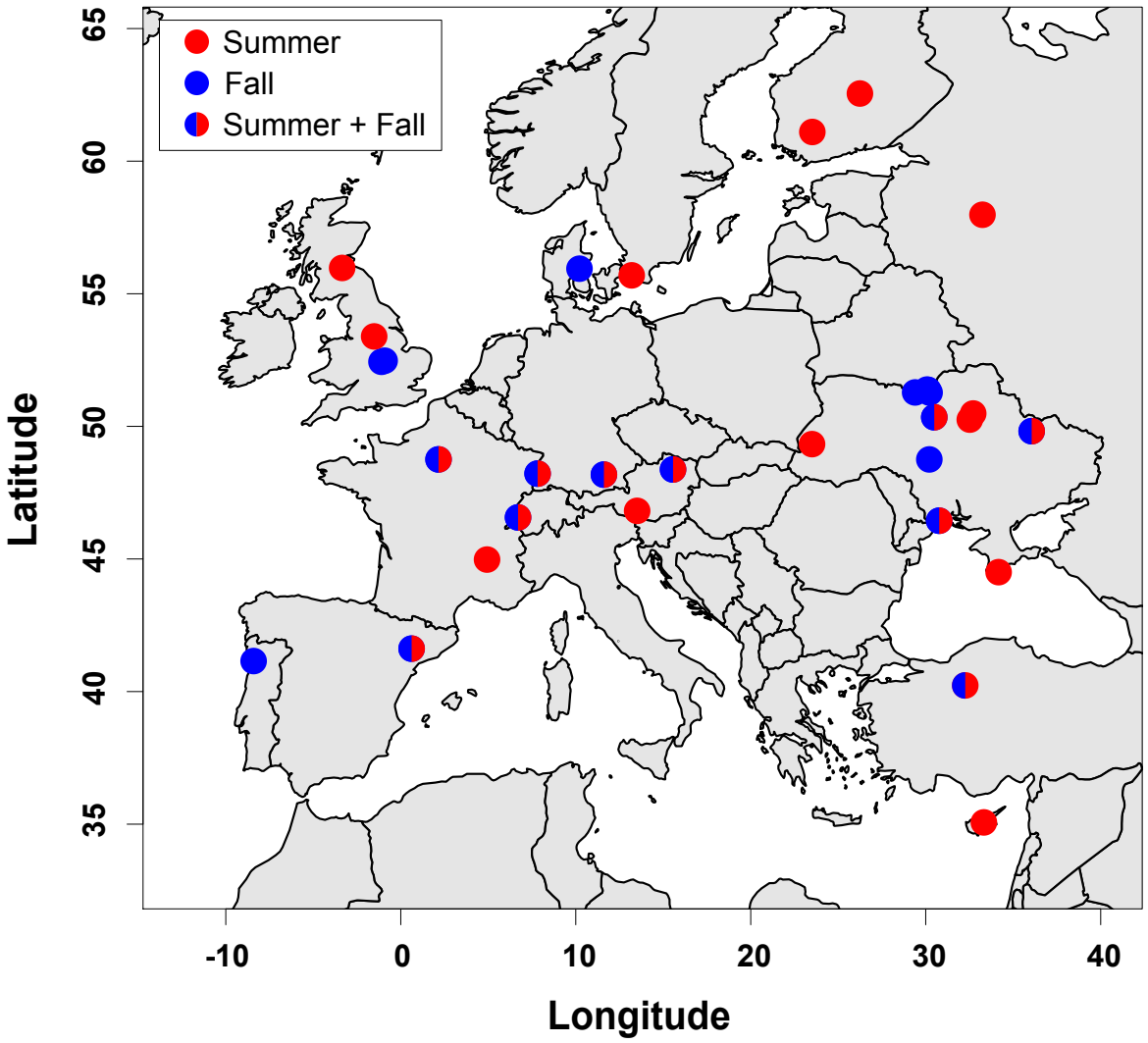
1803 **Figure 1**



1804

1805

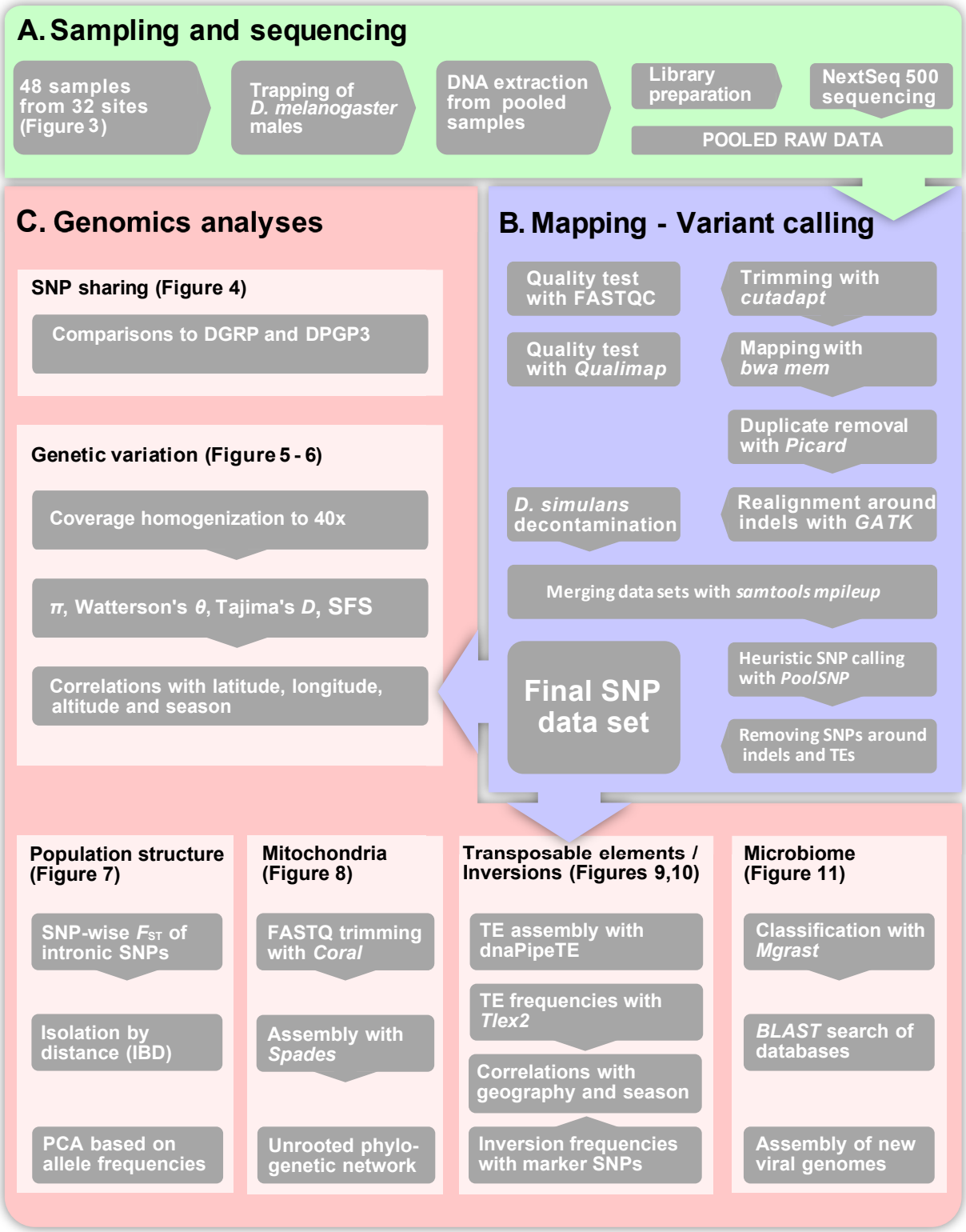
1806 **Figure 2**



1807

1808

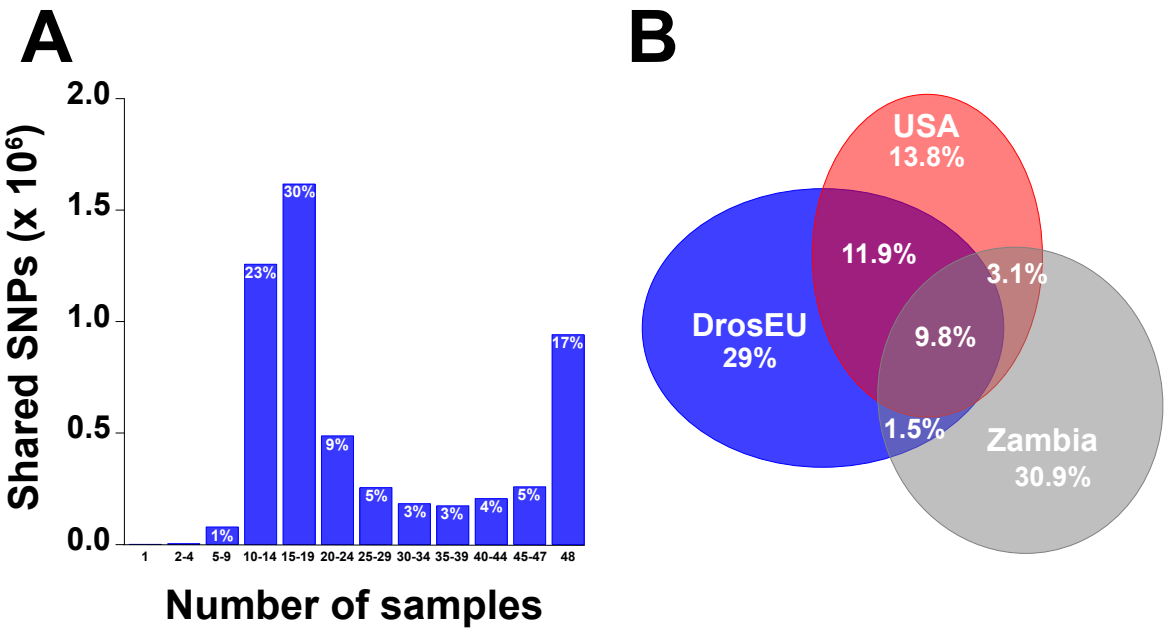
1809 **Figure 3**



1810

1811

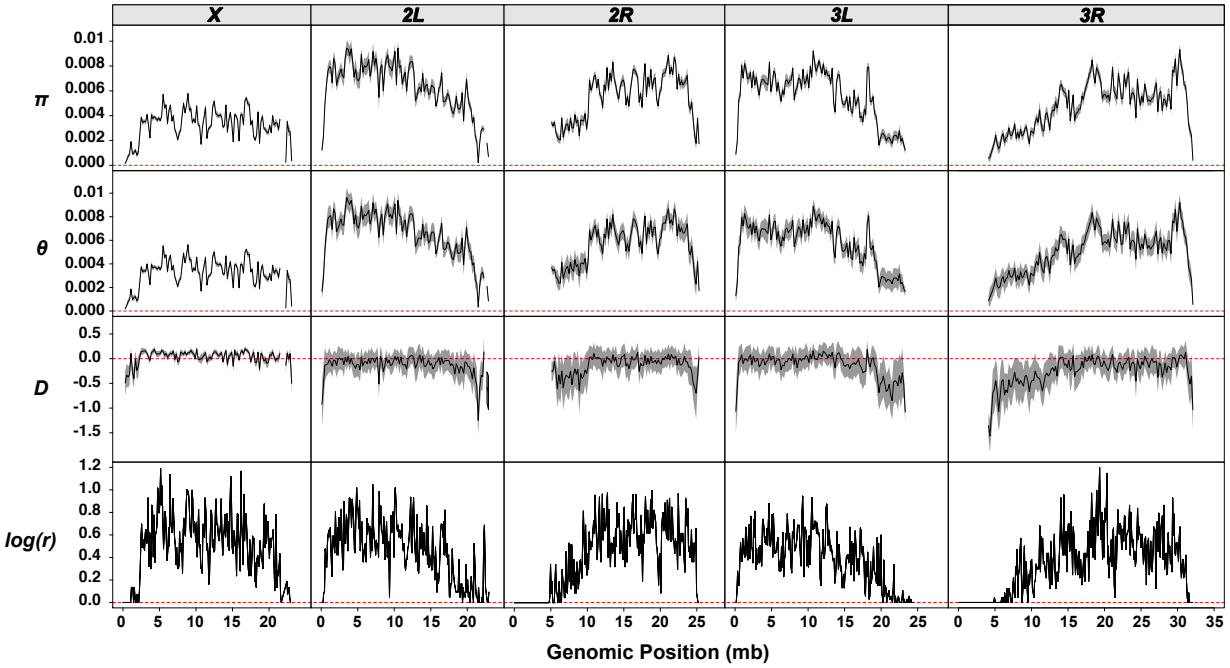
1812 **Figure 4**



1813

1814

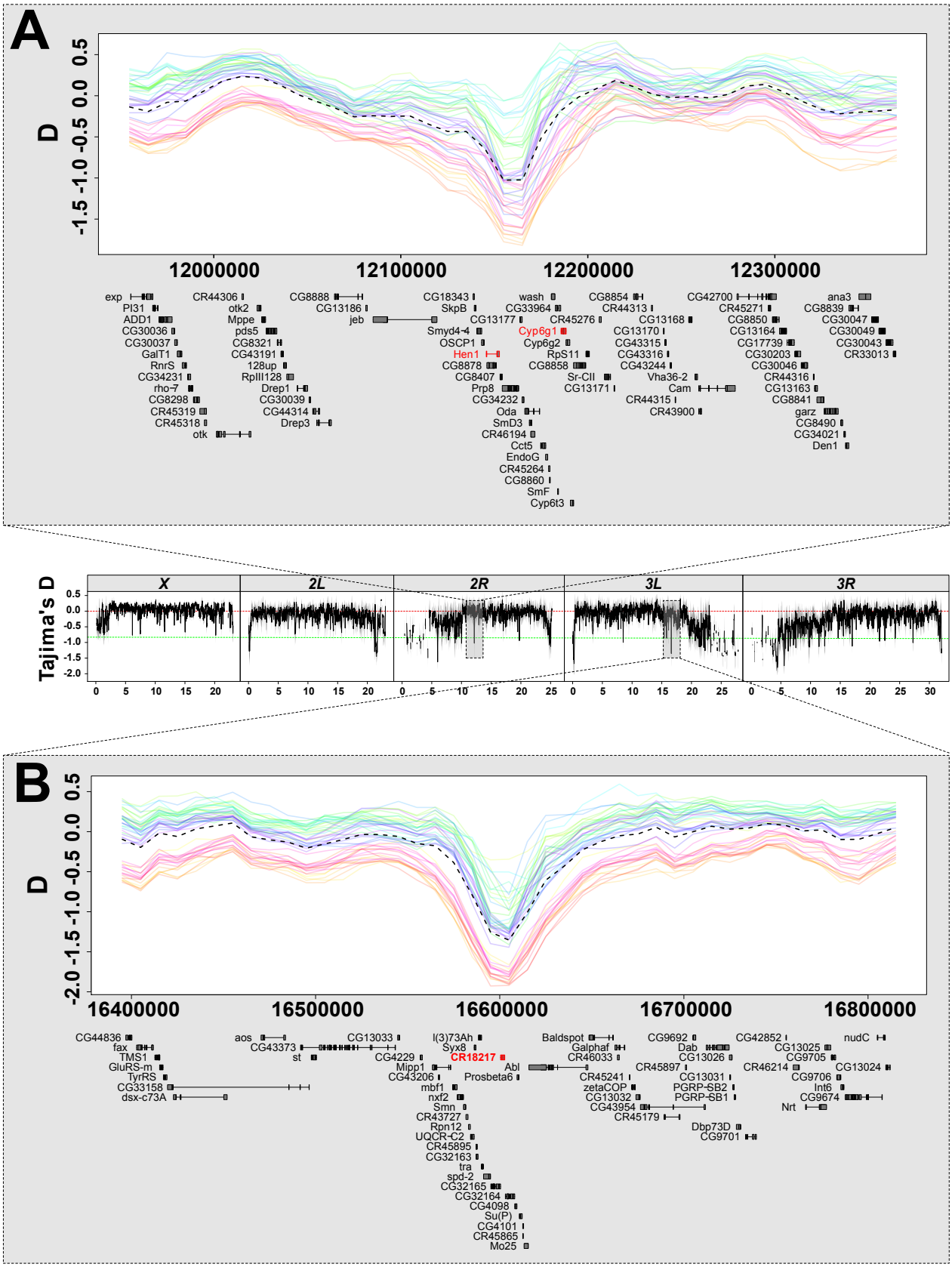
1815 **Figure 5**



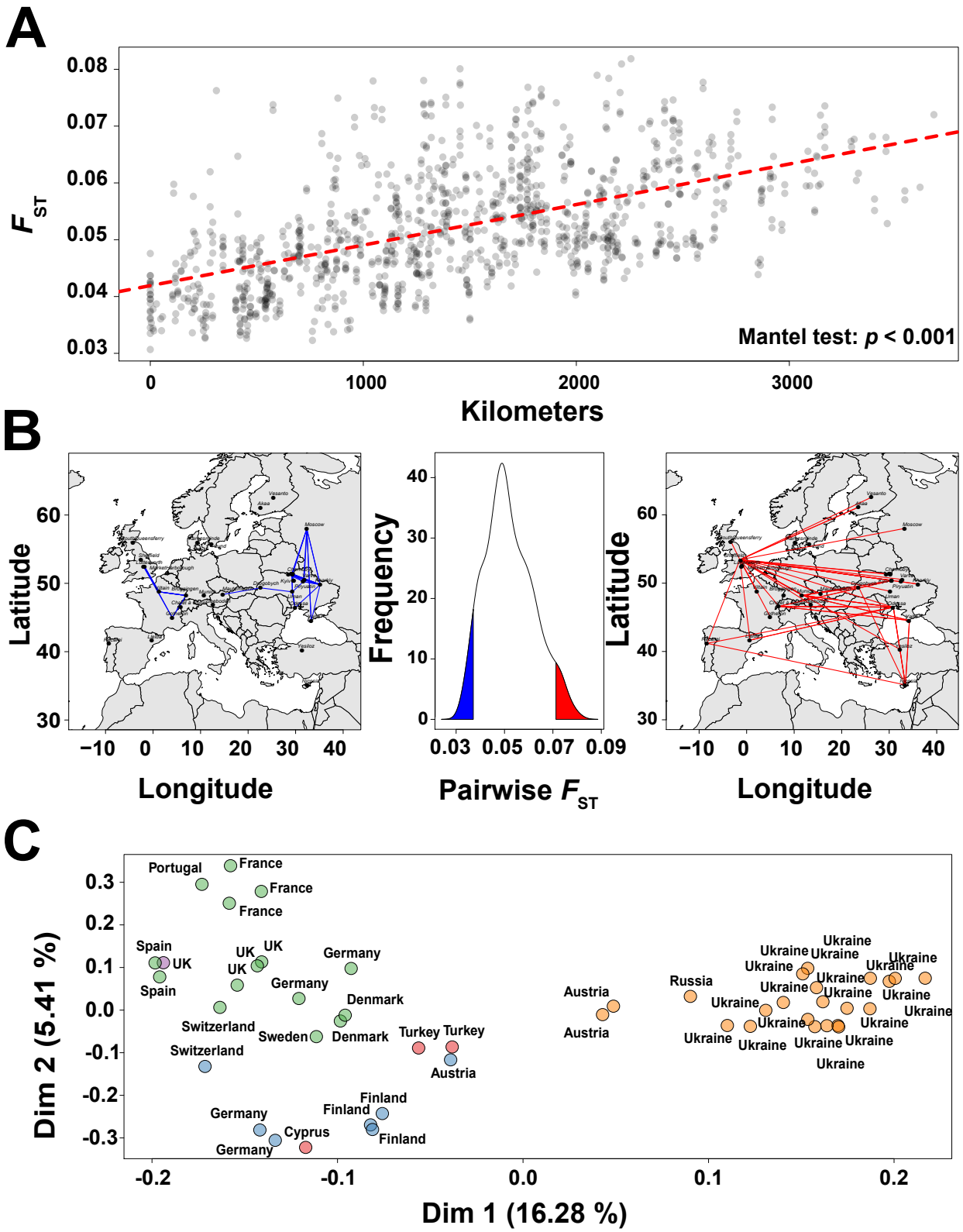
1816

1817

Figure 6



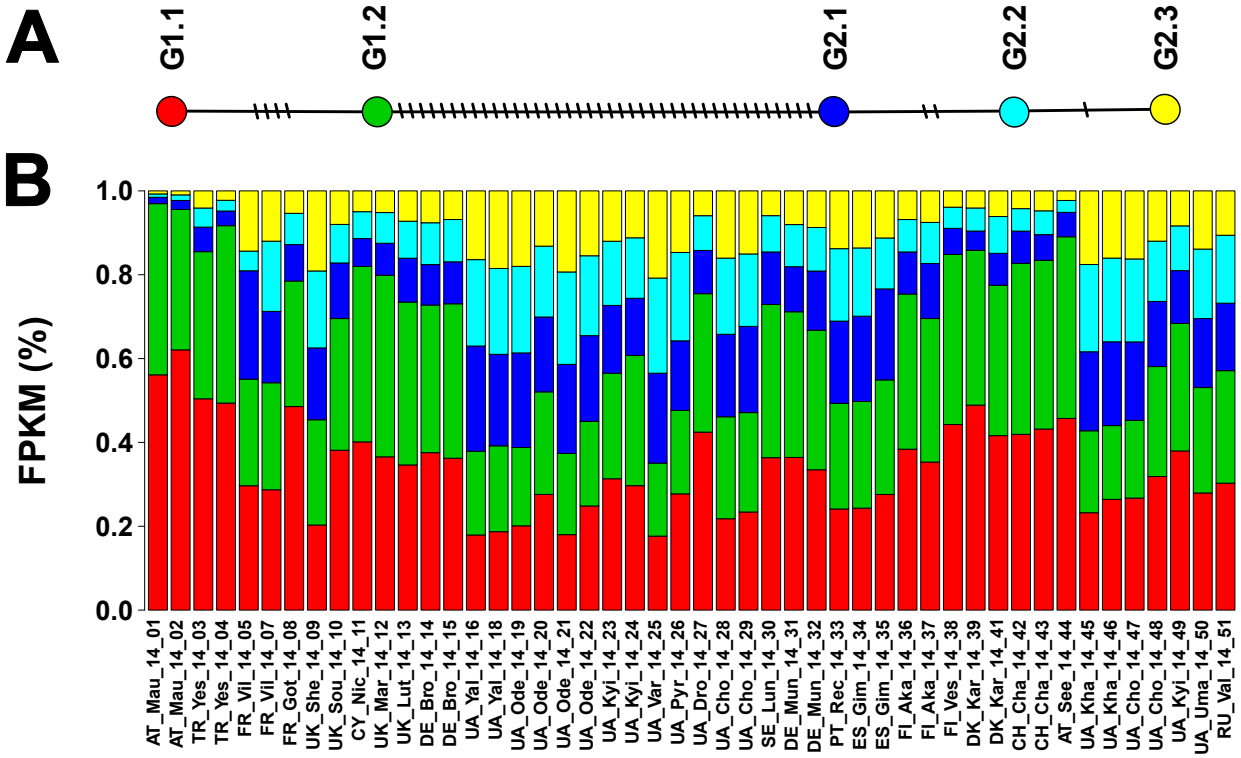
1820 **Figure 7**



1821

1822

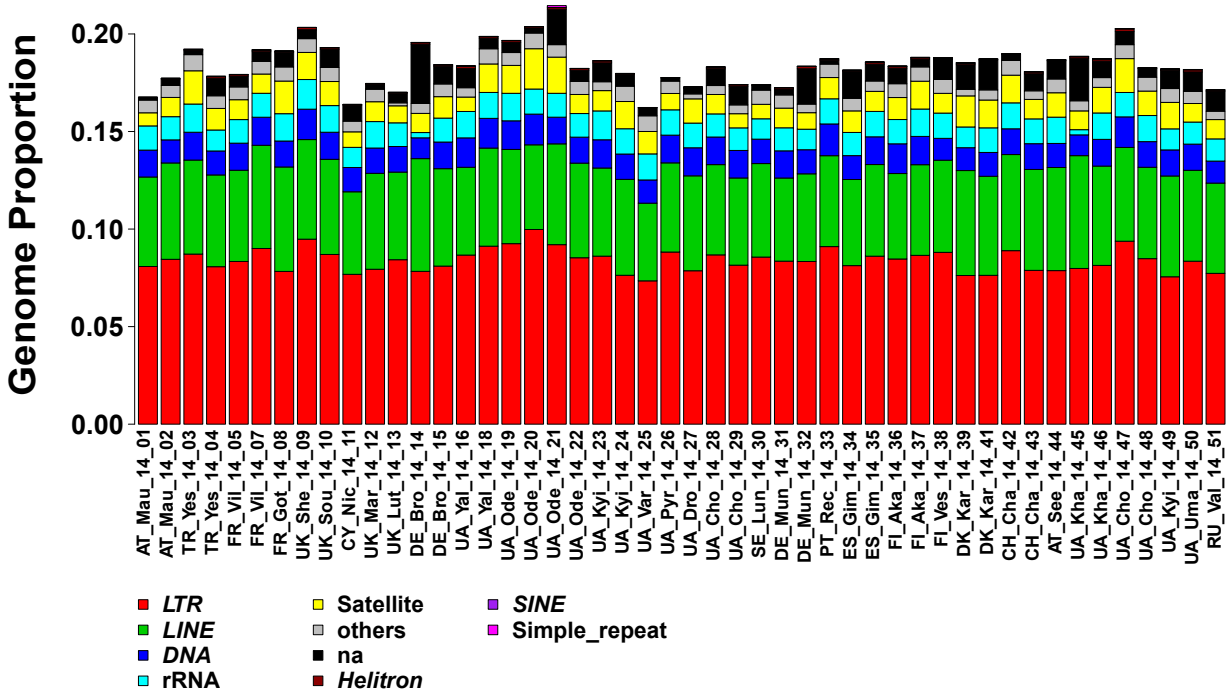
1823 **Figure 8**



1824

1825

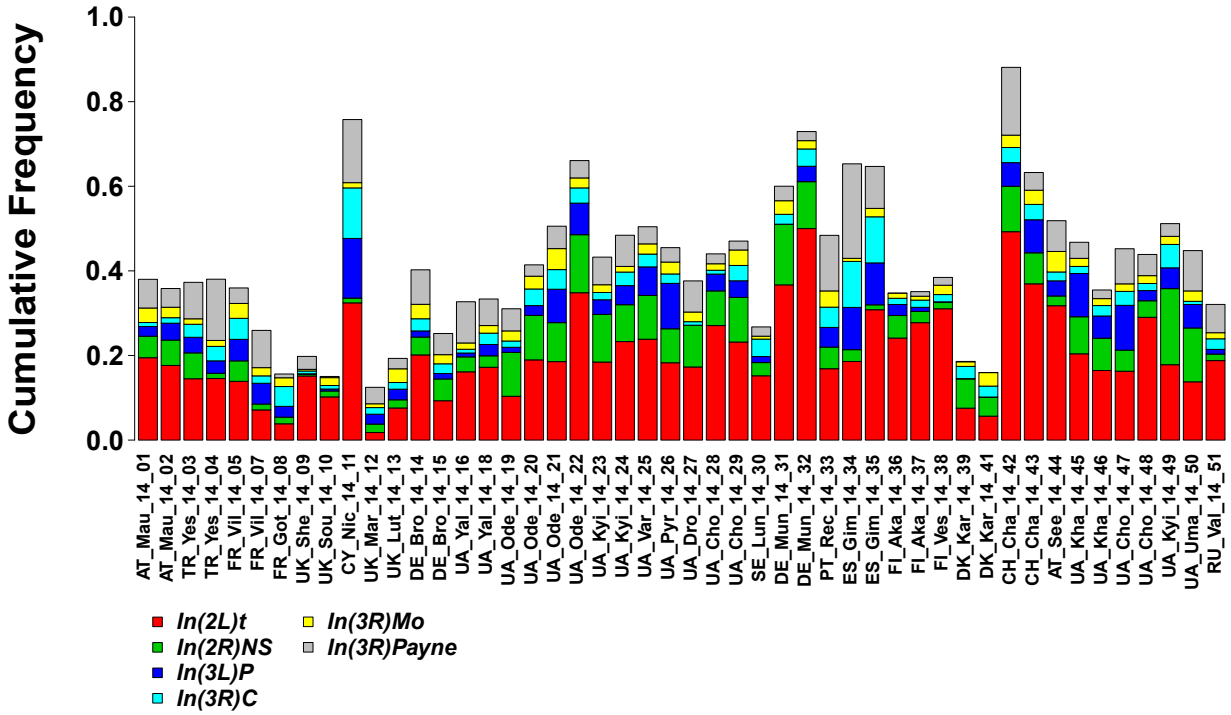
1826 **Figure 9**



1827

1828

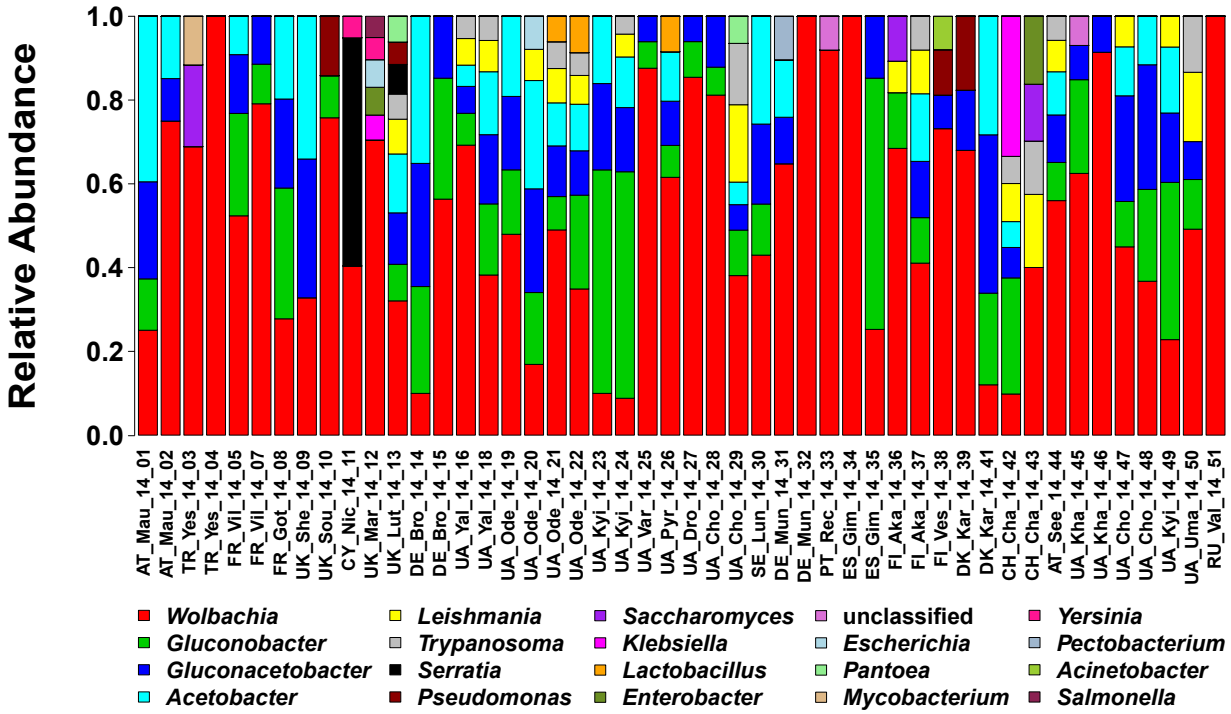
1829 **Figure 10**



1830

1831

1832 **Figure 11**



1833

1834