Demographic History of the Human Commensal *Drosophila melanogaster*

J. Roman Arguello[1,*], Stefan Laurent[2,*], and Andrew G. Clark[3,4]

[1]Department of Ecology and Evolution, University of Lausanne, Switzerland

[2]Max Planck Institute for Plant Breeding Research, Köln, Germany

[3]Department of Molecular Biology and Genetics, Cornell University

[4]Department of Biological Statistics and Computational Biology, Cornell University

*Corresponding authors: E-mails: roman.arguello@unil.ch; laurent@mpipz.mpg.de.

Accepted: January 25, 2019

Abstract

The cohabitation of *Drosophila melanogaster* with humans is nearly ubiquitous. Though it has been well established that this fly species originated in sub-Saharan Africa, and only recently has spread globally, many details of its swift expansion remain unclear. Elucidating the demographic history of *D. melanogaster* provides a unique opportunity to investigate how human movement might have impacted patterns of genetic diversity in a commensal species, as well as providing neutral null models for studies aimed at identifying genomic signatures of local adaptation. Here, we use whole-genome data from five populations (Africa, North America, Europe, Central Asia, and the South Pacific) to carry out demographic inferences, with particular attention to the inclusion of migration and admixture. We demonstrate the importance of these parameters for model fitting and show that how previous estimates of divergence times are likely to be significantly underestimated as a result of not including them. Finally, we discuss how human movement along early shipping routes might have shaped the present-day population structure of *D. melanogaster*.

Key words: demography, *Drosophila*, migration, admixture, population expansion.

Introduction

Drosophila melanogaster is a pre-eminent genetic and evolutionary model. The species originated in sub-Saharan Africa, and only recently expanded its range to inhabit diverse habitats around the globe. At some point early in the species' history, *D. melanogaster* evolved to be a human commensal (David and Capy 1988; Lachaise et al. 1988). Its recent global expansion and well-documented large population sizes have implied a capacity to quickly adapt to local ecological conditions. These insights, together with powerful functional genomic and genetic tools available for *D. melanogaster*, position it as a compelling model species with which to study the molecular mechanisms and evolutionary processes of range expansion and local adaptation. Foundational to understanding recent adaptive differences between populations of *D. melanogaster* is an understanding of its recent demographic history.

Past surveys of genetic diversity have placed the origin of *D. melanogaster* in sub-Saharan Africa (Begun and Aquadro 1993; Lachaise and Silvain 2004), and more recent African sampling has begun to illuminate an increasingly fine-scale understanding of its genetic variation over the continent (Pool et al. 2012). The common understanding is that *D. melanogaster* began to expand north in concert with the recession of the last ice age (David and Capy 1988; Li and Stephan 2006), resulting in a single "out-of-Africa" population bottleneck, possibly in concert with human dispersal (Henn et al. 2012). Current estimates place this divergence between African and European lineages at 12–19,000 years ago (assuming ten generations per year), though its severity and timing have been topics of debate (Thornton and Andolfatto 2006; Stephan and Li 2007). There was an initial conjecture based on elevated phenotypic divergence that some Asian populations might pose an exception, possibly having an older independent colonization unrelated to human movement (referred to as an ancient "Far Eastern race"; David et al. 1976; Lemeunier et al. 1986; David and Capy 1988). However, subsequent modeling using Southeast Asian samples (Kuala Lumpur) was unable to identify genetic signatures of such a scenario (Laurent et al. 2011). As a result,

the factors underlying the morphological characteristics of Asian subpopulations have remained enigmatic.

Although genome-wide approaches to identify and estimate admixture between populations of *D. melanogaster* have been expanding (Kao et al. 2015; Bergland et al. 2016), demographic models for non-African populations have been limited to three general locales: the aforementioned Southeast Asian sample (Laurent et al. 2011), a North American sample (Duchen et al. 2013), and a European sample (the Netherlands [Li and Stephan 2006]). Estimates from these analyses have supported a single out-of-Africa event for Eurasian populations (12,000–19,000 ya; Li and Stephan 2006; Laurent et al. 2011; Duchen et al. 2013) associated with a severe bottleneck. They have additionally provided extant and ancestral population size estimates, which indicate a recent population expansion.

Together, these previous analyses have provided an initial understanding of the divergence patterns, the existence of admixture, and changes in population size among *D. melanogaster* populations. However, from the perspective of demographic parameterization, these efforts have been limited by small sample sizes and by the scope of model parameters that were investigated. For example, many of the previous data sets used for demographic modeling relied primarily on a small number of X-linked polymerase chain reaction-amplified fragments. Additionally, aside from the admixture estimate provided by Duchen et al. (2013), most of the previous demographic models have not included gene flow between populations (i.e., migration and admixture).

Here, we present an analysis of the population structure and demographic history inferred from autosomal polymorphism in the Global Diversity Lines (GDLs) (Grenier et al. 2015). The GDL provide high-quality, validated and uniformly generated genome-wide samples from five geographically diverse *D. melanogaster* populations: Africa, North America, Europe, Asia, and the South Pacific (fig. 1*A*) . We first show that the African and Asian populations are the most diverged among pair-wise comparisons, with the remaining three populations closely related to each other (including the physically distant Tasmanian sample). We then fit demographic models to these populations and demonstrate that the inclusion of gene flow, by long-term migration and by more recent admixture, provides critical improvements to the model fits and yields earlier divergence times. We discuss these results in light of several open questions, including the age and severity of the out-of-Africa bottleneck, the "European-like" south Pacific samples, as well as the implications these models have for inferences of natural selection.

Materials and Methods

Single Nucleotide Polymorphism Data Set

The single nucleotide polymorphism (SNP) data used for this study originated from the GDLs (Grenier et al. 2015), a collection of 84 lines that were derived from five world populations: Beijing, China (15), Ithaca, USA (19), the Netherlands (19), Tasmania (18), and Zimbabwe (13). GDLs were inbred for 12 generations and are largely homozygous, the exception being regions associated with inversions harboring lethal alleles that could not be made homozygous by inbreeding (Grenier et al. 2015). GDLs were fully sequenced to an average depth of 12.5× per line, and independent SNP validation was carried out demonstrating very high-quality calls. These data are publicly available (SRA study SRP050151). We applied the IBD and callability masks to the SNP calls, as described (Grenier et al. 2015). SNPs were limited to autosomes (excluding chromosome 4) and only small intronic (positions 32–65 bp) and 4-fold degenerate positions were used based on genomic annotations generated using SNPeff (Cingolani et al. 2012) and *D. melanogaster*'s r5.57 genome assembly (Grenier et al. 2015). This subset of the full GDLs was outputted into VCF files (supplementary 1 file, Supplementary Material online) using vcftools (v0.1.11; Danecek et al. 2011). For calculating summary statistics, we treated our data sets as haploid by randomly selecting one of two alleles across heterozygous sites (within the heterozygous blocks). Missing SNP genotypes were imputed based on the population-specific allele frequency of the site. SNP diversity estimates generated using vcftools (v0.1.11; Danecek et al. 2011). Divergence statistics were based on the alignment of the GDL SNPs to *D. melanogaster* (dm3), *Drosophila simulans* (droSim2), *Drosophila sechellia* (droSec1), *Drosophila erecta* (droEre2), and *Drosophila yakuba* (droYak2).

Given that we had access to high-quality alignments and posterior probabilities assigned to ancestral states for most our SNP data set (Grenier et al. 2015), we initially aimed to use the unfolded site frequency spectrum (SFS) for our demographic inferences. However, as errors in polarization can greatly impact the high- and low-frequency bins of the SFS (and can thus result in model misestimation), we provided additional analyses of the ancestral state calls. To do this, we used the SNP-dense ancestral-like Zimbabwe data set. Summarizing a large subset of all posterior probabilities indicated that nearly all ancestral calls were above 0.75, with a vast majority >0.95 (supplementary fig. 3*A*, Supplementary Material online). One option to guard against mispolarization would have been to use SNPs above an arbitrary allele frequency threshold. We thus extracted all SNPs with a posterior probability ≥ 0.95 and plotted these values as a function of frequency class (supplementary fig. 3*B*, Supplementary Material online). We observed a steady decline across the frequency bins, with a notable drop for SNPs at frequencies ≥ 0.9. We hypothesize that this decline is attributable to positions hit by more than a single mutation (a violation of the infinite site assumption) as well as to incomplete lineage sorting. To guard against this subtle but impactful bias, we chose to carry out all of our analyses with the folded SFS. Furthermore, recent simulation results indicate that fastsimcoal2 can correctly identify the

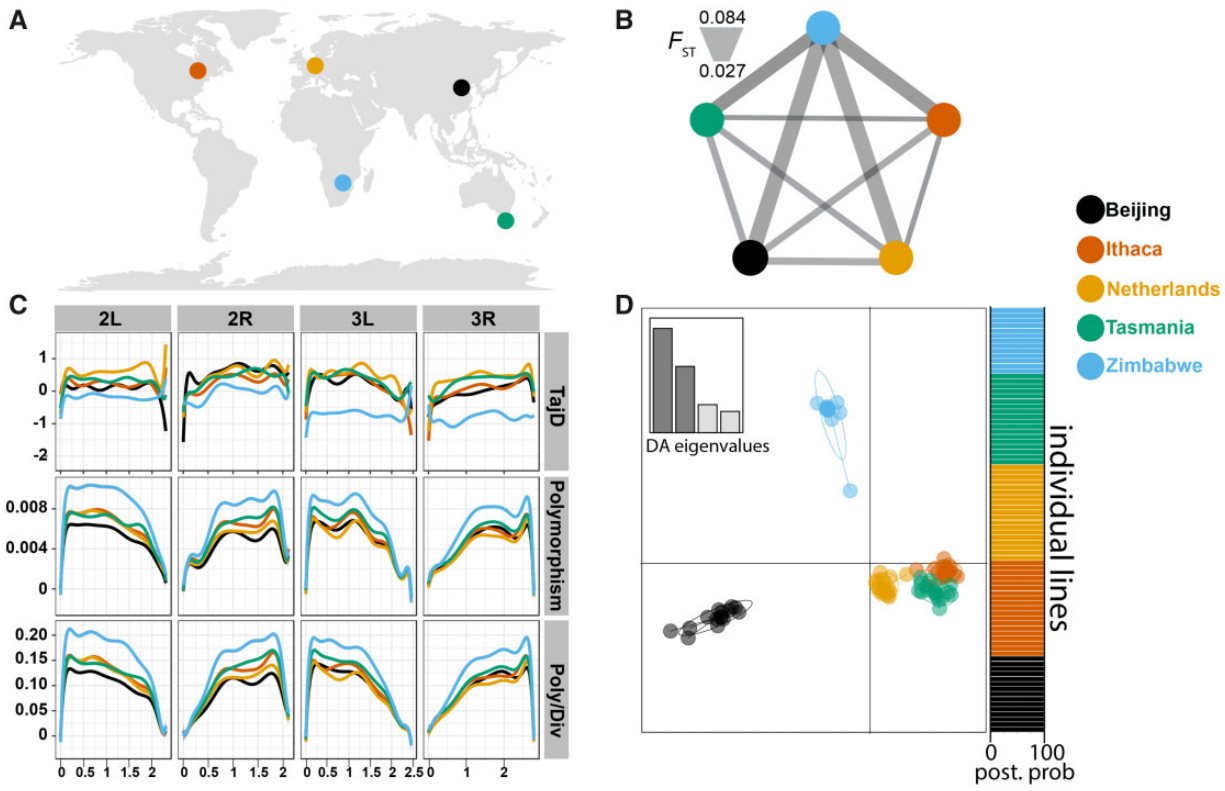


Fig. 1.—Overview and clustering of the data. (*A*) Sampling locales included in the GDLs. (*B*) Population differentiation as measured by genome-wide F_{ST} within a pair-wise network. Thickness of the lines connecting pairs of populations indicate F_{ST} measured between them. (*C*) Summary statistics for genome-wide SNP data (TajD = Tajima's *D*; polymorphism = average number of nucleotide differences per site, π [Nei and Li 1979], Poly/Div = π/divergence, where divergence was measured as the average number of nucleotide substitutions per site between the *Drosophila melanogaster* and *Drosophila simulans*; modified from Grenier et al. [2015]). (*D*) Genetic clustering of "neutral" autosomal SNPs by Discriminant Analysis of Principal Components (Jombart 2008; Jombart and Ahmed 2011).

intensity and direction of gene flow using a folded site frequency spectrum (Gollner et al. 2016). In cases where unfolding the SFS is required, correcting erroneous ancestral state calls can be done following the methods of Hernandez et al. (2007) or Keightley and Jackson (2018).

Data Processing and Demographic Inferences

∂a∂i (Gutenkunst et al. 2009) was used to generate input files and summary statistics for fastsimcoal2 (v2.5.2.21) (Excoffier and Foll 2011; Excoffier et al. 2013). VCF files were converted to ∂a∂i-formatted files using our vcf2dadi_GDL_Neutral_Class.py script (https://gitlab.com/roman.arguello/GDL_demo; last accessed February 28, 2019). To inspect the unfolded SFS, we appended ancestral state calls to the ∂a∂i-formatted file using the input_GDLancestral_GDLstates.pl script (https://gitlab.com/roman.arguello/GDL_demo; last accessed February 28, 2019). For each population, we generated SFS based on the sample size that maximized the number of SNPs by using ∂a∂i's "projection" function within the *pop_Srange.py scripts (supplementary table 3, Supplementary Material online; https://gitlab.com/roman.arguello/GDL_demo; last accessed February 28, 2019). SFS data sets were outputted by ∂a∂i, and the fastsimcoal2 header was manually added. Demographic parameter inferences were calculated with the maximum likelihood framework implemented in fastsimcoal2 (v2.5.2.21; Excoffier and Foll 2011; Excoffier et al. 2013). For each model, we ran 50 replicates, each with 100,000 simulations and 40 expectation–maximization cycles. Model choice was carried out by calculating the Akaike's weight of evidence based on the natural logarithm-transformed maximum likelihood outputted for each model (Johnson and Omland 2004; Excoffier et al. 2013). The complete output and scripts are available within the supplementary 2 file, Supplementary Material online. Within this zipped file are individual directories for all of the models contained in supplementary table 2, Supplementary Material online, and "best_paramater" files that additionally summarize the parameters outputted over all of replicates.

Predictive Simulations

Coalescent simulations were generated using fastsimcoal2 (v2.5.2.21) (Excoffier and Foll 2011; Excoffier et al. 2013).

For each model, we simulated 1,000 data sets under the parameters that generated the maximum likelihood for the best fitting model (see above). Summary statistics and the folded SFS were calculated from these simulations using ∂a∂i within our predictor*.py scripts (supplementary 3 file, Supplementary Material online). The outputted data were then plotted with the scripts found in R files: 1pop_pred_sims.Rmd, 3pop_BNZ_pred_sims.Rmd, 3pop_NIZ_pred_sims.Rmd, and 3pop_NTZ_pred_sims.Rmd (https://gitlab.com/roman.arguello/GDL_demo; last accessed February 28, 2019). The complete input, output, and code for these simulations are available within the supplementary 3 file, Supplementary Material online.

Confidence Intervals

Confidence intervals (CIs) were calculated for all parameters of the four best models (dro03, dro17, NIZ03, and dro21, see also supplementary table 2, Supplementary Material online). We used each of these demographic models, together with their respective estimated parameters, to generate 100 simulated data sets each consisting of 25,000 fragment of 100 bp. The mutation and recombination rates were set to 1.39×10^{-9} and 1×10^{-8}, respectively. We re-estimated demographic parameters for each of the simulated data sets using the same procedure as was applied to the observed data (above) and constructed the CIs as the 0.025 and 0.975 quantiles of the distributions obtained from 100 re-estimated parameter values. The complete output and scripts are available within the zipped supplementary 4 file, Supplementary Material online.

Population Clustering

To investigate the genetic clustering of individual lines with respect to their geographic origins, we applied Discriminant Analysis of Principal Components implemented within the R (v3.4.1) package adegenet (v1.4-2; Jombart 2008; Jombart and Ahmed 2011). The VCF-formatted data file (above) was converted to adegenet-formatted data using the "VCF_2_adegenet_snp_format.pl" script. The R markdown file with the analysis scripts (DAPC_GDL_autosomal_neutral_SNPs.Rmd) is available (https://gitlab.com/roman.arguello/GDL_demo; last accessed February 28, 2019).

Shipping Route Data

To visualize early European shipping routes, we used data available for the years 1662–1855 from the CLIWOC Database 2.1 (A Climatological Database for the World's Oceans; García-Herrera 2007). The IMMA-formatted data file was converted to a csv file and plotted with a script with R (v3.4.1) script plot_CLIWOC21.Rmd file (https://gitlab.com/roman.arguello/GDL_demo; last accessed February 28, 2019).

Results

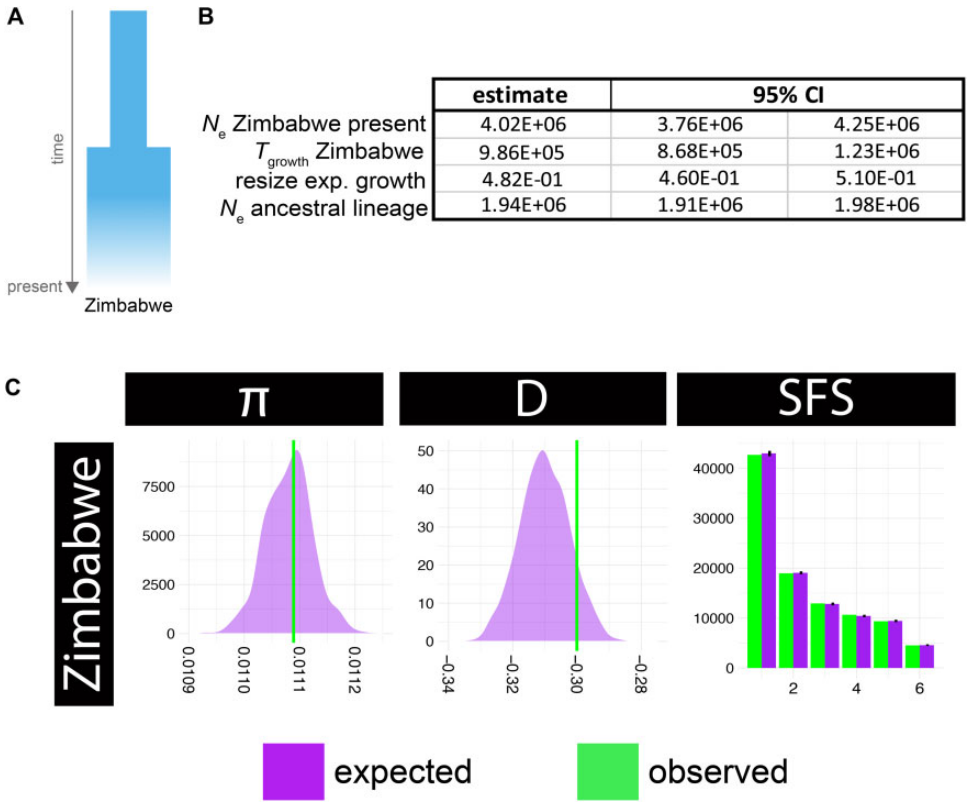
Population Structure of the GDL

Though *D. melanogaster* has colonized much of the globe, population genetic surveys have demonstrated that individuals from even distant locales regularly display low to intermediate levels of population differentiation (Caracristi and Schlötterer 2003; Dieringer et al. 2005; Turner et al. 2008; Pool et al. 2012). Consistent with these previous observations, the five GDL populations also display relatively low levels of population differentiation, with genome-wide pair-wise F_{ST} ranging from 0.027 to 0.086 (fig. 1*B*; see also Grenier et al. 2015). In addition to genetic differentiation, there are notable differences among geographic populations in population genetic summary statistics (fig. 1*C*; see also Grenier et al. 2015), including a significant reduction in nucleotide diversity among all non-African populations (Begun and Aquadro 1993), a pattern consistent with *D. melanogaster*'s range expansion subsequent to its southern African origin (Stephan and Li 2007).

Given the relatively low genetic differentiation, particularly between the non-African lineages, we initially asked to what extent these populations could be defined as distinct clusters. To examine this, we quantified the probability that individuals from a given population could be correctly assigned to its geographic sampling locale using Discriminant Analysis of Principal Components (DAPC; Jombart et al. 2010). Our analyses were carried out on ~167,000 small intronic and 4-fold degenerate autosomal SNPs, which were chosen to minimize the impact of nonneutral evolutionary forces on these (and subsequent) analyses (Parsch et al. 2010; Lange and Pool 2018; see Materials and Methods). This data set clustered strongly by sample locale, with the posterior probabilities of the assignability of each fly line to its geographic label being unambiguous (posterior probabilities nearly 100% for all samples; fig. 1*D*; supplementary table 1, Supplementary Material online).

As expected based on genome-wide F_{ST}, Zimbabwe is the most differentiated among pair-wise comparisons, followed by Beijing (fig. 1*D*). The increased divergence between our Asian samples and the other three derived populations adds to previous observations that Northern Asia samples display some of the most differentiated *D. melanogaster* lineages outside of Africa (Schlötterer et al. 2006; Laurent et al. 2011). Additionally, the clustering of the Tasmanian samples with the North American and European samples affirms the close genetic relationship that this S. Pacific population has with the latter, an observation that was previously reported using an independent genomic data set (Bergland et al. 2016).

Together, these initial analyses indicate that, despite the modest genetic differentiation among the five GDL populations, they display sufficient genetic divergence to enable correct assignment of individual samples to their geographic locality. The population clustering among the GDL, along



	estimate	95% CI	
N_e Zimbabwe present	4.02E+06	3.76E+06	4.25E+06
T_{growth} Zimbabwe	9.86E+05	8.68E+05	1.23E+06
resize exp. growth	4.82E-01	4.60E-01	5.10E-01
N_e ancestral lineage	1.94E+06	1.91E+06	1.98E+06

Fig. 2.—Single population demographic inferences. (*A*) Schematic of the single population demographic model. (*B*) Table of estimates [for] the single population model. Symbols indicate the following: N_e = effective population size, T_{growth} = time of population growth measured in number of generation (assuming ten generation per year). (*C*) Comparison of estimated and predictive simulation values that were calculated under the best fitting population expansion (nucleotide diversity [π], Tajima's *D* [*D*; Tajima 1989], and the nucleotide SFS). Black vertical lines on the simulated SFS bars indicate the 95% CIs.

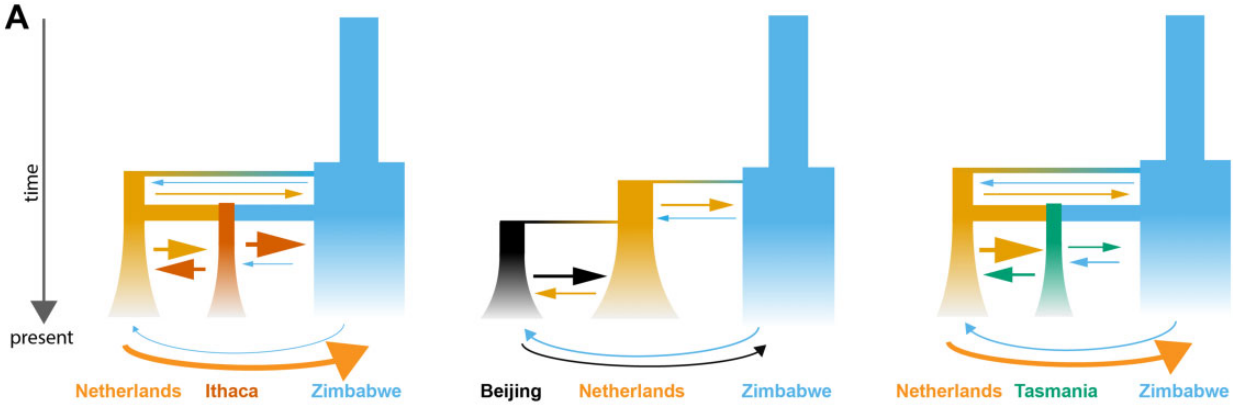
with previous analyses of other *D. melanogaster* samples (Duchen et al. 2013; Kao et al. 2015; Bergland et al. 2016), does not suggest isolation-by-distance, raising questions about the demographic models that could account for these more complicated historical genetic patterns. Among these questions are what models are capable of explaining the elevated Beijing divergence, and how might past gene flow have impacted Tasmania's relationship with samples from North America and Europe?

Demographic Inference: The Population Expansion Signal in the Zimbabwe Population

We chose to start our demographic inference by identifying the best 1-population model for the African sample because this population is central to all subsequent more complex models. Using the same set of ~167,000 autosomal SNPs, we carried out model choice among three models that have previously been investigated: steady state, expansion, and a bottleneck model (Li and Stephan 2006; Thornton and Andolfatto 2006; Duchen et al. 2013; Ragsdale and Gutenkunst 2017). The estimates for the bottleneck model effectively yield a population size expansion similar to the one estimated for the expansion model (the expansion model was parameterized to be nested in the bottleneck model). Thus, owing to a smaller number of parameters, the expansion model was identified as our best model for the African population (fig. 2*A*). Similar results were obtained by Ragsdale and Gutenkunst (Ragsdale and Gutenkunst 2017). The ancestral African population size was estimated to be ~1,900,000 (1,907,055–1,980,494) individuals, with the expansion having a relative increase in size ($N_{ancestral}/N_{present}$) of 0.48 (0.46–0.51) beginning ~99,000 ya (86,779–122,664, assuming ten generations per year; fig. 1*B*). This estimate places the expansion ~1.5–2.5 times earlier than the previous estimates of 37,300 and 60,000 ya (figs. 2, 4*A*, and *B*; Li and Stephan 2006; Laurent et al. 2011).

Demographic Inference: Three Population Models

Having clarified the ancestral-like Zimbabwe population expansion, we next aimed to infer models for the samples that were collected from the regions more recently colonized (conditioning on the ancestral African expansion, above). We were particularly interested in the extent to which gene flow through past migration and more recent admixture has contributed to patterns of genetic diversity. These estimates have largely been excluded from previous inference



B

	INZ			BNZ			NTZ		
	estimate	95% CI		estimate	95% CI		estimate	95% CI	
N_e Ithaca present	5.54E+05	6.36E+04	2.55E+06	-	-	-	-	-	-
N_e Beijing present	-	-	-	1.51E+05	8.60E+04	4.03E+06	-	-	-
N_e Tasmania present	-	-	-	-	-	-	1.64E+06	7.95E+04	2.84E+06
N_e Netherlands present	1.60E+06	1.07E+06	4.17E+06	4.73E+05	2.03E+05	3.89E+06	6.35E+05	4.94E+05	3.81E+06
N_e Zimbabwe present	3.98E+06	3.29E+06	4.75E+06	3.91E+06	3.02E+06	4.69E+06	4.23E+06	3.37E+06	4.82E+06
N_e Beijing bottleneck	-	-	-	1.81E+02	8.20E+01	2.27E+04	-	-	-
N_e Ithaca bottleneck	8.39E+02	1.01E+02	2.20E+03	-	-	-	-	-	-
N_e Tasmania bottleneck	-	-	-	-	-	-	1.07E+03	1.13E+02	1.55E+03
N_e Netherlands bottleneck	3.78E+04	9.27E+03	6.09E+04	3.54E+04	9.15E+01	5.15E+04	6.00E+04	5.90E+03	6.65E+04
N_e ancestral Africa	1.93E+06	1.82E+06	2.02E+06	1.95E+06	1.79E+06	2.00E+06	1.86E+06	1.72E+06	1.93E+06
T_{split} Ithaca	1.79E+02	2.70E+01	4.23E+02	-	-	-	-	-	-
T_{split} Netherlands	2.00E+04	7.24E+03	2.43E+04	6.62E+04	1.17E+04	1.03E+05	2.60E+04	5.05E+03	2.68E+04
T_{split} Beijing	-	-	-	2.18E+04	3.41E+03	3.12E+04	-	-	-
T_{split} Tasmania	-	-	-	-	-	-	3.68E+02	3.83E+01	4.28E+02
2*Nm* Ithaca to Netherlands	6.43E+01	1.64E+01	1.12E+02	-	-	-	-	-	-
2*Nm* Netherlands to Ithaca	6.67E+01	1.46E+01	1.00E+02	-	-	-	-	-	-
2*Nm* Zimbabwe to Netherlands	8.00E+01	1.84E+01	9.19E+01	3.87E+00	1.15E-01	5.56E+00	1.67E+01	8.21E+00	8.51E+01
2*Nm* Netherlands to Zimbabwe	3.01E+00	1.68E+00	8.48E+01	9.16E-01	3.21E-02	5.05E+00	9.43E-02	1.12E+00	4.32E+01
2*Nm* Ithaca to Zimbabwe	1.19E+01	1.43E+00	3.54E+01	-	-	-	-	-	-
2*Nm* Zimbabwe to Ithaca	5.79E+01	1.36E+01	9.02E+01	-	-	-	-	-	-
2*Nm* Beijing to Netherlands	-	-	-	4.56E+00	3.90E+00	5.24E+01	-	-	-
2*Nm* Netherlands to Beijing	-	-	-	1.54E+01	1.54E+01	8.18E+01	-	-	-
2*Nm* Beijing to Zimbabwe	-	-	-	9.16E-01	6.87E-01	1.20E+01	-	-	-
2*Nm* Zimbabwe to Beijing	-	-	-	1.11E+00	7.06E-02	6.26E+00	-	-	-
2*Nm* Tasmania to Netherlands	-	-	-	-	-	-	1.08E+02	2.25E+01	1.20E+02
2*Nm* Netherlands to Tasmania	-	-	-	-	-	-	2.44E+01	3.64E+01	1.72E+02
2*Nm* Tasmania to Zimbabwe	-	-	-	-	-	-	4.26E+00	4.36E-01	2.45E+01
2*Nm* Zimbabwe to Tasmania	-	-	-	-	-	-	3.55E+00	6.36E+00	7.26E+01
2*Nm* ancestral Zimbabwe to Netherlands	1.74E+00	1.62E-01	3.60E+00	7.79E-01	1.04E+00	1.91E+01	4.27E+00	2.36E-01	4.37E+00
2*Nm* ancestral Netherlands to Zimbabwe	1.78E+00	7.50E-01	8.94E+00	2.54E+00	1.55E+00	1.37E+02	1.12E+00	9.70E-01	1.05E+01
resize exp. growth	4.85E-01	4.18E-01	5.74E-01	4.98E-01	4.20E-01	6.25E-01	4.40E-01	3.91E-01	5.39E-01
T_{growth} Zimbabwe	1.13E+05	9.86E+05	2.14E+06	8.81E+04	7.90E+04	2.18E+05	1.28E+05	1.15E+05	2.22E+05
$A_{Zimbabwe}$	1.82E-01	1.57E-01	2.25E-01	-	-	-	3.27E-01	2.88E-01	3.87E-01

Fig. 3.—Best fitting 3-population models and their parameter estimates. (*A*) Schematics for the three best fitting 3-population models. Width of the population branches suggest population sizes (not to scale); arrows indicate direction of migration forward in time, with their sizes suggesting relative rates (not to scale). (*B*) Parameter estimates for the corresponding best fitting models and their 95% CI ranges. Symbols indicate the following: N_e = effective population size, $2Nm$ = scaled migration rate forward in time, T_{split} = population split-time measured in number of generations (ten generation per year), T_{growth} = time of population growth measured in number of generations (ten generation per year), and A = admixture proportion.

work, and we hypothesized that if their inclusion provided improved fits to the data then most of the existing demographic estimates would be significantly impacted. In particular, the split-times of these populations will have likely been underestimated.

We examined three sets of models involving the samples from the more recently colonized localities. Each of the three sets used the Zimbabwe and the Netherlands data sets (abbreviated as Z and N, respectively) but differed by the inclusion of Tasmania, Beijing, or Ithaca (abbreviated as T, B, or I, respectively). Eighteen models were examined in total (supplementary fig. 1 and supplementary table 2, Supplementary Material online). The motivation for this approach was to keep our models relatively tractable while also facilitating comparisons between several previous demographic analyses that used 1) African–European–North American trio of data sets and 2) African–European–Asian trio of data sets (Laurent et al. 2011; Duchen et al. 2013) (see Discussion). The South

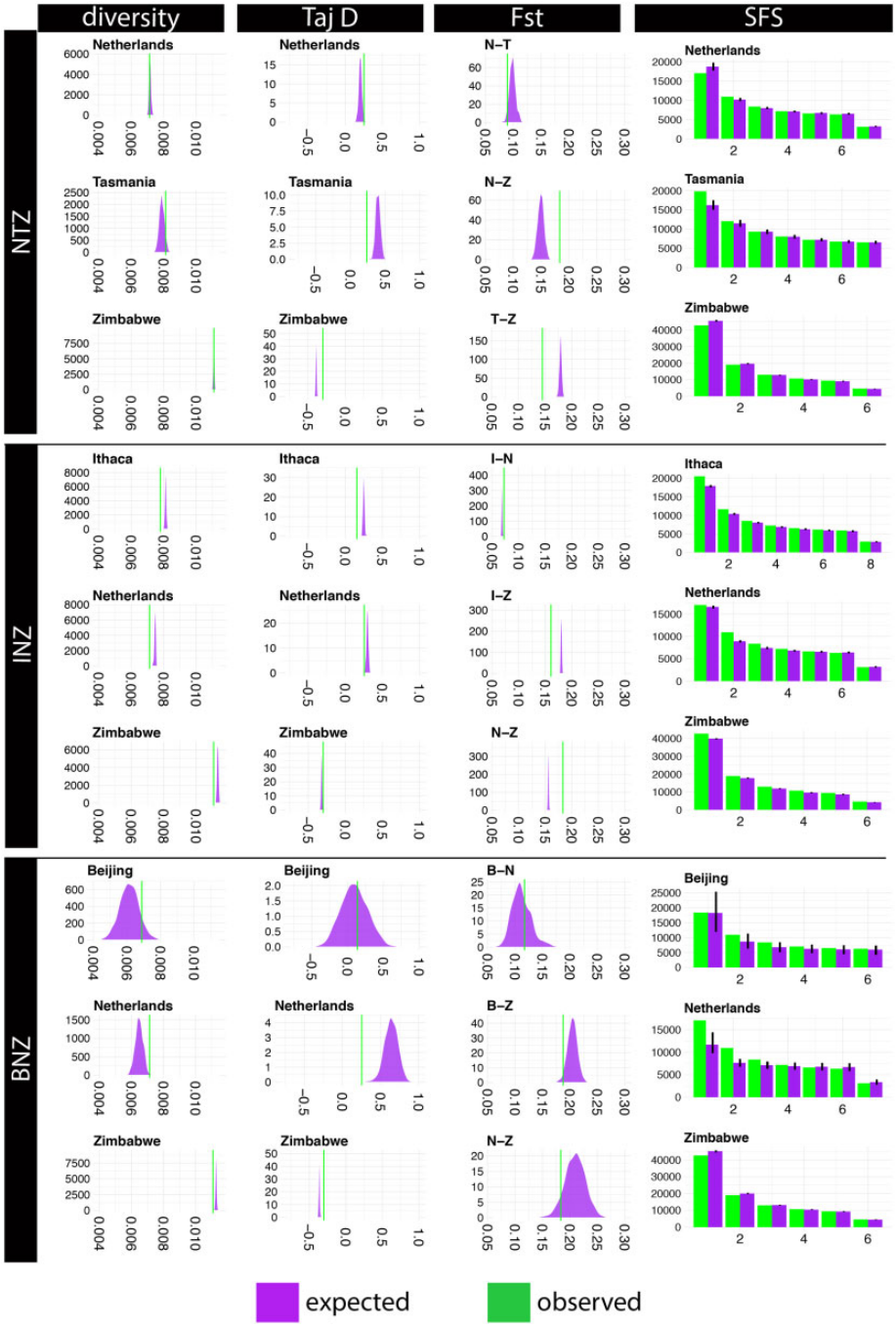


Fig. 4.—3-Population predictive simulations. Comparison of simulated values under the three best fitting 3-population models (from fig. 3A) to the observed values: (nucleotide diversity [π] [Nei and Li 1979], Tajima's D [D] [Tajima 1989], population differentiation [F_{ST}], and the nucleotide SFS). Population names are abbreviated: B = Beijing; I = Ithaca; N = the Netherlands; T = Tasmania; and Z = Zimbabwe. Black vertical lines on the simulated SFS bars indicate the 95% CIs.

Pacific sample provides a novel population data set for demographic inferences. For the models involving Tasmania and Ithaca samples, we constrained the divergence times to be no older than the start of European seafaring exploration, as was done previously (Duchen et al. 2013).

D. melanogaster Experienced a Single Out-of-Africa Event

Among the three 3-population models tested, we found no evidence for additional out-of-Africa colonization events. Indeed, multiple models in which the Asian lineage independently diverged prior to the European lineage provided lower

likelihoods than the serial founding model with gene flow (supplementary fig. 1 and supplementary table 2, Supplementary Material online). This latter result is important because the previous Asian demographic modeling included samples from Kuala Lumpur, a region that experienced significant early European shipping activity and potentially different histories from more isolated regions of Asia (supplementary fig. 2, Supplementary Material online). Instead, by using a north Asian sample located more distant from maritime routes, our analyses indicate that all non-African lineages are derived from the same out-of-Africa event, consistent with previous claims (Baudry et al. 2004; Laurent et al. 2011).

Migration and Admixture Are Crucial Historical Factors for *D. melanogaster*

For each of the 3-population models, the inclusion of migration consistently provided significantly better fits to the data than models without it (fig. 3*A*; see supplementary table 2, Supplementary Material online, for comparisons). Generally, migration was greatest between pairs of populations from the Ithaca, Tasmania, and Zimbabwe populations, with the direction of migration usually asymmetric (fig. 3*A* and *B*). For example, there was considerably more back migration into Africa from European lineages ($2Nm$ ranged from 3.87 to 72.61 vs. 0.09 to 2.90 for the opposite direction), as well as higher rates from European gene pools to the S. Pacific ($2Nm$ = 107.70 vs. 24.40; see fig. 3*B* for all estimates and CIs). Notably, migration was inferred to be overall lower for the best-fitting model involving the Beijing population.

In addition to migration, gene flow is expected to have also impacted patterns of genetic diversity through recent admixture. Admixture has been shown to occur among African *D. melanogaster* populations (Pool et al. 2012), as well as growing evidence for it occurring among populations from similar regions of the world that the GDL represent (Caracristi and Schlötterer 2003; Duchen et al. 2013; Kao et al. 2015; Bergland et al. 2016). Although the genetic signatures of admixture have been identified among these populations, other than the modeling done by Duchen et al. (2013), demographic inferences that include these parameters have largely been ignored (but see Corbett-Detig and Nielsen 2017; Medina et al. 2018).

In line with the previous N. America–Africa estimate (Duchen et al. 2013), we again estimated the African admixture proportion to be 18% (16–23%). We additionally found that admixture between Tasmania and Africa provided a better fit compared with models without it, with the proportion of African admixture estimated to be two times higher than that seen between Africa and N. America (33% [29–39%]; fig. 3*B*). Intriguingly, the same admixture models tested with the BNZ trio did not result in a better fit over the migration-alone model (supplementary table 2, Supplementary Material online). Therefore, along with the reduced migration rates for the BNZ data set (above), these results highlight a second line of evidence that gene flow experienced by the Asian lineage has been significantly lower compared with the New World samples (fig. 3*B*). They also provide key demographic information that likely underlies the elevated population differentiation observed for Beijing (as illustrated in the DAPC and by F_{ST} values; fig. 1).

To determine how well our best-fitting models recapitulate our observed data, we again carried out predictive simulations under the maximum likelihood estimates for each of the best 3-population models. Overall, we observe good matches to our SFS and summary statistics, particularly with respect to diversity levels (fig. 4). However, the predictive simulations did highlight aspects of these populations' histories that we have not captured in our current models. This is most readily observed in the singleton and low-frequency class of the SFS (and reflected in *D*), where our models result in both higher and lower expectations within the Netherlands and Tasmanian data sets. Additionally, these simulations predicted lower F_{ST} values for the N–Z comparison, and higher F_{ST} values for the T–Z comparison in the ZNT model. Likewise, in the case of the NIZ model, they predicted higher F_{ST} values for the I–Z comparison and slightly lower values for the N–Z comparison. We suspect that these misspecifications arise primarily from two sources. First, despite our efforts to enrich our SNP data with neutral variants, it is likely that a subset of these variants are linked to sites under selection (both negative and positive selection). Second, we have focused on relatively simple models that aim to capture the predominant demographic features of the species' history; this simplistic approach likely omits aspects of their histories that have important but more subtle impacts on genetic diversity (see Discussion). Overall however, these migration and admixture estimates highlight the historical and ongoing importance that gene flow has played in shaping global patterns of genetic diversity within *D. melanogaster*.

Inclusion of Gene Flow Impacts Our Demographic Understanding

It is well appreciated that gene flow between diverging populations decreases the coalescence time for alleles drawn from the distinct populations (Wakeley 2000). Estimates based on models that omit past gene flow can therefore result in significantly shallower population divergence times. As most of the previous demographic work on *D. melanogaster* did not include migration, it is of interest to examine how the inclusion of these parameters impact the split-time of the populations. Previously, estimates for the out-of-Africa *D. melanogaster* split-time/bottleneck has been estimated at ~12–19 ka (Li and Stephan 2006; Thornton and Andolfatto 2006; Laurent et al. 2011; Duchen et al. 2013). Though the estimates obtained from models depended to some extent on

the populations included in the analyses, all results indicate that the split-time/bottleneck was likely to be at the upper end of the previous estimates (19,000 ya), and most likely earlier: INZ = 20,040 (7,243–24,345), NTZ = 26,021 (5,054–26,830), BNZ = 66,208 (11,727–102,890) (fig. 3*B*). Additionally, compared with the previous estimates for the Asian split-time (Laurent et al. 2011) of ~22,000 ya (3,409–31,235), we inferred the event to have occurred much earlier, at ~5,000 ya (fig. 3*B*). For the North American split-time, Duchen et al. (2013) constrained their models based on entomological sampling records from the turn of the 19th century. We have similarly constrained our NTZ and INZ models to have a North American split-time to be $\leq$ 500 ya (i.e., around the European discovery of N. America).

Discussion

Geographically diverse population genomic data for *D. melanogaster* provide unique opportunities to investigate the historical processes experienced by a human commensal insect, as well helping to establishing neutral expectations for genetic diversity upon which tests of selection can be based. Benefiting from significantly expanded genomic data sets compared with earlier *D. melanogaster* demographic studies, as well as an expanded set of populations (Grenier et al. 2015), our results demonstrate a central role for both ancient migration and more recent admixture. However, the degree to which populations experienced gene flow varied. Most prominently, the European (the Netherlands) and New World samples (Ithaca and Tasmania) were inferred to have experienced recent admixture, a historical feature not supported for the Asian sample (Beijing). Additionally, except for the Asia-to-Europe migration, the overall estimate of gene flow was reduced for the Asian population (fig. 2*B*; supplementary table 2, Supplementary Material online). In light of the gene flow inferred for INZ and NTZ, the BNZ model helps inform previous reports of elevated genetic and phenotypic divergence of Asian populations of *D. melanogaster* (Lachaise et al. 1988; Schlötterer et al. 2006; Laurent et al. 2011; Scheitz et al. 2013). It suggests a scenario in which increased divergence has been promoted through a relative reduction of both ancient and recent gene flow, and not an earlier independent colonization. Though direct evidence is lacking, these patterns of genetic diversity are consistent with expectations based on early ocean-based human exploration, which likely accelerated *D. melanogaster*'s global colonization. In particular, the stronger connection between Europe, N. America, and the S. Pacific provided by European exploration and shipping routes (supplementary fig. 2, Supplementary Material online) compared with Asia would be consistent with increased opportunity for recent admixture. It also provides an explanation for the close genetic relationships between the physically distant Tasmanian and N. American–European populations (fig. 1*B* and *D*).

The inclusion of gene flow in the *D. melanogaster* models has had the greatest impact on the population split-time estimates for the European and Asian populations. Previous studies had indicated that the European–African split and the Asia–Africa split were comparable, having occurred around 12–15,000 ya (Li and Stephan 2006; Laurent et al. 2011). The CIs of our estimates are partially overlapping with the CIs of previous estimates but, as expected, indicate older divergence times with MLEs >20,000 ya for both population spit times. These results suggest that *D. melanogaster* was likely expanding its northern range before the end of the last glacial period (~12,000 ya). We emphasize that this estimate is not equivalent to the timing of the colonization of Europe and Asia, but rather places the lineages from these locales within the subpopulation that had begun to separate from the ancestral sub-Saharan population. A recent study by Kapopoulou et al. (2018) estimated an ancient divergence between west and southern African populations (Zambia) ~72,000 years. It remains to be investigated from which of these two populations the European lineages have likely diverged and whether the population size expansion and these early population divergences have or have not been influenced by the demography of human populations.

To our knowledge, these models provide the most comprehensive demographic estimates for *D. melanogaster* populations to date. As a result, these estimates provide updated parameters that can be used to establish null expectations for studies of selection designed to identify loci involved in population-specific adaptation. Although the GDL have previously been used to examine population differences in targeted classes of genes and gene families (Arguello et al. 2016; Cardoso-Moreira et al. 2016; Early et al. 2017), these models should broaden the capacity to test for local adaptation and for quantifying rates of change among populations. Additionally, we expect that the parameter estimates derived from these models can help to inform patterns of genetic diversity that will be ascertained in several large-scale ongoing (and future) *D. melanogaster* sequencing efforts in North America, Europe, and elsewhere (Kapun et al. 2018).

The approach that we have used to infer demography for these populations has assumed that the genetic variation is neutral. Our effort to reduce the GDL data set to the most "neutral-like" SNP set has been to analyze only small intronic and 4-fold degenerate SNPs (Parsch et al. 2010). However, the compactness of the *D. melanogaster*'s genome, combined with its large N_e, does raise concern over potential biases introduced if some portion of these sites are linked to regions under selection. We suspect that some of our model misspecifications that were illustrated in our predictive simulations (fig. 4) could in part be explained by linked selection. Further distilling the *D. melanogaster* SNP data set to neutral variants would likely

require additional information on population-specific regions under selection as well as population-specific recombination rates across the genome, which are known to vary among individuals (Comeron 2014). How selection is biasing our inferences is difficult to assess because the specificities of our modeling approach are also likely to have an impact; though we have considered a relatively broad collection of models, they still remain simple despite having a rich parameter space. Lending additional confidence [to our study] is evidence that selection may have relatively less impact on biasing parameters for recent demographic events (such as we focus on here) than for more ancient events when using "neutral-like" data sets (Lange and Pool 2018).

Our aim to minimize the effect of selection on our demographic inferences led us to exclude sex chromosomes, which carry genetic signals of nonneutral processes including sex-specific variation in reproductive success, sex-biased dispersion, and elevated rates of positive selection (Caballero 1995; Charlesworth 2001; Pool and Nielsen 2007; Ellegren 2009; Meisel and Connallon 2013; Charlesworth et al. 2018). We did observe considerably more population separation within X-chromosome DAPC analyses (data not shown), consistent with diversity estimates and theory (Pool and Nielsen 2007; Grenier et al. 2015). Our autosome-based demographic inferences will provide valuable future comparisons to those based on X-linked variation. Will SNP data from the X, for example, support the population split-times and size changes consistent with the best autosomal models provided here? Hinting at possible differences between these genomic contexts, one previous demographic study using polymorphism data from a highly recombining 2.1-Mb region of the X provided evidence for a more complex African demographic scenario (Singh et al. 2013). Additionally, autosomal-X comparisons may help tease apart demographic processes from selective events on the X.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Jen Grenier, Margarida Cardoso Moreira, Srikanth Gottipati, and Sean Hackett for the work establishing, sequencing and SNP-calling for the GDL resource. Greg Ewing, Jeff Jensen, and the Jensen lab provided important comments and suggestion on early aspects of this work. We would also like to thank the Swiss Institute of Bioinformatics' Vital-IT group for computational resources and support. This work was supported in part by Swiss National Science Foundation grant PP00P3_176956 to JRA, and by National Institute of Health grant AI064950 to A.G.C. and Brian P. Lazzaro.

Literature Cited

Arguello JR, et al. 2016. Extensive local adaptation within the chemosensory system following *Drosophila melanogaster*'s global expansion. Nat Commun. 7:ncomms11855.

Baudry E, Viginier B, Veuille M. 2004. Non-African populations of *Drosophila melanogaster* have a unique origin. Mol Biol Evol. 21(8):1482–1491.

Begun D, Aquadro C. 1993. African and North American populations of *Drosophila* are very different at the DNA level. Nature 365(6446):548–550.

Bergland AO, Tobler R, González J, Schmidt P, Petrov D. 2016. Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. Mol Ecol. 25(5):1157–1174.

Caballero A. 1995. On the effective size of populations with separate sexes, with particular reference to sex-linked genes. Genetics 139(2):1007–1011.

Caracristi G, Schlötterer C. 2003. Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. Mol Biol Evol. 20(5):792–799.

Cardoso-Moreira M, et al. 2016. Evidence for the fixation of gene duplications by positive selection in *Drosophila*. Genome Res. 26(6):787–798.

Charlesworth B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. Genet Res. 77(2):153–166.

Charlesworth B, Campos JL, Jackson BC. 2018. Faster-X evolution: theory and evidence from *Drosophila*. Mol Ecol. 27(19):3753–3771.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*. Fly 6(2):80–92.

Comeron JM. 2014. Background selection as baseline for nucleotide variation across the *Drosophila* genome. PLoS Genet. 10(6):e1004434.

Corbett-Detig R, Nielsen R. 2017. A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. PLoS Genet. 13(1):e1006529.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

David J, Bocquet C, Pla E. 1976. New results on the genetic characteristics of the far east race of *Drosophila melanogaster*. Genet Res. 28(3):253–260.

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

Dieringer D, Nolte V, Schlötterer C. 2005. Population structure in African *Drosophila melanogaster* revealed by microsatellite analysis. Mol Ecol. 14(2):563–573.

Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S. 2013. Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. Genetics 193(1):291–301.

Early AM, et al. 2017. Survey of global genetic diversity within the *Drosophila* immune system. Genetics 205(1):353–366.

Ellegren H. 2009. The different levels of genetic diversity in sex chromosomes and autosomes. Trends Genet. 25(6):278–284.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9(10):e1003905.

Excoffier L, Foll M. 2011. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27(9):1332–1334.

García-Herrera R. 2007. CLIWOC Database 2.1 (final release, original IMMA format). In: Jones PD, et al, editors. Climatological observations from ship logbooks between 1750 and 1854 (release 2.1). Greenwich,

UK: PANGAEA. Available from: https://doi.org/10.1594/PANGAEA.611088. Last accessed February 28, 2019.
Gollner S, et al. 2016. Mitochondrial DNA analyses indicate high diversity, expansive population growth and high genetic connectivity of vent copepods (Dirivultidae) across different oceans. PLoS One 11(10):e0163776.
Grenier JK, et al. 2015. Global diversity lines—a five-continent reference panel of sequenced *Drosophila melanogaster* strains. G3(Bethesda) 5(4):593–603.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5(10):e1000695.
Henn BM, Cavalli-Sforza LL, Feldman MW. 2012. The great human expansion. Proc Natl Acad Sci U S A. 109(44):17758–17764.
Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol. 24(8):1792–1800.
Johnson JB, Omland KS. 2004. Model selection in ecology and evolution. Trends Ecol Evol. 19(2):101–108.
Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24(11):1403–1405.
Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27(21):3070–3071.
Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11(1):1–15.
Kao JY, Zubair A, Salomon MP, Nuzhdin SV, Campo D. 2015. Population genomic analysis uncovers African and European admixture in *Drosophila melanogaster* populations from the South-Eastern United States and Caribbean Islands. Mol Ecol. 24(7):1499–1509.
Kapopoulou A, Pfeifer SP, Jensen JD, Laurent S. 2018. The demographic history of African *Drosophila melanogaster*. Genome Biol Evol. 10(9):2338–2342.
Kapun M, et al. 2018. Genomic analysis of European Drosophila melanogaster populations on a dense spatial scale reveals longitudinal population structure and continent-wide selection. Cold Spring Harbor Laboratory. bioRxiv.
Keightley PD, Jackson BC. 2018. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. Genetics 209(3):897–906.
Lachaise D, Silvain J-F. 2004. How two afrotropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster–D. simulans* palaeogeographic riddle. Genetica 120(1–3):17–39.
Lachaise D, et al. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol Biol. 22:159–225.
Lange JD, Pool JE. 2018. Impacts of recurrent hitchhiking on divergence and demographic inference in *Drosophila*. Genome Biol Evol. 10(8):1882–1891.
Laurent SJY, Werzner A, Excoffier L, Stephan W. 2011. Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. Mol Biol Evol. 28(7):2041–2051.
Lemeunier F, David J, Tsacas L, Ashburner M. 1986. Genetics and biology of Drosophila. Vol. 3e. London: Academic Press.
Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2(10):e166.
Medina P, Thornlow B, Nielsen R, Corbett-Detig R. 2018. Estimating the timing of multiple admixture pulses during local ancestry inference. Genetics 210(3):1089–1107.
Meisel RP, Connallon T. 2013. The faster-X effect: integrating theory and data. Trends Genet. 29(9):537–544.
Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 76(10):5269–5273.
Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. Mol Biol Evol. 27(6):1226–1234.
Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. Evolution 61(12):3001–3006.
Pool JE, et al. 2012. Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. PLoS Genet. 8(12):e1003080.
Ragsdale AP, Gutenkunst RN. 2017. Inferring demographic history using two-locus statistics. Genetics 206(2):1037–1048.
Scheitz CJF, Guo Y, Early AM, Harshman LG, Clark AG. 2013. Heritability and inter-population differences in lipid profiles of *Drosophila melanogaster*. PLoS One 8(8):e72726.
Schlötterer C, Neumeier H, Sousa C, Nolte V. 2006. Highly structured Asian *Drosophila melanogaster* populations: a new tool for hitchhiking mapping? Genetics 172(1):287–292.
Singh ND, Jensen JD, Clark AG, Aquadro CF. 2013. Inferences of demography and selection in an African population of *Drosophila melanogaster*. Genetics 193(1):215–228.
Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98(2):65–68.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595.
Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172(3):1607–1619.
Turner TL, Levine MT, Eckert ML, Begun DJ. 2008. Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. Genetics 179(1):455–473.
Wakeley J. 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution 54(4):1092–1101.

Associate editor: Brandon Gaut