The Demographic History of African *Drosophila melanogaster*

Adamandia Kapopoulou[1], Susanne P. Pfeifer[1,2], Jeffrey D. Jensen[1,2], and Stefan Laurent[1,3,*]

[1]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[2]School of Life Sciences, Center for Evolution and Medicine, Arizona State University, Tempe, Arizona

[3]Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany

*Corresponding author: E-mail: laurent@mpipz.mpg.de.

Accepted: August 27, 2018

Abstract

As one of the most commonly utilized organisms in the study of local adaptation, an accurate characterization of the demographic history of *Drosophila melanogaster* remains as an important research question. This owes both to the inherent interest in characterizing the population history of this model organism, as well as to the well-established importance of an accurate null demographic model for increasing power and decreasing false positive rates in genomic scans for positive selection. Although considerable attention has been afforded to this issue in non-African populations, less is known about the demographic history of African populations, including from the ancestral range of the species. While qualitative predictions and hypotheses have previously been forwarded, we here present a quantitative model fitting of the population history characterizing both the ancestral Zambian population range as well as the subsequently colonized west African populations, which themselves served as the source of multiple non-African colonization events. We here report the split time of the West African population at 72 kya, a date corresponding to human migration into this region as well as a period of climatic changes in the African continent. Furthermore, we have estimated population sizes at this split time. These parameter estimates thus represent an important null model for future investigations in to African and non-African *D. melanogaster* populations alike.

Key words: demographic inference, *Drosophila melanogaster*, inversion polymorphisms.

Introduction

Populations of *Drosophila melanogaster* span five continents, making this organism a widely utilized system to study patterns of local adaptation. Yet, this complex underlying demographic history represents unique challenges for disentangling nonneutral from nonequilibrium processes (e.g., Jensen et al. 2005; Teshima et al. 2006; Thornton and Jensen 2007; Pavlidis et al. 2010), and thus numerous studies have worked to better illuminate the correct demographic null model. Considerable effort has been made in understanding the species' expansion into Europe (e.g., Li and Stephan 2006; Thornton and Andolfatto 2006), Asia (e.g., Laurent et al. 2011), and the Americas (e.g., Duchen et al. 2013; Kao et al. 2015).

However, it is only in the past decade that African demographic history has been similarly scrutinized. In one of the earliest studies, Dieringer et al. (2004) surveyed X-chromosomal microsatellite variation from 13 sampling locations across Africa, describing considerable population structure between North, West, and East Africa. Pool and Aquadro (2006) surveyed nucleotide variation at four 1-kb fragments in 240 individuals from sub-Saharan Africa, and described a distinct East-West geographic pattern, suggesting that western Africa may have been recently colonized from the East. Simultaneously, Li and Stephan (2006) examined about 250 noncoding X-chromosome regions from a population sampled in Zimbabwe, suggesting strong evidence of population growth. In a much larger-scale study, Pool et al. (2012) sequenced whole-genomes from 139 wild-derived strains from 22 sampling locations in sub-Saharan Africa. Based on levels of variation and F_{ST}, they qualitatively described a fit to a model in which Zambia represents the species origin, with subsequent population expansion, structuring and gene flow across the continent—though they concluded on the need for proper demographic model fitting in order to better elucidate these patterns. In addition, Singh et al. (2013) examined a 2 Mb region in 20 individuals sampled from Uganda, also finding support for population expansion, but also suggested an associated population bottleneck out of the

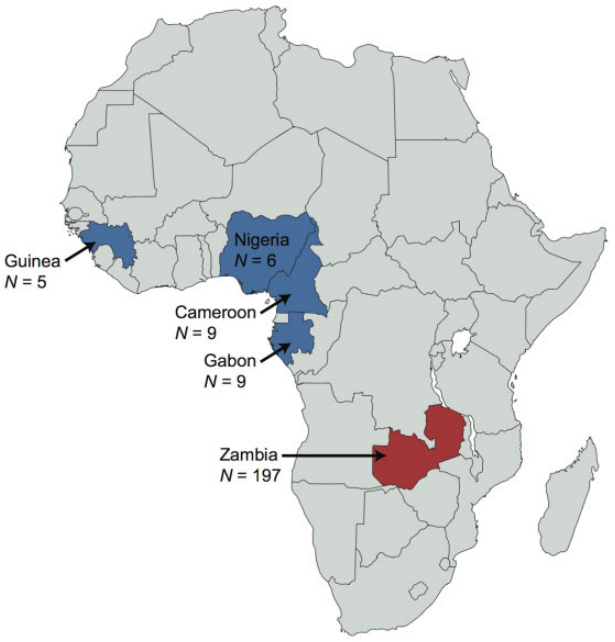

Fig. 1.—Geographic distribution of the five *D. melanogaster* populations. Samples (sample sizes indicated by *N*) were obtained from the Phase 2 (blue) and Phase 3 (red) of the Drosophila Population Genomics Project (Pool et al. 2012; Lack et al. 2015).

initial ancestral range (presumably being Zambia, hundreds of miles to the south).

Following this important work, we here focus our study on Zambia as the likely population of origin, and West Africa as a likely source of multiple widely studied non-African populations (fig. 1). We quantify the demographic history of these regions, including the timing of West African colonization, effective population sizes, and rates of gene flow (supplementary fig. 1, Supplementary Material online). Furthermore, given known segregating inversions as well as the associated difficulties that may arise if they are left unaccounted for, we have carefully curated a data set for the purposes of inferring these underlying neutral demographic parameters, which may serve as the basis for future studies.

Inferring Population History

The levels of genetic differentiation between individuals were assessed using a principal component analysis. The first principal component, explaining 2.7% of the variation, separates the Zambian individuals from the West African individuals, which cluster according to their sampling location (i.e., Cameroon, Gabon, Guinea, and Nigeria; supplementary fig. 2, Supplementary Material online). In contrast, Zambian individuals cluster into two distinct groups based on chromosomal inversions carried by the individuals (supplementary fig. 3, Supplementary Material online). This pattern was well described by Corbett-Detig and Hartl (2012) who noted that polymorphic inversions in *D. melanogaster* affect genomic variation chromosome-wide, with trans-effects beyond the inversions' breakpoints. To avoid the confounding effects of these segregating inversions on subsequent demographic inference, 121 Zambian individuals carrying at least one inversion (i.e., In2RNS, In2Lt, In3R, and In3LOk) were excluded from any further analyses, keeping 76 Zambian lines devoid of any known inversion.

Population structure was then assessed using an admixture model to infer individual ancestry proportions using *sNMF* (Frichot and François 2015), a statistical method to evaluate the ideal number of ancestral populations. The best-fit model (i.e., the model with the lowest minimal cross-entropy) had two ancestry components (fig. 2*a*), strongly supporting the division of individuals from Zambian and West African populations, with evidence of admixture between them (fig. 2*b*). Principal component analysis confirms the two population clusters inferred by *sNMF*, with no additional subgenetic stratification of the Zambian individuals (fig. 2*c*). While there does not appear to be substructure in the Western samples, larger sample sizes may naturally be expected to provide additional resolution.

Given the observed population structure, the demographic history of Zambian and West African populations was investigated using six different two-population demographic models, allowing for both size change as well as gene flow between the populations. Three of the six models assumed that populations remained at a constant size with either no gene flow, symmetric migration, or asymmetric migration between them (supplementary fig. 1, Supplementary Material online). To account for the fact that West African populations exhibit lower nucleotide diversity levels than populations from south-central Africa ($\pi = 0.0086$ in Zambia, $\pi = 0.0077$ in West Africa; and see Pool et al. 2012; Lack et al. 2015), suggesting a potential population bottleneck during their recent colonization from the ancestral range (Haddrill et al. 2005), the remaining three models allowed for population size changes (supplementary fig. 1, Supplementary Material online). The demographic model best fitting the data (fig. 3; supplementary table 1, Supplementary Material online) inferred exponential growth for both the Zambian and West African populations after their split around 70 kya, with ongoing gene flow. In addition, the parameter estimates obtained for the ancestral and present effective population sizes (N_e(anc) = 1,525,061 (95% CI: 1,498,713—1,562,754); N_e(Zambia) = 3,160,475 [95% CI: 2,933,313—3,447,248]) reiterate the higher levels of variation observed in the putative ancestral range of the species.

While the specific parameter values inferred are of particular importance for explicitly modelling an appropriate demographic null in future studies, and represent the first estimates of split times between the ancestral range and West Africa, the qualitative patterns are largely consistent with previous supposition. Namely, the estimated ancestral split times,

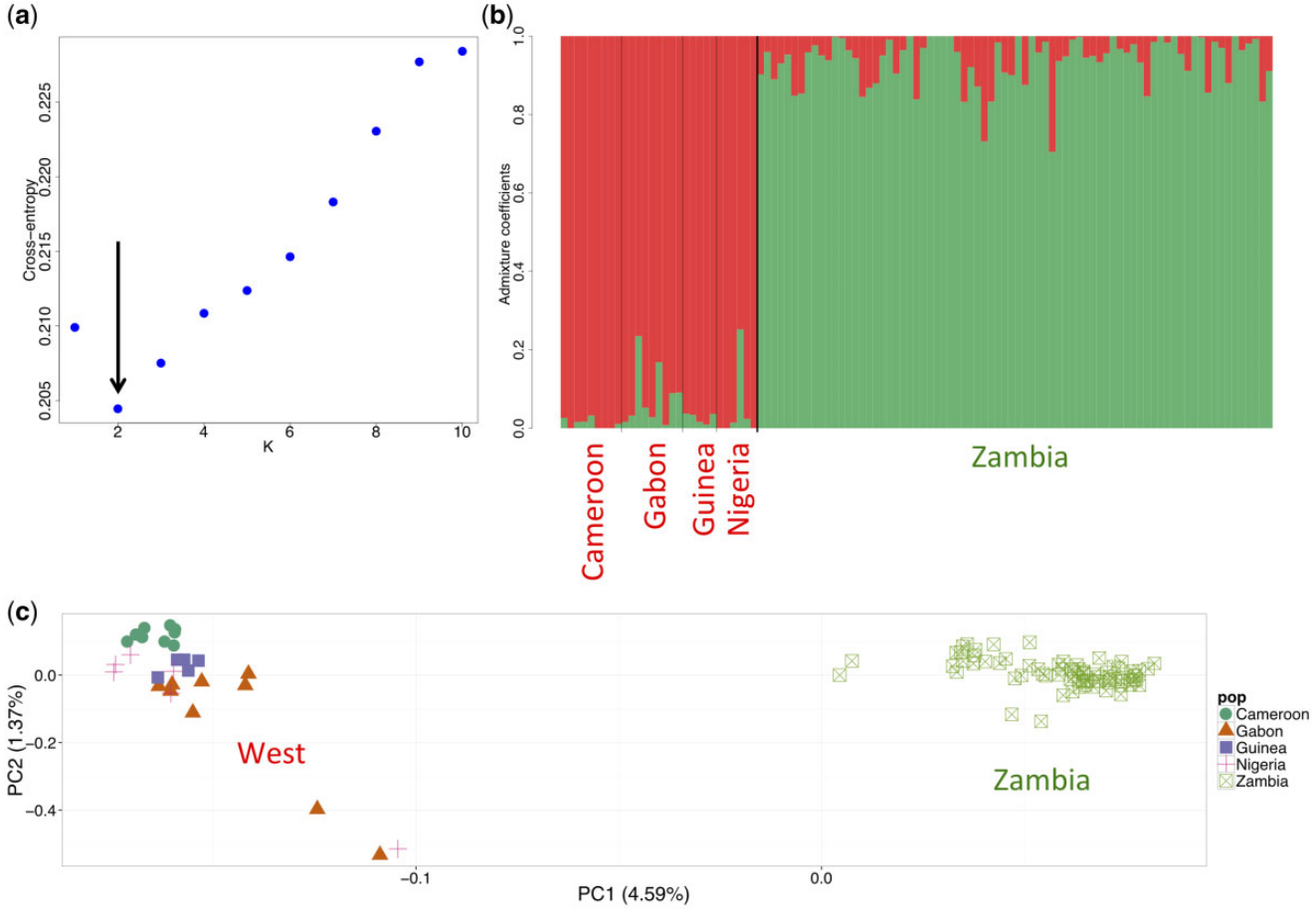

Fig. 2.—Genetic structure of African *D. melanogaster* populations. (*a*) The number of *K* ancestry components best explaining the data was assessed by calculating the cross-entropy corresponding to the model. The best-fit model (i.e., the model with the lowest minimal cross-entropy) had two ancestry components ($K = 2$). (*b*) Individual admixture proportions. (*c*) Principal component analysis (symbols correspond to individuals from different populations; green square: Zambia [$N = 76$ individuals which do not carry the chromosome arm's specific inversion]; green circle: Cameroon [$N = 9$]; orange triangle: Gabon [$N = 9$]; purple square: Guinea [$N = 5$]; red cross: Nigeria [$N = 6$]). Data was thinned to prune for linkage, excluding SNPs with an $r^2 > 0.2$ within a 50 SNP window. Percentages indicate the variance explain by each principle component.

population structure (Pool and Aquadro 2006), and effective population sizes (Laurent et al. 2011), as well as the underlying growth and colonization models themselves (Pool et al. 2012), are all largely in agreement with earlier studies.

Concluding Thoughts

In concordance with Corbett-Detig and Hartl (2012), we find that even when polymorphisms within the inversion breakpoints were not considered in the analysis, the genetic structure associated with inversion polymorphisms persists and is visible when analyzing other markers located on the same chromosomal arm (supplementary fig. 3, Supplementary Material online). By removing these individuals from the analysis, and by carefully curating the data set for neutral sites, we have quantified the demographic histories characterizing these sampling locations. We find evidence for strong growth in populations inhabiting both regions, consistent structure separating West Africa from Zambia, as well as evidence for on-going gene flow particularly in the direction of south/central to west. Thus, this well-fit nonequilibrium demographic model of both the ancestral range of the species as well as the source population of subsequent non-African colonization events, represents a uniquely appropriate null model for future investigations pertaining to the demographic and adaptive histories of both African and non-African populations of *D. melanogaster*.

Materials and Methods

Samples

Publicly available whole-genome sequence data from haploid *D. melanogaster* embryos originating from Guinea ($N = 5$), Nigeria ($N = 6$), Cameroon ($N = 9$), Gabon ($N = 9$), as well as from Zambia ($N = 197$) was obtained from the Phase 2

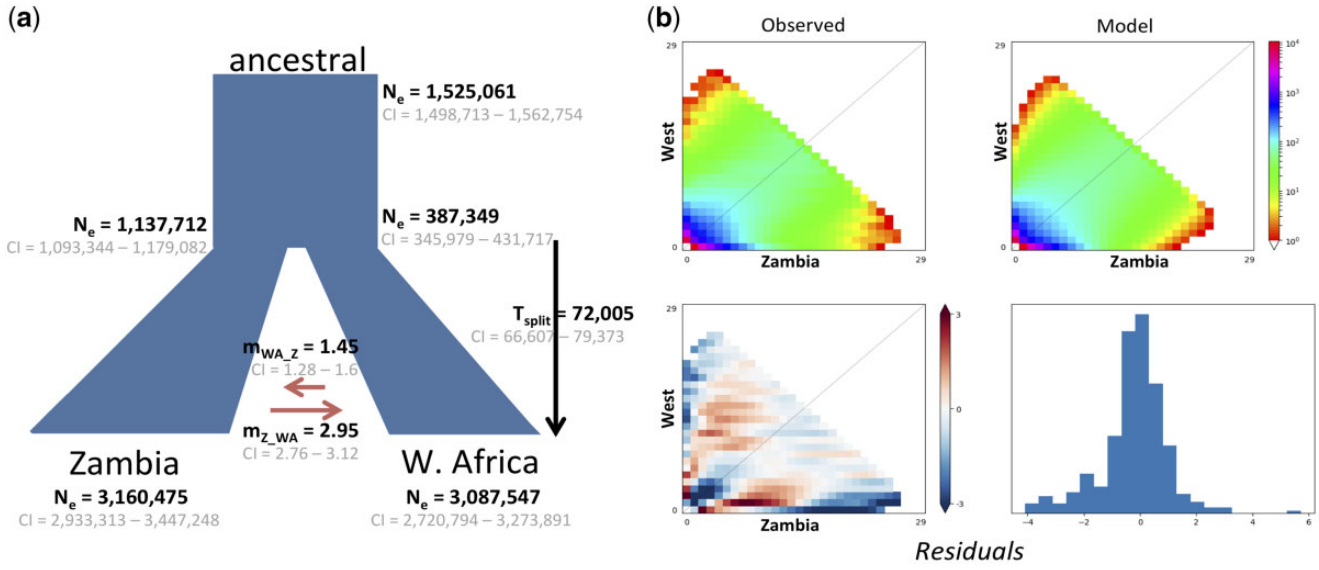

FIG. 3.—Parameter estimates inferred by ∂a∂i under the best fitting demographic model. (*a*) At time T_{split}, the ancestral population splits into two distinct populations, which grow exponentially with asymmetric migration (*m*) between them. The time of the split (T_{split}) was estimated in generation time, which was converted to years, assuming ten generations per year (Laurent et al. 2011). Effective population sizes (N_e) for the ancestral, West African, and Zambian populations were directly estimated by fixing the mutation rate (μ) to 1.3×10^{-9} per base pair per generation (Laurent et al. 2011). 95% confidence intervals (CI) were calculated for each parameter estimate by generating 150 parametric bootstrap replicates of the best fitting model. Note that the mode of the bootstrapped parameter estimates corresponds approximately to the obtained maximum likelihood value estimate. (*b*) Comparison of Joint SFS for the observed data (left) and the best fitting model (right). Below are shown the residuals of the model.

and Phase 3 of the Drosophila Population Genomics Project (DPGP) (Pool et al. 2012; Lack et al. 2015, 2016), respectively (fig. 1). Specifically, genomes previously aligned to a common *D. melanogaster* reference sequence were downloaded from the Drosophila Genome Nexus (DGN) (Lack et al. 2015, 2016) and variants on both arms of chromosome 2 (i.e., chr2L and chr2R) and chromosome 3 (i.e., chr3L and chr3R) were identified using the SNP-sites C program (Page et al. 2016).

As chromosomal inversions may be targeted by natural selection in *D. melanogaster* (Corbett-Detig and Hartl 2012), known inversions were excluded from all demographic analyses (information on inversion breakpoints was obtained from the DGN [Lack et al. 2015]; http://www.johnpool.net/Updated_Inversions.xls; last accessed September 3, 2018). To further minimize the confounding effects of linked selection on demographic inference, the data set was limited to putatively neutral regions of the genome, including 4-fold synonymous degenerate sites (Grenier et al. 2015) as well as the 8th to the 30th base of introns smaller than 65 bp (Parsch et al. 2010). The resulting data set contained 82,149 variants.

Inferring Population Structure

Population structure was investigated using two methods, which cluster individuals based on their genetic similarity using a set of independent SNPs (i.e., SNPs with an $r^2 > 0.2$ within a 50 SNP window were excluded from the data set using PLINK v1.07 [Purcell et al. 2007]). Evidence of population structure was assessed using both a principal component analysis (PCA) as well as the *sNMF* function implemented in the R package LEA v2.0.0 (Frichot and François 2015). The latter implements an admixture model (Pritchard et al. 2000; Patterson et al. 2006) which uses sparse nonnegative matrix factorization to infer individual ancestry proportions based on *K* potential components. Using a cross-validation technique, *K* values ranging from 1 to 10 were examined, and, following (Frichot et al. 2014), the best *K* was selected to minimize the cross entropy.

Demographic Inference

The demographic history of south-western African *D. melanogaster* populations was inferred from the distribution of minor allele frequencies (i.e., the folded joint site frequency spectrum) obtained from the putatively neutral segregating sites using ∂a∂i 1.7.0 (Gutenkunst et al. 2009), a diffusion approximation method. Given the genetic differentiation between populations, six different two-population scenarios (corresponding to samples originating from West Africa—i.e., Guinea, Nigeria, Cameroon, and Gabon, as well as Zambia) were tested, allowing for both population size changes as well as gene flow among the populations (supplementary fig. 1, Supplementary Material online). Thereby, gene flow was modelled either as symmetric or asymmetric, and considered only between the time of the population split and the present.

For every demographic model, ten independent runs were performed using different starting points and the parameter estimates for the best run (i.e., the estimation with the highest likelihood) reported. 95% confidence intervals (CI) were calculated for each parameter estimate by generating 150 parametric bootstrap replicates of the best model. Effective population sizes (N_e) were directly estimated by fixing the mutation rate (μ) to 1.3×10^{-9} per base pair per generation (Laurent et al. 2011). Generation times were converted to years, assuming ten generations per year (Laurent et al. 2011). The best-fitting demographic model was selected based on the Akaike's information criterion (AIC) score (Akaike 1974).

Supplementary Material

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

Acknowledgments

We thank Roman Arguello for helpful discussions and for providing the coordinates of short introns and 4-fold degenerate coding sites for the neutral set of loci. We also thank Athanasios Kousathanas and Anna-Sapfo Malaspinas for useful feedback. This work was supported by grants from the Swiss National Science Foundation and the European Research Council to J.D.J.

Literature Cited

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723.

Corbett-Detig RB, Hartl DL. 2012. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. PLoS Genet. 8(12):e1003056.

Dieringer D, Nolte V, Schlötterer C. 2004. 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.

Frichot E, François O. 2015. LEA: an R package for landscape and ecological association studies. Methods Ecol Evol. 6(8):925–929.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics 196(4):973–983.

Grenier JK, et al. 2015. Global diversity lines—a five-continent reference panel of sequenced *Drosophila melanogaster* strains. G3 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.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. 15(6):790–799.

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170(3):1401–1410.

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.

Lack JB, et al. 2015. The Drosophila genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. Genetics 199(4):1229–1241.

Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE. 2016. A thousand fly genomes: an expanded Drosophila genome nexus. Mol Biol Evol. 33(12):3308–3313.

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.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 2(10):e166.

Page AJ, et al. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genomics 2(4):e000056.

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.

Patterson N, Price AL, Reich D. 2006. Population structure and Eigen analysis. PLoS Genet. 2(12):e190.

Pavlidis P, Jensen JD, Stephan W. 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics 185(3):907–922.

Pool JE, et al. 2012. Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. PLoS Genet. 8(12):e1003080.

Pool JE, Aquadro CF. 2006. History and structure of sub-Saharan populations of *Drosophila melanogaster*. Genetics 174(2):915–929.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155(2):945–959.

Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559–575.

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.

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16(6):702–712.

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.

Thornton KR, Jensen JD. 2007. Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175(2):737–750.

Associate editor: Brandon Gaut