A Large Panel of *Drosophila simulans* Reveals an Abundance of Common Variants

Sarah A. Signor[1,*], Felicia N. New[2], and Sergey Nuzhdin[1]

[1]Department of Molecular and Computational Biology, University of Southern California

[2]Department of Molecular Genetics and Microbiology, University of Florida College of Medicine

*Corresponding author: E-mail: ssignor@usc.edu.

Accepted: December 7, 2017

Data deposition: This project has been deposited at the SRA under accession SRP075682.

Abstract

The rapidly expanding availability of large NGS data sets provides an opportunity to investigate population genetics at an unprecedented scale. *Drosophila simulans* is the sister species of the model organism *Drosophila melanogaster*, and is often presumed to share similar demographic history. However, previous population genetic and ecological work suggests very different signatures of selection and demography. Here, we sequence a new panel of 170 inbred genotypes of a North American population of *D. simulans*, a valuable complement to the DGRP and other *D. melanogaster* panels. We find some unexpected signatures of demography, in the form of excess intermediate frequency polymorphisms. Simulations suggest that this is possibly due to a recent population contraction and selection. We examine the outliers in the *D. simulans* genome determined by a haplotype test to attempt to parse the contribution of demography and selection to the patterns observed in this population. Untangling the relative contribution of demography and selection to genomic patterns of variation is challenging, however, it is clear that although *D. melanogaster* was thought to share demographic history with *D. simulans* different forces are at work in shaping genomic variation in this population of *D. simulans*.

Key words: *Drosophila simulans*, population genetics, selection.

Introduction

The recent influx of high throughput sequencing has enabled an entirely new scale of examination of the effects of demography and selection on patterns of genomic variation. Where before these analyses were often based on a handful of loci sampled from a limited number of individuals, data sets are becoming available that provide genome-wide information from potentially hundreds of individuals (Grenier et al. 2015; Lack et al. 2015, 2016). To these experiments, we can apply the population genetic models that were developed in the era of limited sequencing, including an examination of standing variation, linkage disequilibrium, and selection.

Drosophila melanogaster has been the model species for population genetic research since the advent of the field, providing evidence for pervasive adaptation at the molecular level (Aguade et al. 1989; Begun and Aquadro 1992; Fay et al. 2002; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Andolfatto 2005; Shapiro et al. 2007; Eyre-Walker and Keightley 2009; Sella et al. 2009; Schneider et al. 2011; Kousathanas and Keightley 2013; Garud et al. 2015). *Drosophila simulans*, the closest relative of *D. melanogaster*, is of particular interest to population genetics because both *D. melanogaster* and *D. simulans* are presumed to have a similar demographic history, yet population genetic and ecological work suggests very different signatures of selection and demography. They are both thought to have originated in or around Africa (*D. melanogaster*, southern-central Africa, *D. simulans*, Madagascar) followed by an out-of-Africa expansion ~10,000 years ago, with a subsequent colonization of the Americas in the past few hundred years (David and Capy 1988; Lachaise et al. 1988; Baudry et al. 2004; Li and Stephan 2006; Thornton and Andolfatto 2006). *D simulans* was recorded on the east coast of North America in 1920 (Sturtevant 1920). A large survey of the Pacific coast at the same time did not find evidence of *D. simulans* (Sturtevant 1920).

Differences in the patterns of divergence in the *D. melanogaster* and *D. simulans* lineages have been observed that

suggest interesting differences in biology between the two species (Andolfatto 2005; Begun et al. 2007; Haddrill et al. 2008). *Drosophila melanogaster* shows strong clinal differentiation while *D. simulans* does not (Weeks et al. 2002; Hoffmann and Weeks 2007; Schmidt and Paaby 2008; Machado et al. 2015; Sedghifar et al. 2016). The species differ in their habitat use and seasonal abundance, with some evidence that *D. simulans* may come from a more temperate ancestral climate than *D. melanogaster* (Parsons 1977; McKenzie and McKechnie 1979; David and Van Herrewege 1983; Singh 1989; Milan et al. 2012; Machado et al. 2015; Sedghifar et al. 2016). In general, non-African *D. simulans* have been found to have higher within population diversity compared with *D. melanogaster*, and to be less geographically differentiated (Singh 1989; Machado et al. 2015; Sedghifar et al. 2016). In the past, work on American populations of *D. melanogaster* has suggested a complex demographic history of bottlenecks, admixture, and selection (Glinka et al. 2003; Haddrill et al. 2005; Corbett-Detig and Hartl 2012; Langley et al. 2012; Garud et al. 2015; Lack et al. 2015, 2016; Pool 2015). Past studies on particular loci or gene regions in African, Asian, and American *D. simulans* have found haplotypes at intermediate frequency, though the authors diverge in their conclusions that this is due to population contraction, admixture, directional selection, or balancing selection (Hasson et al. 1998; Hamblin and Veuille 1999; Rozas et al. 2001; Quesada et al. 2003; Sanchez-Gracia and Rozas 2007). Begun et al. (2007) detected an excess of high-frequency polymorphisms in *D. simulans*, suggesting recurrent adaptive evolution.

The importance of different types of selection for adaptation, such as soft versus hard sweeps, and the expectations for their genomic signature is contentious (Hermisson and Pennings 2005; Przeworski et al. 2005; Barrett and Schluter 2008; Macpherson et al. 2008; Karasov et al. 2010; Kelly et al. 2013; Messer and Petrov 2013; Ferrer-Admetlla et al. 2014; Jensen 2014; Garud et al. 2015; Schrider et al. 2015). For example, whether soft sweeps are an important part of adaptation has been debated in the literature for 20 years, including whether or not they can be reliably distinguished from hard sweeps (Hermisson and Pennings 2005, 2017; Przeworski et al. 2005; Pennings and Hermisson 2006a,b, Pritchard et al. 2010; Messer and Petrov 2013; Jensen 2014; Garud et al. 2015; Schrider et al. 2015). It has remained a central question in population genomics, as it addresses fundamental questions about the tempo and mode of adaptation. Disentangling the signatures of different demographic events and selective sweeps is very difficult, as many scenarios will create similar patterns of variation and linkage. Deep sequencing of local populations increases our ability to recognize and differentiate between models of demography and selection, both because measures of haplotype diversity, linkage disequilibrium (LD), and the frequency of polymorphisms will more accurately represent the population value and because it provides more power to detect soft sweeps (Akey et al. 2004; Ross-Ibarra et al. 2008; Ramirez-Soriano and Nielsen 2009; Tennessen et al. 2010, 2012).

We investigate patterns of variation in *D. simulans*, and look for evidence of demography and selection in shaping this variation. We sequenced a new panel of 170 genotypes of *D. simulans* collected from a large organic orchard population in Zuma Beach (CA). We summarize patterns of variation in this population using Tajima's *D*, π, and LD. We find an abundance of common variants, leading to pervasive positive Tajima's *D*. We simulate several demographic and selective scenarios to investigate the potential causes of the patterns of variation in this population. We look for evidence of admixture within the population using two types of tests (PCA and ADMIXTURE). We examine the haplotype structure in the population, to better understand patterns of polymorphism, their frequency, and relationship to LD in the population. The authors intend this panel to be a valuable addition to the *Drosophila* community, complementing the DGRP (Mackay et al. 2012) and the DSPR (King et al. 2012) by enabling GWAS analyses and comparison between two closely related species.

Materials and Methods

All analysis scripts and documentation to reproduce results are located at: https://github.com/signor-molevol/signor_popgen_2016.

Drosophila Strains

The *D. simulans* lines were collected in the Zuma organic orchard in Zuma beach, California on the consecutive weekends of February 11th and 18th of 2012 from a single pile of fermenting strawberries. Single-mated females were collected and their offspring identified as being *D. simulans*. This was followed by 15 generations of full sib mating of their progeny. These strains are available upon request.

Library Preparation

DNA was extracted from 20 whole female flies using the Gentra Puregene kit (Qiagen). Libraries were prepared by RAPiD-Genomics, LLC using standard protocols and sequenced at the USC Genome Center. Prior to sequencing, the samples were pooled according to their molar concentrations for equal representation. Reads were single end 100 bp, sequenced on the illumina Hiseq across 10 lanes per 96 samples. The intended coverage per library was 10×, although ultimately some libraries were over or underrepresented. Raw data have been deposited to the Sequence Read Archive, accession number: SRP075682.

Mapping and SNP Discovery

Libraries were demultiplexed by barcode using standard protocols. Libraries were trimmed and cleaned using SolexaQA (v. 2.2), assembled using BWA mem (v. 0.7.5), and processed with samtools (v. 0.1.19) using default parameters (Cox et al. 2010; Li 2015). The reference genome was *D. simulans* version 2.01 as described in Hu et al. (2013). PCR duplicates were removed using Picard MarkDuplicates (v. 1.89) and GATK (v. 3.3) was used for indel realignment and SNP calling using default parameters (http://picard.sourceforge.net) (McKenna et al. 2010). SNPs were called using Haplotypecaller to call SNPs jointly and GenotypeGVCFs to call SNPs individually for each genome (McKenna et al. 2010). There were few qualitative differences in the downstream results, and those from Haplotypecaller were used. Files were produced using default filtering (phred score of 30), no additional filtering was applied (McKenna et al. 2010). GATK is slightly biased toward the calling of heterozygous SNPs (0–0.005 minimum and maximum) such that it should be the most unbiased toward low-frequency SNPs among the available SNP calling software (Hwang et al. 2015).

Filtering

SNPs that could not be placed on chromosomes (meaning SNPs belonging to fragmented portions of the *D. simulans* assembly) were excluded from the analysis, as were nonbiallelic loci, lines with >25% missing data, and loci with >10% missing data. The final data set included 170 genotypes and 5,998,575 loci. Of these SNPs, there were 1.2 million with no missing data. *Drosophila simulans* does not have segregating inversions (Ashburner and Lemeunier 1976), which is where the majority of heterozygosity is observed in *D. melanogaster* (Mackay et al. 2012; Grenier et al. 2015; Lack et al. 2015; Palmieri et al. 2015). Thus, while residual heterozygosity was observed it is not biased toward particular regions of the genome in *D. simulans* (supplementary fig. S1, Supplementary Material online). The residual heterozygosity did not alter our conclusions (see Sensitivity Analysis). In previous analyses, individuals with a genome-wide IBD >20% are removed from data sets of this type (Garud et al. 2015), however, none were observed in this data set according to PLINK pairwise IBD estimation (Purcell et al. 2007).

Summary Statistics

Nucleotide diversity (π), and Tajima's *D* can both provide insight into the demographic and selective history of a population. π Tajima's *D* and were calculated in 10-kb windows using VCFtools (v0.1.14) (Tajima 1989; Danecek et al. 2011).

Annotating SNPs

Differences in π and Tajima's *D* between gene regions that are presumed to be more or less neutral can be help distinguish between demography and selection. The gff files for the *D. simulans* v2 genome were used to annotate SNPs (exon, intron, 3′-UTR, 5′-UTR) with bedtools intersect (Quinlan and Hall 2010). The vast majority of introns were short, and they were split into groups based on length and presumed differences in selective regime (long introns > 120 bp, introns < 120 bp, bp 8–30 of introns < 65 bp) (Parsch et al. 2010). Annotated features <40 bp were excluded from the analysis, corresponding to <1% of annotated features.

Linkage Disequilibrium

The decay of linkage disequilibrium can help determine the appropriate scale with which to examine haplotype structure in a population, and can also help clarify the selective or demographic forces shaping variation in the population. LD (r^2) was calculated within windows of 10 kb using VCFtools (v1.14) geno-r2 and PLINK (v1.07) r2 (Purcell et al. 2007; Chang et al. 2015) with the LD filter set to 0 to report all pairwise comparisons within a window. We excluded 1-mb regions of low recombination at centromeres and telomeres. However, more was excluded if there were significant reductions in diversity for a broader region as seen in figure 1, calculated as extended negative Tajima's *D* or values of $\pi < 1/2$ the chromosomal average (Sedghifar et al. 2016). For this portion of the analysis only, we filtered out SNPs with very low and high frequency (.05 and .95) to avoid inflated estimates of LD from the effects of sampling (Nuzhdin and Turner 2013). Plots of LD were smoothed by averaging LD values binned according to the distance between SNPs, with two different bin sizes to reflect rapid decay of LD. From 0- to 300-bp bins were 20-bp nonoverlapping windows, while from 300- to 10-kb windows were 150 bp (Garud and Petrov 2016). While LD or Tajima's *D* alone can suggest certain modes of selection or demography, examining their association may reveal additional information. For example, if low LD is associated with high Tajima's *D*, this suggests that balancing selection may be mediating the relationship. The average LD value was compared with the value of Tajima's *D* for each window that was an outlier. Regression lines were fit for the data corresponding to positive *D* and negative *D* separately.

Sensitivity Analysis

It is possible that some patterns in the data could be produced by how it was processed, so we sought to investigate this possibility by evaluating the sensitivity of our analysis to different steps in the data processing. We lowered the phred score threshold from 30 to 20 to examine the effect on the results of Tajima's *D* and LD. To examine the effect of SNP calling method, we used both HaplotypeCaller to call SNPS jointly and GVCF to call them individually (McKenna et al. 2010). To investigate the effect of inbreeding on this population's parameters, 82 of the most inbred lines were reanalyzed (data not shown). The coefficient of inbreeding was

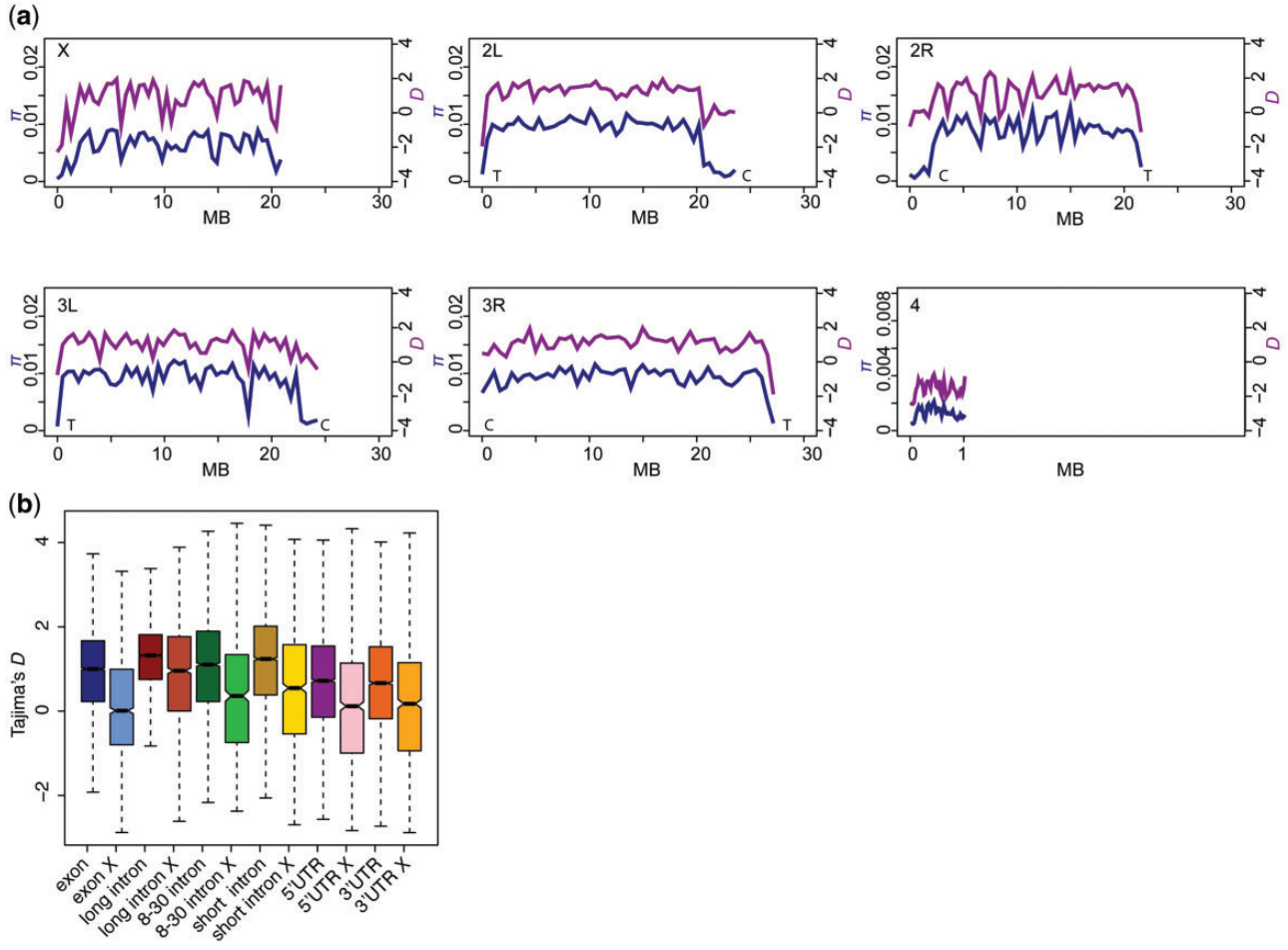

Fig. 1.—Summary statistics. (*a*) For each chromosome, Tajima's *D* (pink) and π (blue) are plotted in 10-kb windows. The genomic coordinates are split into 10-kb windows along the *x*-axis. Each data point represents a single window. Note that the axis for π (blue) is on the left-hand side while Tajima's *D* (pink) is on the right. For clarity, chromosome 4 is on a different *y*-axis for π, due to its small size. Location of centromeres and telomeres for each chromosome are indicated at the lower edges of each graph. All lines were smoothed using local regression in R. (*b*) The median and distribution (boxes are first and third quartile) of Tajima's *D* for each annotation category. Notches at the median are calculated as described in boxplot.stats in R and correspond roughly to the 95% confidence interval for the medians.

calculated using VCFtools (v0.1.14) (Danecek et al. 2011). The quality of variant calls was also compared with the level of heterozygosity for all lines. We examined the effect of masking heterozygous bases and excluding SNPs with missing data. We also used different software to calculate our metrics (data not shown). To evaluate the possibility that the results were due to protocols implemented within VCFtools (Danecek et al. 2011), three windows were run through a custom R script to calculate Tajima's *D* (Steige et al. 2015).

Simulations

We wanted to focus on potential demographic and selection scenarios that could produce the patterns of variation seen in our population, namely the positive Tajima's *D*. For example, a population bottleneck would produce negative values of Tajima's *D*, and thus is not considered here. Populations were simulated using MSMS (Ewing and Hermisson 2010) with a population size of 2×10^6 (Przeworski et al. 2001) for regions of 10^5 bp. We sampled 170 individuals to match the depth of our data and assumed a neutral mutation rate of $\mu = 3\times10^{-9}$ (Przeworski et al. 2001) and a recombination rate of $\rho = 5\times10^{-7}$ cM/bp (Presgraves 2005; Garud et al. 2015). *Drosophila simulans* is predicted to have a higher recombination rate than *D. melanogaster*, thus this value should be conservative for purpose of LD based statistics such as H12 (True et al. 1996). We assume ~10 generations per year (Li et al. 1999; Przeworski et al. 2001; Stephan and Li 2007; Garud et al. 2015).

Demographic Simulations

We simulated demographic models with 1×10^3 replicates. We simulated five potential admixture scenarios, as admixture

has been suggested as a potential explanation for the haplotype structure seen in some populations of *D. simulans* (Hamblin and Veuille 1999). No specific admixture regimes have been suggested for *D. simulans*, thus we simulated potential admixture by isolating the two populations beginning ~250 years ago and admixing in equal proportions ~50 years ago. In addition, we simulated a scenario in which the populations were isolated 5,000 and 500,000 years ago. We also simulated a combination of admixture and population contraction, where the populations separate 5,000 years ago followed by a bottleneck to 1% of the original population size for 30 years in one population. The other population is stable until 80 years before the end of the simulation, at which point there is a 30-year bottleneck to 1% of the original population size followed by admixture of each population in equal proportions. Lastly, we simulated a population in which a 200-year bottleneck to 1% of the original population size preceded splitting. Following this, the populations were isolated for 5,000 years, until 80 years before the end of the simulation when there is a 30-year bottleneck in one population followed by admixture in equal proportions.

We also simulated the simplest scenario, a constant population of 2×10^6 based on predictions for the effective population size of *D. simulans* (Przeworski et al. 2001). Population contraction has also been proposed as a potential demographic force in *D. simulans*, thus we simulated several potential population contractions to explore how the timing and length of contraction would affect the nucleotide frequency spectrum (Wall et al. 2002; SchÖfl and Schlötterer 2006). We simulated a population contraction of 1% for 120 years, 100 years before sampling. We simulated two population contractions to 0.1% of the original population size, one for 60 years and the other 120 years, both ending 100 years before sampling. Lastly, we simulated a 0.1% population contraction for 60 years, ending 40 years before sampling. The authors note that in our exploration of parameter space these demographic scenarios more closely mimicked our observed data, for example, longer bottlenecks resulted in excessive loss of variation compared with our population, less severe bottlenecks did not sufficiently elevate Tajima's *D*, and more severe bottlenecks (0.01%) created much greater variance in Tajima's *D* than seen in our population. However, we are not trying to uncover the exact demographic history of *D. simulans*, but rather understand the general forces that may have been shaping variation in this population.

Simulations of Selection

A summary of the simulations performed can be seen in supplementary table S2, Supplementary Material online, though only some of these simulations will be considered in full. We simulated a hard sweep and an incomplete sweep without demography, $s = 0.1$ and conditioning upon ending frequencies of 1 and 0.5, respectively. We simulated a complete soft sweep without demography with $s = 0.1$ and $\theta = 0.3$ where $\theta = 4N_e\mu$. We also simulated soft sweeps from standing variation with different starting frequencies, either 5×10^{-6} or 1.25×10^{-5}, and different starting times (either 40 or 70 years ago) (supplementary table S2, Supplementary Material online). For the remainder of the simulations, we included demography, in this case population contractions, as they more closely recapitulated the observed data than the other demographic scenarios we simulated. This included a subset of the aforementioned population contractions—either 120 years ending 100 years ago or 60 years ending 40 years ago. The contractions were either to 1% or 0.1% of the original population size. Selection commenced before, during, or at the completion of the contractions. The selection coefficient was always 0.1, though for simulations of balancing selection this applied to the heterozygote. This value was chosen for *s* as it more closely recreated the haplotype structure seen in *D. simulans* than lower values. For all simulations except balancing selection, we assumed codominance, where the homozygous individual carrying two copies of the beneficial allele has twice the fitness of the heterozygous individual. The range of mutations rates for sweeps was between 0 and 0.5, where anything 0.2 and below was primarily hard. The beneficial allele was always placed at the center of the locus. For all MSMS simulations, Tajima's *D* was output using the –oTPi flag within the program, and H12 was calculated for each simulated population using scripts from (Crisci et al. 2016).

Admixture

PCA

We wanted to investigate possible admixture in the population, where a PCA with either a long tail or a clear division into separate groups would suggest admixture (Ma and Amos 2012). To ensure that uncorrected LD did not distort, the PCA SNPs were thinned using VCFtools (v1.14) (Danecek et al. 2011) to intervals of 500 bp. The data set was also filtered to contain only autosomes. The PCA analysis was performed using SNPRelate (Zheng et al. 2012) and standard commands.

Admixture

The possibility of admixture can also be evaluated using maximum likelihood estimation of ancestry. The vcf file containing all chromosomes was converted to bed format using bcftools (v1.19) and PLINK (v1.07) (Purcell et al. 2007; Li et al. 2009). Population structure was estimated by using the ADMIXTURE program, a model based estimation of ancestry that uses maximum likelihood (Alexander et al. 2009). The best model was determined by selecting the value of *K* with the lowest cross-validation error.

H12

To identify selective sweeps across the chromosome arms, we implement a haplotype homozygosity test (H12) (Garud et al. 2015). H12 considers the frequency of the two most frequent haplotypes, where H2 is the haplotype homozygosity for the second most common haplotype only. When a sweep is hard, H2 should be very small, because only one haplotype should be present at a high frequency, that is, H1 will be high. For soft sweeps, H2/H1 will increase as a sweep becomes increasingly soft. For this analysis, heterozygous genotypes were coded as missing data, and sites with missing data were excluded from the analysis. SNPs were analyzed in windows of 400 SNPs with steps of 50 SNPs. We calculated H12 for each of the demographic scenarios described above to determine if our choice of window size was conservative. Haplotypes were grouped only if they matched at every site. We grouped together tracts of elevated H12, choosing the highest value within the tract to represent the peak. We defined the edge coordinates of each peak as the SNPs with the largest and smallest coordinates across the contiguous windows (Garud et al. 2015; Garud and Petrov 2016).

To confirm the expected relationships between the test statistics, we compared H12 and H2/H1 with Tajima's *D*. We plotted the top 50 H12 windows for each chromosome arm, and the H2/H1 values for those top H12 windows, against the average Tajima's *D* for each window of H12. The edge coordinates of the windows were defined as the largest and smallest coordinates across contiguous windows. Tajima's *D* was calculated in windows of 10 kb, thus they were matched to the windows of H12 as the first and last window of Tajima's *D* that overlapped the edge coordinate of the H12 window.

Results

Six Million Variants Identified in *D. simulans*

SNPs were filtered for missing data and a phred score of at least 30, and in the resulting data set the call rate was 96%. We recovered 5,998,575 high-quality biallelic single nucleotide polymorphisms (SNPs), at an average of 5.06 polymorphisms per 100-bp window across the genome. About 15 generations of inbreeding is not expected to remove all variation from individual lines, and selection on deleterious mutations can also strongly counteract inbreeding (Wang et al. 1999). No unexpected relatedness was observed. The vcf files were deposited at https://zenodo.org/communities/genetics-datasets/? page=1&size=20.

π is Consistent with Previous Estimates in *D. simulans*

In comparison with *D. melanogaster* estimates from European, African, Asian, and American populations ($\pi = 0.003$–0.0084) (Langley et al. 2012; Mackay et al. 2012; Grenier et al. 2015; Lack et al. 2016), the Zuma population of *D. simulans* exhibits higher π ($\pi_{avg} = 0.009$, $\pi_{max} = 0.018$), consistent with other estimates in *D. simulans* ~0.01 (fig. 1*a* and table 1) (Begun et al. 2007; Kofler et al. 2015). In European and US populations of *D. melanogaster*, variation on the X chromosome is strongly depleted, even after correction for its reduced population size ($X_{4/3}$) (Langley et al. 2012; Mackay et al. 2012). This varies considerably by population (Grenier et al. 2015), and is most often found in derived non-African populations where complex demographic scenarios are likely responsible (Singh et al. 2007). In *D. simulans*, variation on the *X* is less depleted than in *D. melanogaster*, but *D. simulans* also exhibits statistically significant depletion of variation on the *X* (average π for autosomes = 0.00918, average for *X* = 0.00822 [corrected for reduced N_e]; *t*-test $P < 2.2e\text{-}16$; Begun et al. 2007). The fourth chromosome in *D. melanogaster* does not recombine and has very little variation (Wang et al. 2002; Grenier et al. 2015). There is more recombination on the *D. simulans* fourth chromosome (Wang et al. 2004), and we observe less depletion of variation ($\pi_A = 0.0092$, $\pi_4 = 0.0012$) compared with *D. melanogaster* ($\pi_A = 0.005$–0.0075, $\pi_4 = 0.0004$–0.0015) (Wang et al. 2002; Langley et al. 2012; Mackay et al. 2012) (table 1).

Tajima's *D*

We find a surprising abundance of intermediate frequency variants (positive Tajima's *D*) and relatively fewer regions with an excess of low-frequency variants (negative Tajima's *D*) (fig. 1*a*). Using Tajima's conservative critical values for $n = 175$ (-1.765~2.095) (Tajima 1989), we find that between 5% and 10% of the windows of positive Tajima's *D* are significant (table 1). This is an unusual result given previous studies of *Drosophila* that recovered largely negative values of Tajima's *D* (Parsch et al. 2001; Wall et al. 2002; Braverman et al. 2005; Ometto et al. 2005; Nolte and Schlötterer 2008; Fabian et al. 2012; Langley et al. 2012; Mackay et al. 2012; Campo et al. 2013; Hübner et al. 2013).

Tajima's D Varies between Gene Regions

We calculated Tajima's *D* for each gene region (exon, long intron >120, short intron <120, bp 8–30 of introns <65, 5′-UTR, 3′-UTR) split between the X and autosomes to determine if different regions exhibit different patterns of Tajima's *D* (fig. 1*b*). Tajima's *D* was calculated in windows using gene regions within 10 kb of one another, for example, all introns within a gene that is <10 kb would be included in a single window. Short introns are thought to be evolving the most neutrally, though they will still be effected by demographic events (Parsch 2003; Parsch et al. 2010). The median and distribution of each category is shown in figure 1*b*. Long introns on the autosomes have the largest values of

Table 1
Summary Statistics

Summary Statistics	2L	2R	3L	3R	*X*	4
SNP density (per 100 bp)	5.010	4.850	5.090	5.460	4.870	3.160
π avg.	0.009	0.009	0.009	0.009	0.008218[a]	0.001
π median	0.010	0.009	0.010	0.010	0.006	0.001
π min.	0.000	0.000	0.000	0.000	0.000	0.000
π max.	0.018	0.018	0.017	0.018	0.013	0.003
Tajima's *D* avg.	1.114	1.104	1.075	1.122	0.650	−1.513
Tajimas *D* med.	1.262	1.231	1.169	1.179	0.813	−1.529
Tajima's *D* min.	−2.565	−2.509	−2.380	−2.656	−2.906	−2.648
Tajima's *D* max.	3.394	3.948	4.443	3.229	3.190	−0.075
% significant positive *D*	8.439	10.143	6.966	5.778	8.297	0.000
% significant negative *D*	0.763	0.510	0.289	0.258	4.124	33.330
% positive *D*	86.047	85.178	88.746	94.479	73.381	0.000
% negative *D*	9.033	9.587	7.007	5.263	26.139	100
% neutral *D* (*D*=0)	4.920	5.230	4.250	0.258	0.480	0.000
LD avg.	0.123	0.120	0.113	0.116	0.160	
LD median	0.047	0.050	0.042	0.042	0.070	

Note.—For each chromosome or chromosome arm, we have summarized the population genetic parameters for the Zuma population of *Drosophila simulans*. This includes SNP density, π, Tajima's *D*, and linkage disequilibrium. For the X chromosome, π is scaled by 4/3 to account for its reduced population size.
[a]Scaled by 4/3.

Tajima's *D* (mean 1.24), and it is larger than either 8- to 30-bp introns <65 bp on the autosomes (*P* value < 2.2e-16, Mann–Whitney *U*-test) or exons on the autosomes (*P* value < 2.2e-16, Mann–Whitney *U*-test). The variance in the distribution of Tajima's *D* is different for long introns compared with 8- to 30-bp introns < 65 bp (0.8 and 1.4, respectively, supplementary fig. S2, Supplementary Material online). The difference is significant between long introns and introns <120 bp on the autosomes, though it is less significant than the other comparisons (*P* value = 0.0001479 Mann–Whitney *U*-test). The difference between the X and the autosomes was significant for every category (*P* value < 2.2e-16, Mann–Whitney *U*-test). However, no significant differences are observed between autosomal 5′-UTRs and 3′-UTRs (*P* value = 0.09, Mann–Whitney *U*-test), between X 5′-UTRs and 3′-UTRs (*P* = 0.438, Mann–Whitney *U*-test), or between exons on the X and 5′-UTRs or 3′-UTRs on the X (5′-UTR *P* = 0.4299, 3′-UTR *P* = 0.0815, Mann–Whitney *U*-test).

Comparison with Simulated Populations

The proposed demographic scenarios produce a wide range of Tajima's *D* values (fig. 2*a* and *b*). The parameters chosen to model pure admixture did not produce a positive Tajima's *D*, rather increasing amounts of separation produced increasingly negative Tajima's *D*, thought the effect is not large (fig. 2*a*). Bottlenecks in both admixed populations did not have an appreciable effect on Tajima's *D*, whereas a bottleneck in the population prior to splitting, followed by a bottleneck in one lineage, created a very negative Tajima's *D* (fig. 2*a*). This is a small sample of the many possible demographic scenarios incorporating admixture, thus it does not rule it out as a potential cause of the observed patterns. However, these simulations do not provide support for admixture as an underlying cause.

A constant population size maintained an average Tajima's *D* of zero as expected (fig. 2*b*). Population contraction had the largest effect on Tajima's *D* out of the simulations performed, with contractions to 0.1% ending either 100 or 40 years prior to sampling producing roughly the same distributions (fig. 2*b*). Contraction to 0.1% for 120 years elevated Tajima's *D* above that observed, whereas a 1% contraction for 120 years did not elevate it sufficiently. Notably, none of the demographic scenarios produce the variance seen in the *D. simulans* population (e.g., variance of Tajima's *D* for a contraction of 1% for 120 years = 0.01, *D. simulans* = 0.8) (fig. 2*b*).

Many different demographic and selective scenarios could create the patterns seen in *D. simulans*. We explored simulations of selection and selection with demography, focusing on scenarios that more closely recapitulated our data (fig. 2*c*). Selection without demography, both hard, incomplete, and soft sweeps, produced a negative spread of Tajima's *D* values but did not produce the elevated Tajima's *D* seen in our population (fig. 2*c*). It is possible that other scenarios not modeled, or scenarios such as balancing selection with competing mutations rather than heterozygote advantage, could produce more elevated values of Tajima's *D*. Simulations of selection including demography, in this case a population contraction to 0.1% for 60 years, 40 years prior to sampling, produced elevated Tajima's *D* with means and variances more similar to that observed in our population (fig. 2*c*). Selection from standing genetic variation with a starting frequency of 20 copies, starting during the population contraction, had a

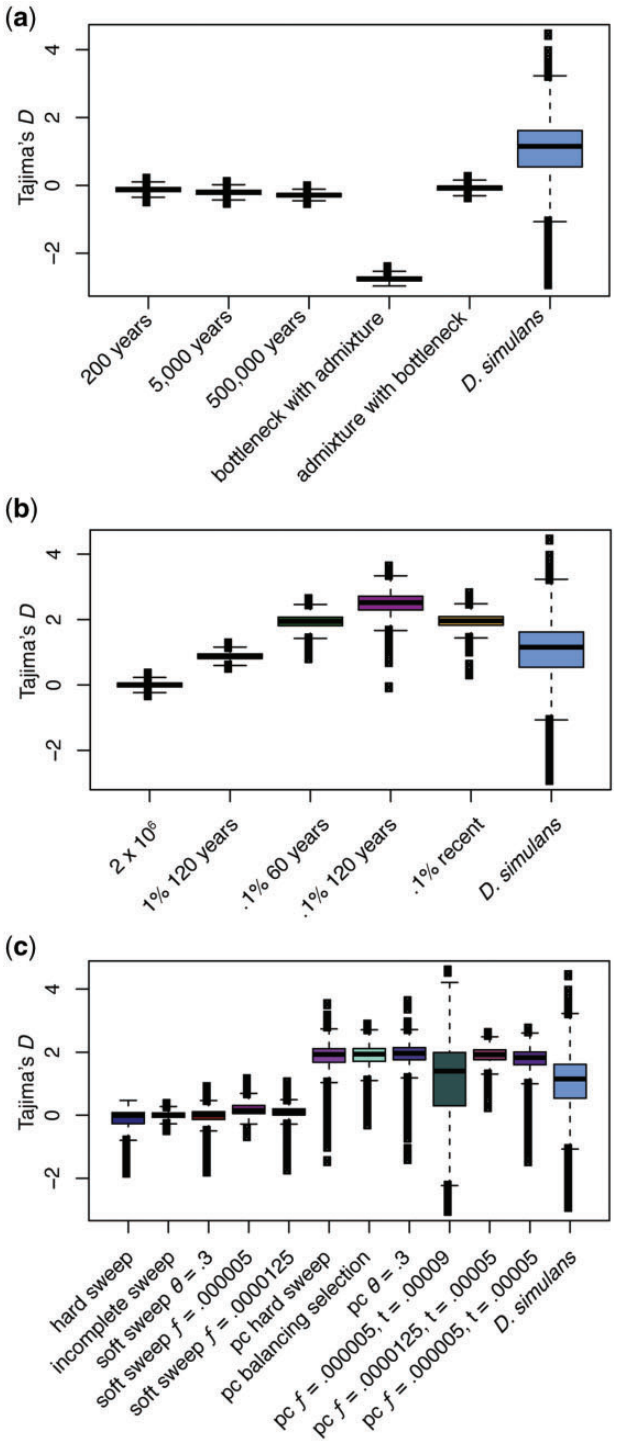

Fig. 2.—Each simulation was performed for a region 10^5 bp and Tajima's *D* was calculated in windows of 10 kb. Recombination was $\rho = 5\times10^{-7}$ and mutation $\mu = 3\times10^{-9}$. Here, Tajima's *D* in demographic simulations is shown compared with *Drosophila simulans*. (*a*) Admixture simulations, including admixture with a bottleneck prior to splitting, and a bottleneck along one lineage prior to admixing (bottleneck with admixture), and admixture with bottlenecks in both populations after isolation (admixture with bottleneck). (*b*) Constant population size and population contractions of different degrees and durations. A recent bottleneck to 0.1% for 60 years resembles the data as much as slightly older bottlenecks, and is used in concert with the simulations of selection in the following figure. (*c*) Tajima's *D* in different simulations of selection compared with *D. simulans*. Not all of the simulations listed in supplementary table S2, Supplementary Material online, are included here, only those that most resemble the *D. simulans* data or illustrate the difference between selection alone and selection with demography. All population contractions were to 0.1% of the population for 60 years, terminating 40 years prior to sampling. For all simulations $s = 0.1$, f refers to the starting frequency of the selected mutation, and t is its time of introduction with units of $4N_e$.

mean of 1 much as *D. simulans*, though a larger variance ($t = 0.00009$, 1.6 vs. 0.8). Selection on standing variation with either 50 or 20 copies with selection concurrent with the end of the population contraction had larger means ($t = 0.00005$, 1.7–1.9) and smaller variances (0.08–0.3). A hard sweep had a lower mean (0.5) but a more similar variance to *D. simulans* (0.6). When demography is incorporated with selection in MSMS the timing of selection must be set, thus it is possible that sweeps that finished more or less recently would present more similar values to *D. simulans*.

Linkage Disequilibrium

We assessed the scale of LD decay in the *D. simulans* data set using r^2 to calculate pair-wise LD at distances from a few base pairs to 10 kb (fig. 3*a*). LD is shown on a log scale to better illustrate smaller values. We found a rapid decay of LD within 200 bp on all autosomes, much as in *D. melanogaster* (Garud et al. 2015) (fig. 3*a*). It should be noted that our estimates of LD are likely somewhat inflated relative to recent estimates from *D. melanogaster*. We do not have detailed knowledge of recombination rate and potential admixture, both of which have been used to filter *D. melanogaster* data sets, excluding >50% of the genome in some cases (Garud et al. 2015; Garud and Petrov 2016; Lack et al. 2016). Admixture may increase or decrease LD depending on the scenario, but regions of reduced recombination could not be excluded here other than at centromeres and telomeres. All other filtering that we performed prior to calculating LD was the same as that performed in *D. melanogaster*. We compared the relationship between LD and positive values of Tajima's *D* and found that it is slightly positive (*P* value < 0. 001; fig. 3*b*). Between LD and negative values of Tajima's *D*, the relationship is negative, meaning that lower values of Tajima's *D* have a higher LD (fig. 3*b*). The purpose of this comparison is to guide the investigation into potential demographic and selective forces shaping the population, for example, if a high Tajima's *D* was associated with a low LD then long-term balancing selection would be a natural area to investigate. This is not the pattern we observe, so we will focus on selective

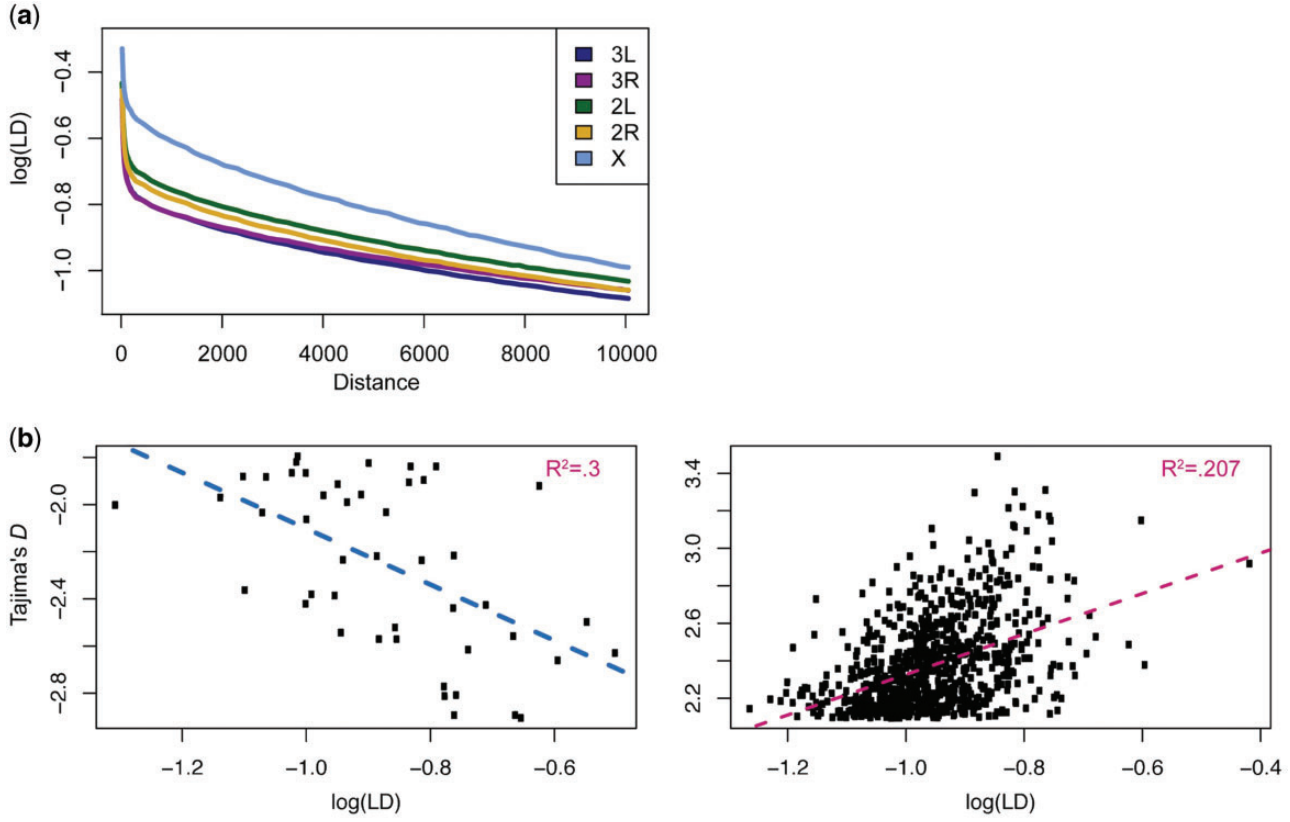

Fig. 3.—Linkage disequilibrium. (*a*) Decay of linkage disequilibrium on the X and autosomes calculated by distance. LD is binned initially in windows of 20 bp until 300 bp and 150 bp thereafter. Shown is the log(LD) of the mean of each of these bins. Regions of low recombination have been filtered out. (*b*) The relationship between LD and outliers of Tajima's *D* with regression lines plotted for negative and positive values of Tajima's *D* separately. This is to highlight the possibility for different relationships between LD and Tajima's *D* depending upon the selective regime—for example, that lower (higher) values of Tajima's *D* may be associated with higher (lower) values of LD. The R^2 value of for each of these lines is shown in the upper right hand corner of each graph.

scenarios such as short-term balancing selection and hard or soft sweeps.

Sensitivity Analysis

To confirm that our results were robust, we performed a number of different sensitivity analyses. The GATK pipeline is known to call SNPs more promiscuously than other pipelines, thus biasing in favor of rare variants. To test the effect of increasing the number of rare variants, we lowered the quality threshold to a phred score of 20 (data not shown). In addition, we called SNPs using a different SNP calling protocol, GenotypeGVCFs, which calls SNPs individually for each genotype (data not shown). We also examined the effect of masking heterozygous bases and excluding SNPs with missing data (supplementary text S1 and fig. S3, Supplementary Material online). None of these calculations had a qualitative effect on our results. We also recalculated Tajima's *D* using R scripts (courtesy of Benjamin Laenen) to determine if there was a bias arising from with the software (VCFtools). We found no differences in the values of Tajima's *D* (data not shown). Lastly, we examined the relationship between heterozygous calls and call quality and found no relationship (supplementary text S1 and fig. S3, Supplementary Material online). These results support the conclusion that Tajima's *D* is robust to biases arising from technical issues. Any residual bias should be toward calling more rare SNPs.

To determine if the level of heterozygosity was affecting our results, we recalculated our summary statistics using subsets of individuals with different levels of inbreeding (data not shown). This will also help us to determine if our results are being driven by an unusual subset of individuals within the panel. We reduced our data set to the 82 most inbred genotypes and recalculated all of the summary statistics. Again, our results were unaffected, supporting the conclusion that the observed patterns are not the result of methodological or computational artifacts.

Admixture

Admixture could cause the elevated Tajima's *D* in our population. A PCA analysis found no evidence of admixture, as

there is no tail to the distribution as would be expected if individuals were sampled from an admixed population (supplementary text S1 and fig. S4, Supplementary Material online) (Ma and Amos 2012). It is possible that every individual is equally admixed and therefore appearing as one population. We also performed an analysis of population structure with the algorithm ADMIXTURE (Alexander et al. 2009). Using the value of *K* (number of populations) with the lowest cross validation score, we determined that the best model fit in ADMIXTURE for these samples is a single ancestral population (supplementary text S1 and fig. S4, Supplementary Material online). Admixture cannot be ruled out, but these results do not explicitly support admixture as being the reason behind the presence of intermediate frequency haplotypes in this population. It is a given that demography contributes to the patterns observed here, including the genome-wide elevated Tajima's *D*, so we will focus instead on outliers determined by simulations of selection and demography.

H12

To investigate potential sweeps occurring in this population, we will use a haplotype diversity test (H12) (Garud et al. 2015; Garud and Petrov 2016). In this test, high values of H12 indicate a sweep, but the relative frequency of the second most frequent haplotype is indicated by the corresponding H2/H1 values (Garud et al. 2015). A high value of H12 with a correspondingly high value of H2/H1 indicates a higher frequency of the second haplotype. A high value of H12 and a low value for H2/H1 suggests that a second haplotype does not segregate at an appreciable frequency (Garud et al. 2015). In the *D. simulans* population, the mean H12 was 0.05, with a median of 0.04 (fig. 4). The highest values of H12 are 0.95, and are found on the X chromosome. The third largest value of H12 is 0.89, and corresponds to the interval containing *Cyp6g1*. *Cyp6g1* has been previously inferred to be involved in selective sweeps in response to insecticides (Schlenke and Begun 2004; Schmidt et al. 2010; Kolaczkowski et al. 2011; Garud et al. 2015; Sedghifar et al. 2016).

Comparison with Simulated Populations

Window size is an important parameter for this test, and is generally based on the extent of LD in a population. We note that while decreasing the window size in the demographic scenarios to 300 SNPs does increase the presence of haplotype structure, it does not do so enough to alter any of our conclusions. In addition, while the results are quantitatively changed by an increase in window size to 500 SNPs in the *D. simulans* population, they are not qualitatively different (data not shown). Admixture did not produce values of Tajima's *D* comparable with that seen in our population, thus we will not discuss it further here. However, the authors note that the scenarios we simulated did not produce appreciable haplotype structure (data not shown). A recent bottleneck to 0.1% did produce appreciable haplotype structure, with a top H12 of 0.21 and an H2/H1 of 0.04 (fig. 5). A recent bottleneck to 0.1% and a bottleneck ending 100 years ago to 0.1% produced similar distributions of Tajima's *D*, however, given the increased haplotype structure in the more recent bottleneck, we will focus on this demographic scenario when demography and selection are combined (fig. 5).

Simulations of selection or selection with demography produced a variety of H12 patterns. For genome-wide H12 values selection without demography produced elevation of H12 higher to that seen in *D. simulans*, despite not producing elevated Tajima's *D* (figs. 2 and 6). Soft sweeps from repeated mutation and hard sweeps had means (0.07, 0.08) higher than *D. simulans* (0.05). However, the median for *D. simulans* was higher (0.04), than for hard sweeps (0.025) or soft sweeps from repeated mutation (0.023). Population contraction with selection from standing variation occurring during the population contraction also had a higher mean ($t = 0.00009$, 0.074), whereas other simulations from standing variation with population contraction were lower ($t = 0.00005$, 0.036, 0.022) (fig. 6). However, for the same three simulations, the most similar median also occurred from selection on standing variation occurring during the population contraction ($t = 0.00009$, 0.035) versus from standing variation occurring at the end of contraction ($t = 0.00005$, 0.02, 0.017). The contributions of selection and demography are difficult to parse, and the authors do not claim to do so here. However, overall of the simulations which we performed, a population contraction combined with selection from standing variation where selection commences during the contraction produced the most similar values of both Tajima's *D* and H12 ($t = 0.00009$). None of the simulations produced sweeps as soft as in *D. simulans*, but the highest values of H2/H1 for the highest H12 were from population contraction with hard sweeps, soft sweeps from repeated mutation, and balancing selection (fig. 6). The timing of selection is important, for example, selection on repeated mutation that occurs after a population contraction generally results in a hard sweep due to lost variation (data not shown, supplementary table S2, Supplementary Material online), and selection on standing variation produces very different patterns depending upon whether it beings during the contraction or concurrent to its end (fig. 6). It is likely that more fully exploring this parameter space would yield additional insight into the patterns of variation seen in *D. simulans*.

H12 Outliers

A recent population contraction to 0.1% for 60 years produced the highest H12 of all of the demographic scenarios, with a maximum value of 0.17. The lowest value of the top 50 outliers for *D. simulans* is 0.24, we will consider these outliers

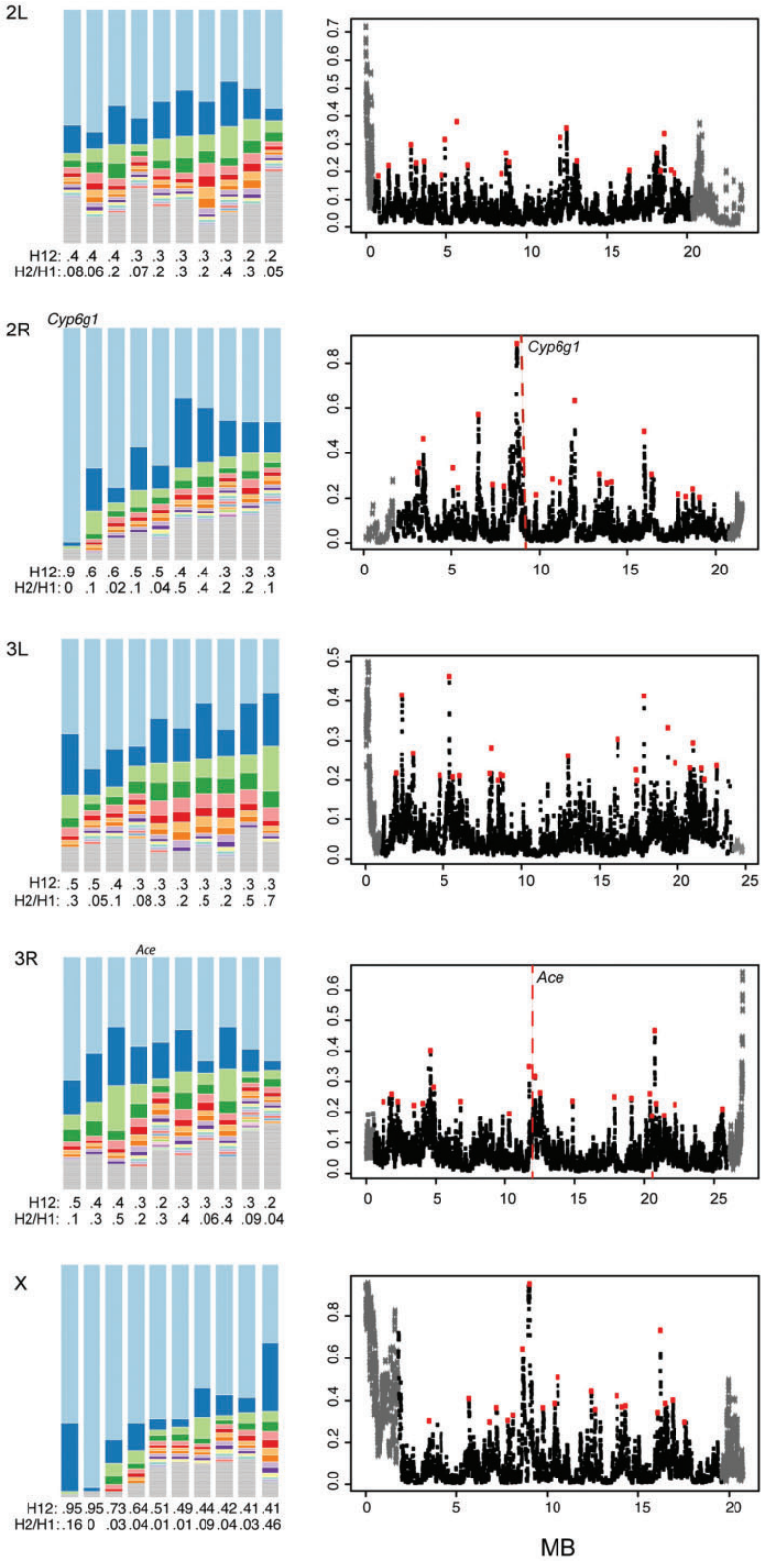

Fig. 4.—Haplotype Diversity. Shown are the results of the H12 scan for each chromosome other than the fourth, due to its overall reduced recombination rate. On the right-hand side are the scans in windows of 400 bases, with window centers separated by 50 SNPs. Gray regions indicate regions of the chromosome arms with reduced recombination that were not included in the final analysis. Red points indicate the top peaks for each chromosome. The dotted red lines indicate two positive controls from *Drosophila melanogaster*, *Cyp6g1*, and *Ace*. On the left-hand side, the haplotype frequency spectra for

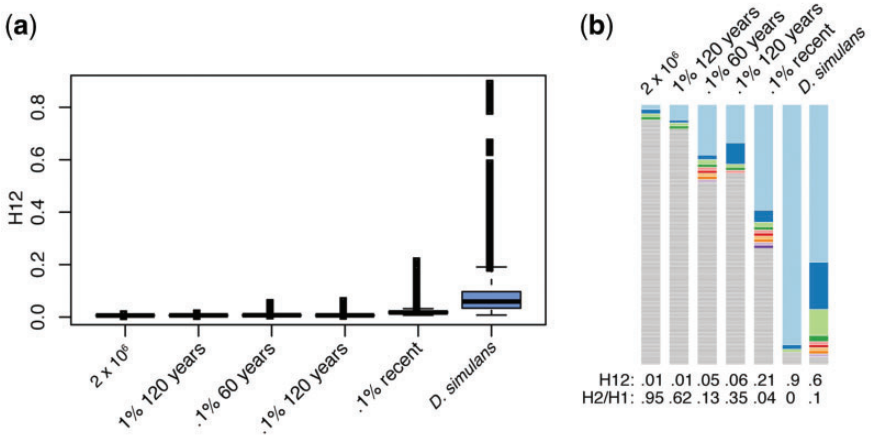

Fig. 5.—Demographic simulations and H12. (*a*) The spread of H12 values in each of the demographic scenarios that were simulated. Demography does not substantially elevate Tajima's *D* though population contraction does have more haplotype structure than the other demographic scenarios. (*b*) An example of the haplotype structure under each of the demographic scenarios, with the highest H12 value for each scenario shown. Again, the most frequent haplotype is shown in light blue and the second most frequent in dark blue. Gray regions indicate singletons.

to be less likely to be caused by demographic factors. To focus on outliers, we selected the 50 intervals with the highest H12 values and annotated all genes contained completely within each interval (supplementary table S1, Supplementary Material online). Some selective sweeps confirmed in previous studies overlapped with genes found in our top 50 intervals (Garud et al. 2015). For instance, *Cyp6g1* has repeatedly been implicated in sweeps relating to insecticide resistance, and among our sweeps, we find *Cyp6g1* within a window with one of the highest H12 scores (Schlenke and Begun 2004; Schmidt et al. 2010; Kolaczkowski et al. 2011; Garud et al. 2015; Sedghifar et al. 2016). *Ace* has also repeatedly been implicated in sweeps relating to insecticide resistance and it is within the top 50 peaks for chromosome 3R, but not the top 50 overall (fig. 4) (Fournier et al. 1992; Mutero et al. 1994; Menozzi et al. 2004; Sedghifar et al. 2016).

Tajima's D and H12

We wanted to understand the relationship between H12 and Tajima's *D*. They may be correlated under neutrality or selection, but it is also possible they are detecting separate phenomena. For example, are the regions of higher LD that would be detected using H12 the same regions with elevated intermediate frequency variation, or are those regions associated with lower haplotype structure? We plotted the top 50 H12 values from each chromosome arm against the average Tajima's *D* within that window. We also plotted the corresponding H2/H1 against Tajima's *D* for the same windows. H12 is less able to detect selection the more haplotypes are present (e.g., as might be found in ancient balancing selection), so one might expect a slight negative relationship between values of H12 and Tajima's *D*, which is also what we find ($P < 0.0001$) (fig. 7 and supplementary fig. S5, Supplementary Material online). However, the majority of the Tajima's *D* values for the top H12 windows are positive, though neither the positive or negative values are extreme. There is also the expected positive relationship between the H2/H1 values from the top 50 H12 and Tajima's *D*. This is the pattern we find on the X chromosome and all of the autosomes other than 3L ($P < 0.0003$–0.002, 3L $P < 0.148$) (fig. 7, table 1 and supplementary text S1 and fig. S5, Supplementary Material online). It is unclear why 3L is an exception to this pattern, although it does have more negative values of Tajima's *D* than the other autosomes (supplementary text S1 and fig. S6, Supplementary Material online).

What is most striking about the *D. simulans* data in comparison with the simulations is the high frequency of the second haplotype. We were not able to reproduce this pattern with any of the demographic or selective scenarios we investigated. Plots of this haplotype structure show that there is a single invariant haplotype at high frequency, followed by a second common haplotype that may or may not be quite diverged. That is the second most common haplotype is generally not a single mutational step away from the most common haplotype, it is often separated by many mutations. Examples of these haplotypes are shown in supplementary figure S7, Supplementary Material online, along with an example of a hard sweep.

Fig. 4.—Continued
the top ten intervals from each chromosome are shown. The height of the top shaded region in light blue indicates the proportion of that haplotype in the population. Colors after that indicate the second, third, and so on most frequent haplotypes while regions in gray indicate singletons. Most sweeps, with the exception of *Cyp6g1*, have second haplotypes sorting at appreciable frequencies.

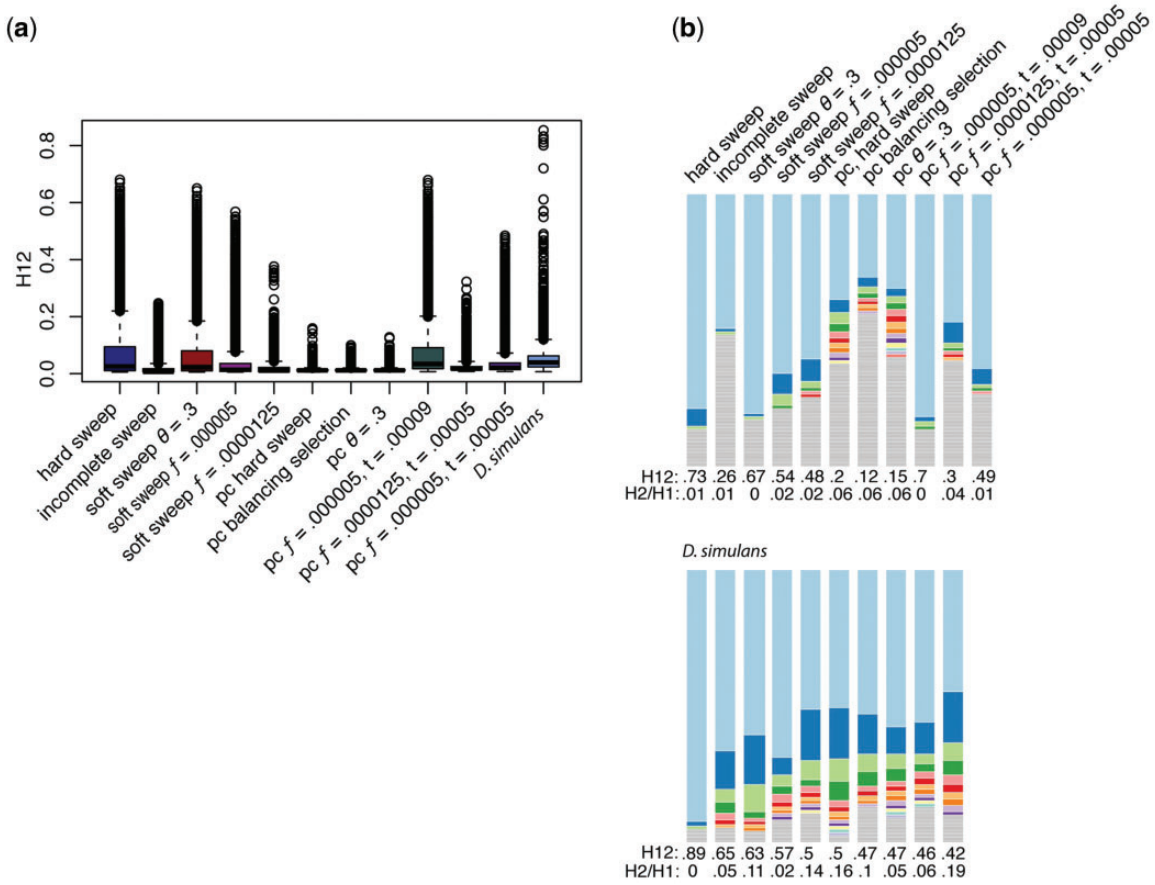

Fig. 6.—Simulations of selection and H12. (*a*) The spread of H12 values for each of the simulations of selection. While demography and selection elevated Tajima's *D* the most, they do not elevate H12 values the most. This may be due to the timing of selection, which had to be modeled differently for combinations of selection and demography. (*b*) Patterns of haplotype diversity in the top values of H12 from each simulation and *Drosophila simulans*. Note that we include the highest values of H12 for *D. simulans* from the autosomes, as these simulations pertain principally to autosomal variation. These selection scenarios are more similar to that observed in *D. simulans* than demography, however, the high frequency of the second haplotype in *D. simulans* was not recapitulated in any simulation.

Discussion

Demographic Scenarios

Our simulations of demographic scenarios suggest that this population of *D. simulans* could have undergone a population bottleneck. Other, unexplored, demographic scenarios could likely produce the same patterns, however, it is most consistent with the scenarios we explored. In addition, Tajima's *D* between different annotation categories (exons and 3′-UTR) do differ, though all are substantially elevated. Long introns are more positive than all other categories, with a reduced variance, suggesting a different selective regime for this category. However, presumably unconstrained bp 8–30 introns $<$ 65 bp are also positive, suggesting that a demographic scenario such as population contraction is involved (Parsch et al. 2010). *Drosophila simulans* is thought to have colonized the Americas ~100 years ago, which would have created a population bottleneck. In addition, there may have been bottlenecks associated with an out-of-Africa migration or the spread of *Wolbachia* strain *w*Ri (Turelli 1984; Turelli and Hoffmann 1995). Ecologically, there is no reason to expect seasonal cycling of population abundance that could lead to demographic signatures of contraction (Behrman et al. 2015). There is a year-round supply of food, as well as temperature changes that are well within the tolerable range for *D. simulans*. Furthermore, *D. simulans* has been collected in various locations in Southern California from January to November (Turelli and Hoffmann 1995; Ballard et al. 2008; San Diego Stock Center). This is not to say that there will be no seasonal selection on traits such as desiccation resistance, but there is likely no seasonal population collapse and expansion or recolonization (Behrman et al. 2015). In addition, *D. simulans* has a lower F_{ST} than *D. melanogaster* when geographical clines are sampled (Machado et al. 2015; Sedghifar et al. 2016). Less clinal variation in *D. simulans* compared with *D. melanogaster* could indicate more recent spread, reflecting the possibility of a recent population contraction and expansion in *D. simulans* (Machado et al. 2015; Sedghifar et al.

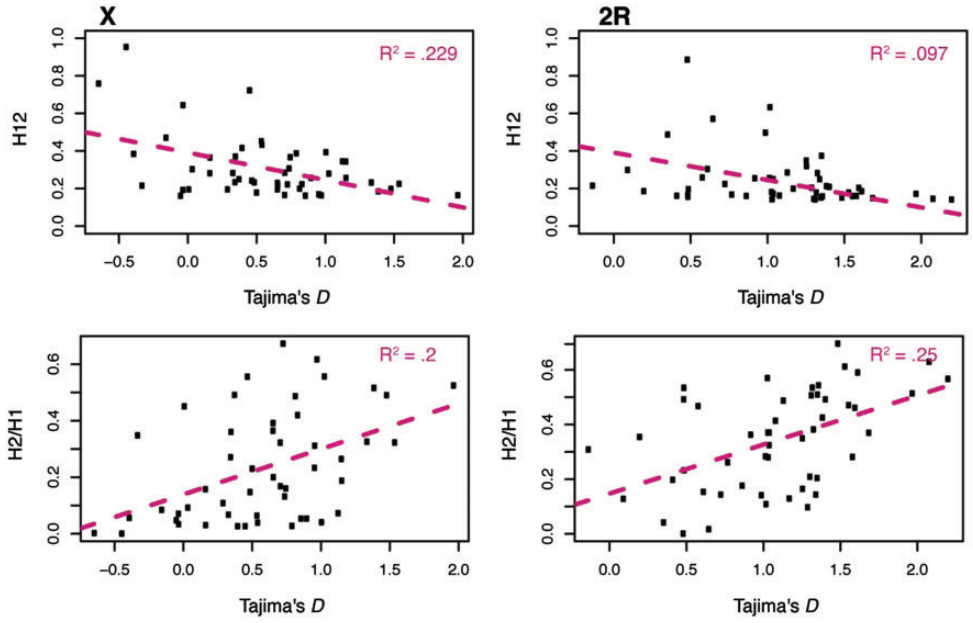

Fig. 7.—H12. (*a*) Tajima's *D* versus H12 and H2/H1. The average Tajima's *D* value was calculated for intervals identified as the top 50 H12 scores for each chromosome. Intervals are defined as the smallest and largest edges of each peak, thus adjacent intervals with high H12 scores are considered a single peak. The slight negative relationship between H12 and Tajima's *D* is significant for all chromosomes ($P < 0.0001$), and is expected given the decreasing ability of H12 to detect sweeps as they become softer (supplementary fig. S5, Supplementary Material online). A high value of H2/H1 and H12 indicates a soft selective sweep, so a positive relationship between Tajima's *D* and H2/H1 indicates that soft selective sweeps are responsible for the high Tajima's *D* values. The relationship is significant at $P < 0.0003$–0.002 for all chromosomes other than 3L (supplementary fig. S6, Supplementary Material online).

2016). It is possible that a different demographic scenario created this elevated Tajima's *D*, though of the simulations we performed it resembled the data more than other demographic scenarios.

A relative excess of intermediate frequency alleles is expected during some admixture scenarios (Gillespie and Turelli 1989). However, we did not find evidence of admixture in our PCA or ADMIXTURE analyses and the simulations of admixture that we performed. There are many possible scenarios for admixture, and it is possible that if the two populations were isolated for longer than 500,000 years that it would have produced a more pronounced change in Tajima's *D*. In addition, selection in each population prior to admixture may have elevated Tajima's *D*, though bottlenecks in the isolated populations did not. The particular scenarios simulated here tended to create more negative Tajima's *D* the longer the populations were separated. It is possible that a different combination of admixture and contraction could produce the observed values of Tajima's *D*. The evidence here does not support admixture, however, it cannot be ruled out as potentially contributing to the observed patterns in *D. simulans*.

Selection

Of the parameter space we explored, selection without demography was able to create high-frequency haplotypes and high H12 values, but did not elevate Tajima's *D* or create the variance in Tajima's *D* seen in our population. Selection alone could be responsible for the patterns observed in *D. simulans*, though the elevated Tajima's *D* across all categories of gene regions (introns and exons) suggests a role for demography. Hard sweeps, balancing selection, and soft sweeps from repeated mutation, combined with a population contraction, all created elevated Tajima's *D* and haplotype structure, though H12 was not as high as in *D. simulans*. Soft sweeps from standing variation produced similar patterns, though a sweep from standing variation that begins during the population contraction rather than at its conclusion increased the variance in Tajima's *D* in a way more similar to *D. simulans*. However, these sweeps tended toward single haplotypes at high frequency compared with *D. simulans*. The *D. simulans* H12 outliers are characterized by a high frequency of the second haplotype, which we were unable to recapitulate in any of our simulations. Overall, the timing of both population contractions and selection appear to be important in producing haplotype structure, but likely many different demographic and selective scenarios could create the observed patterns.

Overall H12 values in *D. simulans* were much higher than that observed in North American *D. melanogaster*, with a maximum of 0.89 in this panel and 0.24 in *D. melanogaster* (Garud and Petrov 2016). Of the top ten potential sweeps, there are more haplotypes at high frequency in North American *D. melanogaster*, with values as high of H2/H1 as 0.54 in *D. melanogaster*, whereas values of H2/H1 are only as

high as 0.3 in *D. simulans*. However, the top 50 sweeps in *D. simulans* contain values of H2/H1 as high as 0.7. North American *D. melanogaster* likely has a different demographic history than *D. simulans*, which certainly contributes to this difference, or the sweeps in *D. melanogaster* may be older. Previous work on clines of Australian and North American *D. simulans* found shared SNPs associated with extreme values of F_{ST} between populations, which suggested local adaptation from standing variation (Sedghifar et al. 2016).

Conclusion

Understanding the effect of demography and selection on variation in the genome is a difficult goal given the complexity of the potential realities of both factors. In general, sampling has been done sparsely across a species range, and while deep sequencing of a single population of a species is becoming more routine, few experiments exist to date of the depth considered here. Our observations are consistent with previous literature on *D. simulans*, which found invariant haplotypes at intermediate frequency in both African and non-African populations (Hasson et al. 1998; Hamblin and Veuille 1999; Rozas et al. 2001; Quesada et al. 2003; Sanchez-Gracia and Rozas 2007). In general, our results are suggestive of a contribution of both demography and selection to the patterns of variation observed in *D. simulans*. It is difficult to determine if this is the type of selection involved as the timing and strength of selection as well as its interaction with demographic factors can create similar patterns in response to different inputs. Of the simulations performed here, that which is most consistent with *D. simulans* is a soft sweep from standing variation combined with a population contraction.

Our insights into the Zuma *D. simulans* population's dynamics reveal a complex population exposed to a myriad of demographic and selective forces. Separating out the effects of demography and selection is one of the major goals of population genetics, and we cannot say definitely what is shaping the genome-wide patterns of diversity we observe in this population. However, the demographic simulations performed here are more suggestive of the conclusion that *D. simulans* has undergone a recent population bottleneck. In addition, among outlier regions of the genome with invariant haplotypes at high or intermediate frequency, the pattern of variation appears to be more consistent with soft sweeps than with hard sweeps, though this is not conclusive. Furthermore, California *D. simulans* populations appear to have a very different demographic and selective history than *D. melanogaster*, the causes of which will be interesting to disentangle in the future.

Supplementary Material

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

Acknowledgments

We thank Z. Polonus, Jeremy Newman, Miguel Ibarra, Alyona Sokolkova, and Benjamin Laenen for technical assistance and J. Butler, G. Mackeral, and J. Crisci for advice on the manuscript. This work was funded by grants GM102227 and MH091561 from the National Institutes of Health.

Literature Cited

Aguade M, Miyashita N, Langley CH. 1989. Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics 122(3):607–615.

Akey JM, et al. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2(10):e286.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9):1655–1664.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437(7062):1149–1152.

Ashburner M, Lemeunier F. 1976. Relationships within the melanogaster species subgroup of the genus *Drosophila* (Sophophora). I. Inversion polymorphisms in *Drosophila melanogaster* and *Drosophila simulans*. Proc R Soc B 193(1111):137–157.

Ballard JWO, Melvin RG, Simpson SJ. 2008. Starvation resistance is positively correlated with body lipid proportion in five wild caught *Drosophila simulans* populations. J Insect Physiol. 54(9):1371–1376.

Barrett R, Schluter D. 2008. Adaptation from standing genetic variation. Trends Ecol Evol. 23(1):38–44.

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 DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356:519–530.

Begun DJ, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5(11):e310.

Behrman EL, Watson SS, O'Brien KR, Heschel MS, Schmidt PS. 2015. Seasonal variation in life history traits in two *Drosophila* species. J Evol Biol. 28(9):1691–1704.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol Biol Evol. 21(7):1350–1360.

Braverman JM, Lazzaro BP, Aguade M, Langley CH. 2005. DNA sequence polymorphism and divergence at the *erect wing* and *suppressor of sable* loci of *Drosophila melanogaster* and *D. simulans*. Genetics 170(3):1153–1165.

Campo D, et al. 2013. Whole-genome sequencing of two North American *Drosophila melanogaster* populations reveals genetic differentiation and positive selection. Mol Ecol. 22(20):5084–5097.

Chang CC, et al. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7.

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

Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485.

Crisci JL, Dean MD, Ralph P. 2016. Adaptation in isolated populations: when does it happen and when can we tell? Mol Ecol. 25(16):3901–3911.

Danecek P, 1000 Genomes Project Analysis Group, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

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

David JR, Van Herrewege J. 1983. Adaptation to alcoholic fermentation in *Drosophila* species: relationship between alcohol tolerance and larval habitat. Comp Biochem Physiol. 74(2):283–288.
Ewing G, Hermisson J. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 26:2064–2065.
Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 26(9):2097–2108.
Fabian DK, et al. 2012. Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. Mol Ecol. 21(19):4748–4769.
Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415(6875):1024–1026.
Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 31(5):1275–1291.
Fournier D, Bride JM, Hoffmann F, Karch F. 1992. Acetylcholinesterase. Two types of modifications confer resistance to insecticide. J Biol Chem. 267(20):14270–14274.
Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genet. 11(2):e1005004.
Garud NR, Petrov DA. 2016. Elevated linkage disequilibrium and signatures of soft sweeps are common in *Drosophila melanogaster*. Genetics 203(2):863–880.
Gillespie JH, Turelli M. 1989. Genotype-environment interactions and the maintenance of polygenic variation. Genetics 121(1):129–138.
Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics 165(3):1269–1278.
Grenier JK, et al. 2015. Global diversity lines–a five-continent reference panel of sequenced *Drosophila melanogaster* strains. G3 (Bethesda) 5(4):593–603.
Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and negative selection on noncoding DNA in *Drosophila simulans*. Mol Biol Evol. 25(9):1825–1834.
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.
Hamblin MT, Veuille M. 1999. Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. Genetics 153(1):305–317.
Hasson E, Wang IN, Zeng LW, Kreitman M, Eanes WF. 1998. Nucleotide variation in the *triosephosphate isomerase* (*Tpi*) locus of *Drosophila melanogaster* and *Drosophila simulans*. Mol Biol Evol. 15(6):756–769.
Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169(4):2335–2352.
Hermisson J, Pennings PS. 2017. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol Evol. 8:700–716.
Hoffmann AA, Weeks AR. 2007. Climatic selection on genes and traits after a 100 year-old invasion: a critical look at the temperate-tropical clines in *Drosophila melanogaster* from eastern Australia. Genetica 129(2):133–147.
Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res. 23(1):89–98.
Hübner S, et al. 2013. Genome differentiation of *Drosophila melanogaster* from a microclimate contrast in Evolution canyon, Israel. Proc Natl Acad Sci U S A. 110(52):21059–21064.
Hwang S, Kim E, Lee I, Marcotte EM. 2015. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 5:17875.
Jensen JD. 2014. On the unfounded enthusiasm for soft selective sweeps. Nat Commun. 5:5281.
Karasov T, Messer PW, Petrov DA, Malik HS. 2010. Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. PLoS Genet. 6(6):e1000924.
Kelly JK, Koseva B, Mojica JP. 2013. The genomic signal of partial sweeps in *Mimulus guttatus*. Genome Biol Evol. 5(8):1457–1469.
King EG, Macdonald SJ, Long AD. 2012. Properties and power of the *Drosophila* synthetic population resource for the routine dissection of complex traits. Genetics 191(3):935–949.
Kofler R, Nolte V, Schlötterer C. 2015. Tempo and mode of transposable element activity in *Drosophila*. PLoS Genet. 11(7):e1005406.
Kolaczkowski B, Kern AD, Holloway AK, Begun DJ. 2011. Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. Genetics 187(1):245–260.
Kousathanas A, Keightley PD. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. Genetics 193(4):1197–1208.
Lachaise D, et al. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol Biol. 22:159–225.
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.
Langley CH, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192(2):533–598.
Li H. 2015. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1–3.
Li H, 1000 Genome Project Data Processing Subgroup, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078–2079.
Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2(10):e166.
Li YJ, Satta Y, Takahata N. 1999. Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes Genet Syst. 74(4):117–127.
Ma J, Amos CI. 2012. Principal components analysis of population admixture. PLoS One 7(7):e40115.
Machado HE, et al. 2015. Comparative population genomics of latitudinal variation in *D. simulans* and *D. melanogaster*. Mol Ecol. 25:723–740.
Mackay TFC, et al. 2012. The *Drosophila melanogaster* genetic reference panel. Nature 482(7384):173–178.
Macpherson JM, et al. 2008. Nonadaptive explanations for signatures of partial selective sweeps in *Drosophila*. Mol Biol Evol. 25(6):1025–1042.
McKenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.
McKenzie JA, McKechnie SW. 1979. A comparative study of resource utilization in natural populations of *Drosophila melanogaster* and *D. simulans*. Oecologia 40(3):299–309.
Menozzi P, Shi MA, Lougarre A, Tang ZH, Fournier D. 2004. Mutations of acetylcholinesterase which confer insecticide resistance in *Drosophila melanogaster* populations. BMC Evol Biol. 4:4.
Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 28(11):659–669.

Milan NF, Kacsoh BZ, Schlenke TA. 2012. Alcohol consumption as self-medication against blood-borne parasites in the fruit fly. Curr Biol. 22(6):488–493.
Mutero A, Pralavorio M, Bride JM, Fournier D. 1994. Resistance-associated point mutations in insecticide-insensitive acetylcholinesterase. Proc Natl Acad Sci U S A. 91(13):5922–5926.
Nolte V, Schlötterer C. 2008. African *Drosophila melanogaster* and *D. simulans* populations have similar levels of sequence variability, suggesting comparable effective population sizes. Genetics 178(1):405–412.
Nuzhdin SV, Turner TL. 2013. Promises and limitations of hitchhiking mapping. Curr Opin Genet Dev. 23(6):694–699.
Ometto L, Glinka S, De Lorenzo D, Stephan W. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol. 22(10):2119–2130.
Palmieri N, Nolte V, Chen J, Schlötterer C. 2015. Genome assembly and annotation of a *Drosophila simulans* strain from Madagascar. Mol Ecol Res. 15(2):372–381.
Parsch J. 2003. Selective constraints on intron evolution in *Drosophila*. Genetics 165(4):1843–1851.
Parsch J, Meiklejohn CD, Hartl DL. 2001. Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. Genetics 159(2):647–657.
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.
Parsons PA. 1977. Larval reaction to alcohol as an indicator of resource utilization differences between *Drosophila melanogaster* and *D. simulans*. Oecologia 30(2):141–146.
Pennings PS, Hermisson J. 2006a. Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol. 23(5):1076–1084.
Pennings PS, Hermisson J. 2006b. Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet. 2:e186.
Pool JE. 2015. The mosaic ancestry of the *Drosophila* genetic reference panel and the *D. melanogaster* reference genome reveals a network of epistatic fitness interactions. Mol Biol Evol. 32:3236–3251.
Presgraves DC. 2005. Recombination enhances protein adaptation in Drosophila melanogaster. Curr Biol 15:1651–1656.
Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. 20(4):R208–R215.
Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. Evolution 59(11):2312–2323.
Przeworski M, Wall JD, Andolfatto P. 2001. Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. Mol Biol Evol. 18(3):291–298.
Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J HumGenet. 81(3):559–575.
Quesada H, Ramírez UEM, Rozas J, Aguade M. 2003. Large-scale adaptive hitchhiking upon high recombination in *Drosophila simulans*. Genetics 165(2):895–900.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842.
Ramirez-Soriano A, Nielsen R. 2009. Correcting estimators of θ and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. Genetics 181(2):701–710.
Ross-Ibarra J, et al. 2008. patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. PLoS One 3(6):e2411.
Rozas J, Gullaud M, Blandin G, Aguade M. 2001. DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics 158(3):1147–1155.
Sanchez-Gracia A, Rozas J. 2007. Unusual pattern of nucleotide sequence variation at the OS-E and OS-F genomic regions of *Drosophila simulans*. Genetics 175(4):1923–1935.
Schlenke TA, Begun DJ. 2004. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proc Natl Acad Sci U S A. 101(6):1626–1631.
Schmidt JM, et al. 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. PLoS Genet. 6(6):e1000998.
Schmidt PS, Paaby AB. 2008. Reproductive diapause and life-history clines in North American populations of *Drosophila melanogaster*. Evolution 62(5):1204–1215.
Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189(4):1427–1437.
SchÖfl G, Schlötterer C. 2006. Microsatellite variation and differentiation in African and non-African populations of Drosophila simulans. Mol. Ecol. 15:3895–3905.
Schrider DR, Mendes FK, Hahn MW, Kern AD. 2015. Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics 200(1):267–284.
Sedghifar A, Saelao P, Begun DJ. 2016. Genomic patterns of geographic differentiation in *Drosophila simulans*. Genetics 202(3):1229–1240.
Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 5(6):e1000495.
Shapiro JA, et al. 2007. Adaptive genic evolution in the *Drosophila* genomes. Proc Natl Acad Sci U S A. 104(7):2271–2276.
Singh ND, Macpherson JM, Jensen JD, Petrov DA. 2007. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster*. BMC Evol Biol. 7:202–216.
Singh RS. 1989. Population genetics and evolution of species related to *Drosophila melanogaster*. Annu Rev Genet. 23:425–453.
Smith N, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature 415(6875):1022–1024.
Steige KA, Laenen B, Reimegård J, Scofield D, Slotte T. 2015. Genomic analysis reveals major determinants of cis-regulatory variation in *Capsella grandiflora*. Proc Natl Acad Sci U S A. 114:1087–1092.
Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98(2):65–68.
Sturtevant AH. 1920. Genetic studies on *Drosophila simulans*. I. Introduction. Hybrids with *Drosophila melanogaster*. Genetics 5(5):488–500.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595.
Tennessen JA, Madeoy J, Akey JM. 2010. Signatures of positive selection apparent in a small sample of human exomes. Genome Res. 20(10):1327–1334.
Tennessen JA, NHLBI Exome Sequencing Project, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090):64–69.
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.
True JR, Mercer JM, Laurie CC. 1996. Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics 142:507–523.
Turelli M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theor Popul Biol. 25(2):138–193.
Turelli M, Hoffmann AA. 1995. Cytoplasmic incompatibility in *Drosophila simulans*: dynamics and parameter estimates from natural populations. Genetics 140(4):1319–1338.

Wall JD, Andolfatto P, Przeworski M. 2002. Testing models of selection and demography in *Drosophila simulans*. Genetics 162(1):203–216.
Wang J, Hill WG, Charlesworth D, Charlesworth B. 1999. Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. Genet Res. 74(2):165–178.
Wang W, Thornton K, Berry A, Long M. 2002. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science 295(5552):134–137.
Wang W, Thornton K, Emerson JJ, Long M. 2004. Nucleotide variation and recombination along the fourth chromosome in *Drosophila simulans*. Genetics 166(4):1783–1794.
Weeks AR, McKechnie SW, Hoffmann AA. 2002. Dissecting adaptive clinal variation: markers, inversions and size/stress associations in *Drosophila melanogaster* from a central field population. Ecol Lett. 5(6):756–763.
Zheng X, et al. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28(24):3326–3328.

Associate editor: Charles Baer