

Inferences of Demography and Selection in an African Population of *Drosophila melanogaster*

Nadia D. Singh,^{*,1} Jeffrey D. Jensen,^{*,‡} Andrew G. Clark,[§] and Charles F. Aquadro[§]

^{*}Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695, [†]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [‡]Swiss Institute of Bioinformatics, Lausanne, Switzerland, and

[§]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853

ABSTRACT It remains a central problem in population genetics to infer the past action of natural selection, and these inferences pose a challenge because demographic events will also substantially affect patterns of polymorphism and divergence. Thus it is imperative to explicitly model the underlying demographic history of the population whenever making inferences about natural selection. In light of the considerable interest in adaptation in African populations of *Drosophila melanogaster*, which are considered ancestral to the species, we generated a large polymorphism data set representing 2.1 Mb from each of 20 individuals from a Ugandan population of *D. melanogaster*. In contrast to previous inferences of a simple population expansion in eastern Africa, our demographic modeling of this ancestral population reveals a strong signature of a population bottleneck followed by population expansion, which has significant implications for future demographic modeling of derived populations of this species. Taking this more complex underlying demographic history into account, we also estimate a mean X-linked region-wide rate of adaptation of 6×10^{-11} /site/generation and a mean selection coefficient of beneficial mutations of 0.0009. These inferences regarding the rate and strength of selection are largely consistent with most other estimates from *D. melanogaster* and indicate a relatively high rate of adaptation driven by weakly beneficial mutations.

THERE is considerable interest in understanding the nature and extent of adaptation in natural populations. *Drosophila melanogaster* has been the focus of many such studies and a variety of approaches to address this fundamental question have been developed with this system. These include divergence-based methods (e.g., Sattath *et al.* 2011), polymorphism-based approaches (e.g., Kim and Stephan 2002; Sabeti *et al.* 2002; Kim and Nielsen 2004; Li and Stephan 2006b; Jensen *et al.* 2008a), and approaches that use both polymorphism and divergence data (e.g., Fay *et al.* 2002; Welch 2006; Andolfatto 2007; Macpherson *et al.* 2007; Shapiro *et al.* 2007). In some cases, these methods aim to localize putative targets of selection (e.g., Harr *et al.* 2002; Glinka *et al.* 2003; Jensen *et al.* 2007a), while in others the ultimate goal is to generally characterize the average rate and strength of adaptation (Wiehe and Stephan 1993; Li and Stephan 2006a; Andolfatto 2007;

Macpherson *et al.* 2007; Jensen *et al.* 2008a). These approaches have enjoyed much success, and it is becoming clear that the signatures of both recent and recurrent selection abound in the *D. melanogaster* genome (for review see Sella *et al.* 2009).

However, demographic history will also leave its signature in the genome, which has the potential to confound inferences of natural selection. Indeed, to effectively investigate models of natural selection, which are of great general interest in evolutionary biology, one must take into account the underlying demographic history of a population. Consequently, understanding demographic history is integral for making incisive inferences about the genetic adaptation of populations (e.g., Akey *et al.* 2004; Nielsen *et al.* 2009; Lohmueller *et al.* 2011).

Many sophisticated methodologies for estimating demographic parameters have been developed, and application of such methods has revealed much about the demographic history of natural populations of *D. melanogaster* (Haddrill *et al.* 2005; Ometto *et al.* 2005; Thornton and Andolfatto 2006; Stephan and Li 2007; Laurent *et al.* 2011). Although it is clear that presumably ancestral East African populations (Pool and Aquadro 2006) show strong signals of nonequilibrium

Copyright © 2013 by the Genetics Society of America
doi: 10.1534/genetics.112.145318

Manuscript received August 24, 2012; accepted for publication October 18, 2012
Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.145318/-/DC1>.

¹Corresponding author: Department of Genetics, North Carolina State University, Campus Box 7614, Raleigh, NC 27695. E-mail: ndsingh@ncsu.edu; 919-515-1761

demography (for review see Stephan and Li 2007), it has mostly been assumed that many of the observed deviations from neutrality have been driven by population expansion (although see Haddrill *et al.* 2005). As a consequence, the parameter space for characterizing the demographic history of African populations has not been thoroughly explored.

Here we present a high-coverage resequencing data set collected from 20 individuals of *D. melanogaster* from Uganda for a continuous 2.1-Mb region of the X chromosome. This locus encompasses the region between the genes *garnet* and *scalloped*; we focus on this region because recombination rates in this interval are estimated to be high (Fiston-Lavier *et al.* 2010). This high rate of recombination enhances our ability to detect multiple independent selective events, and multiple genealogies represented should also improve the quality of demographic inferences. In addition to examining patterns of polymorphism and divergence across this region and across functional categories of sites, we also use these data to model the demographic history of this population. Our results provide strong evidence in support of a severe population contraction followed by expansion in this presumed ancestral population. These findings have significant consequences for demographic inference in derived populations of *D. melanogaster*, which typically assume that ancestral populations of this species can be best modeled with a simple growth model. We then use the estimated demographic parameters to inform models of single and recurrent hitchhiking. Our inferences of selection point to three potential targets of recent selective sweeps in this 2.1-Mb region, and we infer a mean selection coefficient of beneficial mutations and a mean rate of adaptation that are wholly consistent with previous results. These results thus further refine our understanding of the adaptive history of ancestral populations of *D. melanogaster* and highlight the importance of modeling this ancestral nonequilibrium history when subsequently inferring both adaptive and demographic events in derived populations.

Materials and Methods

Strains used

This study used twenty X chromosome extraction lines of *D. melanogaster* derived from a single population sample collected from Namulonge, Uganda, in 2005 by J. Ogwang for John E. Pool (Pool and Aquadro 2006). The chromosome-extraction procedure ensures that each strain is isogenic for a single, wild-derived X chromosome. These strains are: UgX2b, UgX3a, UgX6b, UgX8b, UgX10b, UgX11a, UgX12b, UgX13b, UgX17a, UgX18, UgX19a, UgX20, UgX29b, UgX30a, UgX37, UgX40, UgX45b, UgX52b, UgX54a, and UgX63.

DNA preparation

Genomic DNA was extracted from each line using the Qiagen DNeasy blood and tissue kit. Twenty-five individuals from each line were used for each genomic DNA extract. Genomic

DNA was sheared using sonication, and genomic DNA libraries suitable for Illumina single-end sequencing were prepared following a standard protocol (available upon request). Each library was barcoded with 1 of 10 3-bp barcodes (ACT, ATA, AAG, GGA, TTG, GCG, TAA, TGT, GAT, and AGC). DNA from each genomic library was quantified using PicoGreen, and DNAs from 10 different libraries (each with a different barcode) were combined in equal amounts. This procedure thus yielded two pools of DNA, each of which was composed of DNA from 10 different strains of *D. melanogaster*.

These pooled DNA samples were enriched for the X chromosome region between the genes *garnet* and *scalloped* (coordinates in Flybase release 5.2 are 13621236–15719755) using a custom NimbleGen Comparative Genomic Hybridization Array (OID26736). This array contains 385,000 oligos, each of which map to the target region of interest. The array does not include repetitive sequence or low-complexity sequence, and overall, 94.3% of bases in the reference sequence were covered by at least one oligo. The two pooled DNA samples were individually hybridized to an array at the Cornell Microarray Core following standard NimbleGen hybridization protocol. The resulting DNA, enriched for the 2.1-Mb X chromosome region between *garnet* and *scalloped*, was sequenced using Illumina single-end sequencing (86-bp reads) at the Cornell Life Sciences Core Laboratories Center. Each sample was sequenced twice (each on a different run).

Bioinformatics

Individual reads were mapped to the *D. melanogaster* genome using the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) with default parameters. Only reads mapping uniquely to the genome were retained for analysis. SAMtools (Li *et al.* 2009) was used to generate alignment files in pileup format. These pileup files were used to call bases and identify single nucleotide polymorphisms (SNPs) using the joint genotype for inbred lines (JGIL) (Stone 2012).

Insertions, deletions, and larger-copy-number variants were ignored in this analysis largely because it is difficult to explicitly model ascertainment bias of this type of event given our targeted enrichment strategy. In particular, the use of short oligos on the array limits our ability to capture genomic regions with $>\sim 3$ mismatches relative to the reference sequence. Given levels of polymorphism in *D. melanogaster* ($\sim 1\%$) and oligo length (< 100 bp), we do not expect this to compromise our ability to capture fragments with SNPs, but we do expect that capturing insertion, deletion, and copy number variation will be challenging. As a consequence, we do not include these types of events in our analysis. Although these events would add to the resolution of the models, counting all data of this class as missing should not bias the inference derived solely from alignable single-base changes.

Validation of base calls

To validate base calls, nine noncoding loci roughly evenly distributed across the 2.1-Mb region of interest were PCR amplified and Sanger sequenced in all 20 strains. The sizes

of these loci ranged from 622 to 788 bp. Primer sequences for and approximate physical positions of these nine loci can be found in [Supporting Information, Table S1](#).

Genomic DNA was extracted from each of the 20 lines using the Qiagen DNeasy Blood and Tissue kit. Twenty-five individuals from each line were used for each genomic DNA extract. Genomic DNA from each line was amplified using PCR. We amplified these loci using a touchdown PCR program. Amplifying conditions were as follows: 94°/2 min, 12 cycles of 94°/30 sec (target annealing temperature + 6°)/30 sec, 72°/60 sec with the annealing temperature reduced by 0.5 degrees per cycle, followed by 23 cycles of 94°/30 sec (target annealing temperature)/30 sec, 72°/60 sec. We included a final extension of 72°/7min. The target annealing temperature was 56° for VL1, VL2, VL3, VL6, VL7, and VL9, 52° for VL4 and VL10, and 53° for VL5. All PCR reactions were 8 μ l, and each contained 4 μ l Qiagen 2X PCR Master-Mix, 0.2 μ l of each 10 μ M primer, 2.6 μ l H₂O, and 1 μ l genomic DNA (at 10 ng/ μ l). PCR reactions were enzymatically cleaned with exonuclease I and shrimp alkaline phosphatase and were cycle sequenced in half-strength half-reactions with Big Dye under standard cycling conditions. These sequencing reactions were cleaned and sequenced at the Genome Sciences Laboratory at North Carolina State University. Sequence data were analyzed in Sequencher (v. 4.10.1). Base calls in each individual at each locus were based on at least two sequence reads (one in each direction) from a single PCR amplification.

Summary statistics

Annotations for this 2.1-Mb region are based on annotations of release 5.4 of the *D. melanogaster* genome as well as information from the RedFly database and a curated database of footprinting literature (kindly provided by R. Kulathinal, personal communication). Each base is annotated with respect to several features, including whether it is (a) contained in a transcript, (b) contained in a coding sequence, (c) first codon position, (d) second codon position, (e) third codon position, (f) 5' UTR, (g) 3' UTR, or (h) intron. We further refined the “intron” classification to separate short (≤ 65 bp) and long (> 65 bp) introns based on the release 5.4 GFF file for *D. melanogaster* and trimmed the first and last 10 bp of each intron. We used previously reported definitions of “preferred” codons (Marais *et al.* 2001) to further classify changes at third codon positions (preferred \rightarrow preferred or unpreferred \rightarrow unpreferred).

Pairwise alignments between *D. melanogaster* and *D. sechellia* for the *garret-scalloped* region were generated by parsing the net.axt files from the UCSC genome browser. This net.axt file was based on release 3 of the *D. melanogaster* genome and release 1 of the *D. sechellia* genome. When polarizing the site frequency spectrum (SFS) to *D. sechellia* alone, we used these pairwise alignments. When polarizing the SFS to both *D. sechellia* and *D. yakuba*, we relied on multispecies alignments for this region that were kindly provided by R. Kulathinal (personal communication).

For several annotation classes, we estimated Watterson's θ , π , D_{xy} (pairwise divergence between species), and Tajima's D for the entire 2.1-Mb region using custom Perl scripts (available upon request). We also estimated these quantities (combining all sequence types) in sliding windows across the 2.1-Mb region with a window size of 10 kb and a step size of 5 kb.

Demographic modeling

Demographic modeling was performed with dadi (Gutenkunst *et al.* 2009). To correct for potential mis-inference of the ancestral state, we corrected the SFS for multiple mutations along the lineage to the ingroup following Hernandez *et al.* (2007). Note that this approach also corrects for violations of the infinite-sites model that occur along this lineage as well. Estimates of the Q matrix appropriate for *Drosophila* were taken from Singh *et al.* (2009), and trinucleotide frequencies were derived from the sequence data obtained in the current study.

We considered five demographic models implemented in dadi: neutral (standard neutral model), two epoch (instantaneous size change some time ago), growth (exponential growth beginning some time ago), bottlegrowth (instantaneous size change followed by exponential growth), and three epoch (bottleneck of some duration followed by recovery).

Fitting single hitchhiking models

To identify putatively swept regions of the genome, we utilized the likelihood framework initially proposed by Kim and Stephan (2002), with a number of modifications as implemented by Pavlidis *et al.* (2010). The initial composite likelihood ratio test (CLRT) used the spatial distribution of SNP frequencies and levels of variability and compares the composite likelihood of the data under the standard neutral model to the likelihood under a single selective sweep model. Although this method was an important and widely used advance, Jensen *et al.* (2005) subsequently demonstrated that neutral nonequilibrium models may indeed be frequently identified as putatively swept regions under this model, owing to the ability of particular population bottlenecks to resemble sweep-like patterns of variation in the SFS.

To better deal with these nonequilibrium perturbations, two subsequent strategies have been employed. First, apart from SFS-based approaches, Stephan *et al.* (2006) proposed that recently swept regions produce unusual patterns of linkage disequilibrium (LD), which may indeed be useful in distinguishing sweep models from neutral nonequilibrium models. Jensen *et al.* (2007b) demonstrated that such patterns (as captured in the statistic ω ; Kim and Nielsen 2004) attained sufficient power at reasonable false-positive rates in identifying swept loci even in severely bottlenecked populations.

A second approach to this problem was developed by Nielsen *et al.* (2005) and still relies on the likelihood framework of Kim and Stephan and patterns in the SFS alone—but

rather than comparing these against a neutral equilibrium model, their null effectively becomes the so-called “background SFS.” This test is thus essentially an outlier approach—a model is fitted to the overall SFS of the data in a model-nonspecific manner, and then unusual regions are identified relative to this background pattern.

Thus, these background SFS- and LD-based approaches have both been proposed to improve performance for the identification of recently swept loci in nonequilibrium populations. Pavlidis *et al.* (2010) recently evaluated the performance of both methodologies independently and compared performance with a hybrid approach—that is, using both sweep-like patterns in the SFS and LD, with the background SFS as a null. Results demonstrate that the hybrid approach is superior to either approach independently, and this thus represents the best-performing approach currently available for the identification of putatively swept regions. It is this method that we adopt here, which is available for download at: http://www.bio.lmu.de/~pavlidis/home/?Software:Genome_scans_for_selection

To reduce the false-positive rate, we used the demographic parameter estimates to construct the null distribution for the SFS (Thornton and Jensen 2007). As third codon positions are likely to be less constrained than long intron sites (see below), we used demographic parameter estimates from the third codon position data to generate the neutral null distributions. Finally, following Jensen *et al.* (2008a) we take a parametric bootstrap approach to quantify uncertainty in parameter estimation.

Fitting recurrent hitchhiking models

In addition to detecting adaptive evolution at specific loci as discussed above, a separate literature has grown around the question of characterizing the genomic rate and strength of adaptation (for review see Sella *et al.* 2009)—thus, estimating a recurrent hitchhiking model as opposed to a single hitchhiking model. We here take the approximate Bayesian (ABC) approach of Jensen *et al.* (2008a), which has been demonstrated to accurately estimate the mean rate ($2N\lambda$) and strength (s) of positive selection for data sets of this size, and it allows for the modeling of these parameters to be given by distributions rather than fixed values. Because the level of reduction in variation owing to recurrent hitchhiking depends on the joint parameter ($2Ns\lambda$) (Wiehe and Stephan 1993), both the mean and standard deviations of a number of common polymorphism summary statistics are used to uncouple these parameters (the mean average pairwise difference (π), Watterson’s theta (θ_w) (Watterson 1975), θ_H (Fu 1996), and ZnS (Kelly 1997)). Calculating these summary statistics from the observed data and from simulated data, we implement the regression approach of Beaumont *et al.* (2002) in fitting a local linear regression of simulated parameter values to simulated summary statistics, substituting the observed statistics into a regression equation (see Thornton 2009). The prior distribution on the

strength of selection is a uniform (1.0×10^{-6} , 1.0), as is the prior distribution on the rate of selection (1.0×10^{-7} , 1.0×10^{-1}). Tolerance is set at 0.001. Estimation is based on 10^6 draws from the prior using the recurrent hitchhiking (RHH) machinery of Jensen *et al.* (2008a). Note that the RHH framework is an equilibrium model and that the underlying demographic model is used only to generate the priors. For inferences of selection parameters, we assume exponential distributions of the rate and strength, such that each draw from the prior represents the mean of the distribution. For reference, the empirical values of the key parameters used in this model fitting are presented in Table S2. The estimation and simulation programs are available for download at <http://www.molpopgen.org/software/Jensen-ThorntonAndolfatto2008/>. Nonequilibrium selection simulations were performed in SFS_CODE (Hernandez 2008).

Results

Next-generation sequencing

Sequence reads from each of the two Illumina sequencing runs for each pooled sample were combined prior to analysis. Barcoded adapters were trimmed from each read, as were premature 3′ adapters. The number of reads per (barcoded) strain ranged from 2,272,898 to 10,190,383, with an average of 2,486,628 reads per strain. These reads were mapped to the *D. melanogaster* genome using BWA (Li and Durbin 2009). The percentage of reads mapping uniquely to the *D. melanogaster* genome ranged from 67.4 to 75.2, with an average of 71.5%. Between 63.6 and 71.1% of reads overall mapped uniquely to our target region, with an average of 66.7% across strains. Perhaps more importantly, of reads uniquely mapping to the genome, between 89.1 and 96.4% of reads map to our target region, with an overall average of 93.1% across strains.

Duplicate reads (those with identical start and end coordinates) were removed. Given that the expected coverage based on read counts (ranging from 59× to 268×) in many cases exceeded the read length (86 bp), the removal of duplicate reads is likely to have removed both duplicates arising from PCR as well as duplicate reads arising stochastically through the sequencing itself. This, while not systematically biasing our results in any way, reduces coverage dramatically. Postduplicate read removal, coverage in our target region ranged from 16.9× to 67.4×, with an average coverage across strains of 32.4×. Within each strain, 74–92% of sites were covered by at least 5 reads, 61–90% of sites were covered by at least 10 reads, and 38–87% of sites were covered by at least 20 reads. Between 6 and 16% of sites had zero coverage within strains, which is likely due in part to gaps in the hybridization array (see *Materials and Methods*).

Base calling and validation

Bases were called using JGIL (Stone 2012). This tool is specifically designed for lines that are mostly homozygous due to inbreeding, but that may have regions of residual

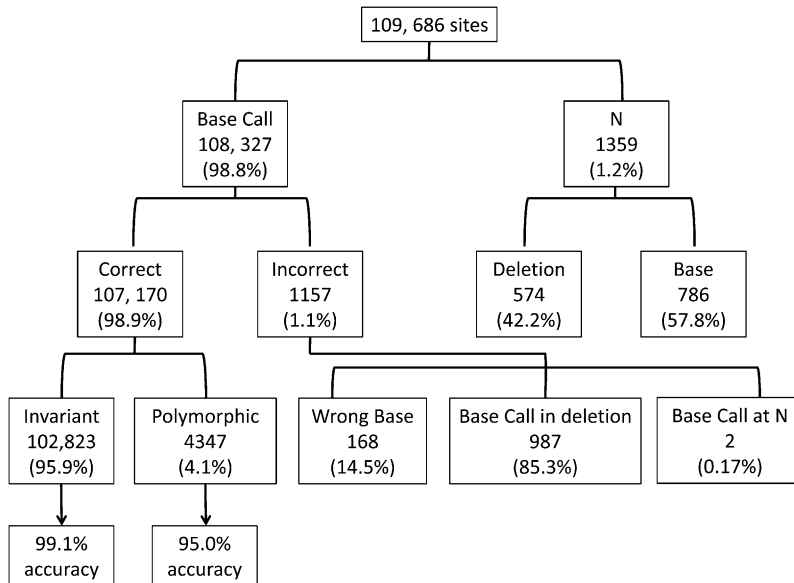


Figure 1 Sanger validation of JGIL base calls. In total 109,686 sites were sequenced, which corresponds to 20 individuals at nine ~700-bp loci across the 2.1-Mb region between *garret* and *scalloped*. Thus, ~5.5 kb of Sanger sequence data were generated for each of the 20 lines. JGIL made base calls 98.8% of the time, 98.9% of which were correct. This corresponds to an accuracy of 99.1% at invariant sites and 95.0% at polymorphic sites. Incorrect base calls were overwhelmingly revealed to be in deletions through Sanger sequencing. Approximately 42% of N's called by JGIL correspond to deleted bases in the Sanger data.

heterozygosity. Over 91% of possible bases were called, with the remaining ~9% left as N's in the sequence. Across the region, 78.7% of sites had base calls in all 20 lines, and 5.7% of sites were left as N's in all 20 lines. The remaining 15.6% of sites were called in between 1 and 19 lines; the full distribution of the number of sites in which bases were called as N's in 0–20 strains is presented in Figure S1.

Overall, 73,002 sites were identified as SNPs, 71,365 of which are biallelic. Of these biallelic polymorphic sites, 46,883 (65.7%) had bases called in all 20 lines. The (folded) SFS of these SNPs is presented in Figure S2.

To assess the accuracy of the JGIL base and SNP calls, we Sanger sequenced nine short loci distributed across the *garret-scattered* region in all 20 Uganda lines. These validation data are summarized in Figure 1. Summed across lines and loci, we Sanger sequenced a total of 109,686 sites. Of these sites, 98.8% had a JGIL base call, and 1.2% sites were called N in JGIL. Of those sites with base calls, 98.9% of bases were called correctly. Both invariant and polymorphic sites are included in this category, and if we measure accuracy in these types of sites separately, our Sanger data suggest a 99.1% accuracy of JGIL base calls at invariant sites and 95.0% accuracy at polymorphic sites. The bulk of this slightly reduced accuracy is due to JGIL failing to call SNPs rather than falsely calling SNPs at invariant sites; of polymorphisms identified by Sanger sequencing that were missed by JGIL, 50% are singletons. However, the site frequency spectra of polymorphisms identified by Sanger sequencing and JGIL are statistically indistinguishable and show comparable deviations from neutrality including an excess of low- and high-frequency-derived alleles (Figure S3).

Approximately 1.1% of JGIL base calls did not match the Sanger data. The vast majority of these cases (85.3%) correspond to sites that are called as bases in JGIL while

the Sanger data indicate that these bases are in deletions. At 14.5% of incorrect base calls, JGIL called a base that did not match the base called from Sanger sequencing. The remaining sites correspond to sites at which the Sanger base calls were unclear but JGIL called a base. We refer to this as an “incorrect” base call by JGIL, although in these cases the JGIL data may in fact be correct.

Approximately 1.2% of sites were called N's by JGIL. Comparison with the Sanger sequencing data suggests that slightly more than half of these bases can be called with Sanger data, and slightly less than half of these sites are in deletions.

Polymorphism and divergence

When estimating summary statistics regarding polymorphism and divergence, we considered only sites for which bases were called in all 20 lines as well as in our outgroup species, *D. sechellia*. We first examined polymorphism across the entire 2.1-Mb region, partitioned by sequence annotation. The sequence classes considered (and numbers of sites included) for this analysis were intergenic (527,975), UTR (97,842), long intron (583,879), short intron (9234), first codon position (113,207), second codon position (113,229), and third codon position (112,835). Estimates of nucleotide diversity (Watterson's estimator and π) and divergence from *D. sechellia* (D_{xy}) are presented in Figure 2.

We also conducted a sliding window analysis in which we examined levels of per-site polymorphism and divergence as well as Tajima's *D* in 10-kb windows (with a slide of 5 kb). These metrics are summarized without regard to annotation, as we consider all sites within the window, independent of the functional annotation of those sites. These data are presented in Figure 3. Importantly, the window break-points are set by the chromosomal coordinates, independent of how many sites within the 10-kb window have bases called in all 20 lines of *D. melanogaster* and *D. sechellia*. As

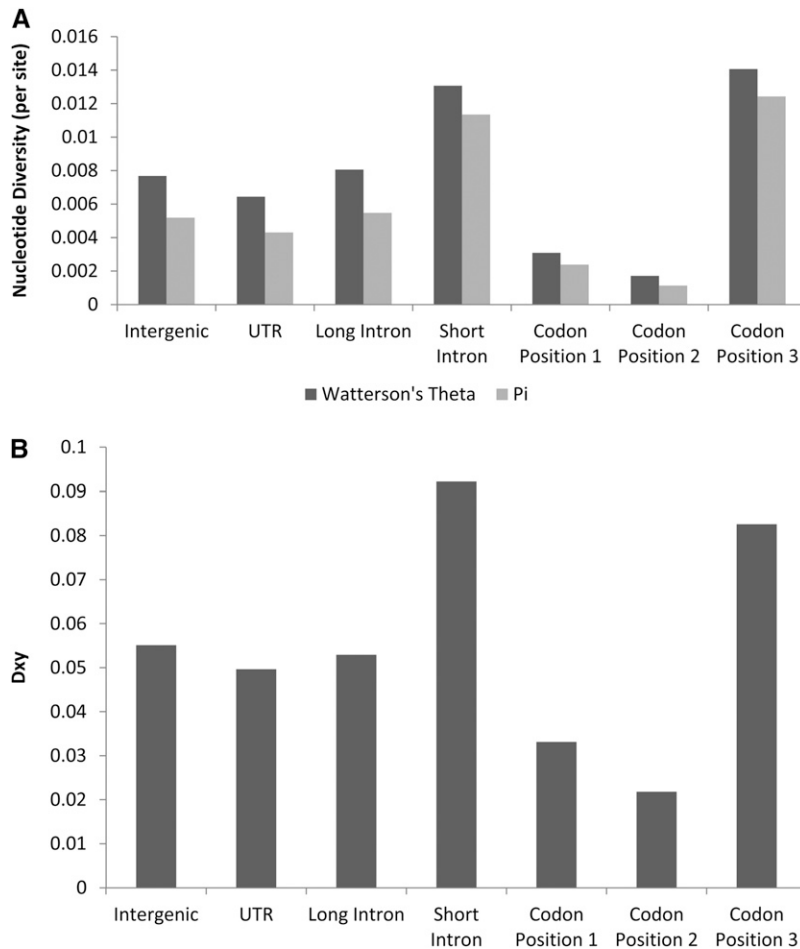


Figure 2 Overall estimates of (A) nucleotide variation, measured as Watterson's θ (dark shading) and π (light shading) across functional annotation categories and (B) divergence (estimated as D_{xy}) across functional annotation categories.

a consequence, after excluding sites with missing data, the number of sites considered in each window (1) <10,000 and (2) varies across windows. The four windows containing fewer than 2000 sites were excluded from the analysis.

Demographic modeling

We fitted several demographic models to our polymorphism data using dadi (Gutenkunst *et al.* 2009). We used the unfolded SFS with *D. sechellia* as our outgroup and considered the site frequency spectra of different functional classes of site. Taking into account both potential constraint experienced by each annotation class as well as the number of sites within that annotation class, we chose to use long intron sites and third codon position sites to fit demographic models. We corrected the long intron site frequency and third codon position site frequency spectra for potential mis-inference of ancestral state using previously described methodology (Hernandez *et al.* 2007). This correction is parameterized by expected divergence, and we thus employed a site-class-specific correction, making use of observed divergence for both site classes from previous work (Singh *et al.* 2009). These divergence values are 0.057 for long introns and 0.093 for third codon positions. The corrected vs. uncorrected site frequency spectra for these two classes of site are presented in Figure S4.

To assess the efficacy of this mathematical correction to the SFS, we instead polarized the SNPs using outgroup data from both *D. yakuba* and *D. sechellia* (Figure S4). Comparison of these spectra with those using the mathematical correction for potential mis-inference of ancestral state reveals a lower proportion of high-frequency-derived alleles in the mathematically corrected SFS. We thus believe that this mathematical correction is the most conservative approach for polarizing the SFS, and use “corrected” site frequency spectra throughout the rest of our analyses.

We evaluated the fit of five different demographic models to the polymorphism data from each of our two types of sites. Using dadi nomenclature, we explored a neutral equilibrium model, a two-epoch model (corresponding to an instantaneous size change some time ago), a growth model (exponential growth beginning some time ago), a bottlegrowth model (instantaneous population reduction some time ago followed by exponential growth), and a three-epoch model (a bottleneck some time ago of some duration followed by recovery). The log-likelihoods of these various models are presented in Table 1 and the site frequency spectra under each of the demographic models compared to the observed site frequency spectra are presented in Figure S5.

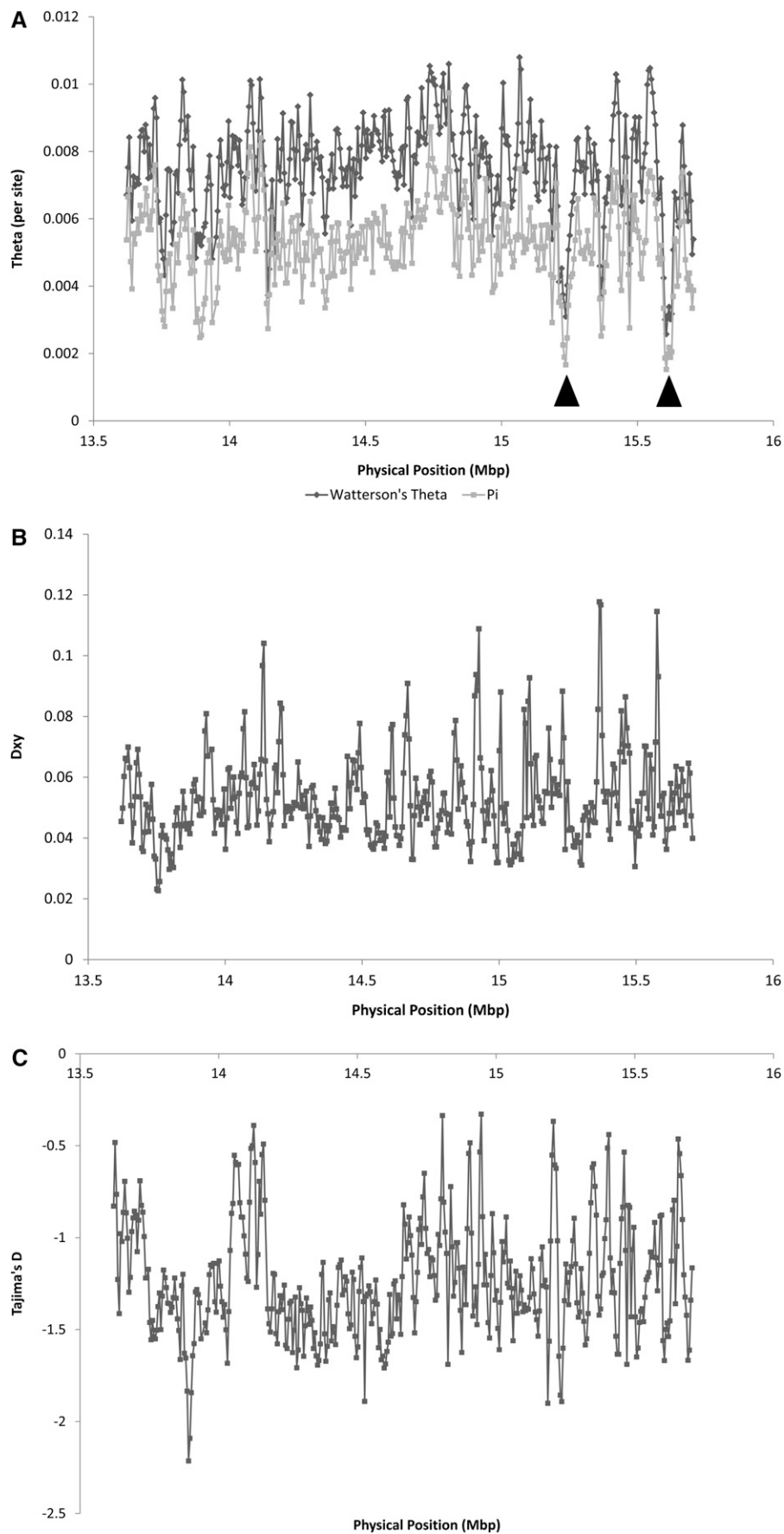


Figure 3 Sliding window plot of (A) nucleotide variation, measured as Watterson's θ (dark shading) and π (light shading), (B) divergence (estimated as D_{xy}), and (C) Tajima's D across the *garnet-scalloped* region. Solid triangles in A denote the two regions of the chromosome with the most depressed levels of nucleotide diversity.

Table 1 Log-likelihoods of several alternative demographic models

	Long introns	Third codon positions
Neutral equilibrium	−1985.45	−206.63
Two epoch	−245.52	−116.00
Growth	−248.11	−116.60
Bottlegrowth	−163.37	−76.35
Three epoch	−172.12	−79.89

Because these models are not nested, it is difficult to statistically compare the fits of the five models to our empirical data. However, it seems clear that for both long intron sites and third codon positions, the bottlegrowth and three epoch models fit the observed data better than the neutral, two-epoch, and growth models. It also seems clear that distinguishing between the three-epoch and bottlegrowth models would be difficult, as both models seem to fit the data equally well.

We assessed the robustness of these findings to various aspects of the data. We first explored the possibility that we undercorrected the SFS of the long intron data by using a divergence value of 0.057. However, using expected divergence at third codon positions (0.093) to correct the SFS of long intron sites does not affect our demographic modeling results (Table S3). In addition, polarizing SNPs using multispecies divergence data rather than a mathematical correction does not change our modeling results either, with the bottlegrowth and three-epoch models clearly better supported than simple growth or two-epoch models (Table S3). Further, given that our Sanger data indicated that half of the polymorphic sites missed by JGIL were singletons as well as the sensitivity of demographic modeling to this site class in particular, we refitted the demographic models to both long intron and third codon position sites excluding singletons. In this case as well, demographic models including a bottleneck followed by growth fitted the data much better than simple growth or simple bottleneck models (Table S3). Finally, selection on codon usage at synonymous sites may potentially compromise demographic inference. Although we believe our results are robust to this given that long introns also provide strong support for a bottleneck followed by growth model, we reanalyzed the third codon position data, focusing only on synonymous changes that were either preferred codon → preferred codon or unpreferred codon → unpreferred codon. These synonymous polymorphisms also show strong support for the three-epoch and bottlegrowth models to the exclusion of a simple growth or simple bottleneck model (Table S3).

Given the overwhelming support for the three-epoch and bottlegrowth models and their robustness to myriad aspects of these data, we chose these two models for further exploration. Comparing between long introns and third codon positions, both models fitted the third codon position data much better than the long intron data (Figure S5). This is likely due to the excess of high-frequency-derived alleles,

Table 2 Demographic parameter estimates under the bottlegrowth model

	nuB ^a	nuF ^b	T ^c	LL
Long introns	0.03	6.11	0.32	−163.37
Masked long introns	0.007	1.81	0.14	−58.94
Third codon positions	0.016	0.76	0.14	−76.64

^a Ratio of population size after instantaneous change to ancient population size.

^b Ratio of contemporary to ancient population size.

^c Time in the past at which instantaneous change happened and growth began (in $2N_e$ generations)

which is more pronounced in the long intron data than it is in the third codon position data (Figure S4 and Figure S5). To explore this possibility, we refitted the bottlegrowth and three-epoch models to the long intron polymorphism data, only considering those polymorphisms that were present at a frequency of 15 or fewer. Demographic parameter estimates and model log-likelihoods for the long intron data, the masked long intron data, and the third codon position data are presented in Table 2 (bottlegrowth model) and Table 3 (three-epoch model).

Inferences of selection

To gain insight into the history of positive selection in this 2.1-Mb region, we fitted models of single and recurrent hitchhiking. We used the Pavlidis *et al.* (2010) framework, which takes into account both the SFS and patterns of linkage disequilibrium to fit single hitchhiking models. We found three regions with strong statistical signals of a recent selective sweep. Given the resolution afforded by our bootstrapping, these putatively swept loci are contained within the following regions on the physical map: 13.922–13.924 Mbp ($P < 0.0001$), 15.248–15.250 Mbp ($P < 0.0001$), and 15.600–15.602 Mbp ($P < 0.0001$). Fitting models of recurrent hitchhiking (Jensen *et al.* 2008a) facilitated estimating the mean rate of adaptive evolution ($2N\lambda$) and the mean selection coefficient (s). The fit of the RHH models to our data are quite good; the fraction of simulations of our best-fitting model falling within the tolerance (defined as 0.001) of the real value for each summary statistic used in the ABC analysis is presented in Table S4. We estimate mean s at 0.0009 [95% credibility interval (CI): 0.0005–0.001] and mean λ at 6×10^{-11} (95% CI: 1×10^{-11} – 9×10^{-11})/site/generation.

Although the Jensen *et al.* (2008a) method is reasonably robust to nonequilibrium demography, we assessed the impact of our inferred population bottleneck on the estimation of these selection parameters. We simulated neutral evolution using the specific demographic models estimated from our third codon position data (Table 2 and Table 3) and reestimated the mean rate of adaptive evolution and mean selection coefficient from these simulated data. This approach yields a rate of adaptation 5 – 7×10^{-9} and a selection coefficient of 6 – 9×10^{-6} depending on the specific demographic model simulated (Table 2 and Table 3). Both parameter estimates from the simulated data differ by 2 orders

Table 3 Demographic parameter estimates under the three epoch model

	nuB ^a	nuF ^b	T _N ^c	T _B ^d	LL
Long introns	0.0028	0.11	0.014	0.01	−172.12
Masked long introns	0.039	0.48	0.66	0.025	−65.28
Third codon positions	0.064	0.47	0.144	0.059	−79.89

^a Ratio of bottleneck population size to ancient population size.^b Ratio of contemporary to ancient population size.^c Length of bottleneck.^d Time since bottleneck recovery.

of magnitude as compared to the estimates from empirical data. The selection coefficient is underestimated by 2 orders of magnitude and the rate of adaptation is overestimated by 2 orders of magnitude. This thus indicates that the demographic model alone does not result in the observed inferences of selection parameters.

To assess the impact of an incorrect underlying demographic model in inferences of recurrent hitchhiking parameters, we used a forward simulation approach. We used our maximum *a posteriori* (MAP) estimates of the selection coefficient and rate of adaptation (0.0009 and 6×10^{-11} , respectively) and simulated selection under our inferred demographic model (Table 2 and Table 3). We then assumed a simple growth model (Li and Stephan 2006a) in our estimation of the selection coefficient and rate of adaptation. When the simple growth model is assumed, both the selection coefficient and the rate of adaption are overestimated by an order of magnitude ($s = 0.01$, $2N\lambda = 4 \times 10^{-10}$).

We examined the three putatively swept regions identified by the single hitchhiking model listed above for additional signatures of positive selection. We also explored the two regions with most depressed nucleotide diversity (coordinates of the physical map: 15,226,236–15,301,236 and 15,596,236, 15,636,236; see Figure 2a) for further evidence of positive selection. It is worth noting that these two regions with notably reduced π contain two of the three windows identified by the single hitchhiking models.

In total, there were 15 protein-coding genes contained within these three genomic regions. We conducted McDonald–Kreitman (McDonald and Kreitman 1991) tests to examine whether patterns of polymorphism and divergence at any of these genes were consistent with recurrent positive selection. Tests of positive selection in a Phylogenetic Analysis by Maximum Likelihood (PAML) framework (Yang 1997) have already been performed for all 1:1 orthologs in the (sequenced) *D. melanogaster* species group (Larracuente *et al.* 2008), and we mined these data to see whether any PAML model (site, branch, and/or branch site) was consistent with positive selection. We note that this PAML analysis is likely to be conservative because many rapidly evolving genes were not included in the Larracuente *et al.* data set due to the difficulty in assigning orthology for rapidly evolving genes. Our results are presented in Table 4. Four of the 15 genes have statistically significant McDonald–Kreitman

Table 4 Tests of recurrent positive selection

Gene	Gene Name	PAML ^a	McDonald–Kreitman ^b
FBgn0030630 ^c		NA	P = 0.0005
FBgn0030631 ^c		No	P = 0.31
FBgn0000028 ^{c,d}	<i>acj6</i>	NA	P = 0.47
FBgn0030672 ^{d,e}		No	P = 0.326
FBgn0030673 ^e		Yes	P = 0.15
FBgn0030674 ^e		Yes	P = 0.00001
FBgn0030675 ^e		Yes	P = 0.42
FBgn0030676 ^e		Yes	P = 0.35
FBgn0011741 ^e	<i>actr13E</i>	Yes	P = 1
FBgn0033391 ^e		Yes	P = 0.02
FBgn0026666 ^e	<i>l(1)G0136</i>	No	P = 1
FBgn0030678 ^d		No	P = 0.087
FBgn0030680 ^c		No	P = 0.01
FBgn0030628 ^c		No	P = 0.83
FBgn0026428 ^c	<i>HDAC6</i>	NA	P = 0.68

^a The PAML results are from Larracuente *et al.* (2008). “NA” indicates that this gene was not included in the initial PAML analysis, “Yes” indicates that PAML results are consistent with positive selection for this gene (in at least the *D. melanogaster* lineage), which could result from statistically significant site models, branch models (for branches including the *melanogaster* lineage) and/or branch-site models (for branches including the *melanogaster* lineage). “No” indicates that although this gene was included in the PAML analysis, no statistically significant signal consistent with positive selection was found.

^b Statistical significance was calculated using Fisher’s exact test. Bold denotes significance at $P < 0.05$.

^c Contained within window of reduced π (region coordinates: 15,226,236–15,301,236).

^d Contained within window identified by single hitchhiking model.

^e Contained within window of reduced π (region coordinates: 15,596,236–15,636,236).

tests, 6 of the 15 genes show some evidence consistent with positive selection from PAML analysis, and 2 genes show patterns of polymorphism and divergence that are consistent with recurrent selection from both PAML and McDonald–Kreitman analyses.

Discussion

Sequencing strategy

To maximize sequence read depth and quality, we used a custom NimbleGen hybridization array to focus our sequencing efforts to the 2.1-Mb high-recombination X chromosome region between the genes *garnet* and *scalloped*. This approach proved highly effective, with >65% of overall reads mapping to our target region on average. Moreover, on average, >90% of reads that map uniquely to the genome in fact map to our target. It is thus clear that our custom Nimblegen comparative genomics hybridization platform was highly effective for targeted sequencing of the 2.1-Mb X chromosome region of interest.

We took advantage of a recently developed base-calling algorithm, JGIL (Stone 2012), specifically designed for inbred lines, to call bases from our sequence data. In our experimental framework, our lines are not inbred but rather are isogenic for the X chromosome due to chromosome extraction. We thus adjusted the algorithm, which uses a prior on how much expected residual heterozygosity there should be given the inbreeding strategy, to take into account that

the chromosome extraction procedure should yield zero residual heterozygosity within lines. Our Sanger sequencing data are consistent with the presumed lack of heterozygosity within these lines, with >99.99% of bases showing unambiguously monomorphic chromatograms. Of the 109,686 sites interrogated by Sanger sequencing, only two yielded ambiguous base calls and in both cases the ambiguity appears driven by noise rather than heterozygosity.

Application of JGIL to our data yielded base calls at 91% of sites overall. To assess error rates on the base-calling algorithm, we Sanger sequenced a handful of loci distributed across the 2.1-Mb region. Our results (Figure 1) indicate that ~99% of JGIL base calls are correct, which is in close agreement with other empirical validation of JGILs performance (Stone 2012). Our accuracy estimates are likely to be biased slightly downward because we specifically chose loci containing at least one *N* in the JGIL base calls for our Sanger sequencing. Our reported accuracy estimates are therefore conservative and when coupled with existing validation data (Stone 2012), it is clear that JGIL is an exceptional tool for inferring bases from next-generation sequence data of appropriately constructed lines.

Polymorphism and divergence

The polymorphism data collected here represent an exciting opportunity to examine patterns of polymorphism across functional categories of sites for a reasonably large data set. We thus examined patterns of nucleotide diversity for first, second, and third coding positions, short and long introns, UTRs, and intergenic regions. Estimates of variation using Watterson's θ as well as π suggest a consistent hierarchy of nucleotide diversity (Figure 2), with second codon positions showing the least polymorphism, followed by the first coding position. As all second codon positions are nonsynonymous and most first coding positions are as well, this is thus consistent not only with the generally depressed variability at these classes of sites, but also with their relative levels of polymorphism. In contrast, short introns and third codon positions show the highest level of nucleotide variability, while UTR sequences, intergenic regions, and long introns show intermediate levels of polymorphism. These relative levels of polymorphism across functional categories of site are broadly consistent with previous reports from a variety of *Drosophila* species including *D. miranda* (Bachtrog and Andolfatto 2006), *D. simulans* (Begun *et al.* 2007; Haddrill *et al.* 2008), and *D. melanogaster* (Andolfatto 2005; Parsch *et al.* 2010; Zeng and Charlesworth 2010).

As divergence is often used as a measure of constraint, previous work on divergence across functional annotation categories in *Drosophila* abound. Our results largely echo previous findings, with the relative hierarchy of divergence across these functional categories of site consistent with that shown in polymorphism. That is, first and second codon position show the least divergence with *D. sechellia*, which is again consistent with expectation given functional con-

straint of these sites as well as being consistent with previous studies in *Drosophila* (Andolfatto 2005; Bachtrog and Andolfatto 2006; Halligan and Keightley 2006; Haddrill *et al.* 2008; Parsch *et al.* 2010). We also recover high levels of divergence for short introns and third codon positions, which are comparatively unconstrained site classes, which also confirms previous results (Andolfatto 2005; Bachtrog and Andolfatto 2006; Halligan and Keightley 2006; Haddrill *et al.* 2008; Parsch *et al.* 2010). Finally, intermediate and similar levels of divergence are found in UTRs, intergenic regions, and long introns, consistent with previous reports (Bergman and Kreitman 2001; Halligan *et al.* 2004; Andolfatto 2005; Halligan and Keightley 2006; Haddrill *et al.* 2008).

Demographic history

Considerable advances in our understanding of the demographic history of *D. melanogaster* have been made in the past 25 years (for review see Stephan and Li 2007). Population level genetic data consistently support an African origin of *D. melanogaster* (e.g., Begun and Aquadro 1993; Glinka *et al.* 2003; Ometto *et al.* 2005; Baudry *et al.* 2006; Pool and Aquadro 2006; Schlotterer *et al.* 2006; Nunes *et al.* 2008). Much of the work to date on inferring demographic parameters in *D. melanogaster* has been framed within the context of demography in derived populations (Haddrill *et al.* 2005; Ometto *et al.* 2005; Thornton and Andolfatto 2006; Laurent *et al.* 2011). However, understanding the demographic history of East African, presumed ancestral populations is imperative for making demographic inferences in derived populations given that the modeling of derived populations necessarily requires making assumptions about the ancestral populations from which they were derived. It seems likely that mis-inference of the ancestral population history will lead to subsequent mis-inference in derived populations, and our results strongly suggest that commonly held assumptions about the history of this ancestral population are incorrect.

Although not exhaustively modeled, it is clear from previous work that African populations of *D. melanogaster* show substantial deviation from equilibrium (Haddrill *et al.* 2005; Ometto *et al.* 2005; Li and Stephan 2006a; Thornton and Andolfatto 2006). The observed excess of rare mutations in African population samples (Glinka *et al.* 2003; Haddrill *et al.* 2005; Ometto *et al.* 2005) is consistent with population growth, and parameters of growth models have been estimated using several data sets (Ometto *et al.* 2005; Li and Stephan 2006a; Laurent *et al.* 2011). Such models have yielded mean estimates of the ratio of current to ancestral population sizes of 2.6–5 and the mean estimated time of expansion ranges from 15,000 to 60,000 years before present (Ometto *et al.* 2005; Li and Stephan 2006a; Laurent *et al.* 2011).

However, it has also been suggested that simple growth models do not adequately describe observed patterns of polymorphism in East African samples of *D. melanogaster* and that multilocus patterns of variability in such samples

are better explained by a simple bottleneck model (Haddrill *et al.* 2005). This likely results from the observation that while growth models reduce variance among loci relative to equilibrium models, bottlenecks inflate variance relative to equilibrium models (Thornton and Andolfatto 2006). Our results are consistent with this latter result, as our data from a Uganda population of *D. melanogaster* appear more consistent with models that include both a population contraction followed by expansion than simple growth models for both long introns and third codon positions (Table 1).

Although it is clear that our polymorphism data fit the bottlegrowth and three epoch models better than the other models implemented, we cannot statistically distinguish between the fits of these two models to our data. However, the two models are qualitatively quite similar in that they both involve a population reduction followed by recovery (Table 2, Table 3). For the third codon position data, the parameter estimates for the ratio of the current population size to the ancient population sizes are within an order of magnitude between the two models (0.47 vs. 0.76; Table 2 and Table 3), and the estimated time since recovery (in units of $2N_e$) are similar in magnitude as well (0.059 vs. 0.14; Table 2 and Table 3). Moreover, the ratios of the bottlenecked population size to the ancient population sizes are similar between the two models (0.064 vs. 0.016; Table 2 and Table 3).

The unmasked long intron data are more difficult to interpret. Notably, the fit of the model to the data are markedly reduced relative to the masked long intron data and the third codon position. Again, we believe this is likely due to an excess of high-frequency-derived mutations, which would be consistent with positive selection at long intron sites, which has been shown previously in *D. melanogaster* and *D. simulans* (Andolfatto 2005; Haddrill *et al.* 2008). The comparatively poor fit of the model to the (unmasked) intron data, which may be driven by positive selection in this functional class of site, may underlie that observed disparity in the parameter estimates between the two demographic models (Table 2, Table 3).

However, if we mask high-frequency sites (see *Materials and Methods*), parameter estimates for these data are within an order of magnitude difference between the two demographic models. Perhaps more importantly, the parameter estimates within each demographic model comparing third codon sites to the masked long intron data are quite similar to each other. We thus believe that within our data set, given the number of polymorphisms within each functional classification of site and the frequency spectra of those polymorphisms, the third codon position data and the masked long intron data are most appropriate for demographic inference. Our data are thus consistent with a model in which the ancestral population was reduced to 0.007–0.064 of its ancestral size followed by recovery 0.14–0.66 ($2N_e$) generations ago to a contemporary size that is 0.47–1.81 times the ancestral population size. Assuming an effective population size of 1,000,000 and 10 generations per year, this corre-

sponds to a recovery beginning 28,000–132,000 years before present, which is broadly consistent with previous work indicating expansion times 14,000–60,000 years before present (Ometto *et al.* 2005; Li and Stephan 2006a; Laurent *et al.* 2011). Note that while linked selection may affect such demographic estimation (Zeng and Charlesworth 2009), the consistency between the demographic models fit to both our long intron and third codon position data suggest a certain level of robustness. Moreover, our findings do not appear to be compromised by mis-inference of ancestral state, selection on codon usage, or incorrect base calls. Thus, the general demographic history of this African population appears to be characterized by a severe population reduction followed by expansion.

It is also worth noting that although we focus on a high recombination region of the X chromosome to minimize the effects of linkage, there will certainly be linkage among sites in this region. Given that this region will be subject to purifying and positive selection, this linkage will limit our ability to make demographic inferences, and an analysis of a single continuous 2-Mb region will not be as powerful as a collection of smaller regions that are completely unlinked. Moreover, demographic events will differentially affect levels of polymorphism and patterns of linkage on the X chromosome vs. the autosomes (assuming a 1:1 sex ratio), which suggests that generalizing our results from X-linked data to the entire genome should be done with caution. However, our data set remains the largest polymorphism data set from an African population to date and our results highlight the need to thoroughly explore the demographic history of presumed ancestral populations of this species. Inclusion of additional loci, particularly the autosomes, will be instrumental in identifying the demographic models most appropriate for this population as well as refining parameter estimates regarding the timing and severity of the bottleneck as well as the nature of the following population expansion.

Positive selection

There is mounting evidence that adaptation is widespread in *Drosophila* (for review see Sella *et al.* 2009), and there is a general interest in both localizing particular targets of selection and estimating the average rate and strength of selective events. We used a variety of approaches to address these questions using this polymorphism data set from a 2.1-Mb region on the X chromosome. Importantly, we focus on a presumed ancestral population of *D. melanogaster* for our inferences regarding the adaptive history of this species. Previous studies (e.g., Pool and Aquadro 2006) have inferred these populations to be ancestral and to have had larger and more stable effective population sizes. Thus, inferences of selection in this population are less likely to be compromised by demographic history. Although there is strong evidence in support of a nonequilibrium demographic history in African populations of *D. melanogaster*, given the size and relative stability of effective population size compared to what is

found in derived populations, we expect that the adaptive history of these ancestral populations is more reflective of general patterns and processes of adaptation in *Drosophila*.

To infer the rate of adaptation and strength of selection in this region, we fitted a recurrent hitchhiking model (Jensen *et al.* 2008a); our data are consistent with a mean selection coefficient of 0.0009 and an adaptation rate of 6×10^{-11} /site/generation. Previous estimates of these parameters for *D. melanogaster* and *D. simulans* have a considerable range, with selection coefficients ranging from 10^{-5} to 0.01 and the rate of adaptation ranging from 3.6×10^{-12} to 7.5×10^{-10} (Li and Stephan 2006a; Andolfatto 2007; Macpherson *et al.* 2007; Jensen *et al.* 2008a). These previous results are often described as being polarized between a low frequency of adaptive events with large selection coefficients vs. a high frequency of adaptive events with small selective effects. However, the boundaries between “large” and “small” selection coefficients are poorly defined and arbitrary at best. Moreover, these two models are not incompatible with each other, and in fact recent evidence from *D. simulans* is suggestive of two classes of beneficial mutations, one of which has a mean selection coefficient of 0.005 and one of which has a mean selection coefficient of 4×10^{-5} (Sattath *et al.* 2011), which nearly spans the entire range of selection coefficients estimated previously.

We thus refrain from qualitatively characterizing our estimates of the rate and strength of adaptive events and instead note that our estimates of selective coefficients are firmly sandwiched between all of the published estimates from *D. melanogaster* (Li and Stephan 2006a; Andolfatto 2007; Jensen *et al.* 2008a), with an order of magnitude increase over the Andolfatto (2007) estimate and an order of magnitude decrease over the Jensen *et al.* (2008a) and Li and Stephan (2006a) estimates. Our estimated rates of adaptation are similarly consistent with previous data from *D. melanogaster*, on the same order of magnitude as the Jensen *et al.* (2008a) and Li and Stephan (2006a) estimates, and an order of magnitude lower than the Andolfatto (2007) estimate. Notably, although previous estimates in *D. melanogaster* were indeed based on X-linked data from an African population (Li and Stephan 2006a; Andolfatto 2007; Jensen *et al.* 2008a), our data set is by far the largest polymorphism data set to be used for this type of inference in *D. melanogaster*, covering roughly an order of magnitude larger data set. Moreover, that our data are from a single continuous region gives us further confidence in our results (Jensen *et al.* 2008b), although again we note that linkage among sites will limit our ability to fully discriminate among distinct selective events. Given the high rate of recombination in this region and the consistency of our results with previous work, we do not believe that linkage among sites has markedly biased our parameter estimation.

It is important to note that background selection will both reduce nucleotide variation and skew the SFS toward rare alleles. As a consequence, not accounting for background selection may artificially inflate recurrent hitchhiking estimates.

Models that estimate the distribution of fitness effects of new mutations (Eyre-Walker and Keightley 2007; Keightley and Eyre-Walker 2007; Boyko *et al.* 2008; Schneider *et al.* 2011; Wilson *et al.* 2011) are consistent with both deleterious and advantageous mutations contributing significantly to segregating polymorphisms, with the vast majority of new mutations being deleterious. However, there is reason to be hopeful that our insights into the adaptive landscape of this Uganda population of *D. melanogaster* are not entirely driven by background selection, as this process should not result in an excess of high-frequency derived alleles and linkage disequilibrium (and in particular the specific linkage disequilibrium pattern captured by the omega_max statistic) (Kim and Nielsen 2004; Jensen *et al.* 2007b; Pavlidis *et al.* 2010), both of which are critical parameters for drawing inference about past beneficial fixations. Moreover, estimates of the selection coefficient of adaptive mutations from models incorporating deleterious mutations are comparable to our own (Schneider *et al.* 2011), which gives us further confidence that our results have not been grossly compromised by failing to consider background selection.

The discrepancies between our estimates and previous estimates may reflect our use of a demographic model including a bottleneck followed by growth, as our results indicate that estimation of the selection coefficient and rate of adaptation are overestimated if an incorrect demographic model is assumed. However, the general concordance of parameter estimates under recurrent hitchhiking models from a variety of data sets using a variety of approaches suggests that these estimates are likely to characterize the adaptive history of the X chromosome of *D. melanogaster*.

We also localized putative targets of selection by fitting single hitchhiking models (Pavlidis *et al.* 2010) to our polymorphism data set. Additional exploratory analyses of the genes contained within the windows identified by this modeling framework as well as those genes contained within windows of notably depressed polymorphism (Figure 2) are consistent with recurrent selection in many cases (Table 4). These genes, particularly the two genes that show signals consistent with positive selection in both a PAML framework and a McDonald–Kreitman framework, are potentially interesting candidates for future studies.

Acknowledgments

The authors gratefully acknowledge R. Gutenkunst for his assistance with dadi and R. Gutenkunst and R. Hernandez for their roles in helping design and implement a *Drosophila*-specific correction for potential mis-inference of ancestral state in dadi. We also thank E. Stone for his efforts in applying JGIL to our sequence data. Comments from two anonymous reviewers greatly improved this manuscript. This work was supported by a Priming Grant from the Cornell Center for Comparative and Population Genomics and a National Institutes of Health grant to C.F.A. (R01-GM036431).

Literature Cited

- Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver *et al.*, 2004 Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2: 1591–1599.
- Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437: 1149–1152.
- Andolfatto, P., 2007 Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res.* 17: 1755–1762.
- Bachtrog, D., and P. Andolfatto, 2006 Selection, recombination and demographic history in *Drosophila miranda*. *Genetics* 174: 2045–2059.
- Baudry, E., B. Virginier, and M. Veuille, 2006 Non-African populations of *Drosophila melanogaster* have a unique origin. *Mol. Biol. Evol.* 21: 1482–1491.
- Beaumont, M. A., W. Y. Zhang, and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. *Genetics* 162: 2025–2035.
- Begun, D. J., and C. F. Aquadro, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature* 365: 548–550.
- Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y.-P. Poh *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* 5: e310.
- Bergman, C. M., and M. Kreitman, 2001 Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* 11: 1335–1345.
- Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4(5): e1000083.
- Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8: 610–618.
- Fay, J. C., G. J. Wyckoff, and C. I. Wu, 2002 Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* 415: 1024–1026.
- Fiston-Lavier, A. S., N. D. Singh, M. Lipatov, and D. A. Petrov, 2010 *Drosophila melanogaster* recombination rate calculator. *Gene* 463: 18–20.
- Fu, Y. X., 1996 New statistical tests of neutrality for DNA samples from a population. *Genetics* 143: 557–570.
- Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. De Lorenzo, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* 165: 1269–1278.
- Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5(10): e1000695.
- Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto, 2005 Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* 15: 790–799.
- Haddrill, P. R., D. Bachtrog, and P. Andolfatto, 2008 Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Mol. Biol. Evol.* 25: 1825–1834.
- Halligan, D. L., and P. D. Keightley, 2006 Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16: 875–884.
- Halligan, D. L., A. C. Eyre Walker, P. Andolfatto, and P. D. Keightley, 2004 Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res.* 14: 273–279.
- Harr, B., M. Kauer, and C. Schlotterer, 2002 Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 99: 12949–12954.
- Hernandez, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* 24: 2786–2787.
- Hernandez, R. D., S. H. Williamson, and C. D. Bustamante, 2007 Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Biol. Evol.* 24: 1792–1800.
- Jensen, J. D., Y. Kim, V. Bauer DuMont, C. F. Aquadro, and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* 170: 1401–1410.
- Jensen, J. D., V. L. B. DuMont, A. B. Ashmore, A. Gutierrez, and C. F. Aquadro, 2007a Patterns of sequence variability and divergence at the diminutive gene region of *Drosophila melanogaster*: complex patterns suggest an ancestral selective sweep. *Genetics* 177: 1071–1085.
- Jensen, J. D., K. R. Thornton, C. D. Bustamante, and C. F. Aquadro, 2007b On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics* 176: 2371–2379.
- Jensen, J. D., K. R. Thornton, and P. Andolfatto, 2008a An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genet.* 4(9): e1000198.
- Jensen, J. D., K. R. Thornton, and C. F. Aquadro, 2008b Inferring selection in partially sequenced regions. *Mol. Biol. Evol.* 25: 438–446.
- Keightley, P. D., and A. C. Eyre-Walker, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177: 2251–2261.
- Kelly, J. K., 1997 A test of neutrality based on interlocus associations. *Genetics* 146: 1197–1206.
- Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513–1524.
- Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765–777.
- Larracuente, A. M., T. B. Sackton, A. J. Greenberg, A. Wong, N. D. Singh *et al.*, 2008 Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* 24: 114–123.
- Laurent, S. J. Y., A. Werzner, L. Excoffier, and W. Stephan, 2011 Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. *Mol. Biol. Evol.* 28: 2041–2051.
- Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.
- Li, H., and W. Stephan, 2006a Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet.* 2: e166.
- Li, H., and W. Stephan, 2006b Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* 171: 377–384.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- Lohmueller, K. E., C. D. Bustamante, and A. G. Clark, 2011 Detecting directional selection in the presence of recent admixture in African-Americans. *Genetics* 187: 823–835.
- Macpherson, J. M., G. Sella, J. C. Davis, and D. A. Petrov, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. *Genetics* 177: 2083–2099.
- Marais, G., D. Mouchiroud, and L. Duret, 2001 Does recombination improve selection on codon usage?: lessons from nematode and fly complete genomes. *Proc. Natl. Acad. Sci. USA* 98: 5688–5692.

- McDonald, J., and M. Kreitman, 1991 Adaptive protein evolution in *Drosophila*. *Nature* 351: 652–654.
- Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. *Genome Res.* 15: 1566–1575.
- Nielsen, R., M. J. Hubisz, I. Hellmann, D. Torgerson, A. M. Andres *et al.*, 2009 Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* 19: 838–849.
- Nunes, M. D. S., H. Neumeier, and C. Schlotterer, 2008 Contrasting patterns of natural variation in global *Drosophila melanogaster* populations. *Mol. Ecol.* 17: 4470–4479.
- Ometto, L., S. Glinka, D. De Lorenzo, and W. Stephan, 2005 Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Mol. Biol. Evol.* 22: 2119–2130.
- Parsch, J., S. Novozhilov, S. S. Saminadin-Peter, K. M. Wong, and P. Andolfatto, 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: 1226–1234.
- Pavlidis, P., J. D. Jensen, and W. Stephan, 2010 Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* 185: 907–922.
- Pool, J. E., and C. F. Aquadro, 2006 History and structure of sub-Saharan populations of *Drosophila melanogaster*. *Genetics* 174: 915–929.
- Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832–837.
- Sattath, S., E. Elyashiv, O. Kolodny, Y. Rinott, and G. Sella, 2011 Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Genet.* 7(2): e1001302.
- Schlotterer, C., H. Neumeier, C. Sousa, and V. Nolte, 2006 Highly structured Asian *Drosophila melanogaster* populations: A new tool for hitchhiking mapping? *Genetics* 172: 287–292.
- Schneider, A., B. Charlesworth, A. Eyre-Walker, and P. D. Keightley, 2011 A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* 189: 1427–1437.
- Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural selection in the *Drosophila* genome? *PLoS Genet.* 5(6): e1000495.
- Shapiro, J. A., W. Huang, C. H. Zhang, M. J. Hubisz, J. Lu *et al.*, 2007 Adaptive genic evolution in the *Drosophila* genomes. *Proc. Natl. Acad. Sci. USA* 104: 2271–2276.
- Singh, N. D., P. F. Arndt, A. G. Clark, and C. F. Aquadro, 2009 Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila*. *Mol. Biol. Evol.* 26: 1591–1605.
- Stephan, W., and H. Li, 2007 The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* 98: 65–68.
- Stephan, W., Y. S. Song, and C. H. Langley, 2006 The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* 172: 2647–2663.
- Stone, E. A., 2012 Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines. *Genome Res.* 22(5): 966–974.
- Thornton, K., and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* 172: 1607–1619.
- Thornton, K. R., 2009 Automating approximate Bayesian computation by local linear regression. *BMC Genet.* 10: 35.
- Thornton, K. R., and J. D. Jensen, 2007 Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* 175: 737–750.
- Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* 7(12): e1002395.
- Welch, J. J., 2006 Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics* 173: 821–837.
- Wiehe, T. H. E., and W. Stephan, 1993 Analysis of a genetic hitchhiking model and its application to DNA polymorphism data in *Drosophila melanogaster*. *Mol. Biol. Evol.* 10: 842–854.
- Wilson, D. J., R. D. Hernandez, P. Andolfatto, and M. Przeworski, 2011 A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet.* 7(12): e1002395.
- Yang, Z., 1997 PAML: am program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13: 555–556.
- Zeng, K., and B. Charlesworth, 2009 Estimating selection intensity on synonymous codon usage in a nonequilibrium population. *Genetics* 183: 651–662.
- Zeng, K., and B. Charlesworth, 2010 Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. *J. Mol. Evol.* 70: 116–128.

Communicating editor: W. Stephan

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.145318/-/DC1>

Inferences of Demography and Selection in an African Population of *Drosophila melanogaster*

Nadia D. Singh, Jeffrey D. Jensen, Andrew G. Clark, and Charles F. Aquadro

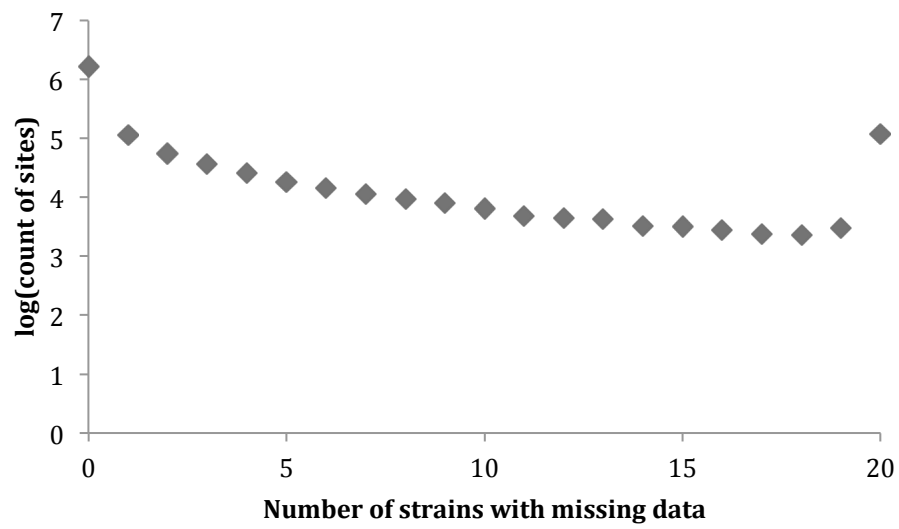


Figure S1 Distribution of missing data across sites. Note the log-scale on the y-axis.

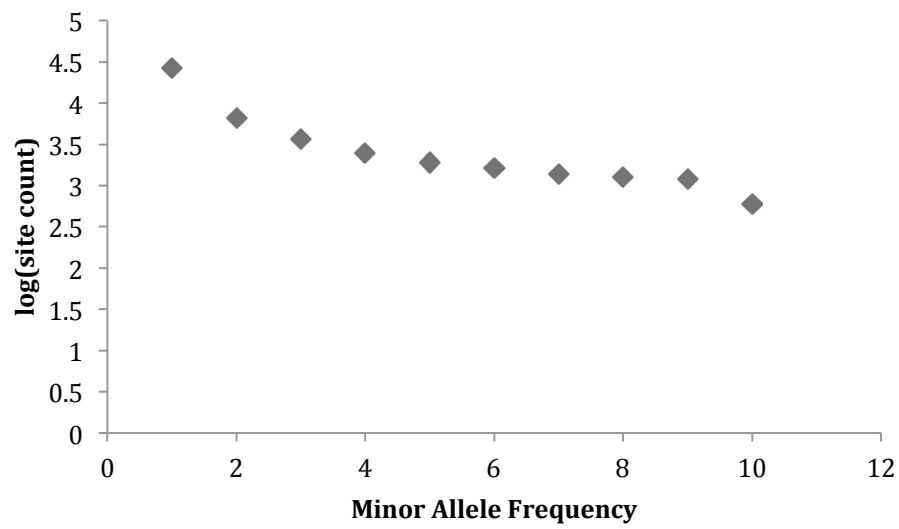


Figure S2 Folded site frequency spectrum of bi-allelic SNPs with called bases in all 20 lines. Note the log scale on the y-axis.

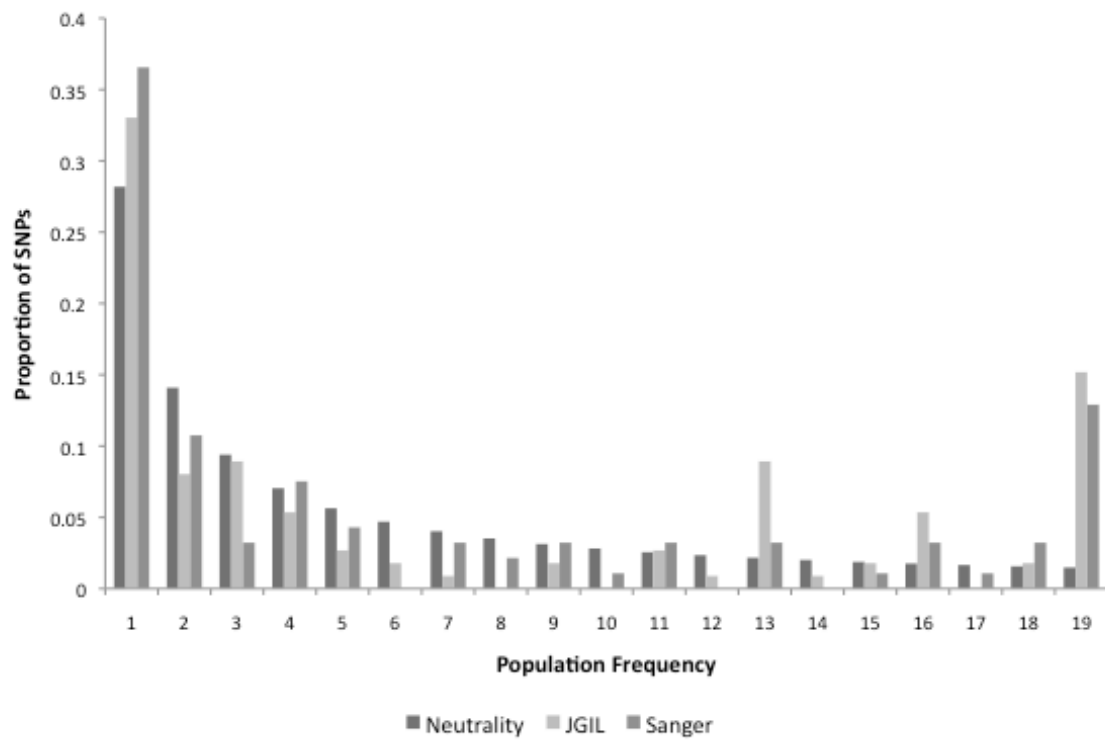
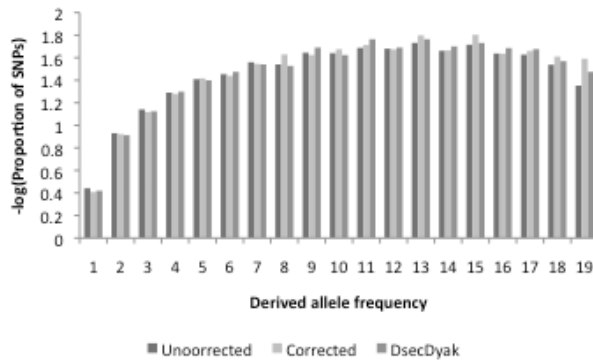


Figure S3 Site frequency spectrum of polymorphic sites based on polymorphisms called by JGIL versus Sanger sequencing with neutral expectation plotted for reference.

a)



b)

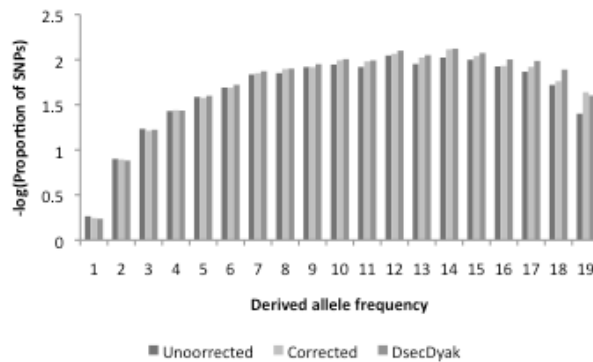
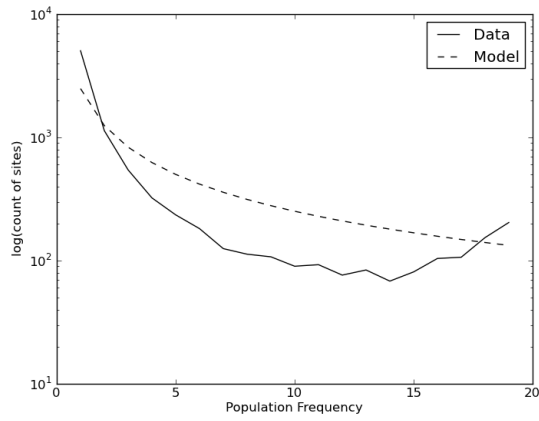
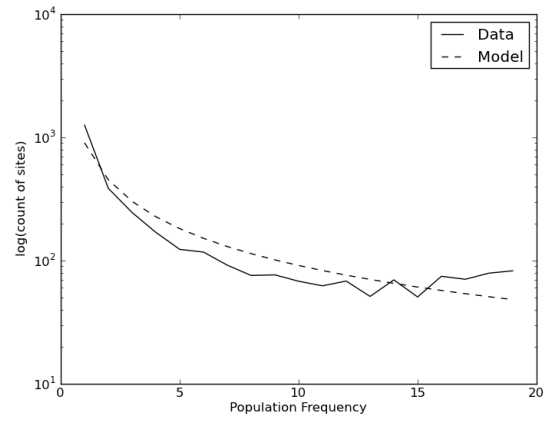


Figure S4 Site frequency spectrum plotted on a negative log scale for a) long introns and b) third codon positions. The uncorrected site frequency spectrum ('uncorrected'), the corrected site frequency spectrum ('corrected', see Methods), and the site frequency spectrum obtained by polarizing SNPs to both *D. sechellia* and *D. yakuba* ('DsecDyak') are presented. Note that the corrected site frequency spectrum shows a lowest proportion of SNPs with high derived allele frequencies.

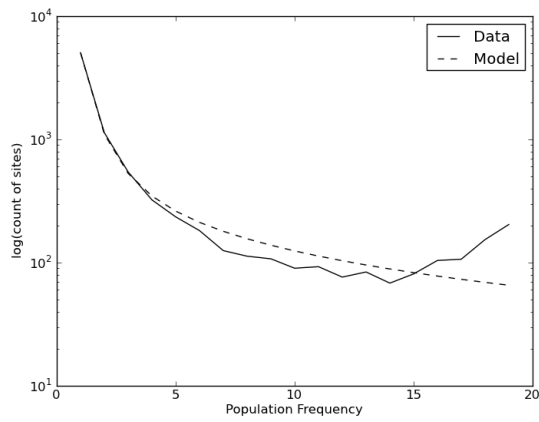
a) LI Neutral



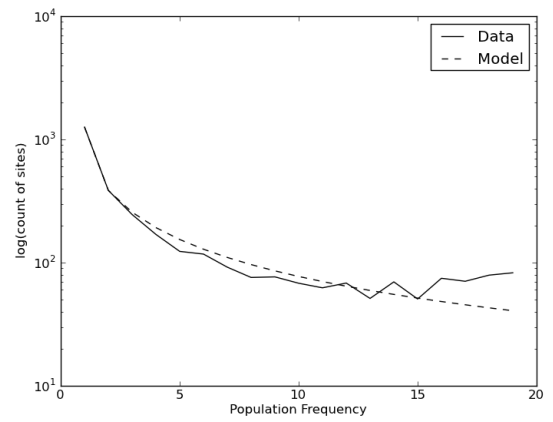
b) Third Neutral



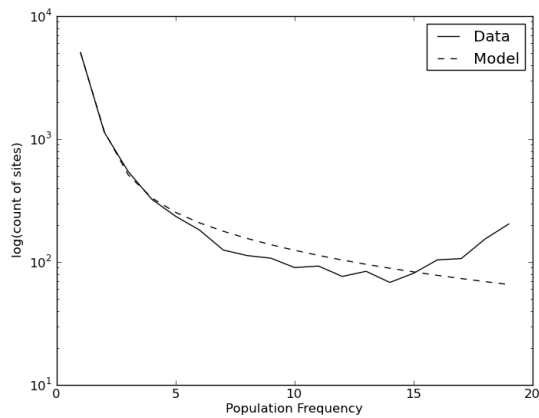
c) LI Growth



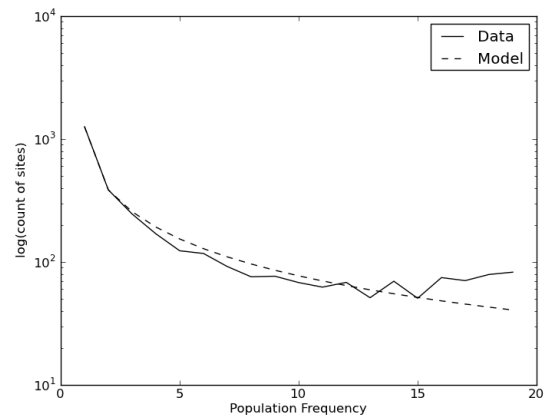
d) Third Growth



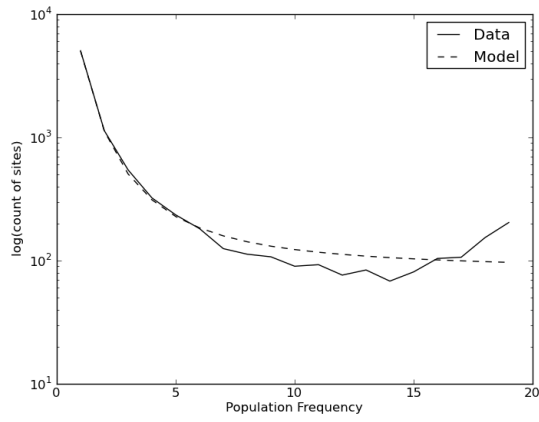
e) LI Two Epoch



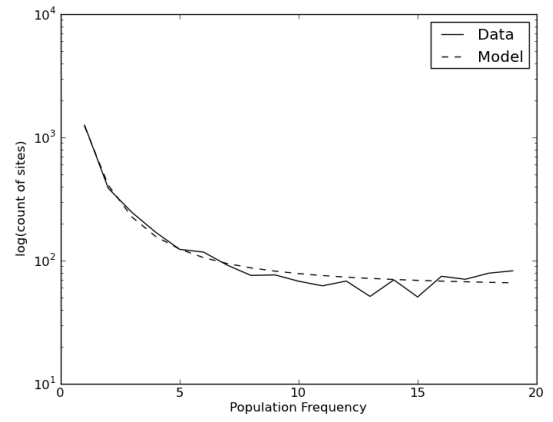
f) Third Two Epoch



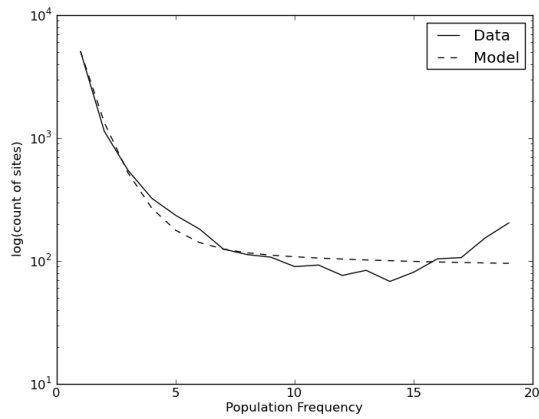
g) LI Bottlegrowth



h) Third Bottlegrowth



i) LI Three Epoch



j) Third Three Epoch

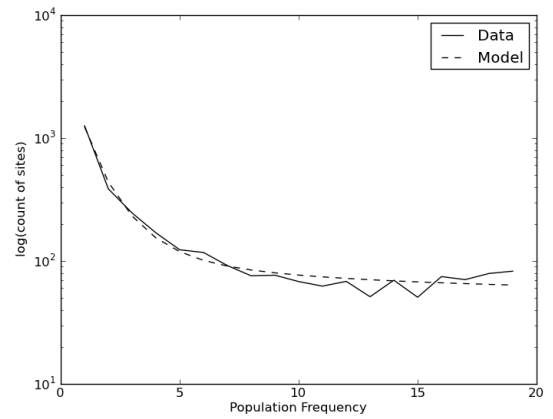


Figure S5 Comparison of observed versus expected site frequency spectra for a) long intron data under a neutral equilibrium model, b) third codon position under a neutral equilibrium model, c) long intron data under a growth model, d) third codon position under a growth model, e) long intron data under a two epoch model, f) third codon position under a two epoch model, g) long intron data under a bottlegrowth model, h) third codon position under a bottlegrowth model, i) long intron data under a three epoch model, j) third codon position under a three epoch model.

Table S1 Primer sequences

Primer	Physical Position (Mbp)	Sequence
VL1L	13.66	CTACATCTTGGAGGTCAACG
VL1R		AGTTGCCTCACTCTCACTTC
VL2L	13.87	AAAAACAGCAGACCAGAATG
VL2R		TAAACAGGACGAAACAGGAC
VL3L	14.08	TCCAGTGGCTCCTACTCAG
VL3R		GACTTCATCATTCTGCTTG
VL4L	14.30	GTCGCGTGTACTTGGTTTC
VL4R		GATTATTCGAAGGGGGAAG
VL5L	14.51	ATGCTTAGTCAGCCGAAATC
VL5R		TATGACCGCCATAAATTCAC
VL6L	14.73	GCAATCTGCAGCTATCGAC
VL6R		TTGCCGCAATCAGAACAC
VL7L	14.94	TGCACAATTCCACTTACAAG
VL7R		GTCGTGTCGAGAGTTGAGTC
VL9L	15.38	ATTGCATTTGCACAGATACG
VL9R		AAGGAGATCGAAGAATGAGG
VL10L	15.59	TGAATTTATGGCACAAGGAC
VL10R		CAGAGGGTTGAAATCGTTC

Table S2 Empirical values of key summary statistics used in the RHH analysis

	Mean	Standard Deviation
π	0.0144	0.091
ϑ_w	0.0136	0.12
ϑ_H	0.0179	0.13
ZnS	0.1284	0.11

Table S3 Log-likelihoods of several demographic models using different subsets of polymorphism data

	Two Epoch	Growth	Bottlegrowth	Three Epoch
Overcorrected long introns ^a	-159.70	-160.72	-121.68	-126.36
Equal change synonymous sites ^b	-81.91	-81.91	-64.75	-66.61
DsecDyak long introns ^c	-317.37	-322.60	-253.87	-226.37
DsecDyak third codon positions ^c	-170.58	-170.59	-96.62	-98.43
Long introns no singletons ^d	-240.13	-242.13	-157.34	-131.20
Third codon position no singletons ^d	-111.51	-111.51	-71.27	-68.68

^aLog-likelihood of several demographic models fit to the long intron polymorphism data corrected for potential mis-inference of ancestral state using expected divergence observed at synonymous sites rather than long introns (see Methods).

^bLog-likelihood of several demographic models fit to equal change synonymous (preferred -> preferred or unpreferred->unpreferred) polymorphism data.

^cLog-likelihood of several demographic models fitted with SNPs polarized to both *D. sechellia* and *D. yakuba*.

^dLog-likelihood of several demographic models fitted with singletons removed.

Table S4 Proportion of simulations falling within tolerance (0.001) of the real value

	Mean	Standard Deviation
π	0.93	0.87
ϑ_w	0.96	0.86
ϑ_H	0.95	0.90
ZnS	0.84	0.79