Surprising differences in the variability of Y chromosomes in African and Cosmopolitan populations of *Drosophila melanogaster*

Amanda M. Larracuente¹,²* and Andrew G. Clark¹

1. Molecular Biology and Genetics, Cornell University, NY 14853
2. Department of Biology, University of Rochester, Rochester, NY 14627
Running title: *D. melanogaster* Y chromosome evolution

Key words/phrases: *Drosophila melanogaster*, Y chromosome evolution, demography, selective sweeps

Corresponding author:

Amanda M. Larracuente

University of Rochester

Department of Biology, River campus

480 Hutchison Hall

Rochester, NY 14627

Phone: 585-276-2183

Fax: 585-275-2070

[alarracu@bio.rochester.edu](mailto:alarracu@bio.rochester.edu)
ABSTRACT

The non-recombining *Drosophila melanogaster* Y chromosome is heterochromatic and has few genes. Despite these limitations, there remains ample opportunity for natural selection to act on the genes that are vital for male fertility and on Y factors that modulate gene expression elsewhere in the genome. Y chromosomes of many organisms have low levels of nucleotide variability, but a formal survey of *D. melanogaster* Y chromosome variation had yet to be done. Here we surveyed Y-linked variation in six populations of *D. melanogaster* spread across the globe. We find surprisingly low levels of variability in African relative to Cosmopolitan (i.e. non-African) populations. While the low levels of Cosmopolitan Y chromosome polymorphism can be explained by the demographic histories of these populations, the staggeringly low polymorphism of African Y chromosomes cannot be explained by demographic history. An explanation that is entirely consistent with the data is that the Y chromosomes of Zimbabwe and Uganda populations have experienced recent selective sweeps. Interestingly, the Zimbabwe and Uganda Y chromosomes differ: in Zimbabwe, a European Y chromosome appears to have swept through the population.

INTRODUCTION

The *Drosophila melanogaster* Y chromosome is highly functionally specialized in male-related activities (Goldstein *et al.* 1982; Hardy *et al.* 1981; Kennison 1981). Although it comprises ~13% of the male genome (40 Mb; Hoskins *et al.* 2002), the *D. melanogaster* Y chromosome has only 13 known protein-coding genes (Carvalho *et al.* 2001; Carvalho *et al.* 2000; Goldstein *et al.* 1982; Krsticevic *et al.* 2010; Vibranovski *et al.* 2008), all thought to be active only in primary spermatocytes in the testis (Hardy *et al.* 1981). The remainder of the Y chromosome is heterochromatic and dense in repetitive elements (Hoskins *et al.* 2002).
Although the *D. melanogaster* Y chromosome is highly repetitive and gene-poor, several lines of evidence suggest that it harbors functionally important variation (Chippindale and Rice 2001; Lemos et al. 2008; Lemos et al. 2010; Rohmer et al. 2004). The Y chromosome has contributed to variation in thermotolerance across *D. melanogaster* populations (Rohmer et al. 2004), has epistatic effects on male fitness (Chippindale and Rice 2001) and has epigenetic effects on the expression of genes across the genome (Jiang et al. 2010; Lemos et al. 2008; Lemos et al. 2010), including specifically in the male germline (Zhang et al. 2000). The functional variation on the Y chromosome has been somewhat of a paradox because, while there is evidence for structural polymorphism (Lyckegaard and Clark 1989), the haploid transmission of the Y chromosome makes it far more difficult to maintain substantial levels of genetic variation (Clark 1987). In one small survey of Y chromosomal variation in a protein-coding gene, *kl-5* (*Dhc-Yh3*), only a single segregating site was discovered in 11 lines of *D. melanogaster* and no variants were found in 10 lines of *D. simulans* (Zurovcova and Eanes 1999). A more recent study in a closely-related species, *Drosophila simulans*, showed low levels of polymorphism in fragments sequenced from the Y-linked genes *kl-2* and *ORY* (Kopp et al. 2006). While no large survey of *D. melanogaster* Y chromosome variation has yet been reported, these small surveys suggest that the Drosophila Y chromosome is nearly devoid of nucleotide variation.

The Y chromosome has an effectively clonal inheritance: it is passed from father to son each generation without recombining. As a result, the Y chromosome is particularly sensitive to the demographic history of populations. Population size fluctuations are a potential cause of reduced Y chromosome variation. *D. melanogaster* originated in sub-Saharan Africa and colonized Europe beginning roughly 16,000 years ago (Thornton and Andolfatto 2006) and reached the Americas only in the last few hundred years (David and Capy 1988). Thus, the
demographic history of Cosmopolitan (non-African) populations includes population bottlenecks corresponding to colonization events. Population size changes can have opposite effects on different regions of the genome, depending on the model parameters (Pool and Nielsen 2007). For example, because of its smaller effective population size, the X chromosome is expected to recover from population size changes faster than the autosomes. Outside of Africa, the ratio of X-to-autosome diversity is lower than the expected \( \frac{3}{4} \) (Andolfatto 2001). A recent, severe bottleneck could potentially explain the dearth of variation on the X chromosome (Pool and Nielsen 2007). These effects should be more pronounced for haploid regions of the genome such as the Y chromosome and mitochondria (Pool and Nielsen 2007)—a theoretical prediction that has yet to be tested with empirical data.

We surveyed Y-linked polymorphism in *D. melanogaster* across the globe to investigate patterns of variation, and to explore the effects of population bottlenecks on patterns of Y chromosome variation and differentiation between populations. We surveyed introns of four Y-linked genes in six populations and show that there indeed are low levels of polymorphism on the Y chromosome. Surprisingly, we observed a striking reduction of variation on the Y chromosome in African relative to Cosmopolitan populations—a pattern that contrasts with polymorphism data for the X chromosome and autosomes. We conclude that while recent natural selection has clearly shaped Y chromosome evolution in African populations, the demographic history of the Cosmopolitan populations is capable of explaining the patterns of diversity on the Y chromosome. Although we cannot formally reject neutrality of the Cosmopolitan Y chromosomes under bottleneck models, other attributes of the data suggest that our ability to detect positive selection on these chromosomes may be limited.

**MATERIALS AND METHODS**

Fly strains
We surveyed *D. melanogaster* isofemale lines from six populations: Zimbabwe (24 lines), Uganda (20 lines), Beijing (17 lines), Tasmania (19 lines), Netherlands (19 lines) and Pennsylvania (24 lines). The lines from the Netherlands were a gift courtesy of Zoltan Bochdanovits; (BOCHDANOVITS and DE JONG 2003). The lines from Tasmania are a gift from Ary Hoffman. Lines from Beijing were provided by Chip Aquadro (BEUN and AQUADRO 1995). The lines from Zimbabwe (ZW) are from Victoria Falls, provided by Bill Ballard. The Uganda lines are a gift from John Pool.

**Re-sequencing of PCR products**

Genomic DNA was isolated from single male flies using Puregene DNA extraction kit (Puregene). We amplified fragments from the introns of four Y-linked genes: *kl-5, kl-3, kl-2*, and *ORY* totaling 7.8 kb in each of 123 fly lines. The fragments correspond to introns in each of the four genes: *kl-5* (*kl-5_int8* forward primer 5’ ACTCTCGACCCACACCTTTG and reverse primer 5’ GCTGCCAAACTGATCCAAAAT; and *kl-5_int10* forward primer 5’ TGTCAGATTGATCCAAAGG and reverse primer 5’ AGATTTGTCTGCAGCTCATC), *kl-3* (*kl-3_int2* forward primer 5’ CGTTTTGGCCATCCTAAAAA and reverse primer 5’ CTCCTTTGATATGGGTGGCAAT; and *kl-3_int6* forward primer 5’ GCCGAAATGGTCTTATGAT and reverse primer 5’ TGGATGCGATTCTCTTGGT), *kl-2* (*kl-2_int1* forward primer 5’ GCAGCAAATAAAAGCGAAGC and reverse primer 5’ TGTAAACCCAATACGCACGA; and *kl-2_int2* forward primer 5’ TTTTAAAATACCAACCTCTCTGCT and reverse primer 5’ AATAAAAGCTGCGGAAACGA), and *ORY* (*ORY_intu1* forward primer 5’ TTATAGCATTCCCTTTTTT and reverse primer 5’ CAGTAAATCCAAAAATTGTATCC; and *ORY_intu2* forward primer 5’ ATTCCGAGTTTACTTTTGTGATACATG and reverse primer 5’
ATCAAGCTGTTATCAAAAGTTCAGC). Primer pairs kl-5_int10 and kl-2_int2 were anchored in 20 bp and 29 bp of an exon, respectively. The coding sequences (totaling 49 bp) were trimmed for the analysis.

Primers were tested in both males and in virgin females to confirm Y-linkage of sequences. Conditions for PCR varied with template and primers and are available on request. Unincorporated nucleotides were removed using Exonuclease I/Shrimp alkaline phosphatase. PCR re-sequencing was done using the ABI Prism Big Dye cycle sequencing kit according to the manufacturer’s protocol and sequencing reactions were purified using a Sephadex column. Both the forward and reverse strand of each PCR product was sequenced using an ABI 3730 automated sequencer.

Raw sequence chromatograms were edited by eye using Sequencher version 4.10.1 (Gene Codes, Ann Arbor, MI) and assembled into contigs where each putative polymorphic site was scrutinized. Fragments were re-sequenced when traces did not provide at least 2X coverage over each SNP. Sequences were exported, concatenated (for each analysis) and formatted using custom PERL scripts. Alignments were done using Sequencher version 4.10.1. Some sequences in some lines repeatedly resulted in either weak PCR bands or ambiguous chromatograms—those sequences were dropped from the analysis.

Because data quality issues encountered in amplifying repetitive regions of the genome are expected to falsely increase diversity, and we find little diversity in our dataset, we do not expect that our data suffer from many errors at the PCR or sequencing level. Nonetheless, because of the repetitive nature of the Y chromosomes, every measure was taken to ensure that orthologous regions of the Y chromosome were amplified in each line: only a single band was amplified in males prior to sequencing, and there was no evidence of “heterozygous” sites in our traces,
indicating that our primers annealed uniquely in the *D. melanogaster* genome. Any traces with ambiguous chromatograms were removed from the dataset.

**Polymorphism analysis**

We estimated $\theta_w$ (population mutation rate per silent nucleotide site), $\theta_\pi$ (measure of nucleotide diversity per silent site), and Tajima’s $D$ (a summary of the frequency spectrum) using the program **compute** in the analysis package version 0.8.0 associated with the **libsequence** version 1.7.0 library (Thornton 2003). Because of the considerable population structure in *D. melanogaster*, summary statistics for each population were calculated separately.

**Recombination**

We estimated the minimum number of recombination events using the RecMin software ([http://www.stats.ox.ac.uk/~myers/RecMin/](http://www.stats.ox.ac.uk/~myers/RecMin/)) as $R_m$ (Hudson and Kaplan 1985) and $R_h$ (Myers and Griffiths 2003).

**X-linked polymorphism**

All X-linked data were obtained from the literature. Briefly, Pool and Aquadro (2006) surveyed 10 lines of flies from a China and Uganda population at 4 non-coding X-linked loci at least 10 kb from coding loci in regions of high recombination (Table S1). Haddrill *et al.* (2005) surveyed the introns of 10 X-linked genes across regions of the X chromosome with high recombination in Zimbabwe, Netherlands and Pennsylvania populations (Table S1). Kolaczkowski *et al.* (2011) surveyed variation across the genome of 19 individuals of a Tasmanian population of *D. melanogaster* using Illumina technology. For the Tasmania population, we used the estimate of pairwise diversity from Kolaczkowski *et al.* (2011; Table S1). Our Y-linked data come from these same populations and in most cases, the same lines.

**Coalescent methods for tests of neutrality**
All coalescent simulations were performed with custom C++ programs using the
libsequence version 1.7.0 library (Thornton 2003). To determine the significance of the
reduction of \( \pi_Y \), or \( \Delta \pi \) (the reduction in variation on the Y, estimated as \( [\pi_X - \pi_Y]/\pi_X \)), we simulated
1x10^5 X-linked and Y-linked genealogies under the infinite sites model making the assumption
that \( \theta_Y = \theta_X/3 \) (i.e. we assume that there is one Y chromosome for every three X chromosomes in a
population). We used the average \( \theta_X \) estimated from Haddrill et al. (2005) for the following
populations: Zimbabwe (0.01327); Netherlands (0.00414); and Pennsylvania (0.00557). For
Uganda and Beijing, we used estimates of \( \theta_X \) from Pool and Aquadro (2006) 0.012 and 0.0036,
respectively; Table S1). Because comparable X-linked data (PCR re-sequence data from non-
coding X-linked loci) are not currently available for Tasmania, we assumed a uniform range of \( \theta_X \)
values that encompass the Cosmopolitan samples that we have considered \( \theta_X \sim U(0.0020-0.0045) \).
We generated P-values using the empirical cumulative distribution function (ecdf) in R. The
false discovery rate (FDR) was estimated using the p.adjust package in R (Benjamini 1995).

**Demographic models**

**Population Bottlenecks**

We used models of simple population bottlenecks to determine whether the Y
chromosome polymorphism and frequency spectra from the Netherlands and Beijing populations
could be explained by past demographic changes. The model describes a population that drops in
size at time \( t_b \) (\( t_b = t_r + d \) where \( t_r \) is the time to recovery from the bottleneck and \( d \) is the duration of
the bottleneck) in the past according to severity parameter \( f (f = N_b/N_0 \) where \( N_0 \) is the \( N_e \) after the
bottleneck moving forward in time, or the current population size) and recovers from the
bottleneck at time \( t_r \) to \( N_0/N_A \), where \( N_A \) is the ancestral population size before the bottleneck
(Figure 3, Figure S2). We also considered models where the bottlenecked population came from
an ancestral population that was expanding in either a stepwise, or exponential manner (Figure S2). The parameter $t_{\text{grow}}$ describes when the ancestral population began expanding in the past, and was estimated in Li and Stephan (2006) using Zimbabwe data. Neutral mutations were placed on the tree according to a Poisson distribution with mean $\theta/2$, where $\theta$ is the neutral mutation rate. For models incorporating expansion in the ancestral population, we used an average $X$-linked $\theta$ estimated by Li and Stephan (2006) of 0.0499. For models that just consider bottlenecks from a stable ancestral African population, we drew $\theta$ for each $X$-linked locus from a uniform distribution bounded by the 95% confidence intervals of $\theta_w$ obtained in Haddrill et al. [2005; $\theta \sim \text{U}(0.01,0.015)]$ estimated from the $X$-linked Zimbabwe population. In all models, we make the assumption that $\theta_Y = \theta_X / 3$, $\rho_Y/\theta = 10$ (Haddrill et al. 2005; Thornton and Andolfatto 2006) and that $\rho_Y = 0$. We performed simulations under six population bottleneck models consistent with parameter values estimated in the literature for Beijing and Netherlands populations (Table S3). The six models can be grouped into two types for each population: the B1 models that consider an older, shorter and more severe bottlenecked European population ($t_r = 0.048$, $d = 0.001$, $f = 0.002$, $N_0/N_A = 0.125$; Table S3; e.g. Li and Stephan 2006) and B2 models that consider a longer bottleneck ($t_r = 0.004$, $d = 0.018$, $f = 0.029$, $N_0/N_A = 1.0$; Table S3; e.g. Thornton and Andolfatto 2006). Each of the two model types then had different scenarios: the founding population came from an ancestral African population that was expanding exponentially, according to a stepwise function, or was of stable size. The beginning of the African expansion was assumed to be about 60,000 years ago (Li and Stephan 2006), and growth rates and times were scaled to match $N_0$ for each population (Table S3). All Beijing population models include a second, recent bottleneck from an ancestral European population using parameters comparable to those estimated by Laurent et al. (2011; B1: $t_{r2} = 0.019$, $d_2 = 0.001$, $f_2 = 0.02$, $N_0/N_{A2} = 0.254$; B2: $t_{r2} = 0.002$, $d_2 = 0.0001$, ...
$f_2 = 0.02$, $N_0/N_A = 1.0$). All time is in scaled units of $4N_e$ generations and was scaled appropriately for the Y chromosome in the simulations ($t_Y = 3t_X$). $N_0$ corresponds to estimates from the literature under the different demographic models (Netherlands B1: $1.075 \times 10^6$; L1 and Stephan 2006; B2: $2.4 \times 10^6$; Thornton and Andolfatto 2006; Beijing B1: $4.14 \times 10^5$; L1 and Stephan 2006; B2: $2.4 \times 10^6$). Each demographic scenario was simulated 10,000 times and two-sided $P$-values for the X- and Y-linked empirical summary statistics were generated using the ecdf function in R. $P$-values were adjusted for false positive rates using a Benjamini-Hochberg FDR rate and the $Q$ values are reported in Tables 3 and S4.

A demographic model for the Pennsylvania population was assigned assuming that North American populations were founded from an ancestral European population a few hundred years ago (B1: $t_{r2} = 0.0006$, $d_2 = 0.0003$, $f_2 = 0.001$, $N_0/N_A = 1.0$; B2: $t_{r2} = 0.0003$, $d_2 = 0.0001$, $f_2 = 0.001$, $N_0/N_A = 1.0$; Table S3). $N_0$ has not been specifically estimated for the Pennsylvania population therefore we assumed a simple bottleneck (where the population recovers to the pre-bottleneck size at $t_{r2}$) for both types of models (Pennsylvania B1: $1.075 \times 10^6$; B2: $2.4 \times 10^6$). It is important to note that these parameter values have not been inferred using X-linked or autosomal data, however the results do not change significantly when the timing of this bottleneck is moved (data not shown). Because North American populations have higher nucleotide diversity than European populations (Haddrill et al. 2005; Caracristi and Schlötterer 2003), North American populations may have a more complicated demographic history than a simple bottleneck from a European ancestor. Caracristi and Schlötterer (2003) hypothesized that East Coast North American populations are admixed with African alleles. To account for this possibility, we also simulated the Pennsylvania population under bottleneck models with varying degrees of African admixture (0%, 5%, 10%, 15% and 20%). Two scenarios are possible: the European ancestor of North American populations was admixed with an African population, or the African alleles
entered the Pennsylvania population after the bottleneck, perhaps because of admixture with Caribbean populations (CARACRISTI and SCHLÖTTERER 2003). Because West Coast populations appear more similar to European populations than East Coast populations, and because the North America-African genetic distance is smaller than European-African genetic distance (CARACRISTI and SCHLÖTTERER 2003), it seems more likely that admixture occurred after the North American bottleneck. Nonetheless, we simulated under both admixture scenarios: African–European admixture (AF-EU) and African-Pennsylvanian admixture (AF-PA), where admixture occurred at two fixed times corresponding to 300 years ago in the AF-EU model and 125 years ago for the AF-PA model (see Figure S3 and Table S6). These simulations were performed in ms (HUDSON 2002) 10,000 times each for 10 X-linked and a single Y-linked locus, with the following assumptions: $\theta_Y = \theta_X/3$, $\rho_Y = 0$ and $\rho_Y/\theta = 10$. Our conclusions for the Pennsylvania population do not change with the inclusion of admixture in the Pennsylvania bottleneck models (Tables 4 and S6). The similarity between Tasmanian and Pennsylvania Y chromosomes indicates that they may originate from a similar European population, therefore we also simulated Tasmanian populations with admixture in the event that the European ancestor had been admixed with an African population (Table S6).

Although the demographic history may differ between Uganda and Zimbabwe populations, we used the Zimbabwe demographic model as an approximation for the Uganda population. Likewise, we used demographic models for Pennsylvania, Netherlands and Beijing populations as approximations for the Tasmania population and chose the most appropriate demographic model (i.e. Pennsylvania, see Table 4; Figure 3; Figure S2) for future simulations. Again, each demographic scenario was simulated 10,000 times and two-sided $P$-values for the X- and Y-linked empirical summary statistics were generated using the ecdf function in R.
were adjusted for false positive rates using a Benjamini-Hochberg FDR rate and the $Q$-values are reported in Tables 3 and 4.

We repeated the neutral coalescent rejection sampling procedure described above except this time incorporating the demographic history described by the “best fitting” model (the one that could describe the most aspects of the data) shown in bold face in Tables 3 and 4 and Tables S3 and S4. The false discovery rate (FDR) was estimated using the p.adjust package in R (Benjamini 1995). $Q$-values < 0.05 (corresponding to an FDR of 5%) are considered significant.

*Population Expansion*

We used a simple model of population expansion to consider the possibility that a history of population growth in Africa could explain the reduction in diversity and skewed frequency spectrum on the Y chromosome in the African lines. The assumptions of the models were the same as described for the population bottleneck models described above (i.e. $\rho_Y/\theta = 10$, $\rho_Y = 0$, $\theta_Y = \theta_X/3$). We ran one model with exponential growth and one with a simple stepwise growth (Figure 3; Figure S2). The model with exponential growth describes a population that began expanding at time $t_{grow}$ in the past at rate $\lambda$ until the present time $t_{rgrow}$. We chose parameter values consistent with those estimated in Li and Stephan (Li and Stephan 2006) for the Zimbabwe population: $\lambda \sim U(15,25)$, $t_{rgrow} = 0$ (assuming $N_{e0} = 2.4 \times 10^6$). The model with stepwise population growth describes a population that expanded to size $N_0/N_A$ at time $t_{grow}$ in the past. Again, we chose parameter values consistent with those estimated in Li and Stephan (Li and Stephan 2006) for the Zimbabwe population: $N_0/N_A \sim U(2.0,11.5)$, $t_{grow} = 0.0833$, $t_{rgrow} = 0$. All time is in scaled units of $4N_e0$ generations and was adjusted for the Y chromosome in the simulations ($t_{rgrow} = 0.2499$). Each demographic scenario was simulated 10,000 times and two-sided $P$-values for the X- and Y-linked empirical summary statistics were generated using the ecdf function in R. We
used these same models to obtain \( P \)-values for the Uganda population, which also shows evidence for population expansion although the true parameter values may differ between Zimbabwe and Uganda. The false discovery rate (FDR) was estimated using the \texttt{p.adjust} package in R (Benjamini 1995).

Estimating the reduction in \( N_e \) on the Y chromosome

To identify the minimum reduction in \( N_e_Y \) required to produce the \( \Delta \pi \) observed, we used an Approximate Bayesian Computation (ABC) approach with rejection sampling conditional on \( \Delta \pi \) and \( \pi_Y \) (Przeworski 2003; Thornton and Andolfatto 2006). We estimate the reduction in effective population size on the Y chromosome compared to the X chromosome (\( N_e_Y/N_e_X \)) as the posterior probability of the scaling factor of \( \theta \), \( k \). We assumed that population recombination rates follow \( \rho_X/\theta = 10 \) and \( \rho_Y = 0 \). The observed data in this case are the Y chromosome polymorphism data collected in this study, the X-linked polymorphism data compiled from the literature for the Zimbabwe (10 loci), Netherlands (10 loci), Pennsylvania (10 loci), Uganda (4 loci) and Beijing populations (4 loci). We assumed that the Tasmanian and Pennsylvanian populations have similar demographic histories because the Americas and Australia were founded around the same time (David and Capy 1988).

The rejection sampling algorithm is as follows:

1. Draw \( \theta_X, \rho \) and \( k \) from prior distributions, where \( \theta \) is the X-linked population mutation rate, \( \rho \) is the X-linked population recombination rate and \( k \) is the scaling factor for \( \theta_Y \) (\( k \) represents \( N_e_Y/N_e_X \)).

2. Estimate \( \theta_Y \) as \( k \theta_X \).
3. Simulate genealogies for $nsam_X$ independent X-linked loci and one Y-linked locus under the neutral coalescent model based on the empirical sample size for each locus where $nsam_X$ is the number of loci in the empirical dataset.

4. Continue if $|\theta_X^{\text{observed}} - \theta_X^{\text{simulated}}| \leq \varepsilon$.

5. Calculate summary statistics for the simulated genealogies according to $\theta_X$ or $\theta_Y$, assuming an infinite sites mutation model.

6. Accept or reject the chosen parameter values conditional on $|\Delta\pi^{\text{observed}} - \Delta\pi^{\text{simulated}}| \leq \varepsilon$, and $|\pi_Y^{\text{observed}} - \pi_Y^{\text{simulated}}| \leq \varepsilon$, and record accepted parameter values.

7. Return to step 1 and continue simulations until 1,000 samples from the joint posterior probability distribution are collected.

The prior distributions used were: $\theta_Y \sim U(0.01, 0.15)$ and $k \sim U(1 \times 10^{-7}, 1)$. The tolerance parameters, $\varepsilon$ and $\varepsilon_\pi$, were set to 80% and 2% of the observed $\theta$, and $\Delta\pi$ or $\pi$, respectively.

The same simulations were performed under a standard neutral model. For the Pennsylvania population admixture models, we used the same rejection sampling scheme, except it was performed using custom Perl scripts and $ms$ (Hudson 2002).

**Positive selection**

We estimated the time to the most recent hard selective sweep in a coalescent framework using Approximate Bayesian Computation with rejection sampling conditional on multiple summaries of each Y-linked dataset. Because there is no crossing over on the $D. melanogaster$ Y chromosome, selective sweeps were modeled as absolute bottlenecks, where at time $t_{\text{sweep}}$, all remaining lineages coalesced. We modeled Zimbabwe and Uganda populations separately. We incorporated population demographic histories into the selective sweep simulations so that the X-
linked and Y-linked loci coalesce in the simulations according to the appropriate demographic model until the sweep, at which time all remaining Y-linked lineages coalesced.

The rejection sampling algorithm is as follows:

1. Draw $\theta_X$ and $t_{\text{sweep}}$ from prior distributions, where $\theta_X$ is the X-linked population mutation rate, $\rho$ is the X-linked population recombination rate and $k$ is the scaling factor for $\theta_Y$ (where $\theta_Y$ is the population mutation rate for the Y chromosome and $k$ represents $Ne_Y/Ne_X$).

2. Estimate $\theta_Y$ as $k\theta_X$.

3. Simulate genealogies for $nsam_X$ independent X-linked loci and one Y-linked locus under the neutral coalescent model based on the empirical sample size for each locus where $nsam_X$ is the number of loci in the empirical dataset.

4. Continue if $|\theta_{\text{observed}} - \theta_{\text{simulated}}| \leq \varepsilon$, for both the X- and Y-linked loci.

5. Calculate summary statistics for the simulated genealogies according to $\theta_X$ or $\theta_Y$, assuming an infinite sites mutation model.

6. Accept or reject the chosen parameter values conditional on $|\Delta\pi_{\text{observed}} - \Delta\pi_{\text{simulated}}| \leq \varepsilon$, $|D_{\text{Taj-Y}_{\text{observed}}} - D_{\text{Taj-Y}_{\text{simulated}}}| \leq \varepsilon$, $|S_{1-Y_{\text{observed}}} - S_{1-Y_{\text{simulated}}}| \leq \varepsilon$ and $|\pi_{Y_{\text{observed}}} - \pi_{Y_{\text{simulated}}}| \leq \varepsilon$, and record accepted parameter values.

7. Return to step 1 and continue simulations until 10,000 samples from the joint posterior probability distribution are collected.

The prior distributions used were: $\theta_X \sim U(0.0005,0.03)$, and $t_{\text{sweep}} \sim U(0,10^{-6})$. The tolerance parameter, $\varepsilon$, was set to 35% of the observed $\Delta\pi$, $D_{\text{Taj}}$, $\pi_Y$ and $S_{1}$ (number of singletons).

RESULTS

Patterns of Polymorphism
When compared to the X chromosome, the Y chromosome harbors less variation than expected under the standard neutral model (Table 1). Contrary to patterns of variation on the X chromosome and autosomes, African populations have significantly less Y-linked variation than Cosmopolitan populations (Table 1; Figure 1; \( P = 0.002 \) considering indels and \( P = 0.008 \) without considering indels, MWU test). Specifically, there is a \(~400\)-fold reduction in pairwise nucleotide diversity on the Y chromosome compared to the X chromosome in Africa (414-fold for Zimbabwe and 383-fold for Uganda), whereas there is a six-fold (Tasmania) to 66-fold (Pennsylvania) reduction in pairwise nucleotide diversity outside of Africa. The estimated reduction in variation on the Y chromosome compared to the neutral expectation (\( \pi/\pi_0 \)) is \(~0.7\)% in African populations and varies between 4.6% and 19.5% outside of Africa (Table 1; Table S2). We scrutinized each segregating site in our dataset: most were sequenced twice on the forward strand and twice on the reverse strand and had independent confirmation with a second DNA isolation and re-sequencing. Interestingly, there is a significant bias in the frequency of indel polymorphism (usually a single base pair) in African compared to Cosmopolitan populations. Among African Y chromosome variants, 80% (4/5) of the segregating sites are indels whereas among Cosmopolitan Y chromosome variants only 8.3% (2/24) of the segregating sites are indels (Figure 2B; \( P = 0.003 \), FET). We consider the full dataset including indels for the remainder of the paper.

**Population structure**

Our analysis of population structure using the pairwise average number of nucleotide substitutions per site (\( D_{xy} \)), \( F_{st} \), \( K_{st}^* \) (measure of population differentiation analogous to \( F_{st} \)) and \( S_{nn} \) (a nearest neighbor statistic), yielded surprising results. The African populations do not cluster together: Zimbabwe instead clusters with the Netherlands population. Moreover, Pennsylvania and Tasmania Y chromosomes are nearly indistinguishable (Figure 2; Table 2). Consistent with several surveys of population differentiation showing considerable population
structure in Asia (BAUDRY et al. 2004; HALE and SINGH 1991; POOL and AQUADRO 2006; SCHLÖTTERER et al. 2006), the Beijing Y chromosome is highly differentiated from the rest of the populations. There is no evidence of sub-structuring within the Beijing population on the Y chromosome (Figure 2B). Outside of the clustering of Zimbabwe and Netherlands Y chromosomes, these results are consistent with the population structure of autosomal, X chromosomal and mitochondrial loci from these or similar populations (BAUDRY et al. 2004; HADDRILL et al. 2005; HALE and SINGH 1991; POOL and AQUADRO 2006; SCHLÖTTERER et al. 2006). Genomic sequences from non-Y-linked loci exclude the possibility that the similarity of Netherlands and Zimbabwe Y chromosomes are explained by contamination in our fly lines (A.G.C. unpublished).

Evidence for intrachromosomal recombination

We used the program RecMin to identify any possible recombination events in our dataset (MYERS and GRIFFITHS 2003). Three different methods ($R_\text{ms}$, $R_h$ and $R_s$) yield the inference that at least one recombination event has occurred in the sampled regions. This event was confirmed and other possible recombination events were identified by eye (Figure 2B). Although the intronic regions of the Y chromosome that we have amplified are repetitive, the PCR and sequence trace reads indicate that we have amplified single, orthologous regions of the Y chromosome in each fly line. Recombination via crossing over does not occur in most Drosophila species males, however gene conversion events have been detected on the $D. \text{simulans}$ Y chromosome between duplicate genes (KOPP et al. 2006). While the gene regions surveyed are not known to be duplicated in $D. \text{melanogaster}$, the introns are repetitive, and it is possible that there are several similar sequences throughout the Y chromosome offering opportunities for intrachromosomal recombination. These recombination events are likely to reflect mitotic recombination in the male germline.
Population demographic history

We find a reduction in variation on the Y chromosome compared to the X chromosome and this is statistically significant under a standard neutral model, assuming equal effective numbers of breeding males and females (Table 1). However, Cosmopolitan populations of *D. melanogaster* clearly violate the assumption of a constant population size: *D. melanogaster* originated in Africa and populated Europe approximately 16,000 years ago and American and Australian populations were founded from a European ancestor only in the last few hundred years (Caracristi and Schlötterer 2003; David and Capy 1988). Moreover, African populations may have a history of population expansion that began before the out-of-Africa migration events (60,000 years ago; Li and Stephan 2006).

We considered the effect of population demographic history on Y chromosome evolution by simulating data under several different models that have been inferred using empirical data from the Zimbabwe, Netherlands and Beijing populations (Laurent et al. 2011; Li and Stephan 2006; Thornton and Andolfatto 2006; Figure 3). We found that ancestral growth in the Zimbabwe population can account for values of Tajima’s *D* similar to what we observe on the Y chromosome, however the reduction in variation on the Y chromosome compared to the X is actually smaller under growth models than under the standard neutral model (Table 3; Table S4). A history of population expansion alone therefore cannot explain patterns of variation on the Y chromosome in Zimbabwe. We see similar results when we assume that the Uganda population experienced the same degree of population expansion (Table 4; Table S5).

All Cosmopolitan populations of *D. melanogaster* have a history of population bottlenecks. We considered several population bottleneck models inferred from empirical data in the literature and asked whether the summary statistics we observe on the Y chromosome could be explained by demographic history alone for the Netherlands and Beijing populations. Population
bottleneck models with different bottleneck parameters, with or without population expansion in the ancestral African population, predict different patterns of Y chromosome variation (summarized in Table 3; Table S4). For example, assuming a severe bottleneck that was fairly short in duration (i.e. B1 models; Table 3; Table S4), while the observed Tajima’s $D$ on the Y chromosome is expected under some of these models, rarely generates a reduction in variation (measured as $\Delta \pi$) as large as what is observed in the Netherlands population. In contrast, considering a strong bottleneck with an order of magnitude longer duration (i.e. B2 models; Table 3; Table S4), so that the population has only recently recovered, makes it likely that one would observe Tajima’s $D$ as low and $\Delta \pi$ of the magnitude seen in the Netherlands and Beijing populations. These results highlight the utility of considering multiple genomic locations in inferring demographic history. Our Y-linked data from the Netherlands and Beijing can thus be explained purely with demographic models without the need to invoke selection. Although a detailed demographic model has not been fitted to X-linked or autosomal data from the Pennsylvania population, we considered an arbitrary model where a recent bottleneck from a European population occurred in the past few hundred years ($t_b = 0.002$). This model can explain all major aspects of the Y chromosome dataset (Table 4; Table S5), however $D_{TajY}$ in this model just barely meets the 5% false discovery rate (FDR) cutoff ($P$-value=0.0086; $Q$-value=0.08). Although it is important to note that the timing of the North American bottleneck was not inferred from empirical data, because this bottleneck is so recent in history, the precise timing has little impact on our results (data not shown). All major aspects of the Tasmania population can be explained by demographic history when assuming that Tasmania experienced the same population bottlenecks as each of the Netherlands, Beijing and Pennsylvania populations (Table 4; Table S5). Because North American populations have more diversity than European populations and may be admixed with African alleles (CARACRISTI and SCHLÖTTERER 2003), we also explored two
admixture scenarios: 1) an admixed European population founded North American populations of *D. melanogaster* (model AF-EU; Figure S3A; Table S6) and 2) the Pennsylvania population admixed an African or African-like population after the North American bottleneck (Figure S3B; Table 6; Caracristi and Schlötterer 2003). We simulated the Pennsylvania population under each admixture scenario with 0%, 5%, 10%, 15%, and 20% admixture with African alleles at two fixed times corresponding to ~300 years ago for the AF-EU models and ~125 years ago for the AF-PA model. Because the Tasmania and Pennsylvania populations may have originated from a similar European ancestor, we also simulated under the same AF-EU admixture models for the Tasmania population. We get qualitatively similar results under all the admixture models when compared to models with no admixture: observed estimates of $D_{Taj}$ and $\Delta \pi$ are expected under each of these models for both the Pennsylvania and Tasmania populations at a 5% FDR (Table 4; Table S6). However, the empirical $\Delta \pi$ and $D_{Taj}$ (X- and Y-linked) in the present dataset appear more likely under models with a small degree of admixture between the Pennsylvania population and an African-like population (e.g. AF-PA $p_{AF}=0.05$; $P$-value $D_{TajX}=0.6520$; $Q$-value $D_{TajX}=0.7964$; $P$-value $D_{TajY}=0.8186$; $Q$-value $D_{TajY}=0.8396$; $P$-value $\Delta \pi=0.7072$; $Q$-value $\Delta \pi=0.8082$; Tables 4 and S6). While $\Delta \pi$ and $D_{TajX}$ are expected under the models with no admixture, a $D_{TajY}$ as low as we observe in Pennsylvania is not expected under the recent bottleneck model, but this does not survive the 5% FDR cutoff (AF-PA $p_{AF}=0$; $P$-value $D_{TajX}=0.5014$; $Q$-value $D_{TajX}=0.7161$; $P$-value $D_{TajY}=0.009$; $Q$-value $D_{TajY}=0.096$; $P$-value $\Delta \pi=0.3418$; $Q$-value $\Delta \pi=0.7933$; Table S6, also see the PA-B2 model in Table 4). Natural selection on the Y chromosome may be implicated when the X-linked data fit well to a plausible demographic model, but the Y data clearly reject the same model. We looked for any incongruence between the X- and Y-linked data in our simulations. Although a model with 20% admixture between the European ancestor of
North American populations and an African population produced summaries completely consistent with the X-linked data and less consistent with the Y-linked data, none of the summaries could be statistically rejected and survive a correction for the false discovery rate (AF-EU $p_{AF}=0.20$; $P$-value $D_{TajX}=0.4988$; $Q$-value $D_{TajX}=0.7161$; $P$-value $D_{TajY}=0.0114$; $Q$-value $D_{TajY}=0.0960$; $P$-value$_{Δπ}=0.0792$; $Q$-value$_{Δπ}=0.2192$; Tables 4 and S6).

**Reduction in $N_e$ on the Y chromosome**

We estimated the reduction in $N_e$ on the Y chromosome as $k (N_{eY}/N_{eX})$ using an Approximate Bayesian Computation approach. We ran simulations both under the best demographic model (bold face in Tables 3 and 4) and a standard neutral model. Historical changes in population size affect the ratio of Y to X chromosomes. Under neutrality the expected $k$ is 1/3, but without incorporating demographic history the estimated $k$ is an order of magnitude (Cosmopolitan) or two orders of magnitude (African) less than this (Table S7). When we account for the demographic history of each population, the maximum a posteriori (MAP) estimate of $k$ is not significantly different from 1/3 in any Cosmopolitan population but is significantly less than 1/3 in both African populations ($MAP\ k_{ZW}=0.0041, P = <10^{-3}$ ; $MAP\ k_{UG}=0.0037, P = <10^{-3}$ ; Figure 4; Table S7).

**Positive selection in Africa**

A history of population expansion in Africa can explain summaries of the frequency spectrum, but cannot explain the ~400-fold reduction in variation on the Zimbabwe and Uganda Y chromosomes compared to the X chromosomes. This drastic reduction in variation could, however, be explained by a history of recent selective sweeps. We simulated selective sweeps using coalescent simulations incorporating population expansion in Africa and assuming constant population size. To estimate the time since the most recent selective sweep in African populations,
we used an Approximate Bayesian Computation approach and conditioned on several aspects of the Y chromosome polymorphism data ($\Delta \pi$, $\pi_f$, $\theta_w$, $S_f$ and $D_{Taj}$). Under a population expansion model, we estimate the time to the most recent selective sweep to be similar in Zimbabwe (MAP $t_{sweep} = 0.00098$ [0.00042-0.00226] $4N_e$ generations; Figure 5) and Uganda (MAP $t_{sweep} = 0.00082$ [0.00033-0.00214] $4N_e$ generations; Figure 5) populations. Assuming that the $N_e$ of African populations is 2.4x10$^6$ (THORNTON and ANDOLFATTO 2006) and 10 generations per year, the last sweep on the Zimbabwe Y chromosome occurred 236 years ago ($0.25* (4N_e t/g)$; 101-542 years) and the last sweep on the Uganda Y chromosome occurred 197 years ago (80-515 years). Because population expansion reduces the difference in variation between the X and Y chromosomes, and because we condition on $\Delta \pi$, under a model of constant population size, the time of the sweep is pushed back for both populations (Zimbabwe MAP $t_{sweep} = 0.00388$ [0.00168-0.00956] $4N_e$ generations or 931 (404-2293) years; Uganda MAP $t_{sweep} = 0.003122$ [0.001271-0.00917] $4N_e$ generations or 749 (305-2226) years; Figure 5). The marginal posterior distribution on $t_{sweep}$ provides 95% credible intervals that overlap between ~400-550 years ago for Zimbabwe and ~300-515 years ago for Uganda.

**DISCUSSION**

While the gene-poor and heterochromatic *D. melanogaster* Y chromosome clearly harbors functional variation, it has very low levels of nucleotide polymorphism. The effect of the Y chromosome on male fitness (CHIPPINDALE and RICE 2001), heat tolerance (ROHMER et al. 2004) and global (LEMOS et al. 2008; LEMOS et al. 2010) or germline-specific (ZHANG et al. 2000) gene expression may be driven in part by variation in heterochromatin content (LEMOS et al. 2010), specific trans-activators of gene expression (ZHANG et al. 2000), or in rDNA copy number (PAREDES et al. 2011), rather than by variation at the single nucleotide level.

**Evidence for admixture between European and African populations?**
Surprisingly, the *D. melanogaster* Y chromosome harbors far less polymorphism than expected in African populations compared to Cosmopolitan populations. While overall the Y chromosome recapitulated the history of these populations based on autosomes, the X chromosome and mitochondria, we found one surprising relationship between Y chromosomes from African and European Y chromosomes: Zimbabwe Y chromosomes are more similar to Y chromosomes from the Netherlands than Uganda. There are two explanations for the similarity between Zimbabwe and Netherlands Y chromosomes: 1) The ancestral African population that founded the European populations could have been similar to Zimbabwe rather than Uganda. However, a shared site between the Uganda and Beijing Y chromosomes is discordant with this interpretation. Moreover, Pool and Aquadro (2006) hypothesized that the ancestral source population for the Out-of-Africa migration to Europe was similar to a Uganda population based on X-linked variation data. 2) The similarity of the Y chromosomes from Zimbabwe and Netherlands may be evidence of admixture between these populations. Several studies have documented recent admixture between Cosmopolitan populations and populations from Congo (Capy et al. 2000) and Zimbabwe (Kauer et al. 2003), Eritrea, Gabon and South Africa (Pool and Aquadro 2006) and between Caribbean/South American populations and West African populations (Caracristi and Schlötterer 2003). The present dataset has too few informative polymorphisms to distinguish between these possibilities. An ancestral European population colonized the Americas and Australia in the last few hundred years (Caracristi and Schlötterer 2003; David and Capy 1988). The similarity between the Tasmanian and Pennsylvanian Y chromosomes suggests that these populations originated from the same, or a similar ancestral population in Europe.

**Recent selection on African Y chromosomes**

Variation is reduced a staggering 400-fold on the Y chromosome compared to the X chromosome in our African population samples, whereas fold reduction varies between 6 and 66-
fold outside of Africa. This is in stark contrast with variation across the autosomes, the X chromosome and mitochondrial loci, which all harbor more variation in African than outside of Africa. While a demographic history of population expansion in Africa can explain some aspects Y chromosome frequency spectra in Zimbabwe and Uganda, it cannot explain the drastic reduction in variation on the Y chromosome. There are three possible explanations for the dearth of Y-linked variation on African Y chromosomes: 1) A higher variance in male reproductive success in Africa relative to outside of Africa reduces the effective population size of the Y chromosome and can lead to lower levels of variability (Charlesworth 2001; Nunney 1993). Indeed, Hutter et al. (2007) found evidence for skewed sex ratios in African populations of D. melanogaster. While this is likely to make some contribution to this pattern, the maximum reduction in $N_e$ on the Y compared to the X resulting from sexual selection is ~9-fold (Caballero 1995; Charlesworth 2001) whereas the estimated reduction in $N_e$ on the Y compared to the X in Africa is 256-fold ($k =0.0039$) and 370-fold ($k =0.0027$) in Zimbabwe and Uganda, respectively (Figure 1; Figure 4; Table S7). A higher variance in male reproductive success in Africa is therefore not likely to be sufficient to explain this pattern. 2) Selection against deleterious mutations could reduce levels of variation on the Y chromosome relative to the X chromosome, but rarely to the degree observed in African populations (Kaiser and Charlesworth 2009). The relatively small number of genes on the Y chromosome offers few targets of purifying selection: we estimate that there are ~19kb of non-synonymous nucleotide sites on the D. melanogaster Y chromosome. For a target of this size in a non-recombining region of the genome, we expect to see a reduction in variation an order of magnitude smaller (Kaiser and Charlesworth 2009) than we observe if background selection were the only force shaping patterns of polymorphism on the Y chromosome. Moreover, the strength of background selection is not expected to vary among these populations and therefore background selection is unlikely to
explain differences in the degree of reduced diversity on the Y chromosome between populations.

3) African Y chromosomes may have experienced selective sweeps in the recent past. We estimate that the time to the most recent selective sweep in Zimbabwe and Uganda is similar—in the last ~200-500 years (Figure 5), assuming an $N_e$ of $2.4 \times 10^6$ (Thornton and Andolfatto 2006). It is unlikely that a single Y chromosome has swept across Africa in the recent past because there are a number of fixed differences between Zimbabwe and Uganda Y chromosomes (Figure 2B). Instead, it seems more likely that the Y chromosomes in Africa experienced local adaptation and thus independent selective sweeps in the recent history of each population. X-linked polymorphism data shows some structure among sub-Saharan African populations, suggesting that there may be reduced gene flow between Uganda and Zimbabwe populations (Pool and Aquadro 2006). It is possible that a European Y chromosome swept through the Zimbabwe population as a result of recent admixture.

What could be driving local Y-linked adaptations? The presence of X-linked meiotic drivers inside Africa could cause the replacement of Y chromosomes and shape patterns of local adaptation. Although the frequency of X-linked male meiotic drive is unknown in D. melanogaster (see Hurst 1996; Reed et al. 2005; Hanks 1968 for possible evidence of such drive in this species), cryptic X-linked meiotic drive is common in other Drosophila species (Gershenson 1928; James and Jaenike 1990; Montchamp-Moreau et al. 2006; Orr and Irving 2005; Tao et al. 2007). Another possibility is that effects of Y chromosome polymorphism on gene expression (Lemos et al. 2008; Lemos et al. 2010) may influence the adaptive evolution of the African Y chromosome. African and Cosmopolitan populations differ in patterns of gene expression (Hutter et al. 2008; Meiklejohn et al. 2003). The D. melanogaster Y chromosome affects the expression of genes across the genome (Lemos et al. 2008; Lemos et al. 2010), and a subset of these genes are similar to genes that differ in expression between
populations (Hutter et al. 2008). For example, local Y-linked adaptation in Africa could be a result of the Y chromosome’s effect on thermotolerance (Rohmer et al. 2004) or the Y chromosome may have an effect on wing flight muscle architecture, for which there is expression divergence between African and Cosmopolitan populations (Hutter et al. 2008).

**Patterns of variation in Cosmopolitan Y chromosomes is consistent with a recent, long bottleneck**

The frequency spectra and the reduction in variation on the Y versus the X chromosome is more consistent with a history of recent, long population bottlenecks (B2) rather than shorter and more severe bottlenecks (B1). Pool and Nielsen (2007) showed that different demographic scenarios can produce sharply different patterns of variation on the X chromosome compared to the autosomes and that their inference can be extended to effectively haploid regions of the genome such as the Y chromosome and mitochondria. Our results largely agree with their simulations: strong/severe bottlenecks with a longer duration can widen the difference between the X and the autosomes, or the Y and the X chromosomes. Our results also agree with the contrasting scenario where population expansion can cause smaller differences in X and autosomal, or Y and X chromosomal variability (Pool and Nielsen 2007).

While natural selection is not necessarily required to explain the reduction of diversity and skew in the frequency spectra of the Cosmopolitan Y chromosome, the data do not exclude the possibility that the *D. melanogaster* Y chromosome may have an adaptive history. First, several differences exist between Pennsylvanian/Tasmanian Y chromosomes and the Netherlands Y chromosomes. Because there is little population structure in Europe, this suggests that the Netherlands Y chromosome may have experienced recent selection. This Y chromosome may have experienced positive selection in the European ancestor prior to sweeping through the Zimbabwe population. Moreover, the reduction in variation on the Pennsylvania Y chromosome
compared to the X chromosome is less likely if there is a moderate amount of admixture between Pennsylvania and an African-like population (20% African alleles), although the observed reduction in variation ($\Delta \pi$) was not statistically significant after controlling for the false positive rate (Tables 4; S6). Therefore, the combination of a lack of recombination and recent population bottlenecks may compromise our ability to detect positive selection on Cosmopolitan Y chromosomes. Additionally, if shorter bottlenecks (similar to the B1 models described here) are truly more representative of the demographic history of these populations, then the observed magnitude of reduced variation on the Y chromosome is not expected under neutrality and may be better explained by a history of selection. A more complete understanding of the demographic history of Cosmopolitan populations is necessary to tease apart adaptive and demographic forces shaping the Y chromosome.

**ACKNOWLEDGEMENTS**

The authors are grateful for the generous donation of flies by C.F. Aquadro, W. Ballard, J. Pool, Z. Bochdanovits, and A. Hoffman. The authors thank D. Bachtrog for help early on in the project and K. Thornton for help with modeling using libsequence, and anonymous reviewers and David Begun for suggestions that improved the manuscript. This work was supported by an NIH grant R01 GM64590 to A.G.C. and A.B. Carvalho.
LITERATURE CITED


James, A. C., and J. Jaenike, 1990 "Sex ratio" meiotic drive in *Drosophila testacea*. Genetics **126**: 651-656.


TABLE 1. Summary statistics of Y chromosome variation in each population

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>m</th>
<th>S</th>
<th>S₁</th>
<th>π</th>
<th>θₜ</th>
<th>Δπ</th>
<th>π/π₀</th>
<th>Dₜaj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zimbabwe</td>
<td>24</td>
<td>7729</td>
<td>3</td>
<td>3</td>
<td>0.00323</td>
<td>0.01039</td>
<td>0.9976***</td>
<td>0.0074</td>
<td>-1.733*</td>
</tr>
<tr>
<td>Uganda</td>
<td>20</td>
<td>7523</td>
<td>2</td>
<td>2</td>
<td>0.00266</td>
<td>0.0075</td>
<td>0.9978***</td>
<td>0.0067</td>
<td>-1.513*</td>
</tr>
<tr>
<td>Beijing</td>
<td>17</td>
<td>7618</td>
<td>7</td>
<td>4</td>
<td>0.01696</td>
<td>0.02772</td>
<td>0.9529**</td>
<td>0.1428</td>
<td>-1.368</td>
</tr>
<tr>
<td>Tasmania</td>
<td>19</td>
<td>7544</td>
<td>7</td>
<td>4</td>
<td>0.01370</td>
<td>0.02667</td>
<td>0.9355**</td>
<td>0.1951</td>
<td>-1.631*</td>
</tr>
<tr>
<td>Netherlands</td>
<td>19</td>
<td>7519</td>
<td>6</td>
<td>5</td>
<td>0.01505</td>
<td>0.02425</td>
<td>0.9637***</td>
<td>0.1103</td>
<td>-1.305</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>24</td>
<td>7172</td>
<td>6</td>
<td>6</td>
<td>0.00849</td>
<td>0.02365</td>
<td>0.9848***</td>
<td>0.0460</td>
<td>-2.048*</td>
</tr>
</tbody>
</table>

n Number of lines surveyed within each population  
m Number of aligned bases considered in the summary statistic calculations  
S Number of segregating sites, including indels  
S₁ Number of singletons, or polymorphisms only found once in the sample  
π Average percent pairwise nucleotide diversity per 100 sites  
θₜ Average percent nucleotide diversity per site  
Δπ is calculated as (πₓ-π₁)/πₓ. Assuming k is 1/3 (equal numbers of breeding males and females), the neutral expectation of Δπ is ~0.667.  
π/π₀ is the reduction in π compared to the neutral expectation of π of the Y chromosome based on coalescent simulations of 10⁵ neutral genealogies  
Dₜaj Tajima’s D statistic to summarize the frequency spectrum  
*P < 0.05; **P ~ 1x10⁻³; ***P ~ 1x10⁻⁴; **
TABLE 2. Genetic differentiation between populations.

<table>
<thead>
<tr>
<th></th>
<th>Pennsylvania</th>
<th>Beijing</th>
<th>Netherlands</th>
<th>Tasmania</th>
<th>Uganda</th>
<th>Zimbabwe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pennsylvania</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beijing</td>
<td>0.7286</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>0.4213</td>
<td>0.5122</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tasmania</td>
<td>0.0088</td>
<td>0.5567</td>
<td>0.3088</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uganda</td>
<td>0.2863</td>
<td>0.4393</td>
<td>0.1389</td>
<td>0.2157</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>0.4055</td>
<td>0.4504</td>
<td>0.0769</td>
<td>0.3106</td>
<td>0.1250</td>
<td>-</td>
</tr>
</tbody>
</table>

Average pairwise \( F_{st} \) estimates between populations are shown. 
** \( P \sim 1 \times 10^{-5} \). Significance of \( \Delta \pi \) and \( D_{Taj} \) is determined from neutral coalescent simulations of X and Y-linked loci under the assumption that \( N_{eY}/N_{eX} \) is 1/3. \( P \) values are based on an empirical cumulative distribution function of \( \Delta \pi \) or \( D_{Taj} \) simulated values.
TABLE 3. Performance of different inferred demographic models for the Netherlands, Beijing and Zimbabwe populations

<table>
<thead>
<tr>
<th>DEMOGRAPHIC MODELS&lt;sup&gt;a&lt;/sup&gt;</th>
<th>MEDIAN $D_{TAJ} (Q)$&lt;sup&gt;b,c&lt;/sup&gt;</th>
<th>$\Delta \pi (Q)$&lt;sup&gt;b,c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X$</td>
<td>$Y$</td>
</tr>
<tr>
<td>Netherlands</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>-0.4642 (0.9257)</td>
<td>-1.4102 (0.9007)</td>
</tr>
<tr>
<td>B2</td>
<td>0.5209 (0.6966)</td>
<td>-1.5108 (0.9007)</td>
</tr>
<tr>
<td>B1-AF</td>
<td>0.349 (0.5778)</td>
<td>-0.1418 (0.4205)</td>
</tr>
<tr>
<td>B2-AF</td>
<td>0.7600 (0.5964)</td>
<td>-1.7177 (0.6966)</td>
</tr>
<tr>
<td>Beijing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>-0.2792 (0.7961)</td>
<td>-0.5183 (0.1413)</td>
</tr>
<tr>
<td>B2</td>
<td>0.4050 (0.9587)</td>
<td>-1.2509 (0.5925)</td>
</tr>
<tr>
<td>B1-AF</td>
<td>-0.2988 (0.5925)</td>
<td>-0.5550 (0.1208)</td>
</tr>
<tr>
<td>B2-AF</td>
<td>0.5572 (0.8599)</td>
<td>-1.4466 (0.7961)</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp</td>
<td>-0.7658 (0.4471)</td>
<td>-0.8480 (0.0812)</td>
</tr>
</tbody>
</table>

The X and Y-linked data from each population were tested under the different types of demographic model (bottlenecks, B1 and B2; bottlenecks with ancestral growth, B1-AF and B2-AF; Exponential growth, Exp).

<sup>a</sup>Results from simulated genealogies for 10 independent X-linked loci, and a single Y-linked locus (assuming $\theta_Y=1/3 \theta_X$) under bottleneck or exponential growth models specified in the Methods and Table S2.

<sup>b</sup>Median $D_{TAJ}$ (for X and Y-linked loci) and $\Delta \pi$

<sup>c</sup>Reported are $Q$-values estimated from two-sided $P$-values (see Table S4 for $P$-values). $Q < 0.05$ corresponds to a FDR <5%. The model chosen for use in subsequent simulations is in bold print.
TABLE 4. Performance of demographic models for Pennsylvania, Tasmania and Uganda populations

<table>
<thead>
<tr>
<th>DEMOGRAPHIC MODELS^a</th>
<th>MEDIAN $D_{Taj} (Q)$^b,c</th>
<th>$\Delta \pi (Q)$^b,c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>0.6559 (0.5138)</td>
<td>-1.159 (0.1605)</td>
</tr>
<tr>
<td>B2</td>
<td>0.8364 (0.5138)</td>
<td>0.1386 (0.0803)</td>
</tr>
<tr>
<td>B1-AF</td>
<td>0.7634 (0.5138)</td>
<td>-1.7325 (0.5965)</td>
</tr>
<tr>
<td>B2-AF</td>
<td>1.031 (0.5138)</td>
<td>0.1390 (0.1422)</td>
</tr>
<tr>
<td>AF-EU $\rho_{AF}=0.05$</td>
<td>0.8321 (0.7161)</td>
<td>0.1143 (0.0960)</td>
</tr>
<tr>
<td>AF-EU $\rho_{AF}=0.20$</td>
<td>0.9673 (0.7161)</td>
<td>0.1334 (0.0960)</td>
</tr>
<tr>
<td>AF-PA $\rho_{AF}=0.05$</td>
<td>-1.5996 (0.7964)</td>
<td>-1.5455 (0.8396)</td>
</tr>
<tr>
<td>AF-PA $\rho_{AF}=0.20$</td>
<td>-0.6201 (0.9638)</td>
<td>-1.0113 (0.6785)</td>
</tr>
<tr>
<td>Tasmania</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA-B2</td>
<td>-</td>
<td>0.1386 (0.8400)</td>
</tr>
<tr>
<td>BEI-B2</td>
<td>-</td>
<td>-1.2509 (0.9538)</td>
</tr>
<tr>
<td>NE-B2</td>
<td>-</td>
<td>-1.5108 (0.9538)</td>
</tr>
<tr>
<td>Uganda</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZW-Exp</td>
<td>-0.7658 (0.9350)</td>
<td>-0.8480 (0.2117)</td>
</tr>
</tbody>
</table>

The X and Y-linked data from each population were tested under the different types of demographic model (bottlenecks, B1 and B2; bottlenecks with ancestral growth, B1-AF and B2-AF; admixture models AF-EU and AF-PA, Exponential growth, Exp). The Tasmania and Uganda populations were tested with models based on other populations (PA- Pennsylvania; BEI- Beijing; NE- Netherlands; ZW- Zimbabwe).

^aResults from simulated genealogies for 10 independent X-linked loci, and a single Y-linked locus (assuming $\theta_\ell=1/3\theta_Y$) under hypothetical bottleneck or exponential growth models specified in the Methods and Table S2. Three possible demographic models corresponding to the population histories of Pennsylvania, Beijing and Netherlands populations were simulated for Tasmania.
Median $D_{Taj}$ (for X and Y-linked loci) and $\Delta \pi$. An empirical estimate of $D_{Taj}$ for the X chromosome in Tasmania has not been published.

Reported are $Q$-values estimated from two-sided $P$-values. $Q < 0.05$ corresponds to a FDR <5%. The model chosen for use in subsequent simulations is in bold print.
FIGURE LEGENDS

FIGURE 1. Levels of nucleotide variation by population. Standard boxplots showing Watterson’s estimator of nucleotide diversity per site from the number of segregating sites ($\theta_w$) and the pairwise nucleotide diversity per site ($\theta$) for each Y-linked fragment sequenced in the four Non-African and two African populations. The horizontal lines interrupting the boxes indicate the median, the whiskers extend to the most extreme data point no more than 1.5 times the interquartile range from the box and open circles indicate outliers. Estimates were calculated both including and excluding insertion-deletion mutations (indels) and only the analysis including indels is considered for the rest of the analysis.

FIGURE 2. Population structure. A. Nucleotide Divergence between populations. An unrooted population distance tree generated by a neighbor-joining analysis of the average number of nucleotide substitutions per site between populations ($D_{xy}$) is diagrammed. B. Y chromosome polymorphism table. The sequenced fragments of $kl$-5, $kl$-3, $kl$-2 and $ORY$ are separated by black lines. The populations are separated by white space.
**FIGURE 3. Simple demographic models.** Diagrammed are schematics describing the models of simple population bottlenecks and population expansion and their parameters. For choice of the best demographic model, and parameter values, see Tables 3 and 4, and S1. **A.** The Netherlands population bottleneck describes a population that was reduced in size to \( f_1 \cdot N_0 \) at time \( t_{b1} \) and recovering to size \( N_0 / N_{A1} \) at time \( t_{r1} \). **B.** The Beijing population bottlenecked twice, where the first was the European bottleneck from an ancestral African population (same as Netherlands bottleneck). **C.** The Pennsylvania population experienced a similar bottleneck as did Beijing, except that the second bottleneck occurred more recently. **D.** The demographic model for the Zimbabwe population describes an ancestral African population that began expanding at time \( t_{grow} \) at rate \( G \) until the present time. All time is in scaled units of \( 4N_e \) generations. We used the Pennsylvania double bottleneck (C) for the Tasmania population and the Zimbabwe expansion model (D) for the Uganda populations. Not shown in the diagram are admixture events (see text and Figure S3 for bottleneck with admixture models).

**FIGURE 4. The estimated reduction in effective population size of the Y chromosome compared to the X chromosome.** The marginal posterior distributions of \( k \) \( (N_{eY}/N_{eX}) \) under the appropriate demographic model (in bold in Tables 3 and 4) for each population are plotted. The dotted line represents the neutral expectation of \( k \) assuming equal numbers of breeding males and females (1/3). Zimbabwe and Uganda populations have a \( k \) significantly lower than 1/3. \( P \)-values were calculated from the empirical cumulative distribution functions of the posterior distribution for \( k \).
FIGURE 5. Inference of the time since the most recent selective sweep in African populations. The marginal posterior distribution of the time since the most recent selective sweep ($t_{\text{sweep}}$) shows evidence for a recent selective sweep in both Zimbabwe and Uganda populations under both models of population expansion (black) and constant population size (grey).
Figure 1

Cosmopolitan

African

Population

BEIJING
NETHERLANDS
PENNSYLVANIA
TASMANIA
UGANDA
ZIMBABWE

Nucleotide Diversity per site (%)
Figure 5