Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population

Pablo Duchen,[1] Daniel Živković, Stephan Hutter, Wolfgang Stephan, and Stefan Laurent
Evolutionary Biology, University of Munich, 82152 Planegg-Martinsried, Germany

ABSTRACT *Drosophila melanogaster* spread from sub-Saharan Africa to the rest of the world colonizing new environments. Here, we modeled the joint demography of African (Zimbabwe), European (The Netherlands), and North American (North Carolina) populations using an approximate Bayesian computation (ABC) approach. By testing different models (including scenarios with continuous migration), we found that admixture between Africa and Europe most likely generated the North American population, with an estimated proportion of African ancestry of 15%. We also revisited the demography of the ancestral population (Africa) and found—in contrast to previous work—that a bottleneck fits the history of the population of Zimbabwe better than expansion. Finally, we compared the site-frequency spectrum of the ancestral population to analytical predictions under the estimated bottleneck model.

TO date, several studies have confirmed that *Drosophila melanogaster* originated in sub-Saharan Africa and spread to the rest of the world (Lachaise *et al.* 1988; David and Capy 1988; Begun and Aquadro 1993; Andolfatto 2001; Stephan and Li 2007). With its cosmopolitan distribution we expect that different populations have evolved and adapted differently to distinct environments, making *D. melanogaster* a perfect study system for both adaptation and population history. Extensive research has been performed to detect signatures of adaptation at the genome level (Sabeti *et al.* 2006; Li and Stephan 2006; Zayed and Whitfield 2008). Such detection usually depends on the underlying demographic scenario, since demographic events can leave similar patterns on the genome as adaptive (selective) events (Kim and Stephan 2002; Glinka *et al.* 2003; Jensen *et al.* 2005; Nielsen *et al.* 2005; Pavlidis *et al.* 2008, 2010a). Therefore, a better understanding of the demography of a population will not only allow us to estimate past and present population sizes and the times of the population size changes but will also decrease the rate of false positives of signatures of adaptation. Here we study the demography of African, European, and North American populations, with an emphasis on the North American population.

There is evidence that *D. melanogaster* colonized North America <200 years ago (Johnson 1913; Sturtevant 1920; Keller 2007). *D. melanogaster* (then known as *D. ampelophila*) was first reported in New York in 1875 by New York State entomologist Lintner (Lintner 1882; Keller 2007). In the year 1879 several articles were published indicating the appearance of *D. melanogaster* in several parts of eastern North America, including Connecticut and Massachusetts (Johnson 1913). At that time the dipteran fauna was very well described. It is therefore unlikely that entomologists would have overlooked *D. melanogaster* for long (Keller 2007). Less than 25 years after its introduction, *D. melanogaster* became the most common dipteran species in North America (Howard 1900). Johnson (1913) suggested that North America could have been colonized from the tropics, since the first specimen of *D. melanogaster* in the new world was first described from Cuba (possibly following routes from Central or South America). However, the same author also suggests that the first individuals could have come in vessels from southern Europe during the Spanish regime or from western Africa during the slave trade.

Even if there is agreement regarding the origin of *D. melanogaster*, the demographic history of North American flies is still poorly understood, and population genetic analyses of both the ancestral and derived populations are

doi: 10.1534/genetics.112.145912
Manuscript received September 14, 2012; accepted for publication October 30, 2012
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.145912/-/DC1.

[1] Corresponding author: Evolutionary Biology, University of Munich,, Grosshaderner Strasse 2, 82152 Planegg-Martinsried, Germany. E-mail: duchen@bio.lmu.de

required to tackle this problem. Begun and Aquadro (1993) and Andolfatto (2001) showed that variation in non-African populations (including North America) is a subset of that found in African populations. They suggested a simple "out-of-Africa" bottleneck scenario. Later, Kauer *et al.* (2002) and Caracristi and Schlötterer (2003) used microsatellite data for 40 X-linked loci to study several populations worldwide. Caracristi and Schlötterer (2003) found that some North American populations present only African alleles, whereas other North American populations present only European alleles. Based on the proportion of shared alleles and F_{ST} values, their study shows that American populations are closer to African populations than to European populations. Baudry *et al.* (2004) and Haddrill *et al.* (2005) analyzed 4 and 10 X-linked loci, respectively, but this time using sequence data. Baudry *et al.* (2004) suggested that rare alleles shared between non-African and African populations might represent immigrants from Africa. This agrees with the hypothesis of admixture between European and African flies suggested by Caracristi and Schlötterer (2003). Furthermore, Haddrill *et al.* (2005) found in their North American sample higher diversity and larger linkage disequilibrium than in their European sample, which is also compatible with an admixture scenario.

To infer the population history of North America, we also revisit the demography of the likely source populations from Africa and Europe. Concerning the demography of African *D. melanogaster* Glinka *et al.* (2003) and Pool and Aquadro (2006) found that African samples have an excess of rare derived mutations when compared to the standard neutral model. This excess can be generated by population expansion or a bottleneck. Li and Stephan (2006) proposed a population expansion model for the African population. However, it is still unclear if Zimbabwe is the center of origin. If Zimbabwe lies outside the center of origin we may expect that a bottleneck model would fit the data of the Zimbabwe population better than expansion, since range expansions are usually associated with bottlenecks and founder effects (Excoffier *et al.* 2009). Therefore, we decided to revisit the expansion scenario proposed by Li and Stephan (2006).

In this study we focus on modeling and inferring the demography of *D. melanogaster* using approximate Bayesian computation (ABC) (Tavaré *et al.* 1997; Pritchard *et al.* 1999; Beaumont *et al.* 2002). First, we revisit the demography of the Zimbabwe population and compare a model of instantaneous population expansion with a population bottleneck. Second, having found the best model for our ancestral population we model the joint demography of Africa, Europe, and North America. Finally, we analyze the observed site-frequency spectrum (SFS) of the Zimbabwe population and compare it to analytical predictions.

Materials and Methods

SNP data

Individuals come from three populations: Zimbabwe in Africa (sample size $n = 12$), The Netherlands in Europe ($n = 12$), and Raleigh in North America ($n = 37$). Sequence data consist of 242 intronic and intergenic X-linked loci from each population. African and European loci were originally target sequenced by Glinka *et al.* (2003), Ometto *et al.* (2005), and Hutter *et al.* (2007), while North American loci were extracted from full-genome sequences (publicly available from the Drosophila Population Genomics Project at http://www.dpgp.org) that were created using Illumina next-generation sequencing (NGS) technology. As a first quality control step for the NGS data, all bases with a Phred quality control score <20 were masked. All 242 orthologous loci extracted from the North American data were then aligned to the European and African sequences using MUSCLE (Edgar 2004) to account for insertion/deletion polymorphism. *Drosophila simulans* has been used as an outgroup sequence. As a second quality control step, the alignments were inspected for singleton polymorphisms private to the North American sample and these positions were removed from further analysis. We believe that a sizable fraction of these singleton polymorphisms are created by sequencing errors. This is reflected by the fact that the average quality score of a base causing a singleton polymorphism is significantly lower than the quality of bases creating variants segregating at higher frequencies (Mann–Whitney *U*-test: $P < 2.2 \times 10^{-16}$) (Supporting Information, Figure S1). From all these loci we computed the mean and the variance of the following summary statistics: the number of segregating sites S_n, Watterson's Θ_{W} (Watterson 1975), the average number of pairwise differences in all pairwise comparisons of n sequences Π_n, Tajima's D (Tajima 1989), the number of haplotypes K (Depaulis and Veuille 1998), the linkage disequilibrium statistic Z_{nS} (Kelly 1997), and the distance of Nei as a measure of population differentiation (Nei and Li 1979). Summary statistics of the North American population after exclusion of singletons are also reported. Additionally, we computed the joint site-frequency spectrum (JSFS) of all three pairs of populations, namely: Africa–Europe, Africa–North America, and Europe–North America (Figure S2). Each JSFS was summarized in four classes according to the Wakeley–Hey model (Wakeley and Hey 1997). These summaries are W1 (private polymorphisms in population 1), W2 (private polymorphisms in population 2), W3 (fixed differences between populations), and W4 (shared ancestral polymorphisms). This group of summary statistics, plus the summaries of the JSFS, constitutes our "observed vector" or "observed data" (Tables 1 and 2).

Demographic models of Africa

We first analyzed the data from the ancestral population in Africa. We tested whether an instantaneous expansion or a bottleneck fits better the observed data. The instantaneous expansion model had three parameters: ancestral population size, current population size, and time of expansion (Figure S3). The bottleneck model includes the severity as an additional parameter, which is defined as the ratio of the bottleneck duration and the population size during the

Table 1 Mean and variance (in parentheses) of observed summary statistics over all 242 fragments

	Africa (n = 12)	Europe (n = 12)	North America (n = 37)	Africa (no singletons)	Europe (no singletons)	North America (no singletons)
No. of segregating sites S_n	17.55 (81.31)	6.35 (29.31)	13.10 (50.22)	10.70 (42.45)	4.11 (18.30)	7.47 (29.57)
Watterson's Θ_W	5.91 (9.40)	2.11 (3.30)	3.22 (3.12)	3.57 (4.72)	1.36 (2.01)	1.83 (1.79)
Π_n	5.13 (9.06)	2.18 (4.81)	2.52 (3.64)	3.92 (6.35)	1.36 (2.56)	2.05 (3.16)
Tajima's D	−0.67 (0.34)	−0.09 (1.43)	−0.77 (1.05)	0.33 (0.43)	−0.006 (1.56)	0.21 (1.15)
No. of haplotypes K	9.46 (5.26)	3.87 (3.71)	10.31 (23.24)	8.09 (9.47)	2.85 (2.62)	6.98 (19.25)
Kelly's Z_{nS}	0.15 (0.01)	0.43 (0.075)	0.21 (0.055)	0.23 (0.03)	0.53 (0.08)	0.38 (0.16)

bottleneck (Figure S3). We fixed the duration of the bottleneck to 1000 generations (Laurent *et al.* 2011).

Demographic models of North America–Europe–Africa

Based on the best model for the ancestral population we tested five different models that included all three populations (Figure 1 and Table S2). These five models are: model A ("no migration"), which comprises Africa as the ancestral population; the colonization of Europe is followed by exponential growth, and the colonization from Europe to North America with subsequent exponential growth. Model B ("migration") matches model A but adds an equal migration rate between all populations starting at the colonization time of North America (we assumed that migration between continents increased significantly with human dispersal, which started a few centuries ago). Model C ("admixture") equals the previous models until North America is founded through an admixture between Africa and Europe followed by exponential growth in North America. In model C, we estimated the proportion of European and African ancestry in the founding population of North America. Model D ("no migration II") has Africa as the ancestral population with North America and Europe splitting independently from Africa. Finally, model E ("migration II") matches model D but adds an equal migration rate between all populations starting at the colonization time of North America. Models A and D have 10 parameters, and models B, C, and E have 11 parameters each (Figure 1). In all models the time of colonization of North America was given a very small prior around 200 years ago (the time of the reported colonization of North America). We also let migration due to human-associated dispersal start at this same time (for models B and E). Model selection was performed with all models. For further analysis we selected only models A to C because of the biological assumptions that were already presented in the Introduction. A thorough explanation of the reasons why we discarded other models is presented in the *Discussion*. A more detailed description of all analyzed models can be found in the supporting information (Table S2) and in Figure 1.

ABC simulations

We simulated 100,000 data sets for each of the models described above following the protocol of Laurent *et al.* (2011). Each simulated data set consisted of 242 loci with individual per locus sample sizes, as well as mutation and recombination rates identical to the ones found in the observed data set. Mutation and recombination rates per site per generation for each locus were taken from Laurent *et al.* (2011). Our primary tool was the coalescent simulator *ms* by Hudson (2002). Each parameter was chosen from uniform prior distributions (see Table S1). Missing nucleotides (mostly present in the North American population) were also simulated at the same positions as they occur in the observed data. We accomplished this by following two steps: (1) from the observed data set we generated a missing-nucleotide table with the relative positions (beginning and end) of each chunk of missing nucleotides and recorded this information for each line and for each fragment and (2) by a simple manipulation of the *ms* output we masked all simulated polymorphisms that occurred at the same relative positions that were indicated in the missing-nucleotide table. From the *ms* output we also excluded all singletons that occurred in the simulated North American population. Following the same procedure as with the observed data set we calculated the summary statistics, the SFS, and the JSFS from the modified *ms* output, taking into account missing data in all calculations. Handling of priors, simulation of missing data, exclusion of singletons, and calculation of summary statistics was coded by ourselves. The software

Table 2 Comparisons between pairs of populations

	Africa–Europe	Africa–North America	Europe–North America
Distance of Nei (with singletons in North America)	0.78 (0.66)	1.12 (1.38)	0.59 (1.15)
Distance of Nei (without singletons in North America)	0.69 (0.44)	1.01 (0.93)	0.53 (0.72)
W1 (private polymorphisms of population 1)	2278	1961	214
W2 (private polymorphisms of population 2)	363	743	924
W3 (fixed differences between populations)	17	86	89
W4 (shared polymorphisms between poulations)	647	990	809

The first two lines denote mean and variance (in parentheses) of Nei's distance, and lines 3 to 6 the observed classes of the JSFS.

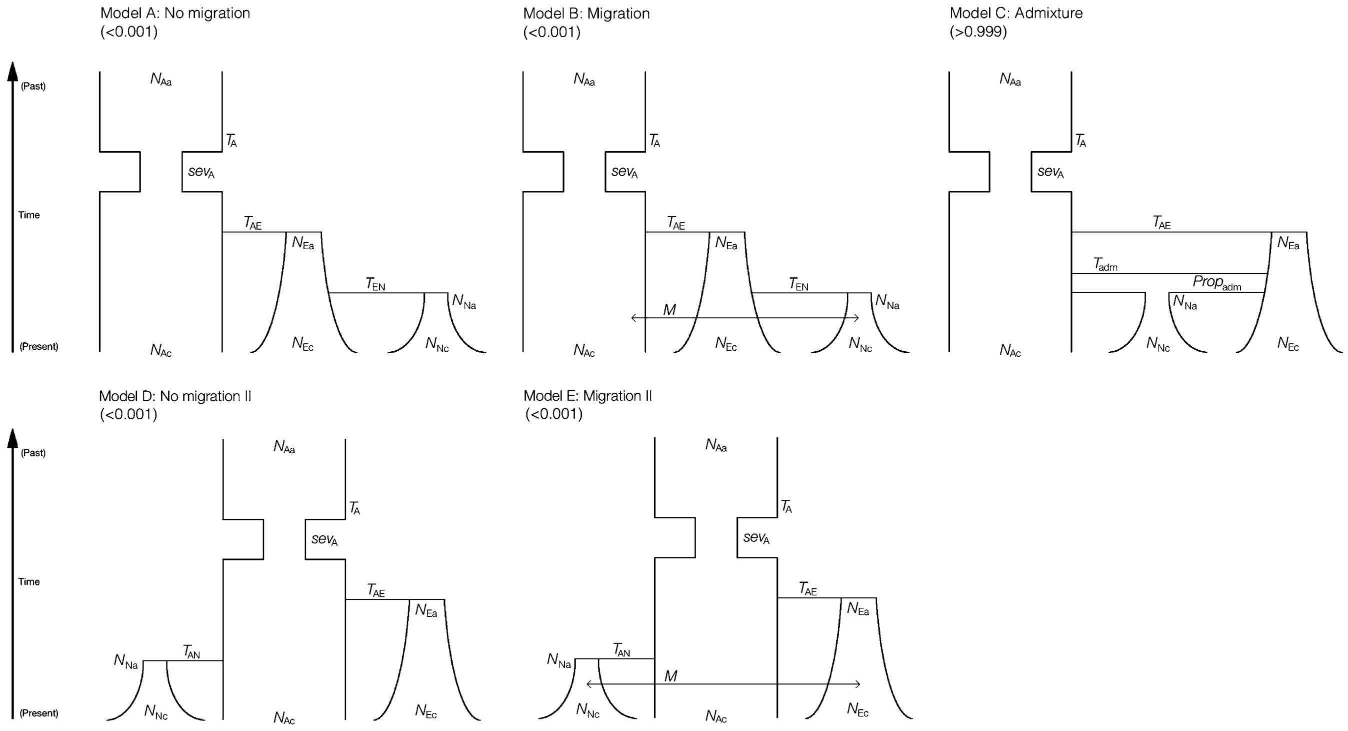

Figure 1 Three-population models. Numbers in parentheses are the posterior probabilities of each model. The symbols are explained in Table 3.

msABC (Pavlidis *et al.* 2010b) is able to perform similar simulations but does not calculate the JSFS. However, we still used msABC to validate our prior distributions. We launched simulations on a 64-bit Linux cluster with 510 nodes (at the Leibniz-Rechenzentrum LRZ, Munich).

Model choice

Model selection was also performed within an ABC framework. Posterior probabilities for each model were calculated according to Fagundes *et al.* (2007). Model selection was done based on the mean and the variance of S_n, mean and variance of Tajima's D and linkage disequilibrium (Z_{nS}). In our analysis (see *Results*) Watterson's Θ_W, Π_n, and K were correlated with S_n and therefore its inclusion did not change the results of the model choice procedure. Model selection was also performed separately using the summaries of the JSFS of all pairs of populations. The model with the highest posterior probability when comparing bottleneck and expansion for the African population as well as the three-population models was chosen as the best fit to the observed data. A validation for using 100,000 simulations for model choice was also performed: we conducted model choice for bottleneck/expansion and between all three-population models A to E for varying numbers of simulations ranging from 10,000 to 200,000 simulations. For the bottleneck *vs.* expansion case we show that starting at 50,000 simulations the posterior probability of the best model does not change significantly when the number of simulations is increased (Figure S4). For the three-population model choice the posterior probability of the best model is always >0.999 if the number of simulations is 10,000 or higher. Therefore a choice of 100,000 simulations for model choice is enough. Model choice performance was assessed by simulating 1000 different pseudo-observed data sets under models A, B, and C (samples for each parameter were taken from the prior distributions as well as from the posterior distributions based on the rejection method). Model choice was performed using the same method as above for each simulated vector of summary statistics. We considered one model to be preferred over the other if the Bayes factor of the models under comparison was above 3.

Parameter estimation

We estimated population parameters of the best African model and of the best three-population model. The number of simulations for parameter estimation was increased to 1,000,000. To validate the use of 1,000,000 simulations for parameter estimation we calculated the mean square error (MSE) of model parameters for varying numbers of simulations, ranging from 100,000 to 1,000,000 simulations (Table S3). Additionally, we also plot the mode and the 95% confidence intervals for varying numbers of simulations (Figure S5). We show that the MSE of each estimate and the estimated mode stay both relatively constant (Table S3 and Figure S5). Therefore, 1,000,000 simulations are enough for parameter estimation. Estimation was based on ABC rejection (Tavaré *et al.* 1997; Pritchard *et al.* 1999) and regression (Beaumont *et al.* 2002) methods. Both methods were performed using Wegmann's ABCtoolbox (Wegmann *et al.* 2009) and checked with Csilléry's abcR (Csilléry *et al.* 2012). First, we pooled all statistics and checked for correlations with the parameters. We did not keep statistics that did not correlate with any parameter, because keeping them does not provide information for the estimation and would only add noise to the final estimates. All these statistics were transformed using partial least squares (p.l.s.) as implemented in Wegmann *et al.* (2009). This transformation is advantageous because it extracts a small number of orthogonal components from a highly dimensional array of summary statistics. The new set of transformed statistics (with reduced dimensionality) reduces the noise produced by uninformative summary statistics. Moreover, the p.l.s.-transformed statistics are completely uncorrelated with one another ensuring the assumption of singularity, which is required for estimating parameters according to the regression method (Beaumont *et al.* 2002).

Predictive simulations

To check for the quality of our parameter estimates we took two approaches: (1) we sampled parameter values from the posterior distributions (based on the regression method) of each parameter estimate and resimulated data sets, and (2) we plotted the distributions of summary statistics directly from the set of the 5000 simulations closest to the observed data (which represents a sample of the joint posterior distribution based on the rejection method). The resulting distributions of summary statistics were compared to the observed ones for both approaches and plots were generated (see *Results*). Both approaches were performed only under the best model, since this is a test to see how well the best model fits the observed data. The same predicitive simulations were also performed for autosomal data (50 intergenic and intronic loci from chromosome 3R) to check how good our best model can predict autosomal summary statistics. For the sake of computational simplicity we assumed a relative effective population size (N_e) ratio of 0.75 for X-linked *vs.* autosomal loci in our simulations. This assumes a 1:1 male/female ratio in all populations even though we have evidence that actual sex ratios might deviate from these expectations (Hutter *et al.* 2007). However, we expect that this simplification should have only minor effects on our ability to predict the autosomal data since even in extreme cases of sex bias the X/A ratio of N_e can never drop below 0.5625 or exceed 1.125 (Hedrick 2011, Chap. 4).

Prediction of the site-frequency spectrum of Zimbabwe

Our available sequence data not only allow us to summarize genetic diversity with S_n, Θ_W, or Π_n, but also allow us to compute the observed SFS of the African population (Figure 2) and compare it to predictions under a given demographic model. Analytical methods for predicting the SFS of one population for arbitrary deterministic changes in population size have been successfully developed (Griffiths and Tavaré 1998; Živković and Wiehe 2008; Živković and Stephan 2011) and are briefly revisited as follows. Let $T_n, \ldots, T_2$ be the time periods during which the genealogy has $n, \ldots, 2$ lineages, respectively. Furthermore, let $\lambda(t) = N(t)/N$ denote the ratio of the population sizes at time t in the past and the present. The probability $p_{n,k}(i)$ that a randomly chosen line of waiting time T_k, $k = n, \ldots, 2$, has i descendants, $i = 1, \ldots, n - 1$, during time T_n (Fu 1995; Griffiths and Tavaré 1998) is

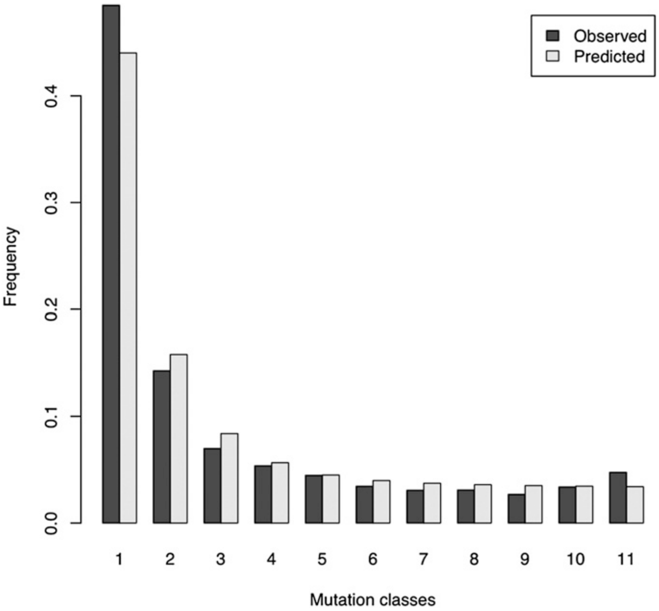

Figure 2 Observed (solid) and predicted (shaded) site-frequency spectrum of the African population. To calculate the frequency classes Equation 4 was used.

$$p_{n,k}(i) = \binom{n-i-1}{k-2} \Big/ \binom{n-1}{k-1}. \qquad (1)$$

The mean waiting times are given by

$$E(T_k) = \sum_{j=k}^{n} (-1)^{j+k} \alpha_{n,j,k} \int_0^\infty \exp\left\{ -\binom{j}{2} \int_0^t \frac{1}{\lambda(u)} du \right\} dt, \qquad (2)$$

where

$$\alpha_{n,j,k} = \frac{(2j-1)n!(n-1)!(k+j-2)!}{(j-k)!k!(k-1)!(n-j)!(n+j-1)!}.$$

The integral in (2) can be solved explicitly for models that consist of multiple instantaneous changes in population size and be evaluated numerically for models that include phases of exponential growth. Let L_i be the total length of branches leading to i descendants, where i represents singletons, doubletons, etc. Then,

$$E(L_i) = \sum_{k=2}^{n-i+1} k p_{n,k}(i) E(T_k). \qquad (3)$$

Assuming an infinitely many sites mutation model (Kimura 1969), the expected unfolded site frequency ξ_i for each class i is given by

$$E(\xi_i) = \frac{E(L_i)}{\sum_{k=2}^{n} k\, E(T_k)}. \qquad (4)$$

We use the ABC parameter estimates obtained for Zimbabwe as an input to the equations shown above, calculate the SFS based on Equation 4, and compare it to the observed SFS.

Results

Observed data

A first examination of the observed summary statistics (Table 1) shows that Africa is the most diverse population (based on the number of segregating sites), followed by North America and Europe. Watterson's Θ_W and Π_n follow the same pattern. Tajima's D is most negative in North America (−0.77), followed by Africa (−0.67) and Europe (−0.09). Linkage disequilibrium (Z_{nS}) is highest in Europe (0.43) compared to North America (0.21) and Africa (0.15). Population differentiation (Table 2) is highest between Africa and North America (distance of Nei = 1.12), followed by Africa–Europe (0.78) and North America–Europe (0.59). All these comparisons are based on the observed data set that included singletons in North America. The resulting statistics of North America after excluding singletons can also be found in Table 1.

The SFS of the African population is shown in Figure 2. Regarding the JSFS (Table 2) we observe an excess of private polymorphisms in Africa when compared to private polymorphisms in Europe (2278 *vs.* 363) and North America (1961 *vs.* 743) (W1 *vs.* W2). We must keep in mind that singletons were excluded from the North American population, and these singletons are mostly private to North America. The opposite pattern is seen when comparing private polymorphisms in Europe to private polymorphisms in North America (214 *vs.* 924). Shared polymorphism (W4) has its lowest value between Africa and Europe (647) when compared to Africa–North America (990) and Europe–North America (809). The number of fixed differences between populations is small in all pairwise comparisons (W3).

African demography

Model choice results show that a population bottleneck in Africa (P = 0.987) fits the observed data better than an expansion (P = 0.013). We used the following statistics for parameter estimation of the best model: mean and variance of S_n, mean and variance of Tajima's D, and mean Z_{nS}. We estimated these parameters (Table 3) using the priors listed in Table 4. After the reduction of dimensionality using partial least squares (see *Materials and Methods*) we kept three components from the original five statistics used. The estimated ancestral and current N_e are 4.9 million and 5.2 million individuals, respectively. The bottleneck severity (Log_{10} scale) was estimated as 0.21, which corresponds to ~620 individuals for a fixed bottleneck duration of 1000 generations. The estimated time of the bottleneck is ~200,000 years ago, assuming 10 generations per year (Table 4 and Figure S6). Predicted distributions of summary statistics for the bottleneck and the expansion models overlap significantly. However, observed Tajima's D as well as the mean and the variance of S_n are reproduced more often by the bottleneck model than by the expansion model (Figure S7). Estimations of the African parameters were also performed using the classes of the folded SFS of Zimbabwe but the results do not vary significantly (data not shown).

Table 3 Parameters used in models A, B, C, D, and E

Abbreviation of parameter	Explanation
N_{Aa}	Ancient population size of Africa
sev_A	Severity of the bottleneck in Africa
T_A	Time of the bottleneck in Africa
N_{Ac}	Current population size of Africa
T_{AE}	Time of split between Africa and Europe
T_{AN}	Time of split between Africa and North America
N_{Ea}	Starting population size of Europe
N_{Ec}	Current population size of Europe
T_{EN}	Time of split between Europe and North America
N_{Na}	Starting population size of North America
N_{Nc}	Current population size of North America
M	Migration rate between all populations
T_{adm}	Time of admixture between Africa and Europe
Prop_{adm}	Proportion of European admixture in North America

Site-frequency spectrum

The SFS of the observed African data has an excess of high-frequency-derived variants (Figure 2, solid bars), while the predicted SFS under a bottleneck does not show such a large excess (Figure 2, shaded bars). Predicted values were calculated using the modes of the parameter estimates under the bottleneck scenario (Table 4) and applying Equation 4. Predicted values fit the observed SFS better than the expansion model of Li and Stephan (2006) for the intermediate-frequency classes, but not for the low-frequency variants. The largest relative discrepancies are found for both models for the high-frequency variants that make the SFS slightly U shaped.

Table 4 Parameter estimates of the African population

Parameter	Prior	Mode	95% quantiles
N_{Ac}	unif(1×10^5, 1×10^7)	4,975,360 individuals	(2.40×10^6, 9.13×10^6)
T_A (in years)	unif(1×10^2, 4×10^5)	237,227 years ago	(0.82×10^5, 3.45×10^5)
N_{Aa}	unif(1×10^5, 1×10^7)	5,224,100 individuals	(1.98×10^6, 9.55×10^6)
sev_A (Log_{10})	unif(−2,2)	0.21	(−0.15, 0.57)

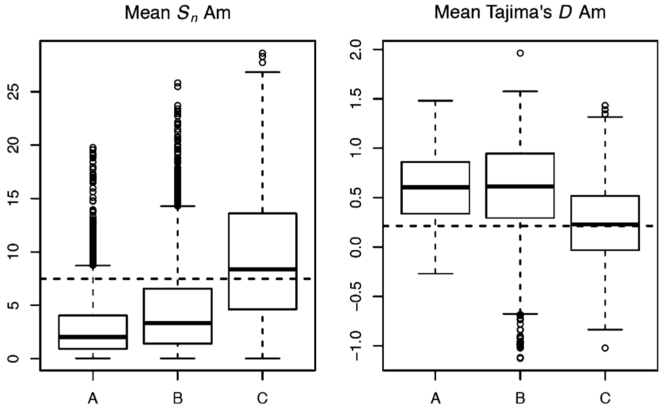

Figure 3 Predicted summary statistics under models A, B, and C for the North American population based on the rejection method. The horizontal dashed line represents the observed value.

North American demography

The model with the highest posterior probability is the admixture model C with $P > 0.999$. Model choice yields the same results when using summary statistics and also when using the JSFS (in both cases the posterior probability of model C is >0.999). Parameters of this model are explained in Table 3. Predictive simulations based on both the regression and rejection methods show that admixture is the only model that can explain the diversity observed in North America (Figure 3 and Figure S8). Admixture can also explain better the observed Tajima's D in America (Figure 3). It is important to remember at this point that diversity in North America is higher than in Europe, although the colonization of North America has been much more recent than the one inferred for Europe. It is thus reasonable to believe that admixture is playing an important role in this case.

Other parameters, such as diversity in Africa and Europe can be explained by both admixture and nonadmixture models (Figure S8 and Figure S9). The accuracy of the model choice procedure shows that the simulated model could be correctly identified in 90% of the cases. The cases in which model C is not preferred occur when one or a combination of the following events happen: (a) the time of split between African and European populations is very young (about 1000 to 2000 years ago), (b) the proportion of European ancestry in the North American population is very high (above 90%), and (c) the founding population of Europe is large (in the order of 100,000 individuals). The results of model choice performance when sampling from the posterior distributions of each parameter do not vary significantly with the ones we provide here (see *Materials and Methods*).

For estimating the parameters of model C, we used the following statistics: mean and variance of S_n in Africa, mean and variance of Tajima's D in Africa, mean K in Africa, mean and variance of Tajima's D in Europe, mean and variance of K in Europe, mean Z_{nS} in Europe, mean S_n in North America, mean and variance Tajima's D in North America, mean and variance of K in North America, mean Z_{nS} in North America, mean distance of Nei Africa–Europe, mean distance of Nei Africa–North America, mean distance of Nei Europe–North America, W1 Africa–North America, W2 Africa–North America, W4 Africa–North America, W1 Africa–Europe, W2 Africa-Europe, and W2 Europe–North America. The abovementioned statistics were chosen after pooling all statistics and checking for correlations between statistics and parameters (see *Materials and Methods*). After dimensionality reduction using partial least squares we kept six components. Parameter estimates (Table 5 and Figure S10) imply that African and European populations split around 19,000 years ago and Europe was founded with around 17,000 individuals. These estimates are in agreement with previous studies (Li and Stephan 2006; Laurent *et al.* 2011). The North American population was founded by ~2500 individuals from which ~85% are of European ancestry and the remaining of African ancestry (Figure 4). The current population sizes of Europe and North America cannot be estimated accurately.

Predictive simulations of model C (Figure S11 and Figure S12) were generated by sampling parameters from the posterior distributions (based on the regression method). These parameters were used to simulate data sets and calculate summary statistics and JSFS statistics (see *Materials and Methods*). The resulting distributions show that all summary statistics can be well predicted by the admixture model (Figure S11 and Figure S12). The only statistics that are overestimated are the number of fixed differences (W3) between Africa and North America or Europe and North America and the distance of Nei between Europe and America. W3 and distance of Nei are related to each other, and an increase in one involves always an increase in the other. An improvement of the model in this aspect is discussed below

Table 5 Joint parameter estimates of the European and North American populations

Parameter	Prior	Mode	95% quantiles
T_{AE} (decimal log generations)	unif(4,7)	5.29 (~19,000 years ago)	(4.69, 5.86)
T_{adm} (decimal log generations)	unif(2,4)	3.16	(2.08, 3.82)
N_{Ec}	unif(1x10^4,1 × 10^7)	3,122,470 individuals	(0.39 × 10^6, 9.55 × 10^6)
N_{Ea} (decimal log)	unif(2,5)	4.23 (~17,000 individuals)	(3.58, 4.83)
N_{Nc}	unif(1x10^4,3 × 10^7)	15,984,500 individuals	(1.11 × 10^6, 28.8 × 10^6)
N_{Na} (decimal log)	unif(2,5)	3.40 (~2500 individuals)	(2.20, 4.79)
Prop$_{\mathrm{adm}}$	unif(0.01,0.99)	0.85	(0.64, 0.97)

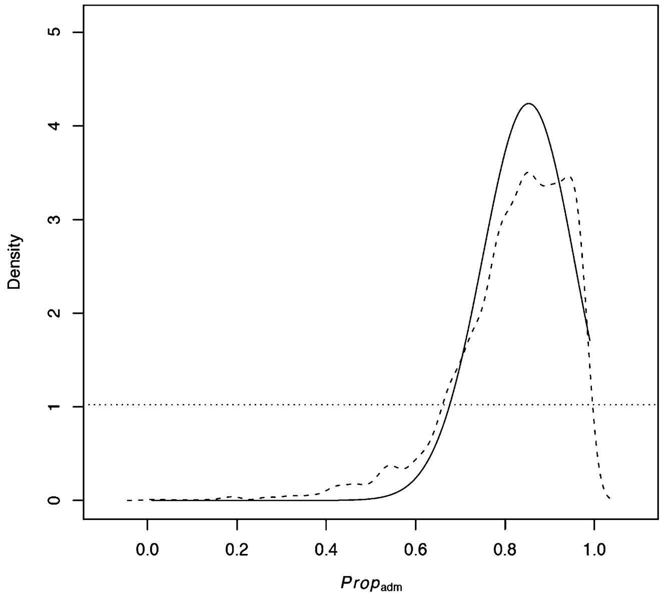

Figure 4 Probability density of the proportion of European admixture based on the regression method (solid line) and rejection method (dashed line). The horizontal dotted line represents the uniform prior distribution.

(see *Discussion*). Model C was also able to predict autosomal summary statistics quite accurately (Figure S13 and Figure S14) even under the simplified assumption of equal sex ratios in all populations (see *Material and Methods*).

Discussion

The demography of the Zimbabwe population was modeled in several studies as a simple expansion process (Glinka *et al.* 2003; Ometto *et al.* 2005; Laurent *et al.* 2011). However, it is still unclear if the Zimbabwe population is the source from which all other *D. melanogaster* populations derive. Based on this scenario we may expect that a bottleneck model would fit the data of the Zimbabwe population better than expansion, since range expansions are usually associated with bottlenecks and founder effects (Excoffier *et al.* 2009). Indeed, what we find here is exactly that pattern: the bottleneck model is significantly preferred over the expansion model.

The predictive simulations of models A, B, and C show that all models are able to explain the diversity observed in Africa and Europe (Figure S8 and Figure S9). However, only model C (including admixture) is able to fully explain the diversity observed in North America. Model A involves a recent foundation of North America from Europe, but North America shows currently greater diversity. This is hard to explain without considering an input from the ancestral population. Model B provides this input from Africa through migration, but to be able to reach the levels of diversity observed in North America we would need unrealistically high rates of migration. However, this would not be compatible with the observed values of population differentiation. Model C is in accordance with the observed data in this aspect. Another aspect that favors the admixture model over the others is that the values of Tajima's *D* in North America and Europe can also be better explained. We do not have an intuitive explanation why a recent admixture event has an influence on Tajima's *D* in one of the parental populations (*i.e.*, the European one).

Among all tested models (Table S2), we selected models A, B, and C for two main reasons. First, there is evidence that North American *D. melanogaste*r has been introduced from Europe (see Introduction) and we have strong biological reasons to believe that North American diversity was generated through admixture and/or migration with African populations. Second, we wanted to keep the models as simple as possible. When we examined the data we observed that the North American population shares polymorphisms mostly with the European population and, to a lesser extent, with the African population. This observation fits the hypothesis of a European contribution. A model in which North America is derived from the African population without any European contribution would not be able to explain the shared polymorphism between North America and Europe in the observed data.

In addition to the three main demographic models (*i.e.*, models A, B, and C), we examined two more models in which the North American population derives directly from the African one. This alternative topology of the population's genealogy was tested without migration (model D, Table S2) and considering a simple migration process, identical to the one used in model B (model E, Table S2). These models represent possible alternative explanations for the high diversity harbored by the North American population. However, when compared to model C, models D and E are less supported by the data set as indicated by their associated posterior probabilities (>0.999, <0.001, and <0.001, respectively).

We note here that our modeling of the dispersal patterns between worldwide populations of *D. melanogaster* is a crude simplification of the real, but unknown migratory processes characterizing this species. It is well possible that more complex demographic models allowing specific, and potentially asymmetric, migration rates between all pairs of populations might be a more accurate representation of reality. However, in our case, these more sophisticated models have the property of having divergence time and specific migration rates as free parameters for several pairs of populations. A recent simulation study showed that the joint estimation of these two parameters in an ABC framework does not yield satisfying results (Tellier *et al.* 2011). Indeed, it is not clear at the present time which summarization of the raw data set would allow for an accurate joint estimation of divergence times and migration rates within an ABC framework. Although more work is needed to develop methods that allow for the estimation of more complex models, the analysis presented in this study shows that the history of the North American population is well characterized by an admixture of alleles coming from European and African populations.

The admixture model C can predict most of the observed summary statistics and JSFS equally well or better than the other models, except for the observed population differentiation (distance of Nei) between Africa and North America, which is better explained by model A or B (Figure S8). This higher simulated population differentiation in model A or B is associated with lower values of diversity in North America than the observed one, which is still a drawback for these models. We investigated this fact by adding more parameters to the model. We tested three variations of model C: model C1 has an extra bottleneck during the colonization of North America from Africa, model C2 has an extra bottleneck during the colonization of North America from Europe, and model C3 has both bottlenecks (Figure S15). While including the additional bottlenecks can account for the observed population differentiation they also reduce diversity below the observed values. Therefore, when compared to the original admixture model, models C1, C2, and C3 were not favored.

Another possible model in which higher values of population differentiation could be expected is a scenario in which samples are considered to be taken from demes in a metapopulation. If we have samples from different demes from different populations we may not expect migration or admixture to take place equally between all sampled demes, which may lead to higher values of population differentiation. Even though population differentiation in African populations is minimal (Yukilevich *et al.* 2010) this hypothesis still needs to be investigated further, with additional analyses of populations from Africa, Europe or North America, which is beyond the scope of this study.

To obtain further insight into the demography of the Zimbabwe population, we compared the SFS of this population with that predicted under a bottleneck. Regarding the input parameters for this prediction we used the modes (as point estimates) of the posterior distributions that were generated by the ABC regression step (see Table 4). Figure 2 shows the observed SFS compared to the predicted SFS under the conditions described above. Li and Stephan (2006) fitted a population expansion model to this same observed SFS (Figure 3 of Li and Stephan 2006). The bottleneck model in our study fits the intermediate-frequency classes better, whereas the population expansion model is more compatible with the classes of the singletons and doubletons. However, for the high-frequency variants both models show relatively large discrepancies. According to Li and Stephan (2006), this may indicate evidence for positive selection, a hypothesis that needs to be further tested. An alternative explanation for the excess of high-frequency variants may be ancestral state misidentification (Hernandez *et al.* 2007). Note that ancestral misidentification does not change our main ABC results, since the summary statistics used (including the folded SFS) are unaffected by polarization.

Although our modeling approach takes into account the combined effects of mutation, genetic drift, and migration we point out that we did not consider any form of natural selection in this analysis. This omission does not reflect that we believe that the impact of selection is minimal in our data set but rather the lack of available methods to estimate demographic and selective forces simultaneously. We think that such methods would greatly improve the interpretation of data sets like the one we present here, since several studies recently reported evidence that, contrary to previous beliefs, negative and positive selection have a substantial impact on the genetic variation harbored by natural populations of *D. melanogaster* (Macpherson *et al.* 2007; Jensen *et al.* 2008). Until such methods are available it is hard to predict to what extent the results presented in this study are affected by a reduction of the evolutionary history of *D. melanogaster* to a strictly neutral nonequilibrium model.

Nonetheless, we stress that the main result of this study, which is the identification of a substantial contribution of the African gene pool to the North American population, cannot be invalidated by including selection in our analysis. The reason for this is that the above-mentioned result relies on the observation that the level of genetic diversity found in the North American population is too high compared to expectations under a model in which the North American population would derive exclusively from the European one.

In conclusion, this study generated the first joint demographic analysis of African, European and North American populations of *D. melanogaster*. We analyzed the African population and found that a bottleneck fits the observed data better than an instantaneous population expansion. Regarding the North American population, we found that an admixture model fits the observed data significantly better than models involving colonization only from Europe or migration. We estimated the population parameters of all populations, from which we highlight the time of split between Africa and Europe (∼19,000 years ago) and the proportion of European and African ancestry in the North American population (85% and 15%, respectively). The time of colonization of North America was given a very small prior because we know it took place ∼200 years ago. In general, having described such a demographic model for North America, Africa, and Europe will be of valuable importance when looking for signatures of adaptation in any of these populations.

Acknowledgments

We thank Laurent Excoffier for his scientific advice on ABC and the modeling of population range expansions. We are also grateful to two reviewers for their valuable comments. This project has been funded by the Research Unit 1078 of the German Research Foundation (grant STE 325/12 to W. S. and HU 1776/2 to S.H.) and the Volkswagen Foundation (grant I/84232 to D.Z.).

Literature Cited

Andolfatto, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. 18: 279–290.

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

Beaumont, M. A., W. 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.

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

Csilléry, K., M. G. B. Blum, and O. François, 2012 abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3: 475–479.

David, J. R., and P. Capy, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. 4: 106–111.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. 15: 1788–1790.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Excoffier, L., M. Foll, and R. J. Petit, 2009 Genetic consequences of range expansions. Annu. Rev. Ecol. Evol. Syst. 40: 481–501.

Fagundes, N. J. R., N. Ray, M. A. Beaumont, S. Neuenschwander, F. M. Salzano *et al.*, 2007 Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. USA 104: 17614–17619.

Fu, Y.-X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. 48: 172–197.

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.

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Stoch. Models 14: 273–295.

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.

Hedrick, P. W., 2011 *Genetics of Populations*. Jones & Bartlett, Boston.

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.

Howard, L. O., 1900 A contribution to the study of the insect fauna of human excrement. Proc. Washington Acad. Sci. 2: 541–604.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Hutter, S., H. Li, S. Beisswanger, D. De Lorenzo, and W. Stephan, 2007 Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. Genetics 177: 469–480.

Jensen, J. D., Y. Kim, V. B. 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., K. R. Thornton, and P. Andolfatto, 2008 An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. PLoS Genet. 4: e1000198.

Johnson, C. W., 1913 The distribution of some species of Drosophila. Psyche 20: 202–205.

Kauer, M., B. Zangerl, D. Dieringer, and C. Schlötterer, 2002 Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster.* Genetics 160: 247–256.

Keller, A., 2007 *Drosophila melanogaster*'s history as a human commensal. Curr. Biol. 17: R77–R81.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics 146: 1197–1206.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893–903.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. 22: 159–225.

Laurent, S. J., 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 W. Stephan, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2: 1580–1589.

Lintner, J. A., 1882 *First Annual Report on the Injurious and Other Insects of the State of New York*. Weed, Parsons, Albany, NY.

Macpherson, J. M., G. Sella, J. C. Davis, and D. Petrov, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. Genetics 117: 2083–2099.

Nei, M., and W.-H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76: 5269–5273.

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.

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.

Pavlidis, P., S. Hutter, and W. Stephan, 2008 A population genomic approach to map recent positive selection in model species. Mol. Ecol. 17: 3585–3598.

Pavlidis, P., J. D. Jensen, and W. Stephan, 2010a Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics 185: 907–922.

Pavlidis, P., S. Laurent, and W. Stephan, 2010b msABC: a modification of Hudson's *ms* to facilitate multi-locus ABC analysis. Mol. Ecol. Res. 10: 723–727.

Pool, J. E., and C. F. Aquadro, 2006 History and structure of sub-Saharan populations of *Drosophila melanogaster.* Genetics 174: 915–929.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, 1999 Population growth of human Y chromosome: a study of Y chromosome microsatellites. Mol. Biol. Evol. 16: 1791–1798.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.*, 2006 Positive natural selection in the human lineage. Science 312: 1614–1620.

Stephan, W., and H. Li, 2007 The recent demographic and adaptive history of *Drosophila melanogaster.* Heredity 98: 65–68.

Sturtevant, A. H., 1920 Genetic studies on *Drosophila simulans*. I. Introduction: hybrids with *Drosophila melanogaster.* Genetics 5: 488–500.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics 145: 505–518.

Tellier, A., P. Pfaffelhuber, B. Haubold, L. Naduvilezhath, L. E. Rose *et al.*, 2011 Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. PLoS ONE 6: e18155.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics 145: 847–855.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182: 1207–1218.

Yukilevich, R., T. L. Turner, F. Aoki, S. V. Nuzhdin, and J. R. True, 2010 Patterns and processes of genome-wide divergence between North American and African *Drosophila melanogaster*. Genetics 186: 219–239.

Zayed, A., and C. W. Whitfield, 2008 A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee *Apis mellifera*. Proc. Natl. Acad. Sci. USA 105: 3421–3426.

Živković, D., and T. Wiehe, 2008 Second-order moments of segregating sites under variable population size. Genetics 180: 341–357.

Živković, D., and W. Stephan, 2011 Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. Theor. Popul. Biol. 79: 184–191.

Communicating editor: D. Begun

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.145912/-/DC1

Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population

Pablo Duchen, Daniel Živković, Stephan Hutter, Wolfgang Stephan, and Stefan Laurent

DOI: 10.1534/genetics.112.145912

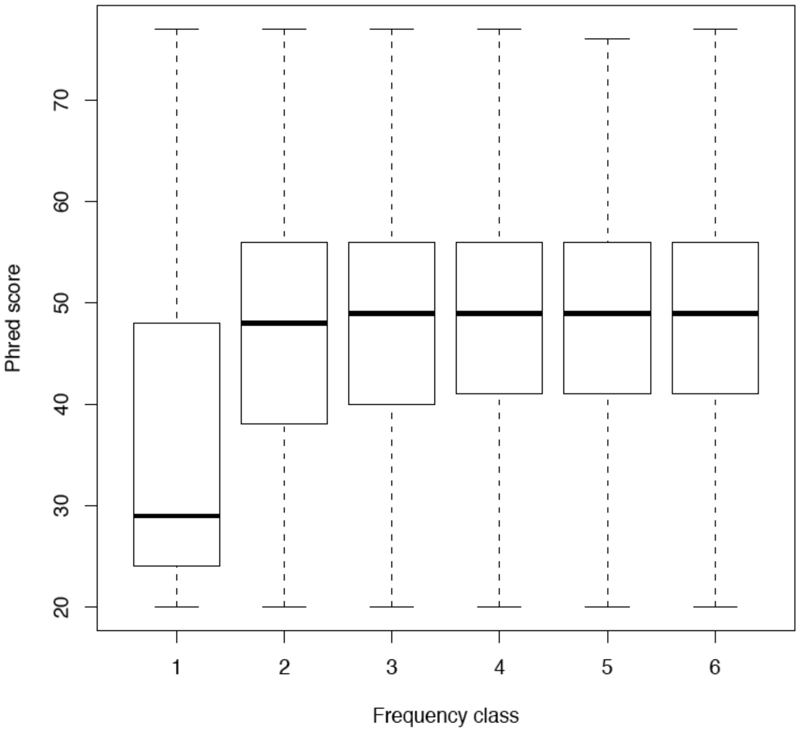

Figure S1 Phred quality scores of individual base calls belonging to the first six classes of the site frequency spectrum in the North American population (calculated from the DPGP1 raw fastq files). The middle bands indicate the median values, the boxes the upper and lower quartiles and the whiskers the minimum and maximum values.

	0	1	2	…	n_2-2	n_2-1	n_2
0	X	W2					W3
1	W1	W4					W1
2							
…							
n_1-2							
n_1-1							
n_1	W3	W2					X

Figure S2 Joint Site Frequency Spectrum (JSFS) classes, according to the Wakeley-Hey model. On left most column we have the sample size n_1 of population 1. On the upper most row we have the sample size n_2 of population 2. The summary statistics proposed by Wakeley-Hey (1997) are represented by the letters W1 to W4.

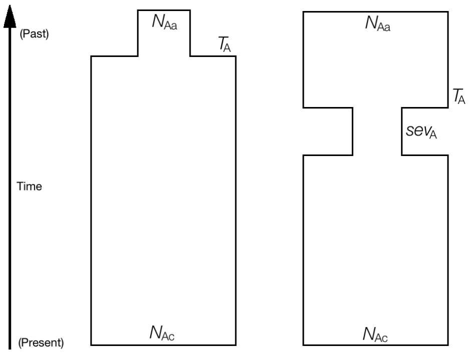

Figure S3 Population expansion (left) versus Bottleneck (right) model in Africa. The posterior probability of the Expansion model is 0.013. The posterior probability of the Bottleneck model is 0.987. Parameters are explained in Table 3.

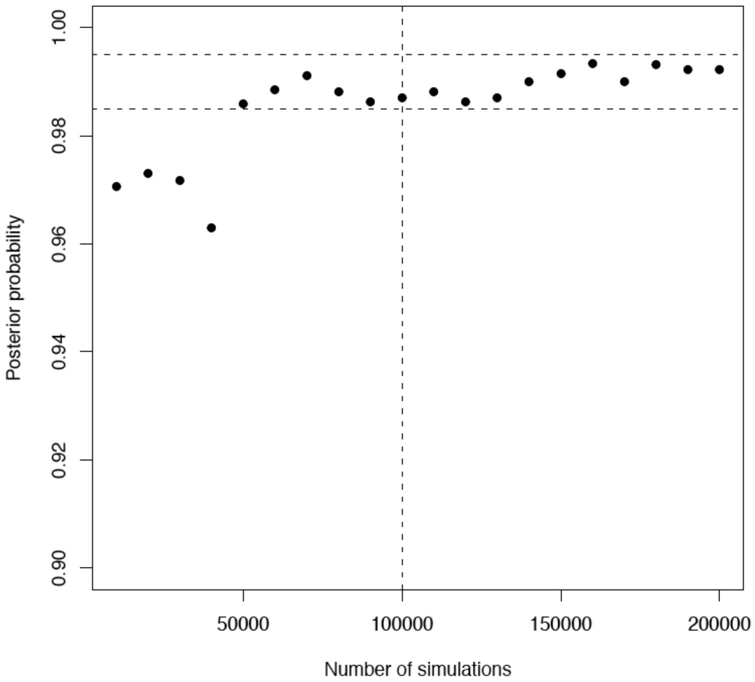

Figure S4 Behavior of the posterior probabilities of the Bottleneck model for different numbers of simulations. In the case of the Admixture model (model C) the posterior probability is always above 0.999 for different numbers of simulations.

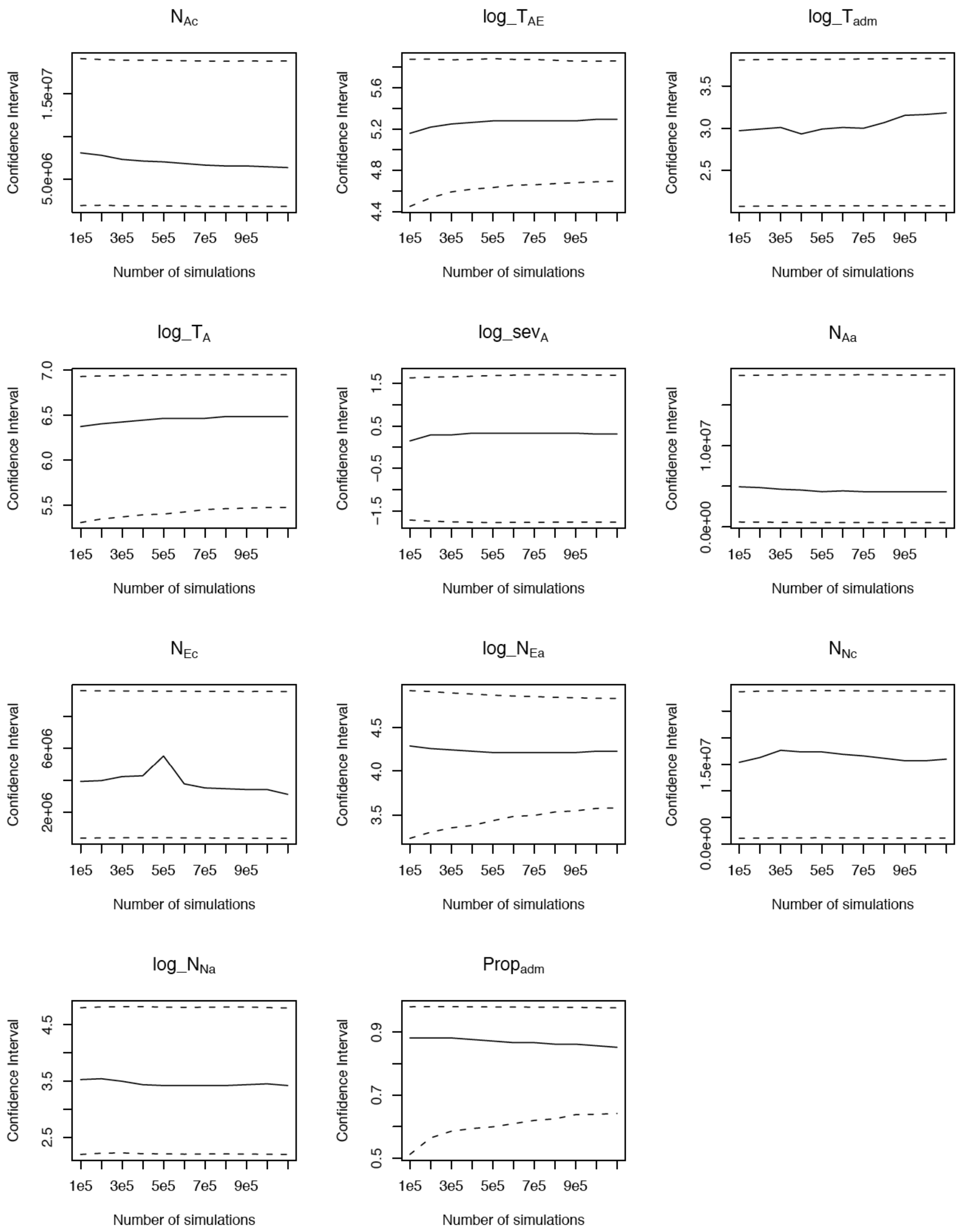

Figure S5 Behavior of the modes and 95% confidence intervals of the estimates of the parameters of the Admixture model (model C) for different numbers of simulations. Solid line: mode, dashed lines: upper and lower confidence intervals.

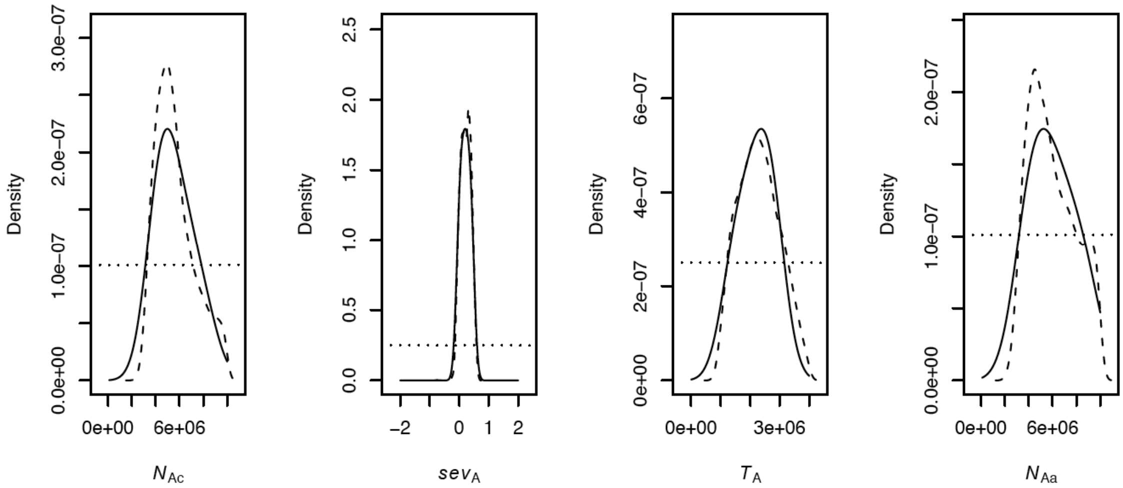

Figure S6 Posteriors of the Bottleneck model in Africa. Posteriors are represented by the rejection method (dashed line) and the regression method (solid line). Parameter abbreviations are explained in Table 3. Mode and confidence interval for each parameter are shown in Table 4.

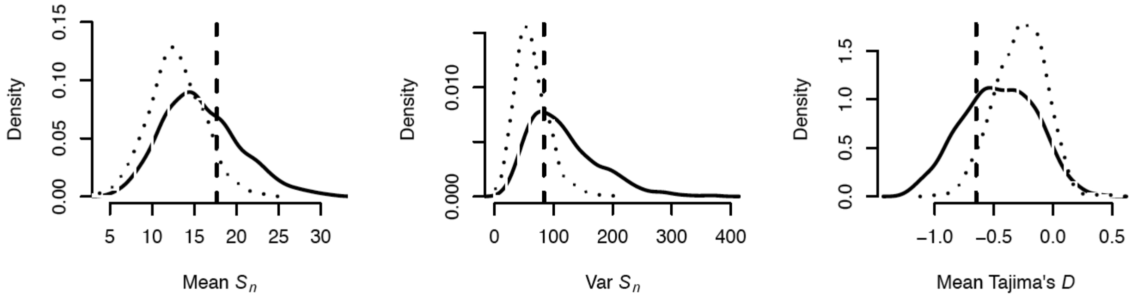

Figure S7 Predictions of the Bottleneck versus Population Expansion in Africa. Solid line: Bottleneck, dotted line: Population expansion, vertical dashed line: observed value. Parameters for predictive simulations are drawn from the posterior distributions generated by the regression method (see Materials and Methods).

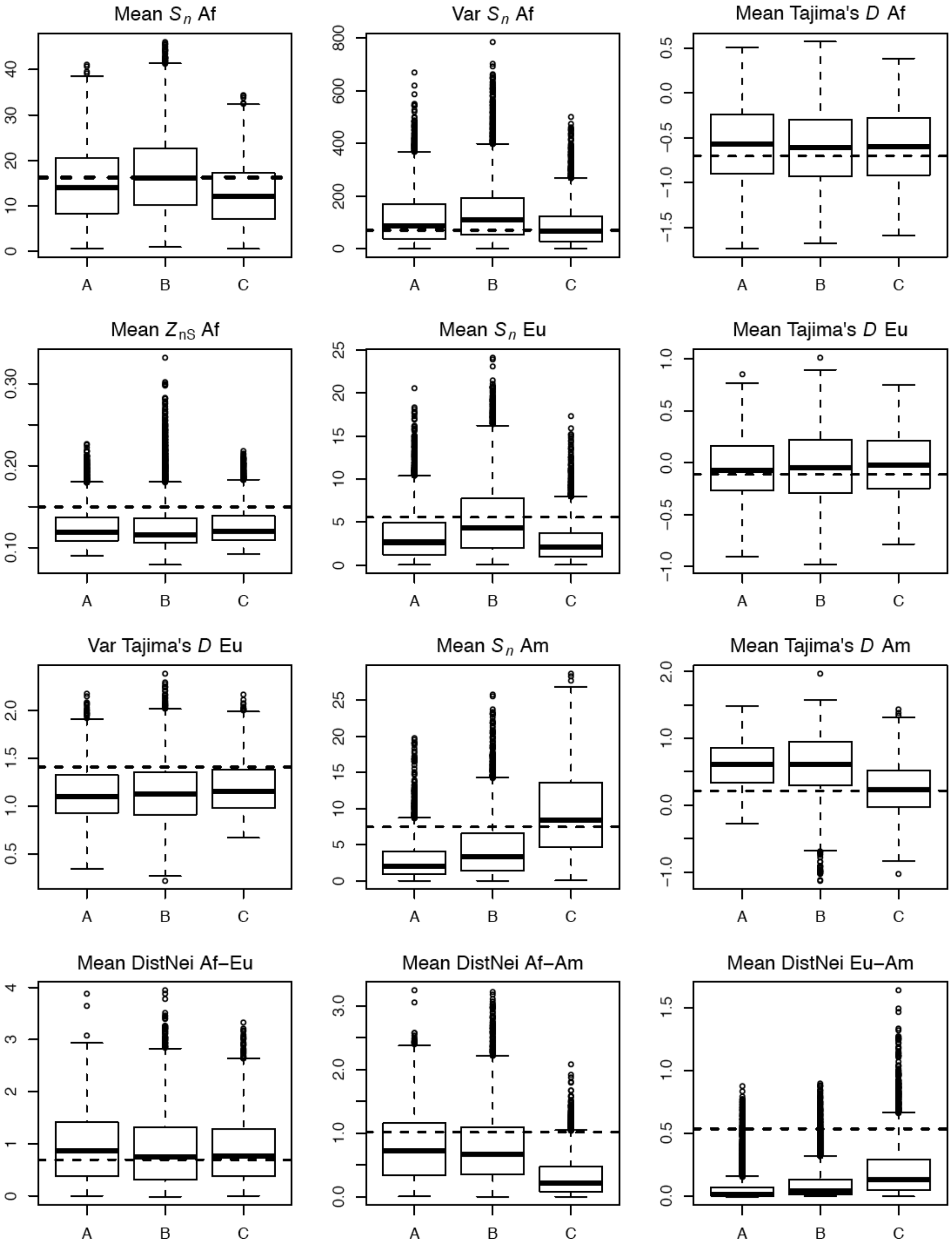

Figure S8 Predictions of summary statistics for models A, B and C based on the rejection method. The horizontal dashed line represents the observed value.

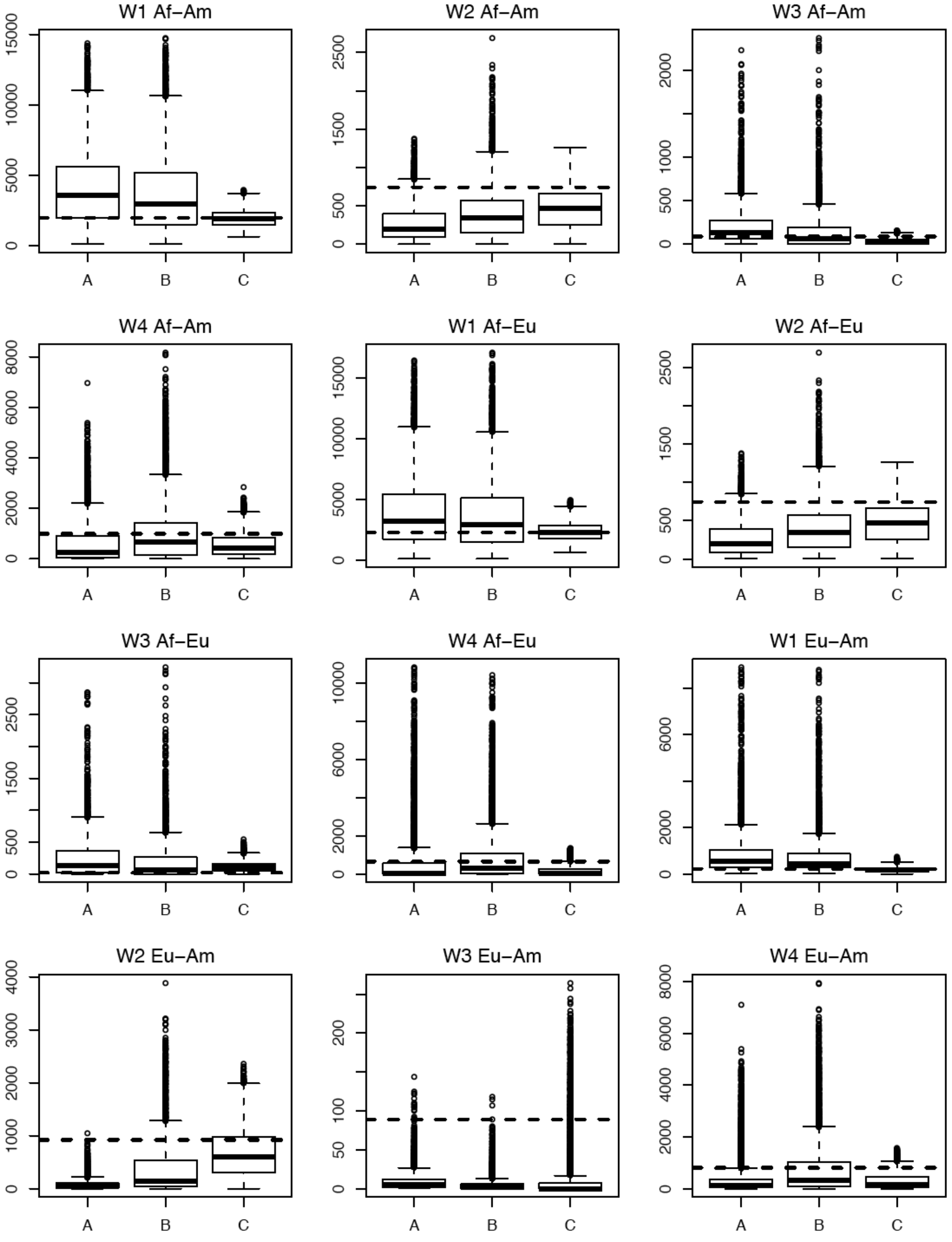

Figure S9 Predictions of the JSFS for models A, B and C based on the rejection method. The horizontal dashed line represents the observed value.

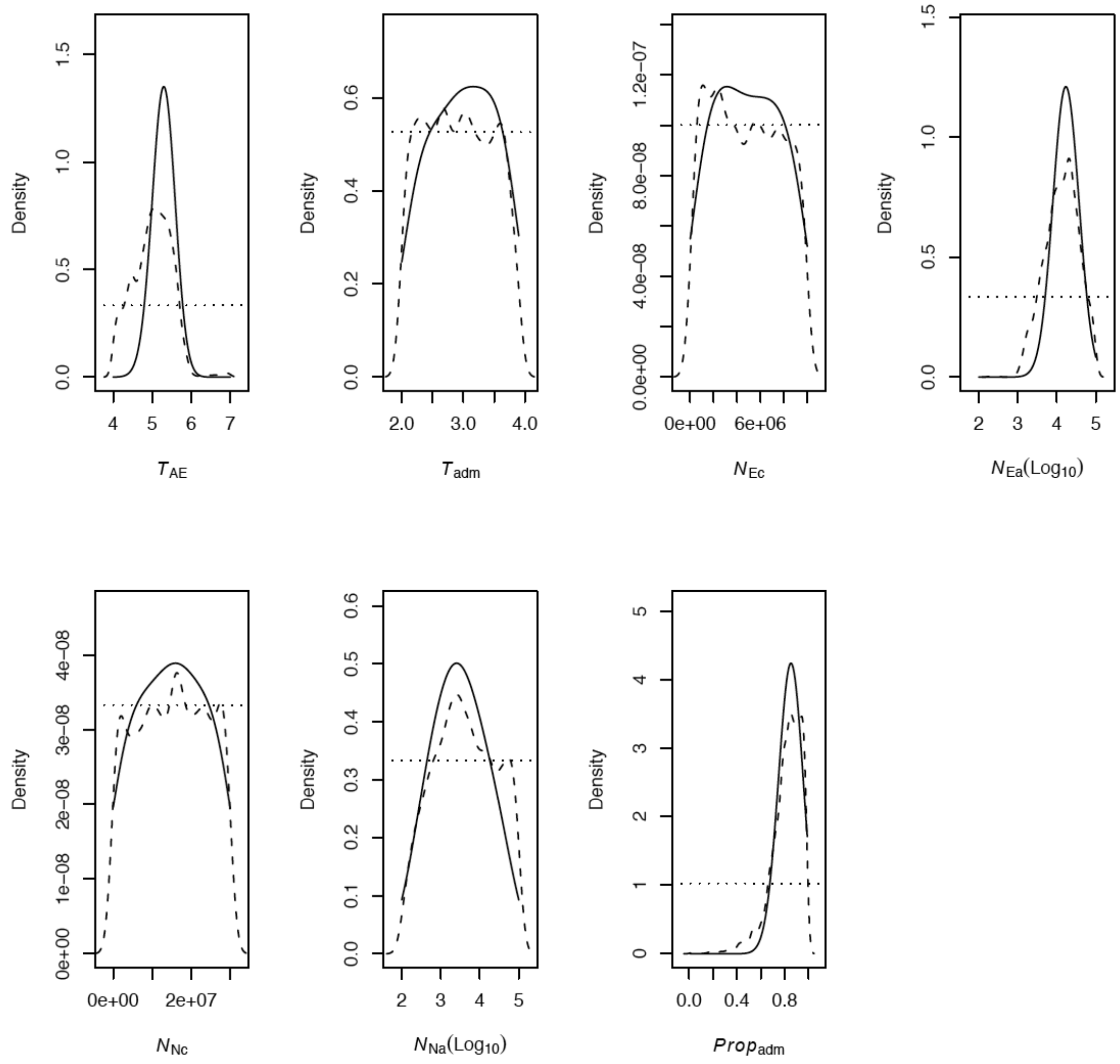

Figure S10 Posteriors of the Admixture model C. Posteriors are represented by the rejection method (dashed line) and the regression method (solid line). Parameter abbreviations are explained in Table 3. Mode and confidence interval for each parameter are shown in Table 5.

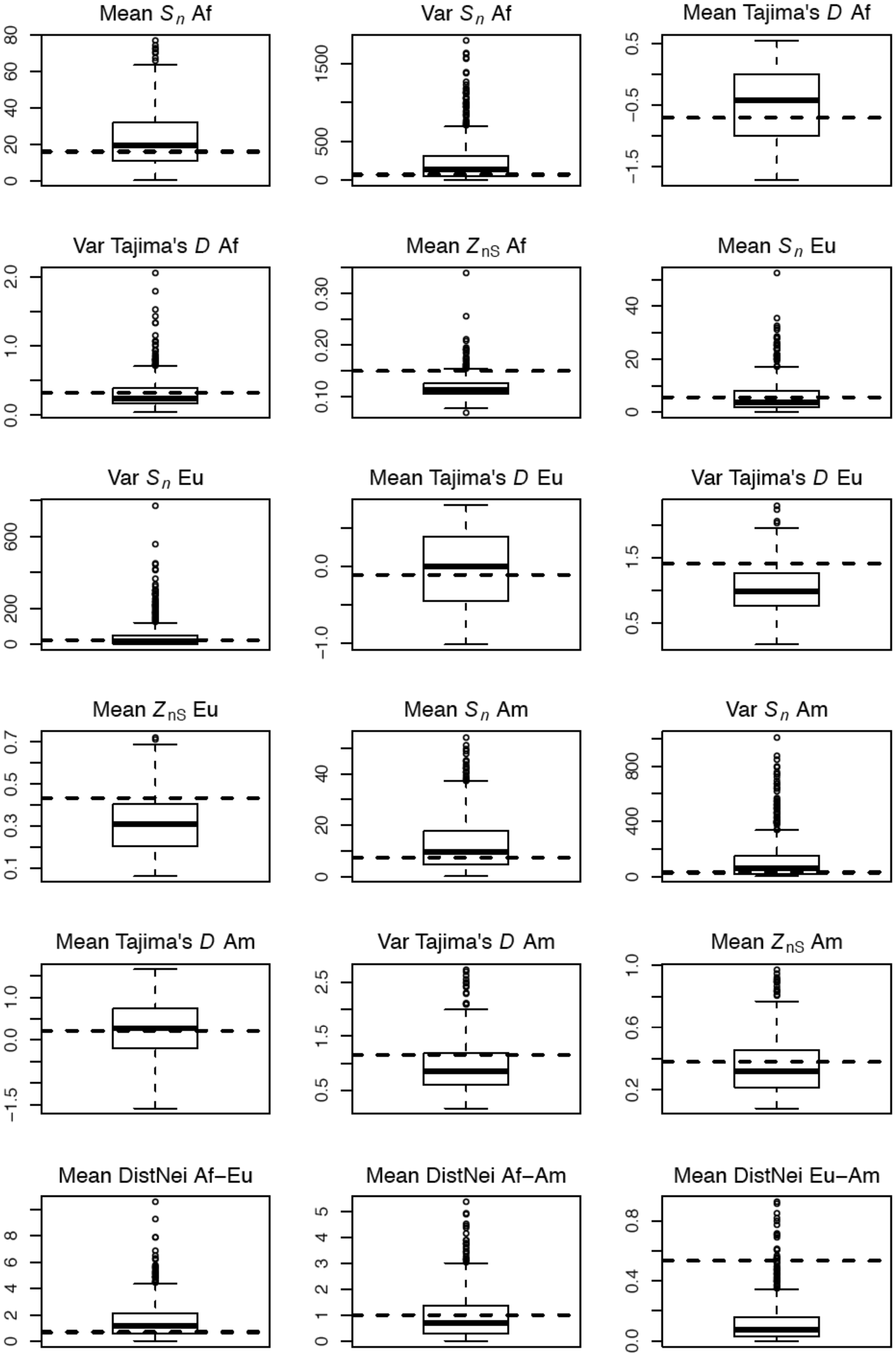

Figure S11 Predicted statistics of model C. Predictions of the mean and variance of S_n, mean and variance of Tajima's D and mean Z_{nS} are shown for each population. Predicted mean Distance of Nei for all pairs of populations are shown as well. Statistics are predicted by sampling parameters from the posterior distributions based on the regression method (see main text for details).

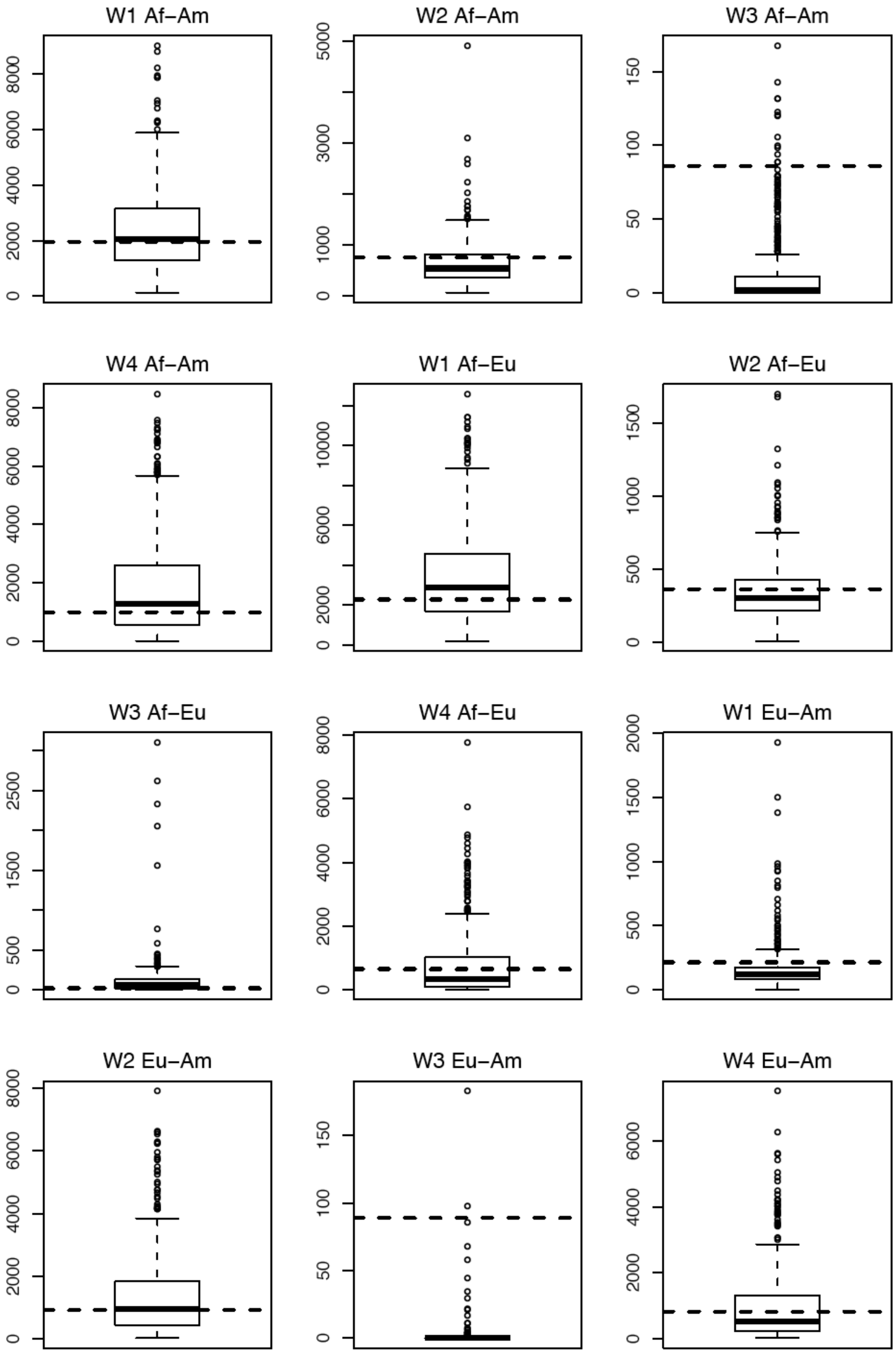

Figure S12 Predicted JSFS of model C. Predictions of each Wakeley-Hey (1997) class are shown. Statistics are predicted by sampling parameters from the posterior distributions based on the regression method (see main text for details).

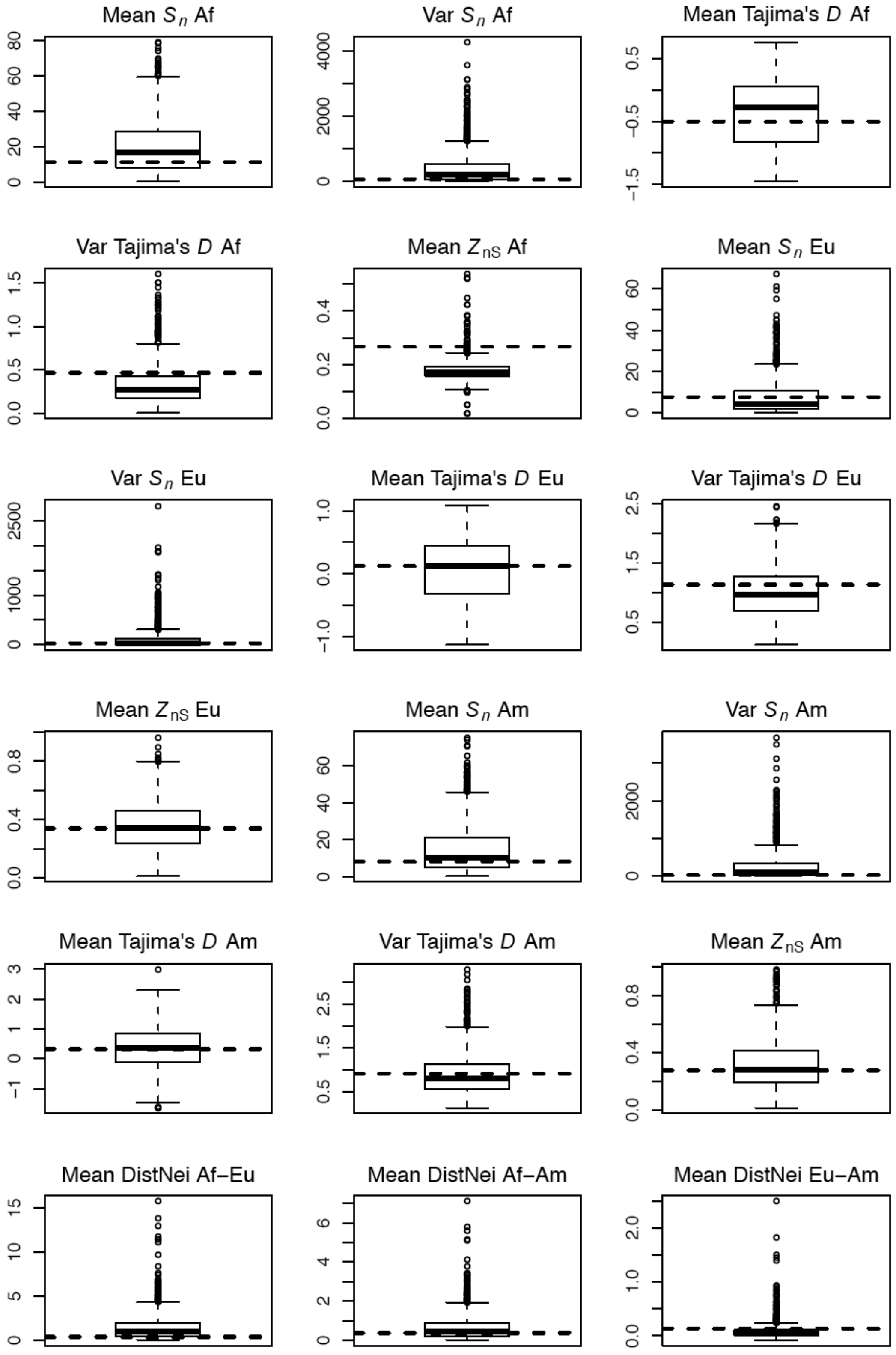

Figure S13 Predicted statistics of model C for autosomal loci (chromosome 3). Predictions of the mean and variance of S_n, mean and variance of Tajima's *D* and mean Z_{nS} are shown for each population. Predicted mean Distance of Nei for all pairs of populations are shown as well. Statistics are predicted by sampling parameters from the posterior distributions based on the regression method (see main text for details).

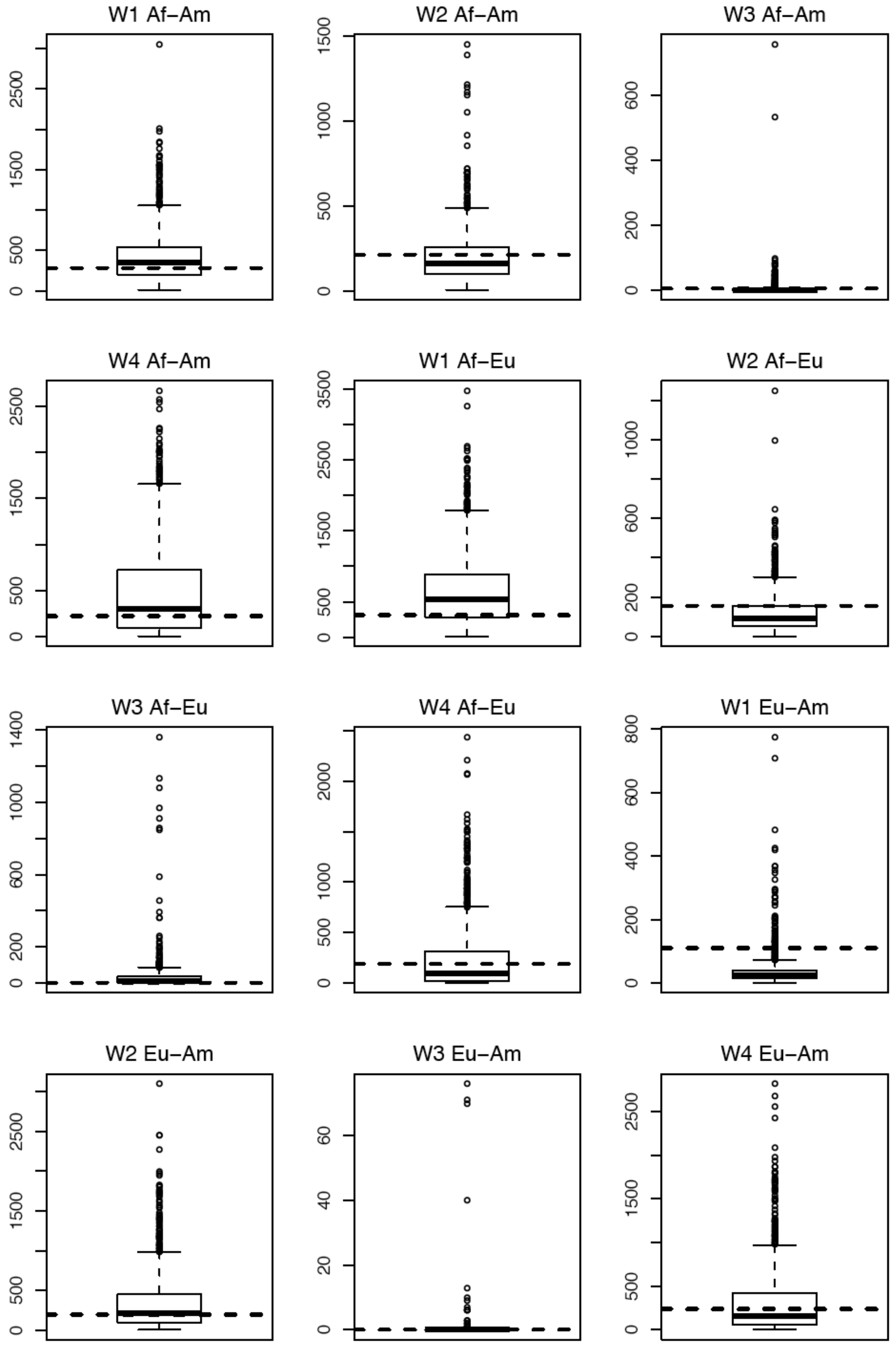

Figure S14 Predicted JSFS of model C for autosomal data (chromosome 3). Predictions of each Wakeley-Hey (1997) class are shown. Statistics are predicted by sampling parameters from the posterior distributions based on the regression method (see main text for details).

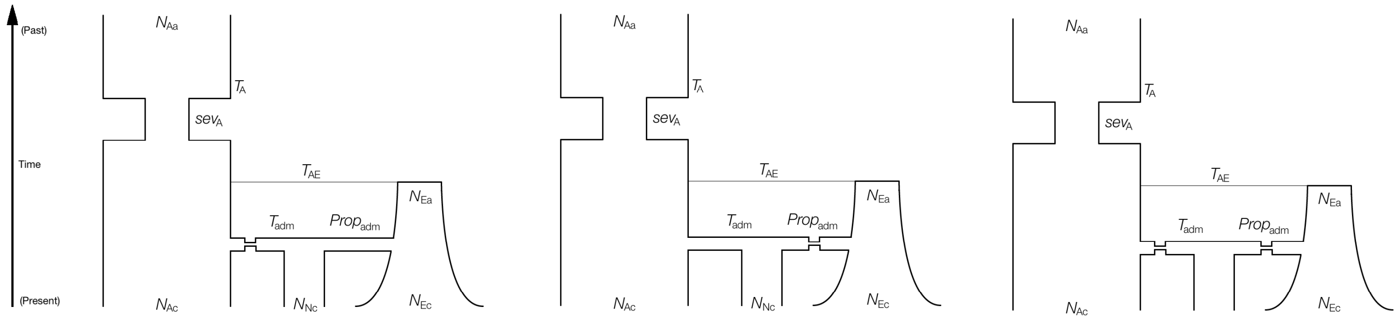

Figure S15 Models C1 (left), C2 (middle) and C3 (right).

Table S1 Parameters and priors used in the one-population models and in models A, B, C, D and E.

Parameter	Prior	Model
Current size Africa: N_{Ac}	unif(1x10^5 ,1x10^7)	Bottleneck and Expansion
Time of bottleneck Africa: T_{A}	unif(1x10^2 ,4x10^5)	Bottleneck and Expansion
Ancient size Africa: N_{Aa}	unif(1x10^5 ,1x10^7)	Bottleneck and Expansion
Severity of bottleneck Africa: sev_{A} (decimal log)	unif(-2,2)	Bottleneck
Time of split Africa-Europe (decimal log): T_{AE}	unif(4,7)	Model A,B,C,D,E
Time of split Europe-North America (decimal log): T_{EN}	unif(4,7)	Model A,B
Time of split Africa-North America (decimal log): T_{AN}	unif(4,7)	Model D,E
Time of admixture (decimal log): T_{adm}	unif(2,4)	Model C
Current size Europe: N_{Ec}	unif(1x10^4 ,1x10^7)	Model A,B,C,D,E
Ancient size Europe (decimal log): N_{Ea}	unif(2,5)	Model A,B,C,D,E
Current size North America: N_{Nc}	unif(1x10^4 ,3x10^7)	Model A,B,C,D,E
Ancient size North America (decimal log): N_{Na}	unif(2,5)	Model A,B,C,D,E
Proportion of European admixture: $Prop_{\mathrm{adm}}$	unif(0.01,0.99)	Model C
Migration rate (decimal log): M	unif(-10,-2)	Model B,E

Table S2 Three-population models covered in this study.

Model	Description	Posterior Probability
A	"No migration" model. Comprises Africa as the ancestral population, colonization of Europe followed by exponential growth, and the colonization from Europe to North America with subsequent exponential growth.	< 0.001
B	"Migration" model, matches Model A but adds an equal migration rate between all populations starting at the colonization time of North America.	< 0.001
C	"Admixture" model, equals the previous models until the North American population is founded through an admixture between Africa and Europe followed by exponential growth in North America.	> 0.999
D	"No migration II" model, North America and Europe split independently from Africa, no migration.	< 0.001
E	"Migration II" model, same as model D plus one single rate of migration starting when the North American population is founded.	< 0.001

Table S3 Mean squared error (MSE) of the ($\log_{10}$) parameter estimates of model C for varying numbers of simulations.

	100000	200000	300000	400000	500000	600000	700000	800000	900000	1000000
N_{Ac}	0.019	0.0101	0.00718	0.00574	0.00443	0.00296	0.00201	0.00149	0.00154	0.00125
T_{adm}	0.242	0.293	0.275	0.306	0.335	0.291	0.309	0.305	0.326	0.322
T_{AE}	0.0693	0.0388	0.0271	0.0214	0.018	0.0147	0.0128	0.011	0.00996	0.00927
T_{A}	0.0447	0.0498	0.043	0.0407	0.0352	0.0318	0.0317	0.025	0.0233	0.0203
sev_{A}	0.0178	0.03	0.03	0.0307	0.03	0.0311	0.0298	0.0291	0.03	0.0326
N_{Aa}	0.00114	0.00243	0.00422	0.00464	0.00661	0.00688	0.00767	0.00835	0.00869	0.00871
N_{Ec}	0.0221	0.0839	0.0831	0.111	0.0804	0.0818	0.069	0.0658	0.0434	0.0366
N_{Nc}	0.000554	0.000369	0.00059	0.000937	0.00103	0.000636	0.00054	0.000316	0.000336	0.000402
N_{Ea}	0.00605	0.00624	0.0075	0.00801	0.0086	0.00942	0.0104	0.011	0.0118	0.0123
N_{Na}	0.471	0.534	0.514	0.505	0.457	0.444	0.443	0.467	0.517	0.509
$Prop_{\text{adm}}$	0.00148	0.00149	0.00169	0.00196	0.00213	0.00222	0.00222	0.00221	0.00219	0.00214