Human Demographic Processes and Genetic Variation as Revealed by mtDNA Simulations

Aida T. Miró-Herrans*1,2 and Connie J. Mulligan2
1Genetics and Genomics Graduate Program, University of Florida
2Department of Anthropology, University of Florida
*Corresponding author: E-mail: aiditere@ufl.edu.
Associate editor: Sohini Ramachandran

Abstract

Humans’ ability for rapid dispersal and adaptation has allowed us to colonize diverse geographic and climatic regions of the planet, creating a complex evolutionary history. This complexity can be understood, at least partially, by modeling the underlying demographic parameters in the evolutionary process. In this study, we analyze a model of human evolution in which population size, gene flow (GF), and time are varied. Specifically, we simulate mitochondrial DNA for 42 demographic scenarios, represented by 42 parameter combinations, to describe the initial dispersal of modern humans out of Africa. The analyses include three values for colonization size (CS; 1%, 10%, and 30% of the African population), seven values for rate of GF (10^{-6}–0.5), and two values for time of colonization (50,000 and 100,000 years ago). We then estimate summary statistics for the simulated data sets to calculate the percent of explained variation by each parameter and to identify which parameter combinations generate distinct differences in genetic variation, that is, which demographic scenarios can be distinguished from each other. On the basis of these results, we make recommendations about which summary statistics to use according to the parameter of interest. Our results show that CS, GF, and their interaction have the largest effect on genetic variation under our model of human evolution. Comparison with empirical data suggests that 1% of the existing African mitochondrial genetic variation left and colonized the rest of the world (i.e., CS = 1%) and bidirectional GF continued at a level of ~10 individuals per generation (i.e., GF = 10^{-3}) after the initial colonization. Our study serves as a model to bridge the gap between the use of simulations for theoretical population genetics and empirical data analysis such as approximate Bayesian computation approaches and is, thus, applicable to the study of molecular evolution in any organism.

Key words: coalescent simulations, human evolution, mtDNA.

Introduction

Evolutionary history can be described as a series of sequential demographic events that created the genetic complexity observed in the organism under study. Reconstructing evolutionary history requires identifying the relevant demographic processes and understanding how these processes have affected patterns of existing genetic variation. To make inferences about human evolutionary processes, such as the first migration of humans out of Africa, it is important to know whether different hypothesized demographic processes are distinguishable based on the genetic variation of the present human population. For instance, can we distinguish between a large and small colonizing population for the initial migration out of Africa? Furthermore, different combinations of demographic parameters, such as population size and gene flow, can interact to generate similar patterns of genetic variation. By looking at small changes in demographic parameters, in combination with each other, we can determine the influence of these parameters on patterns of genetic variation.

With the growth in computational power, simulations now allow us to generate multiple sets of genetic data for complex evolutionary processes. We can compare the simulated data sets to each other to determine how genetic variation changes as demographic parameters change (Carvajal-Rodríguez 2008) and identify which parameter interactions cause detectable differences in genetic variation. Although many studies have compared simulations of evolutionary processes to empirical data to make inferences about the empirical data (Fagundes et al. 2007; Deshpande et al. 2009; Lohmueller et al. 2009; Gronau et al. 2011; Veeramah et al. 2012), few studies have used simulations to investigate the effects of demographic parameters and their interactions on genetic variation (Calafell et al. 2001). Comparing simulated demographic scenarios can help us determine which demographic parameters merit more attention because of their increased effect on genetic variation and can direct the investigation to questions focused on the parameters with greatest effect. Comparisons of simulated scenarios also allow identification of the parameter combinations (and by inference, the demographic scenarios) that can be distinguished from each other based on the genetic variation of each scenario.

For instance, comparing simulated scenarios could improve our understanding of the critical period in human history when anatomically modern humans left Africa and colonized the rest of the planet. Although many studies have focused on estimating specific values for parameters of
interest for the colonization of humans out of Africa (e.g., Fagundes et al. 2007, DeGiorgio et al. 2009, Gronau et al. 2011), the large variances for some of these estimates suggest it would be useful to better understand how specific values for each parameter, and their interactions, affect genetic variation. Three parameters of particular interest for the colonization of humans out of Africa are 1) the size of the colonizing population, 2) the timing of the event, and 3) the amount of subsequent gene flow into and out of Africa. Examining the interaction of these primary parameters using estimates drawn from the literature should give insight into which of the parameters has a larger effect on genetic variation and whether increased efforts to refine the value will lead to increased resolution of other parameters of interest.

Demographic parameters have generally been inferred from statistics that summarize the patterns of genetic variation. Although Lohse and Kelleher (2009) show that likelihood methods provide better estimates of demographic parameters, recent studies show that it is still common practice to use summary statistics (Keinan et al. 2008; Bustamante and Ramachandran 2009; Deshpande et al. 2009). The use of summary statistics has become increasingly popular in methods of Bayesian inference, such as approximate Bayesian computation (ABC) (Beaumont et al. 2002; Marjoram and Tavaré 2006; Beaumont 2010). In brief, the ABC approach compares summary statistics calculated from an empirical data set with summary statistics calculated from simulated scenarios that serve as hypotheses to explain the empirical data. For ABC, and other approaches, it is necessary to know whether different demographic scenarios lead to different summary statistic values, thus reflecting the effect of different demographic parameter combinations on genetic variation. Furthermore, it is essential to determine, and is largely lacking in the current literature, which summary statistics are most informative for a parameter or evolutionary process of interest (Hickerson et al. 2006).

In this study, we simulate mitochondrial DNA (mtDNA) nucleotide sequences for 42 alternative demographic scenarios describing the colonization of modern humans out of Africa. The diverse set of parameter combinations allows us to evaluate the influence of these parameters on genetic variation. Three parameters of primary interest were varied in these scenarios; colonization size (CS), rate of gene flow (GF) between African and non-African populations, and the time of colonization (TC) event. Values for these parameters were chosen from the literature to represent realistic demographic scenarios that could have produced the current human genetic variation. Twelve summary statistics were calculated for each of 42 different demographic scenarios. The summary statistics were used to 1) detect differences between demographic scenarios and determine which scenarios could be distinguished from each other, 2) determine the effects of particular demographic parameters (CS, GF, and TC) on genetic variation, and 3) identify the informativeness of different summary statistics on genetic variation.

Materials and Methods

Models

Forty-two scenarios were modeled to describe the demographic process for the initial colonization of modern humans out of Africa. In this model (fig. 1), two populations split at one of two possible colonization times, with the colonizing population having one of three possible sizes, followed by seven possible proportions of bidirectional GF between the populations. The values for colonization time (TC) were 100,000 and 50,000 years ago, the earliest and most recent estimates for movement of modern humans out of Africa, respectively (Klein 1998; Salas et al. 2002; Macaulay 2005; Mellars 2006; Gonder et al. 2007; Behar et al. 2008). The values for CS were estimated as a proportion of the African population and set at 1%, the lowest estimated value of migrants (Fagundes et al. 2007; Atkinson et al. 2008); 30%, the highest estimated value (Relethford and Harpending 1995; Tenesa et al. 2007); and 10% as an intermediate value (Gronau et al. 2011). The values for GF were $10^{-6}$, $10^{-5}$, $10^{-4}$, $10^{-3}$, $10^{-2}$, 0.1, and 0.5 proportion of migrants (m) per generation from each population. These values for GF were selected to begin at $N_m \ll 1$ and increase by single orders of magnitude until $N_m \gg 1$ to reflect highly structured populations and highly panmictic populations, respectively. The initial population size was fixed at 10,000 (Fagundes et al. 2007; Atkinson et al. 2008) with constant population size.

Simulations

One thousand coalescent simulations for each of the 42 demographic scenarios were generated using SIMCOAL2 (Laval and Excoffier 2004). One hundred sequences of human mtDNA data were simulated for each population with a coding region of 15,446 nucleotides (nt) and a substitution rate of 2.0 substitutions per site per year (Ingram et al. 2000; Atkinson et al. 2008) and a control region of 1,123 nt and a substitution rate of $4.7 \times 10^{-7}$ substitutions per site per year (Howell et al. 2003).

Summary Statistics

Twelve summary statistics (table 1) were calculated for each of the 42,000 simulated mtDNA data sets to capture the genetic variation of the 42 different scenarios. $F_{st}$ and $\Phi_{ct}$ were calculated between the two populations with ARLEQUIN 3.11 (Excoffier et al. 2005). Tau-hat ($\hat{\tau}$) was calculated from the mismatch distribution of the simulated mtDNA coding region with R (R Development Core Team 2010). Number of segregating sites ($S$), Watterson’s $\theta$ ($\theta_W$), nucleotide diversity ($\pi$), Ramos-Onsins and Rozas’ $R_2$, Tajima’s $D$ ($T_D$), number of singleton sites (NSS), number of haplotypes (# Hap), number of singletons (# Single), and homzygosity (Hmzy) were calculated with Sample_stats, a version of the Sample_stats utility distributed with Hudson’s (2002) MS, modified for DNA sequence data (available at http://github.com/ryanraaum/samplestats). The code to calculate Ramos-Onsins and Rozas’ $R_2$ was incorporated.
### Table 1. Summary Statistics Analyzed and Their Definition.

<table>
<thead>
<tr>
<th>Summary Statistic</th>
<th>Reference</th>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{st}$</td>
<td>Wright (1951)</td>
<td>$F_{st}$</td>
<td>Proportion of genetic diversity due to allele differences among populations</td>
</tr>
<tr>
<td>$\Phi_{st}$</td>
<td>Excoffier et al. (1992)</td>
<td>$\Phi_{st}$</td>
<td>Proportion of genetic diversity due to haplotype differences among populations</td>
</tr>
<tr>
<td>Segregating sites</td>
<td>Fu (1995)</td>
<td>$S$</td>
<td>Number of polymorphic sites</td>
</tr>
<tr>
<td>Watterson’s theta</td>
<td>Watterson (1975)</td>
<td>$\theta_W$</td>
<td>$S$ corrected for number of samples</td>
</tr>
<tr>
<td>Nucleotide diversity</td>
<td>Nei (1987)</td>
<td>$\pi$</td>
<td>Average number of nucleotide differences per site</td>
</tr>
<tr>
<td>Ramos-Onsins and Rozas’ $R_2$</td>
<td>Ramos-Onsins and Rozas (2002)</td>
<td>$R_2$</td>
<td>Test of neutrality based on difference between number of singleton mutations and $\pi$</td>
</tr>
<tr>
<td>Tajima’s $D$</td>
<td>Tajima (1989)</td>
<td>$T_D$</td>
<td>Test of neutrality based on difference between $S$ and pair-wise differences</td>
</tr>
<tr>
<td>Number of singletons sites</td>
<td>Balding et al. (2003)</td>
<td>NSS</td>
<td>Number of sites where only one individual has a different allele</td>
</tr>
<tr>
<td>Tau hat</td>
<td>Rogers (1995)</td>
<td>$\hat{\tau}$</td>
<td>Estimate of time measured in mutational units</td>
</tr>
<tr>
<td>Number of haplotypes</td>
<td>Balding et al. (2003)</td>
<td># Hap</td>
<td>Number of different unique allele combinations</td>
</tr>
<tr>
<td>Number of singletons</td>
<td>Balding et al. (2003)</td>
<td># Single</td>
<td>Number of haplotypes that appear only once in the sample</td>
</tr>
<tr>
<td>Homozygosity</td>
<td>Balding et al. (2003)</td>
<td>Hmzy</td>
<td>Probability of two samples in a population having the same haplotype</td>
</tr>
</tbody>
</table>

### Results

#### Partitioning of Genetic Variation by Demographic Parameters

ANOVAs were used to partition the genetic variation summed across all 42,000 simulated data sets into the tested parameters and their interactions and to determine which parameters and interactions had a significant effect in explaining differences between the 42 demographic scenarios as reflected in each summary statistic (table 2). For parameters and interactions that were significant in explaining variation, the actual value of explained variance was then calculated to identify how variation was partitioned among the parameters and interactions within each summary statistic.

All parameters and interactions were significant ($P$ value $< 1.0 \times 10^{-6}$) in explaining the differences in genetic variation between the 42 demographic scenarios using the following summary statistics: $F_{st}$, $\Phi_{st}$, $S$, $\theta_W$, $\pi$, $T_D$, $R_2$, and $\hat{\tau}$ (table 2). TC and TC interactions (i.e., $TC \times CS$ and $TC \times GF$) were not significant for NSS, # Hap, # Single, and Hmzy.

CS, GF, and their interaction (i.e., $CS \times GF$) were significant across all summary statistics and yielded the highest percent of explained variance. CS explained the most variation, ranging from 2.4% to 96.4%, depending on the summary statistic. This was followed by the interaction between CS and GF (3.1–43.5%) and then GF (0.8–86.8%). TC and its interactions with GF and CS explained only a small percent of the variation ($TC: 0.1–1.9\%; TC \times CS: 0.4–5.3\%; and TC \times CS: 0.1–0.8\%).

#### Comparison of Summary Statistics

Percent of explained variation was also compared across summary statistics to determine how each summary statistic partitioned variation into the investigated parameters and interactions (table 2). $F_{st}$ and $\Phi_{st}$ partitioned more variation in GF relative to the other summary statistics, with percent of explained variance of 85.6% and 86.8%, respectively. $\hat{\tau}$, # Hap, # Single, and Hmzy partitioned the most variation for CS relative to the other summary statistics, with values of

![Fig. 1. Alternative scenarios for initial colonization of modern humans out of Africa. Scenarios include all combinations for time of colonization (TC) and colonization size (CS). Each scenario is modeled with seven values of bidirectional gene flow (where GF = 10, 0.1, and 0.5 proportion of migrants per generation).](image-url)

Statistical Analysis

A multifactorial analysis of variance (ANOVA) including all three parameters of interest (CS, GF, and TC) was performed for each summary statistic using R (R Development Core Team 2010). The percent of variation explained by each parameter, and parameter interaction was estimated from the additive component of variance, which was in turn calculated from the expected mean square value. Pair-wise comparisons (Tukey’s tests) were performed for each pair of the 42 demographic scenarios, for a total of 861 comparisons for each of the 12 summary statistics.

with permission from Mlcoalsim (Ramos-Onsins and Mitchell-Olds 2007).
Although developed as a measure to estimate time of a demographic event discussed later (fig. 2 and supplementary fig. S2, Supplementary summary statistics for all 42 demographic scenarios are discussed later, particularly \( C_2 \), although differences in \( \tilde{\tau} \) between demographic scenarios are less distinct, despite \( \tilde{\tau} \) being developed as a measure to estimate time of a demographic event (Rogers 1995). Although \( \Phi_{st} \) and \( F_{st} \) also partition the majority of variation into a single category (GF), these summary statistics show more representation by TC and TC interactions (i.e., TC \( \times \) CS and TC \( \times \) GF) than the other summary statistics: box plots depicting the range of estimates of eight summary statistics for all 42 demographic scenarios are discussed later (fig. 2 and supplementary fig. S2, Supplementary Material online).

Several summary statistics (e.g., NSS, TD, and \( R_2 \)) are significantly different from each other, whereas TC is significantly different for low GF categories (10\(^{-6}\) and 10\(^{-5}\)). \( S \) and \( \theta_W \) are related measurements that also show similar trends in the box plots. Within the CS category of 1%, high and low levels of GF are clearly differentiated (fig. 2b and supplementary fig. S2b and table S1, Supplementary Material online), but the difference between GF categories diminishes with increasing CS. With respect to TC, 50 kya and 100 kya produce significantly different estimates from each other when CS = 10% for GF > 10\(^{-3}\) and when CS = 1% + GF = 10\(^{-3}\). Estimates of \( \pi \) show a similar pattern as \( S \) and \( \theta_W \) estimates (fig. 2c), with smaller differences in \( \pi \) estimates between the scenarios, except that TC produces significantly different \( \pi \) estimates for GF categories of low values within CS = 1%. Estimates of \( R_2 \) and TD show similar patterns in the box plots (fig. 2d and supplementary fig. S2c, Supplementary Material online). For CS = 10%, categories of GF > 10\(^{-3}\) produce significantly greater estimates than categories of GF of 10\(^{-3}\) and less, whereas TC produces differences similar to those of \( \pi \) (supplementary table S1, Supplementary Material online). Estimates of NSS are significantly different for high and low levels of GF for CS = 1% and CS = 10% (fig. 2e and supplementary table S1, Supplementary Material online). TC only produces significantly different estimates for CS = 1% for GF = 10\(^{-3}\).

Most of the discussed summary statistics (except \( F_{st} \) and \( \Phi_{st} \)) show that low GF categories (<10\(^{-4}\)) for CS = 1% and CS = 10% have similar summary statistics relative to high GF categories for CS = 30%. \( F_{st} \) and \( \Phi_{st} \) show a different pattern, thus providing insight into the effects of GF by reflecting the diversity between two migrating populations instead of the overall diversity among the two populations, which the other summary statistics reflect. \( F_{st} \) and \( \Phi_{st} \) show a similar pattern for the three CS categories where \( F_{st} \) and \( \Phi_{st} \) estimates decrease as GF category values increase. The discussed summary statistics also show that, in general, TC produced significantly different estimates for GF categories equal or <10\(^{-5}\) but not for high GF categories. The demographic scenarios with CS = 1% and GF = 10\(^{-3}\) were generally distinct from each other and all other scenarios.
**FIG. 2.** Box plots of estimates of five summary statistics that partition genetic variation more equally between CS, GF, and CS × GF than seen in the other summary statistics. Panels (a) $F_{ST}$, (b) $S$, (c) $\tau$, (d) $R_2$, and (e) NSS. Each panel shows the distribution of 42,000 estimates of the summary statistic calculated for each of 1,000 data sets simulated under 42 parameter combinations as listed at the bottom. The gray bar in panel (a) shows the range of empirical $F_{ST}$ values (0.141–0.235) between African populations versus European and Asian populations (Bowcock et al. 1991).
Discussion

Demographic Parameters

Our results illustrate that migration, whether represented as CS or GF, shows the largest effect on human genetic variation over the time period in which humans colonized the planet. Specifically, our simulations indicate that CS has the largest influence on patterns of genetic variation (percent of variance explained averaged over all summary statistics), and CS, GF, and CS × GF explain most of the genetic variation that has arisen in humans, as simulated in this study. Interactions between CS and GF have varying effects on patterns of genetic variation. For example, the majority of box plots (fig. 2) between CS and GF have varying effects on patterns of genetic variation across all CS categories. This supports findings such as those of Kitchen et al. (2008), where reducing GF as low as zero produced a much larger CS for peopling of the Americas, in contrast to Hey’s (2005) results where CS was 100-fold smaller with much higher levels of GF. However, the extreme values tested here show that large CS with low GF (< 10^{-3}) creates a very different effect relative to small CS and high GF. This can be explained because large CS with low GF leads to high genetic variation as it increases the difference between the populations, whereas small CS and high GF lowers genetic variation as it decreases the difference between the migrating populations.

Summary Statistics

This is one of the first studies to investigate the informative- ness of specific summary statistics in the inference and comparison of demographic processes (Hickerson et al. 2006; Sefc et al. 2007). Specifically, we were interested in determining the extent to which different summary statistics can distinguish between the modeled scenarios. Although one summary statistic alone cannot distinguish between all the scenarios, our results suggest that combining several summary statistics, selected based on the parameters of interest, will allow better resolution when comparing empirical data with simulated data.

Our results clearly show that the summary statistics differentially explain variance depending on the demographic parameter (table 2). In general, summary statistics in which percent of explained variation is distributed across multiple parameters and interactions (such as $S$, $\theta_W$, $\pi$, $T_D$, $R_D$, and NSS) are able to distinguish between more demographic scenarios. In contrast, summary statistics where the majority of variation is concentrated in one main parameter (such as $\hat{r}$, # Hap, # Single, and Hmzy) are useful for studies that focus on one parameter but have more limited utility to distinguish between different demographic scenarios. It is important to note that some demographic scenarios produce genetically similar results that cannot be differentiated regardless of the summary statistic used. This similarity suggests that some evolutionary processes and specific questions, such as the timing of the first human migration out of Africa, may not be resolved based on mtDNA data.

<table>
<thead>
<tr>
<th>Parameter of Interest</th>
<th>Optimal Summary Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colonization size</td>
<td># Single</td>
</tr>
<tr>
<td>Gene flow</td>
<td>$F_a$</td>
</tr>
<tr>
<td>Time</td>
<td>$F_a$</td>
</tr>
<tr>
<td>Colonization size × gene flow</td>
<td>$S$, $\theta_W$, $R_D$, $T_D$</td>
</tr>
<tr>
<td>Gene flow × time</td>
<td>$F_a$, $R_D$, $T_D$</td>
</tr>
</tbody>
</table>

Summary statistics were chosen based on an assessment that identified the statistic with the highest percent of variation explained (table 2) and most extreme differentiation of parameter combinations (fig. 2 and supplementary figs. S1 and S2, Supplementary Material online).

A limitation in methods that compare empirical data with simulated data, such as ABC approaches, is that the analyses require a small number of summary statistics to avoid the situation in which so many simulated scenarios are rejected that it is impossible to make any conclusions about the empirical data (Beaumont et al. 2002; Hamilton 2005; Wegmann et al. 2009). Our results show similarity in box plot profiles across some pairs of summary statistics, most likely because they reflect a similar aspect of the genetic variation, for example, $S$ and $\theta_W$. This similarity offers the possibility to reduce the number of summary statistics used when comparing empirical and simulation data. We make specific recommendations on the optimal summary statistics to use based on the parameters and questions of interest for studies of evolutionary history where CS and GF have played a dominant role, such as in the current analysis (table 3).

Relevance to Human Evolution

A goal in evolutionary studies is to identify changes in demographic parameters given the observed genetic variation. In most cases, it is unclear how small changes in the parameters of interest will influence genetic variation, for example, how will small changes in GF influence observed levels of genetic variation. We use a model of modern human migration out of Africa, with changes in CS, TC, and GF, to investigate the effect of demographic changes on genetic variation.

To address the extent of GF between African and non-African populations after the initial migration out of Africa, values for GF were selected, such that GF < $10^{-4}$ represents population substructure, GF = $10^{-4}$ represents migration equilibrium ($N_e m \approx 1$), and GF > $10^{-4}$ represents panmixia. The box plots of the demographic scenarios (fig. 2 and supplementary fig. S2, Supplementary Material online) illustrate that when CS = 1% (and in some cases when CS = 10%), GF = $10^{-3}$ appears as a transition point where genetic variation decreases significantly (in a sigmoidal curve) as GF increases. Our pair-wise comparisons (supplementary table S1, Supplementary Material online) show that summary statistic estimates when GF < $10^{-3}$ are similar to each other, but have significantly greater estimates when GF > $10^{-3}$, which also have similar estimates to each other. The sharp transition in genetic variation from GF of $10^{-4}$ to $10^{-3}$ reveals a rapid breakdown of population substructure.
to panmixia within only an order of magnitude increase in GF. Thus, it should be possible to distinguish between scenarios on either side of the transition point, that is, $GF = 10^{-3}$, but much more difficult to distinguish scenarios within high or low GF, that is, $GF > 10^{-3}$ or $GF < 10^{-3}$.

$F_{st}$ has been commonly used to measure genetic differences between populations as a means to reflect GF. Our results allow us to assess the amount of GF required to generate different $F_{st}$ estimates. When we compare our simulated $F_{st}$ estimates with $F_{re}$ estimates of African populations versus European populations (0.141) or versus Asian populations (0.235) (Bowcock et al. 1991), we observe areas of overlap between the simulated and empirical estimates of $F_{st}$ (0.141–0.235) (fig. 2a and supplementary table S2, Supplementary Material online). For CS categories of 1% and 10%, we observe the most overlap with empirical $F_{st}$ estimates at $GF = 10^{-3}$, whereas for CS = 30%, the maximum overlap occurs at $GF = 10^{-4}$. Simulations of bottlenecks show that a CS of 30% is not accompanied by a reduction in genetic diversity as we see empirically between African and non-African populations (Relethford 2001; Ramachandran 2005), suggesting that CS = 1% and CS = 10% are better estimates of the size of the colonizing population. Our results show that there is almost twice as much overlap with empirical $F_{re}$ estimates when CS = 1% (fig. 2a and supplementary table S2, Supplementary Material online), providing the strongest support for a scenario in which the migrating population carried 1% of existing African mitochondrial genetic variation (i.e., CS = 1%), and both African and non-African populations experienced bidirectional GF of $10^{-3}$.

It is interesting to speculate on what these values mean in terms of actual individuals, particularly with respect to GF and CS. A GF of $10^{-3}$ represents 10 individuals moving per generation in both directions. Although this value seems large, it may be reasonable that an average of 10 individuals per generation migrated between African and non-African populations, particularly immediately after migration out of Africa when the populations were still geographically close. Deshpande et al. (2009) modeled a serial founder population history for migration out of Africa and colonization of Eurasia, and they identified 0.01 (100 individuals) as the maximum bidirectional exchange rate between adjoining populations. Furthermore, 1% of the population leaving Africa (CS = 1%) to colonize the world seems reasonable under our model of a panmictic population. However, Africa was potentially highly structured before the exit of modern humans out of Africa (Campbell and Tishkoff 2008; Veeramah et al. 2012). A high degree of structure could imply that East Africa, where humans first emigrated to the rest of the world, was a distinct subpopulation, and the CS leaving from this area could be larger than 1%, reflecting published CSSs of 9–18% (Deshpande et al. 2009; Gronau et al. 2011). Future studies should address the effect of African population substructure on the genetic variation of non-African populations, specifically how African substructure affects our ability to accurately model the evolutionary processes that gave rise to non-African genetic diversity.

Although changes in CS and GF create detectable differences in genetic variation in our study, TC shows relatively little effect on patterns of genetic variation. For most of the simulations, demographic scenarios of 50 kya are not distinguishable from those of 100 kya. This result is probably due to the relatively recent times chosen in the current analysis, that is, 50 kya and 100 kya, which were chosen to reflect relevant times for human migration. Although time is an important factor affecting genetic diversity in general, anatomically modern human’s short existence likely explains why colonization time has not played an important role in human genetic diversity. Notably, for values that may most accurately reflect human demographic history (CS = 1% and GF = $10^{-3}$), six summary statistics can distinguish between TC of 50 kya and 100 kya (fig. 2b–e and supplementary fig. S2b and c, Supplementary Material online), suggesting time is only distinguishable under particular conditions.

Our findings demonstrate the utility of comparing simulated demographic scenarios to understand the effect of demographic parameters on genetic variation. Our results show that different demographic parameters have varying effects on contemporary genetic variation. The parameters that generate a larger difference in genetic variation obscure the differences in genetic variation caused by other parameters, such that scenarios with differences in the lesser effect parameters cannot be distinguished from one another. In the case of humans, our comparisons reveal that migration (CS or GF) has such a large effect on genetic variation that scenarios with different times for an event are less likely to be distinguishable, particularly with mtDNA. A better understanding of the three particular parameters addressed in this study (CS, GF, and TC) allows other demographic parameters to be addressed. For example, by narrowing potential values of CS to 1% and GF to $10^{-3}$, new scenarios can be generated to assess the effect of other parameters of interest, such as population growth or the occurrence of GF at specific times. In a similar fashion, comparisons of simulated scenarios can provide insight into the evolutionary history of any system. With increased understanding of the effects of demographic parameters on genetic variation, more accurate inferences about evolutionary histories will be possible.

**Supplementary Material**

Supplementary tables S1 and S2 and figures S1 and S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The authors thank Michael Miyamoto, David Reed, David Hughes, Andrew Kitchen, and two anonymous reviewers for insightful discussion and comments on this manuscript. They thank Ryan Raaum for help with calculating summary statistics and modifications to sample_stat3. This work was supported by a National Science Foundation Graduate Research Fellowship to A.T.M.-H. Part of this work was done using the resources of the University of Florida.
Genetic Institute’s Fisher Bioinformatics Cluster. All research was performed at the University of Florida.

References


