DOI: 10.1534/genetics.108.099275

Methods for Human Demographic Inference Using Haplotype Patterns From Genomewide Single-Nucleotide Polymorphism Data

Kirk E. Lohmueller,*,†,[1] Carlos D. Bustamante† and Andrew G. Clark*

**Department of Molecular Biology and Genetics and †Department of Biostatistics and Computational Biology, Cornell University, Ithaca, New York 14853*

Manuscript received December 5, 2008
Accepted for publication February 19, 2009

ABSTRACT

We propose a novel approximate-likelihood method to fit demographic models to human genomewide single-nucleotide polymorphism (SNP) data. We divide the genome into windows of constant genetic map width and then tabulate the number of distinct haplotypes and the frequency of the most common haplotype for each window. We summarize the data by the genomewide joint distribution of these two statistics—termed the HCN statistic. Coalescent simulations are used to generate the expected HCN statistic for different demographic parameters. The HCN statistic provides additional information for disentangling complex demography beyond statistics based on single-SNP frequencies. Application of our method to simulated data shows it can reliably infer parameters from growth and bottleneck models, even in the presence of recombination hotspots when properly modeled. We also examined how practical problems with genomewide data sets, such as errors in the genetic map, haplotype phase uncertainty, and SNP ascertainment bias, affect our method. Several modifications of our method served to make it robust to these problems. We have applied our method to data collected by Perlegen Sciences and find evidence for a severe population size reduction in northwestern Europe starting 32,500–47,500 years ago.

A major goal of evolutionary genetics is to infer the demographic history of a population. This is traditionally done by fitting a population genetic model to sequence data taken from a sample of individuals. The population genetic model often includes parameters allowing for changes in population size or population structure with or without migration. Such parameters are interesting in their own right, but are critical to define a proper "null model" that can be used to find "unusual" genes that may be targets of positive or negative selection (Jensen *et al.* 2005). Additionally, a proper demographic model is important for assessing genomewide patterns of positive and negative selection (Boyko *et al.* 2008; Lohmueller *et al.* 2008).

Methods have been developed that make full use of sequence data to infer demographic parameters (Griffiths and Tavaré 1994; Kuhner *et al.* 1995). These methods are computationally intensive and are impractical for all but the smallest data sets. Thus, researchers have turned to methods based on summary statistics (reviewed in Marjoram and Tavaré 2006). Summary statistics can be quickly calculated from the data and then be used to infer model parameters using either a likelihood or approximate Bayesian computation (ABC) framework (for example, Wall 2000a; Fagundes *et al.* 2007). The key for successful application of this approach is to find summaries of the data that contain enough information about the demographic parameters of interest. One of the most successfully used summary statistics for population genetic inference, the site frequency spectrum (SFS) (Nielsen 2000; Adams and Hudson 2004; Caicedo *et al.* 2007; Hernandez *et al.* 2007b), is a sufficient statistic for the full data if the single-nucleotide polymorphisms (SNPs) are unlinked. However, in reality, all SNPs are not unlinked. The amount of information in the data lost by ignoring the correlations among SNPs, or linkage disequilibrium (LD), in demographic inference is an open question, but recent theoretical work suggests that it may be nonnegligible (Myers *et al.* 2008).

An additional complication to using the SFS for demographic inference is that many genomewide genetic variation data sets in humans contain SNPs that were discovered through sequencing a small number of individuals. The discovered SNPs were then genotyped in a larger set of individuals, sometimes in a different population than was used for SNP discovery. Since this SNP discovery process will lead to preferential sampling of intermediate-frequency alleles, the SFS computed from SNP genotype data will differ substantially from the true SFS (Nielsen *et al.* 2004; Clark *et al.* 2005). Progress has been made to analytically correct the SFS

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.108.099275/DC1.

[1] *Corresponding author:* 102F Weill Hall, Cornell University, Ithaca, NY 14853. E-mail: kel38@cornell.edu

for ascertainment bias when the SNP discovery process is known (Nielsen *et al.* 2004), but often this is not the case. More problematic is the situation where SNPs were discovered by resequencing individuals in one population, but then are genotyped in a second population. It remains an open question as to how well the SNPs discovered in the first population are representative of genetic diversity in the second population. Several authors have suggested that statistics based on combinations of multiple SNPs, or haplotypes, will be less susceptible to ascertainment bias than single-SNP frequencies or heterozygosities (Conrad *et al.* 2006). However, while this suggestion is encouraging, as yet there has not been extensive investigation into the precise ascertainment conditions under which this is true.

It is known that haplotype patterns and LD can be affected by both recombination and demographic history (Pritchard and Przeworski 2001), making these measures useful statistics for inference. Many recent studies have assumed a demographic model (often the standard neutral model) and then used either LD or haplotype patterns to estimate recombination rates (Wall 2000a; Hudson 2001; Li and Stephens 2003; McVean *et al.* 2004; Myers *et al.* 2005). Other studies have taken the opposite approach and assumed that the recombination rate is known and then used LD or haplotype patterns to estimate demographic parameters (Reich *et al.* 2001; Innan *et al.* 2005; Voight *et al.* 2005; Schaffner *et al.* 2005; Leblois and Slatkin 2007; Tenesa *et al.* 2007). The way in which haplotype information has been used for demographic inference is quite variable among studies. For example, Reich *et al.* (2001) examined how well several different demographic models predicted the observed decay of pairwise LD in humans, rather than estimating the model parameters. Francois *et al.* (2008) and Thornton and Andolfatto (2006) used ABC to estimate model parameters in Arabidopsis and Drosophila, respectively. However, summaries based on the distribution of the number of haplotypes were only one of several summary statistics considered, and it is unclear how much information came from the haplotype information *vs.* the other single-SNP diversity measures. While Anderson and Slatkin (2007) and Leblois and Slatkin (2007) developed methods that use haplotype information exclusively to fit a population split followed by growth model, their model is quite restrictive and allows inference only of one free parameter, the number of founding lineages. Thus, there has not been a systematic investigation as to the utility of haplotype information for inference in general, parameter-rich models, such as those involving population expansions and bottlenecks.

In this article we propose an approximate-likelihood method to estimate parameters in complex demographic models from genomewide SNP genotype (rather than full resequencing) data, using the joint distribution of the number of haplotypes and frequency of the most common haplotype in windows across the genome. We provide extensive simulations evaluating the performance of our method for growth and bottleneck models. These results indicate that a great deal of information regarding demographic history is captured by these two summary statistics. We also extensively test the robustness of our method to many practical problems with genetic data sets in humans. Specifically, we show that for many realistic SNP discovery protocols and levels of population divergence, our method is relatively robust to SNP ascertainment bias. We also found that our method is sensitive to recombination rate variation across the genome (as many haplotype-based summaries will be), and we incorporate a model of recombination rate variation into the inference scheme. Finally, since haplotype phase is often ambiguous, we provide a practical approach to circumvent this problem. We applied our method to genomewide SNP genotype data generated by Perlegen Sciences (Hinds *et al.* 2005). Using the CEU sample (consisting of individuals from Utah with northwestern European ancestry), we find evidence for a recent population bottleneck in northwestern Europe.

METHODS

Summary statistics: We summarize the genomewide data by the joint distribution of two haplotype statistics calculated from windows across the genome. Our method requires that we have a genetic map of the organism in question. Using this map, we divide the genome into windows of fixed genetic map distance, c_{window}. The parameter c_{window} is tunable to the diversity and recombination rates of the organism under study. We chose to divide the genome into n_{window} nonoverlapping windows using genetic map distance so that each window will have the same expected amount of recombination within it and, consequently, the same expected number of haplotypes (Wall 2000a).

In many genomewide SNP data sets, some parts of the genome will have a small number of SNPs while other areas will contain many SNPs. In principle, while this could be due to mutation rate variation across the genome, variations in the time to the most recent common ancestor, or random chance, another likely explanation is ascertainment bias—some parts of the genome were more extensively screened for SNPs than others and consequently have more SNPs. Thus, we do not want our method to use any information about the number of SNPs within a given window. To ensure that all windows of the genome have the same number of SNPs, we select a subset of n_{snp} SNPs for each window of the genome. Again, n_{snp} is a function of the size of the windows as well as the SNP density. Another complicating factor is that SNPs may not have been discovered from the population under study, but instead from a

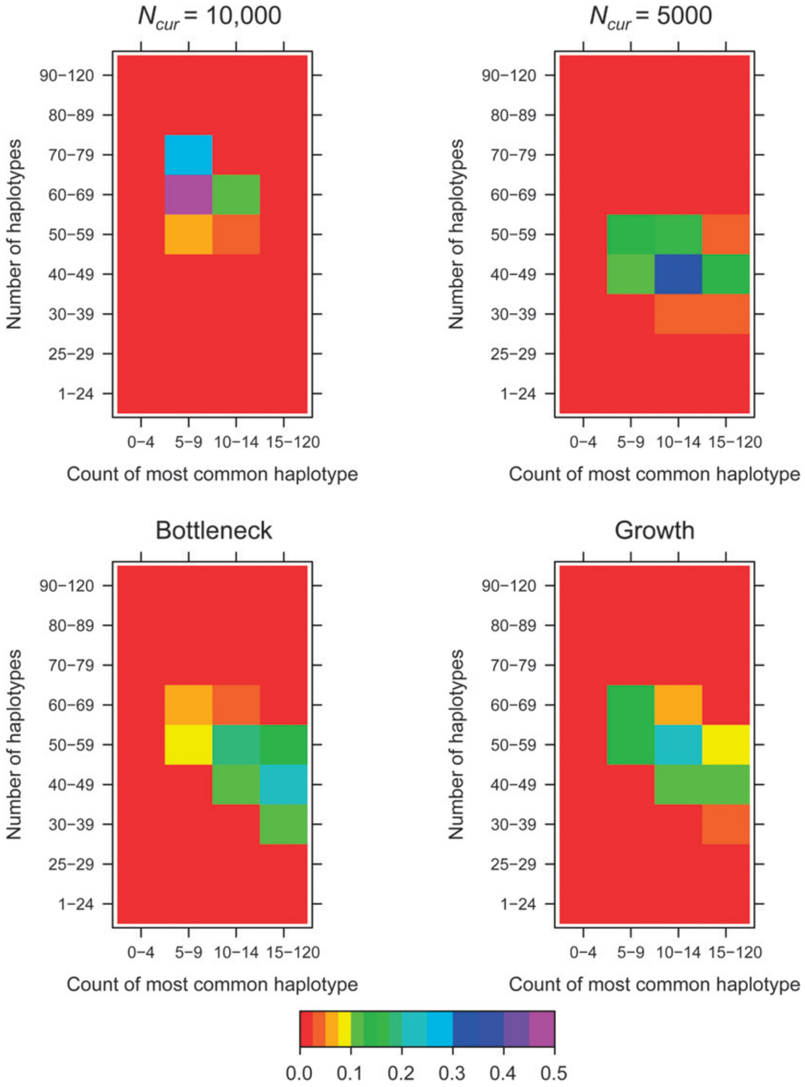


Figure 1.—Examples of the HCN statistic for different demographic models. The color of each cell in a matrix denotes the proportion of simulated windows having the particular number of haplotypes and frequency of the most common haplotype. Approximately 3000 windows were simulated for each demographic model with $n_{\mathrm{snp}} = 25$, $c_{\mathrm{window}} = 0.25$ cM. The parameters for the bottleneck model are $N_{\mathrm{cur}} = N_{\mathrm{anc}} = 10{,}000$, $N_{\mathrm{mid}} = 1000$, and $t_{\mathrm{cur}} = t_{\mathrm{mid}} = 800$ generations. The parameters for the growth model are $N_{\mathrm{cur}} = 10{,}000$, $N_{\mathrm{mid}} = 1000$, and $t_{\mathrm{cur}} = 800$ generations.

second population. Since rare SNPs are more likely to be population specific, and consequently not equally ascertained in all populations, we include only those SNPs with minor allele frequency (MAF) ≥10%.

Having selected a subset of intermediate-frequency SNPs from each window, we can compute the number of distinct haplotypes as well as the count of the most common haplotype in a sample of n chromosomes. The HCN statistic is the genomewide joint distribution of these two statistics. Specifically, let $X = (X_{1,1}, X_{1,2}, \ldots, X_{1,l}, X_{2,1}, \ldots, X_{2,l}, \ldots, X_{k,1}, \ldots, X_{k,l})$, where X_{ij} denotes the number of windows having i haplotypes where the most common haplotype has count j out of n. In principle, $k = l = n$; however, in practice, we bin intervals in the HCN statistic for the inference so that fewer simulation replicates will be needed to obtain an accurate estimate of the expected HCN (see below), and thus there are fewer than n^2 bins in the HCN statistic. Ideally, we wish to integrate over all possible sets of n_{snp} SNPs within each window when constructing the HCN statistic. However, this is not computationally feasible, so we generate 10 random matrices, each using a different randomly selected set of n_{snp} SNPs from each window. We then average these 10 matrices as our final X matrix to be used for inference. This is done to reduce Monte Carlo variance resulting from selecting a single set of SNPs. An example of the HCN statistic for several demographic models is shown in Figure 1.

We chose to use the number of haplotypes as a summary statistic because it is a sufficient statistic for the population mutation rate (θ) in the infinite-alleles model (Ewens 1972) and has been shown by simulation to be informative about population history (Depaulis and Veuille 1998; Innan *et al.* 2005). The count of the most common haplotype was also suggested as a test statistic in the infinite-alleles model (Ewens 1973) and has been found to be correlated with haplotype homozygosity (Zeng *et al.* 2007 and data not shown). The joint distribution of these two statistics performs better at distinguishing among demographic models than using either summary on its own (Figure 1). For example, the number of haplotypes is more informative about overall population size than is the count of the most common haplotype (compare $N_{\mathrm{cur}} = 10{,}000$ to $N_{\mathrm{cur}} = 5000$), as expected, since larger populations have a higher population recombination rate, $\rho = 4N_{\mathrm{e}}c$ than smaller populations, resulting in a larger number of haplotypes per window (Wall 2000a). Note that because we selected n_{snp} SNPs with MAF ≥10% per window, the fact that the larger population has a higher value of θ does not inflate the observed number of haplotypes per window. A recent bottleneck results in an intermediate number of haplotypes, but the stronger signature of the bottleneck is the excess proportion of windows where the most common haplotype is at unusually high frequency. These patterns are due to an elevated rate of coalescence during the bottleneck, which, for some simulated windows, results in there being fewer lineages available to recombine. A recent population expansion also results in an intermediate number of haplotypes, but without an increase in the number of windows where the frequency of the most common haplotype is unusually high.

The HCN statistic contains no information about how different haplotypes within a window are from each other. To add this information, we also considered another summary statistic H_{pair}, the distribution across the genome of the average number of pairwise differences between haplotypes. For all $\binom{n}{2}$ pairs of haplotypes within a given window, we simply counted the number of SNPs (which could range from 0 to n_{snp}) where the two haplotypes differed and counted the average. H_{pair} is the vector giving the number of windows having a given number of average pairwise differences. We show (supporting information, File S1 and Figure S1) that this statistic is not robust to SNP ascertainment bias and do not use it in further analyses.

Demographic models: We consider two different single-population demographic models. These models and their associated parameters are shown in Figure 2. Figure 2A shows a two-epoch model that is used for modeling population growth. Here there are three parameters to estimate: the current population size, N_{cur}; the ancestral population size, N_{mid}; and the time that growth has occurred, t_{cur}. Figure 2B shows a three-epoch model that has five free parameters: the current population size, N_{cur}; the population size during the bottleneck, N_{mid}; the ancestral population size, N_{anc}; the time when the bottleneck started (going backward in time), t_{cur}; and the duration of the bottleneck, t_{mid}. All times are in units of generations. We note that although these models (and all models in population genetics) are arbitrary simplifications of the true demographic history, the hope is that they capture some essential features of population history.

Fitting models to data: Since the observed HCN statistic follows a multinomial distribution, we fit demographic models to the data using an approximate-likelihood approach (see Weiss and Von Haeseler 1998; Wall 2000a; Fearnhead and Donnelly 2002; Plagnol and Wall 2006). We define $\mathbf{p} = (p_{1,1}, p_{1,2}, \ldots, p_{1,l}, p_{2,1}, \ldots, p_{2,l}, \ldots, p_{k,1}, \ldots, p_{k,l})$, where p_{ij} is the probability that a window has i haplotypes where the most common haplotype is at count j. The approximate-likelihood function for the demographic parameters (Θ) can be written as

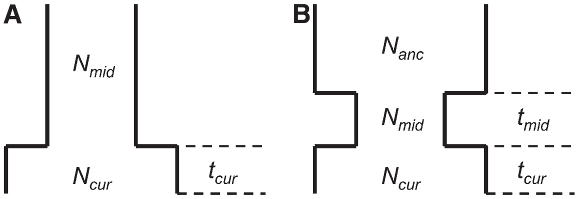


Figure 2.—Demographic models considered. Relevant free parameters are shown in each diagram. (A) Two-epoch model; (B) three-epoch model.

$$L(\Theta) \approx \prod_{i=1}^{k}\prod_{j=1}^{l} p_{ij}^{X_{ij}}. \tag{1}$$

We use the coalescent with recombination (Hudson 1983, 2002) to find p_{ij} for the demographic parameter combination of interest. We simulate z replicates using the demographic parameters of interest (Θ) and $\rho = 4N_{\mathrm{cur}}c_{\mathrm{window}}$. We estimate the matrix $\mathbf{p}$ as the proportion of simulation replicates falling in a particular bin of the HCN statistic. Formally, define the indicator function $I(w, i, j)$ to be equal to 1 if simulation replicate w has i haplotypes and the count of the most common haplotype is j, and equal to 0 otherwise. Then

$$p_{ij} = \frac{\sum_{w=1}^{z} I(w, i, j)}{z}. \tag{2}$$

Since we select n_{snp} SNPs from each window, θ does not explicitly enter into these simulations. Therefore, instead of setting an arbitrary value of θ, and then randomly selecting n_{snp} SNPs, we use the "fixed S approach" (Hudson 1993) to add mutations onto the ancestral recombination graph (ARG). Specifically, n_{snp} mutations are randomly placed onto each simulated ARG such that these SNPs will have MAF ≥10%. To reduce Monte Carlo error, this process is repeated 10 times for each ARG. Each time, we evaluate $I(w, i, j)$ and increment the appropriate bin of $\mathbf{p}$. Note that we record 10 different $\mathbf{p}$ matrices and after the desired number of simulation replicates, we keep the average of the 10 $\mathbf{p}$ matrices as our final matrix. This is an approximate-likelihood function since we are approximating $\mathbf{p}$ using simulations, rather than calculating it exactly, and we also treat all windows of the genome as being mutually independent.

We optimize the approximate-likelihood function described above using a grid search since we are approximating the likelihoods by simulation and the simulation variance may be nonnegligible, misleading deterministic hill-climbing approaches. The number of grid points and number of simulation replicates used to maximize the approximate-likelihood function vary among analyses and are given below.

Since variation in recombination rates at a fine scale can affect the HCN statistic, we have added a model of recombination hotspots into our inference method. We describe the parameters used for specific instances below. Since each window of the genome corresponds to the same genetic map distance (c_{window}), the number of base pairs per window will differ among windows. In our simulations to find **p**, we select the size of the window in base pairs (denoted L) from the observed distribution of physical distances. We then set r, the per base pair recombination rate, to be constant across the window such that rL will give c_{window}. We then simulate an ARG in the normal manner with recombination rate rL, but then, similar to the method used by Li and Stephens (2003), we model hotspots by changing the relationship between physical and genetic distance. Informally, the parts of the window where hotspots occur are assigned fewer base pairs and consequently have a lower probability of a SNP occurring in them than windows with lower recombination rates.

Simulations to evaluate performance: We tested the performance of our method by simulating data under three different demographic models: (1) ancient population growth, (2) recent population growth, and (3) a recent population bottleneck. The parameter values for these models are shown in Figures 3–6. These models were chosen because of their relevance to human demographic history. For each model we simulated data sets (500 for models assuming uniform recombination rates and 100 for models with recombination hotspots), each consisting of 2000 independent 250-kb windows in a sample of 40 chromosomes where $\mu = 1 \times 10^{-8}$/bp/generation. Note that when generating test data sets, we placed a Poisson number of mutations onto the genealogies in the usual fashion (Hudson 1983), rather than using the fixed S approach as we did for the simulations used to estimate **p**. We selected a subset of 20 SNPs ($n_{\mathrm{snp}} = 20$) with MAF ≥10% from each window and constructed the observed HCN statistic for each data set. We repeated this process 10 times for each data set and used the average HCN statistic for inference. For computational efficiency, we performed the coalescent simulations to estimate **p** over a grid of parameters (3780 and 20,580 parameter combinations for the growth and bottleneck models, respectively) once for each demographic and recombination model and stored the values to be used on subsequent data sets. The grid points used for each parameter are shown as the breaks in Figures 3–6. For each grid point we used 10^4 coalescent simulations to approximate the likelihood. Using a representative data set, we then selected at least the top 10^3 grid points and ran an additional 10^5 replicates, and for points near the maximum-likelihood estimates (MLEs), we ran an additional 10^6 replicates. Due to computational constraints, these grids were not as dense as those used to estimate parameters in the Perlegen data.

For all data sets and demographic models, the total amount of recombination within each window simulated was 0.25 cM (*i.e.*, $c_{\mathrm{window}} = 0.25$ cM). However, we also considered a model with five recombination hotspots present at random locations throughout each window. Each hotspot was 2 kb in size. The recombination rate (centimorgans per base pair) of each hotspot was drawn from a gamma distribution (shape = 0.5, scale = 2×10^{-6}). We then rescaled the recombination rate of each hotspot such that 80% of the total amount of recombination in the window occurs within hotspots. We then assumed this model of recombination hotspots when inferring demographic parameters. The test data sets were generated using the program msHOT (Hellenthal and Stephens 2007).

Our method assumes that all windows in the genome are independent of each other. To assess the performance of our method when the windows are not independent, we simulated an additional 500 data sets using the same bottleneck model with $c_{\mathrm{window}} = 0.25$ cM. Here the 2000 windows within each data set were from 300 independent sets of 6 or 7 contiguous windows. All windows were treated as independent in the inference.

While for most of the simulations we assumed that there was no error in estimated recombination rates (*i.e.*, $\hat{c}_{\mathrm{window}} = c_{\mathrm{window}}$), we also determined what effect errors in the estimated genetic map had on our ability to accurately infer demographic parameters. Specifically, we simulated data sets under the same bottleneck model described above, but here instead of having $\hat{c}_{\mathrm{window}} = c_{\mathrm{window}} = 0.25$ cM, we drew c_{window} for each window from a gamma distribution (shape = 10, scale = 0.025). From this distribution, ~10% of windows have $c_{\mathrm{window}} < 0.155$ and ~10% of windows have $c_{\mathrm{window}} > 0.355$. We then inferred demographic parameters when incorrectly fixing $\hat{c}_{\mathrm{window}} = 0.25$ cM. We also correctly incorporated errors into the genetic map by drawing $\hat{c}_{\mathrm{window}}$ for each simulation replicate from the same gamma distribution used to generate the data.

Due to differences between the true HCN statistic and the HCN statistic constructed from phase-inferred haplotypes (Figure S2 and File S1), it is important to incorporate the phasing process into the inference. To do this, we suggest using the same phasing method that was used on the actual data to "phase" the simulated data used to estimate **p**. Unfortunately, many phasing algorithms currently in use are computationally intensive and it would be nearly impossible to run these methods on the millions of coalescent simulation replicates used to find **p**. For this reason, we examined the use of the computationally efficient parsimony phasing algorithm proposed by Clark (1990). If there are no individuals heterozygous at zero or one of the n_{snp} SNPs within a window or if there are genotypes that show no relation to known phased haplotypes, we arbitrarily assigned phase to a random individual and then used these two haplotypes to infer the rest. While this process may seem arbitrary, it can be done consistently both in the observed and in the simulated data sets. To make the method as computationally efficient as possible, we used only one ordering of the individuals. We assessed the performance of this approach by treating the simulated haplotypes in the test data sets as diploid genotypes and "phased" them using the parsimony method. For each simulation replicate to estimate **p** we also phased the simulated data using the same parsimony method.

Since we found that the HCN statistic constructed using SNPs that were discovered in a SNP discovery sample ≥8 chromosomes was very similar to the HCN statistic with complete SNP ascertainment (Figure S3, Figure S4, Figure S5, Table S1, and File S1), we examined how ascertainment bias affected parameter estimates. Specifically, for the bottleneck model described above, we simulated a genotype sample of $n = 40$ and an additional SNP discovery sample of 6 chromosomes. Since the Perlegen SNPs were discovered using a multiethnic panel (Hinds *et al.* 2005), we included a SNP discovery sample of an additional 6 chromosomes from a second population (N_{cur} = 10,000) that 5000 generations ago split from the population that underwent the bottleneck. To construct the HCN statistic from these simulated data sets, we considered only SNPs with MAF ≥10% in the genotype sample that were variable in the 12-chromosome ascertainment panel. To infer parameters, we assumed there was no ascertainment bias (*i.e.*, we used the same lookup tables for **p** that were described above that assumed complete SNP ascertainment).

Analysis of Perlegen data: We applied our method to fit a bottleneck model to the CEU population genotyped by Perlegen Sciences (Hinds *et al.* 2005). We chose to use this population since there is previous evidence of a bottleneck in this population (*e.g.*, Marth *et al.* 2004; Voight *et al.* 2005), and all SNPs that were discovered by the Perlegen resequencing arrays were later genotyped in the CEU sample, without regard to LD status. We note that HapMap phase II specifically did not genotype SNPs that were in high LD in the Perlegen study, and this ascertainment criterion complicates the analysis of those data (see discussion). We considered only autosomal (not X chromosome or mtDNA) SNPs with MAF ≥10% in both the CEU and the African-American samples. Since our simulations of ascertainment bias suggest that SNPs needed to have been discovered from discovery sample sizes >2 chromosomes, we used only those SNPs that were discovered in Perlegen's resequencing arrays of the multiethnic diversity panel (type "A" SNPs). There were 615,415 SNPs that fit both of these criteria. We used Clark's parsimony phasing algorithm to phase haplotypes in both the real data and the simulation replicates to generate **p**. For each population and in each window of the genome, we selected 10 random subsets of n_{snp} SNPs and constructed 10 different HCN statistics. We then used the average HCN statistic for inference.

We then set $c_{\mathrm{window}} = 0.25$ cM and $n_{\mathrm{snp}} = 20$. We used the LDhat genetic map (International Hapmap Consortium 2007) to define windows since the deCODE map (based on pedigrees; Kong *et al.* 2002) does not have sufficient resolution for the scale of 0.25 cM (Myers *et al.* 2005). Since the quality of the genetic map used to divide the genome into windows can affect the inference, we drew $\hat{c}_{\mathrm{window}}$ from a gamma distribution (shape = 10, scale = 0.025) to model errors in the genetic map. This distribution has a mean of 0.25 and a variance of 0.00625. For the CEU data, $n_{\mathrm{window}} = 8833$.

We used a hotspot model similar to that of Schaffner *et al.* (2005; termed the "Schaffner hotspot model"). All hotspots had width of 2 kb. For each simulated window, hotspots occurred at random intervals drawn from a gamma distribution (shape = 0.3, scale = 8500/0.3), giving a mean spacing of 8.5 kb (variance of ~2.41×10^8). Then the recombination rate (cM/2 kb) of each hotspot was drawn from another gamma distribution (shape = 0.3, scale = $\hat{c}_{\mathrm{window}}/0.3L$), where L is the physical size of the simulated window. In practice, L was drawn from the empirical distribution of physical distances for the n_{window} windows. We then rescaled the recombination rate in the hotpots such that 88% of $\hat{c}_{\mathrm{window}}$ occurs within hotspots. The amount of recombination occurring outside of hotspots is then equal to $0.12\hat{c}_{\mathrm{window}}$. Figure S6 and Figure S7 show that this hotspot model matches the mean, the standard deviation, and the overall distribution (tabulated across all windows of the genome) of the observed inter-SNP genetic map distances quite well.

In addition to the Schaffner hotspot model described above, we also directly used the estimated fine-scale LDhat genetic map (International Hapmap Consortium 2007) as a guide to how recombination rates vary within windows (termed the "empirical hotspot" model). To do this, for each simulation replicate to estimate **p**, we selected one of the 8833 windows at random and used the corresponding LDhat genetic map to delimit the relationship between genetic and physical distance for that replicate. We smoothed the map by allowing the recombination rate to change only at points >500 bp and >0.0001 cM apart. We note that this hotspot model does not match the observed inter-SNP genetic distances as well as the Schaffner hotspot model does (Figure S6 and Figure S7).

The grid to optimize the five-dimensional approximate-likelihood function consisted of 85,536 points for the Schaffner hotspot model and 101,088 points for the empirical hotspot model. We used 12,500 simulation replicates for all points, 10^5 replicates for at least the top 4000 points, and finally 10^6 replicates for at least the top 500 points. We found approximate 95% confidence intervals (C.I.'s) for single parameters using asymptotic theory (*i.e.*, the C.I. included points <1.92 log-likelihood units from the maximum), with linear interpolation of the profile-likelihood curve to find points not directly simulated.

RESULTS

Performance on simulated data: Figure 3 shows the distribution of the approximate MLEs of the three growth model parameters for simulated data sets under ancient growth (solid bars) and recent growth (open bars). In both cases, the method is relatively unbiased for all three parameters. For ancient growth, N_{cur} is estimated most accurately and t_{cur} the least. For recent growth, all three parameters are equally accurate, although for any given parameter, the MLE is the true value ~40% of the time. Note that the variance in the distribution of MLEs for t_{cur} is much higher for ancient growth as compared to recent growth (making it the least precise as well as the least accurate). Table 1 shows that for both growth scenarios in 100% of the data sets, the true parameter values were within the asymptotic 95% C.I.'s (<3.9 log-likelihood units) around the MLEs. Additionally, in >95% of the data sets, the one-dimensional 95% C.I.'s for all three parameters from the profile-likelihood curves contain the true parameter values.

We also estimated the five parameters for a bottleneck model in simulated data sets. Figure 4 shows the distribution of the MLEs for the five bottleneck parameters. For the case of uniform recombination and $\hat{c}_{\text{window}} = c_{\text{window}} = 0.25$ cM, the mode of the distribution of the MLEs for each parameter is at the true value of the parameter. The distribution of MLEs is tightest for $N_{\text{mid}}/N_{\text{cur}}$ and t_{cur} and broadest for $N_{\text{anc}}/N_{\text{cur}}$. This suggests that the recent bottleneck greatly alters haplotype patterns such that its timing and severity can be accurately estimated, but so much so that less information about the ancestral, prebottleneck population size ($N_{\text{anc}}/N_{\text{cur}}$) remains. Furthermore, the method appears to be relatively unbiased since it over- and underestimates the true parameter value roughly equally. Table 1 shows that 99.8% of the time, the true parameter values are within the asymptotic 95% C.I.'s (within 5.5 log-likelihood units around the MLEs).

We next evaluated whether our method could accurately estimate demographic parameters in the presence of recombination hotspots (see METHODS for the recombination hotspot model used). Figure 4 shows that when properly modeling recombination hotspots, we are able to accurately estimate the five bottleneck model parameters. Note that the distributions of the MLEs for all parameters have larger variances than for the uniform recombination case. This pattern is due to the extra noise added by recombination hotspots. If a window of the genome has a low number of haplotypes and/or a high count of the most common haplotype, this could be due to demography (which is the only factor considered in the uniform recombination model) or due to SNPs falling in recombination coldspots. Consistent with this observation, Table 1 shows that a smaller proportion of the parameter space (99.65% *vs.* 99.87%) is >5.5 log-likelihood units from the MLEs when there are recombination hotspots, as compared to uniform recombination. Notably, however, the method still appears to be unbiased, and for all 100 data sets, the

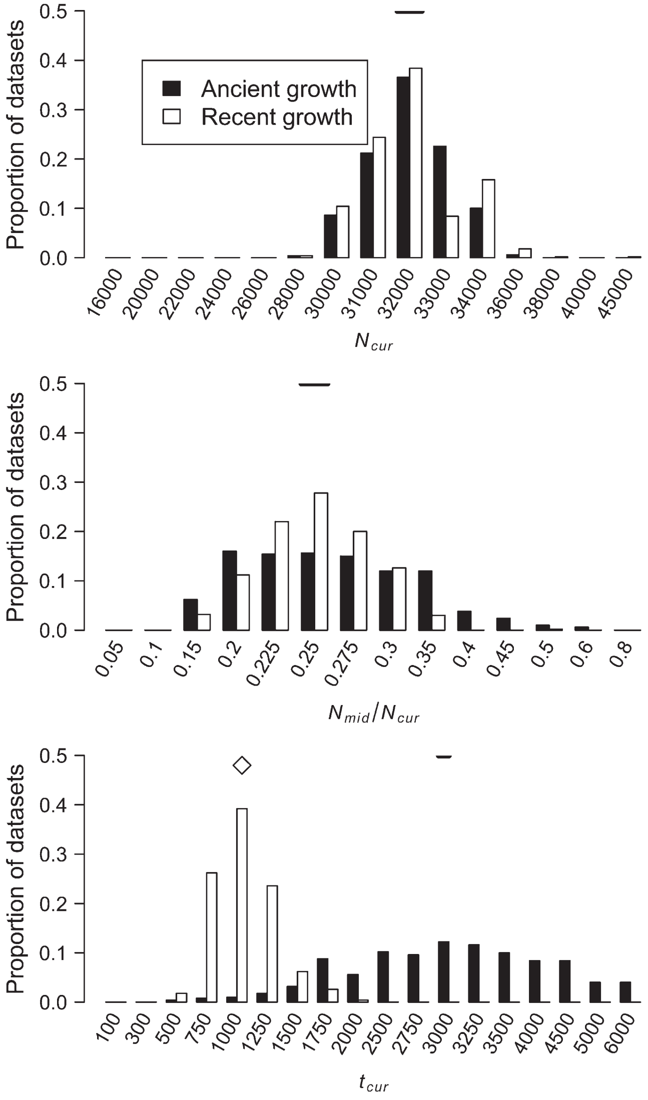


FIGURE 3.—Distributions of MLEs of the three growth model parameters for simulated data sets under ancient growth and recent growth with uniform recombination (see METHODS). The true value of each parameter is denoted by the horizontal line in each section. Since t_{cur} differs between the two growth models, the true value of t_{cur} is denoted by an open diamond for recent growth and a horizontal line for ancient growth.

TABLE 1

Comparison of MLEs to the true parameter values for simulated data sets

Model	Mean $ll_{\text{MLEs}} - ll_{\text{Truth}}$[a]	Max $ll_{\text{MLEs}} - ll_{\text{Truth}}$[b]	% MLEs = truth[c]	Coverage of multi dimensional 95% C.I.'s[d]	Coverage of one-dimensional 95% C.I.'s[e]	% points outside 95% C.I.'s[f]
		Ancient growth				
$\hat{c}_{\text{window}} = c_{\text{window}} = 0.25$ cM	0.631	3.47	3.0	100.0	99.87	94.79
		Recent growth				
$\hat{c}_{\text{window}} = c_{\text{window}} = 0.25$ cM	0.437	3.09	22.6	100.0	99.93	98.84
		Bottleneck				
$\hat{c}_{\text{window}} = c_{\text{window}} = 0.25$ cM	0.363	6.31	47.4	99.8	99.52	99.87
Hotspots[g]	0.505	3.43	17.0	100.0	99.80	99.65
Linkage	0.732	8.30	29.8	99.4	98.48	99.87
$\hat{c}_{\text{window}} = 0.25$ cM, $c_{\text{window}} \sim$ gamma	136.685	179.07	0.0	0.0	8.36	99.98
$\hat{c}_{\text{window}} \sim c_{\text{window}} \sim$ gamma	0.623	6.54	26.4	99.6	99.40	99.72
Clark's phasing algorithm[h]	0.779	4.90	19.8	100.0	99.96	99.43
Ascertainment bias	1.998	12.65	14.6	94.0	97.64	99.86

[a] The average overall data sets of the log-likelihood at the MLEs minus the log-likelihood of the true demographic parameters.

[b] The maximum distance between the log-likelihood at the MLEs and the log-likelihood of the true demographic parameters.

[c] The proportion of data sets where the MLEs for all parameters were the true demographic parameters.

[d] The proportion of data sets where the true parameter values were <3.9 or <5.5 log-likelihood units from the MLEs, for the growth and bottleneck models, respectively.

[e] The proportion of data sets where the true value of each parameter was <1.92 log-likelihood units from the MLE using the profile log-likelihood curve, averaged over three or five parameters for the growth and bottleneck models, respectively.

[f] The fraction of grid points (see RESULTS) having a log-likelihood >3.9 or >5.5 log-likelihood units, for the growth and bottleneck models, respectively, from the MLEs.

[g] Each window has five recombination hotspots, but for the whole window $\hat{c}_{\text{window}} = c_{\text{window}} = 0.25$ cM.

[h] Haplotype phase was inferred in the test data sets and simulations to estimate **p** using Clark's phasing algorithm (see METHODS).

true parameter values are <5.5 log-likelihood units from the MLEs.

The simulations described above assumed that the windows in each data set were independent. In practice, the windows may be contiguous along the genome and thus are not independent. We examined the performance of our method on simulated data sets where some of the windows were linked. Figure 4 shows that the distributions of the MLEs for certain parameters have greater variance than when the windows are

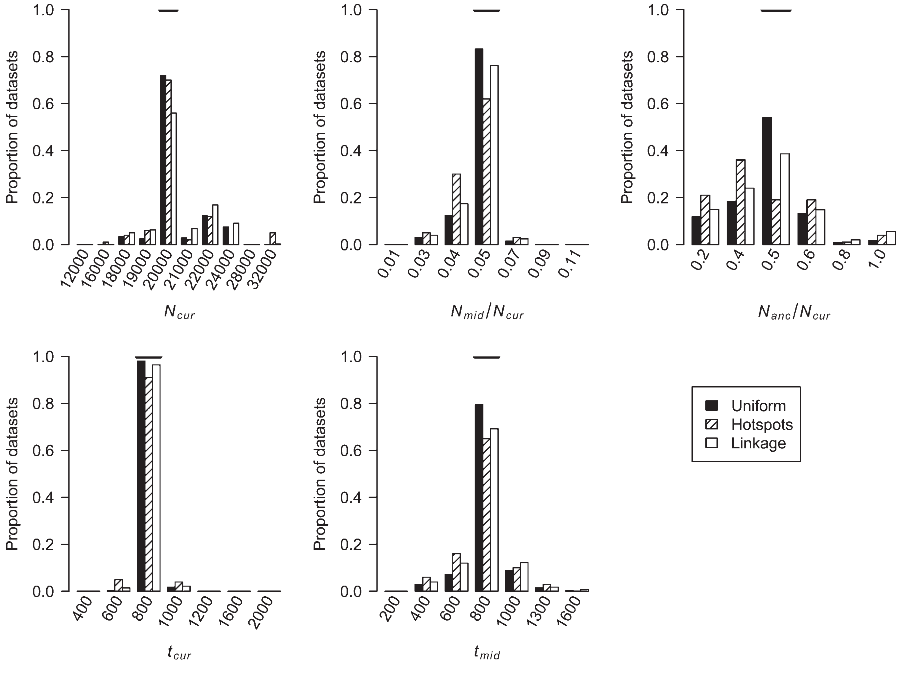


FIGURE 4.—Distributions of MLEs of the five bottleneck model parameters for simulated data sets under uniform recombination, under hotspots, and where some windows in the simulated data sets are linked to one another (see METHODS). The true value of each parameter is denoted by the horizontal line in each section.

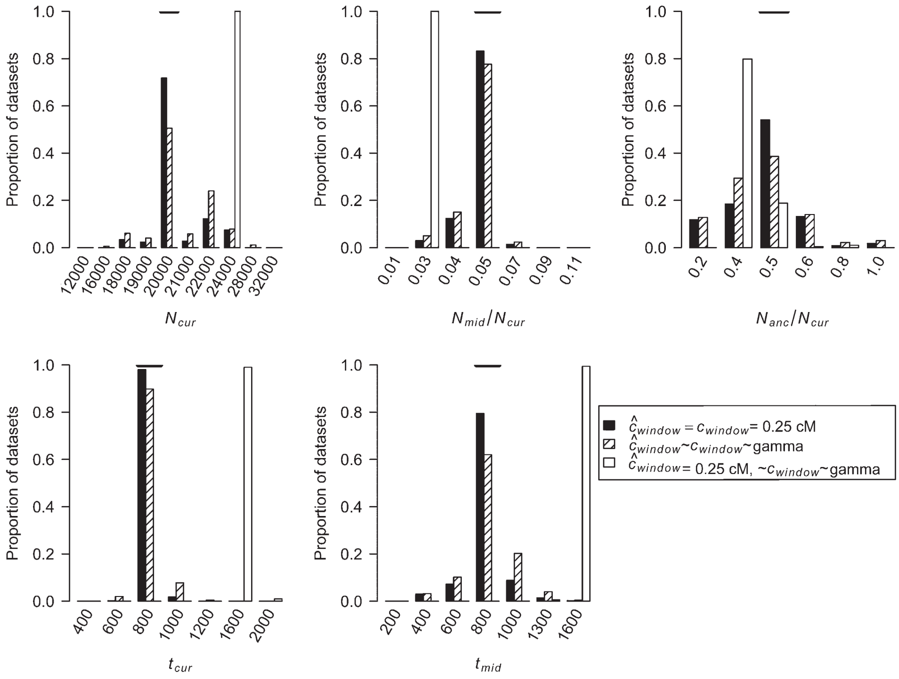


FIGURE 5.—Distributions of MLEs of the five bottleneck model parameters for simulated data sets where there are errors in the genetic map. $\hat{c}_{\text{window}} =$ 0.25 cM, $c_{\text{window}} \sim$ gamma denotes the case where there are errors in the estimated genetic map that are ignored when performing the inference. $\hat{c}_{\text{window}} \sim c_{\text{window}} \sim$ gamma denotes the case where we allow for errors in the genetic map when conducting the inference. The true value of each parameter is denoted by the horizontal line in each section.

unlinked, which is not surprising since linked windows contain less information than independent windows. As shown in Table 1, in all cases except when errors in the genetic map are ignored or there is SNP ascertainment bias (see below), the true parameter values for >99% of the test data sets are within the asymptotic 95% C.I.'s (<3.9 and <5.5 log-likelihood units from the MLEs for growth and bottleneck models, respectively). This result suggests that the asymptotic C.I.'s may actually be conservative, since the true parameter values are contained within the interval >95% of the time. When examining data sets where some of the regions are linked, we find that for 99.4% of the time, the true values are <5.5 log-likelihood units from the MLEs (Table 1). Since in many cases, the 95% C.I.'s for individual parameters based on the profile log-likelihood curve also appeared conservative (Table 1), we assessed their coverage in the data sets with linkage. For each of the five parameters, the true parameter value is <1.92 log-likelihood units from the MLE in at least 96.8% of the data sets. These results suggest that for the level of nonindependence among windows, size of data sets, and parameter grid considered here, the asymptotic 95% C.I.'s remain conservative.

The above simulations assumed that $\hat{c}_{\text{window}} = c_{\text{window}}$. In practice, $\hat{c}_{\text{window}}$ is estimated from a genetic map, based either on patterns of LD or on pedigrees. We next evaluated the performance of our method when c_{window} is drawn from a gamma distribution to mimic errors in the estimated genetic map. We first assumed that $\hat{c}_{\text{window}} = 0.25$ cM when running the inference (*i.e.*, we ignored the errors in the genetic map). Figure 5 shows that our method performs poorly compared to the case where the genetic map is known with certainty. In particular, it overestimates t_{cur} and t_{mid}. Due to the fact that some windows in the simulated data sets will have low recombination rates, these windows will have very few haplotypes and a high frequency of the most common haplotype because $c_{\text{window}} < 0.25$ cM. Since we did not account for this in the inference, the method assumes that these low-diversity windows were due to a stronger (or longer) bottleneck. Table 1 shows that the true parameter values are nowhere near the MLEs in this case. If, however, during the inference, $\hat{c}_{\text{window}}$ for each window is drawn from the same gamma distribution that generated the data, the method performs substantially better (Figure 5), although not quite as well as when $\hat{c}_{\text{window}} = c_{\text{window}} = 0.25$ cM. Likewise, 99.6% of the time, the true parameter values are <5.5 log-likelihood units from the MLEs. Note that, similar to what was seen for the case of recombination hotspots, on average, a smaller proportion of the parameter space (99.72% *vs.* 99.87%) is >5.5 log-likelihood units from the MLEs when $\hat{c}_{\text{window}}$ and c_{window} follow a gamma distribution instead of being fixed at 0.25 cM.

To properly correct for errors introduced from inferring haplotype phase, we decided to phase the simulations used to estimate **p** using the same method as that used on the real data. We evaluated the performance of this strategy using Clark's phasing algorithm (CLARK 1990) on simulated data sets. Figure 6 shows the distribution of the MLEs for the five bottleneck parameters. This strategy works reasonably well and for each parameter the mode of the distribution of the MLEs is at the true parameter value. Note that the distributions of the MLEs for the data sets phased using Clark's phasing algorithm are broader than those when haplotype phase in known with

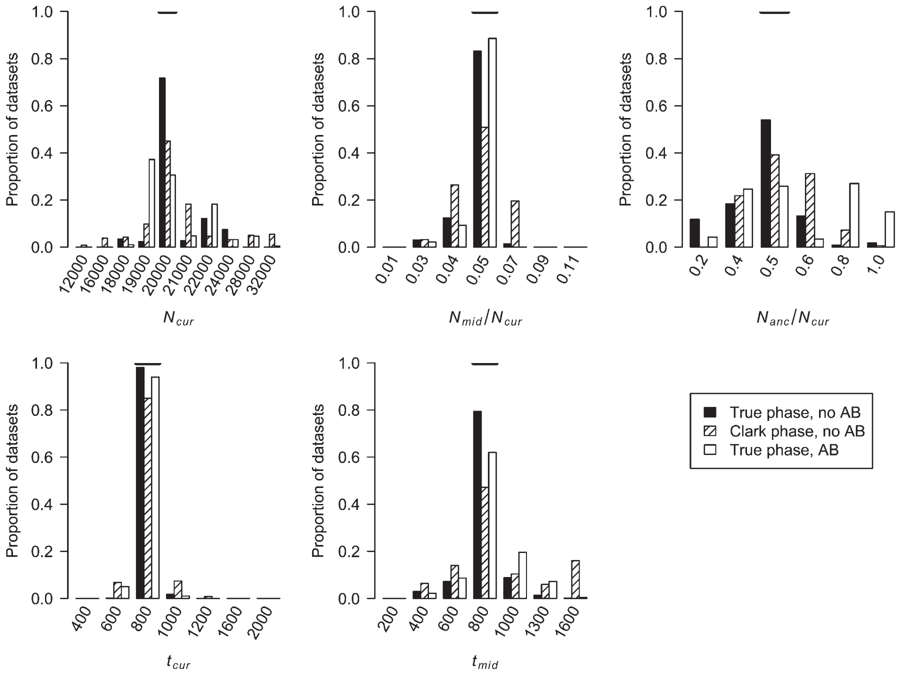


FIGURE 6.—Distributions of MLEs of the five bottleneck model parameters for simulated data sets when phasing genotype data using Clark's phasing algorithm or there is SNP ascertainment bias (AB) (see METHODS). The true value of each parameter is denoted by the horizontal line in each section.

certainty. Additionally, a smaller proportion of the parameter space is excluded (>5.5 log-likelihood units from the MLEs) when using Clark's phasing algorithm as compared to known phase data (99.43% *vs.* 99.87%; Table 1). This finding illustrates that, compared to having phase-known haplotypes, some information is lost when computationally inferring haplotypes. However, the method appears reasonably unbiased, and in all simulated data sets the true parameter values are <5.5 log-likelihood units from the MLEs (Table 1).

To determine if we could accurately estimate bottleneck parameters in the presence of SNP ascertainment bias, we simulated data sets where the $n_{\mathrm{snp}} = 20$ SNPs for each window were picked from those SNPs with MAF $\geq 10\%$ in the genotype sample and were variable in the 12-chromosome SNP ascertainment sample. Figure 6 shows that for $N_{\mathrm{mid}}/N_{\mathrm{cur}}$, t_{cur}, and t_{mid}, our method performs very well even in the presence of SNP ascertainment bias. The distributions of the MLEs for N_{cur} and $N_{\mathrm{anc}}/N_{\mathrm{cur}}$ are more variable than when there is no ascertainment bias, and the modes of their distributions are not at the true parameter values, suggesting that MLEs of these parameters are less reliable in the presence of ascertainment bias. The 95% C.I.'s constructed from the profile-likelihood curves remain conservative for $N_{\mathrm{mid}}/N_{\mathrm{cur}}$, t_{cur}, and t_{mid}, but for N_{cur} the 95% C.I. is no longer conservative (*i.e.*, the true value is within the 95% C.I. only 91.8% of the time). Additionally, the five-dimensional 95% C.I. is also slightly anticonservative (Table 1). For larger data sets (consisting of 10,000 independent regions) the C.I.'s for N_{cur} and $N_{\mathrm{anc}}/N_{\mathrm{cur}}$ and the five-dimensional C.I. become even more anticonservative and the C.I. for t_{mid} becomes slightly anticonservative (not shown), likely due the fact that the ascertainment model is misspecified, which has a stronger effect as the size of the data set increases.

Inference of bottleneck parameters for the Perlegen CEU population: We fit the five-parameter bottleneck model to the Perlegen CEU population. Table 2 shows the MLEs and ~95% C.I.'s for the five parameters when using both the Schaffner model of recombination hotspots and the empirical hotspot model based on the LDhat genetic map. Figure S8 shows that for both hotspot models, the HCNs generated using the MLE parameter estimates match the observed CEU HCN quite well. The current population size is estimated to be ~10,000 when using both recombination models. There was a severe population size reduction (~4.2–6.6% of the current size) lasting 260–552 generations (see Figure S9 for the two-dimensional profile-likelihood surface). The bottleneck ended (forward in time) ~1017–1437 generations ago.

On the basis of the two-dimensional profile-likelihood surface (Figure S10), we estimate that the bottleneck began (t_{cur} + t_{mid}) 1500 generations ago—37,500 years, assuming 25 years/generation when using either hotspot model. Notably, the oldest start time within the asymptotic 95% C.I. (3 log-likelihood units, for 2 d.f.) is 1600 generations, or 40,000 years for the Schaffner hotspot model, and 1900 generations (47,500 years) for the empirical hotspot model.

One potential concern with using SNPs from the Perlegen SNP discovery project is that it contains a lot of missing data, resulting in some fraction of SNPs having been discovered in a smaller sample. To determine what effect this had on the inference of the bottleneck parameters, we examined the depth of the SNP discov-

TABLE 2

Inferred bottleneck parameters for the CEU data set

	Schaffner hotspot model[a]		Empirical hotspot model[b]	
Parameter	MLE	95% C.I.	MLE	95% C.I.
N_{cur}	10,000	9,315–10,665	10,000	9,440–11,454
$N_{\mathrm{mid}}/N_{\mathrm{cur}}$	0.055	0.052–0.064	0.05	0.042–0.066
$N_{\mathrm{anc}}/N_{\mathrm{cur}}$	0.8	0.70–?[c]	0.8	0.66–0.97
t_{cur}	1,100	1,017–1,140	1,200	1,086–1,437
t_{mid}	400	360–475	300	260–552

MLEs and ∼95% C.I.'s from the profile-likelihood curves for the five-parameter bottleneck model are shown.

[a] The hotspot model is similar to that described by Schaffner *et al.* (2005) (see methods).

[b] The hotspot model is based directly on the LDhat genetic map.

[c] We did not estimate an upper boundary on the C.I. since the profile-likelihood surface for $N_{\mathrm{anc}}/N_{\mathrm{cur}}$ is relatively flat from 0.8 to 1.0 and we did not consider points >1.0 (Figure S11).

ery panel for the 615,415 SNPs used in constructing the HCN statistic used for demographic inference in the Perlegen CEU sample. We found that 11.8% of these SNPs were discovered by comparing fewer than eight chromosomes. We removed these SNPs and recomputed the HCN statistic from the Perlegen CEU data (now with 8174 windows) and reestimated the bottleneck parameters using the Schaffner recombination hotspot model. We found identical MLEs for the parameters as in our original analysis.

DISCUSSION

We have proposed a flexible approximate-likelihood method to estimate demographic parameters using haplotype summary statistics. We have shown that accurate estimates of demographic parameters can be made using genomewide SNP data sets of practical size. To the best of our knowledge, this is one of the first studies to estimate parameters in a demographic model in a likelihood framework using haplotype patterns from genomewide SNP genotype data. Furthermore, we have addressed many complications that arise in the analysis of genomewide data, such as recombination rate heterogeneity, errors in the estimated genetic map, haplotype phase uncertainty, and ascertainment bias. Provided that good genetic maps are available, our method could be applied to SNP data from other species, such as dogs and cattle, to estimate domestication bottleneck parameters.

One of the major disadvantages of our approach is its dependence on accurately modeling the distribution of recombination rates across the genome. Our simulations have shown that errors in the genetic map can cause poor performance. Therefore, we suggest applying our method only when there is an accurate genetic map for the species in question. We suggest incorporating a distribution on $\hat{c}_{\mathrm{window}}$ to allow for errors in the estimated genetic map. However, as the quality and resolution of genetic maps continue to improve, the utility and accuracy of our method will also continue to increase. For species where recombination rates vary at the fine scale, it is crucial to incorporate some model of recombination hotspots. Here for the Perlegen CEU data, we have implemented a parametric model as well as empirical estimates based on LD patterns. A similar influence of the assumed recombination rate on the demographic parameter estimates was noted in Thornton and Andolfatto (2006), who used the variance in the number of haplotypes across windows as one of their summary statistics. We recommend, as done in Thornton and Andolfatto (2006), using different recombination models and then comparing the final results to assess how dependent the estimates are on the assumed recombination model. While the dependence of our method on accurate estimates of the recombination rate is not ideal, we point out that many previous methods in molecular evolution and population genetics are dependent on accurate estimates of the mutation rate and will be biased if erroneous estimates are used.

We also assume that the genetic map remains constant over time and is the same across populations. Recombination hotspots do not appear in the same locations in chimps and humans despite a high level of sequence identity (Ptak *et al.* 2005; Winckler *et al.* 2005). It has therefore been speculated that recombination hotspots are not permanent features of the genome and evolve on a timescale of at least tens of thousands of years (Jeffreys *et al.* 2005). However, it appears that the timescale over which many hotspots evolve is older than 100,000 years, and because this is long enough to alter patterns of LD, temporal changes in hotspots on this timescale will not have such a severe impact on our method. Hotspots that evolve over shorter timescales, or are population specific, may have a larger effect on our method. This effect is hard to quantify since the prevalence of rapidly evolving or population-specific hotspots, other than the existence of a few examples (Jeffreys *et al.* 2005; Clark *et al.* 2007), remains largely unknown. Encouragingly, a recent article (Hellenthal *et al.* 2006) found that genotype-specific recombination events would not substantially affect LD patterns, boding well for our method.

We have found that the HCN statistic constructed from computationally phase-inferred data differs from the true HCN. Simply treating the phase-inferred haplotypes as the true haplotypes will likely give biased parameter estimates. Thus, when analyzing data from unrelated individuals, it is important to consider errors induced in the phasing process. We suggest doing this by inferring phase for the coalescent simulations used to estimate HCN. Our simulations suggest that using Clark's phasing algorithm works well for this purpose. However, some information is lost by this procedure (Figure 6; Table 1), and we therefore recommend, where available, using data from trios, where haplotype phase can be inferred with great accuracy, to maximize the information in the data.

It has been suggested (Conrad *et al.* 2006) that haplotype statistics may be less susceptible to SNP ascertainment bias than statistics based upon SNP frequencies. Here we have extensively investigated whether this holds true for the HCN and H_{pair} statistics under a variety of demographic and ascertainment conditions. Encouragingly, we found that the HCN statistic is reasonably robust to SNP ascertainment bias provided that the SNP discovery sample is sufficiently deep. The reason for this is that we focus on subsets of common SNPs, rather than on rare SNPs. However, the H_{pair} statistic was very susceptible to ascertainment bias (Figure S1 and File S1), suggesting that all haplotype statistics are not equally affected by ascertainment bias, and it will be necessary to explicitly evaluate, as we have done here, whether ascertainment bias affects a particular haplotype statistic.

The sizes of the SNP discovery and genotype samples play an important role in determining the effect of ascertainment bias on the HCN statistic. Interestingly, generating the HCN from SNPs ascertained uniformly using two chromosomes of known ethnicity (as done by Keinan *et al.* 2007) would result in a very different HCN statistic from that expected without ascertainment bias (Figure S3, Figure S4, and Figure S5), which would result in biased inference. Although not considered here, it is in principle possible to estimate the expected HCN statistic conditional on this particular ascertainment strategy, and application of such an estimate would reduce this bias.

We have found that SNP discovery sample sizes of at least 12 total chromosomes should be sufficient to result in the HCN statistic from ascertained SNPs to match the expected HCN when considering genotype samples of 40 or 120 chromosomes. Furthermore, we have shown that our method can reliably infer several bottleneck parameters when SNPs were ascertained in this manner. As the SNP discovery sample size increases, performance of our method will continue to improve and become closer to that for the case of no ascertainment bias, since a larger SNP discovery sample will capture more of the SNPs in the genotype sample. These results are especially encouraging since Perlegen's SNP discovery effort used 20–50 chromosomes, where ~12 chromosomes were of African-American ancestry, ~12 chromosomes were of European ancestry, and the remainder were of Mexican-American, Asian-American, and Native American ancestry (Collins *et al.* 1998; Hinds *et al.* 2005).

Furthermore, we have found that the size of the SNP discovery sample is more important than whether or not the SNP ascertainment had been done in a particular population (see Figure S5). This suggests that SNPs ascertained in the Perlegen resequencing survey could be used to estimate demographic parameters in other populations not represented in the SNP discovery panel. It is worth noting that the two populations in our simulation study shown in Figure S5 split 5000 generations ago (125,000 years, assuming 25 years/generation) with no subsequent migration and are thus more differentiated than many actual non-African populations that could be studied empirically.

It is important to note that the type of ascertainment bias studied here is due to the preferential genotyping of common SNPs. In all the analyses presented here, we assume that the genotyped SNPs were selected without regard to physical or genetic distance or LD patterns. Such an assumption is reasonable for the analyses of the Perlegen data presented here since Perlegen attempted to genotype all of the SNPs found in their SNP discovery process (Hinds *et al.* 2005). The assumption is not valid, however, for many of the "SNP chips" that preferentially selected SNPs on the basis of physical distance (in the case of Affymetrix 500K) or LD patterns (in the case of the Illumina platform; Eberle *et al.* 2007). Since the SNPs on these platforms are not a random subset of the total variation, using our method on such data will likely give misleading results. In principle, it should be possible to modify our method to model the SNP selection process in the inference, which would allow our method to be applied to the large-scale SNP genotype data sets that have been collected, such as the HGDP data set (Jakobsson *et al.* 2008; Li *et al.* 2008).

For the analysis of the CEU data, we used two different models of recombination rate variation. Overall, the results using both models are qualitatively similar, suggesting that our method is somewhat robust to minor misspecification of the recombination hotspot model. We also find that the one-dimensional 95% C.I.'s from the profile-likelihood curves overlap for all five parameters estimated (Table 2; Figure S11). Nevertheless, the five-dimensional 95% C.I.'s do differ between the two recombination models, mainly due to the fact that t_{cur} is greater under the empirical hotspot model than under the Schaffner hotspot model.

The time we inferred that the CEU bottleneck began (~37,500 years ago) is too recent to coincide with the accepted dates for the out-of-Africa bottleneck, which is

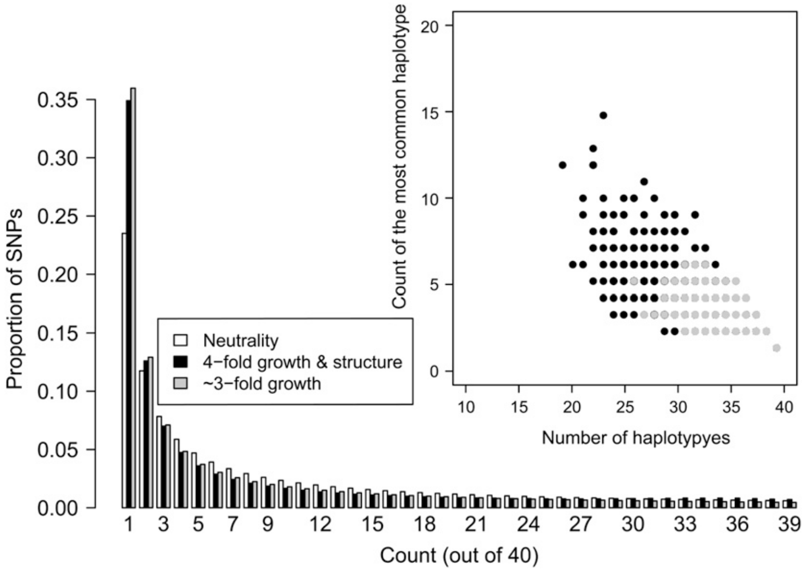


Figure 7.—Comparison of the expected SFS for population growth with ancestral structure to that expected with just population growth. The inset shows the frequency of the most common haplotype *vs.* the number of haplotypes for each simulated window for the same two demographic models. Note that the SFSs for the two models appear similar, but that there is an excess of windows with fewer haplotypes and where the most common haplotype is at high frequency in the growth with ancestral structure model (see discussion).

believed to have occurred 40,000–80,000 years ago (Reed and Tishkoff 2006). Thus, our estimate may coincide with an additional bottleneck associated with the founding of Europe, which likely took place 30,000–40,000 years ago (Barbujani and Goldstein 2004). Alternatively, our estimated start time of the bottleneck may represent an average time over several bottlenecks, including the out-of-Africa bottleneck and a more recent bottleneck, perhaps associated with the Last Glacial Maximum that began ~18,000 years ago (Barbujani and Goldstein 2004). Further work considering multiple European populations and multiple bottlenecks may help resolve this question.

How do our estimates of the bottleneck parameters for the CEU match with published estimates? Voight *et al.* (2005) may not be directly comparable to our study since their analysis considered a southern European sample and ours used individuals with northwestern European ancestry, and differences in haplotype diversity between these two regions have been noted (Lao *et al.* 2008; Auton *et al.* 2009). Nevertheless, our estimates of t_{mid} and $N_{\mathrm{mid}}/N_{\mathrm{cur}}$ fall within their confidence regions. Their bottleneck start times (40,000 years) and current and ancestral population size estimates (~10,000) also agree with ours. Our estimate of the time the bottleneck began is also consistent with that found using the decay of pairwise LD. Reich *et al.* (2001) found evidence for a bottleneck 800–1600 generations ago (our MLE is 1500 generations). We find evidence for a more severe bottleneck than previously estimated (Adams and Hudson 2004; Marth *et al.* 2004; Keinan *et al.* 2007), which could reflect the importance of considering LD-based information in the inference since these studies are all based on the frequency spectrum, and the other study that considers a summary of LD (Voight *et al.* 2005) cannot reject such a severe bottleneck. Alternatively, there could be some other important factors in European population history not captured by these simple bottleneck models that may affect the frequency spectrum and LD patterns differently. Finally, we cannot exclude the possibility that we have overestimated the bottleneck intensity due to greater heterogeneity in the recombination rate than what was included in our hotspot models. In short, improved confidence in the fine-scale genetic map will allow definitive ability to discriminate between these alternative scenarios.

While we have shown that haplotype statistics can be used to estimate demographic parameters from SNP genotype data, and it has been shown that the site frequency spectrum (SFS) will give misleading results when applied to genotype data without a correction for ascertainment bias (Nielsen *et al.* 2004; Clark *et al.* 2005), one important question is whether haplotype summary statistics will provide additional information that is important for inference when full genomewide resequencing data are available and it is possible estimate the SFS accurately. We examined whether the number of haplotypes and the count of the most common haplotype can discriminate between two different demographic models that have similar SFSs. We focused on a demographic model that included ancestral population structure since previous studies found that ancestral structure can result in an excess of long-range LD (Wall 2000b; Plagnol and Wall 2006). Specifically, we found that for certain subsets of the parameter space (the ms command lines giving the parameters used to generate Figure 7 are given as File S1), population growth with ancestral structure can create a similar SFS to that expected under population growth without ancestral structure. Close inspection of Figure 7 reveals a very slight uptick in the proportion of high-frequency derived SNPs in the population growth with structure SFS as compared to the growth without structure SFS, which is the expected signal of ancestral population structure. However, this effect is very subtle and in practice may be attributed to misidentification of the derived allele, rather than ancestral population structure (Hernandez *et al.* 2007a). Note that while the magnitudes of growth in the structure and panmictic cases are different, the growth with structure case still has an excess of low-frequency SNPs (Figure 7), which would often be interpreted as evidence for population growth. The inset in Figure 7 shows the count of the most common haplotype *vs.* the number of haplotypes for 10,000 windows simulated under the two demographic models described above. Note the growth with structure model has an excess of windows where the most common haplotype is at higher frequency and an excess of windows with a fewer number of haplotypes compared to the pure growth model. Thus, this is a case where two demographic models that cannot be readily differentiated on the basis of the SFS can be distinguished easily using haplotype patterns. The reason for this is as follows: population growth results in an excess of low-frequency SNPs, and for the parameters used here, population structure results in an excess of both low-frequency and high-frequency derived alleles. The resulting SFSs in Figure 7 have been affected by both these forces, but the excess of low-frequency SNPs is the predominant feature. Since ancestral population structure results in some genealogies having longer internal branches, any mutations occurring on these branches will be in LD with each other, leading to fewer distinct haplotypes in the sample and the most common haplotype occurring at higher frequency. Additionally, since the SFS treats all SNPs as being independent, haplotype patterns capture more information regarding the local genealogy within a window than the SFS does. In other words, many of the high-frequency derived SNPs in the sample are clustered in certain windows, but this pattern is missed in the SFS since it treats all SNPs as being exchangeable. Notably, summaries of the SFS performed on a local scale may be better at distinguishing these two models (Figure S12).

It is important to point out that the HCN statistic has been designed to be used on ascertained SNP data, where the number of SNPs in a particular window of the genome is affected by the ascertainment process. Consequently, we deliberately did not use the number of SNPs in constructing the HCN statistic. To analyze full-resequencing data in a haplotype framework, a more powerful approach would also make use of information about the number of SNPs in each window (Innan *et al.* 2005). The HCN statistic can be modified to include this information, suggesting that haplotype patterns based on full-resequencing data will be even more informative than described here. Thus statistics, like the HCN statistic, that capture information about the local genealogy of a region of the genome will remain relevant for demographic inference even when ascertainment bias is no longer an issue.

The example above suggests that combining the SFS and the HCN statistic may present a powerful approach to distinguish between complex demographic scenarios. Further work combining the two statistics for demographic inference is ongoing. Another possible extension of our method would be to jointly model two populations in an isolation–migration framework (proposed by Nielsen and Wakeley 2001), where the data are summarized by the HCN statistic for shared and population-specific haplotypes. Finally, instead of using standard coalescent simulations to find the expected HCN statistic for a given demographic scenario, we could approximate the coalescent using the sequentially Markov coalescent (McVean and Cardin 2005; Marjoram and Wall 2006). Doing so would reduce the computational burden of the method and would also allow for greater values of c_{window} to be used.

We thank J. Novembre, R. Hernandez, A. Boyko, A. Auton, J. Degenhardt, and M. Wright for helpful discussions and suggestions throughout the project and A. Auton and two reviewers for constructive comments on a previous version of the manuscript. K.E.L. was supported by a National Science Foundation Graduate Research Fellowship. This work was funded by National Institutes of Health grant RO1 HG003229 to A.G.C. and C.D.B.

LITERATURE CITED

Adams, A. M., and R. R. Hudson, 2004 Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics **168:** 1699–1712.

Anderson, E. C., and M. Slatkin, 2007 Estimation of the number of individuals founding colonized populations. Evolution **61:** 972–983.

Auton, A., K. Bryc, A. R. Boyko, K. E. Lohmueller, J. Novembre *et al.*, 2009 Global distribution of genomic diversity underscores a rich complex history of continental human populations. Genome Res. (in press).

Barbujani, G., and D. B. Goldstein, 2004 Africans and Asians abroad: genetic diversity in Europe. Annu. Rev. Genomics Hum. Genet. **5:** 119–150.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. **4:** e1000083.

Caicedo, A. L., S. H. Williamson, R. D. Hernandez, A. Boyko, A. Fledel-Alon *et al.*, 2007 Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. **3:** 1745–1756.

Clark, A. G., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. **7:** 111–122.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson and R. Nielsen, 2005 Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. **15:** 1496–1502.

Clark, V. J., S. E. Ptak, I. Tiemann, Y. Qian, G. Coop *et al.*, 2007 Combining sperm typing and linkage disequilibrium analyses reveals differences in selective pressures or recombination rates across human populations. Genetics **175:** 795–804.

Collins, F. S., L. D. Brooks and A. Chakravarti, 1998 A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. **8:** 1229–1231.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall *et al.*, 2006 A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. **38:** 1251–1260.

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.

Eberle, M. A., P. C. Ng, K. Kuhn, L. Zhou, D. A. Peiffer *et al.*, 2007 Power to detect risk alleles using genome-wide tag SNP panels. PLoS Genet. **3:** 1827–1837.

Ewens, W. J., 1973 Testing for increased mutation rate for neutral alleles. Theor. Popul. Biol. **4:** 251–258.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

Fagundes, N. J., N. Ray, M. 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.

Fearnhead, P., and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rate. J. R. Stat. Soc. B **64:** 1–64.

Francois, O., M. G. Blum, M. Jakobsson and N. A. Rosenberg, 2008 Demographic history of European populations of *Arabidopsis thaliana*. PLoS Genet. **4:** e1000075.

Griffiths, R. C., and S. Tavaré, 1994 Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

Hellenthal, G., and M. Stephens, 2007 msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics **23:** 520–521.

Hellenthal, G., J. K. Pritchard and M. Stephens, 2006 The effects of genotype-dependent recombination, and transmission asymmetry, on linkage disequilibrium. Genetics **172:** 2001–2005.

Hernandez, R. D., S. H. Williamson and C. D. Bustamante, 2007a Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. **24:** 1792–1800.

Hernandez, R. D., M. J. Hubisz, D. A. Wheeler, D. G. Smith, B. Ferguson *et al.*, 2007b Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. Science **316:** 240–243.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin *et al.*, 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072–1079.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland, MA.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

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

Innan, H., K. Zhang, P. Marjoram, S. Tavaré and N. A. Rosenberg, 2005 Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. Genetics **169:** 1763–1777.

International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449:** 851–861.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature **451:** 998–1003.

Jeffreys, A. J., R. Neumann, M. Panayi, S. Myers and P. Donnelly, 2005 Human recombination hot spots hidden in regions of strong marker association. Nat. Genet. **37:** 601–606.

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.

Keinan, A., J. C. Mullikin, N. Patterson and D. Reich, 2007 Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat. Genet. **39:** 1251–1255.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

Lao, O., T. T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf *et al.*, 2008 Correlation between genetic and geographic structure in Europe. Curr. Biol. **18:** 1241–1248.

Leblois, R., and M. Slatkin, 2007 Estimating the number of founder lineages from haplotypes of closely linked SNPs. Mol. Ecol. **16:** 2237–2245.

Li, N., and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165:** 2213–2233.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science **319:** 1100–1104.

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez *et al.*, 2008 Proportionally more deleterious genetic variation in European than in African populations. Nature **451:** 994–997.

Marjoram, P., and S. Tavaré, 2006 Modern computational approaches for analysing molecular genetic variation data. Nat. Rev. Genet. **7:** 759–770.

Marjoram, P., and J. D. Wall, 2006 Fast "coalescent" simulation. BMC Genet. **7:** 16.

Marth, G. T., E. Czabarka, J. Murvai and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166:** 351–372.

McVean, G. A., and N. J. Cardin, 2005 Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond. B Biol. Sci. **360:** 1387–1393.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

Myers, S., C. Fefferman and N. Patterson, 2008 Can one learn history from the allelic spectrum? Theor. Popul. Biol. **73:** 342–348.

Myers, S., L. Bottolo, C. Freeman, G. McVean and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321–324.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158:** 885–896.

Nielsen, R., M. J. Hubisz and A. G. Clark, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics **168:** 2373–2382.

Plagnol, V., and J. D. Wall, 2006 Possible ancestral structure in human populations. PLoS Genet. **2:** e105.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1–14.

Ptak, S. E., D. A. Hinds, K. Koehler, B. Nickel, N. Patil *et al.*, 2005 Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet. **37:** 429–434.

Reed, F. A., and S. A. Tishkoff, 2006 African human diversity, origins and migrations. Curr. Opin. Genet. Dev. **16:** 597–605.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. **15:** 1576–1583.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. **17:** 520–526.

Thornton, K., and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster.* Genetics **172:** 1607–1619.

Voight, B. F., A. M. Adams, L. A. Frisse, Y. Qian, R. R. Hudson *et al.*, 2005 Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Natl. Acad. Sci. USA **102:** 18508–18513.

Wall, J. D., 2000a A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Wall, J. D., 2000b Detecting ancient admixture in humans using sequence polymorphism data. Genetics **154:** 1271–1279.

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.

Winckler, W., S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald *et al.*, 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. Science **308:** 107–111.

Zeng, K., S. Mano, S. Shi and C. I. Wu, 2007 Comparisons of site- and haplotype-frequency methods for detecting positive selection. Mol. Biol. Evol. **24:** 1562–1574.

Communicating editor: N. Takahata

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.108.099275/DC1

Methods for Human Demographic Inference Using Haplotype Patterns From Genomewide Single-Nucleotide Polymorphism Data

Kirk E. Lohmueller, Carlos D. Bustamante and Andrew G. Clark

doi:10.1534/genetics.108.099275

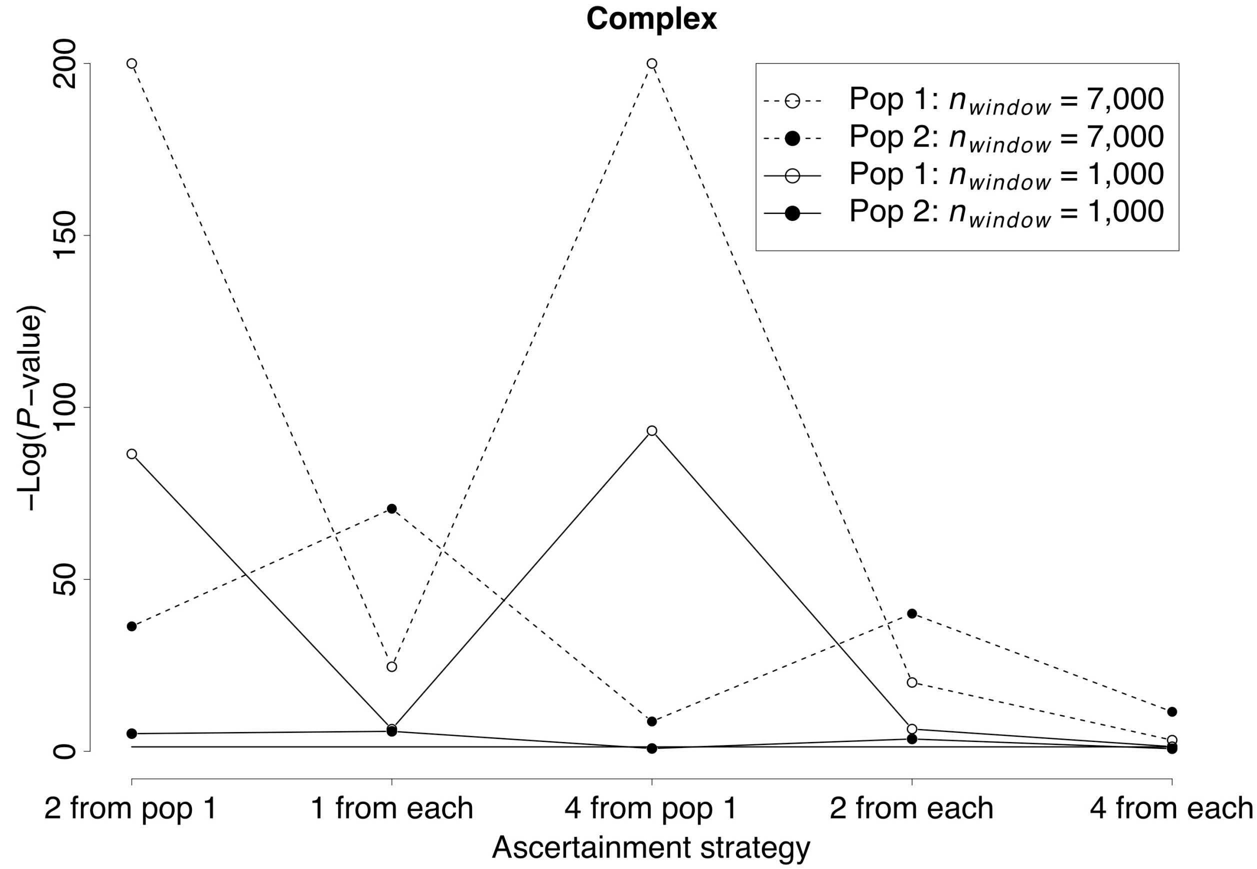


FIGURE S1.—Log$_{10}$ *P*-value of the goodness-of-fit test comparing the H_{pair} statistic under different SNP ascertainment schemes (shown on the x-axis) to that with complete ascertainment for the complex demographic model. Here a sample size of 40 chromosomes from each population is used. The solid horizontal line denotes the 5% significance cutoff. *P*-values <10^{-200} are set to 10^{-200}.

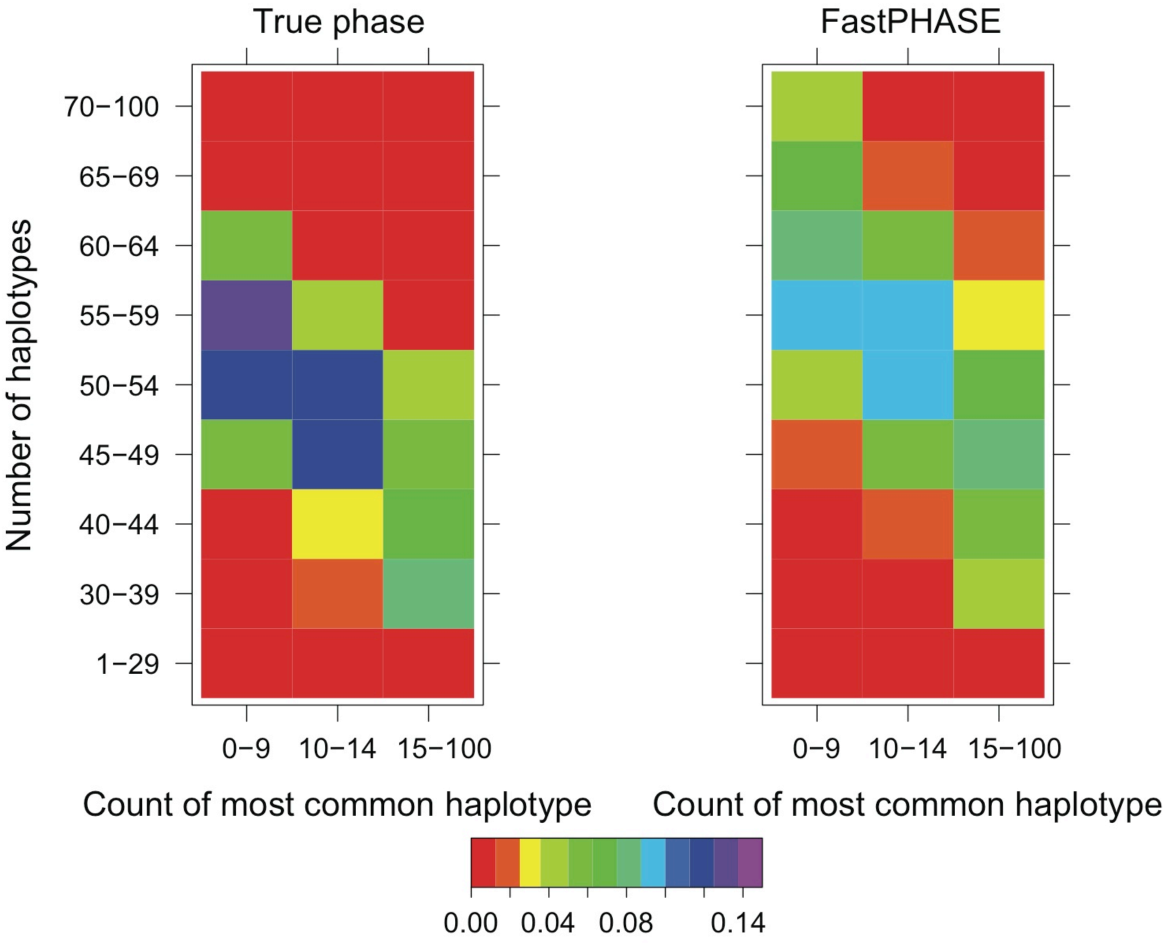


FIGURE S2.—Effect of haplotype phase uncertainty on the *HCN* statistic. The *HCN* for a bottleneck model (see File S1) when haplotype phase is known (left) and inferred using fastPHASE (right).

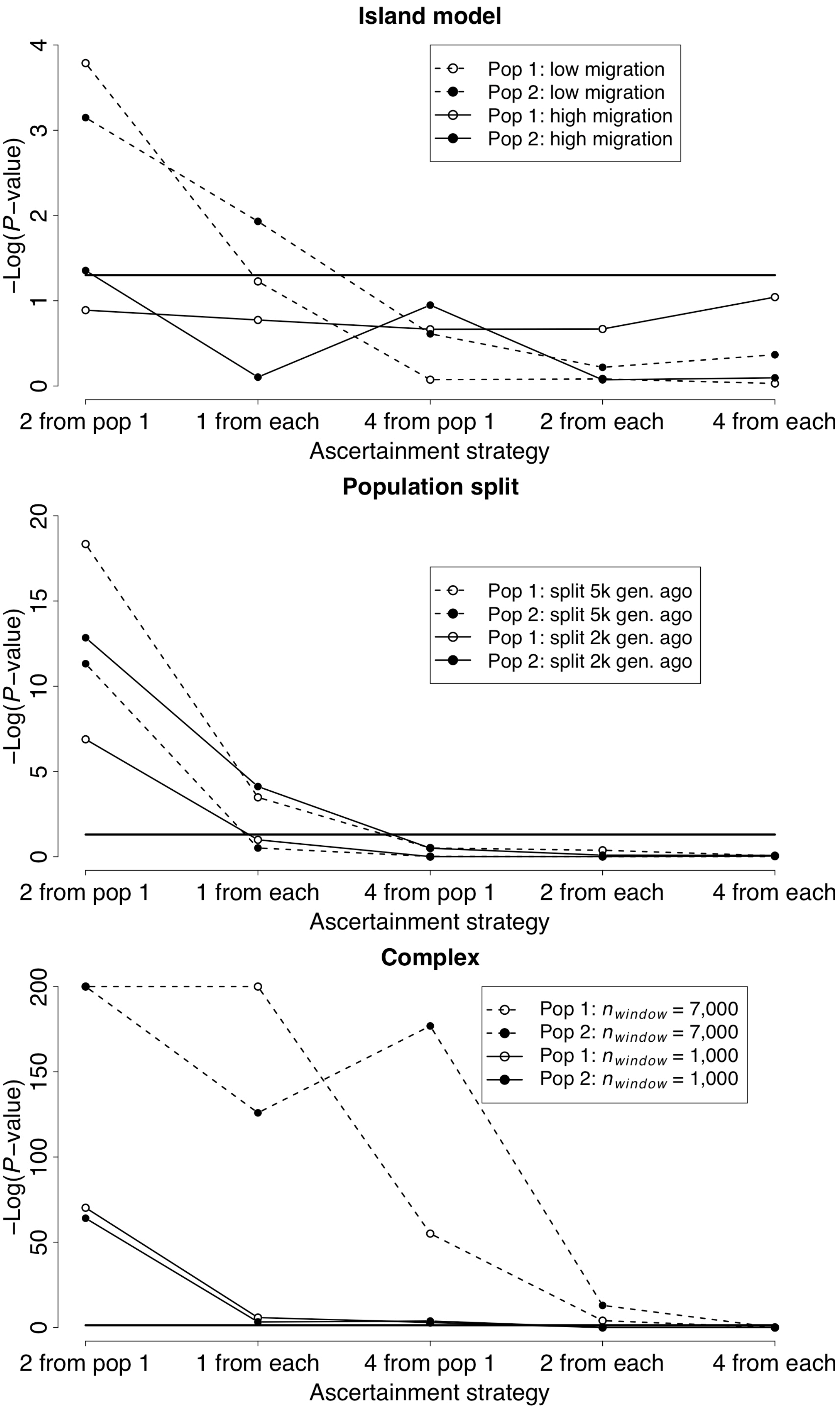


FIGURE S3.—Log$_{10}$ *P*-value of the goodness-of-fit test comparing the *HCN* statistic under different SNP ascertainment schemes (shown on the x-axis) to that with complete ascertainment for three different demographic models (see File S1). Here a sample size of 40 chromosomes from each population is used. The solid horizontal line in each figure denotes the 5% significance cutoff. *P*-values $<10^{-200}$ are set to 10^{-200}.

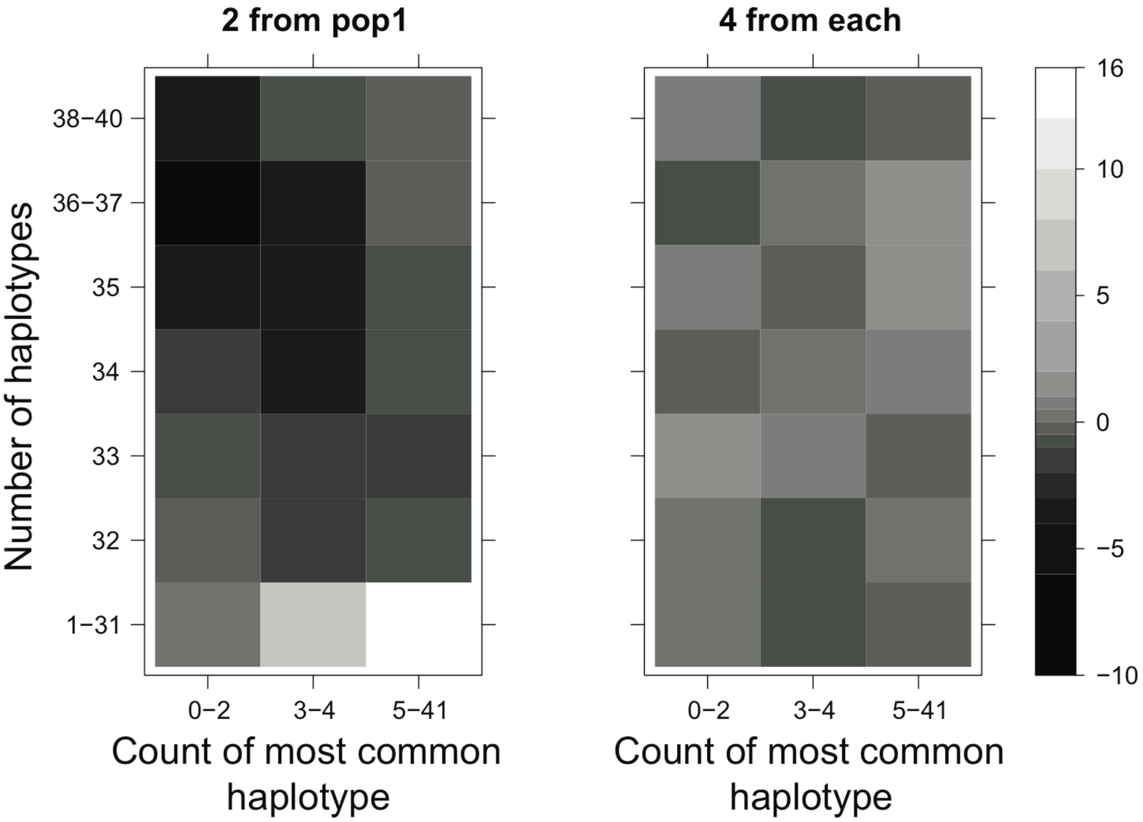


FIGURE S4.—Plot of Pearson's residuals comparing the *HCN* statistic for two different ascertainment strategies to the expected *HCN* having complete SNP ascertainment for the bottlenecked population (population 1) in the complex demographic model. The two SNP ascertainment strategies compared are SNP ascertainment using 2 chromosomes from population 1 ("2 from pop 1") and ascertainment using 4 chromosomes from population 1 and 4 from population 2 ("4 from each"). Darker colors indicate a deficit of windows in the particular cell as compared to complete ascertainment. Lighter colors indicate an excess of windows in the particular cell as compared to complete ascertainment.

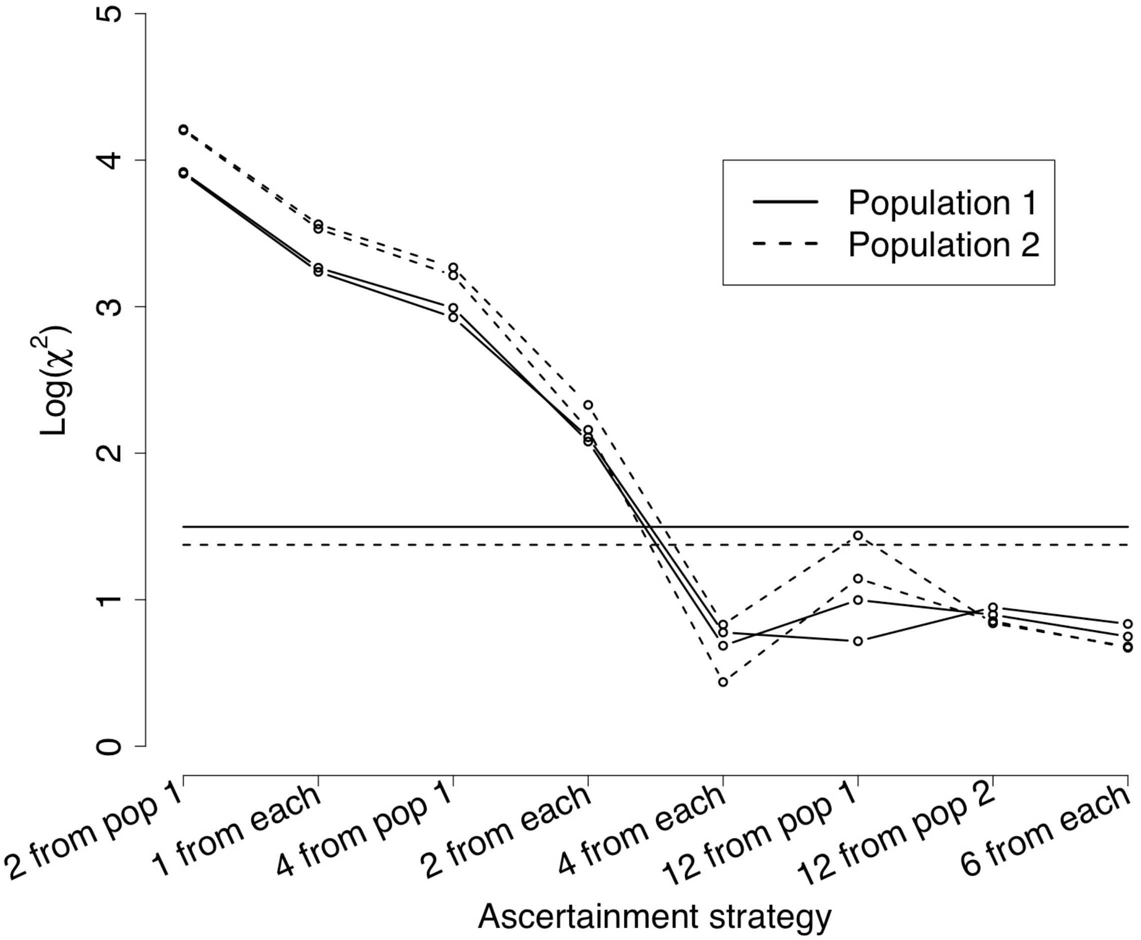


FIGURE S5.—Log$_{10}$ of the χ^2 statistic for the goodness-of-fit test comparing the *HCN* statistic under different SNP ascertainment schemes (shown on the x-axis) to that with complete ascertainment for the complex demographic model. Here a sample size of 120 chromosomes from each population is used. Note that the SNP discovery sample sizes used here differ from those in Figures S1 and S3. The horizontal lines denote the 5% significance cutoff for population 1 (solid) and population 2 (dashed). The two curves for each population are from two entirely independent replicates of the entire process (see File S1) to assess stochastic variance.

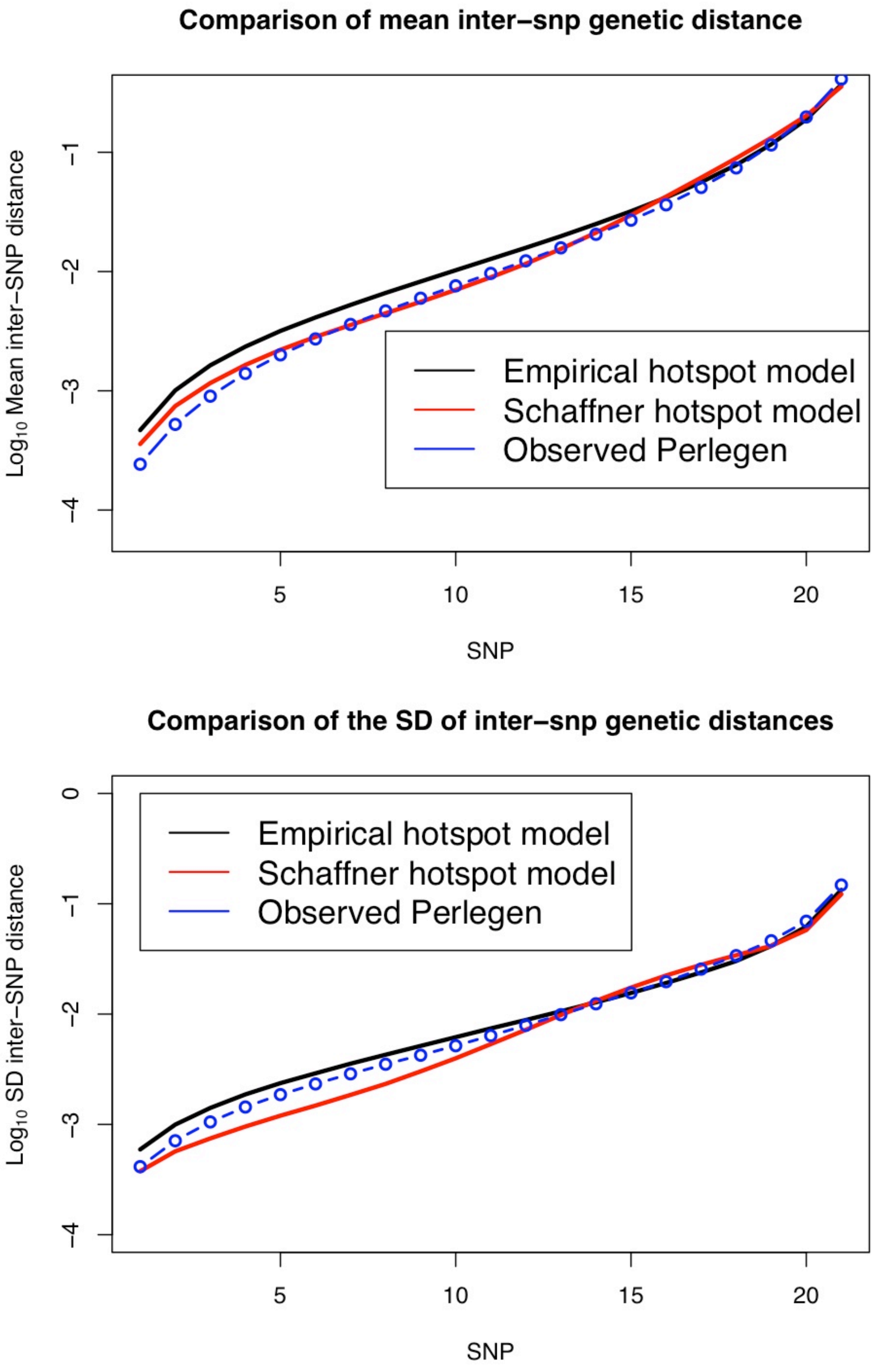


FIGURE S6.—Comparison between the mean and standard deviation (SD) across all 8833 windows of the observed inter-SNP genetic distances (as defined by the LDhat genetic map) and the mean genetic distances simulated using the modified Schaffner hotspot model and the empirical hotspot model (see Methods). The left-most point in the top figure represents the mean of the smallest inter-SNP distance, averaged over all windows, the second point, the second smallest inter-SNP distance, and so on. The actual *HCN* statistic used for inference was averaged over 10 different *HCN* statistics, each of which was generated from a different random sub-set of SNPs from each window (see Methods). Here the observed and simulated inter-SNP genetic distances are based on selecting one random set of SNPs per window. The simulated inter-SNP genetic distances were determined assuming a constant population size, *N*=10,000, and re-scaling genetic distance for each window such that $\hat{c}_{window} = 0.25$ cM.

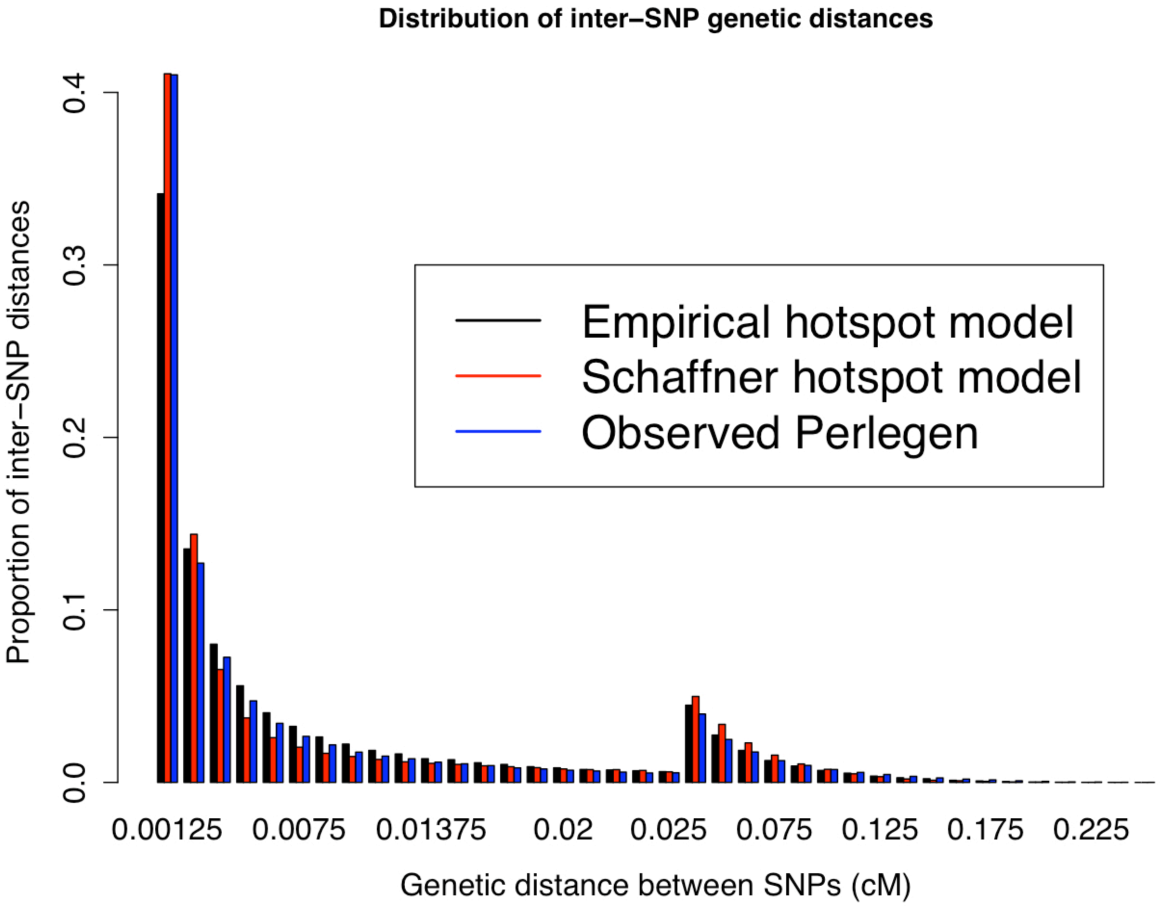
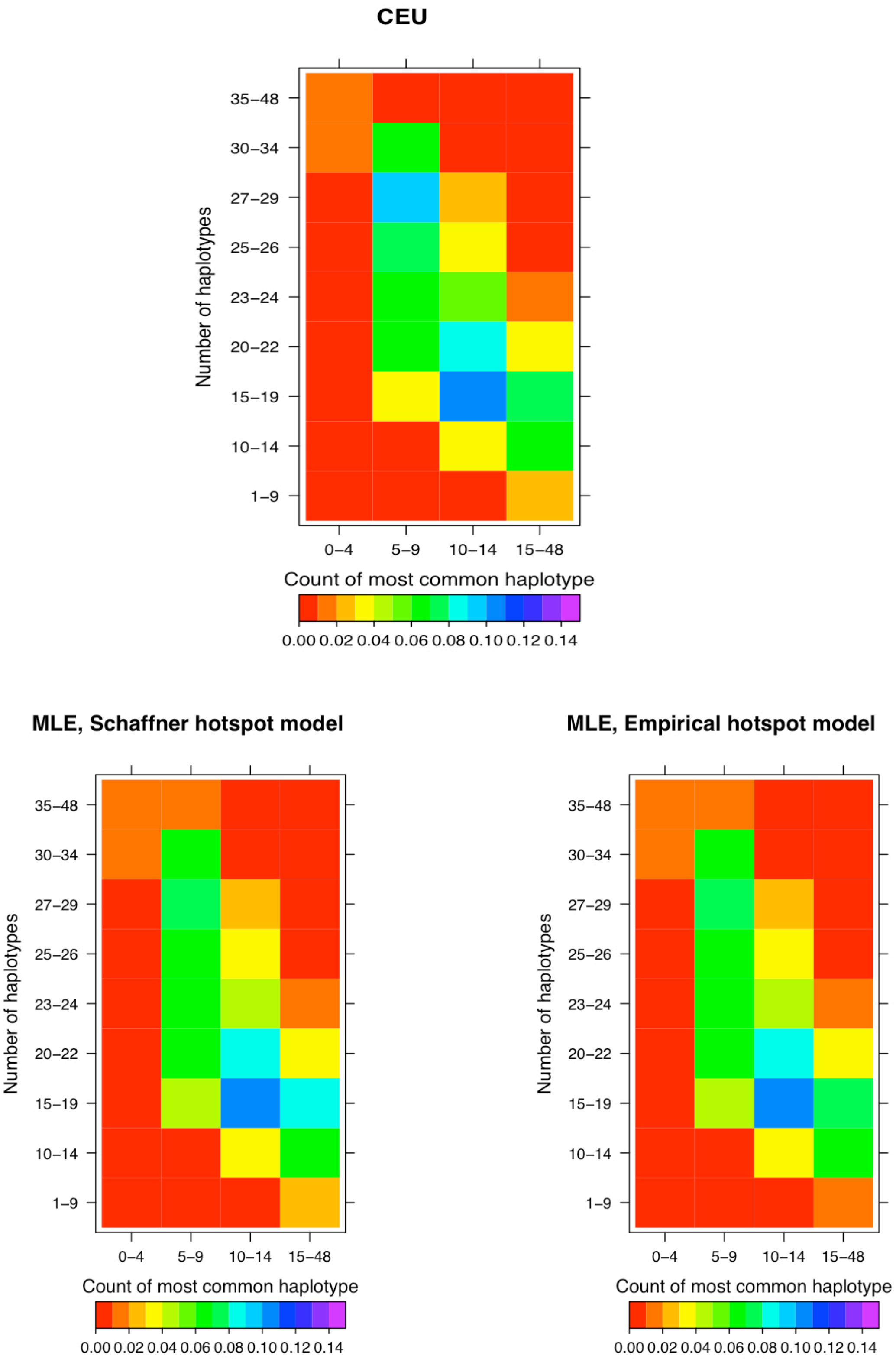


FIGURE S7.—Comparison of the distribution of inter-SNP genetic distances in the Perlegen data (from the LDhat genetic map) with the Schaffner and empirical hotspot models (see Methods). The distribution is tabulated over all 8833 windows across the genome. The increased proportion in the bin after 0.025 cM is due to the change in scale of the bins. As noted in Figure S6, here the observed and simulated inter-SNP genetic distances are based on selecting one random set of SNPs per window. The simulated inter-SNP genetic distances were determined assuming a constant population size, N=10,000, and re-scaling genetic distance for each window such that $\hat{c}_{window} = 0.25$ cM.


CEU

MLE, Schaffner hotspot model

MLE, Empirical hotspot model

FIGURE S8.—Observed *HCN* statistic for the Perlegen CEU sample and the *HCN* statistics for the best-fitting demographic models based on the Schaffner hotspot model and the empirical hotspot model. Windows based on genetic distance were defined using the LDHat genetic map (see Methods). See Table 2 for the parameter values generating the best-fitting *HCN* statistics. Note, the bins shown in the figure were the same ones used when inferring parameters.

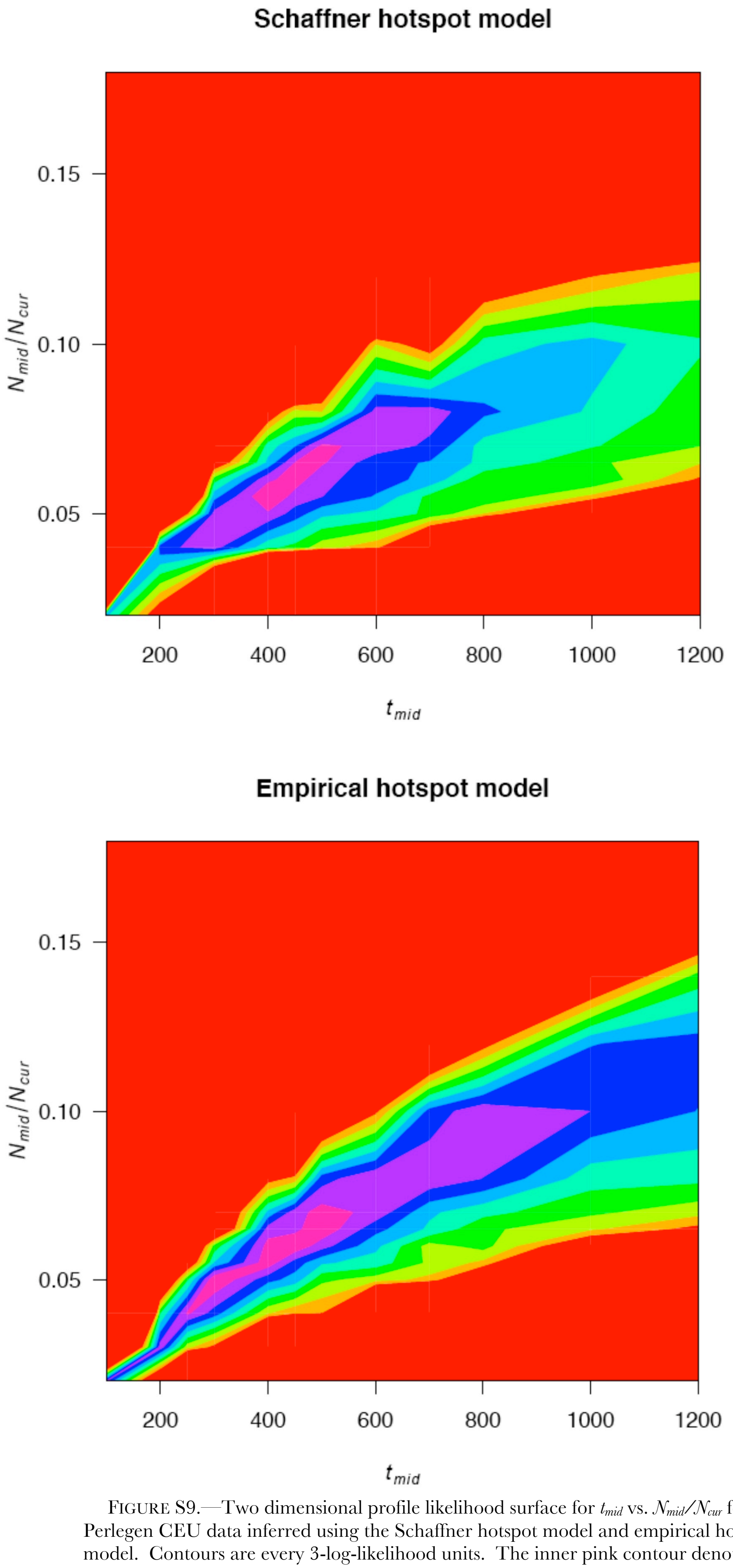
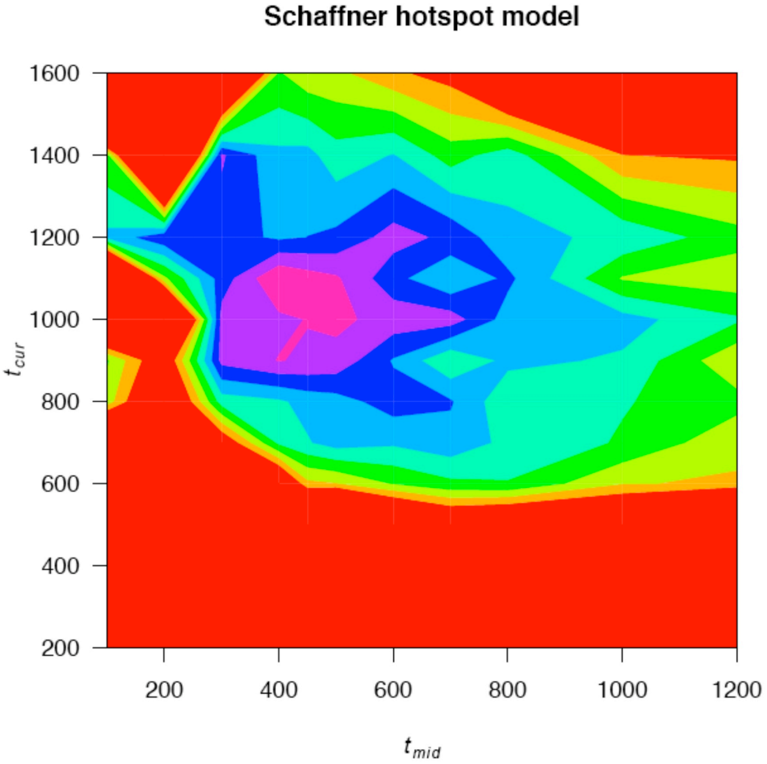


FIGURE S9.—Two dimensional profile likelihood surface for t_{mid} vs. N_{mid}/N_{cur} for the Perlegen CEU data inferred using the Schaffner hotspot model and empirical hotspot model. Contours are every 3-log-likelihood units. The inner pink contour denotes the region of points where the log-likelihood is < 3-log likelihood units from the MLE.



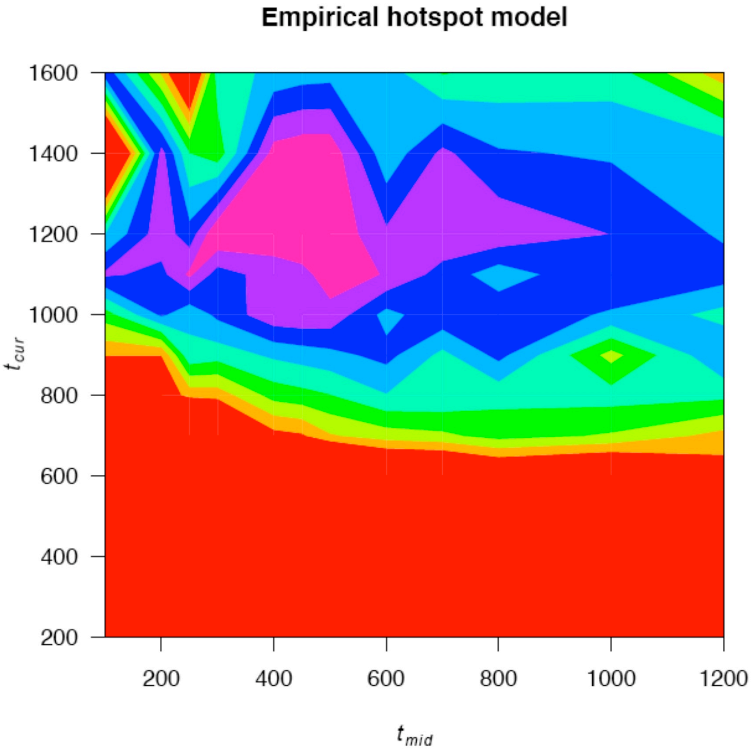


FIGURE S10.—Two-dimensional profile likelihood surface for t_{cur} vs. t_{mid} for the Perlegen CEU data inferred using the Schaffner hotspot model and empirical hotspot model. Contours are every 3-log-likelihood units. The inner pink contour denotes the region of points where the log-likelihood is < 3-log likelihood units from the MLE. Note the jaggedness of the contours is due to the relatively course grid used to estimate parameters combined with Monte Carlo error.

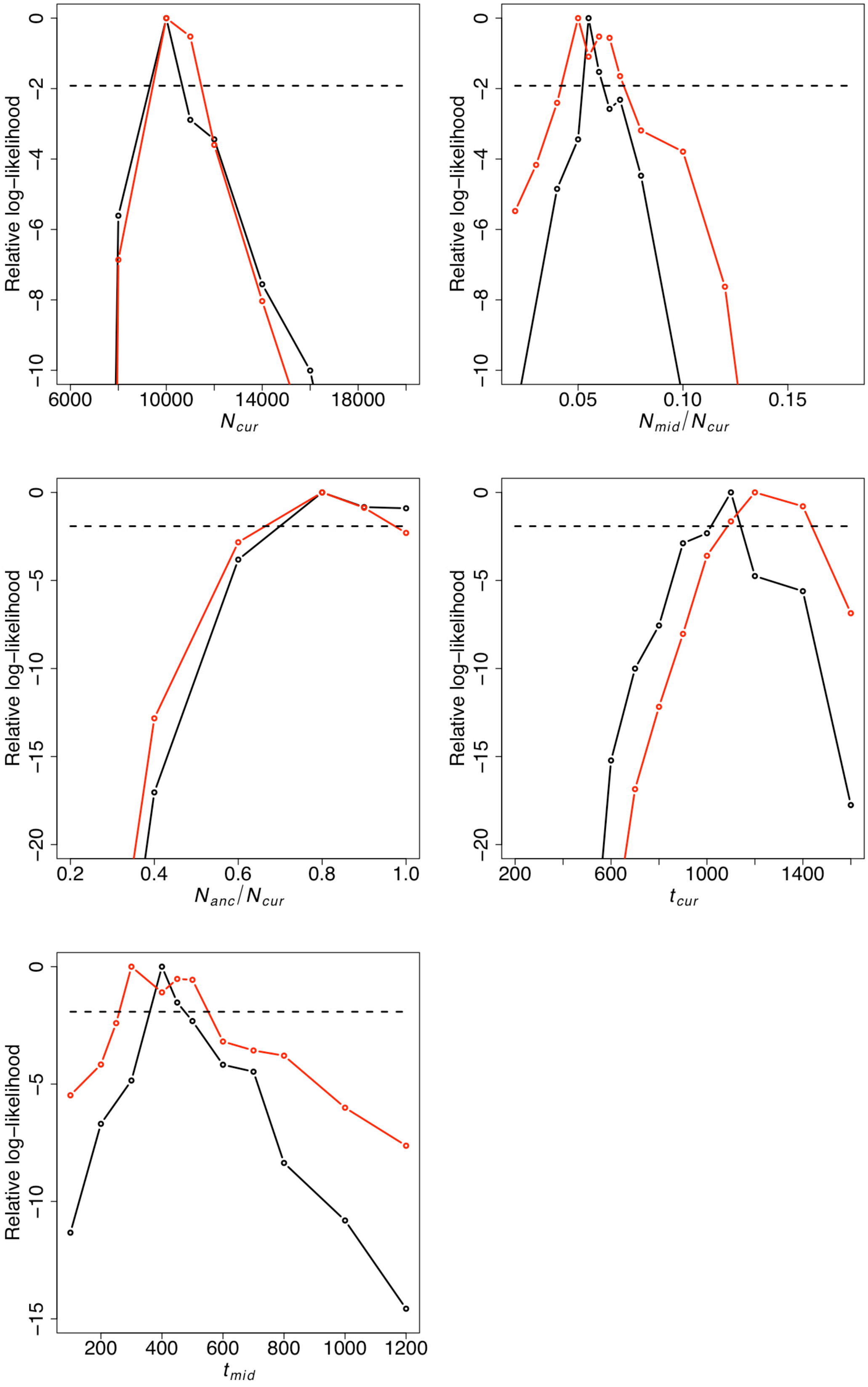


FIGURE S11.—Likelihood profiles for the five CEU bottleneck parameters inferred using the Schaffner hotspot model (black) and the empirical hotspot model (red). The dashed line denotes the approximate 95% confidence interval.

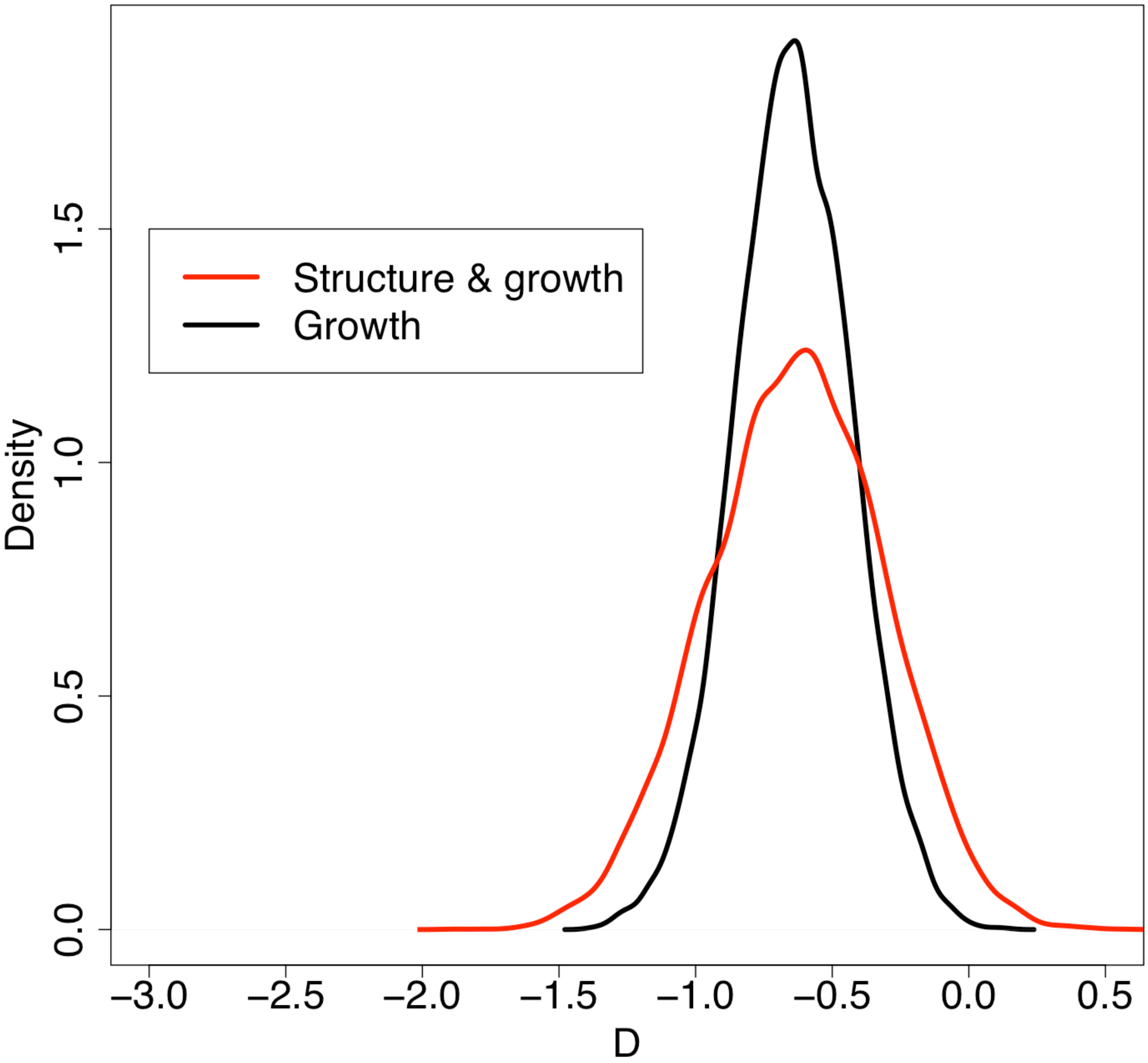


FIGURE S12.—Distribution of Tajima's *D* for 10,000 independent 250 kb windows (c_{window}=0.25 cM) simulated under a growth with ancestral structure model (red) and a growth without ancestral structure model (black). Note that while means of the two distributions are similar, the variance is greater in the growth with ancestral structure case.

File S1

Haplotype phase uncertainty: Since the *HCN* statistic reflects haplotype patterns, and for many genome-wide SNP datasets consisting of unrelated individuals, haplotype phase would need to be computationally inferred, we wanted to determine how this inference affected the *HCN* statistic. To do this, we simulated 1000 windows with c_{window} = 0.25 cM in a sample size of 100 chromosomes from a bottleneck demographic history (N_{cur}=10,000, N_{mid}/N_{cur}=0.1, N_{anc}/N_{cur}=1.0, t_{cur}=800 generations, t_{mid}=800 generations), where n_{snp}=40. For each window, we then randomly paired the chromosomes into diploid individuals and generated diploid genotypes at each SNP. We next inferred haplotypes from these genotypes using a popular phasing method, fastPHASE (Scheet and Stephens 2006), with the default settings. We chose to use fastPHASE since its performance is comparable to one of the better performing phasing algorithms, PHASE, yet is fast enough to be run on genome-wide datasets. Finally, we compared the *HCN* statistic for the phase-known dataset to the phase-inferred dataset.

Figure S2 shows the *HCN* statistic for a bottleneck model when the correct haplotype phase is known with certainty (left) and when haplotype phase is inferred using fastPHASE (right). The *HCN* from phase-inferred haplotypes has a broader distribution than when haplotype phase is known. In particular the *HCN* constructed using the phase-inferred haplotypes has an excess of windows having many haplotypes (green squares in bins "65-90" and "70-100") as compared to the known phase *HCN*. Although it is a bit more subtle, the *HCN* using the phase-inferred haplotypes also has an excess of windows where the most common haplotype is at a high frequency. This can be seen by the yellow square in the phase inferred haplotypes where there was an orange square in the phase-known *HCN*. Thus, inferring haplotype phase will result in an *HCN* statistic that is slightly different from the true phase-known *HCN*.

Ascertainment bias: To evaluate how the *HCN* statistic is influenced by SNP ascertainment bias, we conducted a variety of coalescent simulations under different demographic models and SNP ascertainment strategies. We then compared the *HCN* from the different ascertainment strategies to the *HCN* with complete SNP ascertainment. We also examined whether another haplotype statistic, H_{pair}, is affected by ascertainment bias.

Since we wanted to address the question of whether discovering SNPs in one population and then typing them in a second population is more biased than selecting the SNPs in the genotyped population, we considered demographic models that consisted of two populations. Briefly, we considered a finite island model (where each population has size N_e=10,000) with a low rate of migration between populations ($4N_em$=9) and high rate ($4N_em$ =99), a population split model where the two populations (each of size N_e =10,000) split 2000 or 5000 generations ago, and a complex model where the two populations split 5000 generations ago and there was a bottleneck in population one (N_{mid}/N_{cur}=0.1, t_{cur}=800, t_{mid}=800). The last model can be thought of as a very crude approximation of the contrast between European (as population 1) and African (as population 2) human populations. For each of these demographic models, we simulated a "genotype" sample of 40 chromosomes from each of the two populations as well as a SNP discovery sample consisting of an additional four chromosomes from each population. We then examined five different SNP discovery protocols shown in Table S1a. These ascertainment strategies are reasonable ones for many of the human genome-wide SNP datasets like HapMap where many of the SNPs were discovered by comparing two sequencing reads (as in phase I) or from a polymorphism discovery panel with a few chromosomes from multiple populations (phase II SNPs discovered by Perlegen; Hinds *et al.* 2005; International HapMap Consortium. 2005; International HapMap Consortium 2007). For each ascertainment scheme we simulated 1000 windows 500 kb in size with a uniform recombination rate of 1 cM/Mb (c_{window}=0.5 cM) and μ =1 x 10^{-8} per base-pair per generation. To determine whether ascertainment bias becomes a problem for larger datasets containing more than 1000 windows (n_{window}>1000), we also simulated an additional dataset under the complex demographic history consisting of 7000 windows 250 kb in size with uniform recombination rate of 1cM/Mb (c_{window}=0.25 cM) and μ =1 x 10^{-8} per base-pair per generation. Finally, we considered the case where the genotype sample consisted of 120 chromosomes from each population (to mimic the HapMap CEU and YRI samples) and we had data from 7000 windows 250 kb in size with uniform recombination rate of 1cM/Mb (c_{window}=0.25 cM) and μ =1 x 10^{-8} per base-pair per generation. For this set of simulations, the SNP discovery set consisted of 12 chromosomes per population. Here we considered eight ascertainment strategies shown in Table S1b.

For each demographic scenario and ascertainment scheme, we selected a sub-set of 40 SNPs having MAF >10% (n_{snp}=40). If a window had fewer than 40 SNPs, it was dropped from the analysis. We then generated 10 different *HCN* statistics, each with a set of 10 randomly selected SNPs from each window, and then averaged them to generate the final *HCN* statistic. We compared the average *HCN* statistic to the expected statistic under complete ascertainment using a chi-square goodness of fit test. To generate the expected *HCN* statistic under complete ascertainment, we simulated an additional 10^5 windows each consisting of 40 chromosomes and n_{snp}=40 and averaged over 10 different *HCN* statistics, each using a set randomly selected SNPs for each window. From these simulations, we computed the expected *HCN* statistic. Note that, when conducting the chi-square goodness of fit tests, we binned the *HCN* statistic so that we did not have any expected cell counts $\leq$ 5. For the complex demographic history using 7000 windows (for genotype sample sizes of 40 and 120 chromosomes per population) n_{snp}=20 instead of 40.

We find that for all demographic models examined, except for the island model with a high migration rate, ascertainment of SNPs using two discovery chromosomes from one population results in a different *HCN* statistic than that expected under complete ascertainment (Figure S3). This is shown by the low *P*-values for the goodness of fit tests comparing the *HCN* statistic using SNPs polymorphic in two discovery chromosomes to the expected *HCN* under complete ascertainment. The *HCN* statistic constructed from SNPs ascertained in two chromosomes has an excess of windows having a small number of haplotypes and an excess of windows where the most common haplotype is at higher frequency as compared to the complete ascertainment case (Figure S4).

The reason for this pattern is that SNPs polymorphic in the two chromosome discovery sample must occur on branches of the genealogy where one of the two discovery chromosomes carries the mutant allele and the other does not. These branches are a small fraction of the total area of the genealogy. This fact will result in SNPs that are polymorphic in the discovery sample tending to occur on the same branches of the genealogy more often than expected without ascertainment bias. SNPs that co-occur on the same branches of the genealogy will be in LD with each other, resulting in there being fewer haplotypes and the most common haplotype occurring at higher frequency than in the case of less LD among SNPs. When considering SNPs discovered from two chromosomes from the first population, the *HCN*s in both populations differ from the expected *HCN*, suggesting that SNP discovery using two chromosomes does poorly, regardless of whether those two chromosomes are from the population of interest.

SNP discovery using one chromosome from each population is a slight improvement to SNP discovery using two SNPs from population 1. However, we note that for many of the demographic models considered here (Figure S3), the *HCN* constructed from ascertained SNPs differs significantly from the expected *HCN* under complete ascertainment.

However, SNP discovery using four chromosomes from the first population results in a better fit to the expected *HCN* for most of the demographic models considered. In all cases, except for the complex demographic model, the *HCN*s constructed from ascertained SNPs are quite consistent with the expected *HCN* under complete SNP ascertainment. This finding holds true even for the second population which had no SNP discovery, again illustrating that if the two populations have similar demographic histories, ascertainment sample depth may be more important than which population the SNPs were ascertained from in terms of matching the *HCN* statistic. This pattern, however, does not hold for the complex demographic model. Here SNP discovery using four SNPs from the bottlenecked population (population 1) results in a poor fit to the expected *HCN* statistic. The reason for this is that the four SNP discovery chromosomes from the bottlenecked population are less representative of the diversity in the second population that did not undergo a bottleneck (population 2). If, again for the complex demographic scenario, instead of taking four discovery chromosomes from the first population, we take two discovery chromosomes from each population, the *HCN* statistic from the ascertained SNPs more closely matches the expected *HCN* statistic. However, note that if the number of windows of the genome considered is large (n_{window}=7000), the effects of ascertainment bias are still present.

The *HCN* statistic generated using a four chromosome SNP discovery sample from both of the two populations results in an excellent fit to the expected *HCN* for both populations in all demographic scenarios considered. We also found an adequate fit of the expected *HCN* to the observed *HCN* when considering a larger dataset under the complex demographic model. This finding is especially encouraging since the larger number of windows in the dataset ($n_{winidow}$=7000 as compared to 1000 in previous datasets) will have more power to detect subtle departures in the fit of the model. Thus, for the demographic models considered here using n=40 chromosomes, the *HCN* statistic using SNP discovery sample of ≥ 4 chromosomes from at least two populations is not significantly different from the true *HCN* statistic.

We also examined whether ascertainment bias is a more severe problem when the genotype sample is >40 chromosomes. To do this, we repeated the above approach for the complex demographic model using n=120 chromosomes and considering larger SNP discovery sample sizes (Figure S5). We find that small SNP discovery sample sizes (here <8 chromosomes) result in significant differences between the *HCN* under SNP ascertainment and the expected *HCN*. However, for larger SNP discovery sample sizes, the effect disappears. This holds true even for the population that has no SNP discovery chromosomes (*e.g.* the solid line at "12 from 2"). To assess the amount of evolutionary variance in the whole process, we performed two completely independent sets of simulations for these demographic and ascertainment models. The results of both replicates are shown in Figure S5. Encouragingly, the variance is reasonably low since the two solid curves (and dotted curves) are similar to each other.

We also evaluated whether the H_{pair} statistic was robust to ascertainment bias. As shown in Figure S1, for all demographic models and ascertainment conditions considered, H_{pair} was severely affected by SNP ascertainment bias. Ascertainment bias results in H_{pair} being higher than expected. This finding is analogous to the effect of ascertainment bias on π, the average number of pairwise differences among DNA sequences (Nielsen *et al.* 2004). Ascertainment bias results in an excess of intermediate-frequency SNPs, which results in there being more pairwise differences between haplotypes than low-frequency SNPs do. Thus, by preferentially selecting intermediate-frequency SNPs, H_{pair} becomes inflated.

Interestingly, we find that for the cases where SNPs were ascertained in population 1 exclusively, the fit of the H_{pair} statistic under ascertainment bias to the expected H_{pair} statistic is actually worse in population 1—the population where the SNPs were discovered in—than in population 2. This pattern is seen for both n_{widnow}=1000 and for n_{widnow} =7000 and for both the "2 from pop 1" and the "four from pop 1" ascertainment strategies. One possible explanation for this counter-intuitive pattern is that the ascertained SNPs from population 1 are more likely to be at intermediate frequency in population 1 (as discussed above), but may have drifted to lower or higher frequency in the second population, resulting in those SNPs being more representative of the true frequency spectrum in that population.

Here are the ms commands to generate the *HCN* statistic in Figure 7:

Growth and Structure:

```
./ms 40 10000 -t 400 -r 400 250000 -F 4 -es 0.00625 1 0.1 -eM 0.00625 5 -eN 0.00625 0.5 -eN 0.025 0.125 -ej 0.625 2 1 -eM
0.625 0 -eN 0.625 0.25
```

Growth:

```
./ms 40 10000 -t 400 -r 400 250000 -F 4 -en 0.01925 1 0.303333
```

Literature Cited

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin *et al.*, 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072-1079.

International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449:** 851-861.

International HapMap Consortium, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

Nielsen, R., M. J. Hubisz and A. G. Clark, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics **168:** 2373-2382.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. **78:** 629-644.

TABLE S1

Summary of SNP ascertainment strategies

Abbreviation	Ascertainment sample description
	a. Supplemental Figures 1 & 3; *n*=40
2 from pop 1	2 chromosomes from population 1
1 from each	1 chromosome from population 1 and 1 chromosome from population 2
4 from pop 1	4 chromosomes from population 1
2 from each	2 chromosomes from population 1 and 2 chromosomes from population 2
4 from each	4 chromosomes from population 1 and 4 chromosomes from population 2
	b. Supplemental Figure 5; *n*=120
2 from pop 1	2 chromosomes from population 1
1 from each	1 chromosome from population 1 and 1 chromosome from population 2
4 from pop 1	4 chromosomes from population 1
2 from each	2 chromosomes from population 1 and 2 chromosomes from population 2
4 from each	4 chromosomes from population 1 and 4 chromosomes from population 2
12 from pop 1	12 chromosomes from population 1
12 from pop 2	12 chromosomes from population 2
12 from each	12 chromosomes from population 1 and 12 chromosomes from population 2