Inferring the History of Population Size Change from Genome-Wide SNP Data

Christoph Theunert,*[†,1] Kun Tang,[†,2] Michael Lachmann,[1] Sile Hu,[2] and Mark Stoneking[1]

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2]Group of Human Genetic Variation, CAS-MPG Partner Institute and Key Laboratory for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, People's Republic of China

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: christoph_theunert@eva.mpg.de.

Associate editor: John Novembre

Abstract

Dense, genome-wide single-nucleotide polymorphism (SNP) data can be used to reconstruct the demographic history of human populations. However, demographic inferences from such data are complicated by recombination and ascertainment bias. We introduce two new statistics, allele frequency-identity by descent (AF-IBD) and allele frequency-identity by state (AF-IBS), that make use of linkage disequilibrium information and show defined relationships to the time of coalescence. These statistics, when conditioned on the derived allele frequency, are able to infer complex population size changes. Moreover, the AF-IBS statistic, which is based on genome-wide SNP data, is robust to varying ascertainment conditions. We constructed an efficient approximate Bayesian computation (ABC) pipeline based on AF-IBD and AF-IBS that can accurately estimate demographic parameters, even for fairly complex models. Finally, we applied this ABC approach to genome-wide SNP data and inferred the demographic histories of two human populations, Yoruba and French. Our results suggest a rather stable ancestral population size with a mild recent expansion for Yoruba, whereas the French seemingly experienced a long-lasting severe bottleneck followed by a drastic population growth. This approach should prove useful for new insights into populations, especially those with complex demographic histories.

Key words: genome-wide SNP data, demographic inference, identity by descent, ascertainment bias, population size changes, approximate Bayesian computation.

Introduction

The genomic diversity of a population is shaped by a complex interplay of a large variety of demographic events including population growth or decline, migration of individuals between different populations, and population splits or divergence.

It is, therefore, desirable to estimate past population size changes as a function of time. Methods exist for making such inferences from nonrecombinant sequence data. "Skyline plot" methods are a collection of sophisticated nonparametric methods that infer the population size trajectories by reconstructing the underlying genealogy from mitochondrial DNA (Pybus et al. 2000; Strimmer and Pybus 2001; Drummond et al. 2002, 2005; Minin et al. 2008) and from multiple unlinked loci (Heled and Drummond 2008).

However, the most abundant population genetic data are from the recombinant autosomal genome, especially in the form of single-nucleotide polymorphisms (SNPs) (International HapMap Consortium 2005; Li et al. 2008; Herraez et al. 2009). Many existing methods infer demographic parameters by examining the allele frequency spectra (AFS) (Adams and Hudson 2004; Marth et al. 2004; Keinan et al. 2007; Gutenkunst et al. 2009). Empirical AFS data are first corrected for ascertainment bias and then fit to the best demographic models using maximum likelihood computation. As the likelihoods of AFS can be numerically derived or approximated by simple simulations, these methods are computationally efficient. Nonetheless, the AFS-based methods are sensitive to different sources of ascertainment bias and are usually applied under highly simplified demographic models. Other methods make use of a collection of summary statistics that reflect different aspects of the genetic diversity data and evaluate how well they fit different demographic scenarios (Schaffner et al. 2005; Voight et al. 2005; Fagundes et al. 2007; Wall et al. 2009). Such methods usually involve simulating large amounts of genome-scale SNP data and, therefore, are highly computational intensive. Furthermore, the obscure mutual dependency and the heterogeneous sensitivities of the statistics toward simulation assumptions make it difficult to evaluate the inference accuracy. Improvements in simulation efficiency and novel statistics systematically designed for demographic inference are much needed. Recently, Li et al. developed the pairwise sequentially Markovian coalescent (PSMC) model to infer ancient population size changes from single re-sequenced diploid genomes (Li and Durbin 2011). This method made use of both mutation and recombination information and revealed many details of population size changes without making strong demographic assumptions.

Methods have also been developed based on haplotype or linkage disequilibrium (LD) patterns. Statistics based on LD or haplotype patterns should in theory be less affected by ascertainment bias (Conrad et al. 2006). This is because LD is a property of a genomic region, whereas ascertainment bias influences individual SNPs. Simple effective population sizes have been directly estimated from LD (Sved 1971; Hill 1981). Several studies have investigated LD-based statistics for inferring population size changes (Reich et al. 2001; Hayes et al. 2003; Tenesa et al. 2007). Lohmueller et al. (2009) proposed a method based on examining the window-wise haplotype distribution throughout the genome-wide data. Nevertheless, current LD-based inferences often suffer from limited resolution as either N_e is estimated as an average over long periods of time or the models studied are too simplistic. Slatkin and Bertorelle (2001) reported that the measurements of intra-allelic variability can be used to test neutrality and to infer population growth. Intra-allelic LD may be also well suited for inferring more complex demography, and in this study, we propose two intra-allelic LD-based statistics for population size inference. We show that these statistics are highly informative about ancient population size trajectory and can be used in the framework of approximate Bayesian computation (ABC) (Beaumont et al. 2002) to accurately estimate demographic parameters related to population size change from simulated data, even for relatively complex models. Finally, we applied the ABC-based method to genome-wide SNP data for the Yoruba and French populations from the CEPH-HGDP panel (Li et al. 2008).

Materials and Methods

Overview

In this study, we propose two statistics to infer ancient population size changes under neutrality. It is known that the intra-allelic variability and the allele frequency are two different measurements of allele ages, with the former revealing age at the absolute time scale, for example, in generations (McPeek and Strahs 1999), and the latter at the rescaled coalescent time scale (Slatkin and Rannala 2000). Slatkin and Bertorelle (2001) proposed that the contrast of these two measurements can be used to test neutrality or to make inferences about population growth. Nordborg and Tavare (2002) suggested that the intra-allelic LD can be informative about different aspects of demography, such as ancient population size and population structure. We propose that the intra-allelic LD measurement, when conditioned on allele frequency, may indeed be very informative about complex demographic trajectories. This is because when we compare the allele age in absolute time scale with the age in coalescent scale, their ratio actually measures the N_e in each time interval (supplementary material part I for detailed discussion, Supplementary Material online).

The two statistics we propose here are both related to the haplotype sharing for a given derived mutation. Studies have investigated the extension of the ancestral (identical) haplotypes from a derived mutation, and its use in disease/quantitative trait locus (QTL) mapping and neutrality tests (McPeek and Strahs 1999; Slatkin 2001; Innan and Nordborg 2003; Slatkin 2008). Our statistics are similarly constructed. The first statistic is the extended length of identity by descent (IBD) conditioned on derived allele frequency (AF-IBD). Here we take the literal meaning of IBD, which is the identity of sequences that descend from a single ancestral sequence, without any change in status from either mutation or recombination. IBD quantities usually have to be indirectly estimated, as tracts of IBD cannot be directly observed (Sved 1971; McPeek and Strahs 1999). However, to directly study how AF-IBD varies under different demographic scenarios, we start by assuming that IBD can be directly observed, and we later relax this assumption.

We assume that the genome is continuous, and all recombination and mutation events can be detected and exactly positioned. For any variant s of derived allele frequency j in a sample of n haplotypes ($2 \le j \le n-1$), we denote as $l_{n,j,s}$, the length of the identical haplotype extending from s to either side until the first detectable event (mutation or recombination) occurs (fig. 1*A* and *B*). The AF-IBD for allele frequency j is then defined as the expectation over all variants of frequency j: $\text{AF-IBD}_{n,j} = E(l_{n,j,s})$. To study empirical sequence or SNP data, we propose the statistic AF-IBS, similarly defined as AF-IBD: for a sample of n sequences, for each site s with derived allele frequency j ($2 \le j \le n-1$), we calculate in either direction the distance up to which the carrier chromosomes are identical by state (IBS), that is, up to one site before the first breakpoint, here denoted as $x_{n,j,s}$ (fig. 1*C*). The maximum distance $x_{n,j,s}$ is limited to 500 kb in the simulations; any distance larger than 500 kb in either empirical or simulated data is taken as 500 kb. The AF-IBS of allele frequency j is then taken as the average of $x_{n,j,s}$ over all sites of allele frequency j: $\text{AF-IBS}_{n,j} = \text{Mean}(x_{n,j,s})$.

We first study the properties of AF-IBD under different demographic scenarios, and then we examine the performance of AF-IBD in demographic parameter estimation using an ABC approach on simulated data. We then examine the relationship between AF-IBS and AF-IBD and establish an efficient ABC approach for relating AF-IBS to AF-IBD. Finally, the AF-IBS-based ABC method is evaluated in simulations and applied to the estimation of demographic histories from empirical SNP data for human populations (Li et al. 2008).

Examining AF-IBD under Various Demographic Scenarios

We first examined how AF-IBD behaves under different scenarios of population size changes, by analyzing the mathematics and generating simulations.

For a mutation s of allele frequency j in the n sampled sequences, when the coalescent tree is given, it occurs on the root edge (shown in green in fig. 1*A*) of a subtree J with j lineages (shown in red in fig. 1*A*). The recombination and mutation events (hereafter referred to as "events") can then be superimposed onto the tree with rates ρ and μ base^{-1} generation^{-1}, respectively. As in equation (4) of Slatkin and Bertorelle (2001), $l_{n,j,s}$ follows an exponential distribution with

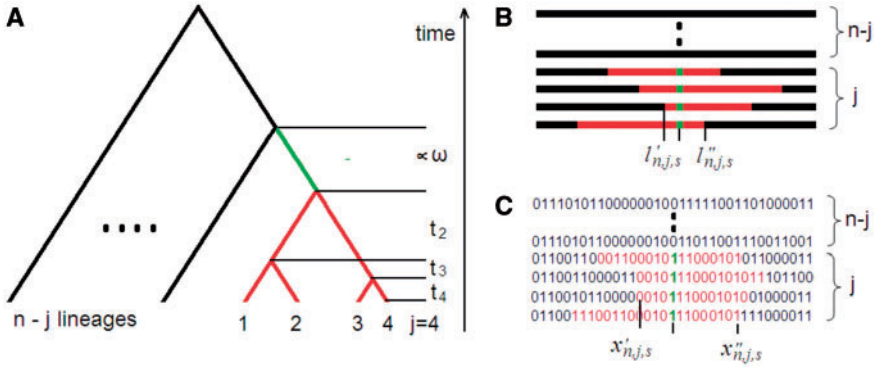

Fig. 1. Illustration of the model, and the statistics AF-IBD and AF-IBS. (*A*) The coalescent of n individuals, where j (here $j = 4$) lineages form a subtree J, colored red, before joining the other lineages by a root edge colored in green. The total length of subtree J, $T_{n,j}$ is the sum of the red edges measured in generations. (*B*) The extension of the ancestral haplotypes in red is shown for multiple sequences from a core mutation of frequency of $j = 4$ (shown in green). Mutation and recombination are taken as equivalent events that terminate the extension of the original ancestral haplotype. The ancestral shared haplotype is, therefore, the overlapping red segment that ends at the first event among all the sequences. The length of this segment is taken as a measurement of $l_{n,j,s}$, which when averaged over all sites of frequency j defines AF-IBD for j. (*C*) As the counterpart of $l_{n,j,s}$ in empirical sequence/polymorphism data, $x_{n,j,s}$ is taken as the length of the shared haplotype extending from the core mutation up to the first observed site that varies among the j haplotypes $x_{n,j,s}$ averaged over all sites of frequency j gives the estimation of AF-IBS for sequence/polymorphism data.

the rate parameter as the event rate integrated over time and lineages in the subtree J:

$$l_{n,j,s} \sim \mathrm{Exp}\left(T_{J,s}(\mu+\rho)\right) \quad (1)$$

where $T_{J,s}$ is the total length of the subtree J in generations, defined by the mutation s.

AF-IBD$_{n,j}$, which is the expectation of $l_{n,j,s}$ over all sites of frequency j out of n, can be integrated over all sites of j out of n as:

$$\mathrm{AF-IBD}_n, j = E[l_n, j, s] = \int_{L=0}^{\infty}\int_{\tau=0}^{\infty} P\left(T_J, s=\tau\right) P\left(l_n, j, s = L | T_J, s = \tau\right) d\tau dL \quad (2)$$

where $E\left(T_{J,s}^{-1}\right)$ is the expectation of the inverted total length of the subtree across all mutations of frequency j. Denote the absolute time as τ and the variable population size as a function of τ, $N(\tau)$. The distribution of AF-IBD can be derived by simulating a large number of coalescent trees as proposed previously (Slatkin and Bertorelle 2001). Details of calculating the distribution of AF-IBD can be found in the supplementary material part II, Supplementary Material online. We refer to this procedure as the tree sampling method, and we used it to study different models of population size changes.

To understand how AF-IBD responds to population size changes, we simulated models of various demographic scenarios including constant size, bottleneck, exponential growth, and complex models. A total of 1,000 coalescent trees were generated for each model. The sample size was set to be 100, so AF-IBD for $j = 2$–99 were calculated. The mutation and recombination rate were both set to the arbitrary value of 2.5×10^{-8} gen^{-1} site^{-1}. All coalescent trees in this study were simulated with the software ms (Hudson 2002). The constant size models assumed different population sizes of 1,000, 5,000, and 10,000. The scenario of expansion was examined by assuming that the population size grew exponentially from an ancestral population size of 500, 1,000, and 10,000 to a present population size of 10,000, 50,000, and 100,000, starting at a time point between 40 and 2,400 generations ago. A series of models of single bottlenecks were simulated with the event occurring sometime between 200 to 3,200 generations ago, with the reduction factor being 0.3, 0.1, or 0.01, and the duration ranging between 10 and 100 generations. Finally, a series of complex models were also simulated with an expansion event following a bottleneck event, or two or three consecutive bottlenecks. Combinations of events of various times of onset, durations, and magnitudes were examined. To quantify the effects of population size changes on AF-IBD, the AF-IBD vectors from various models were compared with that of a standard constant size model with N_e of 10,000.

Parameter Estimation with AF-IBD Using ABC

To further analyze the properties and information content of AF-IBD, we applied ABC. The underlying idea of ABC is that observed and simulated data sets are summarized into several representative values, which are then compared to find the simulations which best match the observed data. We implemented the ABC approach as described previously (Excoffier et al. 2005). The aim was to investigate whether underlying demographic parameters can be estimated if only AF-IBD is used to summarize a data set. Here, we assumed AF-IBD can be calculated from the observed (simulated) data; later we develop a procedure to relate AF-IBD to the statistic AF-IBS, which is directly calculated from the observed data. All data were generated by simulating coalescent trees as described in the previous section. The sample size of 100 was assumed but only AF-IBD for $j = 2, 3 \ldots 41$ were considered as summary statistics for the ABC calculation. Pseudo-observed (i.e., simulated data sets for which we knew the true values of the parameters) were generated for 300 parameter sets from each of three different demographic models. One million ABC simulations, with parameters drawn from the uniform parameter prior distributions, were then compared with the

pseudo-observed data to calculate the posterior parameter distributions. See supplementary note part IV, Supplementary Material online, for details concerning the ABC settings.

The first model assumes a constant size with a single parameter, the effective population size N_e; the second model was a 2-parameter sudden-growth model, in which the ancestral population size is fixed to 10,000 and starts growing exponentially at time T_1 ago until reaching a present day population size of β*10,000; and the third was a 3-parameter single-bottleneck model of a fixed ancestral population size of 10,000, whose population size declines by a factor β at time T_1 and then recovers to 10,000 at time T_2.

The accuracy and performance of this AF-IBD–ABC approach were evaluated by the relative root mean square error (RMSE, which is the square root of the mean square error divided by the true value), the mean absolute error (MAE, a weighted average of the absolute errors, with the relative frequencies as the weight factors), and the 95% and 50% coverage (proportion of times in which the true parameter value is inside the equal tailed 95% or 50% credible interval [CI]). These measurements were calculated by taking the mode of the posterior distribution as a point estimate. In table 1, for both the sudden-growth and the bottleneck model, ancient N_e was fixed to 10,000. In the bottleneck model, the population recovered 100% of its original size after the bottleneck event. Each estimation was based on the comparison between one pseudo-observed AF-IBD and 1 million simulated AF-IBD statistics. The ranges for the uniform prior distributions for each parameter are given as well.

Use of AF-IBS for Sequence or Polymorphism Data

When considering realistic polymorphism data, it is not easy to estimate AF-IBD, as the status of IBD is not directly observable. Although there exist various methods to estimate IBD-based statistics from sequence or SNP data (McPeek and Strahs 1999; Browning and Browning 2011), these methods are too computationally intensive to apply to genome-wide data. We, therefore, use AF-IBS to replace AF-IBD. Other than IBD, IBS can also result from recombination among homologous haplotypes or back mutation, or simply lack of polymorphic sites (Innan and Nordborg 2003). When the SNP density is high, in theory the length of IBS should be mainly accounted for by IBD. Therefore, we test whether AF-IBS has similar sensitivity as AF-IBD toward ancient population size changes.

We simulated sequence data from the same models as for AF-IBD (see model cartoons in supplementary fig. S2, Supplementary Material online). For the sequence simulation, sets of 100 haplotypes of length 2 Mb were simulated for 1,000 replicates for each set of demographic parameters. Simple ascertainment schemes were applied in which only sites variable within a parallel discovery panel of 5, 7, 10, or 15 haplotypes were kept to compose the polymorphism data and used to calculate AF-IBS. Throughout these simulations, the mutation rate and recombination rate for sequence data were assumed to be 2.5×10^{-8} and 1.3×10^{-8} gen^{-1} site^{-1}, respectively, which are the reported genome averages (Human Genome Sequencing Consortium 2001; Yu et al. 2001). Results for different demographic models were then contrasted to a constant size model of $N_e = 10{,}000$ to examine whether AF-IBS shows similar demographic sensitivity.

We then examined how different AF-IBD and AF-IBS are for the same demographic history. We introduce the ratio between AF-IBS and AF-IBD (hereafter referred to as SD ratio) for the same demographic model. The SD ratio is defined as a vector of index j where SD ratio$_j$ = AF-IBS$_{n,j}$/AF-IBD$_{n,j}$ for each frequency j.

Parameter Estimation with AF-IBS Using ABC

We established an ABC method using AF-IBS to estimate demographic parameters and evaluated its performance in the simulated scenarios.

The accuracy and performance of this AF-IBS–ABC approach were similarly evaluated as already described for the AF-IBD–ABC approach (table 2). We simulated 300 random data sets for each of three different models, which have one, three, and five parameters, respectively. The 1-parameter model is similar to the previous constant size model. The 3-parameter model is a sudden-growth model in which an ancestral population size increases instantly by a factor of β at time T_1. The 5-parameter model assumes a population size reduction from an ancestral size at time T_2 by a factor of β_2 and a population expansion at time T_1 by a factor of β_1 to a current size of N_e. For all models, we sampled the pseudoempirical parameters from a uniform prior on each parameter space. For each simulation, 250 10-Mb segments, each composed of 42 haplotypes, were generated with maCS (Chen et al. 2009). For all analyses of AF-IBS–ABC, we used the software recosim (Schaffner et al. 2005) to simulate a random map of variable recombination rates across 10-Mb regions. We used the same recombination parameters as in the "best fit" model of Schaffner et al. (2005), and the basal recombination rate is set according to the autosomal deCODE distribution (Kong et al. 2002). We generated 250 of such 10-Mb maps covering the whole genome, and each simulation takes one of them.

A simple ascertainment scheme was applied to match the SNP densities of all allele frequencies to that of the empirical data, as similarly applied before (Schaffner et al. 2005). Briefly, the empirical allele frequency spectrum was determined for both Yoruba and French, and then the simulated SNPs of a certain derived allele frequency (DAF) were repeatedly removed until the SNP densities in simulations equaled that of the empirical data. AF-IBS was then calculated for the simulated SNP data.

Theoretically, the ABC method based on AF-IBS can be done by randomly generating large amounts of SNP data and calculating their AF-IBS as described earlier. However, this is computationally nonfeasible as the SNP data simulation at genome scale is very time consuming, and the required number of samplings in ABC is usually very large, for example, 10^6. Here we developed a new ABC approach to solve this problem. We note that the simulation of AF-IBD is very efficient as only coalescent trees are sampled. If we calculate AF-IBS from AF-IBD, then the AF-IBS values can be efficiently

Table 1. Measures of Accuracy for AF-IBD–ABC Parameter Estimation.

Model	Parameters	Uniform Prior	RMSE	MAE	95% Coverage	50% Coverage
Constant	N_e	1,000–10,000	0.0498	0.0425	0.97	0.72
Sudden growth	T_1	200–800	0.192	0.1392	0.94	0.68
	β	2–10	0.0851	0.0624	0.97	0.63
Bottleneck	T_1	200–800	0.6761	0.4457	0.93	0.61
	T_2	200–800	0.5412	0.4414	0.95	0.55
	β	0.01–0.3	0.4311	0.3259	0.94	0.54

Table 2. Measures of Accuracy for AF-IBS–ABC Parameter Estimation.

Model	Parameters	Uniform Prior	RMSE	MAE	95% Coverage	50% Coverage
1 parameter	N_e	1,000–20,000	0.080	0.0714	0.96	0.69
3 parameter	T_1	100–2,000	0.195	0.131	0.94	0.57
	β	0.01–0.9	0.179	0.171	0.94	0.52
	N_e	5,000–40,000	0.153	0.131	0.93	0.61
5 parameter	N_e	15,000–50,000	0.257	0.213	0.93	0.59
	T_1	50–2,000	0.391	0.314	0.92	0.49
	β_1	0.01–0.5	0.516	0.467	0.90	0.52
	T_2	10–510	0.314	0.201	0.91	0.48
	β_2	0.1–0.4	0.402	0.357	0.89	0.46

Table 3. Power of Our Approach to Recover the True Model.

	1-Parameter Model	3-Parameter Model	5-Parameter Model
1-parameter model	0.84	0.13	0.03
3-parameter model	0.09	0.78	0.13
5-parameter model	0.06	0.20	0.74

obtained by tree simulations. Noting that AF-IBS and AF-IBD are closely related, and their SD ratios are relatively robust against changes in demographic parameters (supported by our analysis, shown in the Results section), we constructed a SD ratio grid on which AF-IBD can be efficiently converted to the corresponding AF-IBS. The SD ratio grid approach is implemented as follows: First the ratios of AF-IBS/AF-IBD were obtained for a predefined grid of parameter values, by simulating both coalescent trees and SNP data. SD ratios for any arbitrary parameter sets are then imputed based on this grid assuming local linearity along the parameter values (details about the construction of the SD ratio grid and the SD ratio imputation method can be found in the supplementary material part III, Supplementary Material online). Based on this, the ABC method using AF-IBS is briefly summarized as follows: first, 10^6 random parameters sets are sampled from the priors; second, AF-IBD is calculated for each parameter set; third, the AF-IBS/AF-IBD ratio is imputed from the SD ratio grid, and AF-IBS is calculated from AF-IBD; and fourth, the simulated AF-IBS is compared with the empirical AF-IBS to give the best-fitting model.

Model Misspecification

The real populations may have hidden population structures that are not represented by our simple models. It is, therefore, important to evaluate whether such hidden population structure will influence the AF-IBS calculation. We analyzed the AF-IBS behavior under certain model misspecifications. To see the effects of potential hidden population structure, we simulated an ancestral population of size N_e = 10,000 that split into two populations, 200, 500, and 1,000 generations ago (constant size demography). We analyzed the effect on AF-IBS of the two daughter populations having sizes 50/50 or 30/70 percent of the ancestral population, respectively (50 samples each). After that, we additionally simulated gene flow (0.1% and 0.5% per generation) between the two populations.

Empirical data are usually obtained as unphased genotype data, which is subject to an additional statistical calculation of phase reconstruction to infer the haplotype composition. As AF-IBS essentially measures how long a homologous segment extends, it may be sensitive to switching errors during the phase reconstruction. We, therefore, evaluated the effect of errors in the phase reconstruction on our AF-IBS calculations. We applied the program fastPHASE (Scheet and Stephens 2006) to various SNP data sets, simulated under different demographic scenarios (1-, 3-, and 5-parameter models with different parameter sets). The parameter values for the demographic models were chosen to cover a broad range of possible scenarios, with ancestral N_e ranging between 5,000 and 30,000; recent N_e ranging between 5,000 and 40,000; and times of expansion or bottleneck events ranging between 50 and 2,000 generations ago. The parameters for fastPHASE were set to the same values used for the phasing of the empirical data. We then analyzed the ratio of AF-IBS before and after the phasing. The ratios indicate that the phasing errors do have an impact on the AF-IBS calculation, especially for the lower DAFs (supplementary fig. S3, Supplementary Material online). As the effects are similar for different demographic scenarios, we calculated the average ratios across all the simulations. The AF-IBS values calculated for the empirical data were then corrected by multiplying the inverses of these average ratios, for the lower DAFs 2–12. AF-IBS values for higher

DAFs do not seem to be affected by the phase errors and are, therefore, not corrected.

Parameter Estimation for Empirical Data

We applied the ABC method using AF-IBS to the empirical SNP data. The genome-wide SNP data from the CEPH-HGDP panel was used (Li et al. 2008). The data were phased with the fastPHASE program and then corrected for the effects of phasing error, as described before. Statistics were calculated for 42 randomly chosen chromosomes from each population. For the calculation of AF-IBS, we considered only sites at least 5 Mb away from the chromosome ends, which resulted in AF-IBS values for ~490,000 sites, covering a genomic length of ~2.2 billion bp. We tested the same three models as for the pseudoempirical SNP data described earlier. To decide which model performs the best, we performed a model selection using a Bayes factor analysis (Jeffreys 1935; Bertorelle et al. 2010). The same number of simulations was chosen for each model, so that they were a priori equally likely, and we computed the ratio of acceptance rates for each pairwise model comparison. The posterior probability of a given model is then approximated by the proportion of accepted simulations given this model. The approach we used is implemented in the R package "abc" (http://cran.r-project.org/web/packages/abc/index.html, last accessed July 23, 2012). We additionally performed a test based on a logistic regression method (Fagundes et al. 2007), where a multinomial logistic regression is fit with the model being the categorical dependent variable. The regression is local around the observed summary statistics vector (as in the parameter estimation). Finally, the model probability is assessed at the point corresponding to the observed vector of summary statistics. For this method, we used the "calmod" function written by Beaumont MA (available from the "popabc" package at http://code.google.com/p/popabc/, last accessed July 23, 2012). Model selection was based on 1 million simulations for each model.

On the basis of this model choice approach, we additionally analyzed the power of this procedure to accurately recover the true model using the AF-IBS–ABC approach, following previous methods (Fagundes et al. 2007). We used the 300 simulated ascertained and phase-corrected data sets from the prior distribution for each model considered (1-, 3-, and 5 parameters) and analyzed them using the same simulations and pipeline as for the empirical data. Each of the 300 data sets then refers to one of the three models with the highest posterior probability. We then counted how many times our approach was able to identify the true model.

Results

Properties of AF-IBD

When AF-IBD is plotted against the allele frequency, it can be seen that AF-IBD decreases monotonically with increasing allele frequency (fig. 2*A* and supplementary fig. S1A, Supplementary Material online). This is easily understood, as variants of higher allele frequencies are on average older, and their intra-allelic IBD, therefore, has decreased more over time.

When AF-IBD values are compared between constant size models of different population size, we found that the ratio is constant across different allele frequencies, and it is the inverse of the ratio of population size (fig. 2*B*). This is expected given that coalescent rescales with population size.

AF-IBD is essentially contrasting two different measurements of allele age. Each allele frequency defines a time range on the coalescent time scale, for example, in the unit of inverse of effective population size (supplementary equation 1, Supplementary Material online). For the same time range in coalescent scale, when the effective population size is big, then the absolute time span is long, resulting in shorter average IBD length; otherwise, the average IBD length becomes longer. This suggests that smaller AF-IBD indicates a bigger effective population size and vice versa; therefore, the AF-IBD curve along the allele frequency spectrum reflects the details of population changes.

The observations from simulations are consistent with the above statements. We contrasted AF-IBD values for different demographic models with that of a constant size model of N_e = 10,000 (see Materials and Methods). Figure 2C shows the comparisons among four bottleneck models. All ratio curves are elevated above 1, with a single peak at different allele frequencies and magnitudes. The most recent bottleneck has a peak around allele frequency 10 with the highest ratio approximately 2.1; the intermediate-aged bottleneck is shifted to the right to around frequency 15 with a peak height of 1.6; and even the relatively ancient bottleneck event, starting 1,000 generations ago, also resulted in elevated ratios around the frequency 20–30. It is obvious that AF-IBD has higher sensitivity to more recent events than older ones of the same magnitude. On the other hand, strong ancient events can also induce big changes in the relative AF-IBD curve. This can be clearly seen in the fourth model, where the duration of the size reduction was increased to three times that of the third model (fig. 2*C*).

For the scenarios of expansion, figure 2*D* shows that the ratios of AF-IBD started from 0.3–0.4 at the lower allele frequency range, much lower than the value of 1 expected under a constant population size. The ratio curve recovers quickly back to close to 1 for the recent expansion. The increase of the ratios along the allele frequency is progressively slower and to a lower maximum when the expansion starts earlier in time (fig. 2*D*). Finally, the AF-IBD ratio is also sensitive to complex models where multiple events shaped the population size trajectory. Figure 2*E* shows the AF-IBD ratios for two complex models, one defined by a recent weaker bottleneck (200–210 generations ago, 100 times size reduction) following an old strong bottleneck (1,000–1,100 generations ago, 100 times size reduction; colored in black), and the other defined by a recent expansion (population size from 10,000 to 100,000, starting at 500 generations ago) after an intermediate-aged bottleneck (1,000–1,100 generations ago, 100 times size reduction, colored in red). The two curves clearly differ from each other: for the case of two bottlenecks, the ratio starts above 1 and increases to a first turning point around frequency 10, then rises to the second turning point around frequency 40. For the case of expansion following

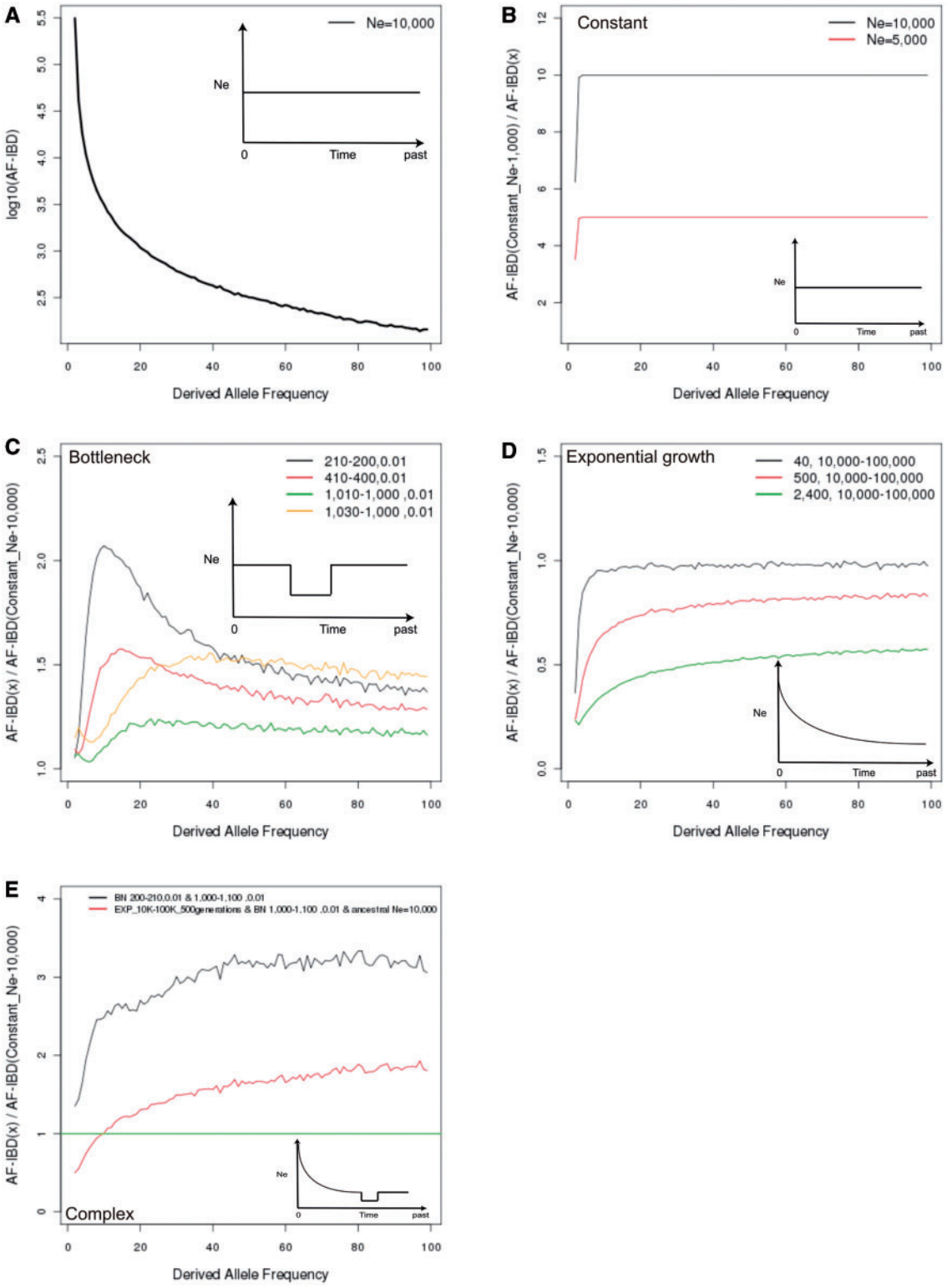

Fig. 2. AF-IBD calculated for several simulated demographic models. AF-IBD was calculated from data sets simulated as coalescent genealogies under various demographic models of interest (see subfigure cartoons) and for a constant size reference model. (*A*) AF-IBD curve calculated from a constant size population of $N_e = 10{,}000$. (*B*) AF-IBD ratios between two different models of constant population size ($N_e = 5{,}000$, $N_e = 10{,}000$) and one constant $N_e = 1{,}000$. (*C*) AF-IBD ratios between various bottleneck models and one constant population size model of $N_e = 10{,}000$. Parameters given in the legend represent the start and end of the bottleneck in generations before present, as well as the reduction factor during the bottleneck. (*D*) AF-IBD ratios between various exponential growth models and one constant population size model of $N_e = 10{,}000$. (*E*) AF-IBD ratios between a two-bottleneck model and one constant size model of $N_e = 10{,}000$, and between a complex bottleneck followed by sudden-growth model and a constant size model of $N_e = 10{,}000$. Parameters given in the legend represent the number of generations for each period of growth lasting to the present day as well as the ancient and present day population sizes. As explained in the main text, different demographic histories have distinct effects on the outcome of AF-IBD, which clearly shows the sensitivity of this statistic to population size changes.

bottleneck, the ratio starts from below 1 as expected for large population size and keeps ascending above 1 until reaching a maximum at the highest frequency. The increase in the AF-IBD ratio is clearly due to the bottleneck.

AF-IBD–ABC

We first tested an ABC framework assuming that AF-IBD can be directly observed. The purpose was to first analyze how accurate underlying demographic parameters, connected to population size changes, can be estimated in the absence of any complications introduced by the type of empirical data (e.g., ascertainment bias). In table 1, we show several calculated measures of precision, which represent the differences between preset parameter values and estimated parameter values. We calculated the RMSE; the MAE, and the 95% and 50% coverage (see Materials and Methods). Results from table 1 show that this method of inference is highly precise for the single parameter constant size model. This can be explained by the underlying mathematical features of AF-IBD. As shown in figure 2*B*, the reverse ratio of AF-IBD for different constant size models coincides with the population sizes. The estimation for the 2- and 3-parameter models, although slightly less accurate, still provides estimates that are sufficiently close to the true values. The reduced accuracy is expected, as the same AF-IBD curve might result from different but equivalent demographic histories. For example, the general effect of a strong but short bottleneck can be very similar to that of a weaker but longer bottleneck. However, in most cases, we could estimate the true underlying parameter values with a high level of accuracy (table 1), demonstrating the validity of the AF-IBD-based ABC approach.

Properties of AF-IBS

We show the comparisons between AF-IBS and AF-IBD for models of three different scenarios: constant size, expansion, and bottleneck. Specifically AF-IBD and AF-IBS of the bottleneck and expansion models were contrasted against those of the constant size model, and the ratios were plotted together (supplementary fig. S2A and S2B, Supplementary Material online). It can be seen that the ratio curve of AF-IBS is close to that of AF-IBD. In the bottleneck scenario, the AF-IBS ratios are shifted slightly below the AF-IBD ratios, but the position of the peak is well conserved. For the expansion scenario, AF-IBS curves are slightly above the AF-IBD curve although the general shape is unchanged. Comparisons for additional population size change models are shown in supplementary figure S1B, Supplementary Material online. Overall, AF-IBS curves for different ascertainment schemes are very similar to each other, which suggest that the AF-IBS ratio is generally robust to the ascertainment bias schemes implemented here.

The IBS/IBD Ratio

We showed in the previous section that the relative AF-IBD curve is very similar to the relative AF-IBS curve for the same demography, despite different ascertainment schemes. This suggests that AF-IBS is related to AF-IBD in a way that is not affected by the changes in population size. We checked the robustness of the SD ratio between AF-IBS and AF-IBD in various demographic scenarios including constant size, bottleneck, and expansion. Supplementary figure S2C, Supplementary Material online, shows the SD ratio curves for AF-IBS. In supplementary figure S2C, Supplementary Material online, the SD ratio starts at a low level and rises steeply above 1.0 for the first few frequency bins. This is an artifact due to the fact that the maximum length of AF-IBS is 0.5 Mb (see Materials and Methods), whereas AF-IBD estimation from tree simulation theoretically can be infinitely long. The subsequent values range between 1.5 and 3, and the curves for the two different models have a similar shape. In fact, we found that the SD ratio curve distributes within a rather defined interval, across a large parameter space, and the values for each bin in general are in a roughly linear relationship with the parameters (data not shown).

AF-IBS–ABC

We constructed a fast ABC pipeline that applies to the observed AF-IBS values. We first checked whether correct estimations can be obtained for simulated pseudo-observed SNP data. Three models—constant size, sudden growth, and expansion—after bottleneck were tested, which contain one, three, and five parameters, respectively. The entire workflow of ABC for AF-IBS is shown in supplementary figure S4, Supplementary Material online.

Figure 3 shows the estimated posterior distributions for some parameters of interest from the 1-, 3-, and 5-parameter models. As presented in table 2, inference based on AF-IBS–ABC is relatively accurate and precise for the 1- and 3-parameter models and still reliable for the most complex 5-parameter model. We also analyzed the power to correctly recover the true model based on the logistic regression procedure. As described earlier, we counted how many times we correctly assigned the true model in a set of 300 simulated data sets from the prior distributions of each model. As presented in table 3, data sets are properly assigned in most of the cases. However, the more complex the model, the less power our approach has. Also, the inferred empirical Bayes factors (see Results) are in good agreement with the ones we simulated.

Application to Genome-Wide Data from the CEPH-HGDP Panel

We then applied our approach to the genome-wide data set of the CEPH-HGDP panel (Li et al. 2008). Figure 4*A* shows AF-IBS for the first 34 DAF bins calculated from 42 randomly chosen chromosomes from each of 11 worldwide populations. As high DAF values reflect old mutations and low DAF values reflect more recent mutations, variation in AF-IBS values indicates population size changes at different times in the past. The AF-IBS values for higher DAF for African populations are clearly smaller than for all non-African populations, indicating much reduced ancient population sizes for non-Africans compared with Africans. Furthermore, populations show continental or areal clustering, which suggests similar demographic histories for populations within the

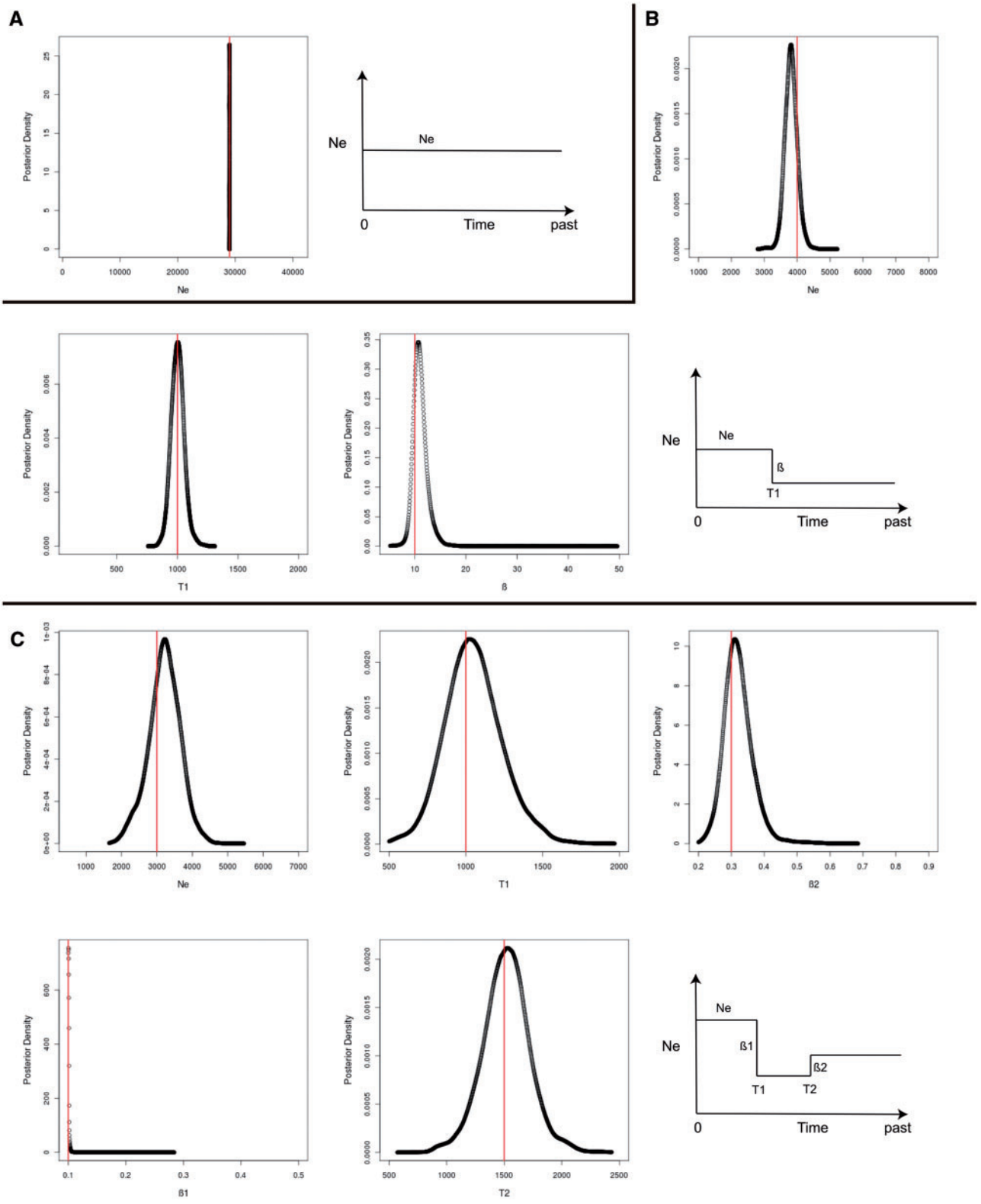

Fig. 3. Approximate Bayesian Computation estimation results for pseudo-observed AF-IBS. The posterior densities from ABC parameter estimation for 1-, 3-, and 5-parameter models are shown. Simulated polymorphism data were used as pseudo-observed data. Vertical red lines represent the true underlying parameter values. For each panel, a cartoon of the underlying model with all parameters that were estimated is shown. (*A*) Results for the single constant size model parameter N_e. (*B*) Results for three parameters of a demographic model of sudden growth. (*C*) Results for five parameters of a model of an ancient bottleneck followed by more recent sudden growth. Prior ranges for each uniformly distributed prior are equivalent to the *x*-axis ranges.

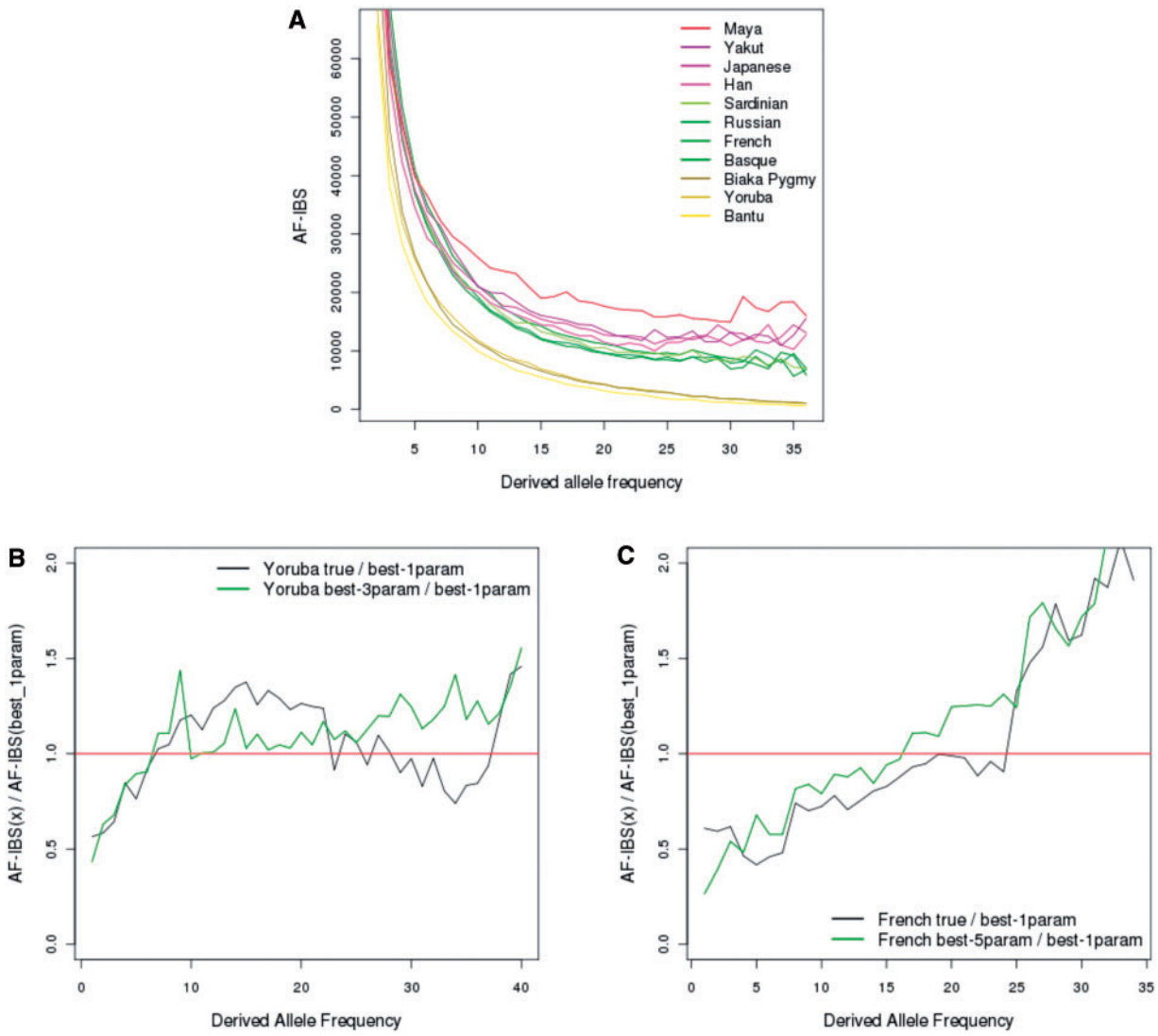

FIG. 4. AF-IBS calculated for several CEPH-HGDP populations. (*A*) AF-IBS calculated for various populations from the CEPH-HGDP panel. (*B*) Two ratios between the observed Yoruba AF-IBS and the AF-IBS of the best constant size model simulation and the ratio between AF-IBS from the best 3-parameter simulation and the best constant size model simulation. (*C*) Ratio between the observed French AF-IBS and the AF-IBS of the best constant size model simulation and the ratio between AF-IBS from the best 5-parameter simulation and the best constant size model simulation.

same cluster. All non-African populations show higher variability in the tails of the curves. This is due to the fact that fewer sites with high DAF are present in these populations, probably because of severe bottlenecks. We then analyzed two representative populations in more detail: Yoruba from Africa and French from Europe.

Model Misspecification

We calculated AF-IBS for a standard constant size model and the models assuming different population structure and migration. The ratios of AF-IBS between the standard model and the models with complete population structures were approximately 1, ranging from 0.96 to 1.14, indicating that hidden population structure does not significantly influence our results. Adding migration between the daughter populations further reduces the difference between the standard and alternative models (supplementary fig. S3, Supplementary Material online).

On the other hand, the phase reconstruction error seems to have an impact on the AF-IBS calculation. We contrasted the AF-IBS values of different scenarios both before and after phasing (supplementary fig. S3, Supplementary Material online). The ratios are rather consistent among different scenarios, starting at approximately 0.8 for the DAF = 1 and recovering back to 1 after DAF = 12. This suggests that we might underestimate AF-IBS for the lower DAFs, due to the phasing errors. We correct such bias by multiplying the empirical AF-IBS values with the phasing error correction ratios (see Materials and Methods).

ABC Analysis for Yoruba

Table 4 lists the results for the estimated demographic models for Yoruba. The logistic regression analysis was done before the actual ABC analysis. Among all model comparisons, the 3-parameter model of sudden expansion was the best fitting model (Bayes factor 4.1 and probability of 0.63), followed

Table 4. ABC Estimation Results for Empirical Data for Yoruba and French.

Population	Model	Parameters	Uniform Prior	Regression Estimate	95% CI
Yoruba	Constant size	N_e	1,000–41,000	8,850	7,825–13,617
	Sudden growth	N_e	5,000–40,000	22,915	21,706–24,110
		T_1	100–2,000	806	685–1,030
		β	0.01–0.9	0.57	0.52–0.63
	Bottleneck + sudden growth	N_e	15,000–50,000	28,310	27,081–32,506
		T_1	50–2,000	1,005	780–1,436
		β_1	0.01–0.5	0.28	0.12–0.35
		T_2	60–2,500	1,302	895–1,498
		β_2	0.11–0.9	0.81	0.73–0.85
French	Constant size	N_e	1,000–41,000	6,311	4,753–8,623
	Sudden growth	N_e	5,000–40,000	5,043*	*
		T_1	100–2,000	351*	*
		β	0.01–0.9	0.21*	*
	Bottleneck + sudden growth	N_e	15,000–50,000	18,300	16,116–22,082
		T_1	50–2,000	1,300	987–1,520
		β_1	0.01–0.5	0.18	0.14–0.25
		T_2	60–2,500	1,580	1,410–1,805
		β_2	0.11–0.9	0.55	0.32–0.68

*For this model, we were not able to reliably infer parameter values from the Fench data.

by the 5- and 1-parameter models. The most likely constant population size was estimated to be approximately 8,850. The inferred parameter ranges for the 3-parameter model suggest a constant recent population size of approximately 22,915 (95% CI: 21,706–24,110) followed by a population-size decrease (backward in time) to approximately 0.57 of the recent N_e (95% CI: 0.53–0.62) to an ancestral size of 13,061 at 806 (95% CI: 685–1,030) generations ago. The 5-parameter model had a probability of 0.25. The inferred parameter ranges suggest a recent population size of approximately 28,000 followed by a bottleneck between 1,005 and 1,302 generations ago, with an ancestral size of approximately 18,600 and a bottleneck size of approximately 8,000. Results from figure 4*B* show that the ratio between the observed AF-IBS and the best 1 parameter model simulations AF-IBS (black line) is approximately 1 for most DAF, which supports a relatively stable ancient population size, followed by a more recent expansion (ratio below 1 for the first bins). We, therefore, conclude that the 3-parameter model of a simple expansion seems to best explain our data.

ABC Analysis for French

To analyze the French data, only the first 2–36 AF-IBS values for 42 randomly chosen chromosomes were used, as there are not enough high-frequency DAF cases to get a reliable genome-wide average for their AF-IBS values. Among all model comparisons, the 5-parameter model of a bottleneck followed by sudden expansion was the best fitting (Bayes factor 3.9 and probability of 0.71), followed by the 3- and the 1-parameter models. Table 4 presents that the most likely constant population size was estimated to be approximately 6,300, which is smaller than for the Yoruba population. Again, note that this estimate cannot directly be compared with usual measurements of N_e. Trying to fit a 3-parameter model of sudden growth did not yield any reliable parameter estimates. As listed in table 2, Euclidian distances of >20 for the best fitting simulations indicated an insufficient fit of our observed data. We also analyzed the 5-parameter model with ABC (table 4). The results suggest a recent population size of approximately 18,300 (95% CI: 16,115–22,082). The ancestral population size was estimated to be 10,065 (95% CI: 5,856–12,444). The timing of the bottleneck was estimated to be between 1,580 and 1,300 (95% CI: 1,410–1,805; 987–1,520) generations ago with a population size of approximately 3,300 (95% CI: 2,562–4,575) during that time. Importantly, the CIs of the parameters seem to be rather narrow compared with the priors, except for the two time parameters. The ratio curve of the best 5-parameter simulation against the best constant size simulation also matches closely with that of the empirical data against the best constant size simulation (fig. 4*C*). We, therefore, conclude that the 5-parameter model of a bottleneck with an expansion is the best fitting model for the French.

Discussion

Using genetic data to make inferences concerning the demography of populations (especially population size changes) has long been of interest (Griffiths and Tavare 1994; Kuhner et al. 1995). As genome-wide SNP and full sequence data are becoming increasingly abundant for human populations and other species, it is of great interest to make efficient use of such data to infer ancestral demographic history with high accuracy. In this work, we introduce two potentially very useful statistics, AF-IBD and AF-IBS, which make use of haplotype configuration changes resulting from both mutation and recombination events. We showed that both have some desirable mathematical properties, which determine their high sensitivity to population size changes even for complex demographic histories over a wide time range.

The high sensitivity of AF-IBD and AF-IBS toward ancient population size changes results from contrasting two types of age estimators: the intra-allelic LD inferring the absolute age and the derived allele frequency surrogating the coalescent scale age. In this study, we use the ABC approach to estimate

the trajectory of population size, by minimizing the distance between the summary statistics calculated from simulated and observed data. On the other hand, if a closed form equation can be found that defines the AF-IBD/AF-IBS as a function, say $G(j,N(\tau))$ (assuming a one-to-one map between N and G), of allele frequency j and $N(\tau)$, it is possible to analytically derive $N(t)$ by solving the reverse function G^{-1}. In the perspective of the coalescent, the AF-IBD/IBS statistics are similar to AFS: they are all conditioned on the derived allele frequency. Although the AFS measures the length of the root edge of a j-node subtree (corresponding to the green edge in fig. 1) by counting the number of mutations, the AF-IBD/IBS measures the total subtree length (the red subtree in fig. 1). In principle, the subtrees should be more informative about the population size changes than their root edges. This is because the subtrees of the same DAF coalesce in the same time interval and are responsive to the same population size changes. The root edges on the other hand do not necessarily overlap in time for a given DAF and thus are less responsive to a particular population size change. Lohmueller et al. (2009) proposed the HCN statistics, which also make use of the haplotype distributions. By summarizing the local haplotype frequency distribution, the HCN essentially makes use of both recombination and mutation events to reflect the properties of the coalescent trees within windows of fixed recombination size. Our statistics are similar to HCN in the use of both mutation and recombination information, but AF-IBD/IBS focus explicitly on the tree defined by the central SNP. The recently proposed PSMC method directly estimates the time of most recent common ancestor (TMRCA) on a pair of genome sequences (Li and Durbin 2011). By evaluating the coalescent density over the stepwise time intervals, this method revealed many details of the population size trajectories. However, the pairwise comparison by design provides less information on very recent history and is sensitive to recent population structure. The AF-IBD/AF-IBS statistics are based on multiple haplotype comparisons and, therefore, may help complement the PSMC for recent history.

In this work, we associate AF-IBS to the AF-IBD statistic by a correction ratio. In fact, it might be possible to express AF-IBS as functions of AF-IBD in explicit mathematical form. The greater AF-IBS than AF-IBD values at higher frequencies are mainly due to the undetected recombination events (supplementary fig. S2, Supplementary Material online). We introduced the SD ratio as one potential way of transforming AF-IBD to AF-IBS. However, we note that this is an approximate way of solving this issue, and there is room for improvement. Nonetheless, the ABC estimation based on AF-IBS already shows promising accuracy on the pseudo-observed SNP data. Our ABC parameter estimation results show that even for quite complicated models such as the 5-parameter model, we can accurately estimate parameters of interest.

Our results show that the AF-IBS ratios are relatively robust against very different ascertainment schemes (supplementary fig. S2, Supplementary Material online). This suggests that possible misspecifications of the ascertainment scheme should not affect our inference very much. Some SNP data are censored for the lower minor allele frequencies. This will certainly cause losses of information for very recent or ancient demographic events. On the other hand, the switching errors during the phase reconstruction from the empirical genotype data do seem to cause a slight underestimation of AF-IBS for lower DAFs. This is not difficult to understand: phasing errors can be seen as a low level of artificial recombination. When this fraction of recombination rate, say ρ_{phase} is added to the term $T_{J,s}(\mu+\rho)$ in equation (1), it tends to reduce AF-IBD/AF-IBS when $T_{J,s}$ is small, which corresponds to lower DAFs. However, the effect of ρ_{phase} can be negligible when $T_{J,s}$ or DAF is big. This problem can be minimized by using phase certain SNP data, such as those genotyped on trio samples.

In the application of the AF-IBS statistic to the CEPH-HGDP Yoruba and French data, we found that neither of the two data sets can be fully explained by the constant size model. The three parameter model with a recent population expansion provides a slightly better fit to the Yoruba data than the more complex 5-parameter model. For the French, we found that the 5-parameter model featuring both a bottleneck and an expansion is needed to explain the observed data. This result is in general agreement with previous studies. Most of the existing studies showed that a simple expansion is sufficient to account for the African demography (Adams and Hudson 2004; Marth et al. 2004; Voight et al. 2005; Keinan et al. 2007), whereas Schaffner et al. suggested a minor bottleneck for the Yoruba ($F = 0.008$, [Schaffner et al. 2005]), and Li et al. showed a mild reduction between 20,000 and 100,000 years ago (Li and Durbin 2011). Moreover, all studies infer that European populations had at least one bottleneck before the recent expansion (Adams and Hudson 2004; Marth et al. 2004; Schaffner et al. 2005; Voight et al. 2005; Keinan et al. 2007; Lohmueller et al. 2009; Wall et al. 2009). For the specific parameter estimation, we summarize the comparisons among different studies in tables 5 and 6. Our result shows that the Yoruba had an ancient population size (N_{anc}) of ~13,000 recovering to a present size (N_{cur}) of ~22,900. This is in good agreement with previous studies (N_{anc} 9,069–12,500; N_{cur} 16,233–31,000, tables 5 and 6). The time of expansion T_{exp} varies considerably among different studies. Although our estimate of 806 generations (~20 thousand years ago [kya]) is close to previous estimates of 27 kya (Adams and Hudson 2004) and 25 kya (Voight et al. 2005), other studies gave much older estimates (186–425 kya, tables 5 and 6). Results from Li et al. revealed two waves of expansions (or bottlenecks depending on the perspectives), one earlier (200–600 kya) and one later (~20 kya) (Li and Durbin 2011). This suggests that different methods may have captured either of the two inferred periods of growth. The more recent expansion given by our result coincides with that of Li and Durbin (2011) and the last glacial maximum.

For the European demography, our estimates of the ancient population size (N_{anc} ~10,000) and current population size (N_{cur} ~18,300) are also similar to those from previous studies of N_{anc} 8,000–10,065 and N_{cur} 10,000–20,000 (tables 5 and 6).The time when the bottleneck starts (T_{bot}) and the time of recovery (T_{exp}) are surprisingly consistent among most studies, although these two values are usually

Table 5. Estimated African Demographic Parameters Compared among Different Studies.

Studies	N_{anc}	N_{cur}	T_{exp} (gen)	T_{exp} (kya)
Adams and Hudson (2004)	10,000	19,000/31,000	1,080	27
Marth et al. (2004)	10,000	18,000	7,500	187.5
Voight et al. (2005)	10,625	21,304	1,000	25
Keinan et al. (2007)	9,069	16,234	7,440	186
Schaffner et al. (2005)	12,500	24,000	17,000	425
Fagundes et al. (2007)	12,772	206,920	—	—
This study	13,601	22,915	806	20.15

Table 6. Estimated European Demographic Parameters Compared among Different Studies.

Studies	N_{anc}	N_{bot}	N_{cur}	T_{bot} (gen)	T_{bot} (kya)	T_{exp} (gen)	T_{exp} (kya)	F
Marth et al. (2004)	10,000	2,000	20,000	3,500	87.5	3,000	75	0.125
Adams and Hudson (2004)	10,000	1,500	20,000	1,500	37.5	—	—	—
Wall et al. (2009)	—	625	—	1,240	31	1,200	30	0.032
Voight et al. (2005)	10,695	1,065.9	—	2,000	50	1,600	40	0.19
Keinan et al. (2007)	8,712	—	—	1,280	32	—	—	0.151
Schaffner et al. (2005)								0.085, 0.02
Lohmueller et al. (2009)	8,000	550	10,000	1,500	37.5	1,100	27.5	0.36
This study	10,065	3,300	18,300	1,580	39.5	1,300	32.5	0.042

considered difficult to estimate. Other than one study (Marth et al. 2004) with older time estimates (T_{bot} ~87.5 kya, T_{exp} ~75 kya), the other studies estimated the T_{bot} to be approximately 31–50 kya and T_{exp} approximately 27.5–40 kya (tables 5 and 6). Our estimations of 39.5 kya and 32.5 kya fall into these two ranges. This bottleneck probably corresponds to the Out of Africa dispersion. Estimates of the population size of the bottleneck (N_{bot}) vary considerably among studies; our estimate of ~3,300 is larger than many such estimates (tables 5 and 6). When the inbreeding coefficient F is calculated (Keinan et al. 2007), our estimate (0.042) is close to previous estimates of 0.085, 0.02 (Schaffner et al. 2005), and 0.032 (Wall et al. 2009), although much smaller than other studies of 0.125–0.364 (tables 5 and 6). Li et al. showed a much reduced population size of approximately 1,200 between 40 and 20 kya. These suggest that our method may have underestimated the intensity of the bottleneck. The precise reason is not clear, but the 95% lower bound of our N_{bot} is approximately 2,500, suggesting a lower bottleneck size is also possible.

We emphasize that this is a preliminary study to demonstrate the usefulness of the AF-IBD-related statistics. There are various ways in which we expect the inference can be improved. For example, we use the mean AF-IBD/IBS as our inference statistics in this study. In fact, we notice that the distribution of each AF-IBD/IBS for a given DAF is also sensitive to population size changes (data not shown). This is easy to understand: subtrees of the same DAF span different lengths of the coalescent time scale, therefore may be perturbed by the fluctuating demography at different times or intensities. The power of the inference methods may be further improved by using the full distributions of AF-IBD/IBS.

Moreover, our current computational approach still offers room for improvement. Although coalescent simulators are capable of simulating a wide range of demographic scenarios within a rather short time, simulating full genomes with an underlying variable recombination map is still computationally quite intensive, especially when every full sequence simulation needs to be ascertained and corrected for phase reconstruction error. Although the simulations we carried out provide support for the overall effectiveness of our approach, further improvements (under development) should improve the accuracy of the parameter estimates, especially for more complex (and hence realistic) models.

In conclusion, we have shown that quite accurate estimates of demographic parameters can be obtained from ascertained genome-wide SNP data, even for complex underlying population histories. Improved inference may also be achieved by applying more elaborate methods of parameter estimation, especially when adding more parameters to underlying demographic models. For example, combining the advantages of ABC and MCMC can lead to improved estimation results (Wegmann et al. 2009). Moreover, with full sequence data sets becoming available, the limitations of SNP data will no longer apply. With further work, it might be possible to find the closed forms of AF-IBS and AF-IBD as functions of population size change $N(\tau)$, and nonparametric methods could potentially be used to infer more realistic demographic trajectories through time.

Supplementary Material

Supplementary material and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgment

Research was done at Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany. This study was supported by the Max Planck Society.

References

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

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Bertorelle G, Benazzo A, Mona S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol.* 19:2609–2625.

Browning BL, Browning SR. 2011. A fast, powerful method for detecting identity by descent. *Am J Hum Genet.* 88:173–182.

Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res.* 19:136–142.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet.* 38: 1251–1260.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161: 1307–1320.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 22:1185–1192.

Excoffier L, Estoup A, Cornuet JM. 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* 169:1727–1738.

Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A.* 104: 17614–17619.

Griffiths RC, Tavare S. 1994. Ancestral inference in population genetics. *Stat Sci.* 9:307–319.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13:635–643.

Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. *BMC Evol Biol.* 8:289.

Herraez DL, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M. 2009. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One* 4(11):e7888; doi:10.1371/journal.pone.0007888.

Hill WG. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genet Res.* 38:209–216.

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

Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Innan H, Nordborg M. 2003. The extent of linkage disequilibrium and haplotype sharing around a polymorphic site. *Genetics* 165: 437–444.

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

Jeffreys H. 1935. Some tests of significance, treated by the theory of probability. *Proc Camb Phil Soc.* 29:83–87.

Keinan A, Mullikin JC, Patterson N, Reich D. 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, Gudbjartsson DF, Sainz J, et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. *Nat Genet.* 31:241–247.

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

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493–496.

Lohmueller KE, Bustamante CD, Clark AG. 2009. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* 182:217–231.

Marth GT, Czabarka E, Murvai J, Sherry ST. 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.

McPeek MS, Strahs A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet.* 65:858–875.

Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol.* 25:1459–1471.

Nordborg M, Tavare S. 2002. Linkage disequilibrium: what history has to tell us. *Trends Genet.* 18:83–90.

Pybus OG, Rambaut A, Harvey PH. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155:1429–1437.

Reich DE, Cargill M, Bolk S, et al. (11 co-authors). 2001. Linkage disequilibrium in the human genome. *Nature* 411:199–204.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.

Scheet P, Stephens M. 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.

Slatkin M. 2001. Simulating genealogies of selected alleles in a population of variable size. *Genet Res.* 78:49–57.

Slatkin M. 2008. A Bayesian method for jointly estimating allele age and selection intensity. *Genet Res.* 90:129–137.

Slatkin M, Bertorelle G. 2001. The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* 158: 865–874.

Slatkin M, Rannala B. 2000. Estimating allele age. *Annu Rev Genomics Hum Genet.* 1:225–249.

Strimmer K, Pybus OG. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol Biol Evol.* 18: 2298–2305.

Sved JA. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol.* 2:125–141.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17:520–526.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A.* 102:18508–18513.

Wall JD, Lohmueller KE, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol.* 26:1823–1827.

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

Yu A, Zhao C, Fan Y, et al. (11 co-authors). 2001. Comparison of human genetic and sequence-based physical maps. *Nature* 409: 951–953.