Published in final edited form as:

Nat Genet. 2015 May ; 47(5): 555–559. doi:10.1038/ng.3254.

Exploring Population Size Changes Using SNP Frequency Spectra

Xiaoming Liu[1] and **Yun-Xin Fu**[2]

[1]Department of Epidemiology, Human Genetics & Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA.

[2]Department of Biostatistics, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA.

Abstract

Inferring demographic history is an important task in population genetics. Many existing inference methods are based on pre-defined simplified population models, which are more suitable for hypothesis testing than for exploratory analysis. We developed a novel model-flexible method called stairway plot, which infers population size changes over time using SNP frequency spectra. This method is applicable for whole-genome sequences of hundreds of individuals. Using extensive simulation we demonstrated the usefulness of the method for inferring demographic history, especially recent population size changes. The method was applied to the whole genome sequence data of nine populations from the 1000 Genomes Project, and showed a pattern of human population fluctuations from 10 to 200 thousand years ago.

Inferring human demographic history using genetic information can shed light on important prehistoric evolutionary events such as population bottleneck, expansion, migration, and admixture, among others. It is also the foundation of many population genetics analyses, as demographic history is one of the most important forces shaping the polymorphic pattern of our genome[1]. Many of the methods available for inferring demographic history with genome-scale data are model-constrained[2–5], that is, researchers need to pre-define a demographic model (for example, a constant-size phase followed by an exponential growth phase beginning at a certain time point) and the number of the parameters to be estimated before estimating the demographic history. Parameters of the models are then estimated by fitting the expected polymorphic pattern (e.g. a SNP frequency spectrum) given a set of parameters to that of the observed data, either through extensive simulation[2] or diffusion approximation[3]. On the other hand, model-flexible methods (sometime also called "model-

Correspondence should be addressed to X.L. (Xiaoming.Liu@uth.tmc.edu)..

AUTHOR CONTRIBUTIONS
X.L. designed the study, developed the method, conducted the analyses and wrote the paper. Y.-X.F. provided critical advice on methodology development and result interpretation.

URLs: A java program package implemented the stairway plot method is freely available for download at https://sites.google.com/site/jpopgen/stairway-plot.

Competing financial interests: The authors declare no competing financial interests.

free" methods), such as the skyline plot[6] and its derivatives[7–13], are not restricted to a specific demographic model and typically explore larger model space than model-constrained methods. Therefore, model-flexible methods can infer significantly more detailed demographic history and may be more suitable for exploratory or hypothesis generating analysis. However, the skyline plot and its derivatives are based on the full-likelihood of DNA sequences, and at the current stage can only be applied to recombination-free loci such as mitochondrial DNA[14,15]. Recently, Li and Durbin[16] proposed a model-flexible method based on the pairwise sequentially Markovian coalescent (PSMC) framework, which specifically models the recombination between two sequences and therefore can analyze autosomes. However, the PSMC method also has its limitations: i) it still requires the users to have a rough idea of the population history in order to determine the number of parameters to estimate; ii) it requires high-quality sequence data for its application; and iii) it tends to produce biased estimation for recent population histories[17].

We developed a new method called stairway plot. It uses a flexible multi-epoch model (**Fig. 1**) as used in the skyline plot methods[7,8], which has worked well in previous demographic inference applications[8,13]. However, instead of calculating the likelihood of the whole sequence, our method calculates the expected composite likelihood of a given SNP frequency spectrum (SFS)[18–20]. Composite likelihood calculation treats each SNP as an independent locus, which significantly reduces the computational burden. This simplified likelihood is a good approximation when the number of SNPs is large and it has worked well in a population parameter estimation application[18]. Therefore, the stairway plot has both the model flexibility of the skyline plot methods and the computational efficiency making it applicable to hundreds of individuals. The number of parameters to be estimated is systematically determined by the standard likelihood ratio test, and can range from 1 to n-1, where n is the number of sequences in the sample. As the method is based on SFS, it has the potential to be applied to pooled sequence data[22] and even species whose reference genome are not yet available[23]. Details of the stairway plot method can be found in Online Methods.

We evaluated the stairway plot using extensive simulation and demonstrated the usage of the method for exploratory demographic inference. Compared to the PSMC method, the stairway plot produced more accurate estimations for recent population size changes. Although it has limited inference accuracy and resolution for more ancient histories, at its applicable range the performances were still comparable to those of the PSMC method. We applied our method to the genomes of nine populations (CEU, GBR, TSI, FIN, CHB, CHS, JPT, YRI, LWK) from the 1000 Genomes Project[24] that are not recently admixed, inferred demographic histories of the populations, and provided interesting hypotheses for future studies, such as ancestors of the FIN population (Finnish in Finland) potentially experienced a recent bottleneck between 10-20 thousand years ago (kya)[25].

RESULTS

Simulation Studies

We validated the stairway plot using extensive coalescent simulations and compared its demographic estimations to those of the PSMC method (see Online Methods). More specifically, for each pre-defined demographic model, we simulated 200 independent samples with ms[26] or MaCS[27] software. For each simulated sample, we used the stairway plot and the PSMC method to infer the demographic history. For the PSMC method, we used the pre-tuned parameters for estimating human population history as suggested by its authors. Along the estimated time span, we calculated the medians and the 2.5 and 97.5 percentiles of the 200 inferred population sizes with the stairway plot and the PSMC method, respectively, and used those percentiles to measure the overall accuracy (by medians) and dispersion (by 2.5 and 97.5 percentiles) of the two methods.

Fig. 2 compared the performances of the stairway plot and the PSMC methods using six different models inspired by previously estimated human population histories. Without loss of generality, one could use the expected number of mutation(s) per base pair (bp) to measure time, and θ per bp to measure population size, where θ=4$N_e\mu$, N_e is the effective population size and μ is the mutation rate per generation. Dividing by μ and 4μ, one can easily convert the above time measure and population size measure to the number of generations and the number of individuals, respectively. Throughout this paper, we assumed a mutation rate of 1.2×10^{-8} per bp per generation[28–30] and a generation time of 24 years[31]. Model 1 (**Fig. 2a**) assumes a constant effective population size of 10,000 individuals. For this model, the medians of inferred histories of both methods fitted well with the true model. Compared to the stairway plot, the PSMC method can infer more ancient history. As to dispersion, that of the stairway plot was smaller (in absolute term) than that of the PSMC method for more recent history, while the opposite was observed for more ancient history. The last two observations were generally true for all models we studied, therefore for the following models we will focus on the accuracy of the two methods for inferring recent histories. Model 2 (**Fig. 2b**) assumes a sudden population size increase at one time point and besides that the population size remains constant, which mimics a previously estimated model for an African population[32]. For this model, the median of the stairway plot's inference fitted almost perfectly with the true model, while that of the PSMC method did not fit very well. Model 3 (**Fig. 2c**) assumes an exponential growth of population size with a rate of 0.004 per generation[32] (i.e. r=0.004). Model 4 (**Fig. 2d**) is another exponential growth model which mimics the estimated recent growth of a population with European ancestry[3]. In both cases, while the stairway plot fits the true model reasonably well, the PSMC is biased upward dramatically. Model 5 (**Fig. 2e**) is based on an estimated human population demographic history[4] with a faster exponential growth rate (r=0.01288). Model 6 (**Fig. 2f**) is a model tested in the PSMC publication[16]. Again, the stairway plot was a better fit to the recent population history than the PSMC.

For inferring more ancient population size changes, we compared the performances of the two methods using four additional models tested in the original PSMC publication plus a population split model (**Supplementary Figure 1**). As we mentioned previously, the stairway plot had a shorter upper limit and a larger dispersion for ancient history inference compared to the PSMC method. The former is a disadvantage for the stairway plot but the latter correctly reflects the uncertainty of our inferences, on the other hand. As to the PSMC method, although it had a smaller dispersion for ancient history inferences, the true histories often fall outside its 95% inference ranges. The stairway plot might produce an artificial bottleneck when the time spans of the last few θ estimations (see Online Methods) overlap with ancient population size fluctuations (see **Supplementary Figure 1e** for an example and Discussion for its recognizable pattern). Overall, within the applicable time spans of the stairway plots, roughly up to the last 10 steps of the plot, the performances of the stairway plot for inferring ancient population size were comparable to the PSMC.

Many factors can affect the inference of the stairway plot. Using simulation we studied the impact of SNP number (or sequence length), sample size and recombination rate. In short, increasing sample size can significantly improve the inference accuracy (median), especially for inferring recent population growth, while the most obvious effects of larger SNP number and recombination rate are reducing the inference dispersion (**Supplementary Figure 2**). The underlying true demographic history determines the information contained in the sample SFS so that the inference results will also affected. There are known caveats related to that; some bottlenecks of the studied population may be missing from the plot due to limitation of inference power. For example, when two bottlenecks are close to each other or a very deep bottleneck following an ancient bottleneck, the stairway plot may not be able to infer the more ancient one (see **Supplementary Figure 3** and more explanation in Discussion).

Application to the 1000 Genomes Project Data

We applied the stairway plot to the whole genome sequences of nine populations (LWK, YRI, CEU, GBR, TSI, FIN, CHB, CHS, JPT) from the 1000 Genomes Project[24]. We restricted our analysis to the genomic regions that are at least 50 kilobase away from any coding regions based on the RefSeq database[33] to avoid potential impacts from natural selection[34]. We also removed regions that are outside the strict mask of the 1000 Genomes Project[24] to reduce artifacts due to mapping errors. Finally, only sites whose ancestral alleles have been inferred with high confidence (see Online Methods) were included for analysis. Because all the SNPs are from intergenic regions and were called with low-depth sequencing, many of the SNPs on the rare spectrum were not observed. We adjusted the SFSs by using the empirical transition probabilities from the SFSs of the high-depth-sequenced exome regions to the SFSs of low-depth-sequenced exome regions, with the assumption that the SFS bias due to low-depth is systematic and universal across the genome (see Online Methods and **Supplementary Note** for details). For each population, 200 bootstrap SFSs were created from the adjusted SFS, and for each bootstrap SFS the stairway plot was used to infer the demographic history. The median inferred population size in each time interval based on the 200 estimations was used to construct a single inferred history of population size. As there were likely artificial bottlenecks observed for all nine populations (**Supplementary Figure 4**), only more recent histories up to 200-300 kya were taken as results. As a higher mutation rate or a lower generation time will lower our time estimation (and on the opposite a lower mutation rate or a higher generation time will heighten our time estimation), we also provided lower and upper estimations for time ranges assuming a (apes-like) generation time of 20 years[35,36] with a mutation rate of 1.4×10^{-8} per bp per generation[37] or a generation time of 30 years[38] with a mutation rate of 1.0×10^{-8} per bp per generation[29,30,39], respectively (in brackets in the following paragraph).

Fig. 3 shows the estimations (see also **Supplementary Figure 4**) and their 95% bootstrap ranges for the nine populations. There are several patterns that are easily observed: (1) Non-African populations all showed severe bottlenecks between 50-70 kya (36-105 kya), which are most likely due to modern human's OOA migration. (2) All non-African populations except the FIN also showed a shallower and more recent bottleneck between 20-30 kya (14-45 kya), and then was followed by size recoveries. The FIN did not show an obvious bottleneck between 20-30 kya, potentially due to limitation of inference power (see Discussion for details), and its size recovery began at around 15 kya (11-23 kya). (4) Compared to the Non-African populations, the two African populations show wider and shallower bottlenecks between 50-70 kya (36-105 kya) and no bottlenecks between 20-30 kya (14-45 kya). (5) Both African populations also show bottlenecks between 100-200 kya (71-300 kya), probably associated with the origination of the anatomically modern human[40]. This bottleneck is not observed in non-African populations, also likely due to limitation of inference power (see Discussion).

DISCUSSION

Here we reported the development of a novel model-flexible method called stairway plot for inferring population demographic histories, which is designed for exploratory or hypothesis generating analysis. There are several other model-flexible methods including the family of skyline plot methods[6–13] and the PSMC method, whose advantages and limitations were briefly discussed in the Introduction. New developments in this area include the diCal method[17] and multiple sequential Markovian coalescent[41] (MSMC). The diCal method extends the PSMC by modeling the configurations of multiple sequences, and showed improvement over the PSMC on inferring recent population histories. However, diCal requires the users to provide haplotypes (i.e. phased sequence data) and a mutation matrix (i.e. relative mutation rates) for the four nucleic bases, which may introduce biases into the estimation if not properly estimated. Besides, the computational intensity limits diCal's application to ~10 sequences. MSMC is another extension of the PSMC method. Instead of modeling all the coalescent events of multiple sequences, it focuses on the first coalescent event and the external branches of coalescent trees. However, due to the modeling and computational complexity, its application is currently limited to roughly 8 phased sequences. Our stairway plot method is based on the composite likelihood of SFS, and therefore has the advantages of efficient computation and the applicability to a broader range of sequence data, such as low-depth sequence[24], pooled sequence[22] and potentially even reference-free transcriptome data[23]. At the current stage, it can be applied to hundreds of unphased sequences. Compared to the PSMC method, the stairway plot can take the advantages of larger sample sizes and provide more accurate inference for recent population histories. However, the stairway plot still has the limitation for inferring ancient histories, for which the PSMC, diCal or MSMC methods may perform better. Therefore, we recommend the complementary usage of the stairway plot with the PSMC, diCal or MSMC.

The application of our stairway plot to nine populations from the 1000 Genomes Project provided some observations worth further and more careful investigation. First, we observed a bottleneck between 10-20 kya in the FIN, which was not observed in other European populations; and vice versa we observed a bottleneck between 20-30 kya in all European populations except the FIN. One explanation of this pattern is that FIN ancestors separated from those of other European populations as earlier as 30 kya. Another possibility is that the FIN may also experience a bottleneck as other European populations, as the shape of its 95% inference ranges suggests a population size decrease around 30 kya. We did some preliminary simulation experiments to investigate the two possibilities (see **Supplementary Note** for details). The results (**Supplementary Figure 3b,c,d**) showed that if a population experienced two continuous bottlenecks, one between 10-20 kya and another between 20-30 kya, our method was not able to infer both bottlenecks. Instead, the plot tended to suggest the more recent bottleneck, which more or less matches the pattern we observed for the FIN. Although we cannot rule out the first explanation, our preliminary analysis suggests the second explanation might be true; that is, the FIN may experience the same bottleneck between 20-30 kya as other European populations, but it may also experience an additional bottleneck between 10-20 kya. Second, we observed that African populations have a bottleneck between 100-200 kya, which is missing in the plots of non-African populations. Again, one possible explanation is that ancestors of all non-African populations separated from those of the African populations as earlier as 200 kya[41], and an alternative explanation is that our method does not have sufficient power to infer that ancient bottleneck with the non-African samples. Because our estimation of population sizes depends on the gene lineages available for coalescence during a period of time, the fewer gene lineages available during the period the less information available for inferring population sizes. As all non-African populations experienced a deep OOA bottleneck between 50-70 kya, many gene lineages of the samples may not survive the bottleneck and be available for inferring more ancient population histories. Although we cannot rule out the first explanation, the simulation experiments we described above supported the alternative explanation, that is, any population having a deep OOA bottleneck did not show an ancient bottleneck between 100-200 kya although the true model has one (**Supplementary Figure 3b,c,d**). However such an ancient bottleneck can be inferred if the population does not have a deep OOA bottleneck (**Supplementary Figure 3a**). Those results also emphasize that interpretations of inferred bottlenecks need to be careful and hypothesis testing is necessary before any conclusions are formulated.

There are many ways the stairway plot can be further improved. As our method models the “average” behavior of many independent coalescent trees, the expectations of coalescent times or $E(t_k)$s are the “building blocks” for the steps observed in the stairway plot. By nature $E(t_k)$ is inversely proportional to $k(k\text{-}1)$ (see Online Methods). Reflecting on the stairway plot, the step size of the plot, which is proportional to $E(t_k)$, is typically much larger when k is small (corresponding to ancient histories) than it is when k is large (corresponding to recent histories) . Put another way, we only model ancient demographic histories using a small number of parameters (or steps as to the plot). When the ancient demographic history is complex, the small number of steps overlapping that complex history may ill-fit the data. A typical result is an artificial bottleneck, which occurs only at the last few (< 10) steps of the plot with a distinguishable pattern of a beginning of population decrease at the second step (θ_3) and a lowest point typically around the third step (θ_4) (see examples in **Supplementary Figure 1e** and **Supplementary Figure 4**). Here we caution users of the stairway plot when such a pattern is observed, the true demographic changes of the population studied may not be correctly reflected. Considering the lower resolution for ancient histories as to the stairway plot, we suggest comparing estimations from various methods (such as the PSMC/MSMC method and diCal) when applicable, and avoiding over-interpretation of the inferred history with the last 10 steps of the stairway plot. One possible improvement for the stairway plot as to the estimation of ancient histories is by integrating the composite likelihood into a Bayesian framework[8,9], which smoothes the θ estimations into continuous probability estimations. A further smoothness can be achieved with a smoothing prior based on a Gaussian Markov random field, in which the smoothness is informed by the data[10]. Another possible improvement for the estimation of the demographic history of a fast growing population, such as for the human population, is by using a different null model. Generally speaking, the underlying null model of the stairway plot is a population of constant size during a certain time period. If an instantaneous size change at a certain point within the period (defined by coalescent times) creates an alternative model with a significantly larger likelihood, the alternative model will replace the null model for further model refinement. This procedure produces a stairway-like inferred population model for a population with a fast size increase or decrease. Assuming an exponential growth model[42] as the null model or a hybrid of the null models of constant size and exponential growth may reduce the number of parameters to be estimated for such populations, and therefore improve the accuracy of estimations. In addition, a more efficient optimization search algorithm for the number and values of θs shall further reduce the computational intensity so that the stairway plot method can be applicable to even larger sample sizes.

ONLINE METHODS

Composite likelihood of a SFS

We assume a random sample of *n* sequences is taken from a population, whose size may instantaneously change at the time points coinciding with coalescent events of the *n* sequences of the gene genealogy (**Fig. 1**). Let t_k be the *k*-coalescent time, then the probability

$$Pr\left(t_k|N_k\right)=\frac{\begin{pmatrix}k\\2\end{pmatrix}}{2N_k}exp\left(-\frac{\begin{pmatrix}k\\2\end{pmatrix}}{2N_k}t_k\right),$$

where N_k is the effective size of the population during t_k. We assume N_k remains constant during t_k, and N_{k-1} or N_{k+1} may be equal to or different to N_k. With a given N_k, a realization of t_k from an independent coalescent tree follows the above distribution. If we summarize a large number of independent coalescent trees, the average of observed t_k will approach its expectation $E(t_k|N_k) = 4N_k/(k(k\text{-}1))$. Let p_i be the probability (or the expectation from a large number of independent coalescent trees) that a nucleic site is a SNP of size *i* ($n\text{-}1 \geq i \geq 1$), then p_i can be expressed as a function of θ_k, where $\theta_k = 4N_k\mu$, and μ is the mutation rate per bp per generation[43]. In more detail,

$$\begin{aligned} p_i= & \mu \sum_{k=2}^{n-i+1} kPr\left(k,i|n\right)E\left(t_k|N_k\right) \\ = & \sum_{k=2}^{n-i+1} kPr\left(k,i|n\right)\tfrac{\theta_k}{k(k-1)} \\ = & \sum_{k=2}^{n-i+1} \tfrac{\Gamma(n-i)\Gamma(n-k+1)}{\Gamma(n-i-k+2)\Gamma(n)}\theta_k \\ & \left(1 \le i \le n-1\right), \end{aligned}$$

where

$$Pr\left(k,i|n\right)=\begin{cases} \dfrac{\begin{pmatrix} n-i-1 \\ k-2 \end{pmatrix}}{\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}} & \text{if} \quad n-i+1 \ge k \ge 2, n>i \ge 1 \\ 0 & \text{otherwise,} \end{cases}$$

For simplicity, we define SNP size 0 as the size of monomorphic sites, and its probability is

$$p_0=1=\sum_{i=1}^{n-1} p_i.$$

Assuming each site is from an independent coalescent tree (i.e. unlinked), the number of SNPs of size *i*, ξ_i, can be modeled with a multinomial distribution and the composite likelihood of observing ξ_0, ξ_1, ... , ξ_{n-1} can be written as

$$L_n=l_n!\prod_{i=0}^{n-1}\frac{p_i^{\xi_i}}{\xi_i!},$$

where

$$l_n=\sum_{i=0}^{n-1}\xi_i.$$

Theoretically, it is possible to use a subset of the SNP sizes for the likelihood calculation with a sacrifice of loss of information contained in those SNP size bins (see **Supplementary Note** for details and potential pitfalls).

When missing data exist, we can separate the whole SNP spectrum into l_n sites with *n* observed alleles, l_{n-1} sites with *n*-1 observed alleles, and l_{n-2} sites with *n*-2 observed alleles, ⋯ . The composite likelihood of the whole data set is

$$L=\prod_{j=1}^{n} L_j.$$

Estimating θs

We used a Java library for numerical optimization called SwarmOps[44] to search for the θs that maximize the composite likelihood of a given SFS. We used a specialized Genetic Algorithm method for real-valued search-spaces called Differential Evolution (DE)[45] if the number of sequences is smaller than 200. Otherwise, we used a Pattern Search (PS) method[44,46]. We used default behavior parameters for DE, and 5000×d and 50×d iterations for DE and PS, respectively, where d is the number of different θs to be estimated.

As there are a total of n-1 different θs that can be estimated, we try to minimize the number of different θs to be estimated by using "break points" to group them. That is, in a ordered serial of θ_2, θ_3, ..., θ_n, break points are inserted into the serials that separate the θs into continuous groups. Any two consecutive θs that are not separated by a break point belong to the same group. We assume the θs within the same group have the same value, while those belonging to different groups may have different values. We also modeled the autocorrelation between the values of adjacent groups of θs following previous successful practices[8].

The procedure for finding the best grouping of θs fitting the observed SFS is as follows: (1) It begins with a single θ, i.e. $\theta_2 = \theta_3 = ... = \theta_n$. Obtain L_1 as the likelihood calculated with this single θ estimation, that is, for a population model of constant size. (2) Increase d by 1; for each point between θ_k and θ_{k+1}, let $\theta_l = \theta_k$ for all $l \leq k$ and $\theta_m = \theta_{k+1}$ for all $m > k$; use SwarmOps to find the estimations of the two θ values that maximized L; calculate L corresponding to that specific break point and the θ estimations; and find the break point with the largest L and designate it as L_2. The procedure stops if $-2\ln(L_1/L_2) < 3.84$, (i.e. a likelihood ratio test with one degree of freedom and $\alpha = 0.05$), otherwise, we accept the new split. (3) increasing d by 1 and repeat the practice; based on the best θ break point(s) associated with L_{d-1}, find an additional break point associated with the largest L and designate it as L_d; and stop when $-2\ln(L_{d-1}/L_d) < 3.84$. As this procedure is not an exhaustive search for the global optimum from the whole parameter space. It is not guaranteed to find the global optimum, especially when the underlying true model is complex. Based on our experiments and observations, the estimation results are typically acceptable approximations for the global optimum (see **Supplementary Note** for results from three example experiments).

Determining the population size at a given time point

Without loss of generality, we use θ to measure population size and mutation per bp to measure time (from the time point when the sample was taken). They can be easily converted to the number of individuals and the number of generations if divided by 4μ and μ, respectively. Given θ_k per bp, the expected length of t_k is $\theta_k/(k(k-1))$. Let

$$T_i=\sum_{k=i}^{n}\frac{\theta_k}{k\,(k-1)}, i=2, 3, \ldots, n,$$

then the stairway plot infers θ at $T_i < T \leq T_{i-1}$ equals θ_{i-1}.

PSMC estimation

The PSMC estimations were conducted using the default parameters tuned for human populations. To measure its dispersion, for each simulated sample or bootstrap sample of multiple individuals, we inferred population size changes using PSMC. Then at each time point along the population history, we calculate the 2.5% and 97.5% percentiles of population size estimations from all inferred histories.

The simulation data

Sequence data were simulated using either ms[26] or MaCS[27] software. Detailed simulation commands can be found in the **Supplementary Note**. If not specified, all sequences were simulated assuming a mutation rate (μ) of 1.2×10^{-8} per bp per generation[28–30] and a recombination of $\rho = 0.8\mu$ per bp per generation. Please note that we used a smaller estimation of recombination, as a recent study suggested that the average recombination rate for humans is about the same as the mutation rate[47].

The 1000 Genomes Project data

The 1000 Genomes Project phase 1 whole genome SNP calls of the nine populations (LWK, YRI, CEU, GBR, TSI, FIN, CHB, CHS, JPT) were downloaded from the 1000 Genomes Project ftp sites. Regions that are within 50 kb from any known coding genes (based on the RefSeq database[33]) and that are outside the 1000 Genomes Project phase 1 strict mask were removed. Sites whose ancestral alleles were not inferred with a high confidence based on the 1000 Genomes Project phase 1 annotation were also removed. The total number of sites in the human genome that passed our filtering is 650,351,035. For each population we calculated SFS only from the retained sites. Because intergenic regions were sequenced with low depth, many of the alleles with low frequencies were not observed. We adjusted the first 20 minor allele frequency bins of each SFS for each population to obtain the most likely true SFS using the empirical transition probabilities that were based on the SFS of the high-depth sequence data of the exome regions and the SFS of low-depth sequence data of the same regions (see **Supplementary Note** for details).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENT

This study was supported by National Institute of Health grants 1U01HG005728 and 2U54HG003273. The authors acknowledge the Texas Advanced Computing Center (TACC, http://www.tacc.utexas.edu) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. The authors acknowledge the 1000 Genomes Project Consortium for generating and providing the DNA sequence resources used in this study. The authors thank Ms. Sara Barton for copyediting the manuscript. The authors thank Drs. Shuhua Xu, Yongtao Guan, Taylor J. Maxwell, Fuli Yu and Eric Boerwinkle for helpful discussions and encouragement.

Reference

1. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012; 336:740–743. [PubMed: 22582263]

2. Schaffner SF, et al. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 2005; 15:1576–1583. [PubMed: 16251467]

3. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009; 5:e1000695. [PubMed: 19851460]

4. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR. Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA. 2009; 106:3871–3876. [PubMed: 19202052]

5. Gravel S, et al. Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA. 2011; 108:11983–11988. [PubMed: 21730125]

6. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 2000; 155:1429–1437. [PubMed: 10880500]

7. Strimmer K, Pybus OG. Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol. Biol. Evol. 2001; 18:2298–2305. [PubMed: 11719579]

8. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 2005; 22:1185–1192. [PubMed: 15703244]

9. Opgen-Rhein R, Fahrmeir L, Strimmer K. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. BMC Evol. Biol. 2005; 5:6. [PubMed: 15663782]

10. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 2008; 25:1459–1471. [PubMed: 18408232]

11. Heled J, Drummond AJ. Bayesian inference of population size history from multiple loci. BMC Evol. Biol. 2008; 8:289. [PubMed: 18947398]

12. Gill MS, et al. Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. Mol. Biol. Evol. 2013; 30:713–724. [PubMed: 23180580]

13. Ho SYW, Shapiro B. Skyline-plot methods for estimating demographic history from nucleotide sequences. Mol. Ecol. Resour. 2011; 11:423–434. [PubMed: 21481200]

14. Atkinson QD, Gray RD, Drummond AJ. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. Proc. R. Soc. B. 2009; 276:367–373.

15. Gignoux CR, Henn BM, Mountain JL. Rapid, global demographic expansions after the origins of agriculture. Proc. Natl. Acad. Sci. USA. 2011; 108:6044–6049. [PubMed: 21444824]

16. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; 475:493–496. [PubMed: 21753753]

17. Sheehan S, Harris K, Song YS. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. Genetics. 2013; 194:647–662. [PubMed: 23608192]

18. Liu X, Fu Y-X, Maxwell TJ, Boerwinkle E. Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. Genome Res. 2010; 20:101–109. [PubMed: 19952140]

19. Nielsen R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics. 2000; 154:931–942. [PubMed: 10655242]

20. Hudson RR. Two-locus sampling distributions and their application. Genetics. 2001; 159:1805–1817. [PubMed: 11779816]

21. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. R. Soc. Lond. A. 1933; 231:289–337.

22. Boitard S, Schlötterer C, Nolte V, Pandey RV, Futschik A. Detecting selective sweeps from pooled next-generation sequencing samples. Mol. Biol. Evol. 2012; 29:2177–2186. [PubMed: 22411855]

23. Gayral P, et al. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet. 2013; 9:e1003457. [PubMed: 23593039]

24. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]

25. Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A. Genetic markers and population history: Finland revisited. Eur. J. Hum. Genet. 2009; 17:1336–1346. [PubMed: 19367325]

26. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002; 18:337–338. [PubMed: 11847089]

27. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. Genome Res. 2009; 19:136–142. [PubMed: 19029539]

28. Kong A, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012; 488:471–475. [PubMed: 22914163]

29. Campbell CD, et al. Estimating the human mutation rate using autozygosity in a founder population. Nat. Genet. 2012; 44:1277–1281. [PubMed: 23001126]

30. Conrad DF, et al. Variation in genome-wide mutation rates within and between human families. Nat. Genet. 2011; 43:712–714. [PubMed: 21666693]

31. Scally A, Durbin R. Revising the human mutation rate: Implications for understanding human evolution. Nat. Rev. Genet. 2012; 13:745–753. [PubMed: 22965354]

32. Boyko AR, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008; 4:e1000083. [PubMed: 18516229]

33. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012; 40:D130–135. [PubMed: 22121212]

34. Lachance J, et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell. 2012; 150:457–469. [PubMed: 22840920]

35. Matsumura S, Forster P. Generation time and effective population size in Polar Eskimos. Proc. Biol. Sci. 2008; 275:1501–1508. [PubMed: 18364314]

36. Langergraber KE, et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc. Natl. Acad. Sci. USA. 2012; 109:15716–15721. [PubMed: 22891323]

37. Awadalla P, et al. Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. Am. J. Hum. Genet. 2010; 87:316–324. [PubMed: 20797689]

38. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. 2005; 128:415–423. [PubMed: 15795887]

39. Michaelson JJ, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012; 151:1431–1442. [PubMed: 23260136]

40. Garrigan D, Hammer MF. Reconstructing human origins in the genomic era. Nat. Rev. Genet. 2006; 7:669–680. [PubMed: 16921345]

41. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 2014; 46:919–925. [PubMed: 24952747]

42. Polanski A, Kimmel M. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics. 2003; 165:427–436. [PubMed: 14504247]

43. Fu YX. Statistical properties of segregating sites. Theor. Popul. Biol. 1995; 48:172–197. [PubMed: 7482370]

44. Pedersen, MEH. PhD Thesis. School of Engineering Sciences, University of Southampton, UK; 2010. Tuning & Simplifying Heuristical Optimization.. University of Southampton

45. Storn R, Price K. Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. 1997; 11:341–359.

46. Davidon WC. Variable metric method for minimization. SIAM J. Optimiz. 1991; 1:1–17.

47. Kong A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010; 467:1099–1103. [PubMed: 20981099]

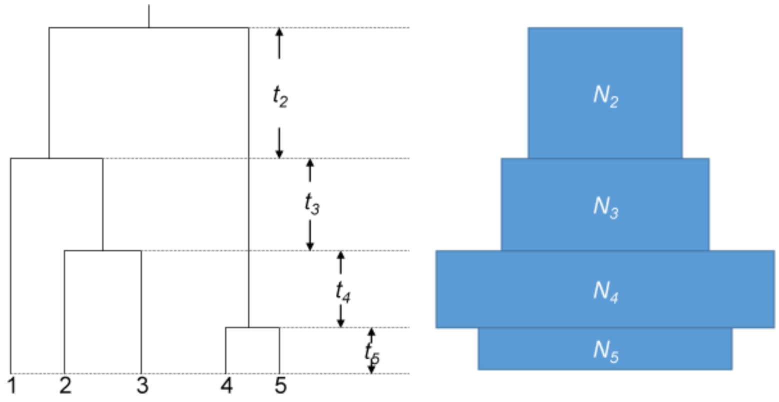

Figure 1.

Illustration of the multi-epoch model. A coalescent tree with corresponding coalescent times is shown on the left and an illustration of the population size (width of the rectangle) changes as multi-epochs with each epoch coinciding with a coalescent event is shown on the right.

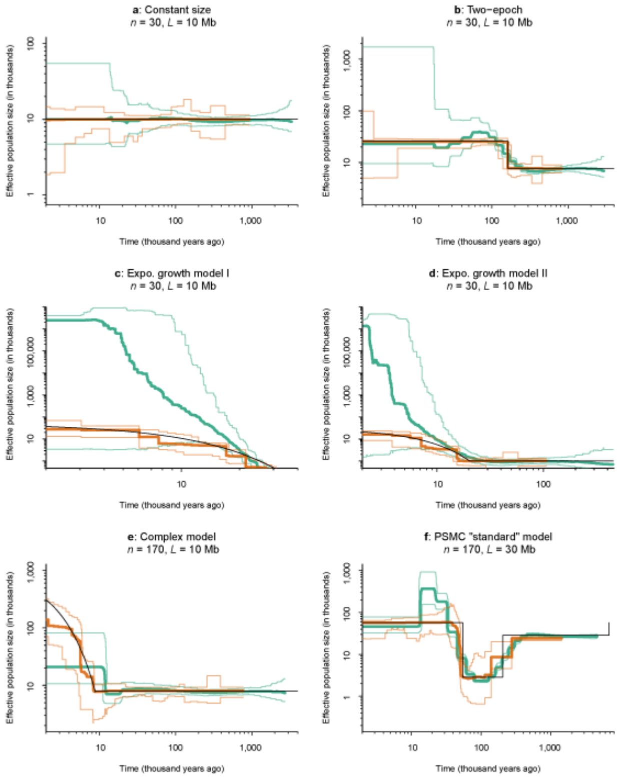

Figure 2.

Comparing the inferred histories of the stairway plot and the PSMC method using simulated samples based on six different models. Subfigures: a: Constant size model; b: Two-epoch model; c: Exponential growth model I; d: Exponential growth model II; e: Complex model; f: PSMC “standard” model. We assumed a mutation rate of 1.2 × 10^{-8} per bp per generation and a generation time of 24 years. Thin black lines: true models. Thick orange lines: medians of the inferred histories of the stairway plot. Thin orange lines: 2.5 and 97.5 percentiles of the inferred histories of the stairway plot. Thick green lines: medians of the inferred histories of the PSMC method. Thin green lines: 2.5 and 97.5 percentiles of the inferred histories of the PSMC method. *n*: number of simulated sequences. *L*: length of simulated sequences.

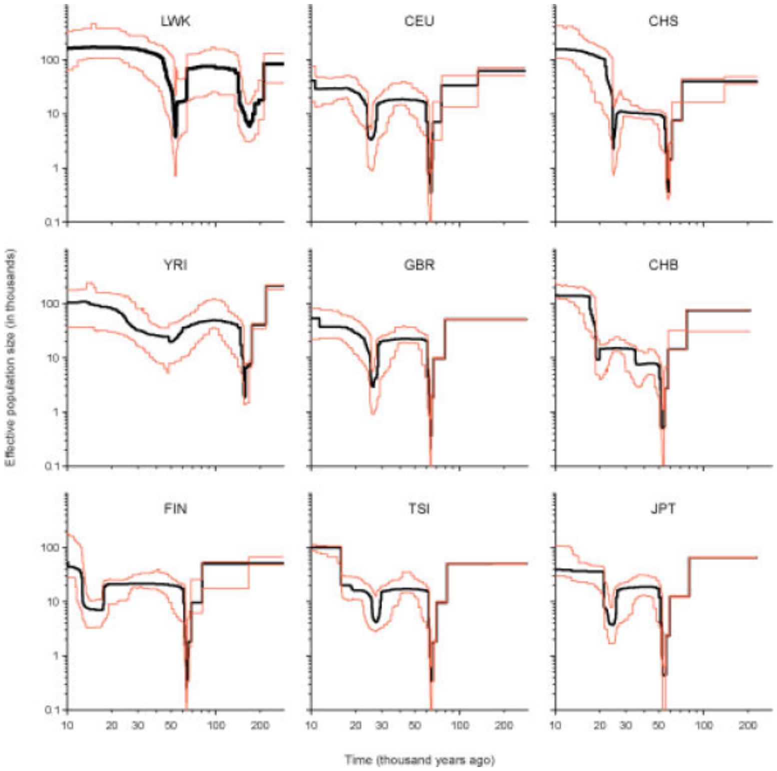

Figure 3.

Inferred histories of nine populations. We assumed a mutation rate of 1.2 × 10^{-8} per bp per generation and a generation time of 24 years. Within each sub-figure, the black lines and two orange lines represent the medians and the 2.5 and 97.5 percentiles, respectively, of the stairway plot's estimations from 200 bootstrap SFSs.