Exploring population size changes using SNP frequency spectra

Xiaoming Liu¹ & Yun-Xin Fu²

Inferring demographic history is an important task in population genetics. Many existing inference methods are based on predefined simplified population models, which are more suitable for hypothesis testing than exploratory analysis. We developed a novel model-flexible method called stairway plot, which infers changes in population size over time using SNP frequency spectra. This method is applicable for whole-genome sequences of hundreds of individuals. Using extensive simulation, we demonstrate the usefulness of the method for inferring demographic history, especially recent changes in population size. We apply the method to the whole-genome sequence data of 9 populations from the 1000 Genomes Project and show a pattern of fluctuations in human populations from 10,000 to 200,000 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 the human genome. Many of the methods available for inferring demographic history with genome-scale data are model constrained; that is, researchers need to predefine 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 parameters to be estimated before estimating the demographic history. The parameters of the models are then estimated by fitting the expected polymorphic pattern (for example, a SNP frequency spectrum) given a set of parameters to that of the observed data, either through extensive simulation or diffusion approximation. In contrast, model-flexible methods (sometimes also called ‘model-free’ methods), such as the skyline plot and its derivatives, are not restricted to a specific demographic model and typically explore larger model space than model-constrained methods. Therefore, model-flexible methods can infer substantially 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. Recently, Li and Durbin 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 users to have a rough idea of the population history 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.

We developed a new method called stairway plot. It uses a flexible, multi-epoch model (Fig. 1) as is implemented in the skyline plot methods, which has worked well in previous demographic inference applications. However, instead of calculating the likelihood of the whole sequence, our method calculates the expected composite likelihood of a given SNP frequency spectrum (SFS) and determines 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.

We evaluated the stairway plot using extensive simulation and demonstrated usage of the method for exploratory demographic inference. In comparison to the PSMC method, the stairway plot produced more accurate estimations for recent changes in population size. Although it has limited inference accuracy and resolution for more ancient histories, at its applicable range, the stairway plot performs comparably to the PSMC method. We applied our method to the genomes of nine populations (CEU (Utah residents (CEPH) with Northern and Western European ancestry), GBR (British in England and Scotland), TSI (Toscani in Italia), FIN (Finnish in Finland), CHB (Han Chinese in Beijing, China), CHS (Southern Han Chinese), JPT (Japanese in Tokyo, Japan), YRI (Yoruba in Ibadan, Nigeria) and LWK (Luhya in Webuye, Kenya)) from the 1000 Genomes Project that are not recently

¹Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA. ²Department of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA. Correspondence should be addressed to X.L. (xiaoming.liu@uth.tmc.edu).

Received 3 November 2014; accepted 26 February 2015; published online 6 April 2015; corrected after print 14 August 2015; doi:10.1038/ng.3254
admixed, inferred the demographic histories of the populations and generated interesting hypotheses for future studies, such as the hypothesis that the ancestors of the FIN population potentially experienced a recent bottleneck between 10,000 and 20,000 years ago.

RESULTS

Simulation studies

We validated the stairway plot using extensive coalescent simulations and compared its demographic estimations to those of the PSMC method (Online Methods). More specifically, for each predefined demographic model, we simulated 200 independent samples with ms25 or MacS26 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 pretuned parameters to estimate human population history, as suggested by its authors. Along the estimated time span, we calculated the medians of the inferred population histories with the stairway plot and the PSMC method, respectively, and used these estimations to measure the overall accuracy (by medians) and dispersion (by 2.5 and 97.5 percentiles) of the methods.

The performances of the stairway plot and PSMC methods using six different models inspired by previously estimated human population histories are compared in Figure 2. Without loss of generality, one could use the expected number of mutation(s) per base pair to measure time and \( \theta \) per base pair to measure population size, where \( \theta = 4N_e \mu \), \( N_e \) is the effective population size and \( \mu \) is the mutation rate per generation. Dividing by \( \mu \) and \( 4\mu \), 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 manuscript, we assumed a mutation rate of \( 1.2 \times 10^{-8} \) mutations per base pair per generation and a generation time of 24 years.

Model 1 (Fig. 2a) assumed a constant effective population size of 10,000 individuals. For this model, the medians of the inferred histories of both methods fitted well with the true model. In comparison to the stairway plot, the PSMC method could infer more ancient history. On one hand, the stairway plot had a shorter upper limit and a larger dispersion for the PSMC method for more recent history, whereas the opposite was observed for more ancient history. These two observations were generally true for all models we studied; for the following models, we therefore focus on the accuracy of the two methods for inferring recent histories.

Model 2 (Fig. 2b) assumed a sudden increase in population size at one time point, aside from which the population size remained constant, mimicking a previously estimated model for an African population. For this model, the median of the stairway plot’s inference fitted almost perfectly with the true model, whereas that of the PSMC method did not fit very well. Model 3 (Fig. 2c) assumed an exponential increase in population size at a rate of \( r = 0.004 \) per generation \((r = 0.004)\), where \( N_0 \) is the current population size and \( t \) is the time before the present, in units of \( 4N_0 \) generations. Model 4 (Fig. 2d) was another exponential growth model that mimicked the estimated recent growth of a population with European ancestry. In both cases, whereas the stairway plot fit the true model reasonably well, the PSMC method was dramatically biased upward. Model 5 (Fig. 2e) was based on an estimated human population demographic history with a faster exponential growth rate \((r = 0.01288)\). Model 6 (Fig. 2f) was a model tested in the PSMC publication. Again, the stairway plot was a better fit to the recent population history than the PSMC method.

Figure 1

Illustration of the multi-epoch model. Left, a coalescent tree with corresponding coalescent times. Right, an illustration of how the population size (width of each rectangle) changes over multiple epochs, with each epoch coinciding with a coalescent event.

Figure 2

Comparing the inferred histories of the stairway plot and the PSMC method using simulated samples on the basis of six different models. (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 \( 1.2 \times 10^{-8} \) mutations per base pair per generation and a generation time of 24 years. Thin black lines, the true models. Thick orange lines, the 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. \( n \) is the number of simulated sequences, and \( L \) is the length of the simulated sequences.
the PSMC method had smaller dispersion for inferences of ancient history, the true histories often fell outside its 95% inference ranges. The stairway plot might produce an artificial bottleneck when the time spans of the last few \( \theta \) estimations (Online Methods) overlap with ancient fluctuations in population size (see Supplementary Fig. 1e for an example). Overall, within the applicable time spans of the stairway plots, roughly up to the last ten steps of each plot, the performances of the stairway plot for inferring ancient population size were comparable to those for the PSMC method.

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 substantially improve the inference accuracy (median), especially in inferring recent population growth, whereas the most obvious effect of larger SNP number and recombination rate is a reduction in the inference dispersion (Supplementary Fig. 2). The underlying true demographic history determines the information contained in the sample SFS, such that the inference results will also be affected. There are known caveats related to this; some bottlenecks of the studied population may be missed from the plot owing to limitation of inference power. For example, when two bottlenecks are close to each other or a very deep bottleneck follows an ancient bottleneck, the stairway plot may not be able to infer the more ancient one (Supplementary Fig. 3).

Application to 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 and JPT) from the 1000 Genomes Project\(^{53}\). We restricted our analysis to genomic regions that were at least 50 kb away from any coding region according to the RefSeq database\(^{52}\) to avoid potential effects from natural selection\(^{53}\). We also removed regions that were outside the strict mask of the 1000 Genomes Project\(^{53}\) to avoid artifacts due to mapping errors. Finally, only sites whose ancestral alleles had been inferred with high confidence (Online Methods) were included for analysis. Because all the SNPs were 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 exome regions sequenced to high depth to the SFSs of the exome regions sequenced to low depth, with the assumption that the SFS bias due to low-depth sequencing was systematic and universal across the genome (see the 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 nine populations (Supplementary Fig. 4), only more recent histories up to 200,000–300,000 years ago were taken as results. As a higher mutation rate or a shorter generation time would lower our time estimation (and, on the opposite end, a lower mutation rate or a longer generation time would increase our time estimation), we also provided lower and upper estimations for the time ranges, assuming an (ape-like) generation time of 20 years\(^{34,35}\) with a mutation rate of 1.4 \( \times 10^{-8} \) mutations per base pair per generation\(^{36}\) and a generation time of 30 years\(^{37}\) with a mutation rate of 1.0 \( \times 10^{-8} \)mutations per base pair per generation\(^{28,29,38}\), respectively (the results of these estimations are shown in parentheses in the following paragraph).

The estimations (see also Supplementary Fig. 4) and their 95% bootstrap ranges for the nine populations are shown in Figure 3. There were several patterns that were easily observed. (i) Non-African populations all showed severe bottlenecks between 50,000 and 70,000 years ago (36,000–105,000 years ago), which are most likely due to the out-of-Africa migration of modern humans. (ii) All non-African populations except the FIN population also showed a shallower and more recent bottleneck between 20,000 and 30,000 years ago (14,000–45,000 years ago) that was then followed by a recovery in population size. (iii) The FIN population did not show an obvious bottleneck between 20,000 and 30,000 years ago, potentially because of limitation of inference power, and its recovery of size began at around 15,000 years ago (11,000–23,000 years ago). (iv) In comparison to the non-African populations, the two African populations showed wider and shallower bottlenecks between 50,000 and 70,000 years
(36,000–105,000 years ago) and no bottlenecks between 20,000 and 30,000 years ago (14,000–45,000 years ago). (v) Both African populations also showed bottlenecks between 100,000 and 200,000 years ago (71,000–300,000 years ago), probably associated with the origin of the anatomically modern human. This bottleneck was not observed in non-African populations, also likely owing to limitation of inference power.

**DISCUSSION**

Here we report 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 and the PSMC method, whose advantages and limitations were briefly discussed herein. New developments in this area include the diCal method and multiple sequential Markovian coalescent (MSMC). The diCal method extends the PSMC method by modeling the configurations of multiple sequences and showed improvement over the PSMC method in inferring recent population histories. However, diCal requires users to provide haplotypes (phased sequence data) and a mutation matrix (relative mutation rates) for the four nucleic bases, which may introduce biases into the estimation if these are not properly estimated. Furthermore, the computational intensity of diCal limits its 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, because of the modeling and computational complexity, its application is currently limited to roughly eight phased sequences. Our stairway plot method is based on the composite likelihood of a given SFS and therefore has the advantages of efficient computation and applicability to a broader range of sequence data, such as low-depth sequence data, pooled sequence data and potentially even reference-free transcriptome data. At the current stage, it can be applied to hundreds of unphased sequences. In comparison 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 a limitation on inferring ancient histories, for which the PSMC, diCal or MSMC methods may perform better. Therefore, we recommend complementary use of the stairway plot with 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 10,000 and 20,000 years ago in the FIN population, which was not observed in other European populations; in opposition to this, we observed a bottleneck between 20,000 and 30,000 years ago in all European populations except the FIN population. One explanation of this pattern is that the FIN ancestors separated from those of other European populations as early as 30,000 years ago. Another possibility is that the FIN population may also have experienced the same bottleneck as the other European populations, as the shape of its 95% inference ranges may suggest the more recent bottleneck, which more or less matches the pattern we observed for the FIN population. Although we cannot rule out the first explanation, the simulation experiments we described above support the alternative explanation; that is, any population having a deep out-of-Africa bottleneck did not show an ancient bottleneck between 100,000 and 200,000 years ago, although the true model has one (Supplementary Fig. 3b–d). However, such an ancient bottleneck can be inferred if the population does not have a deep out-of-Africa bottleneck (Supplementary Fig. 3a). These results also emphasize that interpretations of inferred bottlenecks should be undertaken with care and that 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 $E(t_k)$ values, are the building blocks for the steps observed in the stairway plot. By its nature, $E(t_k)$ is inversely proportional to $k(k - 1)$ for the kth coalescent event (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 with respect to the plot). When the ancient demographic history is complex, the small number of steps overlapping that complex history may fit the data poorly. A typical result is an artificial bottleneck, which occurs only at the last few (<10) steps of the plot with a distinguishable pattern comprising a beginning to a population decrease at the second step ($\theta_1$) and a lowest point typically around the third step ($\theta_2$) (see examples in Supplementary Figs. 1e and 4). Here we caution users of the stairway plot that, when such a pattern is observed, the true demographic changes in the population studied may not be correctly reflected. Considering the lower resolution for ancient histories with respect to the stairway plot, we suggest comparing the estimations from various methods (such as the PSMC, MSMC and diCal methods), when applicable, and avoiding overinterpretation of the inferred history with the last ten steps of the stairway plot. One possible improvement for the stairway plot with respect to the estimation of ancient histories is achieved by integrating the composite likelihood into a Bayesian framework, which smoothens the $\theta$ estimations into continuous probability estimations. 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. Another possible improvement for estimation of the demographic history of a fast-growing population, such as the human population, is achieved 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 change in size at a certain point within the period (defined by coalescent times) creates an alternative model with a substantially 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 increase or decrease in size. Assuming an exponential growth model 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 the estimations. In addition, a more efficient optimization search algorithm for the number and values of \( \theta \) will further reduce the computational intensity so that the stairway plot method can be applicable to even larger sample sizes.

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

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors acknowledge the Texas Advanced Computing Center (TACC; https://www.tacc.utexas.edu) at the University of Texas at Austin for providing high-performance computing 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 S. Barton for copyediting the manuscript. The authors thank S. Xu, Y. Guan, T.J. Maxwell, F. Yu and E. Boerwinkle for helpful discussions and encouragement. This study was supported by US National Institutes of Health grants U10HG005728 and 2U54HG003273 and by Chinese NSF grant 9121120.

AUTHOR CONTRIBUTIONS

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

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336, 740–743 (2012).
2. Schaffner, S.F. et al. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15, 1576–1583 (2005).
3. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5, e1000695 (2009).
4. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA 106, 3871–3876 (2009).
5. Gravel, S. et al. Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA 108, 11983–11988 (2011).
6. Pybus, O.G., Rambaut, A. & Harvey, P.H. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155, 1429–1437 (2000).
7. Strimmer, K. & Pybus, O.G. Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol. Biol. Evol. 18, 2298–2305 (2001).
8. Drummond, A.J., Rambaut, A., Shapiro, B. & Pybus, O.G. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185–1192 (2005).
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. 5, 6 (2005).
10. Minin, V.N., Bloomquist, E.W. & Suchard, M.A. Smooth skyline by a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25, 1459–1471 (2008).
11. Heled, J. & Drummond, A.J. Bayesian inference of population size history from multiple loci. BMC Evol. Biol. 8, 289 (2008).
12. Gill, M.S. et al. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol. Biol. Evol. 30, 713–724 (2013).
13. Ho, S.Y.W. & Shapiro, B. Skyline-plot methods for estimating demographic history from nucleotide sequences. Mol. Ecol. Resour. 11, 423–434 (2011).
14. Atkinison, Q.D., Gray, R.D. & Drummond, A.J. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. Proc. Biol. Soc. 276, 367–373 (2009).
15. Gignoux, C.R., Henn, B.M. & Mountain, J.L. Rapid, global demographic expansions after the origins of agriculture. Proc. Natl. Acad. Sci. USA 108, 6044–6049 (2011).
16. H. & Durbin, R. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496 (2011).
17. Sheehan, S., Harris, K. & Song, Y.S. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution model. Mol. Ecol. 19, 303–317 (2010).
18. Liu, X., Fu, Y.-X., Maxwell, T.J. & Boerwinkle, E. Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. Genome Res. 20, 101–109 (2010).
19. Nielson, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154, 931–942 (2000).
20. Hudson, R.R. Two-locus sampling distributions and their application. Genetics 159, 1805–1817 (2001).
21. Boitard, S., Schlötterer, C., Nolte, V., Pandey, R.V. & Futschik, A. Detecting selective sweeps from pooled next-generation sequencing samples. Mol. Biol. Evol. 29, 2177–2186 (2012).
22. Gayral, P. et al. Reference-free population genomics from next-generation transcriptome data and the vertebrate-riverbedgap Early. Proc. Natl. Acad. Sci. USA, 13, 1336–1346 (2009).
23. Hudson, R.R. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337–338 (2002).
24. Chen, G.K., Marjoram, P. & Wall, J.D. Fast and flexible simulation of DNA sequence data. Genome Res. 19, 136–142 (2009).
25. Kong, A. et al. Rate of de novo mutations and the importance of father’s age to disease risk. Nature 488, 471–475 (2012).
26. Campbell, C.D. et al. Estimating the human mutation rate using autozygosity in a founder population. Nat. Genet. 44, 1277–1281 (2012).
27. Conrad, D.F. et al. Variation in genome-wide mutation rates within and between human families. Nat. Genet. 43, 712–714 (2011).
28. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. 13, 745–753 (2012).
29. Boyko, A.R. et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 6, e1000838 (2008).
30. Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40, D130–D135 (2012).
31. Lachance, J. et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150, 457–469 (2012).
32. Matsumura, S. & Forster, P. Generation time and effective population size in Polar Eskimos. Proc. Biol. Soc. 275, 1501–1508 (2008).
33. Langegraber, K.E. et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc. Natl. Acad. Sci. USA 109, 15716–15721 (2012).
34. Awadalla, P. et al. Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. Am. J. Hum. Genet. 87, 316–324 (2010).
35. Fenner, J.N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. 128, 415–423 (2005).
36. Michaelsson, J.J. et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151, 1431–1442 (2012).
37. Garrigan, D. & Hammer, M.F. Reconstructing human origins in the genomic era. Nat. Rev. Genet. 7, 669–680 (2006).
38. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46, 919–925 (2014).
39. Polanski, A. & Kimmel, M. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165, 427–436 (2003).
ONLINE METHODS

Composite likelihood of an 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 the coalescent events of the n sequences of the genealogy (Fig. 1). Let \( t_k \) be the kth coalescent time, then we get the probability

\[
\Pr(k | N_k) = \frac{\Gamma(n)}{2N_k^n} \exp \left( \frac{1}{2} \right) \Gamma(k),
\]

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 from \( N_k \). 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 the observed \( t_k \) values will approach its expectation \( E(t_k | N_k) = 4N_k/(k(k-1)) \). Let \( p_k \) be the probability (or the expectation from a large number of independent coalescent trees) that a nucleic site is a SNP of size \( i \) \((n-1 \geq i \geq 1)\), then \( p_k \) can be expressed as a function of \( \theta_k \), where \( \theta_k = 4N_k \mu_k \) and \( \mu \) is the mutation rate per base pair per generation.\(^{30}\)

In more detail,

\[
p_k = \mu \sum_{k=2}^{n-i+1} kP(k,i,n)E(t_k | N_k) = \sum_{k=2}^{n-i+1} kP(k,i,n) \frac{\theta_k}{k(k-1)}
\]

where

\[
Pr(k, i | n) = \begin{cases} \frac{n-i}{k-1} & \text{if } n-i+1 \geq k \geq 2, n-i \geq 1 \\ 0 & \text{otherwise} \end{cases}
\]

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 (unlinked), the number of SNPs of size \( i \), \( \xi_i \), can be modeled with a multinomial distribution and the composite likelihood of observing \( \xi_0, \xi_1, \ldots, \xi_{n-1} \) can be written as

\[
L_n = \prod_{i=0}^{n-1} \binom{n}{i} p_i^{\xi_i}
\]

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 the information contained in those SNP size bins (see the Supplementary Note for details and potential pitfalls).

When missing data exist, we can separate the whole SNP spectrum into \( l_i \) sites with \( n \) observed alleles, \( l_{i-1} \) sites with \( n-1 \) observed alleles and \( l_{i-2} \) sites with \( n-2 \) observed alleles, etc. The composite likelihood of the whole data set is

\[
L = \prod_{j=1}^{n} l_j
\]

**Estimating \( \theta \) values.** We used a Java library for numerical optimization called SwarmOps\(^{43}\) to search for the \( \theta \) values 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)\(^{44}\) if the number of sequences was smaller than 200. Otherwise, we used a Pattern Search (PS) method.\(^{45,46}\) We used default behavior parameters for DE, and 5,000 \( \times \) and 50 \( \times \) iterations for DE and PS, respectively, where \( d \) is the number of different \( \theta \) values to be estimated.

As there are a total of \( n-1 \) different \( \theta \) values that can be estimated, we tried to minimize the number of different \( \theta \) values to be estimated by using ‘break points’ to group them. That is, in an ordered serial of \( \theta_2, \theta_3, \ldots, \theta_n \), break points were inserted into the serials that separated the \( \theta \) values into continuous groups. Any two consecutive \( \theta \) values that were not separated by a break point belonged to the same group. We assumed that the \( \theta \) values within the same group had the same value, whereas those belonging to different groups might have different values. We also modeled the autocorrelation between the values of adjacent groups of \( \theta \) values following previously successful practices.\(^8\)

The procedure for finding the best grouping of \( \theta \) values fitting the observed SFS was as follows: (1) the process begins with a single \( \theta \), i.e., \( \theta = \theta_1 = \ldots = \theta_n \). Obtain \( L_1 \) as the likelihood calculated with this single \( \theta \) estimation, that is, for a population model of constant size. (2) Increase \( d \) by 1; for each point between \( \theta_2 \) and \( \theta_{n-1} \), let \( \theta_i = \theta_i \) for all \( i \leq k \) and \( \theta_{i+1} = \theta_{i+1} \) for all \( m > k \); use SwarmOps to find the estimations of the two \( \theta \) values that maximize \( L \); calculate \( L \) corresponding to that specific break point and the \( \theta \) estimations; and find the break point with the largest \( L \) and designate it as \( L_2 \). The procedure stops if \( -2 \ln(L_1/L_2) > 3.84 \) (a likelihood-ratio test with one degree of freedom and \( \alpha = 0.05 \)); otherwise, we accept the new split. (3) Increase \( d \) by 1 and repeat the practice; on the basis of the best \( \theta \) breakpoint(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 \( -2 \ln(L_1/L_d) < 3.84 \). 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. On the basis of our experiments and observations, the estimation results are typically acceptable approximations for the global optimum (see the Supplementary Note for the results from three example experiments).

**Determining the population size at a given time point.** Without loss of generality, we use \( \theta \) to measure population size and mutations per base pair to measure time (from the time point when the sample was taken). These values can be easily converted to the number of individuals and the number of generations if divided by \( 4\mu \) and \( \mu \), respectively. Given \( \theta_k \) per base pair, the expected length of \( t_k \) is \( \theta_k/(4k(k-1)) \). Let

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

then the stairway plot infers \( \theta \) at \( T_i < T < T_{i-1} \) equals \( \theta_{i-1} \).

**PSMC estimation.** 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 calculated the 2.5 and 97.5 percentiles of the population size estimations from all inferred histories.

**Simulation data.** Sequence data were simulated using either ms\(^{25}\) or MaCS\(^{26}\) software. Detailed simulation commands can be found in the Supplementary Note. If not specified, all sequences were simulated assuming a mutation rate \( \mu \) of \( 1.2 \times 10^{-9} \) per base pair per generation\(^{25-29}\) and a recombination rate \( \rho = 0.8\mu \) per base pair per generation. Please note that we used a smaller estimation of the recombination rate, as a recent study suggested that the average recombination rate for humans is about the same as the mutation rate.\(^{42}\)

1000 Genomes Project data. 1000 Genomes Project phase 1 whole-genome SNP calls for the nine populations (LWK, YRI, CEU, GBR, TSJ, FIN, CHB, CHS and JPT) were downloaded from the 1000 Genomes Project ftp sites. Regions that were within 50 kb of any known coding genes (according to the RefSeq database)\(^{32}\) and that were outside the 1000 Genomes Project phase 1 strict mask were removed. Sites whose ancestral alleles were not inferred with high confidence according to the 1000 Genomes Project phase 1
annotation were also removed. The total number of sites in the human genome that passed our filtering was 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 the Supplementary Note for details).

42. Fu, Y.X. Statistical properties of segregating sites. *Theor. Popul. Biol.* **48**, 172–197 (1995).
43. Pedersen, M.E.H. *Tuning and Simplifying Heuristical Optimization*. PhD thesis, Univ. Southampton (2010).
44. Storn, R. & Price, K. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **11**, 341–359 (1997).
45. Davidon, W.C. Variable metric method for minimization. *SIAM J. Optim.* **1**, 1–17 (1991).
46. Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
Corrigendum: Exploring population size changes using SNP frequency spectra

Xiaoming Liu & Yun-Xin Fu

Nat Genet. 47, 555–559 (2015); published online 6 April 2015; corrected after print 14 August 2015

In the version of this article initially published, the authors neglected to acknowledge one of the funding sources for their study. The acknowledgements should have recognized support from Chinese NSF grant 91231120 in addition to the other funding sources listed. The error has been corrected in the HTML and PDF versions of the article.