Genomic inference of a human super bottleneck in the Early Stone Age

Wangjie Hu[1,†], Ziqian Hao[1,†], Pengyuan Du[1], Yi-Hsuan Pan[2,*], Haipeng Li[1,3,*]

[1]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China.

[2]Key Laboratory of Brain Functional Genomics of Ministry of Education, School of Life Science, East China Normal University, Shanghai 200062, China.

[3]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China.

†These authors contributed equally.

*Corresponding Authors: yxpan@sat.ecnu.edu.cn; lihaipeng@picb.ac.cn

Short title: Human super bottleneck

Abstract

The demographic history has been a foundation of human evolutionary studies for more than a century. In this study, we developed a novel method referred to as the fast infinitesimal time coalescent (FitCoal) process. This method allows the accurate calculation of the composite likelihood of a site frequency spectrum and provides the precise inference of recent and ancient demographic history. Genomic sequences of the 1000 Genomes Project and the Human Genome Diversity Project – Centre d'Etude du Polymorphisme Humain panel were analyzed. Results showed that all ten African populations had a population super bottleneck, a small effective size of approximately 1,280 breeding individuals between 813 and 930 thousand years ago, and a 20-fold rapid growth at the end of the bottleneck. The super bottleneck caused a loss of 65.85% in current human genetic diversity, but it may have separated our ancestors from other hominins. Further analysis confirmed the existence of the super bottleneck in all 40 non-African populations. Our results provide new insights into human evolution in the Early Stone Age.

Inferring demographic history from genomic information has played an important role in population genetics. It uncovers prehistoric evolutionary events and deepens our understanding about the evolution of human and other species[3-7]. Multiple methods have been developed to infer demographic history with a predefined demographic model[8-12]. These methods require prior knowledge about the species being investigated and estimation parameters by fitting in summary statistics such as site frequency spectrum (SFS). In contrast, model-free methods do not need a predefined model when inferring demography[13-18]. As SFS plays an essential role in demographic inference, many efforts have been attempted to derive its analytical formula under certain demographic models[19-21].

To precisely infer recent and ancient demography, we developed the fast infinitesimal time coalescent (FitCoal) process (Fig. 1) that calculates expected branch length for each SFS type under arbitrary demographic models. It is effective for a wide range of sample sizes in the calculation of the composite likelihood of a given SFS[8,9]. To infer the demographic history, FitCoal first maximizes the likelihood with the constant size model and then increases the number of inference time intervals and re-maximizes the likelihood until the best model is found. As inference time intervals are variable to avoid long fixed ancient time intervals, the precision in the inference of ancient demographic events can be improved. FitCoal does not need prior information on demography, and its accuracy can be confirmed by simulation.

With African hominid fossils, the origin of anatomically modern humans has been determined to be approximately 200 thousand years (kyr) ago[22]. Although the demographic history of anatomically modern humans has been intensively studied[14,15,17,23-26], it is conceivable that many new insights remain to be explored. In this study, we used FitCoal to analyze genomic sequences of the 1000 Genomes Project phase 3 (1000GP)[1] and the Human Genome Diversity Project–Centre d'Etude du Polymorphisme Humain panel (HGPD-CEPH)[2]. Results revealed a super bottleneck in all 10 African populations between 813 and 930 kyr ago with an effective population size of 1,280. According to the coalescent theory and simulations,

the super bottleneck cannot be directly inferred in the non-African populations because fewer coalescent events remained in the non-African populations during the bottleneck period. Instead, a hidden effect of the bottleneck was found in all 40 non-African populations. Our results suggest that our ancestors experienced a super bottleneck and the effective size of our ancestors remained small for a very long period of time (~117,000 years). The super bottleneck may have separated our ancestors from other hominins[27,28].

Fast infinitesimal time coalescent process

As determination of expected branch length for each SFS type is essential for theoretical population genetics and demographic inference[8,9], we developed the fast infinitesimal time coalescent (FitCoal) process to accomplish the task (Fig. 1). For FitCoal analysis, each of millions of time intervals Δt was set extremely small, and the population size was assumed to be constant within each infinitesimal time interval. The probabilities of all states were calculated backward in time. During each Δt, the branches were categorized according to their state. For each state, the branch length was multiplied by its probability and population size and then transformed to calculate the expected branch length of each SFS type. Because the expected branch length of a SFS type is equal to the sum of the expected branch length of this type during each time interval, the latter can be rescaled and tabulated, making the calculation of the expected branch lengths extremely fast under arbitrary demographic histories. Hereafter, tabulated FitCoal is referred to as FitCoal for short, unless otherwise indicated.

FitCoal demographic inference

After the expected branch lengths were obtained, the composite likelihood of the SFS observed in a sample was calculated[8,9,15,29]. As each single nucleotide polymorphism (SNP) was treated independently, FitCoal did not need phased haplotype data. When inferring demography, the likelihood was maximized in a wide range of demographic scenarios. The FitCoal likelihood surface is smooth (Fig. S1), so it is efficient to maximize the likelihood. FitCoal considered both instantaneous populations size changes[14-16] and long-term exponential changes of population in

order to generate various demographic scenarios.

Demographic inference on simulated data

The accuracy of FitCoal was validated by simulation and comparing its demographic inferences with those of PSMC[14] and stairway plot[15] (Fig. 2). Six different demographic models, examined in the study by Liu and Fu[15], were considered by simulating 200 independent data sets under each model. The medians and 95% confidence intervals of demography were then determined by FitCoal with the assumption that a generation time is 24 years[15,30] and the mutation rate is 1.2×10^{-8} per site per generation for human populations[15,31-33]. Our results (Fig. 2) confirmed that SFS allows precise recovery of the demographic history[34].

FitCoal was found to precisely infer demographic histories (Fig. 2), and the inference accuracy was improved by increasing sample size and length of sequence (Fig. S2). In general, the confidence intervals of FitCoal inferred histories were narrower than those of PSMC and stairway plot, except for those with insufficient information on ancestral populations (Fig. 2c). The proportion of the most recent change type inferred from the six different models mentioned above also showed that FitCoal can distinguish instantaneous and exponential changes (Table S1).

Since a demographic event may affect every SFS type, demographic history can be inferred using a subset of SFS. Results of simulation confirmed that FitCoal accurately determined demographic history based on truncated SFSs (Fig. S3, 4), thus reducing the impact of other factors, such as positive selection (Fig. S5) and sequencing error, on FitCoal analysis.

Demographic inference of African populations

To infer the demographic histories of African populations, seven African populations in the 1000GP[1] were analyzed by FitCoal. Only non-coding regions, defined by GENCODE[35], were used in order to avoid the effect of purifying selection. To avoid the potential effect of positive selection[36], high-frequency mutations were excluded from the analysis.

Results showed that all seven African populations had a super bottleneck around 914 (854–1,003) kyr ago and that this bottleneck was relieved about 793 (772–815) kyr ago (Fig. 3a-c, S6; Table S2). The average effective population size of African

131 populations during the bottleneck period was determined to be 1,270 (770–2,030).
132 Although traces of the bottleneck were observed in previous studies, the bottleneck
133 was ignored because its signatures were too weak to be noticed[1,2,14,16,17]. After the
134 bottleneck was relieved, the population size was increased to 27,080 (25,300–29,180),
135 a 20-fold increase, around 800 kyr ago. This population size remained relatively
136 constant until the recent expansion.

137 To avoid the potential effects of low sequencing depth (~ 5x) of non-coding
138 regions in the 1000GP on the analysis, the autosomal non-coding genomic
139 polymorphism of HGDP-CEPH data set with high sequencing coverage (~35x) was
140 used. In total, populations with more than 15 individuals each were examined. Results
141 showed that the super bottleneck occurred between 859 (856–864) and 1,257 (1,042–
142 1,527) kyr ago in all three African populations in HGDP-CEPH (Fig. 3d-f, S7; Table
143 S3), and the average population size during the bottleneck period was 1,300 (908–
144 1,670). This number was very similar to that (1,270) estimated from the data of
145 1000GP.

146 After the bottleneck was relieved, the population sizes of the two HGDP-CEPH
147 agriculturalist populations were increased to 27,300 and 27,570 (Fig. 3e, S7; Table
148 S3), consistent with the 1000GP estimate of 27,280. However, the Biaka, a
149 hunter-gatherer population, had a larger population size of 35,330, suggesting a deep
150 divergence between this and other agriculturalist populations[37-39]. The Biaka
151 population was found to have a recent population decline (Fig. 3d, S7), as previously
152 observed[2]. These results suggest that hunter-gatherer populations were widely spread
153 and decreased when agriculturalist populations were expanded.

154 To provide a precise inference of the super bottleneck, the results from the two
155 data sets were combined. After analyzing the inferred time of instantaneous change of
156 ten populations, the super bottleneck was inferred to last for about 117,000 years,
157 from 813 (772–864; s.e.m.: 11.02) to 930 (854–1,042; s.e.m.: 23.52) kyr ago. The
158 effective size during the bottleneck period was precisely determined to be 1,280 (767–
159 2,031; s.e.m.: 131). A loss of 65.85% in current genetic diversity of human
160 populations was estimated because of the bottleneck.

161

162 **Demographic inference of non-African populations**

No super bottleneck was directly observed in all 19 non-African populations in 1000GP (Fig. 3a-c, S6; Table S4). The ancestral population size of these populations was determined to be 20,260 (18,850–22,220), similar to that determined in previous studies[2,14,16,17]. The population size started to decline around 368 (175–756) kyr ago in 1000GP non-African populations, suggesting that African and non-African divergence occurred much earlier than the out-of-African migration[1,2,14,16,17,24]. European and South Asian populations were found to have a relatively weaker bottleneck than East Asian populations, and the bottleneck severity was found to correlate with their geographic distance to Africa, consistent with the observed correlation between heterozygosity and geographic distance[40,41]. A weak bottleneck was observed in American populations, probably because of recent admixture[1]. All 1000GP non-African populations were found to increase in size recently.

The super bottleneck was also not directly detected in all 21 HGDP-CEPH non-African populations (Fig. 3d-f, S7; Table S5). The ancestral population size of these populations was determined to be 20,030 (19,060–21,850), very similar to that (20,260) estimated from 1000GP. These populations started to decline 367 (167–628) kyr ago. A positive correlation was also observed between the severity of out-of-African bottleneck and their geographic distance to Africa. The Middle East populations had the weakest bottleneck, while the Maya, an American population, had the strongest bottleneck. Similar to 1000GP non-African populations, most HGDP-CEPH non-African populations were found to increase in size recently, except an isolated Kalash population, consistent with previous studies[2,42].

Super bottleneck in the Early Stone Age

Although a super bottleneck was detected in all 10 African populations, such bottleneck was not directly detected in all 40 non-African populations. To investigate this phenomenon, simulations were performed with three 1000GP demographic models, designated Bottleneck I, II, and III (Fig. 4). Bottleneck I simulated the average inferred demographic history of African populations with the super bottleneck, and Bottleneck II and III simulated the demography of non-African populations without and with the super bottleneck. Both Bottleneck I and II were inferred correctly as a corresponding bottleneck was found in all simulated data sets (Table S6). However, no super bottleneck was detected in Bottleneck III simulations.

The super bottleneck was found to cause a population size gap between the true model and inferred demographic history, suggesting a hidden effect of the super bottleneck on non-African populations. Simulations were then extended to HGDP-CEHP populations with Bottleneck models IV–VI, and similar results were obtained (Fig. S8; Table S7). When simulations were performed on three artificial models (Bottleneck VII–IX) with various demographic parameters, a population size gap was still detected (Fig. S9; Table S8). These results suggest a hidden effect of the super bottleneck on non-African populations.

The population size gap was found in both 1000GP and HGDP-CEPH data sets (Fig. 3a,d). The average population sizes of non-African populations were determined to be 20,260 and 20,030, respectively, while those of African agriculturalist populations were 27,080 and 27,440, respectively in these two data sets. The observed population size gap was 7,020, probably due to the hidden effect of the super bottleneck on non-African populations.

The reasons were then investigated why the super bottleneck had different effects on African and non-African populations. Results showed that non-African populations had the out-of-African bottleneck, but African populations lacked such bottleneck. Therefore, the standard coalescent time of non-African populations was larger than that of African populations (Fig. 3c,f). As African populations had more coalescent events occurred during the bottleneck period, the bottleneck was more readily detected. The mathematical proof on this issue was described in the supplemental material.

Discussion

In this study, we develop FitCoal, a novel model-flexible tool for demographic inference. Key characteristic features of FitCoal are that the composite likelihood can be rapidly calculated based on expected branch lengths and that inference time intervals are variable during the demographic inference. Since coalescent events become rare when tracing backward in time, the length of time interval is usually set to increase progressively[14-17]. Although this strategy can capture recent demographic events, it may miss ancient ones. As FitCoal uses variable time intervals, it can give an accurate inference for both recent and ancient demographic events.

The most important discovery with FitCoal in this study is that human ancestors had a super bottleneck. The ancient population size reduction around 930 kyr ago was likely to be due to the formation of geographically isolated populations driven by culture or ecological factors. In addition, we found that our ancestors had a very small effective size of approximately 1,280 breeding individuals during the bottleneck period. This number is comparable in the same magnitude in the effective size of mammals threatened by extinction[43]. We also detected an instantaneous recovery in all ten African populations with a 20-fold population growth during a short time period around 813 kyr ago (Fig. 3, S10). The earliest archaeological evidence for human control of fire was found in Israel 790 kyr ago[44,45]. As the control of fire profoundly affected social evolution[46] and brain size[47], it may be associated with the big bang in population size at the end of the super bottleneck.

The super bottleneck, which started about one million years ago, might be associated with a speciation event in the early human evolution[27,28]. The questions about where the small ancient population dwelt, how they survived for such a long time, and how they diverged from other hominin groups remain to be investigated. As Neandertals and Denisovans diverged with the modern human between 270 and 440 kyr ago[48,49], it is conceivable that they also had the super bottleneck. In the future, a more detailed picture of human evolution in the Early Stone Age may be revealed when more genomic sequences of African populations and archaic hominins and more advanced population genomics methods become available.

Methods

Standard coalescent time and time in generations. The population size is denoted $N(\cdot)$, representing the demographic history. Time τ represents one-point scaled time since the time in a generation is scaled by $2N(0)$. Time t is usually scaled by $2N(t)$ generations[20,34,50,51]. To distinguish it from the one-point scaled time τ, time t is designated as the standard coalescent time.

Fast infinitesimal time coalescent (FitCoal) process. The FitCoal calculates the expected branch length for each type of site frequency spectrum (SFS) under arbitrary demographic history $N(\cdot)$. We assume that a sample is obtained by randomly taken n

sequences from the population. The sample is designated to be state l $(l = 2, \cdots, n)$ at time t if it has exactly l ancestral lineages at this time. The probability of state l at time t is denoted $p_l(t)$. In a coalescent tree, a branch is designated to be type i if it has exactly i descendants. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}p_l(t) = \begin{cases} \binom{l+1}{2}p_{l+1}(t) - \binom{l}{2}p_l(t) & l = 2, \cdots, n-1 \\ -\binom{l}{2}p_l(t) & l = n \end{cases}.$$

When Δt is extremely small (Fig. 1), there is at most one coalescent event during t and $t + \Delta t$, leading to

$$p_l(t + \Delta t) = \begin{cases} \binom{l+1}{2}\Delta t p_{l+1}(t) + \left(1 - \binom{l}{2}\Delta t\right)p_l(t) & l = 2, \cdots, n-1 \\ \left(1 - \binom{l}{2}\Delta t\right)p_l(t) & l = n \end{cases}.$$

The branch length is in units of generations. The expected branch length of state l during t and $t + \Delta t$ is calculated as $\int_t^{t+\Delta t} 2\,N(t)p_l(t)l\mathrm{d}t$. The probability that a branch of state l is of type i is $\frac{\binom{n-i-1}{l-2}}{\binom{n-1}{l-1}}$ (ref.[20]). The expected branch length of type i of state l during t and $t + \Delta t$ is $\int_t^{t+\Delta t} 2\,N(t)p_l(t)l\frac{\binom{n-i-1}{l-2}}{\binom{n-1}{l-1}}\mathrm{d}t$. Therefore, the expected branch length $BL_i(N(\cdot))$ of type i is

$$\sum_{l=2}^{n-i+1}\int_0^{\infty} 2\,N(t)p_l(t)l\mathrm{d}t\frac{\binom{n-i-1}{l-2}}{\binom{n-1}{l-1}}.$$

A FitCoal time partition is denoted by $\{t_0, t_1, \cdots, t_m\}$, where $0 = t_0 < t_1 < \cdots < t_m$. We have $p_l(t_0) = \begin{cases} 1 & l = n \\ 0 & \text{else} \end{cases}$. For a large positive number m, if t_m is large and $(t_k - t_{k-1})$ is small for $k = 1, \cdots, m$, then

$$p_l(t_k) = \begin{cases} \left(1 - \binom{l}{2}(t_k - t_{k-1})\right)p_l(t_{k-1}) & l = n \\ \left(1 - \binom{l}{2}(t_k - t_{k-1})\right)p_l(t_{k-1}) + \binom{l+1}{2}(t_k - t_{k-1})p_{l+1}(t_{k-1}) & \text{else} \end{cases},$$

where $k = 1, \cdots, m$.

The expected branch length of type i is calculated as

$$BL_i(N(\cdot)) = \sum_{l=2}^{n-i+1} l\frac{\binom{n-i-1}{l-2}}{\binom{n-1}{l-1}}\left(\sum_{k=1}^{m} 2N(t_{k-1})p_l\,(t_{k-1})(t_k - t_{k-1})\right).$$

To determine the time partition, we required that the coalescent probability was less than 10^{-4} during t_{k-1} and t_k $(k = 1, \cdots, m)$, the probability of common ancestor (*i.e.*, the probability of state 1) at t_m was larger than $(1 - 10^{-6})$. When the sample size was 10, the number of infinitesimal time intervals was 1,571,200. When the sample size was 200, the number of infinitesimal time intervals was 7,038,398.

Thus, each Δt was extremely small for precise calculation of expected branch length, and the time was partitioned to obtain $p_l(t)$ in order to calculate the expected branch length of type i.

Tabulated FitCoal. The expected branch length of each type can be calculated for arbitrary time intervals according to the procedure described above. Considering another tabulated time partition $\{t_0, t_1, \cdots, t_m\}$ $(0 = t_0 < t_1 < \cdots < t_m)$, the expected branch length of a type is equal to the sum of the expected branch length of this type during each tabulated time interval, thus the latter can be rescaled and tabulated.

The scaled expected branch length $BL_{i,t}$ of type i during 0 and t is $BL_{i,t} = \sum_{l=2}^{n-i+1} \int_0^t p_s\,(l) l \frac{\binom{n-i-1}{l-2}}{\binom{n-1}{l-1}} \mathrm{d}s$, where $i = 1, \cdots, n-1$. For the tabulated time partition $\{t_0, t_1, \cdots, t_m\}$, BL_{i,t_0}, BL_{i,t_1}, $\cdots$, and BL_{i,t_m} are tabulated. When $n = 10$, $m = 231$. When $n = 200$, $m = 529$.

$BL_{i,t}$ is used to calculate the expected branch lengths under arbitrary demographic histories. When $\tilde{t} \in [t_{k-1}, t_k)$,

$$BL_{i,\tilde{t}} \approx \frac{t_k - \tilde{t}}{t_k - t_{k-1}} BL_{i,t_{k-1}} + \frac{\tilde{t} - t_{k-1}}{t_k - t_{k-1}} BL_{i,t_k}.$$

If $N(t)$ is a piecewise constant, that is, there exist a demographic time partition $\{\tilde{t}_0, \tilde{t}_1, \cdots, \tilde{t}_{\tilde{m}}\}$, $N(t) = N_k$ for $t \in [\tilde{t}_k, \tilde{t}_{k+1})$, $k = 0, \cdots, \tilde{m}$. Then, the expected branch length of type i is calculated as

$$BL_i(N(\cdot)) = \sum_{k=1}^{\tilde{m}} 2N_k \left(BL_{i,\tilde{t}_k} - BL_{i,\tilde{t}_{k-1}}\right).$$

When $N(t)$ is complex, the population size can be approximated by a piecewise constant function.

Composite likelihood. The mutation rate per base pair per generation is denoted μ, and $\vec{\xi} = (\xi_i)$ is the observed number of SNPs of n sequences with σ base pairs, where $i = 1, \cdots, n-1$. The expected SFS is $\vec{\lambda} = (\lambda_i)$, where $\lambda_i = \mu\sigma BL_i(N(\cdot))$. Following the Poisson probability and previous studies[8], the composite likelihood is calculated as follows:

$$L_{\mu,\iota}(\vec{\xi}, N(\cdot)) = \prod_{i=1}^{n-1} \frac{\lambda_i^{\xi_i} e^{-\lambda_i}}{\xi_i!}.$$

The likelihood is extended to missing data and truncated SFS (see Supplemental materials).

Demographic inference. The number of demographic time intervals is variable. FitCoal first fits the observed SFS using a constant size model with one demographic time interval, and the number of time intervals is increased by one at a time to generate more complex models. The Local Unimodal Sampling (LUS) algorithm[52] is used to maximize the likelihood and estimate demographic parameters. A log-likelihood promotion rate is used to determine the best model to explain the observed SFS, and 20% is used as the threshold.

A series of demography with m pieces is denoted by a set $S(m)$, where $S(m)$ contains all of the following m pieces of population size:

$$N(t|N0 > 0, N_{(m)}, t_{(m)}, c_{(m)}) = \begin{cases} N_m N0 & t \geq t_m \\ N_k N0 & t_k \leq t < t_{k+1}, c_k \in \mathcal{C}, k = 1, \cdots, m-1, \\ \frac{(t_{k+1}-t_k)N_{k+1}N_k N0}{(t-t_k)N_k+(t_{k+1}-t)N_{k+1}} & t_k \leq t < t_{k+1}, c_k \in \mathcal{E}, k = 1, \cdots, m-1 \end{cases}$$

where $N_{(m)} = (N_1, \cdots, N_m) \in N[m]$, $t_{(m)} = (t_1, \cdots, t_m) \in t[m]$, $c_{(m)} = (c_1, \cdots, c_m) \in c[m]$, $N[m] = \{(N_1, \cdots, N_m) | N_1 = 1, N_i > 0 \text{ for } i > 1\}$, $t[m] = \{(t_1, \cdots, t_m) | 0 = t_1 <\cdots< t_m\}$, $c[m] = \{(c_1, \cdots, c_m) | c_m \in \mathcal{C}, c_i \in \mathcal{C} \cup \mathcal{E} \text{ for } = 1, \cdots, m-1\}$, $\mathcal{C} = \{\text{constant}\}$, and $\mathcal{E} = \{\text{exponential}\}$.

The set $S(m)$ was used as the wide-range parameter space to determine the maximum likelihood. To find the best demographic history to explain the observed SFS, the following procedures were used:

(1) The number of inference time intervals (or pieces) m is initially set to 1, and the maximum likelihood $\max L_1$ is determined with the constant size model (model in $S(1)$).

(2) Increase m by 1. For each change of type $c_{(m)}$, parameters $N_{(m)} = (N_1, \cdots, N_m)$ and $t_{(m)} = (t_1 = 0, t_2, \cdots, t_m)$ are searched to maximize the likelihood by LUS algorithm to fit the observed SFS. The maximum likelihood $\max L_m$ is calculated with models in $S(m)$ with all possible change types.

(3) Repeat step (2) until $(1 + \text{threshold}) \cdot \log(\max L_m) < \log(\max L_{m-1})$ is obtained. The best model corresponding $\max L_{m-1}$ is determined to explain the observed SFS.

(4) To avoid local optima, steps (1) – (3) are repeated K times to find the best model. $K = 10$ when analyzing simulated samples, and $K = 200$ when analyzing the

observed SFSs of the 1000GP and HGDP-CEPH populations.

To determine the threshold of log-likelihood promotion rate, a large number of simulations were performed (Table S9). For each model, 200 replicates were conducted, and the number of inference time intervals in the estimated demographic history was determined for each replicate. If the estimated number of inference time intervals was larger than the true number of inference time intervals, overfitting was recorded. When the former was smaller than the latter, underfitting was considered. The thresholds of 10%, 20%, and 30% were used. When 10% was used, the maximum overfitting rate was 2%. When 20% was used, all cases examined were inferred correctly. When 30% was used, the underfitting was observed in one of 20 examined models. Therefore, 20% was used as the threshold of log-likelihood promotion rate in subsequent analyses.

Data simulation. Data were simulated using ms[53] and MaCS[54] software. Unless otherwise specified, a generation time was assumed to be 24 years[15,30], the mutation rate μ was set for 1.2×10^{-8} per base per generation[15,31-33], and the recombination rate was $r = 0.8\mu$. For each model, 200 SFSs were simulated to calculate the median and 2.5 and 97.5 percentiles. When verifying the inferred demographic histories, 80,000 DNA fragments with the length of 10kb each were used for simulation, taking into the consideration of small fragments split by sequencing mask in 1000GP and HGDP-CEPH data sets. High frequency alleles of SFS (10% mutation types for Bottleneck I, II, III, VII, VIII, IX, and 15% for Bottleneck IV, V, VI) were removed when assessing models to verify the super bottleneck. Detailed simulation command lines and demographic inference are presented in the Supplementary information.

1000 Genomes Project data. Sequences of autosomal SNPs in 1000GP phase 3 (ref.[1]) were downloaded from the 1000GP ftp server (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/), and 26 populations were analyzed, including seven African populations (ACB, ASW, ESN, GWD, LWK, MSL, and YRI), five European populations (CEU, FIN, GBR, IBS, and TSI), five East Asian populations (CDX, CHB, CHS, JPT, and KHV), five South Asian populations (BEB, GIH, ITU, PJL, and STU), and four American populations (CLM, MXL, PEL, and PUR). The 1000 GP strict mask was used to exclude artifacts of SNP calling. Noncoding regions except pseudogenes, defined by GENCODE release 35

(ref.[35]), were examined to avoid potential effects of purifying selection. Sites without high-confidence ancestral allele inference, according to 1000GP annotations, were excluded. The number of bi-allelic sites that passed the filtering was 826,649,529. To avoid the effect of positive selection, high frequency mutations were excluded, and the truncated SFS was used to infer demographic history (Fig. S11; Table S10). The average proportion of excluded high-frequency SNPs for all 1000GP populations was 4.40%.

HGDP-CEPH data. In total, 24 populations were analyzed, including three African populations (Biaka, Mandeka, and Yoruba), five European populations (Adygei, Basque, French, Russian, and Sardinain), four Middle East populations (Bedouin, Druze, Mozabite, and Palestinian), three East Asian populations (Han, Japanese, and Yakut), eight Central and South Asian populations (Balochi, Brahui, Burusho, Hazara, Kalash, Makrani, Pathan, and Sindhi), and an American population (Maya). Only bi-allelic SNPs locating in GENCODE non-coding regions[35] except pseudogenes that passed HGDP-CEPH filtering were used. HGDP-CEPH accessible mask was also used to filter SNPs[2]. The number of sites that passed filtering was 791,999,125. Missing data were allowed to avoid artifacts due to imputation. The proportion of sites with two or more missing individuals was less than 3% for all populations (Table S11). Each population had two SFSs, with one calculated from sites with no missing data, and another from sites with one missing individual. Similarly, truncated SFSs were used to avoid the effect of positive selection (Fig. S12, S13; Table S12). The average proportion of excluded high-frequency SNPs for all HGDP-CEPH populations was 7.18%.

Data availability.

The authors declare that all data supporting the findings of this study are included in this paper and its supplementary information file.

Code availability.

FitCoal is a free plug-in of the eGPS software[55] and can be downloaded and run as an independent package. FitCoal and its documentation are available via Zenodo at https://zenodo.org/record/4765447#.YKDt7aG-vuq, our institute website at http://www.picb.ac.cn/evolgen/, and eGPS website http://www.egps-software.net/.

Acknowledgement

We thank Daniel Zivković for sharing his codes to calculate the expected branch length, and Xiaoming Liu for sharing his simulated results. This work was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040800), the National Natural Science Foundation of China (nos. 31100273, 31172073, 91131010), and National Key Research and Development Project (No. 2020YFC0847000).

Author contributions: W.H., Z.H., Y.H.P., and H.L. conceived and designed the research; W.H., Z.H., and H.L. wrote the code; W.H., Z.H., P.D., and Y.H.P. analyzed the data; W.H., Z.H., P.D., Y.H.P., and H.L. wrote the paper.

Competing interests: All authors declare that they have no competing interests.

References

1 Altshuler, D. M. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

2 Bergstrom, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).

3 Hu, J. Y. *et al.* Genomic consequences of population decline in critically endangered pangolins and their demographic histories. *Natl. Sci. Rev.* **7**, 798-814 (2020).

4 Zeng, L. *et al.* Out of southern East Asia of the brown rat revealed by large-scale genome sequencing. *Mol. Biol. Evol.* **35**, 149-158 (2018).

5 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449 (2014).

6 Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).

7 Jinam, T. A. *et al.* Evolutionary history of continental southeast Asians: "early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. *Mol. Biol. Evol.* **29**, 3513-3527 (2012).

8 Li, H. & Stephan, W. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet.* **2**, e166 (2006).

9 Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).

10 Griffiths, R. C. & Tavaré, S. Monte Carlo inference methods in population

genetics. *Math. Comput. Modell.* **23**, 141-158 (1996).

11 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).

12 Schaffner, S. F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576-1583 (2005).

13 Heled, J. & Drummond, A. J. Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* **8**, 289 (2008).

14 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).

15 Liu, X. M. & Fu, Y. X. Exploring population size changes using SNP frequency spectra. *Nat. Genet.* **47**, 555-559 (2015).

16 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919-925 (2014).

17 Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303-309 (2017).

18 Liu, X. & Fu, Y. X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome biology* **21**, 280 (2020).

19 Jouganous, J., Long, W., Ragsdale, A. P. & Gravel, S. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics* **206**, 1549-1567 (2017).

20 Fu, Y. X. Statistical properties of segregating sites. *Theor. Popul. Biol.* **48**, 172-197 (1995).

21 Zivković, D. & Wiehe, T. Second-order moments of segregating sites under variable population size. *Genetics* **180**, 341-357 (2008).

22 White, T. D. *et al.* Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature* **423**, 742-747 (2003).

23 Stoneking, M. & Krause, J. Learning about human population history from ancient and modern genomes. *Nat. Rev. Genet.* **12**, 603-614 (2011).

24 Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302-310 (2017).

25 Manica, A., Amos, W., Balloux, F. & Hanihara, T. The effect of ancient population bottlenecks on human phenotypic variation. *Nature* **448**, 346-348 (2007).

26 Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* **102**, 15942-15947 (2005).

27 Stringer, C. The origin and evolution of Homo sapiens. *Philos Trans R Soc Lond B Biol Sci* **371**, 20150237 (2016).

28 Coyne, J. A. & Allen Orr, H. *Speciation*. (Sinauer Associates, Inc., 2004).

29 Hudson, R. R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805-1817 (2001).

30 Scally, A. & Durbin, R. Revising the human mutation rate: implications for

understanding human evolution. *Nat. Rev. Genet.* **13**, 745-753 (2012).

31 Campbell, C. D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, 1277-1281 (2012).

32 Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712-714 (2011).

33 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).

34 Bhaskar, A. & Song, Y. S. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann. Stat.* **42**, 2469-2493 (2014).

35 Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766-D773 (2019).

36 Fay, J. C. & Wu, C.-I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405-1413 (2000).

37 Hsieh, P. *et al.* Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome Res.* **26**, 291-300 (2016).

38 Schlebusch, C. M. & Jakobsson, M. Tales of human migration, admixture, and selection in Africa. *Annu. Rev. Genomics Hum. Genet.* **19**, 405-428 (2018).

39 Skoglund, P. *et al.* Reconstructing prehistoric African population structure. *Cell* **171**, 59-71 (2017).

40 Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* **102**, 15942-15947 (2005).

41 Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* **15**, R159-160 (2005).

42 Ayub, Q. *et al.* The Kalash genetic isolate: ancient divergence, drift, and selection. *Am. J. Hum. Genet.* **96**, 775-783 (2015).

43 Li, H. *et al.* Large numbers of vertebrates began rapid population decline in the late 19th century. *Proc. Natl. Acad. Sci. USA* **113**, 14079-14084 (2016).

44 Goren-Inbar, N. *et al.* Evidence of hominin control of fire at Gesher Benot Ya'aqov, Israel. *Science* **304**, 725-727 (2004).

45 Alperson-Afil, N. *et al.* Spatial organization of hominin activities at Gesher Benot Ya'aqov, Israel. *Science* **326**, 1677-1680 (2009).

46 Foley, R. & Gamble, C. The ecology of social transitions in human evolution. *Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci.* **364**, 3267-3279 (2009).

47 Du, A. *et al.* Pattern and process in hominin brain size evolution are scale-dependent. *Proc. Biol. Sci.* **285**, 20172738 (2018).

48 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722 (2010).

49 Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010).

50 Myers, S., Fefferman, C. & Patterson, N. Can one learn history from the allelic spectrum? *Theor. Popul. Biol.* **73**, 342-348 (2008).

51 Chen, H. A computational approach for modeling the allele frequency spectrum of populations with arbitrarily varying size. *Genomics Proteomics Bioinf.* **17**, 635-644 (2019).

52 Pedersen, M. E. H. Tuning and simplifying heuristical optimization. *PhD thesis, Univ. Southampton* (2010).

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

54 Chen, G. K., Marjoram, P. & Wall, J. D. Fast and flexible simulation of DNA sequence data. *Genome Res.* **19**, 136-142 (2009).

55 Yu, D. *et al.* eGPS 1.0: comprehensive software for multi-omic and evolutionary analyses. *Natl. Sci. Rev.* **6**, 867-869 (2019).

56 Harpending, H. C. *et al.* Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. USA* **95**, 1961-1967 (1998).

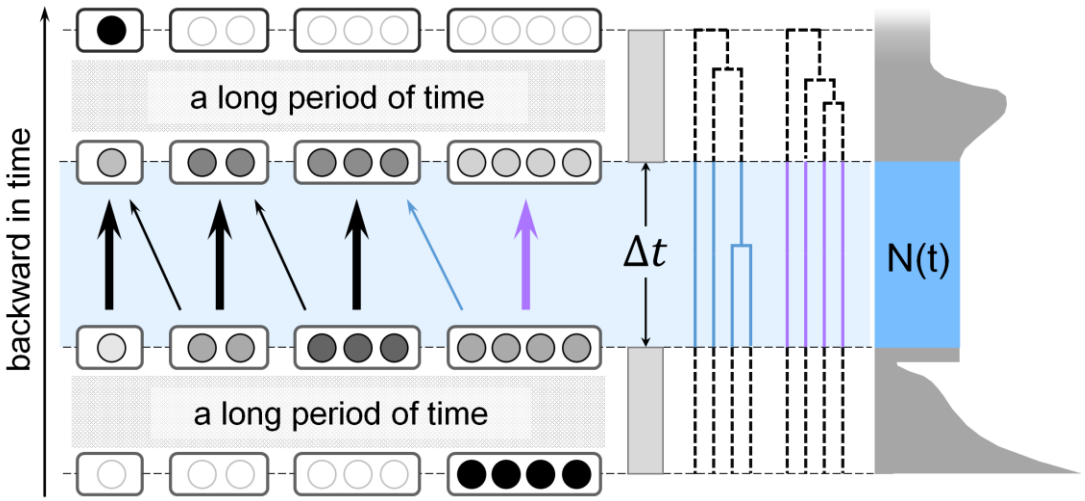

Figure 1. Illustration of the fast infinitesimal time coalescent (FitCoal) process. The left panel shows the backward process in which four lineages coalesce into one after passing through millions of infinitesimal time intervals. The highlighted area shows the backward transformation process of different states with tiny probability changes in an infinitesimal time interval (Δt). Thick arrows indicate high transformation probabilities, and thin arrows indicate low transformation probabilities. Each state is indicated with a rounded rectangle, in which one circle indicates one lineage. The rounded rectangles with black filled circles are the states with probability 1. The rounded rectangles with empty circles are the states with probability 0. The probabilities between 0 and 1 are indicated by grey circles. The middle panel shows branches of different states. The right panel shows the demographic history of a population. The width of shadowed area indicate the effective population size, *i.e.*, the number of breeding individuals[56]. It is assumed that the effective population size remains unchanged within Δt.

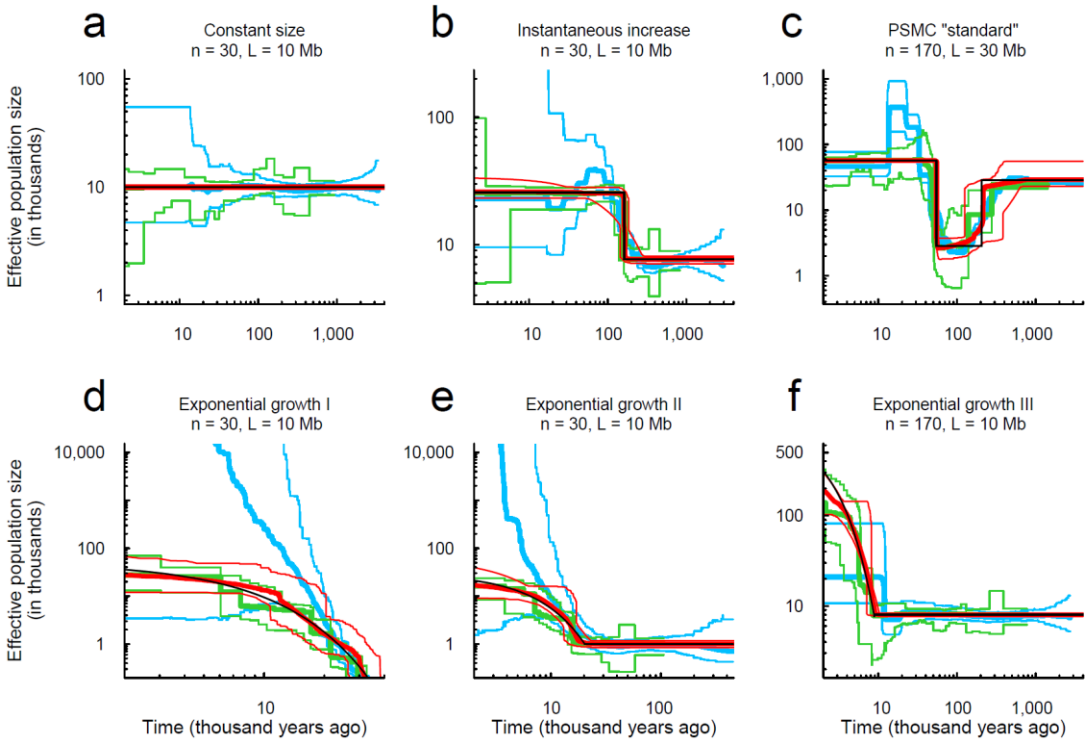

Figure 2. Demographic histories estimated by FitCoal, stairway plot, and PSMC using simulated samples. (a) Constant size model. (b) Instantaneous increase model. (c) PSMC "standard" model. (d) Exponential growth I model. (e) Exponential growth II model. (f) Exponential growth III model. These six models are the same as those of the previous study by Liu and Fu[15]. Thin black lines indicate true models. Thick red lines indicate the medians of FitCoal estimated histories; thin red lines are 2.5 and 97.5 percentiles of FitCoal estimated histories. Green and blue lines indicate the results of stairway plot and PSMC, respectively, of the previous study[15]. The mutation rate is assumed to be 1.2×10^{-8} per base per generation, and a generation time is assumed to be 24 years. *n* is the number of simulated sequences, and *L* is the length of simulated sequences.

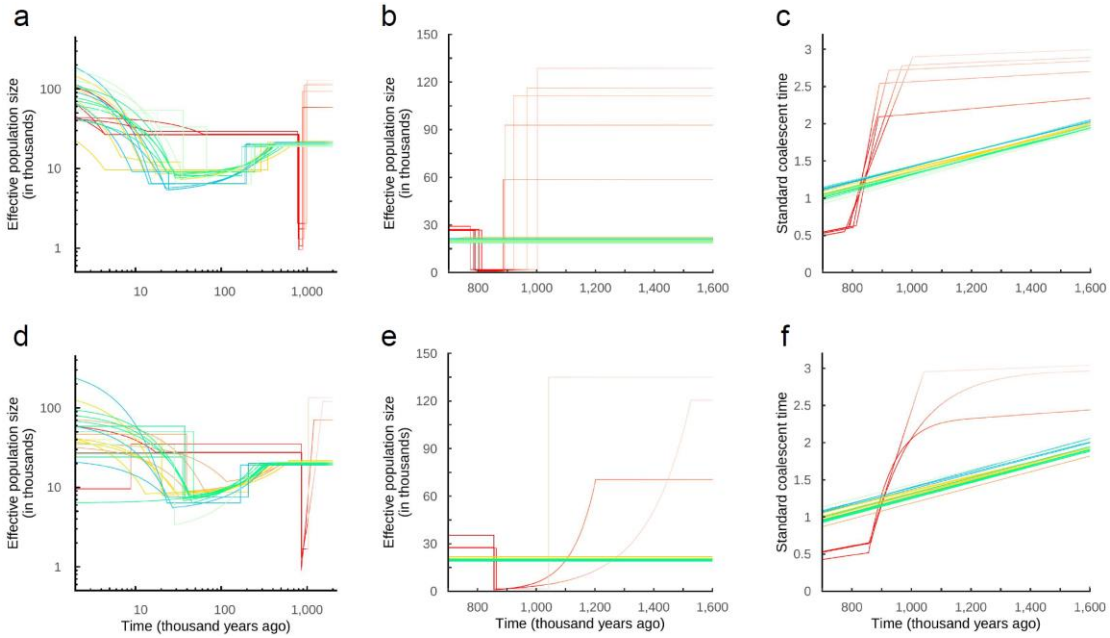

Figure 3. FitCoal estimated histories of human populations using 1000GP and HGPD-CEPH genomic data sets. (a) Estimated histories of 26 populations in 1000GP. (b) Linear-scaled estimation of histories of 1000GP populations during the super bottleneck period. (c) Calendar time *vs* standard coalescent time of estimated histories of 1000GP populations. (d) Estimated histories of 24 HGPD-CEPH populations. (e) Linear-scaled estimation of histories of HGPD-CEPH populations during the super bottleneck period. (f) Calendar time *vs* standard coalescent time of estimated histories of HGPD-CEPH populations. Various color lines indicate the following: red, African populations; yellow, European populations; brown, Middle East populations; blue, East Asian populations; green, Central or South Asian populations; and dark sea green, American populations. The mutation rate is assumed to be 1.2×10^{-8} per base per generation, and a generation time is assumed to be 24 years.

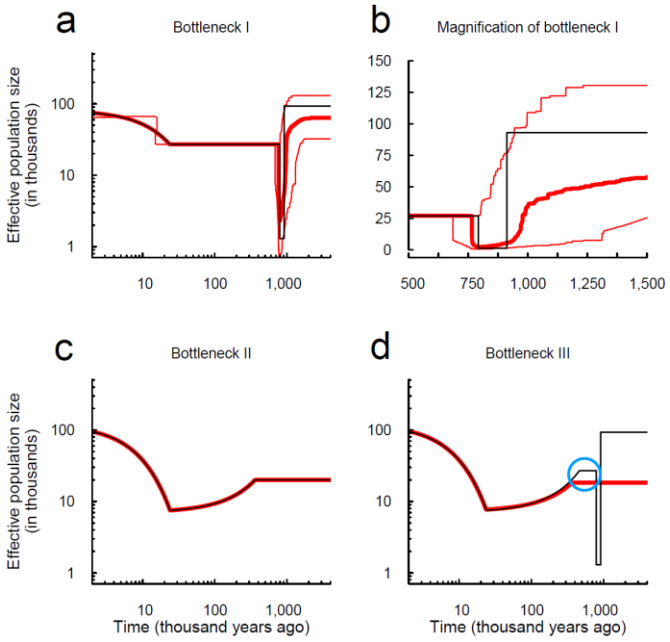

Figure 4. Verification of the super bottleneck. (a) Bottleneck I model, mimicking the demography of 1000GP African population and its estimated histories. (b) Linear-scaled Bottleneck I model during the super bottleneck period. (c) Bottleneck II model, mimicking the estimated demography of 1000GP non-African population and its estimated histories. (d) Bottleneck III model, mimicking the true demography of 1000GP non-African population and its estimated histories. Thin black lines indicate models. Thick red lines denote the medians of FitCoal estimated histories; thin red lines represent 2.5 and 97.5 percentiles of FitCoal estimated histories. Blue circle indicates the population size gap. The mutation rate is assumed to be 1.2×10^{-8} per base per generation, and a generation time is assumed to be 24 years. The number of simulated sequences is 202 in Bottleneck I and 200 in Bottleneck II and III. The length of simulated sequence is 800 Mb.