nature genetics

Inferring human population size and separation history from multiple genome sequences

Stephan Schiffels & Richard Durbin

The availability of complete human genome sequences from populations across the world has given rise to new population genetic inference methods that explicitly model ancestral relationships under recombination and mutation. So far, application of these methods to evolutionary history more recent than 20,000–30,000 years ago and to population separations has been limited. Here we present a new method that overcomes these shortcomings. The multiple sequentially Markovian coalescent (MSMC) analyzes the observed pattern of mutations in multiple individuals, focusing on the first coalescence between any two individuals. Results from applying MSMC to genome sequences from nine populations across the world suggest that the genetic separation of non-African ancestors from African Yoruban ancestors started long before 50,000 years ago and give information about human population history as recent as 2,000 years ago, including the bottleneck in the peopling of the Americas and separations within Africa, East Asia and Europe.

Human genome sequences are related to each other through common ancestors. Estimates of when these ancestors lived provide insight into ancestral population sizes and ancestral genetic separations as a function of time. In the case of non-recombining loci, such as the maternally inherited mitochondrial DNA or the paternally inherited Y chromosome, the time to common ancestors can be estimated from the total number of differences between sequences[1–4]. For autosomal sequences, which account for the vast majority of heritable sequence, the reconstruction of genealogical relationships is more complicated owing to ancestral recombination events that separate many different genealogical trees in different locations of the genome. Although inferring this pattern is more challenging, it provides, in principle, much more information about our past than non-recombining loci, as only a few samples yield many effectively independent genealogies, allowing the inference of a distribution of times to common ancestors with high resolution.

Reconstruction of the full underlying ancestral recombination graph is challenging because the space of possible graphs is extremely large. A substantial simplification was proposed by McVean and Cardin, who model the generating process of genealogical trees as Markovian along the sequence[5,6]. An application of this idea was presented by Li and Durbin in the pairwise sequential Markovian coalescent (PSMC) model[7], a method that focuses on modeling the two genome sequences in one diploid individual. Because only two sequences are modeled, the coalescent event joining the sequences at the most recent common ancestor almost always occurred more than 20,000 years ago, meaning that PSMC can only infer population size estimates beyond 20,000 years ago. Also, with only two sequences, there is only limited scope for the analysis of population separations.

For more than two sequences, extending PSMC in the natural way by enumerating all possible trees with their branch lengths along the sequences would be very computationally costly, even under the Markovian model. A recent simplification was suggested by Song and colleagues[8–10], which is based on approximating the conditional sampling process for adding an $(n+1)$th sequence to the distribution of genealogies connecting n sequences. Here we propose an alternative approach that we call multiple sequential Markovian coalescent (MSMC), which overcomes the increase in complexity by introducing a different simplification. We characterize the relationship at a given location between multiple samples through a much reduced set of variables: (i) the time to the most recent common ancestor of any two sequences, that is, the first coalescence, along with the identities of the two sequences participating in the first coalescence (**Fig. 1a**), and (ii) the total length of all singleton branches in the tree, that is, branches that give rise to variants of minor allele count 1 if affected by a mutation. Given a demographic model, we can keep track of the likelihood distribution for these variables on the basis of the observed mutation data as we move along the sequences. As detailed in the Online Methods, we derive approximate transition and emission rates using the sequentially Markovian coalescent (SMC′)[5,6] framework. This approach allows efficient maximum-likelihood estimation of the free parameters, which include the piecewise constant population size as a function of time. If sequences are sampled from different subpopulations, we use additional free parameters for coalescence rates within and across population boundaries, which allow us to infer how subpopulations separated over time. We compare further the conditional sampling approach and our approach to determining first coalescence.

Because MSMC focuses on the first coalescence event for any pair of haplotypes, the inference limits are set by the distribution of first coalescence times (t), the mean of which scales inversely with the square

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. Correspondence should be addressed to S.S. (stephan.schiffels@sanger.ac.uk) or R.D. (rd@sanger.ac.uk).

Received 30 October 2013; accepted 30 May 2014; published 22 June 2014; doi:10.1038/ng.3015

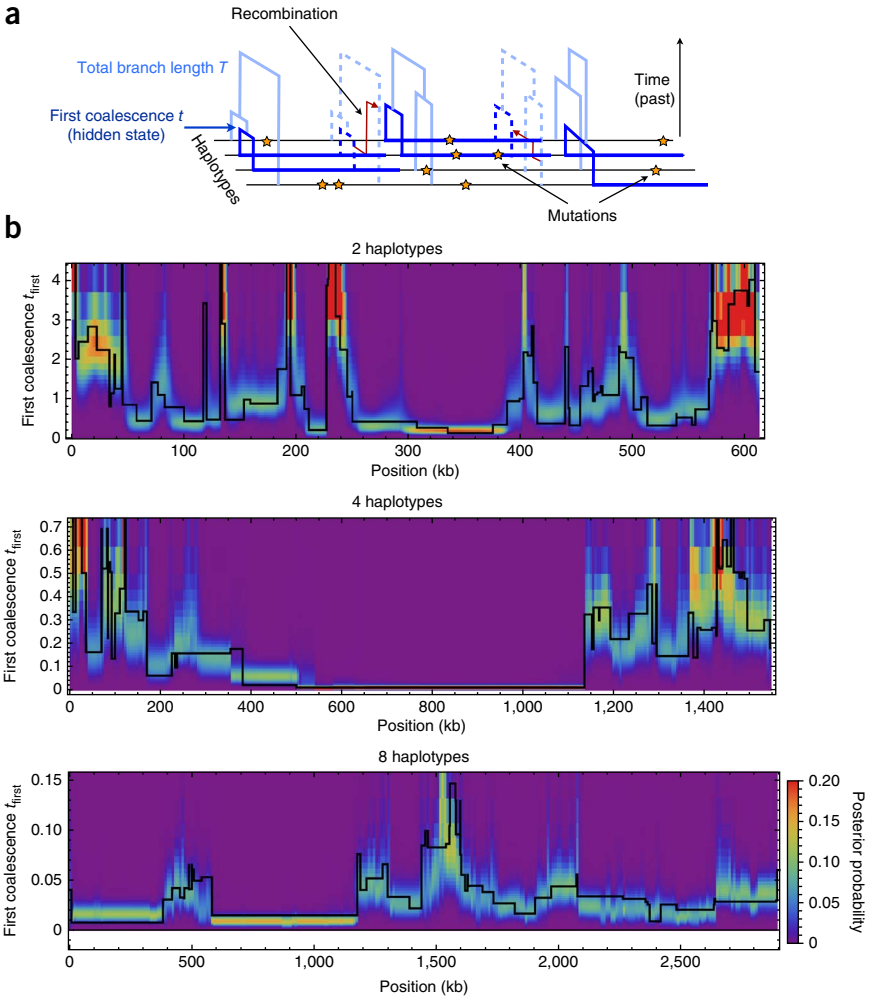

Figure 1 MSMC locally infers branch lengths and coalescence times from observed mutations. (**a**) Schematic of the model. Local genealogies change along the sequences by recombination events that rejoin branches of the tree, according to the SMC′ model[5,6]. The pattern of mutations depends on the genealogy, with few mutations on branches with recent coalescences and more mutations in deeper branches. The hidden states of the model are the time to the first coalescence and the identity of the two sequences participating in the first coalescence. (**b**) MSMC can locally infer its hidden states, shown by the posterior probability with color. In black, we plot the first coalescence time as generated by the simulation. This local inference works well for two, four and eight haplotypes. As more haplotypes are used, the typical time to the first coalescence event decreases, whereas the typical segment length increases.

of the sample size (M), $<t> = 2/(M(M-1))$, in units of $2N_0$ generations (**Fig. 1b** and Online Methods), where N_0 is the long-term average effective population size. Here we demonstrate application of our model on up to 8 haplotypes, which allows us to study changes in population size occurring as recently as 70 generations ago. As a special case of MSMC for two haplotypes, we provide a new implementation of PSMC that we call PSMC′ because it uses the SMC′ model, which accounts for recombination events between segments with the same time to coalescence[6]. PSMC′ accurately estimates the recombination rate (**Supplementary Fig. 1**), which is not the case for PSMC[7].

We apply our method to 34 individuals from 9 populations of European, Asian, African and Native American ancestry, sequenced using Complete Genomics technology[11]. Our results give detailed estimates of population sizes and population separations over time between about 2,000 and 200,000 years ago, including the out-of-Africa dispersal of modern humans, the split between Asian and European populations and the migration into the Americas.

As with other inference methods based on coalescent theory, MSMC can only infer scaled times and population sizes. To convert these estimates into real time and size, all scaled results need to be divided by the mutation rate. Scaled times must be further multiplied by the generation time. Here we will use a generation time of 30 years and a rate of 1.25×10^{-8} mutations per nucleotide per generation, as supported by recent publications[12–16]. As there is current debate about these values[17–19], we consider alternate scalings as well.

RESULTS

Testing the performance of MSMC with simulated data

We implemented the MSMC model as described in the Online Methods, with mathematical derivations given in the **Supplementary Note**. To test the model, we used coalescent simulations with different demographic scenarios (see the **Supplementary Note** for simulation protocols). We first tested whether MSMC could locally decode the time to the first coalescence. Posterior probability over the hidden state for two, four and eight haplotypes is shown in **Figure 1b**. The model recovered the true hidden state with good local resolution. The typical length of segments of constant hidden state increased with sample size (**Fig. 1b**), whereas the typical time to the first coalescence decreased.

We then tried to infer back the simulation parameters from the simulated sequence data, applying MSMC to two different demographic scenarios (**Fig. 2**). First, we simulated a single population under a series of population growths and declines. MSMC recovered the resulting zigzag pattern in population size with good resolution (**Fig. 2a**), where two haplotypes (similar to in PSMC) yielded good estimates between 50,000 and 2 million years ago and eight haplotypes gave estimates as recent as 2,000 years ago, with a small bias toward smaller population sizes in the more distant past. We were able to test a reduced data set with reduced parameters using 16 haplotypes (**Supplementary Fig. 2e**), which suggested that the bias observed with 8 haplotypes in the distant past increased further with more haplotypes. We also tested other simulated histories with sharp changes in population size (**Supplementary Fig. 2a,b**). As expected from the experience with PSMC, very rapid changes in population size were smoothed out over an interval around the true time at which the change occurred.

Next, we simulated two population split scenarios where a single ancestral population split into two equally sized populations 10,000 or 100,000 years ago. For each population split scenario, we inferred effective coalescence rates across the two populations and within populations as a function of time (Online Methods). We provide a

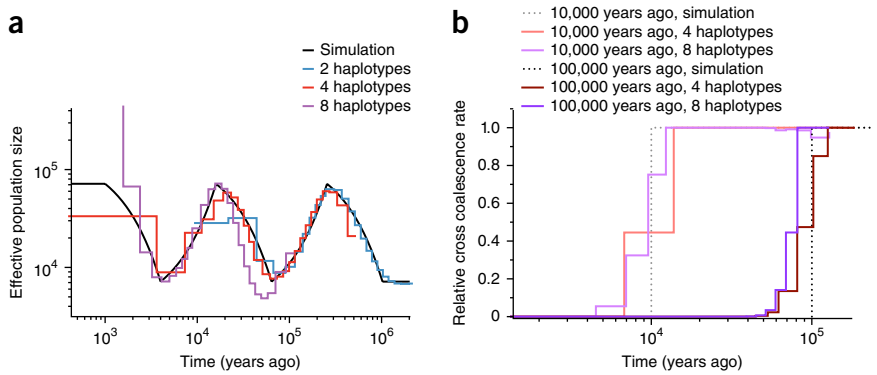

Figure 2 Testing MSMC on simulated data. (**a**) To test the resolution of MSMC applied to two, four and eight haplotypes, we simulated a series of exponential population growths and declines, each changing the population size by a factor of ten. MSMC recovers the resulting zigzag pattern (on a double-logarithmic plot) in different times, depending on the number of haplotypes. With two haplotypes, MSMC infers the population history from 40,000 to 3 million years ago, whereas, with four and eight haplotypes, it infers the population history from 8,000 to 30,000 years ago and from 2,000 to 50,000 years ago, respectively. (**b**) Model estimates from two simulated population splits 10,000 and 100,000 years ago. The dotted lines plot the expected relative cross coalescence rate between the two populations before and after the splits. Maximum-likelihood estimates are shown in red (four haplotypes) and purple (eight haplotypes). As expected, four haplotypes yield good estimates for the older split, whereas eight haplotypes give better estimates for the more recent split.

measure for the genetic separation of populations constituting the ratio between the cross-population and within-population coalescence rates, which we term the 'relative cross coalescence rate'. This parameterization effectively models population separations and substructure in a simpler way than a standard forward-in-time island migration model, which would require a larger state space for structured genealogies, as shown in ref. 10 (see also the **Supplementary Note**). Although not standard, we suggest that our parameter is a more direct measurement of what can be derived about genetic exchange between historical populations from modern samples. The relative cross coalescence rate should be close to 1 when the two populations are well mixed and 0 after they have fully separated. For four and eight haplotypes, our MSMC estimates correctly showed this (**Fig. 2b**), although the instantaneous split time in the simulation was spread out over an interval around the actual split time. As expected, eight haplotypes yielded better estimates for the scenario of a more recent split at 10,000 years ago, whereas four haplotypes yielded better estimates for the older split at 100,000 years ago.

We also tested a population split with subsequent migration (**Supplementary Fig. 2c**), for which we inferred a higher relative cross coalescence rate across the two populations after the split, as would be expected. We further tested the robustness of our method under changes in population size before and after the split (**Supplementary Fig. 2d**), the consequences of the approximation to the singleton branch length (**Supplementary Fig. 3** and **Supplementary Note**) and heterogeneities in recombination rate (Online Methods and **Supplementary Fig. 4**), finding no substantial effect on our estimates.

MSMC requires, in principle, fully phased haplotypes as input, although we could partially allow for unphased data at a subset of sites (see the Online Methods for details). To test the possibilities of extending our unphased approximation to entirely unphased samples, we simulated data sets for two individuals with one or both unphased. Population size estimates based on unphased data were still relatively accurate (**Supplementary Fig. 5**), although biases occurred at the two ends of the analyzed time range. Estimation of relative cross coalescence rate on the basis of partially or fully unphased data was less accurate and more biased in the distant past. Because of this, when applying MSMC to real data, we left unphased sites in the analyses of population size estimates but remove them from the analyses of the population split (**Supplementary Fig. 6**).

Inference of population size from whole-genome sequences

We applied our model to the genomes from one, two and four individuals sampled from each of nine extended HapMap populations[20]: YRI (Nigerian), MKK (Kenyan), LWK (Kenyan), CEU (Northern and Western European), TSI (Italian), GIH (North Indian), CHB (Chinese), JPT (Japanese) and MXL (Mexican admixed with European) (details in **Supplementary Table 1**). We statistically phased all genomes using a reference panel (Online Methods) and tested the impact of potential switch errors by comparing these sequences with family trio–phased sequences that were available for CEU and YRI populations (Online Methods and **Supplementary Fig. 6**).

The results from two individuals (four haplotypes) are shown in **Figure 3a**. In all cases, the inferred population history from four haplotypes matched the estimates from two haplotypes where their inference range overlapped (60,000–200,000 years ago; see thin lines for CEU and YRI in **Fig. 3** and **Supplementary Fig. 7a**). We found that all non-African populations that we analyzed showed a remarkably similar history of population decline from 200,000 years ago until about 50,000 years ago, consistent with a single non-African ancestral population that underwent a bottleneck at the time of the exodus from Africa around 40,000–60,000 years ago[21–23]. The separation of estimates for non-African and African ancestral population sizes began much earlier at 150,000–200,000 years ago, clearly preceding this bottleneck, as already observed using PSMC[7]. We quantify this separation further by directly estimating the relative cross coalescence rate over time. In contrast, we saw only a mild bottleneck in the African population histories, with an extended period of relatively constant population size more recent than 100,000 years ago. Between 30,000 and 10,000 years ago, we saw similar expansions in population size for the CEU, TSI, GIH and CHB populations. For the Mexican ancestors, we saw an extended period of low population size after the out-of-Africa bottleneck, with the lowest value around 15,000 years ago, which was particularly pronounced when we filtered out genomic regions of recent European ancestry due to admixture (dashed line in **Fig. 3**; Online Methods). This extended bottleneck is consistent with estimates of the time that the Native American ancestors crossed the Bering Strait and moved into the Americas[21,24–26]. We repeated all analyses based on four haplotypes on a replicate data set, using sequences for different individuals that were available for all populations except MXL. All results were well reproduced, and differences were only present in the most recent time intervals (**Supplementary Fig. 8**).

Analyzing eight haplotypes from each population except for the MXL and MKK populations (Online Methods), we could see recent changes in population size with higher resolution than with four haplotypes (**Fig. 3b**). Results from eight haplotypes were compatible to those from four haplotypes beyond 10,000 years ago,

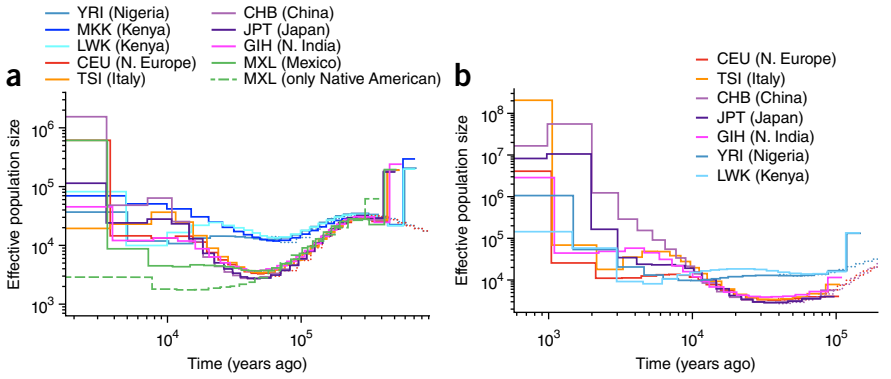

Figure 3 Inference of population size from whole-genome sequences. (**a**) Population size estimates from four haplotypes (two phased individuals) from each of nine populations. The dashed line was generated from a reduced data set of only the Native American components of the MXL genomes. Estimates from two haplotypes for CEU and YRI are shown for comparison as dotted lines. N, Northern. (**b**) Population size estimates from eight haplotypes (four phased individuals) from the same populations as in **a** but excluding MXL and MKK. In contrast to estimates with four haplotypes, estimates are more recent. For comparison, we show the result from four haplotypes for CEU, CHB and YRI as dotted lines.

with systematically slightly lower estimates of population sizes in the range of 10,000–30,000 years ago, as expected from **Figure 2a** (**Supplementary Fig. 7b**). We obtained new insights for periods more recent than 10,000 years ago, during the period of Neolithic expansion. Focusing first on Asian populations, we saw that the ancestors of the CHB population rapidly increased in number from 10,000 years ago to reach effective population sizes of over a million 2,000 years ago. The GIH ancestors also increased in number early, from around 10,000 years ago, with population growth occurring a little more slowly than with the CHB ancestors and numbers leveling off at around 4,000 years ago up until recent times. The JPT ancestral population appeared to have split from the CHB ancestral population by 9,000 years ago and only slowly increased in size up until 3,000 years ago, since which time it also increased in size very rapidly. In Europe, Northern and Western European ancestors (CEU) experienced a relatively constant effective population size between 10,000 and 2,000 years ago, only rapidly increasing in number since 2,000 years ago; Southern European ancestors (TSI) had a consistently higher effective ancestral population size, appearing to show a more complex history of increase and decrease in number between 10,000 and 3,000 years ago and then increasing in number earlier than the CEU ancestors. As discussed previously[7], such a pattern of increase and decrease in population size can result from admixture with previously separated populations, consistent with multiple waves of peopling of Europe with a substantial genetic component from earlier waves in Southern Europe[27]. In Africa, the YRI (Yoruba) ancestral population expanded earliest, around 6,000 years ago, consistent with the introduction of agriculture and the Bantu expansion[28]. This was followed by expansion of the LWK (Luhya) ancestral population, for which, before 6,000 years ago, there was a long 'hump' in ancestral population size extending back beyond 50,000 years ago, again possibly reflecting admixture within the last few thousand years between populations initially separated tens of thousands of years ago.

Divergence from and within African populations

MSMC lets us explicitly study the genetic separation between two populations as a function of time by modeling the relationship of multiple haplotypes, half of which are from one population and half of which are from the other. From analyzing four haplotypes for each pair of populations, we found that all relative cross coalescence rates between any non-African population and the Yoruba were very similar and exhibited a slow, gradual decline beginning earlier than 200,000 years ago and lasting until about 40,000 years ago (**Fig. 4a**). This similarity in rates gives additional information beyond estimates of population size and is consistent with all non-African populations diverging as a single population from the Yoruban ancestors.

To understand whether the gradual decline in relative cross coalescence rate between the YRI and non-African ancestors was due to the inability of MSMC to detect rapid changes (**Fig. 2b**) or due to true ongoing genetic exchange, we compared its results with simulated clean split scenarios at three different time points in the past (**Fig. 4c**). This comparison showed that no clean split could explain the inferred progressive decline in relative cross coalescence rate. In particular, the early beginning of the decline would be consistent with an initial formation of distinct populations before 150,000 years ago, whereas the late end of the decline would be consistent with a final split around 50,000 years ago. This comparison suggests a long period of partial divergence with ongoing genetic exchange between the Yoruban and non-African ancestors that began beyond 150,000 years ago, with population structure within Africa, and lasted for over 100,000 years. The median point of this divergence was around 60,000–80,000 years ago, at which time there was still substantial genetic exchange, with half the coalescences between populations and half within. We also observed that the rate of genetic divergence was not uniform but could be roughly divided into two phases. First, up until about 100,000 years ago, the two populations separated more slowly, whereas after 100,000 years ago, the rate of genetic exchange decreased faster. We note that the fact that the relative cross coalescence rate did not reach 1, even around 200,000 years ago (**Fig. 4c**), might be owing to later admixture of archaic populations such as the Neanderthals into the CEU population after its split from the YRI ancestral population[29].

We also saw extended divergence patterns in eight-haplotype analysis between the ancestors of the three African populations (**Fig. 4a**), with the LWK and YRI ancestral populations being closest and the MKK ancestral population showing a very slow increase in relative cross coalescence rate going back in time with the YRI and LWK ancestral populations. These declines in rate were all more gradual than those shown in **Figure 4b** between out-of-Africa populations, suggesting that the separations of African populations were also not clean splits but gradual separations, perhaps reflecting complex ancestral structure with admixture. In addition, we saw a different separation history between CEU and MKK ancestral populations compared to the LWK and CEU ancestral populations, which in turn was very similar to our estimates of the YRI-CEU separation. Our results suggest that the Maasai ancestors were mixing extensively with non-African ancestors until about 80,000 years ago, much later than the separation of the YRI and non-African populations. This observation is consistent with a model in which the Maasai ancestors and non-African ancestors formed sister groups, which together separated from West African ancestors and continued to extensively mix until much closer to the time of the actual out-of-Africa migration. Nonzero estimates of relative cross coalescence rate between the MKK

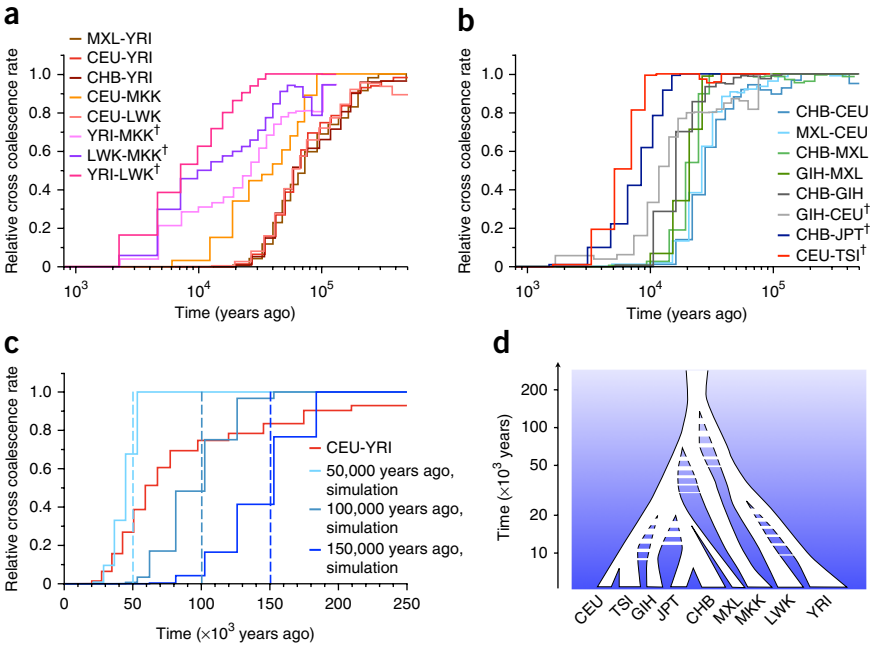

Figure 4 Genetic separation between population pairs. (**a**) Relative cross coalescence rates in and out of Africa. African–non-African pairs are shown in red, and pairs within Africa are shown in purple. (**b**) Relative cross coalescence rates between populations outside Africa. European–East Asian pairs are shown in blue, Asian-MXL pairs are shown in green, and other non-African pairs are shown in other colors, as indicated. The pairs that include MXL are masked to include only the putative Native American components. In **a** and **b**, the most recent population separations are inferred from eight haplotypes, that is, four haplotypes from each population, and corresponding pairs are indicated by a cross. (**c**) Comparison of the African–non-African split with simulations of clean splits. We simulated three scenarios, at split times 50,000, 100,000 and 150,000 years ago. The comparison demonstrates that the history of relative cross coalescence rate between African and non-African ancestors is incompatible with a clean split model and suggests it progressively decreased from beyond 150,000 years ago to approximately 50,000 years ago. (**d**) Schematic of population separations. Timings of splits, population separations, gene flow and bottleneck are shown along a logarithmic axis of time.

and CEU ancestors after 50,000 years ago are probably confounded by more recent admixture from non-African populations back into East African populations, including the Maasai[30,31].

Divergences outside Africa

As expected, the oldest split among out-of-Africa populations was between European and East Asian (CHB and MXL) populations, most of which occurred between 20,000 and 40,000 years ago (**Fig. 4b**). Intriguingly, there might be a small component (10% or less) of this separation extending much further back toward 100,000 years ago, which is not compatible with a single out-of-Africa event around 50,000 years ago. Next oldest was a separation between Asian (CHB and GIH) and American (MXL) populations around 20,000 years ago. This was the most rapid separation we saw, compatible with a clean split. Passing over the GIH separations for now, we found in eight-haplotype analyses separations between the JPT and CHB ancestors around 8,000–9,000 years ago, which is compatible with the divergence in population size history described above, and between the CEU and TSI ancestors around 5,000–6,000 years ago, both also relatively sharp. Four-haplotype analyses of the same separations are shown in **Supplementary Figure 9**.

The pattern of divergence of the North Indian ancestors (GIH) from East Asian and European ancestors was more complex. We observed continued genetic exchange between the GIH ancestors and both of these groups until about 15,000 years ago, suggesting that, even though East Asians and Europeans separated earlier, there was contact between both of these populations and the GIH ancestors after this separation or, equivalently, that there was ancient admixture in the ancestry of North Indians. This deviation from a tree-like separation pattern was independently confirmed by *D* statistics from an ABBA-BABA test[32] (Online Methods and **Supplementary Table 2**), which also indicated that the GIH population is genetically closer to the CEU population than to the CHB and MXL populations. This finding is consistent with the slower and later decline in relative cross coalescence rate between CEU and GIH populations compared to between CEU and CHB populations (**Fig. 4b**). These results suggest that the GIH ancestors remained in close contact with the CEU ancestors until about 10,000 years ago but received some historic admixture component from East Asian populations, part of which is old enough to have occurred before the split with the MXL ancestors.

DISCUSSION

We have presented here both a new method, MSMC, and new insight into the demographic history of human populations as they separated across the globe. We have shown that MSMC can give accurate information about the time dependence of demographic processes within and between populations from a small number of individual genome sequences. As with PSMC, it does this without requiring a simplified model with specific bottlenecks, hard population splits and fixed population sizes as are required by previous methods based on allele frequencies[33–35] or more general summary statistics[36–39]. However, MSMC extends PSMC by an order of magnitude to more recent times and also allows us to explicitly model the history of genetic separations between populations. Because MSMC measures the time to the first coalescence between all pairs of haplotypes, the analyzed time range decreases quadratically with the number of haplotypes. This should be compared with the more naive approach of combining data from PSMC run on different individuals, which would increase information at most linearly, as individuals' histories are not independent.

Although MSMC substantially advances the methodology from PSMC to multiple samples and much more recent times, we have also seen that its practical application appears to be limited to about 8 haploid sequences, both because of the approximations involved and because of computational complexity (see **Supplementary Fig. 2e** for estimates based on 16 haplotypes). It is intriguing, however, to imagine that larger numbers of samples, in principle, contain information about even more recent population history, potentially up until a few generations ago. The basic idea of looking at first coalescence events, as presented here, may lead to new developments that complement

MSMC in this direction, for example, by focusing on rare mutations such as doubletons in large samples[40].

The alternative conditional sampling approach mentioned in the introduction[9,10], implemented in the software diCal, also allows for higher resolution at more recent times. However, when we applied diCal to our zigzag simulation of changes in population size, it did not appear to infer population history more recent than about 20,000 years ago (**Supplementary Fig. 10**). Also, there is currently no way to address the relationship between populations as characterized by MSMC through relative cross coalescence rate. Both of these points may improve in future developments of this method.

Applied to 34 whole-genome sequences from 9 populations, MSMC gives a picture of human demographic history within the last 200,000 years, beginning with the genetic separations of the Yoruban and non-African ancestors and extending well into the Neolithic (**Fig. 4d**). We find strong evidence that the Yoruban–non-African separation took place over a long time period of about 100,000 years, starting long before the known spatial dispersal into Eurasia around 50,000 years ago. Because we model directly an arbitrary history over time of the relative cross coalescence rate between populations, we can see more clearly a progressive separation than earlier analyses based on a single separation time with some subsequent migration[7,17,33,41]. However, the Yoruban population does not represent all of Africa. We now see that the separation of the Maasai population from the out-of-Africa populations occurred within the last 100,000 years. The older part of the separation from the Yoruban ancestors might therefore be a consequence of ancient population structure within Africa, although the direct picture of relationships between African populations is complicated by more recent extensive exchange that we see between all three of the Yoruban, Maasai and Luhya populations within the last 100,000 years. This scenario still does not rule out a possible contribution from an intermediate modern human population that dispersed out of Africa into the Middle East or the Arabian peninsula but continued extensive genetic exchange with its African ancestors until about 50,000 years ago[17,42,43].

Our results are scaled to real times using a mutation rate of 1.25×10^{-8} mutations per nucleotide per generation, as proposed recently[16] and supported by several direct studies of mutation[14–16]. Using a value of 2.5×10^{-8}, as was common previously[44,45], would halve the times. This would bring the midpoint of the out-of-Africa separation to an uncomfortably recent 30,000–40,000 years ago, but, more disconcertingly, it would bring the separation of Native American ancestors (MXL) from East Asian populations to 5,000–10,000 years ago, inconsistent with the paleontological record[25,26]. We suggest that the establishment and spread of the Native American populations might provide a good time point for calibrating population genetic demographic models. We note that the extended period of divergence between African and non-African ancestors that we observe reconciles the timing of the most recent common ancestor of African and non-African mitochondrial DNA around 70,000 years ago[1,18] with the lower autosomal nuclear mutation rate used here, which in simple-split models would suggest a separation around 90,000–130,000 years ago[7,17,33,41,46]. Given that we observe extensive cross coalescence at nuclear loci around 60,000–80,000 years ago, sharing a common ancestor during that time for mitochondrial DNA, which acts as a single locus with reduced effective population size, is entirely likely.

URLs. D Programming language, http://www.dlang.org/; MSMC at GitHub, https://github.com/stschiff/msmc.

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

We thank A. Scally for useful comments and discussion, in particular on interpreting population divergence estimates, and the Durbin group for general discussion. S.S. thanks A. Fischer for helpful support with HMM implementation details. We thank J. Kidd, S. Gravel and C. Bustamante for making ancestry tracts for the MXL individuals available to us. S.S. acknowledges grant support from an EMBO (European Molecular Biology Organization) long-term fellowship. This work was funded by Wellcome Trust grant 098051.

AUTHOR CONTRIBUTIONS

R.D. proposed the basic strategy and designed the overall study. S.S. developed the theory, implemented the algorithm and obtained results. S.S. and R.D. analyzed the results and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Behar, D.M. *et al.* The dawn of human matrilineal diversity. *Am. J. Hum. Genet.* **82**, 1130–1140 (2008).
2. Fu, Q. *et al.* Complete mitochondrial genomes reveal neolithic expansion into Europe. *PLoS ONE* **7**, e32473 (2012).
3. Balaresque, P. *et al.* A predominantly neolithic origin for European paternal lineages. *PLoS Biol.* **8**, e1000285 (2010).
4. Atkinson, Q.D., Gray, R.D. & Drummond, A.J. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol. Biol. Evol.* **25**, 468–474 (2008).
5. McVean, G.A.T. & Cardin, N.J. Approximating the coalescent with recombination. *Phil. Trans. R. Soc. Lond. B* **360**, 1387–1393 (2005).
6. Marjoram, P. & Wall, J.D. Fast "coalescent" simulation. *BMC Genet.* **7**, 16 (2006).
7. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
8. Paul, J.S., Steinrücken, M. & Song, Y.S. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics* **187**, 1115–1128 (2011).
9. Sheehan, S., Harris, K. & Song, Y.S. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* **194**, 647–662 (2013).
10. Steinrücken, M., Paul, J.S. & Song, Y.S. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theor. Popul. Biol.* **87**, 51–61 (2013).
11. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
12. 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).
13. Matsumura, S. & Forster, P. Generation time and effective population size in Polar Eskimos. *Proc. Biol. Sci.* **275**, 1501–1508 (2008).
14. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
15. Campbell, C.D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, 1277–1281 (2012).
16. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
17. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
18. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013).
19. Sun, J.X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161–1165 (2012).
20. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
21. Henn, B.M., Cavalli-Sforza, L.L. & Feldman, M.W. The great human expansion. *Proc. Natl. Acad. Sci. USA* **109**, 17758–17764 (2012).
22. Mellars, P. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* **313**, 796–800 (2006).
23. Mellars, P. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc. Natl. Acad. Sci. USA* **103**, 9381–9386 (2006).

24. Eriksson, A. *et al.* Late Pleistocene climate change and the global expansion of anatomically modern humans. *Proc. Natl. Acad. Sci. USA* **109**, 16089–16094 (2012).
25. O'Rourke, D.H. & Raff, J.A. The human genetic history of the Americas: the final frontier. *Curr. Biol.* **20**, R202–R207 (2010).
26. Goebel, T., Waters, M.R. & O'Rourke, D.H. The late Pleistocene dispersal of modern humans in the Americas. *Science* **319**, 1497–1502 (2008).
27. Botigué, L.R. *et al.* Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc. Natl. Acad. Sci. USA* **110**, 11791–11796 (2013).
28. Berniell-Lee, G. *et al.* Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Mol. Biol. Evol.* **26**, 1581–1589 (2009).
29. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
30. Pickrell, J.K. *et al.* Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. USA* **111**, 2632–2637 (2014).
31. Pagani, L. *et al.* Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am. J. Hum. Genet.* **91**, 83–96 (2012).
32. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
33. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. USA* **108**, 11983–11988 (2011).
34. Marth, G.T. *et al.* The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372 (2004).
35. Keinan, A. *et al.* Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.* **39**, 1251–1255 (2007).
36. Schaffner, S.F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576–1583 (2005).
37. Garrigan, D. *et al.* Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. *Genetics* **177**, 2195–2207 (2007).
38. Plagnol, V. & Wall, J.D. Possible ancestral structure in human populations. *PLoS Genet.* **2**, e105 (2006).
39. Fagundes, N.J. *et al.* Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. USA* **104**, 17614–17619 (2007).
40. Mathieson, I. & McVean, G. Demography and the age of rare variants http://arxiv.org/abs/1401.4181 (2014).
41. Harris, K. & Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.* **9**, e1003521 (2013).
42. Armitage, S.J. *et al.* The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia. *Science* **331**, 453–456 (2011).
43. Petraglia, M. *et al.* Middle Paleolithic assemblages from the Indian subcontinent before and after the Toba super-eruption. *Science* **317**, 114–116 (2007).
44. Takahata, N. & Satta, Y. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc. Natl. Acad. Sci. USA* **94**, 4811–4815 (1997).
45. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
46. Gronau, I. *et al.* Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).

ONLINE METHODS

Sequence data. Whole-genome sequences were generated by Complete Genomics[11] and are freely available on their website ("69 genomes" public data set). Homozygous and heterozygous consensus calls are taken directly from the Complete Genomics tabular format file "masterVar" in the ASM directory of each sample. All regions with a quality tag less than "VQHIGH" were marked as missing data. Furthermore, sites for which more than 17 of the 35 overlapping 35-mers from the reference sequence could be mapped elsewhere with 0 or 1 mismatch were marked as missing (this is similar to the filtering in ref. 7). The anonymous names of all individuals are summarized in **Supplementary Table 1**. The number of called sites and the diversity in each population are summarized in **Supplementary Table 3**. In the MKK samples, we detected cryptic relatedness (**Supplementary Table 4**) by comparing the heterozygosity within and between samples. Specifically, the heterozygosity between haplotype pairs across individuals NA21732 and NA21737 was lower by a factor of one-fourth compared to the heterozygosity within samples. This would be consistent with these two individuals being first cousins. We therefore removed that population from the eight-haplotype analysis. We found that MXL individuals had higher heterozygosity across than within samples, which might be due to different ancestry proportions.

Phasing. All segregating sites that were present in the 1000 Genomes Project integrated variant call set were phased with SHAPEIT2 software[47] using the 1000 Genomes Project reference panel[48]. In addition, we generated full chromosomal haplotype phases from available trio data for YRI and CEU to assess the impact of the accuracy of the population phasing method versus trio phasing. Current switch error rates when phasing against the 1000 Genomes Project (phase 1) reference panel are estimated to be about 300 kb[49], which limits the possibility of observing identity-by-descent (IBD) segments at times older than a few thousand years ago, at the lower limit of our analysis based on eight haplotypes.

Unphased sites affect population size estimates and relative cross coalescence rate estimates differently (**Supplementary Fig. 5**). From a direct comparison with trio-phased data, we found that population size estimates were better if unphased sites were included in the analysis (taken as multiple observations in the hidden Markov model (HMM)), whereas, for relative cross coalescence rate estimates, unphased sites generated severe artifacts in recent time intervals (**Supplementary Fig. 6**). The reason for this is the high sensitivity on shared long haplotypes between individuals, which are not detected if unphased sites are scattered across the sequence. We therefore removed unphased sites for the population split analysis, which in particular improved recent population divergence estimates after 50,000 years ago. To remove unphased sites in an unbiased way, we also removed all blocks of homozygous calls that ended in an unphased heterozygous site, marking them as missing data. This was necessary because otherwise the removal of unphased sites would generate long segments of homozygous calls that would be interpreted as long IBD segments of very recent time to the most recent common ancestor.

Admixture masks for MXL. Admixture masks for NA19735 were downloaded from the 1000 Genomes Project website as part of the phase I analysis results. Admixture tracts for the remaining MXL individuals were generated by Kidd *et al.*[50] and kindly made available to us. For our population split analysis with four haplotypes, we only used the two individuals with the highest Native American component as MXL representative, NA19735 and NA19670. We then filtered out all segments in the genome that were not annotated as homozygous Native American.

***D* statistics (ABBA-BABA test).** We used the following samples: NA20845 (GIH), NA12891 (CEU), NA18526 (CHB), NA19735 (MXL) and NA19238 (YRI). For the *D* statistic, we counted sites at which YRI was homozygous and defined the respective allele as 'ancestral' (A). We then required that the individual specified in the third column carry at least one derived allele (B). We also required that, of the individuals specified in the first and second columns, one be homozygous for allele A and the other carry at least one derived allele (B). This left two configurations per setup named ABBA and BABA, respectively (**Supplementary Table 2**). The *D* statistic is then $D = (n_{ABBA} - n_{BABA})/(n_{ABBA} + n_{BABA})$. We computed *P* values on the basis of a two-tailed binomial test (**Supplementary Table 2**). We performed this test for the two triples CEU-GIH-CHB and CEU-GIH-MXL, using YRI as the outgroup. For the reverse scenarios, treating CEU as the donor population and CHB or MXL as the first population in the test, *D* scores were higher, suggesting that GIH is genetically closer to CEU than to CHB and MXL.

MSMC model. Multiple sequential Markovian coalescent (MSMC) is an HMM along multiple phased haplotypes. Its hidden state is a triplet (t, i, j), where t is the first coalescence time of any two lineages, and i and j are the labels of the two lineages participating in the first coalescence. As detailed in the **Supplementary Note**, we need additional local information about the genealogical tree, namely, the singleton branch length T_S, which sums the lengths of all branches that give rise to variants of minor allele count 1 if mutations occur on these branches. The singleton branch length can be estimated in a separate step before the main inference. We neglect the dependency of the transition rate on the singleton branch length, which is an approximation that is validated by simulations (**Supplementary Fig. 3**).

Transition rate. We use the SMC framework[5,6,51], SMC′, to derive transition rates between hidden states. As shown in the **Supplementary Note** with the help of additional illustrations, the transition probability to state (t, i, j) from state (s, k, l) can be derived in a relatively straightforward way and separated into the probability that no change of the first coalescence time occurs, either because no recombination event happens or because the recombination event does not affect the first coalescence

$$q_1(t) = e^{-M\rho t} + (1 - e^{-M\rho t})\frac{1}{t}\frac{1}{M}\int_0^t 1 + (M-3)\exp\left(-M\int_u^t \lambda(v)\mathrm{d}v\right)\mathrm{d}u \quad \text{if} \quad (t,i,j) = (s,k,l)$$

and the probability to change from time s to time t due to a recombination event is

$$q_2(t\,|\,s) = (1 - e^{-M\rho s})\frac{1}{s}\frac{1}{M}2\lambda(t)\begin{cases} \int_0^t \exp\left(-M\int_u^t \lambda(v)\mathrm{d}v\right)\mathrm{d}u & \text{if } t < s \\ \exp\left(-\binom{M}{2}\int_s^t \lambda(v)\mathrm{d}v\right)\int_0^s \exp\left(-M\int_u^s \lambda(v)\mathrm{d}v\right)\mathrm{d}u & \text{if } t > s \end{cases}$$

Here M denotes the number of haplotypes, ρ is the scaled recombination rate and $\lambda(t)$ is the scaled inverse population size as a function of time. When samples from different subpopulations are analyzed, the transition rate can be modified to allow for different coalescence rates within and across populations (**Supplementary Note**).

For only two haplotypes, MSMC is very similar to PSMC[7], with subtle differences due to the different underlying model SMC′ (ref. 6) versus SMC[5]. We therefore call our special case of two haplotypes PSMC′ to distinguish it from PSMC.

Emission rate. For the emission rate, we assume that we are given local estimates of the singleton branch length of the tree, T_S (**Supplementary Note**), which is treated as a 'soft' constraint; that is, we always consider $\max(T_S, Mt)$ for T_S. For more than two haplotypes, an observation of alleles given some state (t, i, j) could be classified into the following categories, each of which has its own emission probability $e(t;T_S)$, derived in the **Supplementary Note**.

(1) No mutation: $e(t;T_S) = 1 - \mu T_S$
(2) Singleton within the pair $[i, j]$: $e(t;T_S) = \mu t$
(3) Singleton outside the pair $[i, j]$: $e(t;T_S) = \mu t$
(4) Higher frequency variant: $e(t;T_S) = 1 - \mu T_S$
(5) Double mutation: $e(t;T_S) = 0$
(6) Missing data: $e(t;T_S) = 1$

If sites have multiple observations from ambiguous phasing results, we average the emission probability over those observations.

Parameter inference. To infer parameters with MSMC, we discretize time into segments, following the quantiles of the exponential distribution. The boundaries of the segments are

$$T_i = -\log\left(1 - \frac{i}{n_T}\right)\Big/\binom{M}{2}$$

doi:10.1038/ng.3015

where n_T is the number of segments and M is the number of haplotypes, which determines the expected time to first coalescence. All times are given in units of $2N_0$ generations, where $2N_0$ is fixed from Watterson's estimator. The number of segments used in the population size analysis was $n_T = 40$ and in the estimation of relative cross coalescence rate was $n_T = 30$. The denominator reflects the quadratically decreased time to first coalescence with increased sample size M. The limits on the inference are given by the second and the second-to-last boundaries, which correspond to the 2.5%- and 97.5%-quantile boundaries of the distribution of first coalescence times.

As in PSMC, we join segments to reduce the parameter search space. The segmentation we used was $10 \times 1 + 15 \times 2$ for the population size inference, that is, the 30 rightmost segments were joined to pairs of 2. For relative cross coalescence rate estimates, we used a pattern of $8 \times 1 + 11 \times 2$. For plots on a logarithmic time axis, we chose the left cutoff to be at $T_1/4$ and omitted the last time interval, as its value is shown also by the joined second-to-last time interval.

In each segment, the coalescence rate is kept constant. If all samples are from the same population, we have one parameter per segment, which is the scaled inverse population size λ_I, where i enumerates the segments in time. If samples from two populations are analyzed, we have three parameters per segment: λ_i^1, λ_i^{12} and λ_i^2. Here λ_i^1 and λ_i^2 denote the coalescence rates within the two populations, and λ_i^{12} denotes the coalescence rate across populations. In principle, there are two further free parameters: the scaled recombination rate ρ and the scaled mutation rate θ. In practice, we fix θ from Watterson's estimator for θ. The recombination rate can be inferred relatively well from PSMC′ (**Supplementary Fig. 1**), and we fix it to that estimate for more than two haplotypes to reduce the parameter search space. We note that inference results on the coalescence rates are relatively independent of the recombination rate.

Maximum-likelihood estimates of all free parameters are generated iteratively by means of the Baum-Welch algorithm[52,53] with a coarse-grained Q function and numerical maximizations using Powell's direction set method. For the coarse-graining of the Q function, we use precomputed transition-emission matrices and evaluate the local variables only every 1,000 bp. This optimization makes it possible to analyze whole human genomes without the need to first bin the data into windows of 100 bp, as in ref. 7.

For only one population analyzed, an estimate of the scaled population size in time interval i is directly given as the inverse of the coalescence rate: $N_i/N_0 = 1/\lambda_i$. The scaling parameter N_0 is fixed from the scaled mutation rate: $N_0 = \theta/(4\mu)$, where $\mu = 1.25 \times 10^{-8}$ is the per-site, per-generation mutation rate. Relative cross coalescence rate estimates are obtained by dividing the cross-population coalescence rate by the average within-population coalescence rate: $\gamma_i = 2\lambda_i^{12}/(\lambda_i^1 + \lambda_i^2)$. In the maximization step of the Baum-Welch algorithm, we constrained the optimization to $\gamma_i \leq 1$.

Implementation. We implemented the model and inference algorithm in the D programming language (see URLs). The source code, together with executables for common platforms, is freely available from GitHub (see URLs). We also provide additional information on the input file format and necessary preprocessing in the **Supplementary Note**.

47. Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
48. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
49. Delaneau, O., Howie, B., Cox, A.J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).
50. Kidd, J.M. *et al.* Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *Am. J. Hum. Genet.* **91**, 660–671 (2012).
51. Chen, G.K., Marjoram, P. & Wall, J.D. Fast and flexible simulation of DNA sequence data. *Genome Res.* **19**, 136–142 (2009).
52. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabalistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, UK, 1998).
53. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).