LETTER

doi:10.1038/nature10231

Inference of human population history from individual whole-genome sequences

Heng Li[1,2] & Richard Durbin[1]

The history of human population size is important for understanding human evolution. Various studies[1–5] have found evidence for a founder event (bottleneck) in East Asian and European populations, associated with the human dispersal out-of-Africa event around 60 thousand years (kyr) ago. However, these studies have had to assume simplified demographic models with few parameters, and they do not provide a precise date for the start and stop times of the bottleneck. Here, with fewer assumptions on population size changes, we present a more detailed history of human population sizes between approximately ten thousand and a million years ago, using the pairwise sequentially Markovian coalescent model applied to the complete diploid genome sequences of a Chinese male (YH)[6], a Korean male (SJK)[7], three European individuals (J. C. Venter[8], NA12891 and NA12878 (ref. 9)) and two Yoruba males (NA18507 (ref. 10) and NA19239). We infer that European and Chinese populations had very similar population-size histories before 10–20 kyr ago. Both populations experienced a severe bottleneck 10–60 kyr ago, whereas African populations experienced a milder bottleneck from which they recovered earlier. All three populations have an elevated effective population size between 60 and 250 kyr ago, possibly due to population substructure[11]. We also infer that the differentiation of genetically modern humans may have started as early as 100–120 kyr ago[12], but considerable genetic exchanges may still have occurred until 20–40 kyr ago.

The distribution of the time since the most recent common ancestor (TMRCA) between two alleles in an individual provides information about the history of change in population size over time. Existing methods for reconstructing the detailed TMRCA distribution have analysed large samples of individuals at non-recombining loci like mitochondrial DNA[13]. However, the statistical resolution of inferences from any one locus is poor, and power fades rapidly upon moving back in time because there are few independent lineages probing deep time depths (in humans, no information is available from mitochondrial DNA beyond about 200 kyr ago, when all humans share a common maternal ancestor[11]). In contrast, a diploid genome sequence contains hundreds of thousands of independent loci, each with its own TMRCA between the two alleles carried by an individual. In principle, it should be possible to reconstruct the TMRCA distribution across the autosomes and the X chromosome by studying how the local density of heterozygous sites changes across the genome, reflecting segments of constant TMRCA separated by historical recombination events. To explore whether we could use this idea to learn about the detailed TMRCA distribution from a diploid whole-genome sequence, we proposed the pairwise sequentially Markovian coalescent (PSMC) model, which is a specialization to the case of two chromosomes of the sequentially Markovian coalescent model[14] (Fig. 1a). The free parameters of this model include the scaled mutation rate, the recombination rate and piecewise constant ancestral population sizes (see Methods). We scaled results to real time, assuming 25 years per generation and a neutral mutation rate of 2.5×10^{-8} per generation[15]. The consequences of uncertainty in the two scaling parameters will be discussed later in the text.

To validate our model, we simulated one-hundred 30-megabase (Mb) sequences with a sharp out-of-Africa bottleneck followed by a population expansion, and inferred population-size history with PSMC (Fig. 2a). PSMC was able to recover the parameters used in the simulation and the variance of the estimate was small between 20 kyr ago and 3 Myr ago. More recently than 20 kyr ago or more anciently than 3 Myr ago, few recombination events are left in the present sequence, which reduces the power of PSMC. Therefore, the estimated effective population size (N_e) in these time intervals was not as accurate and had large variance. To test the robustness of the model, we introduced variable mutation rates and recombination hotspots in the simulation (Supplementary Information). The inference was still close to the true history (Fig. 2b) and a uniform rate of single nucleotide polymorphism (SNP) ascertainment errors did not change our qualitative results either (Supplementary Fig. 2). The simulations did, however, reveal a limitation of PSMC in recovering sudden changes in effective population size. For example, the instantaneous reduction from 12,000 to 1,200 at 100 kyr ago in the simulation was spread over several preceding tens of thousands of years in the PSMC reconstruction.

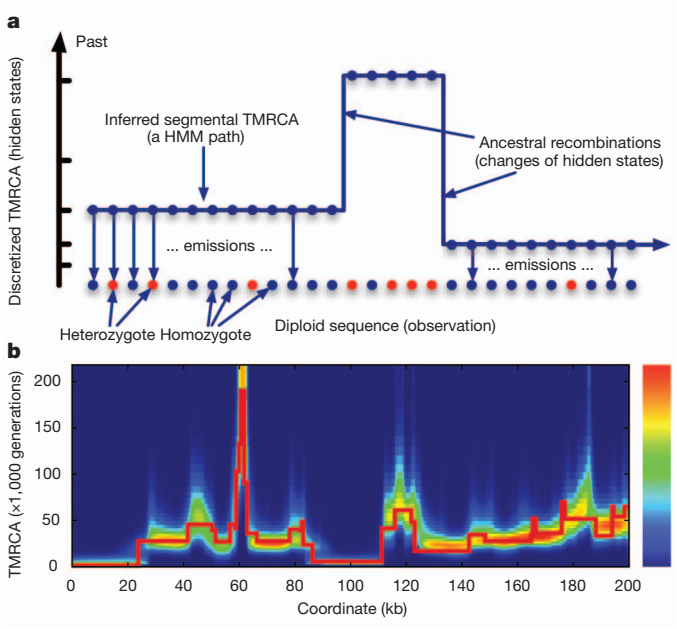

Figure 1 | Illustration of the PSMC model and its application to simulated data. a, The PSMC infers the local time to the most recent common ancestor (TMRCA) on the basis of the local density of heterozygotes, using a hidden Markov model in which the observation is a diploid sequence, the hidden states are discretized TMRCA and the transitions represent ancestral recombination events. **b**, We used the ms software to simulate the TMRCA relating the two alleles of an individual across a 200-kb region (the thick red line), and inferred the local TMRCA at each locus using the PSMC (the heat map). The inference usually includes the correct time, with the greatest errors at transition points.

[1]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.

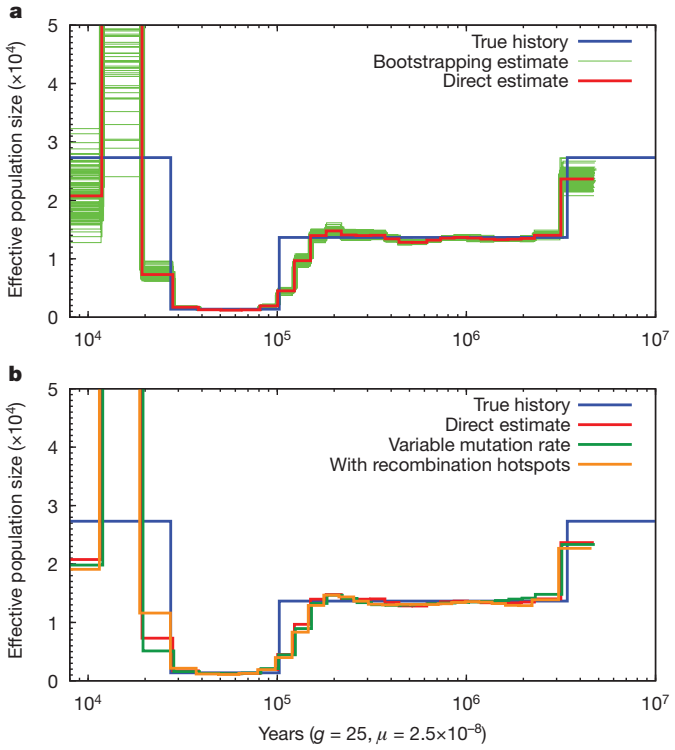

Figure 2 | PSMC estimate on simulated data. **a**, PSMC estimate on data simulated by msHOT. The blue curve is the population-size history used in simulation; the red curve is the PSMC estimate on the originally simulated sequence; the 100 thin green curves are the PSMC estimates on 100 sequences randomly resampled from the original sequence. **b**, PSMC estimate on data with a variable mutation rate or with hotspots. g, generation time; μ, mutation rate.

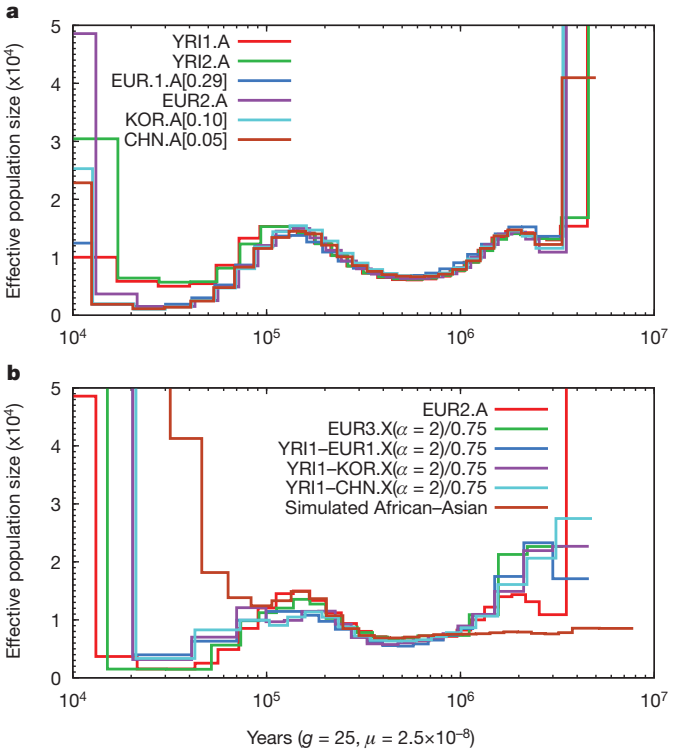

Figure 3 | PSMC estimate on real data. **a**, Population sizes inferred from autosomes of six individuals. 5%, 10% and 29% of heterozygotes are assumed to be missing in CHN.A, KOR.A and EUR1.A, respectively. **b**, Population sizes inferred from male-combined X chromosomes and the simulated African–Asian combined sequences from the best-fit model in ref. 21. Sizes inferred from X-chromosome data are scaled by 4/3. The neutral mutation rate on X, which is used in time-scaling, is estimated with the ratio of male-to-female mutation rate, α, equal to 2 (see Methods).

We applied the PSMC model to real data from recently published genome sequences (see Table 1, which defines the acronyms for samples used elsewhere in the text and figures). Figure 3a shows that all populations are very similar in their estimated N_e history between 150 and 1,500 kyr ago. The Yoruba (YRI) genome differentiates from non-African populations around 100–120 kyr ago (at 110 kyr ago, $N_e^{YRI} = 15{,}313 \pm 559$ and $N_e^{CHN} = 12{,}829 \pm 485$). This evidence of early population differentiation is potentially consistent with the archaeological evidence of anatomically modern humans found in the Near East around 100 kyr ago[12]. European and East Asian populations are nearly identical in estimated N_e before 11 kyr ago. From a peak of 13,500 at 150 kyr ago, the N_e dropped by a factor of ten to 1,200 between 40 and 20 kyr ago, before a sharp increase, the precise magnitude of which we do not have the power to measure. We also observed a less marked bottleneck in YRI from a peak of 16,100 around 100–150 kyr ago to 5,700 at 50 kyr ago, recovering earlier[16] than the out-of-Africa populations, with an increase back to 8,700 by 20 kyr ago, coinciding with the Last Glacial Maximum. All populations showed increased N_e between 60 and 200 kyr ago, about the time of origin of anatomically modern humans[17]. An alternative to an increase in actual population size during this time would be that there was population structure involving separation and admixture[11,16] (Supplementary Fig 5).

We also saw an increase in estimated N_e before 1 million years (Myr) ago in all populations, with a sharp increase before 3 Myr ago. Although it is tempting to read into this the transition from the previously estimated larger N_e at the time of the split from the chimpanzee[18], our method may also be subject to artefacts in this region, due to regions of balancing selection or to clustered false heterozygotes related to segmental duplications (Supplementary Fig. 3).

Analysis of a European female X chromosome (EUR3.X) yielded a history similar to that from autosomes scaled by 0.75, as expected for the X chromosome (Fig. 3b). We did not observe a more severe

Table 1 | Properties of the input sequences

Label	Description	Coverage	Number of called bases (bp)	Number of heterozygotes (bp)	Heterozygosity (×1,000)
YRI1.A (ref. 10)	NA18507 autosomes	×40	2.14×10^9	2.17×10^6	1.013
YRI2.A (ref. 9)	NA19239 autosomes	×29	2.11×10^9	2.21×10^6	1.051
EUR1.A (ref. 8)	Venter autosomes	×9	2.13×10^9	1.23×10^6	0.578
EUR2.A (ref. 9)	NA12891 autosomes	×38	2.11×10^9	1.67×10^6	0.791
KOR.A (ref. 7)	SJK autosomes	×20	2.13×10^9	1.47×10^6	0.690
CHN.A (ref. 6)	YH autosomes	×30	2.19×10^9	1.52×10^6	0.694
YRI3.X (ref. 9)	NA19240 X chromosome	×38	1.06×10^8	7.16×10^4	0.673
EUR3.X (ref. 9)	NA12878 X chromosome	×35	1.10×10^8	4.80×10^4	0.436
KOR–CHN.X	SJK–YH combined X chromosome	-	1.02×10^8	3.97×10^4	0.390
YRI1–EUR1.X	NA18507–Venter combined X chromosome	-	0.83×10^8	5.56×10^4	0.670
YRI1–KOR.X	NA18507–KOR combined X chromosome	-	1.00×10^8	6.69×10^4	0.669
YRI1–CHN.X	NA18507–YH combined X chromosome	-	1.06×10^8	6.95×10^4	0.657

Coverage equals the average number of reads covering HapMap3 loci. A base is said to be called if it passes all filters described (see Methods). The relatively lower coverage for EUR1.A leads to higher sampling bias at heterozygotes, which leads to underestimated heterozygosity, but this can be corrected by adjusting the neutral mutation rate in scaling (Supplementary Information, section 1.2).

bottleneck on the X chromosome[19]. To investigate the relationship between African and non-African populations, we combined X chromosomes from YRI and a non-African to construct a pseudo-diploid genome. From Fig. 3b, we can see that although African and non-African populations might have started to differentiate as early as 100–120 kyr ago, they largely remained as one population until approximately 60–80 kyr ago, the time point at which the YRI1–EUR1.X curve clearly leaves EUR3.X. This supports the recent analysis of the relationship between the Neanderthal genome and that of modern humans[20], which concluded that West Africans and non-Africans descended from a homogeneous ancestral population in the last 100,000 years, with subsequent minor admixture out of Africa from Neanderthals, rather than an alternative explanation involving ancient (>300,000-year-old) sub-structure separating West African and non-African populations.

From Fig. 3b, it is also notable that there is a low N_e between African and non-African populations until approximately 20 kyr ago, indicating that there were substantial genetic exchanges between these populations long after the initial separation. Complete separation would correspond to very large or effectively infinite N_e, as seen more recently than 20 kyr ago. To explore whether the inferred recent gene flow is a modelling artefact, we simulated complete divergence at 60 kyr ago according to the model in ref. 21, and saw increased rather than reduced N_e in the period 20–60 kyr ago (brown line in Fig. 3b). To explore further, we extracted segments from YRI1–KOR.X that coalesced more recently than 50 kyr ago, according to PSMC. These comprised 220 segments covering 31.2 Mb (>20% of the X chromosome). We observed 1,363 base-pair (bp) differences in 20.7 Mb of call-able sequence in these segments, corresponding to an average divergence time of 37.4 kyr ago. In contrast, if we apply the same process to the simulated data from the model of ref. 21, the segments that PSMC identifies as having diverged more recently than 50 kyr ago cover only 0.4% of the simulated chromosome. The human–macaque divergence in the 220 segments was only 4% lower than the chromosome average, so regional variability in mutation rates cannot explain these results. In summary, the existence of long segments of low divergence between YRI1 and KOR supports the inference from PSMC that there was substantial genetic exchange between West African and non-African populations up until 20–40 kyr ago, and is not consistent with a simple separation approximately 60 kyr ago.

The time frame proposed above for continued genetic exchange between Africans and non-Africans is more recent than the archaeologically documented time of the out-of-Africa dispersal, because there are modern human fossils in both Europe and Australasia that date to >40 kyr ago[22]. Further analysis of additional non-African genomes indicates that this genetic exchange occurred primarily before the separation of Europeans and East Asians (Supplementary Information, section 4.3). An important caveat to this conclusion is the uncertainty of the per-year mutation rate of 1.0×10^{-9} ($2.5 \times 10^{-8}/25$). Although this mutation rate agrees well with the rates estimated between primates averaged over millions of years (Supplementary Information, section 3.1), generation intervals as high as 29 years per generation over the last few thousand years[23], and present mutation rates lower than 2.5×10^{-8} per generation[9], are possible in principle. These factors could make our recent date estimates too recent, although it seems unlikely that such inaccuracies would be consistent with a date of final genetic exchange as far back as 60 kyr ago. Our analyses also cannot exclude the possibility that the divergence time inferred from X chromosomes may not be representative, owing to sex-biased demographic processes[19], highlighting the importance of repeating this analysis on autosomal data once haploid whole-genome sequences become available[24]. Notably, a recent study using an orthogonal type of data (analysis of allele frequencies) also inferred that gene flow between Africans and non-Africans continued well after the initial out-of-Africa migration: in the case of that study, until 17–26 kyr ago[25]. An important goal for future work is to determine whether these recent dates reflect real history, and if so, to obtain more detail about the timing and scale of the events involved.

In this paper we have introduced a method to infer the history of effective population size from genome-wide diploid sequence data. It is relatively straightforward to apply, with less potential ascertainment bias than existing methods that use selective genotyping data or resequencing data from a few loci. Furthermore, our method is computationally tractable and typically uses much more primary sequence data than the existing methods, which allows us to estimate population size at each time going back in history, rather than assuming a parametric structure of times, divergences and size changes. The results described above concerning the timing and depth of the out-of-Africa bottleneck are broadly consistent with previous studies, although our results are more detailed (Supplementary Information, section 4.2). The hypothesis that there was significant ongoing genetic exchange throughout the bottleneck is surprising in light of current views about human migrations; however, it is not inconsistent with the archaeological literature, and should motivate further research. There is the potential to extend this type of sequentially Markovian coalescent hidden Markov model approach to data from several individuals, which would access more recent times, but this will require inference over a substantially more complex hidden-state-space of trees on the haplotypes, with each Markov path representing an ancestral recombination graph[14]. In addition, there is the potential to apply the method to investigate the population-size history of other species for which a single diploid genome sequence has been obtained (Supplementary Information, section 2.2).

METHODS SUMMARY

Illumina short reads were obtained from the NCBI Sequence Read Archive and capillary reads from TraceDB. Reads were aligned to the human reference genome with BWA[26]. The consensus sequences were called by SAMtools[27] and then divided into non-overlapping 100-bp bins, with a bin being scored as heterozygous if there is a heterozygote in the bin, or as homozygous otherwise. The resultant bin sequences were taken as the input of the PSMC estimate. Coalescent simulation was done by ms[28] and cosi[21] software. The simulated sequences were binned in the same way.

The free parameters in the discrete PSMC–HMM model are the scaled mutation rate, recombination rate and piecewise constant population sizes. The time interval spanned by each size parameter was manually chosen. The expectation-maximization iteration started from a constant-sized population history. The expectation step was done analytically; Powell's direction set method was used for the maximization step. Parameter values stabilized by the twentieth iteration and these were taken as the final estimate. All parameters were scaled to a constant that is further determined under the assumption of a neutral mutation rate, 2.5×10^{-8}.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 April 2009; accepted 20 May 2011.
Published online 13 July 2011.

1. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411,** 199–204 (2001).
2. Marth, G. T., Czabarka, E., Murvai, J. & Sherry, S. T. 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).
3. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2,** e105 (2006).
4. Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genet.* **39,** 1251–1255 (2007).
5. Fagundes, N. J. R. *et al.* Statistical evaluation of alternative models of human evolution. *Proc. Natl Acad. Sci. USA* **104,** 17614–17619 (2007).
6. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
7. Ahn, S.-M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19,** 1622–1629 (2009).
8. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
9. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
10. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

11. Behar, D. M. *et al.* The dawn of human matrilineal diversity. *Am. J. Hum. Genet.* **82,** 1130–1140 (2008).
12. Mellars, P. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* **313,** 796–800 (2006).
13. 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).
14. McVean, G. A. T. & Cardin, N. J. Approximating the coalescent with recombination. *Phil. Trans. R. Soc. B* **360,** 1387–1393 (2005).
15. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
16. 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).
17. Wall, J. D. & Hammer, M. F. Archaic admixture in the human genome. *Curr. Opin. Genet. Dev.* **16,** 606–610 (2006).
18. Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* **3,** e7 (2007).
19. Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nature Genet.* **41,** 66–70 (2009).
20. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
21. Schaffner, S. F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15,** 1576–1583 (2005).
22. Mellars, P. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. *Nature* **439,** 931–935 (2006).
23. 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).
24. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnol.* **29,** 59–63 (2010).
25. 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).
26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
27. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
28. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to D. Bentley (Illumina) and J. Wang (Beijing Genomics Institute) for early access to the sequencing data. We thank A. Coghlan for the idea of bootstrapping, and N. Patterson, M. Przeworski, D. Reich, and members of the Durbin research group for discussions and critiques. This work was funded by Wellcome Trust grant WT077192.

Author Contributions R.D. proposed the basic strategy and designed the overall study. H.L. developed the theory, implemented the algorithm and analysed results. R.D. and H.L. wrote the manuscript.

Author Information The PSMC software package is freely available at http://github.com/lh3/psmc. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.D. (rd@sanger.ac.uk) or H.L. (lh3@sanger.ac.uk).

METHODS

Read alignment. Alignment for individuals from the 1000 Genomes Project (NA12878, NA12891, NA19239 and NA19240) was obtained from the project FTP site. Illumina sequence reads for NA18507, YH and SJK were obtained from the NCBI Sequence Read Archive (AC:ERA000005, SRA000271 and SRA008175, respectively) and Sanger sequencing reads for Craig Venter were obtained from NCBI TraceDB. These sequence reads were mapped by BWA[26] (0.5.5) against the human reference genome build 36, including unassembled contigs and the genome of Epstein–Barr virus (AC:NC_007605), with pseudoautosomal regions on the Y chromosome masked. For Illumina short reads, BWA option '-q 15' was applied to enable trimming of low-quality bases at the 3′ end. Base qualities of SJK reads were overestimated and were therefore recalibrated using GATK[29] after alignment, with known SNPs in dbSNP-129 discarded. For capillary reads, the BWA-SW algorithm with the default options was used.

Calling the consensus sequence. The diploid consensus sequence for an autosome was obtained by the 'pileup' command of the SAMtools software package[27], and then processed with the following loci marked as missing data: 1) read depth is more than twice or less than half of the average read depth estimated on HapMap3 genotyping loci; 2) the root mean squared mapping quality of reads covering the locus is below 25; 3) the locus is within 10 bp around predicted short insertions or deletions; 4) the inferred consensus quality is below a threshold (20 for Illumina data and 10 for capillary data); 5) fewer than 18 out of the 35 overlapping 35-bp oligonucleotides from the reference sequence can be mapped elsewhere with zero or one mismatch.

The X-chromosome consensus was derived in a similar way but with pseudoautosomal regions filtered as missing data. The X chromosomes of males are haploid and therefore the few heterozygotes that were called were discarded as errors. The pseudo-diploid X chromosomes of males were combined by marking a difference as a heterozygote.

The consensus sequences were further divided into 100-bp non-overlapping bins with each bin represented as 'missing' (marked '.') if ≥90 bases were filtered or uncalled; as heterozygous ('1') if >10 bp were called and there was at least one heterozygote; or as homozygous ('0') otherwise. The sequence of bin values was taken as the input of the PSMC inference.

Coalescent simulation. One-hundred sequences of 30 Mb were simulated by ms software[28] with piecewise constant history, as shown in Fig. 2a. To simulate variation in mutation rate, the local mutation rate averaged in a 20-kb window between human and macaque was calculated from the EPO cross-species alignment obtained from Ensembl v50. In the simulation, the local coalescent trees were simulated with ms but mutations were generated on the basis of the relative local mutation rate on a 30-Mb segment randomly drawn from the human–macaque alignment. The program msHOT was used to simulate sequences with recombination hotspots. The location and size of hotspots were randomly drawn from the hotspot map obtained from HapMap (release 21); the scaled recombination rate in hotspots was tenfold higher than that in non-hotspot regions.

The cosi software package was used to simulate sequences under the best-fit model from ref. 21. This model considers variable recombination rates, recombination hotspots and migration between African and non-African populations.

Overview of the PSMC model. In the PSMC–HMM, the observation is a binary sequence of '0', '1' and '.', as described above. The emission probability from state t is $e(1|t) = e^{-\theta t}$, $e(0|t) = 1-e^{-\theta t}$ and $e(.|t) = 1$; the transition probability from s to t is:

$$p(t|s) = (1-e^{-\rho t})q(t|s) + e^{-\rho s}\delta(t-s)$$

where θ is the scaled mutation rate, ρ is the scaled recombination rate, $\delta(\cdot)$ is the Dirac delta function and

$$q(t|s) = \frac{1}{\lambda(t)}\int_0^{min\{s,t\}} \frac{1}{s} \times e^{-\int_u^t \frac{dv}{\lambda(v)}} du$$

is the transition probability conditional on there being a recombination event, where $\lambda(t) = N_e(t)/N_0$ is the relative population size at state t. The discrete-state HMM is constructed by dividing coalescence-time into intervals and integrating emission and transition probabilities in the intervals, which can be done analytically given a piecewise-constant function, $\lambda(t)$. The stationary distribution of TMRCA can also be analytically derived. Details are in Supplementary Information.

Scaling to real time. The estimated TMRCA is in units of $2N_0$ time, and $\lambda(t)$ is scaled to N_0 as well. The value of N_0 cannot be determined from the model itself. To estimate N_0, a neutral mutation rate $\mu_A = 2.5 \times 10^{-8}$ on autosomes[15] was used and thus $N_0^A = \theta/4\mu_A$. Given the ratio of male-to-female mutation rate[30] $\alpha = 2$, the neutral mutation rate of X chromosomes was derived as $\mu_X = \mu_A\,[2(2+\alpha)]/[3(1+\alpha)] = 2.2 \times 10^{-8}$. If heterozygotes are missed uniformly at a probability p, this is equivalent to reducing the neutral mutation rate from μ to $\mu' = \mu\,(1-p)$. False negatives due to the lack of coverage can thus be corrected. Generations were converted to years under the assumption of 25 years per generation.

Parameter estimate with PSMC. Given a maximum TMRCA in the $2N_0$ scale of T_{max}, and a number of atomic time intervals n, let the boundaries of these intervals be $t_i = 0.1\exp[i/n\,\log(1+10T_{max})]-0.1$, $i = 0,\ldots, n$. To reduce the complexity of the search space, blocks of adjacent atomic intervals were combined to have the same population-size parameter via a user-specified pattern. On autosome and simulated data, $T_{max} = 15$, $n = 64$ and the pattern is '1*4 + 25*2 + 1*4 + 1*6', which means that the first population-size parameter spans the first four atomic time intervals, each of the next 25 parameters spans two intervals, the twenty-seventh parameter spans four intervals and the last parameter spans the last six time intervals. On X-chromosome data, $T_{max} = 15$, $n = 60$ and the pattern is '1*6 + 2*4 + 1*3 + 13*2 + 1*3 + 2*4 + 1*6'.

In the expectation-maximisation (EM) parameter estimate, the initial population-size parameters were all set as 1, representing a constant-sized history, the scaled mutation rate was calculated to match the observed heterozygosity and the initial value of the scaled recombination rate was arbitrarily set as one-quarter of the mutation rate. At the maximization step, Powell's direction set method was used to minimize the Q function in the EM algorithm numerically. Parameters at the twentieth EM iteration were taken as the final results.

Bootstrapping was applied by breaking the consensus sequences into 5-Mb segments and randomly sampling a set of segments with replacement, such that the total length of the sampled segments was close to the size of the human reference genome.

Further discussion of methods and parameters is given in Supplementary Information.

29. McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
30. Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. & Yasunaga, T. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb. Symp. Quant. Biol.* **52,** 863–867 (1987).