nature genetics

Bayesian inference of ancient human demography from individual genome sequences

Ilan Gronau[1], Melissa J Hubisz[1], Brad Gulko[2], Charles G Danko[1] & Adam Siepel[1]

Whole-genome sequences provide a rich source of information about human evolution. Here we describe an effort to estimate key evolutionary parameters based on the whole-genome sequences of six individuals from diverse human populations. We used a Bayesian, coalescent-based approach to obtain information about ancestral population sizes, divergence times and migration rates from inferred genealogies at many neutrally evolving loci across the genome. We introduce new methods for accommodating gene flow between populations and integrating over possible phasings of diploid genotypes. We also describe a custom pipeline for genotype inference to mitigate biases from heterogeneous sequencing technologies and coverage levels. Our analysis indicates that the San population of southern Africa diverged from other human populations approximately 108–157 thousand years ago, that Eurasians diverged from an ancestral African population 38–64 thousand years ago, and that the effective population size of the ancestors of all modern humans was ~9,000.

During the past several decades, investigators from various disciplines have produced a broad outline of the events that gave rise to major human population groups drawing from genetic, anthropological and archaeological evidence[1]. The general picture that has emerged is that anatomically modern humans arose roughly 200 thousand years ago (kya) in eastern or southern Africa, that a small tribe began to expand throughout Africa ~100 kya, that a major migration out of Africa occurred ~40–60 kya and that the descendants of these migrants subsequently populated Europe, Asia and the remaining inhabitable regions of the world, possibly with some introgression from archaic hominids[2,3]. This outline is supported by analyses of mitochondrial and Y-chromosomal data[4,5], autosomal microsatellite markers[6,7], sequences for selected autosomal loci[8–11] and genome-wide genotyping data[12]. Nevertheless, much remains unknown about early human demography. Indeed, current estimates of key parameters such as the date of the migration out of Africa often vary by factors of two or three.

We attempted to investigate these issues using recently released whole-genome sequences for individual humans[13–17]. Although individual genome sequencing studies so far have emphasized the technical feasibility of sequencing, discovery of new genetic variants and identification of disease causing mutations, these data are also potentially informative about human evolution. We examined the published whole-genome sequences of six individuals from six different population groups (**Table 1**). One of these individuals is a member of the Khoisan-speaking hunter-gatherer populations of southern Africa known collectively as the San[17]. Along with other indigenous groups from central and southern Africa[18,19], the San population shows the highest known levels of genetic divergence from other human populations and therefore should be highly informative about ancient human demography. For reasons of statistical power, our demographic analysis focused on the timing of early divergence events between major population groups, in particular between the San population and other groups (the 'San divergence'; **Fig. 1**) and between the Eurasians and other African groups (the 'African-Eurasian divergence').

In analyzing these data, we used a Bayesian statistical approach, based on coalescent theory, that was originally developed for individuals belonging to closely related but distinct species such as human, chimpanzee and gorilla[20,21]. This approach (as implemented in the computer program MCMCcoal) derives information about ancestral population sizes and population divergence times from the patterns of variation in the genealogies at many neutrally evolving loci given a

Table 1 Individual genomes analyzed here

Genome[a]	Population	Technology[b]	Reads[c]	Redundancy[d]	Coverage[e]	Depth[f]	HQC[g]	Reference
Venter	European	Sanger	800 bp PE	7.5	0.912	8.4	0.577	13
NA18507	Yoruban	Illumina	35 bp PE	40.6	0.900	41.1	0.672	14
YH	Han Chinese	Illumina	35 bp PE	36.0	0.896	25.4	0.671	15
SJK	Korean	Illumina	36,75 bp	29.0	0.903	19.7	0.672	16
ABT	Bantu	SOLiD	49 bp	>30.0	0.874	21.4	0.641	17
KB1	San	Illumina[h]	76 bp	23.1	0.901	23.6	0.621	17

[a]Genome identifiers are surnames of sequenced individuals (Venter), identifiers for Coriell DNA samples (NA18507) or abbreviations introduced in published papers (YH, SJK, ABT and KB1). [b]Sequencing technology: Sanger, Sanger (capillary) sequencing; Illumina, Illumina Genome Analyzer; SOLiD, SOLiD system by Applied Biosystems. [c]Average read length in bp and whether or not paired-end (PE) reads were used. [d]Sequencing redundancy, or fold coverage, as reported in the previously published paper (see refs. column for the paper used). [e]Fraction of the genome covered by uniquely aligned reads according to the pipeline used here. [f]Average number of uniquely aligned reads at positions having at least one uniquely aligned read excluding duplicate reads. [g]Fraction of the genome covered by aligned reads that passed data quality filters. [h]KB1 was sequenced using both the 454 and Illumina methods, but the current analysis used the more abundant Illumina data.

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, USA. [2]Graduate Field of Computer Science, Cornell University, Ithaca, New York, USA. Correspondence should be addressed to A.S. (acs4@cornell.edu).

Received 22 February; accepted 16 August; published online 18 September 2011; doi:10.1038/ng.937

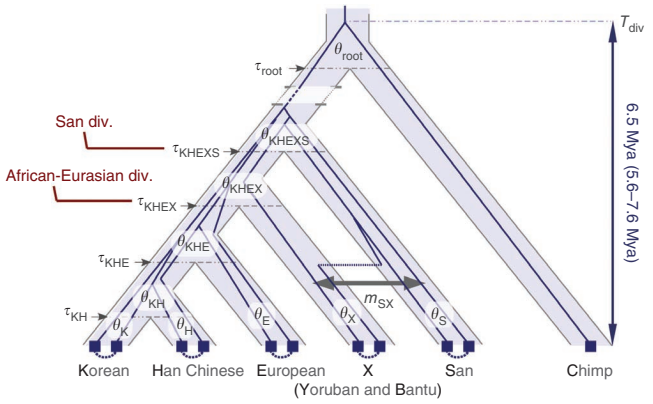

Figure 1 Population phylogeny and genealogies. The population phylogeny assumed in this study with one diploid genome per population (**Table 1**) and a haploid chimpanzee outgroup. We included the Yoruban and Bantu individuals in the analysis as alternative African ingroups (denoted X) because their relationship to one another was uncertain (**Supplementary Note**). The free parameters in our model include the five population divergence times (τ) and the ten effective population sizes (θ), all expressed in units of expected mutations per site. We also considered various 'migration bands' (gray double-headed arrow) to allow for gene flow between populations, treating the (constant) migration rates along bands as free parameters. The two parameters of primary interest were the San (τ_{KHEXS}) and African-Eurasian (τ_{KHEX}) divergence times (div.). We obtained absolute divergence times (in years) and effective population sizes (in numbers of individuals) by assuming a human-chimpanzee average genomic divergence time of 5.6–7.6 Mya and a point estimate of 6.5 Mya.

population phylogeny and a set of sequence alignments. Essentially, it exploits the fact that even small numbers of present-day genomes represent many ancestral genomes that have been shuffled and assorted by the process of recombination. Because the sequences provide only very weak information about the genealogy at each locus, the method integrates over candidate genealogies using Markov chain Monte Carlo (MCMC) methods and pools information across loci in obtaining an approximate posterior distribution for the parameters of interest.

A major challenge in carrying out a population genetic analysis of the available individual whole-genome sequences is that biases may result from differences in power and accuracy in single nucleotide variant detection stemming from differences in sequencing technologies, depth of coverage and bioinformatic methods (**Table 1** and **Supplementary Table 1**). To address this problem, we developed our own pipeline for genotype inference that realigns all raw sequence reads in a uniform manner, empirically recalibrates base call quality scores, calls genotypes using our own reference-genome–free Bayesian genotype inference algorithm (BSNP) and applies a series of rigorous data-quality filters (**Supplementary Fig. 1**). We validated this pipeline using alternative array- and sequence-based calls for two genomes and found that our calls were similar to the others in overall accuracy while avoiding biases from the use of the reference genome in genotype inference (**Supplementary Figs. 2–8**, **Supplementary Table 2** and **Supplementary Note**). We also found that our pipeline eliminated inconsistencies in heterozygosity and SNP density shown by the published genotype calls for these genomes.

A second challenge is that MCMCcoal relies on two assumptions that do not apply here: (i) an absence of gene flow between populations and (ii) the existence of haploid samples from each individual. Using the MCMCcoal source code as a starting point, we developed our own program, called G-PhoCS (Generalized Phylogenetic Coalescent Sampler or 'G-fox'), that relaxes these assumptions. To allow for gene flow, we introduced 'migration bands' that allow for continuous migration at constant rates between designated populations. Following previous isolation-with-migration methods[22,23], we altered the sampling procedure so that it would explore genealogies that crossed population boundaries within these bands (**Fig. 1**). To allow the use of unphased diploid genotype data, we devised a method that integrates over all possible phasings of heterozygous genotypes when computing genealogy likelihoods. Notably, this method makes use of both chromosomes per individual, effectively doubling the size of the dataset. We carried out a series of simulations to test whether G-PhoCS is capable of recovering known parameters from a dataset like ours and found that the parameters of primary interest—the San and African-Eurasian divergence times—can be estimated without bias and with reasonably narrow credible intervals even when the genotypes are unphased and gene flow is present (**Fig. 2**, **Supplementary Figs. 9–11** and **Supplementary Note**). We observed reduced power for recent divergence times, current effective population sizes and migration direction.

Next, we analyzed alignments of the six individual genomes and chimpanzee reference genome at 37,574 1-kb 'neutral loci' excluding protein-coding and conserved noncoding regions. We defined these loci to minimize intralocus recombination but to ensure frequent

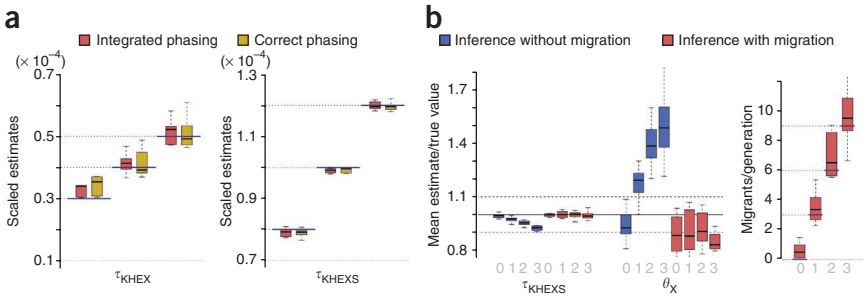

Figure 2 Results of the simulation study. Simulations assumed a population tree like the one shown in **Figure 1** and plausible divergence times, population sizes and migration scenarios (**Supplementary Note**). (**a**) Accuracy of estimated African-Eurasian (τ_{KHEX}) and San (τ_{KHEXS}) divergence times without migration. Dotted lines indicate the values assumed for the simulations, and each boxplot summarizes the posterior mean estimates in six separate runs of G-PhoCS. Results are shown for correctly phased data (gold) and integration over unknown phasings (red). A random phasing procedure produced substantially poorer results (**Supplementary Fig. 10**). Most estimates fall within 10% of the true value, except for the smallest assumed divergence times, where weak information in the data leads to an upward bias. (**b**) Accuracy of the estimated San divergence time (τ_{KHEXS}) and the Yoruban-Bantu population size (θ_X) in simulations with four levels of constant-rate migration (denoted 0, 1, 2 and 3 in order of increasing strength) from population S to population X. Ratios of the estimated to true values are shown when migration is not allowed (blue) and is allowed (red) in the model. Each boxplot summarizes 12 runs. Notice that there is a pronounced bias when migration is present but is not modeled, but this bias is eliminated when migration is added to the model. Simulated and estimated migration rates (measured in expected number of migrants per generation) are shown at right (see **Supplementary Figs. 9–11** for the complete results).

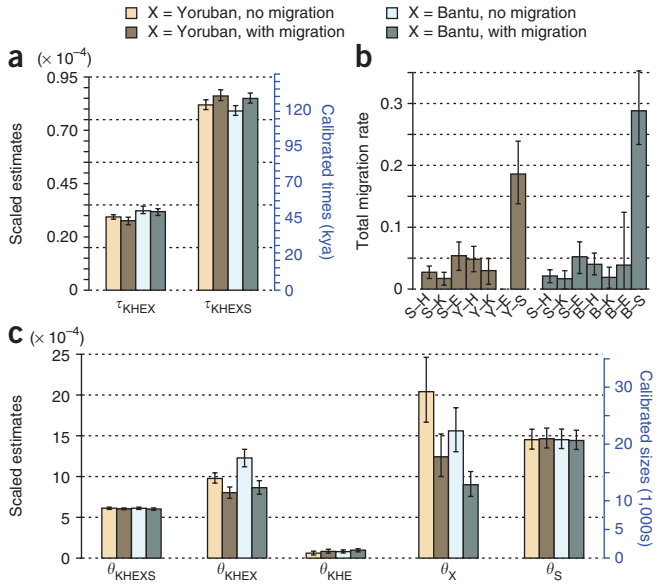

Figure 3 Parameter estimates from real data. Estimates of population divergence times (**a**), migration rates (**b**) and effective population sizes (**c**) obtained for various scenarios. In **a** and **c**, both mutation-scaled (left) and calibrated (right) *y* axes are shown (with a calibration of T_{div} = 6.5 Mya). Results are shown for scenarios with either the Yoruban or Bantu ingroup X and with or without a migration band between X and the San ingroup. Panel **b** shows estimated migration rates for 14 different migration bands. Only the Yoruban-San (Y-S) and Bantu-San (B-S) migration scenarios are strongly supported. In all panels, each bar represents the mean estimate and 95% credible interval (error bars) of a single representative run of the program (see **Supplementary Tables 4** and **5** and **Supplementary Figs. 12** and **13** for complete results).

recombination between loci. We assumed the five-population phylogeny shown in **Figure 1** using as an 'African ingroup' either the Yoruban or the Bantu group. We evaluated 16 alternative scenarios with various migration bands and performed two replicate runs per scenario (**Supplementary Table 3**), cross-checking all results to ensure convergence. To convert estimates of divergence time (τ) and population size (θ) from mutations per site to years (T) and effective numbers of individuals (N), respectively, we assumed a human to chimpanzee average genomic divergence time of T_{div} = 5.6–7.6 million years (Mya) with a point estimate of T_{div} = 6.5 Mya[2,24] (Online Methods). Consistently across runs, a calibration of T_{div} = 6.5 Mya implied a mutation rate of ~2.0×10^{-8} per generation per site, which is in good agreement with independent estimates[25]. Unless otherwise stated, all parameter estimates are reported as posterior means (with 95% credible intervals) in calibrated form based on T_{div} = 6.5 Mya. For estimates of N, we also assumed an average generation time of 25 years.

Assuming no gene flow, we estimated a San divergence time of 125 (121–128) kya with the Yoruban ingroup and 121 (117–124) kya with the Bantu ingroup (**Fig. 3a**). If gene flow is allowed between the San and the African ingroups, these estimates increase slightly to 131 (127–135) kya and 129 (126–133) kya, respectively. Thus, our best estimate of the San divergence time is ~130 kya or 108–157 kya across calibration times (**Table 2**). Of the several migration scenarios considered, those involving the San and the Yoruban or Bantu ingroups were the only ones showing pronounced evidence of gene flow within the limitations of our model (**Fig. 3b**). Notably, we detected the strongest migration signal for the Bantu and San populations, for which gene flow has been reported previously[17].

Our estimates of the African-Eurasian divergence time were also highly consistent across runs, with mean values of ~50 kya and a full range of 38–64 kya (**Table 2**). These estimates showed almost no influence from migration (**Fig. 3a**). We observed only slight differences between the divergence times for the Yoruban (~47 kya) and Bantu (~53 kya) ingroups. Our power for more recent events was reduced, but, notably, we estimated 31–40 kya (26–47 kya across calibrations) for the European–east Asian divergence (**Supplementary Table 4**), a date that is more easily reconciled with the fossil record in Europe than estimates of ~20 kya based on allele frequency data[11,12]. Our estimates of effective population size (θ) are consistent with a population expansion in Africa—we observed a steady increase from θ_{KHEXS} to θ_{KHEX} and then to θ_X and θ_S (**Figs. 1** and **3c**)—whereas the estimates for the Eurasian populations indicate a pronounced bottleneck. Most estimates of θ were unaffected by gene flow, except those measurements for the ingroup populations and their immediate ancestors, which behaved in the expected manner. We estimated the effective size of the MRCA population, N_{KHEXS}, with high confidence at ~9,000 (or ~7,500–10,500 for T_{div} = 5.6–7.6 Mya), a measurement which was highly robust to the choice of ingroup and migration scenario.

Although our estimates of several demographic parameters, including the African-Eurasian divergence time[7,9] and the ancestral effective population sizes[8,9,18], show reasonable agreement with numerous recent studies (**Supplementary Note**), only a few previous multilocus studies have included representatives from the San population. Furthermore, these studies have generally produced estimates of the San divergence time that are considerably less precise than our genome-wide estimate of 126–133 kya (or 108–157 kya across calibrations); estimates have ranged from 71–142 kya[6], to 78–129 kya (assuming T_{div} = 6.5 Mya)[2] to 145–215 kya (not including large credible intervals)[18]. Notably, our point estimate of ~130 kya suggests that the San divergence occurred ~2.5 times as long ago as the African-Eurasian divergence, that major human population groups diverged at least ~80,000 years before the out-of-Africa migration and that the San divergence is more than one-third as ancient as the human-Neanderthal divergence (estimated at 316–341 kya for T_{div} = 6.5 Mya using somewhat different methods[2]). Still, human effective population sizes are sufficiently large that these divergence times are small relative to the time required for lineages to find common ancestors in ancestral populations. Indeed, of the mutations differentiating a San individual from a Eurasian individual, only about 25% are expected to have arisen since the San divergence. Thus, the ancient divergence of the San population does not alter the essential fact that far more human variation occurs within population groups than between them[26].

Table 2 Estimated divergence times with migration

Divergence event	Ingroup (X)	Raw estimates[a]	Calibrated estimates[b]		
			T_{div} = 5.6 Mya	T_{div} = 6.5 Mya	T_{div} = 7.6 Mya
San (τ_{KHEXS})	Yoruban	0.91 (0.89–0.94)	113 (110–116)	131 (127–135)	153 (149–157)
San (τ_{KHEXS})	Bantu	0.90 (0.88–0.93)	111 (108–114)	129 (126–133)	151 (147–155)
AE (τ_{KHEX})	Yoruban	0.33 (0.31–0.34)	40 (38–42)	47 (44–49)	55 (51–57)
AE (τ_{KHEX})	Bantu	0.37 (0.35–0.38)	46 (43–47)	53 (50–55)	62 (59–64)

Raw and calibrated estimates for the San (τ_{KHEXS}) and African-Eurasian (AE) (τ_{KHEX}) divergence times. Separate results are shown for the Yoruban and Bantu representatives of the African ingroup population X. In all cases, a migration band between the San and the African ingroup X was included in the model.
[a]Raw estimates (shown as mean and (95% Bayesian credible intervals)) are given in units of expected mutations per site × 10^{-4}. [b]Calibrated estimates are given in thousands of years (kya) for three different human-chimpanzee calibrations (T_{div} = 5.6, 6.5 and 7.6 million years (Mya)).

In principle, our estimates could be influenced by various complex features of human evolution not adequately considered in our model. However, in a series of follow-up analyses, we could find no evidence that our estimates were strongly influenced by intralocus recombination, mutation rate variation, changes in population size along lineages or our choice of prior distributions (**Supplementary Note**). Moreover, it is doubtful that the scenario hypothesized in the recent analysis of the Neanderthal genome—with low levels of gene flow from Neanderthals to ancestral non-Africans[2]—would substantially change the San divergence time while leaving the African-Eurasian divergence time well within the feasible range. Nevertheless, it should be possible to characterize the demographic history of early humans in greater detail as additional genome sequences become available.

Our methods represent an important step toward coalescent-based inference of demographic parameters from whole-genome sequences. This approach has a number of potential advantages compared with methods based on approximate Bayesian computation[27], summary likelihood approaches[8,10] and the site frequency spectrum[11]. By explicitly representing genealogical relationships at neutrally evolving loci, the coalescent-based approach can more accurately capture the correlation structure of the data, which may lead to improvements in parameter estimation[27]. Moreover, it allows for simple and direct estimation of the posterior distributions of any genealogy-derived quantities of interest, such as times to most recent common ancestors or rates of migration over time. Unlike a recently published method that analyzes individual genomes in isolation[28], our approach simultaneously considers multiple populations and allows direct estimation of divergence times and migration rates. However, by circumventing the critical issue of recombination through the analysis of short loci assumed to be in linkage equilibrium, our methods fail to exploit the information about demography that is provided by patterns of linkage disequilibrium (for example, in the length distribution of shared haplotypes)[10] and instead relies on a relatively weak signal from mutation to drive the inference procedure (our dataset contains only 1.9 polymorphic sites per locus). Therefore we see an opportunity for improved methods for multi-population coalescent-based demographic inference that consider both mutation and recombination and allow entire chromosomes to be analyzed. Recent progress in this area[29,30] suggests that, with clever approximations and careful algorithm design, it may be possible to develop methods that scale to dozens of complete genomes.

METHODS

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

URLs. G-PhoCS, http://compgen.bscb.cornell.edu/GPhoCS/; UCSC Genome Browser, http://genome.ucsc.edu/; MCMCcoal, http://abacus.gene.ucl.ac.uk/software/MCMCcoal.html.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This research was supported by a Packard Fellowship (to A.S.), National Science Foundation grant DBI-0644111 and National Institutes of Health training grant T32HD052471 (to C.G.D.). We thank S. Schuster, W. Miller, D. Reich, G. Coop, J. Hey, J. Wall, R.S. Wells, A. Keinan, A.G. Clark, S.C. Choi, C.D. Bustamante, B. Henn and others for helpful discussions and feedback.

AUTHOR CONTRIBUTIONS

A.S. conceived of and designed the study. I.G. implemented G-PhoCS and applied it to both simulated and real data. B.G. implemented BSNP and applied it to the individual genomes. I.G., M.J.H., B.G., C.G.D. and A.S. performed additional statistical analyses. I.G. and A.S. wrote the paper with review and contributions by all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Cavalli-Sforza, L.L. & Feldman, M.W. The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.* **33** (suppl.) 266–275 (2003).
2. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
3. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
4. Cann, R.L., Stoneking, M. & Wilson, A.C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
5. Gonder, M.K., Mortensen, H.M., Reed, F.A., de Sousa, A. & Tishkoff, S.A. Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.* **24**, 757–768 (2007).
6. Zhivotovsky, L.A., Rosenberg, N.A. & Feldman, M.W. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.* **72**, 1171–1186 (2003).
7. Liu, H., Prugnolle, F., Manica, A. & Balloux, F. A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* **79**, 230–237 (2006).
8. Voight, B.F. *et al.* Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl. Acad. Sci. USA* **102**, 18508–18513 (2005).
9. Fagundes, N.J. *et al.* Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. USA* **104**, 17614–17619 (2007).
10. Wall, J.D., Lohmueller, K.E. & Plagnol, V. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* **26**, 1823–1827 (2009).
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. 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. *Nat. Genet.* **39**, 1251–1255 (2007).
13. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
14. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
15. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
16. 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).
17. Schuster, S.C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947 (2010).
18. 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).
19. Tishkoff, S.A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
20. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656 (2003).
21. Burgess, R. & Yang, Z. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.* **25**, 1979–1994 (2008).
22. Nielsen, R. & Wakeley, J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896 (2001).
23. Hey, J. Isolation with migration models for more than two populations. *Mol. Biol. Evol.* **27**, 905–920 (2010).
24. Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103–1108 (2006).
25. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2003).
26. Lewontin, R.C. The apportionment of human diversity. in *Evolutionary Biology* (eds. Dobzhansky, T.H., Hecht, M.K. & Steere, W.C.) 6 (Appleton-Century-Crofts, New York, New York, USA, 1972).
27. Beaumont, M.A., Zhang, W. & Balding, D.J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
28. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
29. 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).
30. Paul, J.S., Steinrucken, M. & Song, Y.S. An accurate sequentially markov conditional sampling distribution for the coalescent with recombination. *Genetics* **187**, 1115–1128 (2011).

ONLINE METHODS

Genotyping pipeline. Our pipeline for genotype inference consisted of five major stages: (i) alignment of reads to the reference genome; (ii) empirical recalibration of quality scores; (iii) position-specific indexing of aligned reads; (iv) Bayesian genotype inference; and (v) application of filters (**Supplementary Fig. 1**). Sequence reads were mapped to the human reference genome (UCSC assembly hg18) using version 5.0.5 of the Burrows-Wheeler Aligner (BWA)[31] and version 0.1.7 of SAMtools[32]. Exact duplicate reads were removed using 'samtools rmdup' to avoid amplification biases. The raw quality scores were empirically recalibrated using the Genome Analysis Toolkit[33]. For each base in each individual genome, a maximum a posteriori genotype call was computed using a Bayesian algorithm for genotype inference (BSNP) that made use of aligned reads, basecall quality scores and mapping quality scores but avoided the use of the reference allele or previously identified variants. Orthologous sequences from the chimpanzee reference genome (panTro2) were extracted from genome-wide hg18-panTro2 alignments from UCSC.

Filtering. Our filters included both data-quality filters, designed to mitigate the effects of sequencing and alignment error, and comparative filters, designed to avoid the effects of natural selection, hypermutability or misalignment with chimpanzee. The data quality filters excluded sites with low coverage, sites adjacent to indels, sites in clusters of apparent SNPs or those in recent transposable elements or simple repeats. The comparative filters excluded sites in regions of poor human-chimpanzee synteny, recent segmental duplications, hypermutable CpG dinucleotides and sites either within or flanking protein-coding exons, noncoding RNAs and conserved noncoding elements. We ensured that our results were robust to parameters used to implement these filters (**Supplementary Note**).

Genotype validation. We compared our genotype calls with published calls for two individuals (Venter and NA12891 (ref. 34)) for whom both array-based and alterative sequence-based calls were available. For both individuals, we also considered genotype calls obtained by running the program MAQ[35] on our alignments. This approach allowed us to evaluate the performance of both the entire alignment pipeline and the genotype inference step alone (**Supplementary Figs. 2–6** and **Supplementary Table 2**). In addition, we computed key summary statistics (such as numbers of variant sites, heterozygosity and pairwise genomic distances) for the individual genomes in our set and checked that they were concordant with published estimates and with the assumption of a molecular clock (**Supplementary Figs. 7** and **8** and **Supplementary Note**).

G-PhoCS. The G-PhoCS program is derived from the MCMCcoal source code[20,21], but extensive changes to the code and sampling procedure were needed to accommodate migration and the use of unphased diploid genotypes (**Supplementary Note**). Some additional modifications allowed for reductions in running time. We generally ran the program with a burn-in of 100,000 iterations followed by 200,000 sampling iterations (**Supplementary Table 6**). Various analyses indicated that this was sufficient to allow for convergence of the Markov chain. Each run took about 30 days to complete on an Intel(R) Xeon(R) E5420, 2.50 GHz CPU (**Supplementary Table 7**).

Determining alignment blocks for analysis. We defined the 37,574 'neutral loci' by identifying contiguous intervals of 1,000 bp that passed our filters and then selecting a subset with a minimum inter-locus distance of 50,000 bp, ensuring that recombination hotspots (regions with recombination rates >10 cM/Mb[36]) fell between rather than within loci. The locus size and minimum interlocus distance were determined by an approximate calculation similar to one previously used[21]. We assumed a mean recombination rate of 10^{-8} per bp per generation, an average generation time of 25 years and minimum and maximum average genomic divergence times (among the humans) of 200,000 and 500,000 years, respectively. Thus, the expected number of recombinations on the lineages leading to two human chromosomes in a 1-kb interval is at most $2 \times 500{,}000 \times 10^{-8} \times 1000/25 = 0.4$, and the expected number in a 50-kb interval is at least $2 \times 200{,}000 \times 10^{-8} \times 50{,}000/25 = 8$. We conducted a series of validation experiments to ensure that our estimates were robust to modest amounts of intralocus recombination (**Supplementary Note**).

Model calibration. An estimate of a mutation-scaled version of the human-chimpanzee average genomic divergence time was obtained from the model parameters using the relationship, $\tau_{div} = \tau_{root} + \frac{1}{2}\theta_{root}$, where τ_{root} and θ_{root} represent the mutation-scaled human-chimpanzee speciation time and ancestral effective population size, respectively. This leads to an estimated mutation rate per year of $\mu = \tau_{div}/T_{div}$, which can be used to convert all other mutation-scaled divergence times to years ($T = \tau/\mu$). We assumed a generous range of T_{div} = 5.6–7.6 Mya, as previously suggested[24], based on the relative divergence levels of the chimpanzee and orangutan genomes from the human genome, an upper bound of 20 Mya for the orangutan divergence time and other constraints from the fossil record. We followed researchers from a previous study[2] in choosing a 'best guess' of T_{div} = 6.5 Mya. To obtain effective population sizes in numbers of diploid individuals (N) we used the relationship $\theta = 4N\mu g$, where g is the average generation time in years, and estimated N by $\theta/(4\mu g)$ (we assume g = 25 for human populations). We used τ_{div} for calibration because it is robustly estimated by G-PhoCS across a wide variety of different modeling assumptions, unlike τ_{root} and θ_{root}, which depend on the assumed model of mutation rate variation across loci. We obtained estimates of $\tau_{div} = 4.54 \times 10^{-3}$ across many different runs, with 95% confidence intervals of 4.45×10^{-3} to 4.63×10^{-3}.

Validation of parameter estimates. We performed a series of validation analyses using both simulated and real data to examine the influence on our estimates of several factors, including: (i) the choice of prior distributions; (ii) the mutation rate variation across loci; (iii) intralocus recombination; (iv) recent population expansions and bottlenecks; and (v) parameters and thresholds defining our data quality and comparative filters (**Supplementary Note**).

31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
33. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
34. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
35. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
36. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).