Bayesian inference of ancient human demography from individual genome sequences

Ilan Gronau, Melissa J. Hubisz, Brad Gulko, Charles G. Danko, Adam Siepel

SUPPLEMENTARY INFORMATION

Supplementary Figures

S1 Illustration of genotype inference pipeline ........................................... 3
S2 Comparison with array-based calls ......................................................... 4
S3 Power for genotype inference ............................................................... 5
S4 Histograms of read depths and posterior probabilities .............................. 6
S5 Genome-wide comparison for NA12891 .................................................. 7
S6 Genome-wide comparison for Venter ...................................................... 8
S7 Genome-wide comparison of seven genomes .......................................... 9
S8 Distance-based phylogenies ..................................................................... 10
S9 Parameter estimates without gene flow in simulated data ......................... 11
S10 Estimation accuracy in simulated data without gene flow ......................... 12
S11 Estimation accuracy in simulated data with gene flow ............................ 13
S12 Summary of parameter estimates for real data ....................................... 14
S13 Estimated migration rates for real data ................................................. 15

Supplementary Tables

S1 Summary of Source Datasets for Genomes ............................................... 17
S2 Sensitivity and False Call Rates for NA12891 and Venter (HapMap/Array-genotyped Sites) ......................................................... 18
S3 G-PhoCS Runs on Real Data ................................................................. 19
S4 Population Divergence Times Estimated in the G-PhoCS Analysis ................ 20
S5 Effective Population Sizes Estimated in the G-PhoCS Analysis ................... 21
S6 G-PhoCS Setup Parameters .................................................................... 22
S7 G-PhoCS Running Time ......................................................................... 23

Supplementary Note

S1 Genotyping Pipeline
S1.1 Pipeline Design .................................................................................. 24
S1.2 Filters .................................................................................................. 27
S1.2.1 Filtering Conventions ........................................................................ 27
S1.2.2 Data-quality Filters .......................................................................... 27
S1.2.3 Comparative Filters ......................................................................... 28
S1.3 BSNP Method for Genotype Inference ................................................. 28
S1.3.1 General Approach ........................................................................... 28
S1.3.2 Basic Model .................................................................................... 29
S1.3.3 Allowing for Correlated Errors ................................................. 31
S1.3.4 Maximum a Posteriori Genotype Calls ........................................... 33
S1.3.5 More General Prior Implemented in BSNP ....................................... 33

S2 Validation of Genotype Calls ........................................... 33
S2.1 Comparison with Arrays ........................................... 34
  S2.1.1 NA12891 ........................................... 34
  S2.1.2 Venter ........................................... 35
S2.2 Genome-wide Comparison with Alternative Sequence-based Methods ............ 36
  S2.2.1 NA12891 ........................................... 36
  S2.2.2 Venter ........................................... 36

S3 Application of Pipeline to Individual Genomes ................................ 37
S3.1 Numbers of Variants ........................................... 37
S3.2 Trees Based on Average Genomic Divergence ................................... 38

S4 G-PhoCS: A Generalized Phylogenetic Coalescent Sampler ................. 39
S4.1 A High-level Overview of MCMCcoal ....................................... 39
S4.2 Modeling Gene Flow in G-PhoCS ....................................... 43
S4.3 Accommodating Diploid Genomes ....................................... 45
S4.4 New Framework for Computing Genealogy Priors ............................ 48

S5 Applying G-PhoCS to Data ........................................... 49
S5.1 MCMC Setup ........................................... 49
S5.2 Migration Bands ........................................... 49
S5.3 Assessing Convergence and Mixing ....................................... 50
S5.4 Parameter Estimates and Credible Intervals ................................... 51
S5.5 Running Time ........................................... 51

S6 Simulation Study ........................................... 52
S6.1 Generating Simulated Data ........................................... 52
S6.2 Applying G-PhoCS to Simulated Data Sets ................................... 53
S6.3 Simulations without Gene Flow ........................................... 53
S6.4 Simulations with Gene Flow ........................................... 54

S7 G-PhoCS Analysis of Individual Human Genome Sequences ........... 54
S7.1 Neutral Loci ........................................... 54
S7.2 Parameter Calibration ........................................... 55
S7.3 Design of G-PhoCS Analysis of Individual Genome Sequences ........... 56
S7.4 Demographic Parameter Estimates ....................................... 56
S7.5 Yoruban/Bantu Relationship ........................................... 56
S7.6 Comparison with Published Estimates ....................................... 57

S8 Validation of G-PhoCS Analysis ........................................... 58
S8.1 Prior Distributions ........................................... 59
S8.2 Variation in Locus-specific Mutation Rates ................................... 61
S8.3 Intralocus Recombination ........................................... 64
S8.4 Complex Changes in Effective Population Size ............................ 67
S8.5 Filter Settings ........................................... 69
Figure S1: **Illustration of genotype inference pipeline.** Numbers correspond to steps as detailed in Section S1.1.
Figure S2: **Comparison with array-based calls.** Error rates for (A) NA12891 in HapMap sites and (B) Venter in sites at which array data is available. Separate results are shown for sites homozygous for the reference allele (HomRef), sites homozygous for a different allele (HomDif), and heterozygous sites (Het). Notice the difference in the vertical scales in (A) and (B).
Figure S3: **Power for genotype identification as a function of read depth.** Power is computed for (A) NA12891 at HapMap sites and for (B) Venter at sites for which array data is available. Separate curves are shown for HomRef, HomDif, and Het sites. Note that the low read-depth portion of (A) and the high read-depth portion of (B) are supported by sparse data.
Figure S4: **Histograms of read depths and posterior probabilities.** Read depths are shown in (A & C) and posterior probabilities in (B & D) for the NA12891 (top; A & B) and Venter (bottom; C & D) genomes. Results are shown for 3.35 million HapMap sites that pass our data quality filters (excluding the coverage filter), but the genome-wide distributions are similar. Panels B & D represent the posterior probability of the most probable allele, as reported by BSNP. The dashed red lines in A & C indicate the threshold used for the coverage filter. Notice that the distribution of posterior probabilities is strongly peaked near 1.0, even for the lower-coverage Venter genome.
Figure S5: **Comparison of variants detected for NA12891.** (A) Venn diagram of variants (either HomDifs or Hets) detected by MAQ, BSNP, and 1000G. (B–D) Venn diagrams of variants detected by all three pairs of methods, distinguishing between HomDifs and Hets.
Figure S6: Comparison of variants detected for Venter. (A) Venn diagram of variants (either HomDifs or Hets) detected by MAQ, BSNP, and in the original paper by Levy et al. (2007). (B–D) Venn diagrams of variants detected by all three pairs of methods, distinguishing between HomDifs and Hets.
Figure S7: **Genome-wide comparison of seven genomes.** Rates of occurrence of HomDif and Het variants in the seven individual genomes considered, for both the published SNP sets and the results of our genotype inference pipeline (denoted “BSNP”). Rates are defined as the number of sites of each type (HomDif or Het) divided by the total number of sites considered. Only sites that passed the data quality filters for all genomes were included. The Het rates can be interpreted as estimates of heterozygosity per site ($\pi$). Notice the improved consistency in heterozygosity among members of the same population group for BSNP relative to the published SNP sets. The Watson genome was excluded from our final analysis, due to low coverage.
Figure S8: **Distance-based phylogenies.** Phylogenies estimated by the (A) neighbor-joining and (B) UPGMA methods for the seven genomes and a chimpanzee outgroup. On the right, scale bars are shown in substitutions per site and absolute time, assuming a 6.5 Mya human-chimpanzee divergence time. The relatively low implied mutation rate \((46 \times 10^{-4} \div 6.5 \times 10^6 = 7.1 \times 10^{-10}\) substitutions/site/year) reflects the exclusion of CpG sites from the analysis. The divergence time of 825 kya estimated for the Neanderthal by Green et al. (2010) is shown for comparison. Note that these estimates represent average genomic divergence times, not population divergence times (i.e., they are sums of population divergence times and average coalescent times). The Watson genome was excluded from our final analysis, due to low coverage.
Figure S9: Parameter estimates for simulated data sets without gene flow. Each bar describes the mean and 95% CI for a specific parameter in a certain run. Mutation-scaled estimates are shown in units of $10^{-4}$. The 36 runs correspond to two G-PhoCS replicates for each of the 18 data sets simulated without gene flow (Supplementary Note; Section S6.3). They are shown as a sequence of pairs of replicates, in the order of the ids. All estimates reflect unphased data and phasing integration. The values assigned to the parameters in the simulations determine the baseline values (red labels, solid horizontal lines), and dashed lines indicate ±10% from the baseline. Eurasian parameters are shown in the left column and African parameters in the right column.
Figure S10: **Estimation accuracy in simulated data without gene flow.** Eighteen simulated data sets without gene flow are considered. Parameter estimates were obtained using three different phasing strategies: “phasing integration” (red), “correct phasing” (gold), and “random phasing” (green). A single run per data set was used for each strategy. (A) Estimates of the San divergence time ($\tau_{KHEXS}$) and African-Eurasian divergence time ($\tau_{KHEX}$) summarized across these experiments. Each box plot describes the distribution of estimates in six $G$-PhoCS runs generated under the same “true” value for the parameter of interest (indicated by horizontal lines). (B,C) Estimation accuracy for the demographic parameters that were held constant across all simulations. Each box plot describes the distribution of estimates across all 18 $G$-PhoCS runs for a specific phasing strategy, normalized by their true value used in simulation. Demographic parameters are partitioned into African (B) and Eurasian (C) parameters.
Figure S11: **Estimation accuracy in simulated data with gene flow.** Forty eight simulated data sets with different levels of gene flow are considered. (A,B) Comparison of estimates obtained assuming no migration (blue) vs. those obtained allowing for an S→X migration band (red). Each box plot represents the estimates (normalized by their true value) obtained from 12 data sets, all simulated with the same migration rate. The migration rates labeled 0, 1, 2, and 3 correspond to total migration rates of 0.0, 0.1, 0.2, and 0.3, respectively. Demographic parameters are partitioned into African (A) and Eurasian (B) parameters. The parameters $\theta_X$ and $\tau_{\text{KHEXS}}$ (marked with dashed boxes) show the most pronounced bias when migration is not modeled in *G-PhoCS*. (C) Estimates of expected number of migrants per generation (given by $m_{SX} \times \theta_X$) and the total migration rate (given by $m_{SX} \times \tau_{\text{KHEXS}}$), shown in comparison to the true values of these quantities (horizontal lines).
Figure S12: Parameter estimates for the 32 G-PhoCS runs on individual human genome sequences. Each bar describes the mean and 95% CI for a specific parameter in a certain run. The y-axes show both the raw ($\times 10^4$; left) and calibrated (right) estimates, with $T_{\text{div}} = 6.5$ Mya. The calibrated divergence times are given in kya, and the calibrated effective population sizes are given in thousands and assume an average generation time of 25 years. Estimates are ordered by run id, as specified in Table S3. The estimates observed in run #16 (with the Yoruban sample and migration band S$\rightarrow$Y) determine the baseline values (red/dark blue labels, solid horizontal lines), and dashed lines indicate $\pm 10\%$ from the baseline. Eurasian parameters are shown in the left column and African parameters in the right column.
Figure S13: Estimated migration rates for the 28/32 *G-PhoCS* runs in which gene flow was modeled. Each bar describes the posterior mean and 95% CI for the “total migration rate” in a specific run. The “total migration rate” is the migration rate per generation multiplied by the number of generations spanning the migration band (e.g., $m_{SB} \times \tau_{KHEB}$, for the S→B band). This calibration-free statistic provides a reasonable measure of the extent to which migration is detected along each band. The runs are labeled by migration band and ordered by run-id, as specified in Table S3.
| Genome   | Read Src$^a$ | Identifiers                                                                 | Reads$^b$ | Nucs$^c$ | MNucs$^d$ | MPos$^e$ | SNP Src$^f$                                      |
|----------|--------------|-----------------------------------------------------------------------------|-----------|----------|-----------|----------|-------------------------------------------------|
| Venter   | NCBI$^g$     | —                                                                           | 0.032     | 27.900   | 23.612    | 2.827    | ftp://ftp.jcvi.org/pub/huref/                    |
| NA18507  | NCBI SRA     | SRX000600–603, SRX001539–1540                                              | 4.056     | 146.389  | 114.599   | 2.787    | ftp://ftp.sanger.ac.uk/pub/rd/NA18507/         |
| YH       | EBI$^h$      | ERA000005                                                                  | 3.334     | 117.059  | 70.511    | 2.776    | http://yh.genomics.org.cn/download.jsp          |
| SJK      | NCBI SRA     | SRX002757–2761                                                             | 1.730     | 81.515   | 54.972    | 2.797    | ftp://ftp.kobic.kr/pub/                        |
|          |              |                                                                             |           |          |           |          | KOBIC-KoreanGenome/                            |
|          |              |                                                                             |           |          |           |          | http://usegalaxy.org                           |
|          |              |                                                                             |           |          |           |          | http://usegalaxy.org                           |
|          |              |                                                                             |           |          |           |          | http://usegalaxy.org                           |
|          |              |                                                                             |           |          |           |          | ftp://jimwatsonsequence.cshl.edu/              |
|          |              |                                                                             |           |          |           |          | jimwatsonsequence/                             |
|          |              |                                                                             |           |          |           |          | ftp://ftp-trace.ncbi.nih.gov/1000genomes/      |
|          |              |                                                                             |           |          |           |          | ftp/pilot_data/release/2009_12/pilot2/         |
|          |              |                                                                             |           |          |           |          |                                               |
|          |              |                                                                             |           |          |           |          |                                               |

$^a$Source of raw sequence reads: NCBI = National Center for Biotechnology Information; SRA = Sequence Read Archive; EBI = European Bioinformatics Institute; PSU = Penn State University.

$^b$Total number of reads in initial data set (billions).

$^c$Total number of nucleotides in sequence reads (billions). Excludes “N”s.

$^d$Number of nucleotides that mapped uniquely to the reference genome, excluding “N”s (billions).

$^e$Number of genomic positions with at least one uniquely aligned (non-“N”) nucleotide (billions).

$^f$Source of “published SNPs” used for comparison with our genotype calls.

$^g$Downloaded from ftp://ftp.ncbi.nih.gov/pub/TraceDB/PersonalGenomics/Venter.

$^h$Downloaded from ftp://ftp.era-xml.ebi.ac.uk/vol1/ERA000/ERA000005/fastq.

$^i$At the time of our analysis (early 2010), several of the data files available from NCBI were corrupt, so these data were instead obtained by request from Webb Miller’s group at Penn State University.

$^j$Excluded from final analysis, due to low coverage.

$^k$Downloaded from ftp://ftp.ncbi.nih.gov/pub/TraceDB/PersonalGenomics/Watson. This data set is also available from the SRA.

$^l$Sequenced by the The 1000 Genomes Project Consortium (2010) and used for validation only.
Table S2: Variant Sensitivity and False Call Rates for NA12891 (HapMap Sites) and for Venter (Array-genotyped Sites).

| Genome   | Method | Sens.\(^a\) | FCR\(^b\) | Sens.\(^c\) | FCR\(^d\) | Sens.\(^e\) | FCR\(^f\) |
|----------|--------|--------------|-----------|--------------|-----------|--------------|-----------|
| NA12891  | BSNP   | 0.9887       | 0.0052    | 0.9802       | 0.0115    | 0.9894       | 0.0076    |
|          | 1000G  | 0.9686       | 0.0059    | 0.9835       | 0.0139    | 0.9436       | 0.0031    |
|          | MAQ    | 0.9876       | 0.0054    | 0.9767       | 0.0118    | 0.9894       | 0.0091    |
| Venter   | BSNP   | 0.9564       | 0.0013    | 0.8485       | 0.0038    | 0.9982       | 0.0898    |
|          | Levy et al.\(^g\) | 0.8964 | 0.0012 | 0.8365 | 0.0035 | 0.9553 | 0.0190 |
|          | MAQ    | 0.9481       | 0.0013    | 0.8196       | 0.0038    | 0.9982       | 0.1053    |

\(^a\)Fraction of variant sites (Hets or HomDifs), according to HapMap/array data, that are correctly identified as variant sites (without regard to the type of variant).

\(^b\)Fraction of predicted variant sites that are actually HomRef sites, according to HapMap/array data.

\(^c\)Fraction of Het sites, according to HapMap/array data, that are correctly identified as Het sites (without regard to the type of Het, although the type is almost always correct).

\(^d\)Fraction of predicted Het sites that are actually HomRef or HomDif sites, according to HapMap/array data.

\(^e\)Fraction of HomDif sites, according to HapMap/array data, that are correctly identified as HomDif sites (without regard to the type of HomDif, although the type is almost always correct).

\(^f\)Fraction of predicted HomDif sites that are actually HomRef or Het sites, according to HapMap/array data.

\(^g\)Genotypes published by Levy et al. (2007).
Table S3: G-PhoCS Runs on Real Data

|       | X = Yoruban |       | X = Bantu |
|-------|-------------|-------|-----------|
| id    | migration band $^a$ | id    | migration band $^a$ |
| 1, 2  | none        | 17, 18| none      |
| 3, 4  | S $\rightarrow$ H | 19, 20| S $\rightarrow$ H |
| 5, 6  | S $\rightarrow$ K | 21, 22| S $\rightarrow$ K |
| 7, 8  | S $\rightarrow$ E | 23, 24| S $\rightarrow$ E |
| 9, 10 | Y $\rightarrow$ H | 25, 26| B $\rightarrow$ H |
| 11, 12| Y $\rightarrow$ K | 27, 28| B $\rightarrow$ K |
| 13, 14| Y $\rightarrow$ E | 29, 30| B $\rightarrow$ E |
| 15, 16| S $\rightarrow$ Y | 31, 32| S $\rightarrow$ B |

$^a$Each run assumes (at most) one migration band from an African population (S, Y, or B) to a non-African population (H, K, or E).
Table S4: Population Divergence Times Estimated in the *G-PhoCS* Analysis

|       | X = Yoruban |       |       | X = Bantu |       |       |
|-------|-------------|-------|-------|-----------|-------|-------|
|       | no migration | mig S-Y | no migration | mig S-B |
| τKH   | 0.03 | 0.04 | 0.06 | 0.05 | 0.02 | 0.01 | 0.01 | 0.03 |
|        | (0 – 0.06) | (0.01 – 0.07) | (0.03 – 0.09) | (0 – 0.11) | (0.01 – 0.05) | (0 – 0.02) | (0.01 – 0.04) |
|        | 4 | 5 | 8 | 8 | 3 | 1 | 1 | 1 |
|        | (0 – 9) | (2 – 10) | (4 – 12) | (1 – 15) | (1 – 7) | (0 – 3) | (0 – 2) | (2 – 6) |
| τKHE  | 0.28 | 0.27 | 0.25 | 0.25 | 0.28 | 0.29 | 0.27 | 0.27 |
|        | (0.26 – 0.29) | (0.25 – 0.28) | (0.24 – 0.27) | (0.22 – 0.27) | (0.26 – 0.3) | (0.28 – 0.31) | (0.27 – 0.28) | (0.25 – 0.28) |
|        | 40 | 38 | 36 | 35 | 40 | 42 | 39 | 38 |
|        | (37 – 41) | (36 – 41) | (34 – 38) | (31 – 38) | (38 – 43) | (40 – 44) | (38 – 40) | (36 – 40) |
| τKHEX | 0.34 | 0.34 | 0.33 | 0.32 | 0.37 | 0.39 | 0.37 | 0.38 |
|        | (0.33 – 0.35) | (0.33 – 0.36) | (0.31 – 0.34) | (0.31 – 0.33) | (0.36 – 0.4) | (0.37 – 0.4) | (0.35 – 0.38) | (0.36 – 0.42) |
|        | 49 | 49 | 47 | 46 | 53 | 55 | 53 | 55 |
|        | (48 – 50) | (47 – 51) | (44 – 49) | (45 – 47) | (52 – 57) | (53 – 58) | (50 – 55) | (51 – 59) |
| τKHEXS | 0.87 | 0.87 | 0.91 | 0.92 | 0.84 | 0.84 | 0.9 | 0.9 |
|        | (0.85 – 0.89) | (0.85 – 0.9) | (0.89 – 0.94) | (0.89 – 0.94) | (0.82 – 0.87) | (0.82 – 0.86) | (0.88 – 0.93) | (0.87 – 0.91) |
|        | 125 | 125 | 131 | 131 | 121 | 120 | 129 | 129 |
|        | (121 – 128) | (122 – 128) | (127 – 135) | (128 – 135) | (117 – 124) | (117 – 124) | (126 – 133) | (125 – 133) |
| τroot | 45.25 | 44.97 | 45 | 45.02 | 44.67 | 44.78 | 45.02 | 44.96 |
|        | (45.05 – 45.42) | (44.72 – 45.21) | (44.75 – 45.23) | (44.77 – 45.26) | (44.26 – 45.15) | (44.5 – 45.07) | (44.83 – 45.22) | (44.74 – 45.18) |
|        | 6478 | 6439 | 6442 | 6446 | 6395 | 6412 | 6446 | 6437 |
|        | (6450 – 6504) | (6402 – 6472) | (6407 – 6476) | (6410 – 6480) | (6337 – 6465) | (6371 – 6453) | (6419 – 6474) | (6406 – 6468) |
| Tdiv  | 45.4 | 45.4 | 45.4 | 45.4 | 45.4 | 45.4 | 45.39 | 45.39 |
|        | (45.25 – 45.55) | (45.24 – 45.55) | (45.25 – 45.55) | (45.25 – 45.55) | (45.25 – 45.56) | (45.26 – 45.56) | (45.24 – 45.54) | (45.23 – 45.54) |
|        | 6500 | 6501 | 6499 | 6500 | 6501 | 6501 | 6498 | 6498 |
|        | (6479 – 6522) | (6478 – 6522) | (6479 – 6522) | (6478 – 6521) | (6479 – 6521) | (6479 – 6523) | (6477 – 6520) | (6476 – 6520) |

*a*Estimates of population divergence times (with 95% CIs) obtained in the analysis of the human individuals in *G-PhoCS* runs 1,2,15,16, 17,18,31,32, as described in Table S3. Raw estimates (top row) are scaled by the mutation rate × 10^4, and calibrated values (bottom row) are given in kya and are obtained by assuming an average human/chimpanzee divergence of T_{div} = 6.5 Mya.

20
### Table S5: Effective Population Sizes Estimated in the *G-PhoCS* Analysis

|          | X = Yoruban $^a$ |          | X = Bantu $^a$ |
|----------|------------------|----------|----------------|
|          | no migration     | mig S-Y  | no migration   | mig S-B |
| $\theta_K$ |                  |          |                |        |
|          | 2.62             | 2.54     | 2.42           | 2.38   |
|          | (2.37 – 2.85)    | (2.28 – 2.8) | (2.21 – 2.63) | (2.03 – 2.66) |
|          | 3800             | 3600     | 3500           | 3400   |
|          | (3400 – 4100)    | (3300 – 4000) | (3200 – 3800) | (2900 – 3800) |
| $\theta_H$ | 2.07             | 2.9      | 4.06           | 3.74   |
|          | (0.24 – 4.64)    | (0.74 – 5.04) | (1.91 – 6.38) | (0.26 – 7.61) |
|          | 3000             | 4100     | 5800           | 5400   |
|          | (900 – 11500)    | (2600 – 12500) | (4800 – 15200) | (1000 – 16700) |
| $\theta_E$ | 4.07             | 5.24     | 6.78           | 6.1    |
|          | (0.61 – 8.05)    | (1.84 – 8.75) | (3.39 – 10.6) | (0.69 – 11.7) |
|          | 5800             | 7500     | 9700           | 8700   |
|          | (900 – 11500)    | (2600 – 12500) | (4800 – 15200) | (1000 – 16700) |
| $\theta_X$ | 20.45            | 20.31    | 12.42          | 12.13  |
|          | (16.66 – 24.6)   | (16.36 – 24.37) | (9.98 – 15.19) | (7.95 – 14.79) |
|          | 29300            | 29100    | 17800          | 17400  |
|          | (23900 – 35200)  | (23400 – 34900) | (14300 – 21700) | (14000 – 21200) |
| $\theta_S$ | 14.56            | 14.58    | 14.65          | 14.68  |
|          | (13.36 – 15.82)  | (13.35 – 15.78) | (13.48 – 15.89) | (13.47 – 15.92) |
|          | 20900            | 20900    | 21000          | 21000  |
|          | (19100 – 22600)  | (19100 – 22600) | (19300 – 22800) | (19300 – 22800) |
| $\theta_{KH}$ | 3.14             | 2.86     | 2.45           | 2.43   |
|          | (2.45 – 3.64)    | (2.31 – 3.42) | (2.05 – 2.88) | (1.36 – 3.34) |
|          | 4500             | 4100     | 3500           | 3500   |
|          | (3500 – 5200)    | (3300 – 4900) | (2900 – 4100) | (2000 – 4800) |
| $\theta_{KHE}$ | 0.64             | 0.71     | 0.82           | 0.82   |
|          | (0.5 – 0.86)     | (0.48 – 0.99) | (0.62 – 1.06) | (0.61 – 1.29) |
|          | 900              | 1000     | 1200           | 1200   |
|          | (700 – 1200)     | (700 – 1400) | (900 – 1500) | (900 – 1800) |
| $\theta_{KHEX}$ | 9.82             | 9.88     | 8.02           | 8.16   |
|          | (9.19 – 10.48)   | (9.28 – 10.56) | (7.28 – 8.7) | (7.37 – 8.88) |
|          | 14100            | 14100    | 11500          | 11700  |
|          | (13200 – 15000)  | (13300 – 15100) | (10400 – 12500) | (10600 – 12700) |
| $\theta_{KHEXS}$ | 6.12             | 6.12     | 6.08           | 6.07   |
|          | (6.03 – 6.2)     | (6.04 – 6.2) | (5.99 – 6.16) | (5.99 – 6.15) |
|          | 8800             | 8800     | 8700           | 8700   |
|          | (8600 – 8900)    | (8600 – 8900) | (8600 – 8800) | (8600 – 8800) |
| $\theta_{root}$ | 0.31             | 0.86     | 0.8            | 0.75   |
|          | (0.14 – 0.55)    | (0.54 – 1.26) | (0.45 – 1.12) | (0.42 – 1.14) |
|          | 400              | 1200     | 1100           | 1100   |
|          | (200 – 800)      | (800 – 1800) | (600 – 1600) | (600 – 1600) |

$^a$Estimates of effective population sizes (with 95% CIs) obtained in the analysis of the human individuals in *G-PhoCS* runs 1,2,15,16,17,18,31,32, as described in Table S3. Raw estimates (top row) are scaled by the mutation rate $\times 10^4$, and calibrated values (bottom row) in number of diploid individuals are obtained by assuming an average human/chimpanzee divergence of $T_{div} = 6.5$ Mya and an average generation time of 25 years.
| Parameter                                      | Simulation Study | Data Analysis<sup>a</sup> |
|-----------------------------------------------|------------------|---------------------------|
| burn-in iterations                            | 100,000          |                           |
| sample iterations                             | 200,000          | 150,000–200,000           |
| sample-skip                                   | 10               |                           |
| prior for scaled locus-specific mutation rate  |                  | Dirichlet(α = 1)          |
| prior for all θ parameters                    | Γ(α = 1, β = 10000) |                           |
| prior for all m parameters                    | Γ(α = 0.002, β = 0.00001) |                     |
| prior for τ<sub> KH</sub>                     | Γ(α = 1, β = 20000) | Γ(α = 1, β = 30000)      |
| prior for τ<sub> KHE </sub>                   | Γ(α = 1, β = 20000) | Γ(α = 1, β = 30000)      |
| prior for τ<sub> KHEY </sub>                  | Γ(α = 1, β = 20000) | Γ(α = 1, β = 25000)      |
| prior for τ<sub> KHEYS </sub>                 | Γ(α = 1, β = 10000) |                           |
| prior for τ<sub> root </sub>                  | Γ(α = 1, β = 1000) |                           |
| fine-tune for coalescent time update           | 0.3              |                           |
| fine-tune for migration time update            | 0.3              |                           |
| fine-tune for θ update                        | 0.04             |                           |
| fine-tune for m update                        | 0.02             |                           |
| fine-tune for τ update                        | 0.0000008        |                           |
| fine-tune for locus-rate update               |                  | 0.5                       |
| fine-tune for scaling update                  | 0.003            |                           |

<sup>a</sup>Values used in data analysis are indicated only when different from ones used in simulated setup.
| Data Set                     | Number of Loci | Patterns<sup>a</sup> | Iterations/Hr<sup>b</sup> | Time<sup>c</sup> |
|-----------------------------|----------------|----------------------|---------------------------|------------------|
| simulated, unphased<sup>d</sup> | 20,000         | 4.5                  | 1100-1200                 | 11 days          |
| simulated, phased<sup>e</sup>  | 20,000         | 3.6                  | 1200-1300                 | 10 days          |
| genomic, with NA18507<sup>f</sup> | 37,574         | 6.86                 | 450                       | 28 days          |
| genomic, with ABT<sup>g</sup>    | 37,574         | 7.89                 | 400                       | 31 days          |

<sup>a</sup>Average number of phased site patterns per locus (see Section S4.3).

<sup>b</sup>Number of MCMC iterations per hour on an Intel(R) Xeon(R) E5420, 2.50 GHz CPU.

<sup>c</sup>Total running time required to obtain all 300,000 MCMC iterations.

<sup>d</sup>Statistics for these are averaged across all different simulated datasets.

<sup>e</sup>Statistics for these are averaged across all different simulated datasets, assuming accurate phasing.

<sup>f</sup>Statistics for the 5 human samples, excluding ABT, and chimpanzee sequence.

<sup>g</sup>Statistics for the 5 human samples, excluding NA18507, and chimpanzee sequence.
Supplementary Note

S1 Genotyping Pipeline

S1.1 Pipeline Design

Our pipeline for genotype inference consists of five major stages: (1) alignment of reads to the reference genome; (2) empirical recalibration of quality scores; (3) position-specific indexing of aligned reads (in a “pileup” file); (4) Bayesian genotype inference; and (5) application of filters (Figure S1). Below we briefly summarize each stage in the pipeline. In subsequent sections, we describe the filters and genotype inference procedure in more detail.

This pipeline was applied to the raw reads for each genome separately (see Table S1), except that the final filtering step considered all genomes simultaneously. The human reference genome (UCSC assembly hg18, equivalent to NCBI build 36) was used for read mapping and sequence annotations but it was not directly considered in genotype inference. Notice that the Watson genome (Wheeler et al., 2008) was considered along with the six individual human genomes that are described in the main text. The Watson genome was included in an early version of our analysis, but excluded from the final version, due to concerns about high false negative rates for variant detection stemming from low sequencing redundancy (see Section S3.1). Nevertheless, we have retained this genome in the supplementary sections describing our genotyping methods and validation studies.

1. **Alignment.** Sequence reads were mapped to the hg18 reference genome using version 5.0.5 of BWA (Li et al., 2008) and version 0.1.7 of SAMTools (Li et al., 2009a). Briefly, the reference genome was indexed using “bwa index -a bwtsw”, fastq-formatted reads were aligned using either “bwa aln” (for short reads) or “bwa dbwtsw -s 4 -z 2” (for Venter and Watson), and sorted, merged BAM files were generated using “bwa samse” (for single reads) or “bwa sampe” (for paired-end reads) followed by “samtools import”, “samtools sort”, and “samtools merge”. Finally, exact duplicate reads were removed using “samtools rmdup” to avoid amplification biases. Prior to the alignment step, the reads for YH were converted to the Sanger quality scale using “maq sol2sanger” and those for KB1 were converted using “maq ill2sanger.” The sequences and quality scores for Watson and Venter were originally in separate files and were converted to fastq format using a custom script. This involved adding 33 to all quality scores and truncating them at 126 to match standard conventions. The downloaded files for NA18507, SJK, and NA12891 required no preprocessing. Multithreading was used in the alignment step (-t option to bwa).

The SOLiD data for ABT required a slightly modified procedure, due to their representation in “colorscape” rather than sequence space. In this case, the reference genome was indexed using the “-c” option to “bwa index”, the raw *.csfasta and *.qual files were merged using solid2fastq.pl (available with BWA), and reads were aligned using the command “bwa aln -t8 -c -n 4 -k 4 -l 24”. (This allows up to four differences in color from the reference genome, which is needed to accommodate up to two nucleotide mismatches per read; the default settings allowed only one mismatch with our data, which led to a clear bias against variant detection in ABT.) Version 5.0.7 of BWA was needed in this case, and it required a slight modification to correctly handle gaps in paired-end data. (A bug fix has been submitted to the authors.) Minor modifications to solid2fastq.pl were also required, to correct problems in the handling of quality scores of −1 (provided for basecalls of “N”), with comments, and with trimming of first colorspace values. The single and paired end reads for ABT were separated for alignment and BAM generation and then were combined during the merging step.

Notice that reads with non-unique mappings to the genome were not explicitly removed at this stage.
in the pipeline. BWA assigned these reads mapping quality scores of zero, which caused them to be ignored in subsequent stages of the pipeline.

2. **Recalibration.** The raw quality scores were next empirically recalibrated using a method implemented in the Genome Analysis Toolkit (http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit). We used version 1.0.3362 (dated May 17, 2010) of the GenomeAnalysisTK.jar bytecode distribution (from ftp://ftp.broadinstitute.org/pub/gsa/GenomeAnalysisTK/), and executed it with version 1.6.0.07 of the Java Development Kit.

The recalibration procedure involves two steps. First, statistics on apparently miscalled bases are gathered from the initial BAM files, conditional on specified covariates, excluding sites likely to be polymorphic. We excluded all positions at which a preliminary run of our genotype caller (based on uncalibrated data) indicated a non-hg18 genotype with \( \geq 50\% \) probability. As covariates we used the raw quality scores, the machine cycle number (read position), and the dinucleotide context. We also excluded reads with mapping quality scores of zero and positions with very large numbers of reads (> 1000). The full command for this step was of the form:

```
java -Xmx16g -jar GenomeAnalysisTK.jar -l INFO -R hg18.fasta -B mask,VCF,<VAR_in> -I <BAM_in> --max_reads_at_locus 1000 -T CountCovariates -cov QualityScoreCovariate -cov CycleCovariate -cov DinucCovariate -recalFile <RECAL_out> --filterZeroMappingQualityReads --default_platform <PLATFORM>
```

Here <BAM_in> specified the input BAM file and <VAR_in> specified a VCF file listing the variant positions to be excluded. The argument <PLATFORM> was “illumina” for all genomes except Watson, in which case “454” was used. (There is a “solid” option, but it is designed for colorspace data, and at this stage our ABT data had already been converted to nucleotide space. There is no designated option for Sanger reads.) We omitted the cycle number covariate (-cov CycleCovariate) for Watson and Venter because the long read lengths and low sequencing redundancy led to sparse data per read position.

Based on these statistics, empirical error probabilities are computed for each aligned base in a BAM file, using the same covariates considered in the first step. From these error probabilities, recalibrated quality scores are computed as \(-10 \log_{10} P(\text{error})\). The commands for this step were of the form:

```
java -Xmx16g -jar GenomeAnalysisTK.jar -l INFO -R hg18.fasta -I <BAM_in>.bam -T TableRecalibration -outputBam <BAM_out>.bam -recalFile <RECAL_in> --filterZeroMappingQualityReads --default_platform <PLATFORM>
```

Here <RECAL_in> is the file produced in the first step (specified by <RECAL_out> and <BAM_out> specifies a new BAM file, in which the original quality scores are to be replaced with recalibrated scores.

3. **Indexing.** The “pileup” format used by the SAMtools package summarizes, at each genomic position, all aligned bases and the corresponding basecall quality scores and mapping quality scores. A pileup file was produced for each genome using the command “samtools pileup -csf”.

4. **Genotype inference.** Genotype distributions were generated by running version 2.05a of the BSNP program (see detailed methods in Section S1.3) on each pileup file. The full BSNP command was of the form:
Here -i <PILEUP_in> is the input file, <SNP_out> specifies an output file listing an inferred distribution of genotypes and most likely genotype at each position, <AUX_out> specifies an output file with auxiliary information, -ph is the prior for the G+C content, -ka is the prior for the transition/transversion rate ratio -p0 is the prior for heterozygotes, and -th is the value of the $\theta$ parameter for correlated errors (see Section S1.3). <SUM_out> specifies an output file containing a summary of the arguments used and the induced prior over genotypes.

The <SNP_out> file includes a floating point number (a probability) for each of ten possible genotypes at each genomic position, and would be unmanageably large without some means of reduction. Furthermore, much of this data is uninformative, because, each genome is homozygous for the reference allele (“HomRef”) with high probability at the vast majority of genomic positions. Therefore, BSNP only writes a line to the <SNP_out> file for positions at which the genome in question has at least some minimal probability of a genotype other than a HomRef. This probability threshold is determined by the -mp option, and was set to 0.001 in our analysis. As an additional safeguard, BSNP will also write output if MAQ predicts a non-HomRef allele with some minimal quality score (selected to be 1; -mq option). The positions of the “implicit HomRef” genotypes (those at which these thresholds are not exceeded), along with their depths of coverage, are written to the <AUX_out> file. The remaining arguments to BSNP are more technical in nature and are detailed in the BSNP documentation.

5. **Filtering and extraction.** The genotype inferences produced by BSNP (i.e., the full posterior distributions and most likely genotypes) were stored in a relational database, with one table per genome and one row per genomic position. The human genotypes were augmented with genotypes from the chimpanzee reference genome (panTro2), extracted from genome-wide hg18-panTro2 alignments from UC Santa Cruz (http://genome.ucsc.edu). Genomic intervals identified by various automatic filters were stored in the same database, with one table per filter and one row per interval. The filter tables contained optional fields for a “type” (e.g., transposon family), “score” (e.g., divergence from consensus repeat), or “genome” (particular genome to which filter applied) associated with each interval. As detailed in Section S1.2, both “data quality” and “comparative” filters were applied, some of which were based on the sequences and alignments themselves, some of which used comparative genomic data (for example, from the chimpanzee), and some of which used auxiliary genome annotations (genes, repeats, etc.). All genotypes and filters were defined in reference genome (hg18) coordinates. Only autosomal (chromosomes 1–22) positions were considered in all analyses.

Organizing the genotype calls and filters in this way allowed for fast and flexible extraction of genotype data and alignments for genomic positions passing any designated set of filters. The alignments analyzed by G-PhoCS were extracted using a perl script (extractAlign.pl) that accepted as input a file specifying the coordinates of “neutral loci” (see Section S7.1), a list of genotype tables (corresponding to genomes) and a list of filter tables (with optional “score” thresholds and “type” designations). Other scripts were used to extract various summary statistics of interest for positions passing designated filters, such as the rates at which non-HomRef genotypes occurred, the pairwise divergence between genomes, and the rates differences with respect to HapMap genotypes. These scripts were designed to fill in the “implicit HomRef” genotypes, as needed. They made use of programs from the UCSC Genome Browser libraries (http://hgdownload.cse.ucsc.edu/downloads.html), such as overlapSelect and featureBits, to efficiently find intersections between genomic intervals. Our genotype tables, filter tables, and extraction scripts are available by request.
S1.2 Filters

S1.2.1 Filtering Conventions

Drawing on several previous studies (Patterson et al., 2006; Caswell et al., 2008; Burgess and Yang, 2008), we developed a series of conservative data-quality filters to mitigate the effects of sequencing and alignment error in our analysis. These filters were applied for all downstream analyses. In addition, we developed a set of comparative filters for use in our evolutionary analyses. These filters were designed to eliminate sites under selection, hypermutable sites, and sites of uncertain orthology with chimpanzee. The comparative filters were used for the G-PhoCS analysis and in the estimation of distance-based phylogenies (Figure S8) but were not used when collecting summary statistics for individual genomes or in the validation of our genotype calls.

As described above, each filter identified a set of genomic intervals (in hg18 coordinates) that could be excluded in subsequent analyses. The filters could optionally define “type” or “score” information along with each interval, so that subsets of intervals (for example, exceeding a designated score threshold) could be used if desired. In general, each filter was implemented by a stand-alone program that output a series of genomic intervals, which could then be loaded in the database. In some cases, filters were applied in downstream analyses in a “genome-specific” manner, so that they eliminated data only from particular genomes, while in others they were treated as “universal”, so that data from all genomes was eliminated at the designated positions. Moreover, filters could be applied in a “hard” manner, causing sites to be completely excluded from consideration, or in a “soft” manner, causing genotype calls to be masked and treated as missing data (typically by replacing them by “N”s). Genome-specific filters were always soft, while universal filters were typically hard, but in certain cases—as with short, frequently occurring intervals such as those corresponding to CpGs—they were applied in a soft manner. We were careful to treat all filters as universal when comparing genomes (as in Figures S7 and S8) so that exactly the same set of sites was considered in all cases.

S1.2.2 Data-quality Filters

The following data-quality filters were implemented. We indicate for each one its mode of application in the G-PhoCS analysis (genome-specific/universal, hard/soft).

1. **Clustered SNPs.** (universal, hard) Cases of two or more variant sites in a single genome within a 15 bp window were masked. (Tight clusters of SNPs are likely to reflect misalignment or sequencing error.)

2. **Simple repeats.** (universal, soft) Regions in the genome identified as simple repeats by Tandem Repeats Finder (TRF) (Benson, 1999), according to the “Simple Repeats” track in the UCSC Genome Browser, were excluded. (These regions are likely to be enriched for misalignment, and may also exhibit unusually high evolutionary rates.)

3. **Recent transposable elements.** (universal, hard) Regions annotated by RepeatMasker (http://www.repeatmasker.org) as transposable elements with ≤20% divergence from their consensus sequences were excluded. (These regions are likely to be enriched for misalignment.)

4. **Indels.** (genome-specific, soft) Positions within 3 bp of any genomic coordinate at which ≥20% of aligned reads contain an indel were masked. (These regions are likely to be enriched for misaligned bases, due to local optimizations in alignment gap placement. The reason for the threshold of 20% is that sporadic indels due to sequencing error are of less concern than true [possibly heterozygous] indels, which will tend to produce systematic misalignments that are shared across reads, leading to erroneous genotype calls.)

27
5. **Effective Coverage.** (genome-specific, soft) Regions with effective coverage <5 were excluded. For the Watson genome, which has the lowest sequencing redundancy, we used a cutoff of <4 in order to maintain reasonable coverage across the genome. The effective coverage at each position $j$ was calculated as:

$$C(j) = \sum_{i=1}^{\text{depth}(j)} \left(1 - 10^{-m_{ij}/10}\right) \times \left(1 - 10^{-q_{ij}/10}\right)$$

where $m_{ij}$ and $q_{ij}$ refer to the mapping and base-call quality scores for read $i$ at position $j$. Regions with low effective coverage will tend to have high false negative rates in variant detection, particularly at heterozygous positions.

### S1.2.3 Comparative Filters

The following comparative filters were used in the evolutionary analyses:

1. **Chimp synteny.** (universal, hard) Regions not showing conserved synteny in human/chimpanzee alignments, according to the UCSC “syntenic net.” These regions are likely to be enriched for misalignments with chimpanzee.

2. **Recent segmental duplications.** (universal, hard) Regions corresponding to recent segmental duplications in the human genome were excluded, based on the “Segmental Dups” track in the UCSC Genome Browser (see Bailey et al., 2001). These regions are likely to be enriched for misalignments with chimpanzee.

3. **CpGs.** (universal, soft) Positions corresponding to hypermutable CpG dinucleotides in any genome, including chimp, were excluded from all genomes. These sites are likely to produce violations of various assumptions of our models (e.g., the molecular clock assumption, the assumption of constant mutation rates across sites).

4. **Exons of protein-coding genes.** (universal, hard) Intervals corresponding to exons of protein-coding genes (including UTRs) and 1000 bp flanking each exon were eliminated, based on the RefSeq, ENSEMBL, and UCSC Genes tracks in the UCSC Genome Browser.

5. **Noncoding RNAs.** (universal, hard) Intervals corresponding to exons of known noncoding RNA genes and 1000 bp flanking each exon were eliminated, based on the RNA Genes track in the UCSC Genome Browser.

6. **Conserved noncoding elements.** (universal, hard) Intervals corresponding to phastCons elements and 100 bp of flanking sequence were eliminated, based on the 44-way Conservation track in the UCSC Genome Browser (Eutherian subtrack).

We also experimented with “uniqueness” filters, designed to eliminate regions of the genome highly similar to other regions, but we found these filters were unnecessary as long as reads with zero mapping quality were eliminated from all analyses.

### S1.3 BSNP Method for Genotype Inference

#### S1.3.1 General Approach

As in several other recent papers (Marth et al., 1999; Li et al., 2009b; Shen et al., 2010), we take a Bayesian approach to genotype inference. Our method, called BSNP, allows a full posterior distribution over genotypes to be inferred at each position, conditional on the aligned bases, their basecall quality scores, and
the mapping quality scores produced by BWA. BSNP is similar in many respects to the SOAPsnp method (Li et al., 2009b), although it was developed independently. The main differences are that we intentionally avoid making use of the reference allele or previously identified variants (as in dbSNP) in our prior, we use mapping quality scores as well as basecall quality scores, and we do not explicitly consider other covariates (such as the sequencing cycle) in the model, but assume that they are reflected in the quality scores. (Our quality score recalibration step helps to ensure that this is true.)

While BSNP is not radically different from other available tools in its technical details, it is based on a subtle but important difference in motivation. Our goal with BSNP is to infer the correct genotype at each position, in as unbiased a manner as possible. By contrast, standard “SNP detection” methods are designed to identify as many differences as possible from the reference genome, while controlling the false positive rate for variant detection. This distinction is important because, for our purposes, it is critical to avoid making use of the reference allele or previously identified variants (as in dbSNP) in our prior, we use mapping quality scores produced by BW A. BSNP is similar in many respects to the SOAPsnp method (Li et al., 2009b), although it was developed independently. The main differences are that we intentionally avoid having some genomes (say, ones with lower coverage) “pulled” more strongly toward the reference genome than others (say, ones with higher coverage), as can happen when a method is optimized to minimize overall genotyping error. In our setting, this type of bias would tend to systematically distort the apparent genealogical relationships between the samples, leading to skewed estimates of key parameters.

### S1.3.2 Basic Model

Consider a particular genomic position \( j \) with an associated set of aligned bases from sequencing reads for a particular individual. Let \( G \) represent the individual’s true diploid genotype at position \( j \), with \( G \in \{ AA, CC, GG, TT, AC, AG, AT, CG, CT, GT \} \). At this stage, we have no means for distinguishing between the two alleles of \( G \) so we consider the ten unordered pairs of bases, arbitrarily listing the bases of heterozygotes in alphabetical order (e.g., AG rather than GA). Let \( X = (X_1, \ldots, X_n) \) denote the base calls of the aligned reads, let \( q = (q_1, \ldots, q_n) \) denote the corresponding base-call quality scores (on the phred scale), and let \( m = (m_1, \ldots, m_n) \) denote the corresponding mapping quality scores (as produced by BWA and also represented in phred units). It is convenient to transform these quality scores to error probabilities; hence, let \( Q = (Q_1, \ldots, Q_n) \) and \( M = (M_1, \ldots, M_n) \) be probabilities of error, such that \( Q_i = 10^{-q_i/10} \) and \( M_i = 10^{-m_i/10} \) for \( i \in \{ 1, \ldots, n \} \). Finally, let \( H = (H_1, \ldots, H_n) \) be the (unobserved) haploid bases that representation of \( H_i \).

Our model assumes conditional independence of the \( H_i \) variables given \( G \), and conditional independence of the \( X_i \) variables given the corresponding \( H_i, Q_i, \) and \( M_i \) variables (see Figure A). Thus, the conditional likelihood at position \( j \), \( L_j(G) \), is given by:

\[
L_j(G) = P(X|G, Q, M) = \prod_{i=1}^{n} P(X_i|G, Q_i, M_i) = \prod_{i=1}^{n} \sum_{H_i} P(X_i|Q_i, M_i, H_i)P(H_i|G).
\] (2)

It follows that the unconditional likelihood, \( L_j = P(X|Q, M) \), and the posterior distribution, \( P(G|X, Q, M) \), are given by:

\[
L_j = \sum_G L_j(G)P(G), \quad P(G|X, Q, M) = \frac{P(G)L_j(G)}{L_j} \tag{3}
\]

where \( P(G) \) is a prior distribution over genotypes.

Various options are available for the prior distribution. For simplicity and to avoid biases that might influence downstream analyses, we use an uninformative prior that depends only on a per-base heterozygosity rate, \( \pi \), and treats all heterozygous genotypes and all homozygous genotypes as equally probable.
Figure A: Graphical model used by BSNP for genotype inference at each genomic position. $G$ denotes the true diploid genotype, $X_1, \ldots, X_n$ are the reported nucleotides in $n$ aligned sequence reads, $H_1, \ldots, H_n$ are the corresponding haploid alleles (the true values of the nucleotides that were sequenced), $Q_1, \ldots, Q_n$ are the base-call error rates (as implied by the reported basecall quality scores), and $M_1, \ldots, M_n$ are the mapping error rates (as implied by the reported mapping quality scores). The solid circles represent observed random variables and the empty circles represent unobserved (latent) random variables. The two sets of error rates are represented by smaller, black nodes to show that they are not treated as random in this analysis, but are simply conditioned on. The objective of the inference procedure is to compute a posterior distribution for $G$ given the $X$, $Q$, and $M$ variables, integrating over possible values of the $H$ variables.

Specifically, for a genotype $G$ defined by two bases $b_1$ and $b_2$ (e.g., $b_1 = b_2 = A$ for an AA homozygote and $b_1 = A$, $b_2 = G$ for an AG heterozygote), the prior is given by:

$$P(G = b_1 b_2) = \begin{cases} \frac{(1 - \pi)}{4} & b_1 = b_2 \\ \frac{\pi}{6} & b_1 \neq b_2 \end{cases}$$  \hspace{1cm} (4)

For our analyses, we assume $\pi = 10^{-3}$. A somewhat more general prior is actually implemented in BSNP, as described in Section S1.3.5.

The conditional likelihood depends on two quantities, the haplotype probability $P(H_i|G)$ and the data probability $P(X_i|Q_i, M_i, H_i)$ (equation 2). Assuming the two alleles at each position are sampled with equal probability during the sequencing process, the haplotype probability is simply:

$$P(H_i|G = b_1 b_2) = \frac{1}{2} \delta_{H_i, b_1} + \frac{1}{2} \delta_{H_i, b_2}$$  \hspace{1cm} (5)

where $\delta$ is the Kronecker delta function. Notice that this expression reduces to $P(H_i|G = b_1 b_1) = \delta_{H_i, b_1}$ in the case of a homozygote.

The remaining term, the data probability, must allow for both sequencing and mapping error. Here, we introduce an additional latent variable, $Z_i$, indicating whether ($Z_i = 1$) or not ($Z_i = 0$) the $i$th read is correctly mapped. If the read is incorrectly mapped, the base $X_i$ is irrelevant and it is treated as if there were missing data; hence, $P(X_i|H_i, Q_i, Z_i = 0) = 1$. If, on the other hand, the read is correctly mapped, the basecall may be correct or incorrect, with probabilities given by $(1 - Q_i)$ and $Q_i$, respectively:

$$P(X_i|H_i, Q_i, Z_i = 1) = (1 - Q_i) \delta_{X_i, H_i} + \frac{1}{3} Q_i (1 - \delta_{X_i, H_i})$$  \hspace{1cm} (6)
(assuming all three erroneous bases are equally likely). Allowing for both mapping error and basecall error,

\[
P(X_i|H_i, Q_i, M_i) = P(Z_i = 0|M_i)P(X_i|H_i, Q_i, Z_i = 0) + P(Z_i = 1|M_i)P(X_i|H_i, Q_i, Z_i = 1)
\]

\[
= M_i + (1 - M_i) \left[ (1 - Q_i)\delta_{X_i, H_i} + \frac{1}{3}Q_i(1 - \delta_{X_i, H_i}) \right]
\]

\[
= \begin{cases} 
1 - Q_i + Q_iM_i & X_i = H_i \\
\frac{1}{3}Q_i + M_i - \frac{1}{3}Q_iM_i & X_i \neq H_i
\end{cases}
\] (7)

The conditional likelihood is given by a sum over all haplotypes \(H_i\) of the product of equations 5 and 7. Notice, however, that \(P(H_i|G)\) has a degenerate form—it is zero for most values of \(H_i\). As a result, the sum simplifies to:

\[
P(X|G = b_1b_2, Q, M) = \prod_{i=1}^{n} \sum_{H_i} P(H_i|G = b_1b_2)P(X_i|H_i, Q_i, M_i)
\]

\[
= \prod_{i=1}^{n} \left[ \frac{1}{2}P(X_i|H_i = b_1, Q_i, M_i) + \frac{1}{2}P(X_i|H_i = b_2, Q_i, M_i) \right]
\]

\[
= \left( \frac{1}{2} \right)^n \prod_{i=1}^{n} \left[ P(X_i|H_i = b_1, Q_i, M_i) + P(X_i|H_i = b_2, Q_i, M_i) \right].
\] (8)

Notice also that the factor of \((\frac{1}{2})^n\) applies to all genotypes and therefore cancels in the calculation of posterior probabilities.

To avoid numerical underflow, the unconditional likelihood and posterior probabilities (equation 3) are computed in log space. In particular, if \(l_j(G)\) is the log of the conditional likelihood,

\[
l_j(G = b_1b_2) = \log P(X|G = b_1b_2, Q, M)
\]

\[
= -n \log 2 + \sum_{i=1}^{n} \log[P(X_i|H_i = b_1, Q_i, M_i) + P(X_i|H_i = b_2, Q_i, M_i)]
\]

and \(l' = \max_G l_j(G)\), then the unconditional log likelihood \(\ell_j\) is computed as:

\[
\ell_j = \log \left( \sum_G \exp(l_j(G))P(G) \right) = l' + \log \left( \sum_G \exp(l_j(G) - l')P(G) \right)
\] (9)

and the posterior probabilities are computed as:

\[
P(G|X, Q, M) = \frac{\exp(l_j(G) - l')P(G)}{\sum_g \exp(l_j(g) - l')P(g)}.
\] (10)

S1.3.3 Allowing for Correlated Errors

The model above assumes that errors arise independently in the reads that are aligned to a particular genomic position, but this assumption is unrealistic for various reasons, including error-inducing contexts that are shared across reads, or errors that arise during PCR amplification (Huang and Madan, 1999; Li et al., 2008, 2009b). Moreover, this erroneous independence assumption can lead to errors in genotype inference. In particular, it tends to cause homozygous genotypes to be miscalled as heterozygotes. For example, suppose two of seven reads aligned at a particular genomic position match the reference genome, with the other five
representing an alternative allele. An approach that assumes independence of errors will tend to predict a heterozygous genotype in this case, because two errors of the same type are unlikely to occur by chance. However, if errors are highly correlated, a homozygous genotype (with errors) may be more probable than a heterozygous genotype here.

The processes leading to correlated errors are difficult to model explicitly, but heuristic methods have been used to mitigate their effects in genotype inference. The general idea of these methods is to capture correlation of errors by artificially boosting the probabilities of successive sequencing errors. A complication with applying this approach in our setting is that, unlike methods such as MAQ (Li et al., 2008), our model does not maintain an explicit distinction between bases that arise from errors and bases that do not. However, it is not too costly to introduce such a distinction for the purposes of a heuristic for error correlation.

Because our main quantity of interest is the conditional likelihood, \( L_j(G) = P(X|G) \) — the unconditional likelihoods and posterior probabilities follow via equation 3—we need only consider the erroneous bases conditional on a candidate true genotype, \( G = b_1b_2 \). These can be naturally defined as the elements of \( X \) that match neither \( b_1 \) nor \( b_2 \). A minor problem in that some of these bases may result from mapping errors rather than sequencing errors, but we finesse this issue by simply thresholding the mapping scores. Thus, we define the “putatively erroneous bases” with respect to \( G = b_1b_2 \) as all \( X_i \) such that \( X_i \neq b_1 \) and \( X_i \neq b_2 \) and \( m_i > T \). We found that simply eliminating nonunique mappings (\( T = 0 \)) was adequate for our purposes.

Let \( X^{(err)} = X_1^{(err)}, \ldots, X_k^{(err)} \) be the putatively erroneous bases with respect to \( G = b_1b_2 \). Following Li et al. (2008), we arrange these such that bases \( X_1^{(err)}, \ldots, X_k^{(err)} \) have decreasing phred scores \( q_1^{(err)}, \ldots, q_k^{(err)} \) and increasing error probabilities \( Q_1^{(err)}, \ldots, Q_k^{(err)} \). Next, we modify each error probability by raising it to a power \( \theta^i-1 \) (where \( 0 < \theta \leq 1 \)), such that the first probability is unchanged but subsequent ones are increased. In our setting, this is straightforward to achieve because our model already incorporates explicit error probabilities. Notice that assuming the \( i \)th error has probability \( Q_i^{(err)} \theta^{i-1} \) is equivalent to scaling the phred score by a factor of \( \theta^{i-1} \), i.e., we can simply use \( \tilde{q}_i^{(err)} = \theta^{i-1} q_i^{(err)} \) in place of the original \( q_i^{(err)} \). After this transformation of the phred scores corresponding to putatively erroneous bases \( 2, \ldots, k \), the calculation of the conditional likelihood—and then of the unconditional likelihoods and posterior probabilities—proceeds as usual.

Notice that this procedure has no effect when there are \( \leq 1 \) putatively erroneous bases. Its main effect is that it tends to increase the conditional likelihoods of homozygous genotypes in cases of \( >1 \) apparently erroneous bases, while leaving those for heterozygous genotypes largely unchanged (because typically at most two alleles are represented in the data). Thus the posterior probabilities for homzygotes tend to increase and those for heterozygotes tend to decrease, making homozygous calls more likely.

To determine an appropriate value for \( \theta \), we performed a small cross-validation study using sites in NA12891 for which HapMap genotypes were available. We estimated genotypes for each site using BSNP for several values of \( \theta \) between 0.5 and 1.0, and for each value we measured error rates, sensitivity for non-HomRef variants, and false call rates, both for all sites and separately for heterozygotes and homozygous equal to (HomRefs) and different from (HomDifs) the reference allele. We found that a value of \( \theta = 0.85 \) allowed for a good balance between sensitivity and specificity in variant detection. Interestingly, this is the same value obtained by Li et al. (2008), despite several differences in our models. We observed a small but significant improvement in performance with \( \theta = 0.85 \) compared with the case of \( \theta = 1 \), corresponding to no allowance for correlated errors. In particular, BSNP had a clear tendency to overcall heterozygotes with \( \theta = 1 \), which was eliminated at lower values of \( \theta \).
S1.3.4 Maximum a Posteriori Genotype Calls

BSNP reports both a full posterior distribution over genotypes and a single maximum a posteriori (MAP) genotype at each genomic position. Our original intention was to use the full posterior distributions wherever possible in downstream analyses, either by integrating over possible genotypes or by working with posterior expected values of quantities of interest (such as pairwise divergences between genomes). However, we found, in practice, that the posterior distributions were strongly peaked near 1, indicating strong support from the data for a particular genotype. This was true even for the genomes with somewhat lower coverage, such as Watson and Venter (see Figure S4). As a result, conditioning on the MAP genotype calls produced nearly identical results to those obtained by integrating over posterior distributions or using posterior expected values, in several analyses that we tried both ways (results not shown). Therefore, we have used the simpler approach of conditioning on MAP genotype calls throughout this paper.

S1.3.5 More General Prior Implemented in BSNP

As described above, in our analysis we have used a prior that simply assigns all homozygous genotypes one probability and all heterozygous genotypes another. However, it is straightforward to use a more informative prior, considering factors such as the local G+C content, the transition/transversion bias, or context effects from flanking nucleotides. The prior that is implemented in BSNP allows for a nonuniform base composition and a transition-transversion bias, as follows:

\[
\begin{align*}
P(G = b_1 b_2) &= \left\{ \begin{array}{ll}
(1 - \pi) f(b_1; \phi) & b_1 = b_2 \\
\pi \kappa f(b_1; \phi) f(b_2; \phi) \cdot \frac{4}{1 + 2\phi \kappa (1 - \phi)} & b_1 \neq b_2, b_1 \leftrightarrow b_2 \text{ transition} \\
\pi f(b_1; \phi) f(b_2; \phi) \cdot \frac{4}{1 + 2\phi \kappa (1 - \phi)} & b_1 \neq b_2, b_1 \leftrightarrow b_2 \text{ transversion}
\end{array} \right. 
\end{align*}
\]

where \( \kappa \) is the transition-transversion rate bias and \( f(b; \phi) \) is the background probability of base \( b \), assuming G+C content \( \phi \) and equal probabilities of G/C and A/T bases (strand symmetry), i.e., \( f(G; \phi) = f(C; \phi) = \phi/2 \) and \( f(A; \phi) = f(T; \phi) = (1 - \phi)/2 \). The quantity \( 4/(1 + 2\phi \kappa (1 - \phi)) \) is needed to normalize the probabilities of the heterozygotes.

This prior reduces to the one we have used when \( \kappa = 1 \) and \( \phi = 0.5 \) (see the BSNP command-line invocation in Section S1.1). We performed limited experiments with more informative priors and found that they had very little effect on our genotype accuracy, presumably because our sequence data is sufficiently deep that the data quickly overwhelms the prior.

S2 Validation of Genotype Calls

In this section, we examine the accuracy of our methods for genotype inference, focusing on an individual of European descent (Coriell id NA12891), who has been both genotyped by the HapMap project and sequenced to deep coverage in the “Pilot 2” component of the 1000 Genomes (1000G) Project (The 1000 Genomes Project Consortium, 2010) (see statistics in Table S1). Thus, our genotype inferences can be compared both with array-based genotype calls for a fraction of the genome, and with sequence-based calls for the entire genome. In addition to our own genotype calls (denoted “BSNP”), we considered two other sets of sequence-based calls: one released by the 1000G project and one obtained by running the program MAQ (Li et al., 2008) on our alignments. All three sets of sequence-based genotype calls were benchmarked against the array-based calls from HapMap (the best available “gold standard”), and in addition were compared with one another genome-wide. This approach allowed us to evaluate the performance of our entire alignment pipeline (by comparing BSNP/MAQ vs. 1000G), and of just the genotype inference step of the pipeline (by comparing BSNP vs. MAQ). Note that the same sequence reads were used in all cases.
We also examine genotyping accuracy in the Venter genome, for which some array-based genotyping data is available (Levy et al., 2007). This case allows us to observe the effects of lower sequencing coverage (see Table 1, main text) on genotype error rates. In this case, no genotype calls are available from the 1000G project, but we have the published variants from Levy et al. (2007). As with NA12891, we considered both BSNP and MAQ calls based on our alignments.

The genotype calls that we analyzed were obtained as follows. The BSNP calls were produced using our pipeline, as described in Section S1. The MAQ calls were generated as a by-product of the same pipeline, because the SAMtools “pileup” command applies the MAQ algorithm at each genomic position in generating a consensus base (see http://samtools.sourceforge.net/cns0.shtml). These genotype predictions (with corresponding quality scores) were captured and stored in our database along with the BSNP calls. The 1000G calls for NA12891 were downloaded in VCF format from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2009_12/pilot2/. Positions not included in the VCF file were considered to be implicit HomRef predictions. Merged HapMap phase I, II and III SNP calls were downloaded from the HapMap website (http://hapmap.ncbi.nlm.nih.gov; July 7th 2009 release). All data files were obtained for CEU Europeans, then genotypes at which sample NA12891 was defined at both alleles were extracted. For the Venter genome, we used the reported results for the Affymetrix NspI and StyI arrays (Levy et al., 2007). Raw data was obtained for two individual replicates of each array from ftp://ftp.jcvi.org/pub/data/huref/. Data from all four individual hybridizations was combined, and discordant SNP calls were dropped from further analysis. For both NA12891 and Venter, we considered only autosomal sites that passed our data quality filters. The array comparison considered autosomal sites that both passed these filters and were represented on the arrays (3,338,073 for NA12891, 375,050 for Venter). The genome-wide comparisons considered all autosomal sites that passed the data quality filters (1.851 billion for NA12891, 1.557 billion for Venter).

S2.1 Comparison with Arrays

S2.1.1 NA12891

We first report overall rates of error in NA12891 with respect to HapMap for the BSNP, MAQ, and 1000G methods. We report error rates for all sites, and for the subsets of sites that are homozygous for the reference allele (HomRef), homozygous for a different allele (HomDif), or heterozygous (Het), according to HapMap. For HomRefs, all three methods have quite similar error rates, with BSNP and MAQ at 0.31%, and 1000G at 0.34% (Figure S2A). For HomDifs, BSNP and MAQ are nearly identical (1.06%) but 1000G’s error rate is substantially higher, at 1.86%, while for Hets, 1000G performs best, with 1.66% error, compared with 1.99% for BSNP and 2.34% for MAQ. Overall, BSNP performs slightly better than the other methods, with an error rate of 0.78% compared with 0.85% for MAQ and 0.86% for 1000G.

In interpreting these results, it is important to bear in mind that the sites represented in HapMap are highly nonrepresentative of the genome as a whole. Roughly 99.9% of all sites are HomRefs, while among the HapMap sites we have considered, this fraction is only 63.0% (versus 16.4% for HomDifs and 20.5% for Hets). As a result, overall error rates at HapMap sites can be a misleading measure of accuracy, by penalizing errors in variant sites (Hets and HomDifs) too much and errors in HomRef sites too little. At the same time, it is important for our purposes to detect as many variants as possible, while maintaining a high level of accuracy at HomRef sites. To illustrate the difficulty of arriving at an appropriate measure of accuracy, consider the trivial strategy of predicting a HomRef at every position: this would result in an excellent level of accuracy genome-wide (~0.1% error), but would obviously be useless for a population genetic analysis.

To address this problem, we also measured sensitivity and false call rates (FCRs) for variant sites. Sensitivity (the fraction of variant sites that are correctly identified) has the benefit of not depending on the
prevalence of variant sites, i.e., the frequency with which they occur in the test data. The absolute value of the FCR (the fraction of predicted variant sites that are called incorrectly) does depend on prevalence, but relative FCRs for different methods should be insensitive to prevalence, as long as the subsets of variant and non-variant sites considered are reasonably representative of the genome. Therefore, sensitivity and FCRs can be useful measures by which to compare the detection power and accuracy of alternative genotype inference methods, even when the benchmark data set has a nonrepresentative composition.

As shown in Table S2, sensitivity levels and FCRs for NA12891 are generally quite good across methods, with sensitivity approaching 99% and FCRs at ~0.5%. Not surprisingly, sensitivity levels tend to be lower and FCRs higher for Hets than for HomDifs. Nevertheless, Het sensitivity levels remain near 98% and Het FCRs near 1%. BSNP and MAQ are quite similar by all measures, but BSNP appears to have a slight edge in terms of both sensitivity and FCR. 1000G shows slightly higher sensitivity and FCR for Hets, but substantially reduced sensitivity and FCR for HomDifs. This tendency to trade HomDifs for Hets might reflect the use of the reference allele in the prior. It could also stem from differences in the handling of correlated errors, in alignment methods, or in other aspects of the pipeline.

It is worth noting that the estimated FCRs may be substantially inflated by the presence of errors in the HapMap benchmark set. While absolute error rates are low in HapMap, errors will tend to be strongly over-represented among sites at which the sequence-based and array-based genotype calls are discordant. Indeed, Hoberman et al. (2009) recently examined 52 sites at which their sequence-based method for predicting heterozygous genotypes disagreed with HapMap (with high confidence), and found that in 51 of 52 cases (98%), their prediction was supported by an independent genotyping experiment. Manual inspection of our discordant cases reveals many instances in which the sequence data strongly supports a genotype other than the one reported in HapMap (e.g., with many reads in agreement, unambiguous alignments, and high quality scores). We cannot easily estimate our true FCR but the reported values should be considered loose upper bounds. Note that errors in HapMap will also influence our estimates of sensitivity, but they should have a proportionally much smaller effect in this case.

S2.1.2 Venter

The results for the Venter genome were generally similar to those for NA12891 (Figure S2B and Table S2). Again, BSNP and MAQ show quite similar error rates in HomRefs (about 1.1% error) and HomDifs (1.8%). These error rates are slightly but not dramatically higher than those for NA12891. However, the reduced sequencing coverage has a pronounced effect on the error rate for Hets. Here BSNP has an error rate of 15.2% and MAQ has an error rate of 18.0%, leading to overall error rates of 4.2% for BSNP and 5.0% for MAQ. The published results from Levy et al. (2007) appear to favor the reference allele more strongly than the other methods, and therefore have lower error rates for HomRefs (0.1%) but substantially higher rates for HomDifs (4.5%) and intermediate rates for Hets (16.4%), for an overall error rate of 5.4%. Across methods, the sensitivity levels for Venter are somewhat lower than those for NA12891, at ~95%, but the FCRs remain excellent, at ~0.1% (perhaps due to longer sequence reads and higher average basecall quality). As expected, the sensitivity of Hets is considerably reduced (to <85%), but Het FCRs remain good (<0.4%). As with NA12891, BSNP generally performs slightly better than the other methods in terms of its tradeoff between sensitivity and FCRs. The published calls from Levy et al. (2007) appear to be conservative, with reduced sensitivity levels and FCRs, particularly for HomDifs. Both BSNP and MAQ have somewhat elevated FCRs for HomDifs (~10%), presumably because many Hets in low coverage regions are being misidentified as HomDifs.

As in previous studies (Wang et al., 2008; Wheeler et al., 2008), the power for genotype identification depends heavily on read coverage (Figure S3), especially for Hets. This is particularly an issue with the lower-coverage Venter genome. The problem of reduced power in Venter is mitigated somewhat by the use of the coverage filter, but cannot be entirely eliminated with only 7.5x average coverage (see Figure
Interestingly, power is somewhat better in Venter than in NA12891 at comparable levels of coverage, presumably due to higher sequencing accuracy and reduced alignment error from longer reads. Another observation is that power for HomRefs is slightly better than power for HomDifs in NA12891 at low read depths, despite that our genotyping model does not directly consider the reference allele. This probably reflects a slight alignment bias favoring reads that match the reference allele. This difference is not evident with Venter, presumably because this alignment bias is mostly eliminated with longer reads.

We note that our aim in this analysis is not to perform a definitive comparison of genotype inference methods. A full comparison should consider other recently published genotype inference methods (e.g., Marth et al., 1999; Li et al., 2009b; Hoberman et al., 2009; Shen et al., 2010) and consider additional benchmark sets, representing other sequencing technologies and population groups. Nevertheless, we think this limited validation study is sufficient to establish that BSNP performs at least as well as other state-of-the-art methods for genotype inference, while maintaining advantages important for our purposes—for example, by avoiding a direct dependence on the reference allele, and not being tied to any particular sequencing technology. In addition, our benchmarking experiments indicate that our absolute accuracy in genotype inference is reasonably high, allowing >95% of variants to be detected with low false call rates.

### S2.2 Genome-wide Comparison with Alternative Sequence-based Methods

#### S2.2.1 NA12891

We next compare the three sequence-based methods—BSNP, MAQ, and 1000G—genome-wide in NA12891, focusing on the 1.851 billion sites that pass our data-quality filters. In this case there is no “gold standard” to use in estimating absolute error rates, but useful insights can be obtained by comparing the alternative sets of predictions and assessing their degree of concordance.

The three methods predict similar numbers of variant sites—2.173 million for BSNP, 2.137 million for MAQ, and 2.097 million for 1000G. The slight differences in these numbers are consistent with the apparent differences in power observed in the previous section (cf. column 2 of Table S2). Among the 2.237 million sites predicted to be non-HomRef variants by at least one method, all three methods predict a variant at 2.040 million sites (91.2%), and at least two methods predict a variant at 2.144 million sites (95.8%; see Figure S5A). The remaining “singleton” cases are approximately equally divided between BSNP and 1000G (46,233 and 46,376 sites, respectively). These appear to consist primarily of predicted Het alleles, and presumably reflect the use of slightly different criteria in distinguishing between Hets and HomRefs with sequencing errors. Recall that BSNP and MAQ have been applied to the same alignments, and that BSNP has somewhat higher sensitivity for Hets (Table S2); it is probably for this reason that few singletons are attributed to MAQ.

Pairwise comparisons between methods (Figure S5B–D) reveal a fairly high degree of concordance not only in the identification of variant alleles, but in the assignment of variants as Hets or HomDifs. Not surprisingly, the MAQ and BSNP methods (panel B) show the greatest level of agreement, while 1000G and BSNP (panel C) and MAQ and 1000G (panel D) are slightly less concordant. Consistent with the findings reported above, most disagreements between methods have to do with distinguishing between Hets and HomRefs, with BSNP and 1000G being somewhat more aggressive than MAQ in calling Hets (although in slightly different ways). The numbers of HomDif/HomRef disagreements are particularly small, but they are considerably larger in the 1000G/BSNP and MAQ/1000G comparisons than in the MAQ/BSNP comparison, presumably because they tend to reflect differences in alignment.

#### S2.2.2 Venter

We also examined the Venter genome and found qualitatively similar results (Figure S6). As with NA12891, the MAQ and BSNP methods showed a high degree of concordance, agreeing at 93.9% of putative variant
sites, with differences occurring primarily at sites considered Hets by BSNP and HomRefs by MAQ (panels A & B). The SNP calls from Levy et al. (2007) showed somewhat poorer agreement with MAQ and BSNP than did the 1000G calls, above, but all three methods still agreed at 84.3% of putative variant sites (without regard to the type of variant). Interestingly, the published SNPs from Levy et al. contain a substantial number of singleton Hets, probably reflecting differences in the alignment methods.

In general, we find that, while there are clear differences between methods, they occur at a small minority of putative variant sites, and typically derive from borderline Het cases. The overall degree of concordance of MAQ and BSNP with 1000G (for NA12891) and Levy et al. (for Venter) is encouraging, considering the use of independent alignment pipelines. Together, the array-based validation and the whole-genome comparisons suggest that error rates in genotype inference are modest, and are unlikely to be a dominant factor in our population genetic analyses.

S3 Application of Pipeline to Individual Genomes

We applied our genotype inference pipeline to the seven individual genomes under study, comparing our genotype calls for each genome with the published SNP sets (see Table S1), and with one another.

S3.1 Numbers of Variants

We first counted the number of Het and HomDif variants identified in each genome, normalizing these counts by the number of sites under consideration (i.e., excluding sites that did not pass our filters). For this analysis, we applied our data quality filters but not our comparative filters. The normalized counts of Het variants represent estimates of the heterozygosity per site ($\pi$) for each genome. Our estimates of $\pi$, at 5.6–6.1 \times 10^{-4} for Europeans, 6.6–6.7 \times 10^{-4} for East Asians, and 8.8–9.5 \times 10^{-4} for Africans (Figure S7), are considerably more consistent within major population groups than are estimates based on the published SNP sets—see, for example, Venter vs. Watson, and NA18507 vs. ABT and KB1 (Figure S7)—highlighting the value of a uniform pipeline for genotype inference. In addition the absolute values of our estimates are reasonably concordant with values previously estimated from genomic sequence data. For example, Sachidanandam et al. (2001) estimated an average of $\pi = 7.5 \times 10^{-4}$ based on sequence reads from individuals drawn from multiple population groups, and Reich et al. (2001) estimated $\pi = 7.0 \times 10^{-4}$ by resequencing a series of short genomic intervals in 44 unrelated CEPH individuals from Utah. Unlike the Het rates, the estimated HomDif rates are fairly uniform across genomes. These rates have no straightforward population genetic interpretation, and may be influenced by the fact that the reference genome is a chimera, representing several population groups.

Our larger estimates of $\pi$ for Africans than for East Asians and Europeans are consistent with many previous reports, but our estimated values do not reflect the typical observation of slightly elevated heterozygosity in Europeans relative to East Asians (e.g., Nei and Takezaki, 1996; Relethford and Jorde, 1999; Storz et al., 2004; Frazer et al., 2007). However, our two European genomes—Venter ($\pi = 6.1 \times 10^{-4}$) and Watson ($\pi = 5.6 \times 10^{-4}$)—both show somewhat reduced heterozygosity compared with our European validation sample, NA12891 ($\pi = 7.2 \times 10^{-4}$), suggesting that limited power for heterozygotes due to reduced sequencing redundancy has led to underestimates of $\pi$. Indeed, the reduction of the estimate for Venter relative to that for NA12891 is almost exactly what would be expected with a 15% false negative rate for Hets (see Table S2). Comparing NA12891 with the two East Asian genomes, both of which have fairly high sequencing redundancy, we find that its estimated heterozygosity is indeed higher by 8–10%. The largest estimate of $\pi$ is for the San genome, KB1, consistent with previous reports of elevated heterozygosity in this genome (Schuster et al., 2010).

Assuming the true value of $\pi$ for Watson is also similar to that for NA12891, the false negative rate
for heterozygotes in this genome would have to be >22% to explain the observed counts. Such a rate is plausible with an average read depth of only 5.6 (versus 8.4 for Venter; Table 1, main text). Along with the reduced heterozygosity in Watson, we observed a slight excess of HomDifs, suggesting that some Hets are being miscalled as HomDifs, probably in regions of sparse data (recall that our prior favors homozygotes but does not consider the reference allele). Because of concerns about high error rates in Watson, we chose to exclude this genome from our G-PhoCS analysis. While the Venter genome also shows some tendency for Hets to be misidentified as HomDifs, the effect is considerably less pronounced in this case.

S3.2 Trees Based on Average Genomic Divergence

As an alternative way of summarizing our genome-wide genotype inferences, we built distance-based phylogenies for the seven genomes under consideration. In this case, we applied both our data quality and comparative filters. We constructed a matrix of pairwise genomic distances, then built phylogenies from this matrix using both the neighbor-joining and unweighted pair group with arithmetic mean (UPGMA) methods. Here, the estimated pairwise distances were simply meant to reflect average genomic divergence times, rather than population divergence times (i.e., they included both divergence times and ancestral coalescence times). The chimpanzee genome sequence was included in the analysis as well, to allow the human tree to be rooted.

Because the genomes are diploid, a method was needed to summarize the divergence of two separate pairs of chromosomes in the absence of phasing information. We used the following simple estimator for the divergence of two genomes $X$ and $Y$. Let $X_i = a_i b_i$ be the genotype at position $i$ in $X$, and let $Y_i = c_i d_i$ be the genotype at the same position in $Y$. The divergence between $X$ and $Y$ was estimated as:

$$d(X, Y) = \frac{1}{L} \sum_{i=1}^{L} \left[ 1 - \frac{1}{2} \max \left( \delta_{a_i,c_i} + \delta_{b_i,d_i}, \delta_{a_i,d_i} + \delta_{b_i,c_i} \right) \right]$$

(12)

where $\delta$ is the Kronecker delta function and $L$ is the number of sites under consideration. Notice that the expression inside the sum reduces to 0 in the case of matching homozygotes, to 1 in the case of mismatching homozygotes, and either to $\frac{1}{2}$ or 1 in the case of a homozygote and a heterozygote (depending on whether or not they share an allele). However, when measuring the distance between two heterozygotes it allows them to be aligned in the most parsimonious way. The quantity $d(X, Y)$ can therefore be interpreted as a conservative estimate of the expected number of mutations per site between individual chromosomes drawn at random from individuals $X$ and $Y$.

The neighbor-joining tree (Figure S8A) positions the two East Asians and the two Europeans as sister taxa, groups the East Asians and Europeans together in a clade, and places KB1 as a outgroup to the other human individuals, all as expected. The only major ambiguity is the position of ABT, which groups with NA18507 but with a very short leading branch, reflecting the fact that ABT is essentially equidistant to the Eurasian and West African genomes. The leaves of the tree are nearly equidistant from the root, indicating good support for the assumption of a molecular clock. The slightly elongated branch to the Watson genome appears to be explained by an excess of HomDifs rather than Hets, due to low coverage (as described above)—because HomDifs require more mutations to explain than Hets, overcalling of HomDifs leads to overestimates of divergence.

If a molecular clock is assumed to hold and the tree is inferred using the UPGMA method (Figure S8B), the same topology is found, but in this case it is possible to place approximate dates at ancestral nodes of the tree. We calibrate these dates using the average genomic divergence of human and chimpanzee, which we assume occurred about 6.5 Mya (Section S7.2). This results in average divergence times of about 275 kya for the two East Asians, 330 kya for the two Europeans, 350 kya for the Europeans and East Asians, 425 kya for the Yoruban and Bantu, 450 kya for the Eurasians and Africans, and 500 kya for the San and the
other individuals. In comparison, the corresponding divergence time for the Neanderthal recently reported by Green et al. (2010) is about 825 kya.

**S4 G-PhoCS: A Generalized Phylogenetic Coalescent Sampler**

As described in the main text, G-PhoCS is an MCMC-based program for inferring population evolution parameters, which is based on the MCMCcoal program by Rannala and Yang (2003) (see also Burgess and Yang, 2008). The program is derived directly from the MCMCcoal source code, which is written in C (v1.2, downloaded January, 2010)\(^1\). However, fairly extensive changes to the code were required to accommodate our extensions to the model, and large parts of the code have been rewritten.

Below we describe our model and inference procedure in detail. In Section S4.1, we briefly review the main features of MCMCcoal. In Sections S4.2 and S4.3, we outline the two main novelties introduced in G-PhoCS, viz.: (1) modeling of gene flow between populations, and (2) handling of unphased diploid genotypes in the input. Finally, in Section S4.4, we describe some techniques we have developed for improving the efficiency of the likelihood calculations in G-PhoCS.

**S4.1 A High-level Overview of MCMCcoal**

MCMCcoal assumes a known population phylogeny (tree) \(T\), in which each population \(p\) (current and ancestral) is associated with a parameter \(\theta_p\) designating its effective population size, and each ancestral population is associated with a divergence time \(\tau_p\). The \(\theta\) parameters can be viewed as defining the expected waiting time until the coalescence of two lineages (the actual expected time is \(\theta^{-1}\)). Hence, both the \(\tau\) and \(\theta\) parameters can be scaled in the same time units. Since inference is driven by mutation, it is convenient to scale time by the expected number of mutations per site, so that one time unit equals the expected time for a substitution to occur at a single site (see Section S7.2 for details about parameter calibration).

MCMCcoal takes as input a set of multiple sequence alignments, \(\{X_i\}\), where \(X_i\) is the alignment associated with locus \(i\). Each sequence in such an alignment represents a haploid sample mapped to some current population (leaf) in \(T\). The main objective of the MCMC algorithm is to sample model parameters according to their joint posterior density \(P(\{\theta_p\}, \{-\tau_p\}; T, \{X_i\})\). In order to achieve this goal, MCMCcoal introduces a layer of “latent” (unobserved) locus-specific genealogies \(\{G_i\}\), which it samples together with the demographic parameters. The algorithm maintains a version of all latent variables \((\{\theta_p\}, \{-\tau_p\}, \{G_i\})\) and iteratively proposes updates to them, accepting or rejecting these updates according to the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). The algorithm therefore consists of two main components: the computation of the complete data density function \(P(\{X_i\}, \{G_i\}, \{\theta_p\}, \{-\tau_p\}; T)\), and the update scheme for the latent variables \((\{\theta_p\}, \{-\tau_p\}, \{G_i\})\).

**The complete data density function**

Efficient computation of the complete data density function is facilitated by the following factorization:

\[
P(\{X_i\}, \{G_i\}, \{\theta_p\}, \{-\tau_p\}; T) = \left( \prod_p P(\theta_p) \right) \left( \prod_p P(\tau_p) \right) \left( \prod_i P(G_i \mid T, \{\theta_p\}, \{-\tau_p\}) P(X_i \mid G_i) \right),
\]

where the second product includes only the ancestral populations.

\(^1\)This version of MCMCcoal has since been superceded by a new program, called BP&P (Yang and Rannala, 2010). The improvements in BP&P appear to be mostly concerned with uncertainty in species delimitation and are therefore orthogonal to our goals in this study.
Figure B: The coalescent model of MCMCcoal. A locus-specific genealogy $G_i$ is shown embedded in a population “tube-tree” $T$ for three current populations (A, B, and C) and two ancestral populations (AB and ABC). Each of the five populations is associated with a size parameter ($\theta_A$, $\theta_B$, $\theta_C$, $\theta_{AB}$, and $\theta_{ABC}$), and the two ancestral populations are also associated with population divergence times ($\tau_{AB}$ and $\tau_{ABC}$). Assuming no gene flow between populations, the branches of all locus-specific genealogies are confined within the tube-tree structure (top). The contribution of a locus $i$ to the complete data density, $P(G_i, X_i \mid T, \{\theta_p\}, \{\tau_p\})$, can be expressed as a product of a genealogy prior, $P(G_i \mid T, \{\theta_p\}, \{\tau_p\})$, and a locus data likelihood, $P(X_i \mid G_i)$. The locus data likelihood (bottom right) depends only on the genealogy and the observed data, not the population model, and is computed using Felsenstein’s pruning algorithm. The genealogy prior (bottom left), in turn, depends only on the genealogy and the population model, but not on the observed data. This quantity can be expressed in terms of the partitioning of the tube tree into timed intervals that is induced by the genealogy $G_i$. The sufficient statistics for each interval are (1) whether or not it ends in a coalescent event, (2) its elapsed time, and (3) the number of lineage of $G_i$ living during this time. The contribution of each interval to the genealogy prior is determined by the expression in Equation (16). In this example, the total contribution of the three intervals corresponding to population B is: $$\left(\frac{2}{\theta_B}\right)^2 \exp \left\{ -\frac{12t_1+6(t_2-t_1)+1(\tau_{AB}-t_2)}{\theta_B} \right\}.$$
This factorization reflects several independence assumptions. First, independent priors are assumed for the model parameters, \( \{\theta_p\} \) and \( \{\tau_p\} \). These priors are set as diffuse Gamma distributions defined by the user. Second, the genealogies at distinct loci are assumed to be independent given the model parameters (as should be true approximately with sufficient interlocus recombination). Third, each locus-specific genealogy \( G_i \) induces conditional independence of the corresponding alignment, \( X_i \), and the model parameters, \( \{\theta_p\}, \{\tau_p\} \). Hence, the contribution of each locus to the complete data density function is composed of: (a) the genealogy prior, \( P(G_i \mid T, \{\theta_p\}, \{\tau_p\}) \), which does not depend on the data; (b) the locus data likelihood, \( P(X_i \mid G_i) \), which does not depend on the demographic parameters and is computed using Felsenstein’s pruning algorithm (Felsenstein, 1973, 1981) (see Figure B and Sections S4.3 and S4.4).

### Updating the latent variables

Each update step (iteration) of the full set of latent variables, \( \{\theta_p\}, \{\tau_p\}, \{G_i\} \), is divided into a series of Metropolis-Hastings updates of subsets of variables. Some of these component updates are local, meaning that they only affect an individual locus-specific genealogy, \( G_i \), while others are global, meaning that they affect the model parameters, \( \{\theta_p\} \) and \( \{\tau_p\} \), which influence all loci. Specifically, the update steps are:

1. **Update coalescent times (local):** This step slightly perturbs the time of an individual coalescent event without changing the topology of the genealogy or any other coalescent times. It is applied to each coalescent event in each genealogy.

2. **Update genealogy structure (local):** This step alters an individual genealogy via a subtree-prune-and-regraft operation. It is applied to each subtree of each genealogy.

3. **Update \( \theta_p \) (global):** This step slightly perturbs \( \theta_p \). It is applied to each population \( p \).

4. **Update \( \tau_p \) (global):** This step slightly perturbs \( \tau_p \), for each ancestral population \( p \). This step requires modifying each genealogy \( G_i \), as needed to accommodate the proposed change in \( \tau_p \). The modification of these genealogies is accomplished by a rubber-band operation, which “squeezes” some parts of the genealogy and “stretches” others to satisfy the constraints of the population model.

5. **Rescaling (global):** This step slightly perturbs all model parameters \( \{\theta_p\}, \{\tau_p\} \) and all coalescent times across all genealogies by the same multiplicative factor (sampled close to 1). Rannala and Yang (2003) describe this as a mixing step, because it improves mixing of the Markov chain.

Each type of update is associated with a user-defined fine-tune parameter, which controls the overall “boldness” of the proposed moves. The user can adjust these parameters so that steps are large enough for the sampler to mix well but small enough to allow for a reasonable acceptance ratio. The MCMCcoal manual suggests aiming for acceptance ratios of \( \sim 30\% \).

### Variable mutation rates across loci

Burgess and Yang (2008) implemented two approaches for modeling variable mutation rates across loci within MCMCcoal. Both approaches use an additional variable per locus to model its local mutation rate. This variable functions as a scaling factor for the edges in \( G_i \) when computing \( P(X_i \mid G_i) \). These locus-specific rates are normalized to have an average of one, making the scale of all model parameters determined by the average (rather than a constant) mutation rate.

The first model for rate variation, called the “fixed rates” model, requires that these locus-specific rates are estimated separately (say, from sequence divergence to an outgroup species) and supplied by the user. These user-supplied rates are not modified by the algorithm during the sampling process. The second model,
called the “random rates” model, defines a Dirichlet prior on these rates, with a single $\alpha$ parameter that applies to all loci. The algorithm then samples the locus-specific rates along with the other latent variables, updating them separately at each locus on each sampling iteration. Thus, the “fixed rates” model essentially conditions on pre-defined locus-specific mutation rates, while the “random rates” model integrates these rates out of the analysis. Both of these models are implemented in G-PhoCS and we have compared their performance in our data analysis (Section S8.2).

Figure C: The coalescent-with-migration model of G-PhoCS. Here the simple three-population demographic model from Figure B is augmented with two migration bands, B→A and C→B. Genealogical lineages are now allowed to cross population boundaries within these bands (top). As in the original model, the contribution of locus $i$ to the complete data density, $P(G_i, X_i | \{\theta_p\}, \{\tau_p\}, \{m_b\})$, is expressed as a product of a genealogy prior $P(G_i | T, \{\theta_p\}, \{\tau_p\}, \{m_b\})$ and a locus data likelihood $P(X_i | G_i)$. The locus data likelihood (bottom right) is unaffected by the migration model and the presence of migration events in the genealogy. The genealogy prior (bottom left) continues to be computed using the partition (induced by $G_i$) of the population tree into a series of time intervals. However, each interval now contributes to both a coalescent prior (Equation (16)) and a migration prior (Equation (18)). The six intervals of population B in this example contribute a total of $\left(\frac{2}{\theta_B}\right)^3 \exp \left\{ -\frac{12(t_1+6(t_m_1-t_1)+12(t_2-t_m_2)+6(t_m_2-t_2)+1(t_3-t_m_3)+0(\tau_{AB}-t_3))}{\theta_B} \right\}$ to the coalescent prior and a total of $m_{CB} \cdot \exp \left\{ -m_{CB} \cdot (4t_1+3(t_m_1-t_1)+4(t_2-t_m_1)+3(t_m_2-t_2)+2(t_3-t_m_2)+1(\tau_{AB}-t_3)) \right\}$ to the migration prior of band C→B.
S4.2 Modeling Gene Flow in G-PhoCS

The migration model

We introduce gene flow into the MCMCcoal model through migration bands. A migration band is associated with a directed pair of (ancestral or current) populations $S \to T$, where $S$ is the source population and $T$ is the target population (in forward time). A population tree can be augmented with any number of migration bands, at the user’s discretion. The life span of each band is the time interval during which both $S$ and $T$ exist.\(^2\) Gene flow along a migration band is modeled as a constant-rate process (Beerli and Felsenstein, 1999; Nielsen and Wakeley, 2001). Specifically, each migration band $S \to T$ is associated with a rate parameter $m_{ST}$, which is a mutation-scaled version of an instantaneous migration rate $M_{ST}$, with $m_{ST} = \frac{M_{ST}}{\mu}$. $M_{ST}$ is defined as the proportion of individuals in population $T$ that arose by migration from population $S$ per generation. Thus, in the context of the backward-in-time coalescent process, migration can be modeled by allowing lineages in the target population $T$ to migrate into the source population $S$ by a Poisson process with rate $m_{ST}$. Each migration rate $m_{ST}$ is associated with an independent Gamma prior, which is defined by the user. In the MCMC sampling algorithm, it is necessary to augment each genealogy $G_i$ with dated migration events indicating these inter-population lineage crossings (Figure C). These events are incorporated into the genealogy prior in this model according to coalescent-with-migration process described by Beerli and Felsenstein (1999) and Nielsen and Wakeley (2001) (see also Section S4.4). Importantly, however, the locus data likelihood, $P(X_i|G_i)$, is unaffected by these events, as it depends only on the genealogy $G_i$, not the manner in which $G_i$ is nested within the population phylogeny.

Modifications to the sampler

The update steps of the Markov Chain required quite extensive modifications in order to accommodate genealogies with migration events. These modifications are summarized below:

1. **Update coalescent times:** This step is essentially unchanged, but now leaves migration times as well as other coalescent times unchanged.

2. **Update migration times:** This is a new, local update step, in which existing migration times are perturbed without changing the topology of the genealogy or the times of other events.

3. **Update genealogy structure:** The “regraft” portion of the original subtree-prune-and-regraft operation had to be altered to account for migration events. The new method for proposing candidate regrafting updates considers the current values of all demographic parameters, including the migration rates. Specifically, a lineage is simulated from the root of the pruned subtree back in time, possibly with migration, until it eventually coalesces with an existing part of the genealogy (Figure D; a similar update step was used by Beerli and Felsenstein, 1999). Migration events may be eliminated during the pruning step, and new events may be re-sampled during the regrafting step. Indeed, this is the only step in the sampler that alters the set of migration events associated with the latent genealogies. Notably, because this new procedure makes direct use of the prior distribution, it no longer depends on a user-defined fine-tune parameter. We observed a fairly high acceptance ratio for this step ($\sim$50%).

4. **Update $\theta_p$:** This step is unaffected by the addition of migration.

5. **Update $m_b$:** In this step, the migration rates for all migration bands are perturbed, as with the $\theta_p$ parameters.

\(^2\)In principle, the model can be modified to allow inference of the start and end time of each migration band, but, at the time scales of interest to us, there is almost no information about such parameters in the data.
6. **Update** $\tau_p$: Migration events induce complex dependencies between the coalescent processes on different lineages, which must be considered during the rubber-band operation. We address these dependencies using a simple rejection-sampling scheme, to minimize modifications to the original algorithm. During the proposal of a divergence time update, the rubber-band operation is applied as in the original algorithm, and if any conflicts occur due to a migration events in one of the sampled genealogies, the suggested update is rejected (Figure E). This simple approach seems to be adequate with moderate levels of gene flow. However, in cases in which gene flow is very frequent—i.e., with high rates of gene flow and/or many loci—it forces the use of small update steps for $\tau$, which can slow convergence of the Markov chain.

Figure D: Genealogy structure update step of *G-PhoCS*. A branch of the genealogy (red in upper left panel) is detached, leaving a hanging subtree (upper right panel). A lineage is then simulated back in time from the root of that subtree according to the current population parameters, until it coalesces with another lineage in the genealogy (bottom panel). Possible migrations, along designated migration bands, are considered during this procedure (for example, through band C→B in this case).

Figure E: Population divergence time update step of *G-PhoCS*. When an update to a population divergence time is proposed, the times of events in parent and child populations are scaled accordingly using the “rubber-band” procedure described by Rannala and Yang (2003). With migration, this scaling step can lead to conflicts with events in other populations. If such a conflict occurs, the suggested update is rejected. In this example, an increase in $\tau_{AB}$ “pushes back” migration event $m_2$ via the rubber band, but this event is bounded by a coalescent event in population C (at time $t_6$), which is unaffected by the rubber-band.
7. **Rescaling:** In this step, the times of migration events and the migration rates are scaled along with the other times and parameters. The migration rates are scaled by the inverse of the factor applied to \( \{ \theta_p \} \) and \( \{ \tau_p \} \), since \( 1/m \) defines the expected waiting time for a migration event.

### S4.3 Accommodating Diploid Genomes

Each of the diploid individuals in our analysis contains two haploid samples with information relevant to our analysis, but in our raw data these haploid samples are intertwined and represented only as a sequence of site-wise diploid genotypes. It is of course possible to disentangle these haploid sequences using automatic phasing tools, but this process is imperfect, particularly in cases in which high-quality genotype reference panels are not available, as with our Southern African genomes. Moreover, differences in phasing accuracy between the genomes in our study could produce biases in our analysis.

To avoid these problems, we took the approach of integrating over all possible phasings in our analysis, considering them to be equally likely *a priori*. We devised a method for performing this phasing integration that is fairly efficient, at least for small numbers of individuals. In simulation studies, this method appears to work essentially as well as conditioning on the true phase for the purposes of our analysis (Section S6.3), and its computational cost is low.

**The theory and practice behind phasing integration**

In this section we describe how the locus likelihood \( P(X_i | G_i) \) is computed when an alignment \( X_i \) contains unphased diploid genotypes, consisting of both homozygous (homs) and heterozygous (hets) genotypes. For simplicity, we omit the subscript \( i \) in this discussion, and assume that the alignment \( X \) spans \( k \) diploid samples and \( n \) sites. (An alignment including both haploid and diploid samples can easily be accommodated within this setting as well.)

Let the *phasing* \( P \) of a locus be a \( k \times n \) binary matrix defining the manner in which each position in each diploid genome is partitioned between the two associated haploid samples. The locus-likelihood can be calculated by summing over all possible phasings as follows:

\[
P(X | G) = \sum_{P \in \{0,1\}^{k \times n}} P(X | P, G) P(P),
\]

where we assume the prior \( P(P) \) does not depend on the genealogy. Notice that the conditional likelihood \( P(X | P, G) \) can be computed in the ordinary way for haploid samples, because the phasing \( P \) allows a haploid genotype to be unambiguously assigned to each leaf of the genealogy at each position in the alignment. In practice, the summation only needs to consider phasing assignments of hets, because the conditional likelihood is invariant to the assignments for homs. Nevertheless, the number of distinct phasings of \( X \) is still exponential in the total number of hets in \( X \), which will in general make integration by simple enumeration infeasible.

Let us further assume a uniform prior distribution \( P(P) \). Using conditional independence of the alignment columns given the genealogy and phasing, and uniformity of the phasing prior, equation 14 can be rewritten as:

\[
P(X | G) = \prod_j \left( \frac{1}{2^{|H_j|}} \sum_{P^j \in \{0,1\}^{|H_j|}} P(X^j | P^j, G) \right),
\]

where \( X^j \) denotes the \( j \)th column of \( X \), and \( P^j \) denotes a binary phasing vector for the hets at column \( j \) (denoted \( H_j \)).
Thus, the computation of the locus-likelihood decomposes into a series of computations of site-wise phased likelihoods, $P(X^j \mid P^j, G)$. The computational complexity of the procedure is now exponential only in $\max_j |H_j|$, the maximum number of hets in a single alignment column. Each site-wise phased likelihood computation can be performed using Felsenstein’s pruning algorithm (Felsenstein, 1973, 1981) under the Jukes-Cantor substitution model (Jukes and Cantor, 1969), as in the haploid case assumed by $MCMCcoal$. Although highly simplified, the Jukes-Cantor model provides a reasonable approximation for likelihood computations in the time scales of human evolution. In addition, assuming this simple and symmetric model allows a significant reduction in running time through careful alignment pre-processing (details below).

Alignment preprocessing

The high degree of symmetry in the Jukes-Cantor substitution model leads to large equivalence classes of alignment “site patterns”, all having the same conditional likelihood for any choice of genealogy. Substantial savings in computational cost are realized in $MCMCcoal$ by identifying these equivalence classes and evaluating the conditional likelihood once per class. This task is slightly more complicated with unphased diploid alignments, but it can still be solved fairly easily. We omit the full technical details and provide an example in Figure F. After identifying these equivalence classes (i.e., the collection of distinct unphased site patterns), we compute for each unphased pattern the set of phased patterns associated with it. The size of this set is exponential in the number of hets in the unphased pattern.

An additional reduction in the number of effective phased site patterns can be achieved by considering the inherent symmetry between two chromosome samples of the same individual. These two chromosomes become distinct only when the first het is phased, meaning that for each diploid sample we can arbitrarily phase one het per locus, without altering the results of the inference (allowing for the fact that the genealogies will essentially be “integrated out” of the analysis during the course of the analysis). We refer to this as a symmetry breaking scheme for $X$. It turns out that different schemes can result in different total numbers of phased site patterns, and the choice of the symmetry breaking scheme can significantly impact the overall running time of $G$-$PhoCS$. We describe the strategy we take for breaking symmetries in greater detail below.

Breaking diploid symmetries

To find a symmetry breaking scheme for $X$, we must choose a single column in which to make an arbitrary phasing decision for each diploid individual. It turns out that choosing a good symmetry breaking scheme can be a nontrivial task. There are two main factors to consider. First, there is no use in arbitrarily phasing a het in a column if another column in the alignment has the same site pattern (since all phased instances have to be considered anyway). Second, the reduction in number of phased patterns realized by arbitrarily phasing a het depends on the total number of hets in the column in question. In particular, phasing a het in a column with $k$ hets results in the elimination of $2^{k-1}$ phased site patterns. Thus, it is advantageous to choose columns with many hets. In our implementation we use a greedy approach that iteratively chooses, for each individual, a column maximizing the number of unphased hets subject to the constraint that it has a unique site pattern.

The following (somewhat contrived) example demonstrates that different symmetry breaking strategies can produce dramatic differences in the total number of phased site patterns. Consider an alignment of $k$ diploid sequences and assume that it contains $k + 1$ columns with hets. In particular, column $i$ ($i \in [1..k]$) has a single het in sequence $i$ and a hom in every other sequence, while column $k + 1$ has $k$ hets, one for each of the $k$ sequences. For each individual $i \in [1..k]$, we can break the symmetry between its two chromosomes in that locus by arbitrarily phasing its hets either in column $i$ or in column $k + 1$. Assume we choose to break the symmetry of all individuals by arbitrarily phasing their hets in column $k + 1$. This would result in $2^{k-1}$ phased site patterns for each of the columns $1..k$, and a single phased site pattern for column $k + 1$, for a total of $2^{k+1}$ phased patterns. Now assume instead that we choose to break the symmetry of every individual $i$ by
arbitrarily phasing column \( i \). That would result in one phased instance for each of the columns \( 1..k \), and \( 2^k \) phased instances for column \( k + 1 \), for a total of \( 2^k + k \) phased site patterns. Thus, an exponential reduction in the number of phased patterns can be achieved by carefully selected symmetry breaking scheme.

Figure F: Alignment preprocessing in G-PhoCS. Shown is a locus-specific alignment of length 20 including three diploid individuals (b, c, and d) and one haploid sequence (a). Heterozygous genotypes are indicated by IUPAC ambiguity characters and marked in light red, while homozygous genotypes are indicated by \{A, C, G, T\} and marked in dark red. Missing data is indicated by green “N”s. In the first stage of preprocessing, columns are grouped according to unphased site patterns, taking into consideration symmetries implied by the Jukes-Cantor substitution model. Notice the non-trivial grouping of columns 5 and 7. In the second stage, all possible phased site patterns are enumerated. The number of phased site pattern associated with each unphased pattern is exponential in the number of heterozygotes it contains. In the third stage, a “symmetry breaking” site is chosen for each diploid individual. This reflects the fact that there is no “ground truth” in phasing; all that matters is the relationship of individual sites to one another. The chosen site should have a unique unphased site pattern (i.e., it should not be grouped with other sites in the first stage), and a large number of distinct phases. An optimal symmetry breaking scheme in this case chooses site 9 for individuals c and d (reducing the number of phased patterns associated with it from 8 to 2) and site 16 for individual b. The total number of patterns for which likelihoods must be computed is reduced from 21 to 13. In this case, the optimal solution is not unique.
S4.4 New Framework for Computing Genealogy Priors

One important cost-saving strategy used by MCMCcoal is to cache local information used in the pruning algorithm and recompute only the components that are affected by each update. However, MCMCcoal does not make use of a similar strategy in calculating the genealogy priors \( P(G_i \mid \{\theta_p\}, \{\tau_p\}) \), which it recomputes from scratch each time. We noticed that the genealogy prior can be decomposed into local factors, and implemented data structures in G-PhoCS that store this local information and thus allow for faster updates. This approach leads to a significant (~20%) reduction in running time compared with MCMCcoal. Importantly, this benefit is realized whether or not gene flow is assumed. In the discussion below, we describe this factorization for the case without gene flow, and then describe the adjustments required to accommodate migration. More details on the derivation of the genealogy prior can be found in Beerli and Felsenstein (1999), Nielsen and Wakeley (2001), and Rannala and Yang (2003).

In the absence of gene flow

Coalescence is modeled as a Poisson process with a constant rate determined by the effective population size. Given a set of \( n \) lineages coalescing back in time in a population of (mutation-scaled) effective size \( \theta \), the (mutation-scaled) time until the next coalescent event is exponentially distributed with rate \( (\frac{n^2}{2}) \frac{2}{\theta} = \frac{n(n-1)}{\theta} \).

Given that a coalescent event has occurred, it is equally likely to have involved any of the \( \binom{n}{2} \) possible pairs of lineages. Therefore, the probability of observing a time interval \( I \) of length \( \tau \) in a population with effective size \( \theta_p \) with a constant number of lineages \( n \) is given by:

\[
P(I \mid \theta_p) = \left( \frac{2}{\theta_p} \right)^{\text{coal}(I)} \cdot \exp \left\{ -\frac{(n^2 - n)}{\theta_p} \tau \right\},
\]

where \( \text{coal}(I) \in \{0, 1\} \), indicates whether \( I \) ends in a coalescent event or not.

Therefore, the genealogy prior can be computed by partitioning each population into timed intervals according to times of coalescent events. G-PhoCS maintains for each population \( p \) its set of intervals \( \mathcal{I}NT(p) \), and for each interval it holds the three sufficient statistics that determine its contribution to the genealogy prior: \( \text{coal}(I) \), \( \tau(I) \), and \( n(I) \) (Figure B). The genealogy prior is then factored as follows:

\[
P(G_i \mid \{\theta_p\}, \{\tau_p\}) = \prod_p \left( \prod_{I \in \mathcal{I}NT(p)} P(I \mid \theta_p) \right).
\]

When modeling gene flow

Migration is also modeled by a Poisson process with a separate constant rate for every migration band. For each migration band \( b \), G-PhoCS maintains the set of intervals \( \mathcal{I}NT(b) \) in its target population (in the life span of the migration band; Figure C). A timed interval \( I \in \mathcal{I}NT(b) \) of length \( \tau \) with \( n \) lineages contributes the following amount to the genealogy prior:

\[
P(I \mid m_b) = m_b \exp \left\{ -m_b n \tau \right\},
\]

where \( \text{mig}(I, b) \in \{0, 1\} \) indicates whether \( I \) ends in an (incoming) migration event in band \( b \). The adjusted genealogy prior is then given by:

\[
P(G_i \mid \{\theta_p\}, \{\tau_p\}, \{m_b\}) = \left( \prod_p \left( \prod_{I \in \mathcal{I}NT(p)} P(I \mid \theta_p) \right) \right) \left( \prod_b \left( \prod_{I \in \mathcal{I}NT(b)} P(I \mid m_b) \right) \right).
\]
S5 Applying G-PhoCS to Data

This section presents various issues we encountered when applying G-PhoCS to data.

S5.1 MCMC Setup

The Markov chain explored by the algorithm is influenced by the user-defined arguments listed below. The settings used in our simulated test cases and data analysis are detailed in Table S6.

**Finetune parameters for update steps.** As described in Section S4, the update steps of the MCMC are controlled by a set of finetune parameters, which control the acceptance rates for the various updates. These were set in order to achieve acceptance ratios of 20–50%.

**Parameter priors.** Uninformative Gamma priors were used for all demographic parameters in the model. A single prior was used for all $\theta$ parameters, and another was used for all migration rates. The priors for the migration rates were diffuse, with a low mean and large variance, as were the priors for the population sizes (see Table S6). Different priors were used for each of the population divergence times $\tau$. In the case of the “random rates” model for rate variation across loci, a diffuse Dirichlet prior was used for the locus-specific rates (Section S8.2). We experimented with various choices of uninformative prior distributions and observed no significant effect on the resulting estimates (see Section S8.1). However, we did find that the priors for divergence times could significantly impact the time required for convergence of the Markov chain, with poorly chosen priors producing very long burn-in periods. Therefore, after a series of preliminary runs, we allowed these distributions to be set such that the prior means were in the neighborhood of the expected posterior means, to avoid many days of additional computational time. The Gamma priors that were chosen are still sufficiently diffuse not to strongly influence the posterior. Furthermore, runs with different sets of priors were shown to lead to very similar estimates (see Section S8.1).

**Parameter initialization.** Each demographic parameter is initialized to a value chosen uniformly at random in the interval $[0.8X, 1.2X]$, where $X$ is the prior mean. We tested other starting points and observed no notable influence of the starting point on the resulting estimates, but, as discussed above, poor initializations could substantially lengthen the burn-in time.

**Numbers of iterations.** The burn-in period was generally set to 100,000 iterations. We found that this was sufficient to establish convergence for all of the main demographic parameters of interest (see Section S5.3). In most cases, convergence was observed well before this threshold (at $\sim30,000$ iterations). After the burn-in, we collected samples of the demographic parameters every 10 iterations, to reduce the correlation between samples. Sample collection typically proceeded for an additional 200,000 iterations. We note that for some of our data analysis, we had to reduce the sampling period slightly (never $<150,000$ iterations), due to running-time constraints (see Section S5.5).

S5.2 Migration Bands

The migration model implemented in G-PhoCS is quite general and, in principle, could be used to make inferences about complex migration scenarios. Indeed, we found in simulation experiments that the algorithm could recover fairly complex scenarios—involving multiple migration bands, with unidirectional or bidirectional migration—in cases in which the simulated sequence data was highly informative about the parameters of interest, for example, due to high migration rates, large divergence times, and/or high mutation rates. However, the signal for migration appears to be quite weak for the human populations in our study,
owing to low mutation rates (relative to the time scale in question) and large amounts of ancestral coalescence. Experiments conducted with simulated data confirm that it is difficult to detect complex migration scenarios in this setting. However, it does appear to be possible to accommodate migration sufficiently well to avoid obtaining biased estimates of divergence times (Section S6). Therefore, we view the migration bands in our model primarily as a means for allowing for violations to a pure “isolation” model in the estimation of divergence times, rather than as tools for accurately measuring rates of inter-population gene flow.

Because of our limited ability to characterize complex migration scenarios, we used the strategy of performing a large number of different G-PhoCS analyses with various choices of simple migration scenarios. We considered one scenario with no migration bands, and several scenarios with a single migration band each. In this way, we were able to obtain a “marginal” indication of the amount of migration along each band. We measured these marginal migration rates (or “total migration rates”) as a product of the estimated scaled migration rate and life span of the band, which gives an approximate estimate of the probability that each lineage in the target population passes through the migration band into the source population during its history (looking backward in time). In our simulation experiments, we were able to estimate these rates reasonably accurately (Figure S11C).

Experiments with simulated data also indicated that, on a human population genetic timescale, the directionality of migration was nearly impossible to detect by our methods. Nevertheless, divergence times were accurately estimated even when an incorrect migration direction was assumed during inference. We also noticed that directing a band from an “older” population to a “younger” one (e.g., S→Y, in our five population model; Figure 1, main text) led to increased power in detecting migration, regardless of the true direction. Therefore, in our final analysis presented in Section S7 we direct each band according to this guideline.

**S5.3 Assessing Convergence and Mixing**

Obtaining reliable estimates by MCMC requires ensuring both that the Markov chain has converged (i.e., the burn-in period is sufficiently long), and that it is mixing adequately well to explore the parameter space and produce a sufficiently large number of effectively uncorrelated samples. This is a particularly challenging task in our setting, because of our complex model, large data sets, and long running times. We used various strategies to make sure that our methods were adequate for obtaining reliable approximations of the marginal posterior distributions for the parameters of primary interest.

We began by performing numerous preliminary analyses (on both real and simulated data) to establish a conservative burn-in period for our sampler. The traces of the Markov chain in these runs were monitored visually using Tracer v1.5 (Rambaut and Drummond, 2007) to observe convergence for each parameter in the model. Some parameters, such as $\tau_{KHEXS}$ and $\theta_{KHEXS}$, appear to converge rapidly, while others, such as $\tau_{KHE}$ and $\theta_{KHE}$, require much longer periods to converge, evidently due to a weak signal in the data. We selected 100,000 iterations as a long enough burn-in for all parameters of primary interest to have converged, while keeping running times short enough to allow a variety of demographic scenarios and data sets to be examined (Figure G).

Mixing is a serious challenge in our setting, in large part because the consideration of a very large number of loci requires conservative updates for the model parameters, especially the population divergence times (see discussion of “rubber-band” in Sections S4.1 and S4.2). Mixing was monitored using the Auto-Correlation Time (ACT) statistic provided by Tracer, which is the number of iterations that two samples have to be from each other for them to be uncorrelated. The ACT varies extensively across the different parameters. For example, for $\tau_{KHEXS}$, $\theta_{KHEXS}$, $\theta_{S}$, and $\theta_{X}$ it was typically less than 1,000; for $\tau_{KHEX}$ and $\theta_{KHEX}$ it was typically around 10,000; and for parameters in the Eurasian subtree it was typically around 30,000. Differences in the quality of mixing were also clearly visible from the parameter traces (Figure G).
Because mixing was poor in some cases, we made use of multiple runs of the sampler to help confirm that reliable estimates were being obtained. In most cases, we observed a high degree of consistency of mean estimates and 95% credible intervals across runs, despite poor mixing of some parameters within each run (see Figures S9 and S12). These multiple runs could be conveniently performed in parallel on a computer cluster, while each run of an individual Markov chain was necessarily sequential.

Figure G: Examples of MCMC traces for four demographic parameters in our five-population analysis. Each trace plot (obtained using Tracer) displays the traces for two separate runs: run #15 (black) and run #16 (blue) (as described in Table S3). The shaded area represents the burn-in period of 100,000 iterations. Note that $\tau_{KHEXS}$ and $\theta_{KHEXS}$ exhibit somewhat faster convergence and better mixing than $\tau_{KHEX}$ and $\theta_{KHEX}$.

**S5.4 Parameter Estimates and Credible Intervals**

The marginal posterior distribution for each demographic parameter was approximated based on the samples retained from the MCMC algorithm, using standard methods. For point estimates of each parameter, we used the sample mean of the retained samples, and for credible interval (CI) we used the 95% highest posterior density interval, which is a credible set that contains 95% of the sampled values (as defined in Tracer). These 95% CIs are useful in summarizing the uncertainty associated with our parameter estimates (given our modeling assumptions and data), but in some cases the estimated intervals probably underestimate the true uncertainty in the posterior distribution, due to incomplete mixing of the Markov chain. This is especially an issue with some of the parameters for which the signal in the data is weak, such as those associated with the Eurasian populations.

**S5.5 Running Time**

*G-PhoCS* is designed to be efficient enough to analyze tens of thousands of loci. Many of the properties that enable this efficiency are directly inherited from *MCMCcoal*, while others are due to our modifications (Section S4.4). Table S7 provides examples of typical running times obtained for both simulated and real data. By comparing running times for correctly phased simulated sequences with running times for the
unphased diploid version of these sequences, we estimate that the phasing integration procedure increases the computational cost by $\sim 10\%$. With the real data, we observe a further increase due to increased numbers of site patterns per locus, mostly stemming from the introduction of “N”s by our “soft” filters (Section S1.2).

S6 Simulation Study

In order to test the accuracy of G-PhoCS in estimating demographic parameters, we conducted an extensive simulation study. We present here a summary of this study using several test cases that demonstrate our main findings.

S6.1 Generating Simulated Data

Each simulated data set was obtained by generating locus-specific genealogies using the MS coalescent simulator (Hudson, 2002), and evolving sequences along these genealogies under a Jukes-Cantor substitution model (Jukes and Cantor, 1969) using SeqGen (Rambaut and Grassly, 1997). In the simulations described here we considered five diploid individuals (each modeled using two separately generated haploid samples) and a single haploid outgroup (representing the chimpanzee sequence), for a total of 11 chromosomes per locus. The coalescent process was simulated within the population phylogeny depicted in Figure 1 (main text). All demographic parameter were set to be in a region of parameter space consistent with human population evolution. Most parameters were given fixed values for all simulations, while a few parameters of particular interest were varied, namely the San and African-Eurasian divergence times and migration rates. The fixed parameters were set as follows:

| Parameter | Value |
|-----------|-------|
| $\theta_E$ | $5 \times 10^{-4}$ |
| $\theta_K$ | $5 \times 10^{-4}$ |
| $\theta_H$ | $5 \times 10^{-4}$ |
| $\theta_X$ | $15 \times 10^{-4}$ |
| $\theta_S$ | $15 \times 10^{-4}$ |
| $\theta_C$ | $10 \times 10^{-4}$ |
| $\theta_{KH}$ | $5 \times 10^{-4}$ |
| $\theta_{KHE}$ | $1 \times 10^{-4}$ |
| $\theta_{KHEX}$ | $10 \times 10^{-4}$ |
| $\theta_{KHEXS}$ | $5 \times 10^{-4}$ |
| $\theta_{root}$ | $10 \times 10^{-4}$ |
| $\tau_{KH}$ | $0.1 \times 10^{-4}$ |
| $\tau_{KHEX}$ | $variable$ |
| $\tau_{KHEXS}$ | $variable$ |
| $\tau_{root}$ | $41 \times 10^{-4}$ |

Several data sets were generated for each demographic scenario. Each data set included 20,000 loci (sequence alignments) of length 1000 bp. The simulated haploid sequences were used to produce unphased diploid genomes similar to those in the real data. The correctly phased haplotype version of each data set was retained for comparison (see Section S6.3). Below is an example of command line arguments used when generating a simulated data set with the following settings for the variable demographic parameter: $\tau_{KHEX} = 0.5 \times 10^{-4}$, $\tau_{KHEXS} = 1.0 \times 10^{-4}$, $m_{S\rightarrow X} = 2000$.

MS

```
11 20000 -T -I 6 2 2 2 2 2 1
-n 1 0.0005 -n 2 0.0005 -n 3 0.0005 -n 4 0.0015 -n 5 0.0015 -n 6 0.0010
-m 4 5 2000
-ej 0.00001 3 2 -en 0.00001 2 0.0005
-ej 0.00002 2 1 -en 0.00002 1 0.0001
-em 0.00005 4 5 0.0
-ej 0.00005 4 1 -en 0.00005 1 0.0010
-ej 0.00010 5 1 -en 0.00010 1 0.0005
-ej 0.00010 6 1 -en 0.00010 1 0.0010
> genealogies.txt
```

SeqGen -mHKY -l 1000 < genealogies.txt > sequences.txt

52
S6.2 Applying *G-PhoCS* to Simulated Data Sets

*G-PhoCS* was applied to each data set as described in Section S5.1 and Table S6. In some cases multiple runs were used to assess convergence and quality of mixing of the MCMC (see Section S5.3). Typical running times are detailed in Table S7.

S6.3 Simulations without Gene Flow

The simulations without gene flow considered nine demographic scenarios, with San divergence times of $\tau_{KHEXS} \in \{0.8, 1.0, 1.2\} \times 10^{-4}$ and African-Eurasian divergence times of $\tau_{KHEX} \in \{0.3, 0.4, 0.5\} \times 10^{-4}$. Two data sets were generated per scenario (see list of data set id’s below), and two runs of the sampler were performed for each of the data sets, for a total of $9 \times 2 \times 2 = 36$ runs. Parameter estimates and 95% CIs for these 36 runs are presented in Figure S9. In order to quantify the estimation accuracy for each demographic parameter, we calculated the ratios of estimated (*G-PhoCS* sample means) to true values and summarized them for the 18 different data sets. A single *G-PhoCS* run per data set was used for this purpose.

In addition to the analysis of diploid data using phasing integration, *G-PhoCS* runs were also executed on the correctly phased version of each data set and a randomly phased version. A summary of parameter estimation accuracy across these $18 \times 3$ runs is provided in Figure S10.

| id  | $\tau_{KHEX}$ | $\tau_{KHEXS}$ | id  | $\tau_{KHEX}$ | $\tau_{KHEXS}$ | id  | $\tau_{KHEX}$ | $\tau_{KHEXS}$ |
|-----|---------------|----------------|-----|---------------|----------------|-----|---------------|----------------|
| 1, 2 | 0.00003       | 0.00008        | 7, 8 | 0.00003       | 0.00010        | 13, 14 | 0.00003       | 0.00012       |
| 3, 4 | 0.00004       | 0.00008        | 9, 10 | 0.00004       | 0.00010        | 15, 16 | 0.00004       | 0.00012       |
| 5, 6 | 0.00005       | 0.00008        | 11, 12 | 0.00005       | 0.00010        | 17, 18 | 0.00005       | 0.00012       |

The San and African-Eurasian divergence times

The “phasing integration” strategy implemented in *G-PhoCS* achieves very good accuracy in estimating the two divergence times of primary interest (Figure S10A), and the estimation accuracy appears to improve as the divergence time increases. In general, the estimates of the African-Eurasian divergence time appear to be somewhat less reliable than the estimates of the San divergence time. Convergence and mixing are also better for the San divergence time than for the more recent population divergences. Estimates obtained under the phasing integration implemented in *G-PhoCS* are as accurate as those obtained in the analysis of correctly phased data. By contrast, the naive random phasing procedure produces a clear upward bias in the divergence times. These inflated divergence times arise because, in the presence of phasing errors, additional mutations are required to reconcile the observed data with plausible genealogies, and these additional mutations accumulate on external branches of the tree.

Other demographic parameters

Panels B and C in Figure S10 describe the estimation accuracy for the remaining 12 demographic parameters in our model. Accuracy is summarized using the ratio between the estimates and the value assigned to each parameter in the simulations. As above, the accuracy achieved by phasing integration is as good as that from analyzing correctly phased data, and often significantly better than that obtained by random phasing. Accuracy appears to be reasonable for all parameters pertaining to the deeper parts of the population tree (Figure S10B), but the parameters corresponding to the Eurasian subtree are estimated with substantially lower accuracy. The biases in estimates in this subtree (e.g., under-estimation of $\theta_K$) can partly be explained by our choice of priors (see Section S8.1), but also seem to reflect slight sampling biases. Both of these effects become apparent when there is little data supporting inference of a parameter. This result gives some indication of bounds on the times of demographic events this type of analysis has power to address. Another important observation is the high estimation accuracy of the average root divergence $\tau_{div}$, which plays an
important role in our calibration (see Section S7.2). Note that $\tau_{\text{div}}$ is not directly included in our model, but is given by $\tau_{\text{div}} = \tau_{\text{root}} + \frac{1}{2} \theta_{\text{root}}$ (see Section S7.2).

S6.4 Simulations with Gene Flow

In order to test the influence of gene flow on our parameter estimates, we simulated data under the six demographic scenarios with San divergence times of $\tau_{\text{KHEXS}} \in \{0.8, 1.0, 1.2\} \times 10^{-4}$ and African-Eurasian divergence times of $\tau_{\text{KHEX}} \in \{0.4, 0.5\} \times 10^{-4}$, allowing migration from the San population to the African ingroup population ($S \rightarrow X$). We considered migration at four levels of intensity, denoted 0, 1, 2, and 3, corresponding to total migration rates of 0.0, 0.1, 0.2, and 0.3, respectively. The total migration rate is a measure of migration scaled by the total time in which the migration band is in effect ($\tau_{\text{KHEX}}$ in our case). This measure provides an approximation for the probability a lineage in the target population ($X$) will migrate (back in time) into the source population ($S$). For each of these $6 \times 4$ demographic scenarios, two data sets were simulated, and then for each of the 48 synthetic data sets, $G$-PhoCS was executed in two modes: one not modeling gene flow, and one modeling gene flow along the $S \rightarrow X$ band. Figure S11 contains a summary of parameter estimates from the complete set of $6 \times 4 \times 2 \times 2 = 96$ runs.

When migration is included in the simulation but is not modeled in $G$-PhoCS, there is a clear trend toward under-estimation of $\tau_{\text{KHEXS}}$, which grows stronger as the migration rate increases. Other demographic parameters are also affected, particularly $\theta_X$. When migration is modeled, these biases are largely eliminated, and we do not observe any other notable estimation bias. We measure migration intensity both by the total migration rate (see above) and by the number of migrants per generation (Figure S11C). Both measures of migration are slightly over estimated in these experiments, however, we do not observe a significant migration signal in data simulated without gene flow. We conclude that $G$-PhoCS has good power to detect the presence or absence of migration and can accurately estimate divergence times in both cases, but it has somewhat reduced power to infer migration rates. See Section S5.2 for further discussion of migration, including a qualitative summary of other experiments conducted with $G$-PhoCS.

S7 $G$-PhoCS Analysis of Individual Human Genome Sequences

S7.1 Neutral Loci

The coalescent-based analysis performed by $G$-PhoCS assumes the input alignments represent a set of putative “neutral loci” in which recombination occurred at negligible rates during the sample history, and between which recombination occurred at sufficiently high rates that the genealogies are approximately uncorrelated. Such loci were computed in our case by identifying contiguous intervals of 1000 bp that passed the (hard) filters described in Section S1.2, then selecting a subset of these intervals that ensured a minimum inter-locus distance of 50,000 bp. In selecting these intervals, we also made sure that recombination hot spots (regions with recombination rates >10 cM/Mb; International Hapmap Consortium, 2007) fell between rather than within loci. A total of 37,574 autosomal loci were identified in this way, with a median inter-locus distance of 58,500 bp. Each of the resulting 37,574 loci contains a 1000 bp alignment of 6 diploid sequences and one chimp haplotype sequence, with ambiguity characters “Y”, “R”, “K”, “S”, “W”, “M” indicating het genotypes and “N” indicating masked bases (due to soft filters).

The locus size of 1 kbp and the minimum inter-locus distance of 50 kbp were determined by an approximate calculation similar to the one used by Burgess and Yang (2008). We assume a mean recombination rate of $10^{-8}$ per bp per generation, an average generation time of 25 years, a minimum average genomic divergence time of $\sim200,000$ years, and a maximum average genomic divergence time (among the humans) of $\sim500,000$ years. Thus, the expected number of recombinations on the lineages leading to two human genomes in a 1 kbp interval is at most $2 \times 500,000 \times 10^{-8} / 25 \times 1000 = 0.4$ and the expected number in a
50 kbp interval is at least $2 \times 200,000 \times 10^{-8}/25 \times 50000 = 8$. Notably, Burgess and Yang (2008) found that their analysis was not highly sensitive to small amounts of intralocus recombination. See Section S8.3 for a similar analysis of our data.

### S7.2 Parameter Calibration

As with other coalescent-based models, the demographic parameters used by *G-PhoCS* are all scaled by the mutation rate. As a result, *G-PhoCS* effectively infers ratios between model parameters, leaving their absolute scale (in time or number of individuals) undefined. To obtain absolute values, it is necessary to calibrate the model using external information, such as an estimated mutation rate or an estimated divergence time for an outgroup species. In our analysis, we use the chimp genome as an outgroup and calibrate the model based on the average genomic divergence time between human and chimpanzee, denoted $T_{\text{div}}$. Although the mutation-scaled human/chimpanzee average genomic divergence, $\tau_{\text{div}}$, is not a parameter in our demographic model, it is given by a simple function of the demographic parameters: $\tau_{\text{div}} = \tau_{\text{root}} + \frac{1}{2} \theta_{\text{root}}$ ($\frac{1}{2} \theta_{\text{root}}$ being the average time until coalescence of two lineages in the root population). We use this quantity for calibration because it is robustly estimated by *G-PhoCS*, in contrast to $\tau_{\text{root}}$ and $\theta_{\text{root}}$, whose inferred values depend strongly on our modeling assumptions, particularly those relating to mutation rate variation (see detailed experiments in Section S8.2). The average divergence $\tau_{\text{div}}$ is consistently estimated at $4.54 \times 10^{-3}$ across many different runs, with 95% credible intervals in the range $4.45–4.63 \times 10^{-3}$ (see Figure S12 and Section S8).

We assume a generous range of $T_{\text{div}} = 5.6–7.6$ Mya, as suggested by Patterson et al. (2006), 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 various other constraints from the fossil record. We follow Green et al. (2010) in choosing a “best guess” value of $T_{\text{div}} = 6.5$ Mya. Green et al. used a somewhat larger upper limit for $T_{\text{div}}$ in their analysis (8.3 Mya) but this would require assuming an orangutan average genomic divergence time of more than 20 Mya, well outside the range considered reasonable in most studies (e.g., Goodman, 1999; Chen and Li, 2001; Hobolth et al., 2007; McVicker et al., 2009; Orangutan Genome Sequencing Consortium, 2010). The assumption of a larger value for $T_{\text{div}}$ would of course increase our estimates of the San and African-Eurasian divergence times. Notice that the calibration time is by far the largest source of uncertainty in our estimated dates, greatly exceeding the statistical uncertainty in our parameter estimates.

An estimate of $\tau_{\text{div}} = 4.54 \times 10^{-3}$ and a divergence time of $T_{\text{div}} = 6.5$ Mya together imply an average mutation rate of $\mu = \tau_{\text{div}}/T_{\text{div}} \approx 0.7 \times 10^{-9}$ mutations per site per year, or $1.4 \times 10^{-8}$ mutations per site per generation, assuming an average generation time of 20 years for hominins (Chen and Li, 2001). However, this rate ignores CpG mutations, which are excluded by our filters. CpGs account for about one third of all mutations observed in our alignments, so an adjusted estimate of the per generation mutation rate would be slightly more than $2 \times 10^{-8}$ mutations per site. This adjusted estimate agrees well with independent estimates of $1.8–2.5 \times 10^{-8}$ (Nachman and Crowell, 2000; Kondrashov, 2003). It is slightly higher than recently reported estimates of $1.0–1.3 \times 10^{-8}$ (The 1000 Genomes Project Consortium, 2010; Lynch, 2010; Roach et al., 2010), but, considering the many sources of uncertainty in these studies, we do not regard this difference as a serious concern. It is difficult to reconcile per-generation mutation rate estimates as low as $1 \times 10^{-8}$ with the observed levels of human/chimpanzee genomic divergence.

Our estimate of $\mu$ implies a calibration factor of $\mu^{-1} = 1.431 \times 10^{9}$ for all population divergence times. In order to obtain effective population sizes in numbers of diploids, we must additionally assume an average generation time. We use the fairly standard assumption of an average human generation time of 25 years, resulting in a calibration factor of $(4 \times 25 \times \mu)^{-1} = 1.431 \times 10^{7}$ for all effective population sizes.
S7.3  Design of G-PhoCS Analysis of Individual Genome Sequences

Two subsets of the six human individuals were considered, both including all three non-African samples and the San sample, and each with a different African ingroup (the Yoruban or Bantu). We assumed the population phylogeny in Figure 1 (main text). (See Section S7.5 for discussion of the ambiguous relationship between the Yoruban and Bantu individuals.) For each subset, we considered several simple migration scenarios: one with no gene flow, one with a S→X migration band, and six others, each having a single migration band with an African source population (X or S) and a non-African target population (K, H, or E). The directionality of these migration bands, from “old” population to “young” population, was chosen to maximize our power for migration detection (see Section S5.2). As additional validation of this design, we considered several scenarios with multiple migration bands, but these runs did not produce clear support for any complex migration scenario. Two separate G-PhoCS runs were performed for each of these 2 × 8 = 16 scenarios, for a total of 32 runs (Table S3). The general settings used in these runs are given in Table S6. Convergence was assessed according to the guidelines described in Section S5.3. A summary of running times is given in Table S7.

S7.4  Demographic Parameter Estimates

Estimates of all demographic parameters (raw and calibrated) for all 32 G-PhoCS runs are summarized in Figure S12. Parameter estimates for the 2 × 6 scenarios with migration bands between an African population (X or S) and a non-African population (K, H or E) are similar to the ones without gene flow. Accordingly, at most a weak signal for migration was detected in these scenarios (Figure S13). A comparison across runs indicates high quality mixing for the demographic parameters pertaining to the African populations and populations ancestral to them, and lower quality sampling for parameters pertaining to the Eurasian subtree. This confirms results observed in the simulation experiments (Figure S9) as well as convergence and mixing trends observed in individual traces (Section S5.3). Importantly, the estimates for τdiv are highly consistent across all 32 runs, with narrow credible intervals, indicating a strong statistical signal for the parameter used in our calibration (see Section S7.2).

The 2 × 2 runs with S→X migration bands were the only ones to show a notable migration signal. In addition, San-Bantu migration is estimated to occur at substantially higher rates than San-Yoruban migration (Figure S13). Parameter estimates for these four runs are presented alongside estimates obtained in the four runs with no migration bands in Tables S4 and S5. The increased migration rate for the Bantu appears to explain why, when gene flow is ignored in the model, the estimate of the San divergence time is smaller with a Bantu ingroup than with a Yoruban ingroup. Accounting for gene flow in the model appears to correct for this difference, and leads to similar estimates for the San divergence in both cases. Not surprisingly, allowing for S→X gene flow also has significant effects on several of the population size parameters—such as θY, θB, θKHEX, and θKHEXS—because the migration bands influence the populations in which coalescent events occur, and the θ parameters are essentially measures of (inverse) coalescence rate. Notably, the African-Eurasian divergence time is quite consistent across runs, and does not appear to be strongly affected by migration. The Bantu population divergence τKHEB is somewhat deeper than the Yoruban divergence τKHEB, but we cannot rule out the possibility that these estimates correspond to a single event (see Section S7.5).

S7.5  Yoruban/Bantu Relationship

Estimates obtained in separate analyses of the Yoruban and Bantu samples (and allowing gene flow between these populations and the San) show high levels of concordance in model parameters common to both analyses (such as τKHEXS, θKHEXS, and θS), but do not reveal the phylogenetic relationship between these two African populations. There appear to be two plausible scenarios: (1) the two populations diverged from other
human populations in two distinct events (first the Bantu and then the Yoruban); and (2) the two populations share a common parent population (YB). We attempted to resolve this question by executing G-PhoCS runs including both the Yoruban and Bantu samples in either of the two topological configurations mentioned above (with and without the San sample), but had difficulties with convergence and were unable to obtain clear evidence favoring either scenario. We conclude that these genome sequences are not sufficiently informative to unravel the relationship between the Bantu and Yoruban using our methods. Nevertheless, this question seems to have little influence on our main parameters of interest, the San and African-Eurasian divergence times.

S7.6 Comparison with Published Estimates

In this section, we briefly summarize relevant estimates from the literature, focusing on African-Eurasian and European/East Asian divergence times and the effective sizes of ancestral populations. We note that the literature on this topic is vast and difficult to review comprehensively. We focus here on recent studies based on relatively large, multilocus autosomal data sets, involving both African and non-African samples. As a rule, these studies have not included San representatives; the few (autosomal) studies that have included the San are summarized in the main text (Zhivotovsky et al., 2003; Garrigan et al., 2007; Green et al., 2010).

- Schaffner et al. (2005) carried out one of the first attempts to fit a demographic model to genome-wide data, using published SNPs available at the time for West African, European, and East Asian populations (3738 markers from 54 autosomal regions and 250 markers from 16 X-linked regions). Their general approach was to carry out coalescent simulations under a broad range of parameter settings and select the parameters that provided the best fit to observed data in terms of various summary statistics of interest. The goal of this work was not so much to estimate the “true” demographic history of the sample as to infer a demographic model that would fit the data adequately well to serve as an improved null model in studies of variation, linkage disequilibrium, and selection. As a result, the authors did not focus much on issues of calibration, and semi-arbitrarily fixed key parameters such as the mutation rate, and the times for the out-of-Africa migration and the European/East Asian population divergence. Nevertheless, they reported estimates of 12,500 for the effective size of the ancestral W. African/Eurasian population, 24,000 for the W. African effective population size, and 7,700 for the Eurasian effective population size, as well as estimates of parameters describing bottlenecks, population expansions, and recombination hotspots. They found evidence for small amounts of African-European migration. Evidence for African-East Asian migration was much weaker.

- Voight et al. (2005) more directly estimated demographic parameters, based on complete sequencing of 50 unlinked autosomal noncoding regions in samples of Hausa from Cameroon, Italians, and Chinese. They used coalescent simulations and summary likelihood methods to fit various demographic models to their data, analyzing the three populations separately. They estimated an ancestral population size of about 10,700 (9,450–12,300) under a growth model, and found strong support for bottlenecks in the Eurasian samples, but were not able to date these bottlenecks with any confidence. They did not estimate population divergence times.

- Liu et al. (2006) fitted a dynamic genetic model with an explicit representation of geographic distances to autosomal microsatellite data for 52 human populations (783 markers). They estimated an expansion of modern humans 56 kya from a small founding population with an effective size of only ~1,000. Their model predicted an out-of-Africa migration event 45–55 kya.

- Fagundes et al. (2007) used an approximate Bayesian computation (ABC) approach to analyze 50 autosomal noncoding loci of ~500 bp each, representing African, Asian, and Native American samples. They compared various explicit multi-population demographic models and found that a simple
African replacement model with exponential growth fit the data significantly better than alternative multiregional evolution of assimilation scenarios. They reported estimates of 51 kya (95% CI 40–71 kya) for the out-of-Africa migration and 12,772 (6,604–20,211) for the ancestral African effective population size, among others.

- Keinan et al. (2007) used a subset of HapMap SNPs, carefully selected to avoid ascertainment biases, to study the allele frequency spectra of Europeans and East Asians. Using a likelihood approach based on allele frequencies, they estimated recent bottlenecks 18 ± 3 kya in Europeans and 16 ± 2 kya in East Asians, and a European/East Asian population divergence time of 17 ± 3 kya.

- Gutenkunst et al. (2009) introduced a flexible likelihood method for fitting demographic models based on the joint allele frequency spectrum for multiple populations applied it to 5 Mb of noncoding DNA resequenced in 68 individuals from Yoruban, Han Chinese, European, and Mexican-American populations. They reported estimates of 7,300 (6,300–9,200) for the ancestral effective population size (expanding to 12,300 ~220 kya), 140 (60–310) kya for the W. African/Eurasian divergence time, and 21.2 (17.6–23.9) kya for the European/East Asian divergence time. They speculated that their unusually ancient estimate of the W. African/Eurasian divergence time could have resulted from their inclusion of migration in their model, although other methods that have considered migration have also estimated much more recent dates. Their estimate of the European/East Asian divergence time agreed fairly well with the one reported by Keinan et al. (2007), based on similar methods but quite different data.

- Wall et al. (2009) performed a summary likelihood analysis of essentially the same data set considered by Gutenkunst et al. (2009), but with explicit consideration of possible admixture between modern and archaic humans (Neandertals, Homo erectus, and Homo floresiensis). This was an extension of a previous study by Plagnol and Wall (2006). They reported an African/Eurasian divergence time of 80 (60–93) kya based on their East Asian samples and of 120 (86–148) kya based on their European samples (which they analyzed separately). They simply fixed the ancestral population size at 10,000 for this analysis. They reported evidence for ancient admixture in all three populations.

- Laval et al. (2010) carried out an ABC analysis of 20 noncoding autosomal regions from 213 individuals. They estimated an ancestral African population size of ~13,800 an out-of-Africa migration date of 60.0 (47.5–85.0) kya, and a European/East Asian divergence time of 22.5 (17.5–35.0) kya.

With a few exceptions (such as the rather deep Gutenkunst et al. and Wall et al. estimates of the African-Eurasian divergence time) these studies are reasonably concordant with ours, especially given the pronounced differences among them in methods and data sets. The main discrepancy seems to be in the estimate of the European/East Asian divergence time, which others have consistently estimated to be considerably more recent (~20 kya) than our estimate of 30–45 kya. Interestingly, our more ancient estimate is easier to reconcile with fossil evidence for anatomically modern humans in Europe ~40 kya. Nevertheless, our methods are not ideal for dating such a recent event, because the signal from mutation on this time scale is quite weak, and genotyping error and failures of MCMC convergence could produce biases in our estimates. Additional work will be required to explain the discordance between our estimate of the European/East Asian divergence time and those based on allele frequency data.

### S8 Validation of G-PhoCS Analysis

This section presents a series of analyses designed to examine the robustness of our results to key modeling assumptions. We address the issues of influence of the prior, mutation rate variation across loci, intralo-
cus recombination, non-constant population sizes, and parameters/thresholds defining our data-quality and comparative filters.

S8.1 Prior Distributions

We performed a series of tests, using both simulated and real data, to assess the effect of the choice of prior distributions on our parameter estimates. First, we applied G-PhoCS to the 18 simulated data sets generated without gene flow, using four different prior distributions for the demographic parameters. The first distribution (prior A) is the one used in our original analysis (see Section S6.3 and Table S6). The second and third distributions (priors B and C) are different from the original in that all five divergence times are distributed $\Gamma(\alpha = 1, \beta = 10000)$ and $\Gamma(\alpha = 1, \beta = 100000)$, respectively. The fourth prior (prior D) is different from the first in that all $\theta$ parameters are distributed $\Gamma(\alpha = 1, \beta = 1000)$. Note that all of these priors are fairly noninformative (with large 95% credible intervals relative to their means) but their means (given by $\alpha/\beta$) differ substantially. For estimates based on prior A, we used the original 18 runs reported in Section S6.3 and Figure S10, and for each of the other three priors we executed an additional set of 18 G-PhoCS runs. The results of these simulations are summarized in Figure H.

Figure H: Accuracy of G-PhoCS estimates of demographic parameters for the 18 simulated data sets without gene flow, based on four different prior distributions for demographic parameters (priors A, B, C, and D). (A) Estimates of the two population divergence times of primary interest, $\tau_{KHEX}$ and $\tau_{KHEXS}$, with each boxplot summarizing 6 G-PhoCS runs. (B) Estimates of demographic parameters for the African portion of the phylogeny. (C) Estimates of demographic parameters for the Eurasian portion of the phylogeny. In (B) and (C), all parameter estimates are normalized by the true value assigned to them in the simulations, and each boxplot summarizes 18 runs.
Based on these examples, the choice of prior distribution appears to have little effect on the estimates of the two divergence times of primary interest ($\tau_{\text{KHEX}}$ and $\tau_{\text{KHEXS}}$). Moreover, the choice of prior for divergence times (priors B and C) has only a minor effect on all parameters. Changing the prior for effective population size does alter estimates of current effective population sizes (e.g., $\theta_X$, $\theta_K$, and $\theta_H$) and recent population divergence times (e.g., $\tau_K$). The observed differences are consistent with the reduced signal in the data for recent demographic events and the increased effective population size implied by prior D.

In addition to our experiments with simulated data, we applied G-PhoCS to our real data set (all 37,574 neutral loci) using several alternative prior distributions. We considered both versions of our data set (with either the Yoruban or the Bantu sample as population X), allowing for S→X migration. We tried priors A and C (above), as well as an additional prior (prior E) in which all $\theta$ and $\tau$ parameters are assigned a prior distribution of $\Gamma(\alpha = 0.01, \beta = 100)$. This distribution has the same mean as the one used in prior B, but higher variance (by a factor of $10^4$). The estimates obtained in these four runs are compared to the ones obtained in the two original runs (runs 16 and 32; see Table S3) in Figure I. Consistent with our simulation results, the prior does not appear to have a significant effect on the estimates of any of the main demographic parameters of interest, in particular the two main population divergence times, $\tau_{\text{KHEX}}$ and $\tau_{\text{KHEXS}}$.

We conclude that the choice of prior distribution has a negligible effect on the main demographic parameters of interest in our analysis. There is a relatively minor effect on some of the other parameters, predominantly the ones pertaining to the more recent portion of the population history.

Figure I: G-PhoCS estimates of demographic parameters based on our real data set (37,574 neutral loci), with three different priors. Two versions of the data set were considered, with either the Yoruban or the Bantu sample as population X, and an S→X migration band in both cases. Estimates are shown for selected demographic parameters. Each bar describes the mean and 95% CI observed for a specific parameter in a certain run (priors A, C, and E). The $y$-axes show both the raw ($\times 10^4$; left) and calibrated (right) versions, with $T_{\text{div}} = 6.5$ Mya. The calibrated effective population sizes are given in thousands and assume an average generation time of 25 years. The calibrated population divergence times are given in kya. The estimates from the original runs determine the baseline values (red/dark blue labels, solid horizontal line), and the dashed lines indicate ±10% from the baseline.

S8.2 Variation in Locus-specific Mutation Rates

In coalescent-based models, mutation provides the link between time and the observed patterns of variation in the sequences. If the mutation rate is constant across loci, it simply acts as a global scaling factor,
which can be estimated using an appropriate external calibration (see Section S7.2). However, variable mutation rates across loci provide an additional modeling challenge. Sequence data from closely related samples (such as human individuals) contains only weak information about the mutation rate. Introducing an outgroup sequence (chimpanzee in our case) is a first step in dealing with this issue. Each locus-specific mutation rate can be estimated based on the divergence of the outgroup. However, it can be difficult to differentiate between the variance in outgroup divergence caused by variation in mutation rates and that caused by variation in coalescent times in the population at the root of the phylogeny. This is especially challenging when using chimpanzee as an outgroup, because the human/chimpanzee ancestral population appears to have a fairly large effective size relative to the population divergence time between humans and chimpanzees, so that ancestral coalescence makes a large contribution to the variance in genomic divergence.

Mutation rate variation across loci can be modeled in \textit{G-PhoCS} in three different ways:

- **Constant rate across all loci**: All loci are assumed to have the same mutation rate.

- **Fixed locus-specific rates**: The ratios between mutation rates in the different loci are computed externally and are provided to \textit{G-PhoCS} by the user. \textit{G-PhoCS} uses these ratios as fixed locus-specific rates (normalized to have an average of 1).

- **Random locus-specific rates**: \textit{G-PhoCS} samples locus-specific rates as additional latent variables in the probabilistic model. These rates are normalized to have an average of 1. A Dirichlet prior distribution is used with a single user-defined $\alpha$ for all loci.

Each of these three models has certain limitations. The “constant rate” model is clearly limited in a genome-wide setting, because mutation rates are known to vary across the genome. The “fixed rates” model is dependent on the data and methodology used to estimate locus-specific rates. The “random rates” model makes only weak use of the outgroup, and can over-model variation in mutation rate at the expense of variation in ancestral coalescent times (see below). It is thus important to verify that the model we used in our analysis (random rates) does not significantly affect our findings. We performed two types of validation: simulation experiments to gauge the influence of the choice of rate variation model, and reanalysis of real data to observe differences in parameter estimates. Both analyses indicate that mutation rate variation has only a minor influence on our estimates.

**Simulation experiments**

We designed experiments with simulated data to study the possible effects the “random rates” model might have on estimates we obtained in our initial analysis presented in Section S7.4. We took the approach of simulating data with constant rates and performing inference with various models of rate variation, in order to see what types of model-induced biases would arise. We simulated four data sets based on the following demographic parameter values:

- **San divergence time**: $\tau_{\text{KHEXS}} \in \{0.8, 0.9, 1.0, 1.1\} \times 10^{-4}$.

- **African-Eurasian divergence time**: $\tau_{\text{KHEX}} = 0.4 \times 10^{-4}$.

- **Root divergence and effective population size**: $\tau_{\text{root}} = \theta_{\text{root}} = 30 \times 10^{-4}$.

\textit{G-PhoCS} was executed on each of these data sets in two modes: assuming a constant rate across loci, and assuming random rates across loci (with a Dirichlet($\alpha = 1$) prior). The results (Figure J) show that the “random rates” model tends to over-estimate variation in mutation rates across loci at the expense of variation in coalescence times in the root population, as indicated by the under-estimated values of $\theta_{\text{root}}$. This phenomenon is not driven by the prior distribution of $\theta_{\text{root}}$, as similar results were obtained with different
priors. However, despite the bias in $\theta_{\text{root}}$, the average divergence at the root, $\tau_{\text{div}}$, continues to be estimated with high accuracy, suggesting that our calibration is robust to the choice of model for mutation rate variation. The model maintains an accurate total divergence by compensating for under-estimated coalescence times with over-estimated population divergence times.

Interestingly, the bias in estimates of $\theta$ parameters due to mismodeling of rate variation cascades down the population tree. The over-estimation of $\theta_{\text{KHEXS}}$ leads to an under-estimation of $\theta_{\text{KHEX}}$ and $\theta_{\text{S}}$, and then, in some cases, to a slight over-estimation of $\theta_{\text{Y}}$. The biases exhibit a kind of decaying “zig-zag” pattern, as alternating excesses and deficiencies of variance propagate from one level to the next. Our main parameter of interest, the San divergence time ($\tau_{\text{KHEXS}}$), appears to be slightly under-estimated (by $\sim$5%) by this over-modeling of rate variation, evidently to compensate for the over-estimation of the corresponding population size parameter, $\theta_{\text{KHEXS}}$. Thus, our use of the “random rates” model could potentially have led to a slight under-estimation of the San divergence time, particularly if the true amount of rate variation is small and/or the true human/chimpanzee ancestral population size is large, but our model for rate variation is unlikely to have inflated the San divergence time.

Figure J: Estimates from simulated data under different models for rate variation across loci. Four simulated data sets are considered. For each data set $G$-PhoCS was run in two modes: constant mutation rate across all loci, and randomly sampled rates. Each bar describes the mean and 95% CI observed for a specific parameter in a certain run. All estimates are scaled by a factor of $10^4$. The eight runs are shown as four sets of (“constant”, “random”) pairs, in increasing order of $\tau_{\text{KHEXS}}$. The values assigned to the parameters in the simulations determine the baseline values (red labels, solid horizontal lines), and dashed lines indicate $\pm$10% from the baseline.

Validation on genomic data
To further examine the effects of rate variation on our estimates, we performed several additional $G$-PhoCS runs on our genomic neutral loci, using all three models of rate variation across loci. For the “fixed rates” analysis, a mutation rate scale was computed for each of the 37,574 neutral loci in our analysis by maximum likelihood, based on alignments of human, rhesus macaque, and orangutan. We extracted these alignments from the 44-way genome-wide alignments available from the UCSC Genome Browser (reference assembly hg18). Regions with poorly conserved synteny between the human sequence and any of the other two sequences were identified using the UCSC syntenic net tracks, and removed from the analysis. A scaling factor was computed for each locus using the phyloFit function in RPHAST (Hubisz et al., 2011), under a neutral model estimated from fourfold degenerate sites in coding regions (as used for the 44-way Conservation track in the UCSC Genome Browser, assembly hg18; see Pollard et al., 2010). To ensure that these
estimates were reliable, 6,931 loci for which more than 100 bp (out of 1000) were eliminated due to lack of synteny were not considered in this analysis.

We ran G-PhoCS using each of the three models for mutation rate variation across loci on the set of 30,643 loci that passed all of our filters. This analysis was done with five human individuals, including the Yoruban sample and excluding the Bantu sample. An S→Y migration band was allowed. Figure K presents estimates obtained in these three runs alongside the estimates from run #16 of the original analysis (based on the “random rates” model and all 37,574 loci; see Table S3). Estimates obtained under the “fixed rates” model appear to be very similar to ones obtained under “constant rates.” This is not surprising because the fixed rates we computed had quite low variance (0.0435, when normalized by the mean). Under these two models, a relatively large effective size is estimated for the root population (θroot), and the root population divergence drops to ~4.5 Mya (assuming Tdiv = 6.5 Mya). Estimates of the San divergence time (τKHEXS) differ by <1%, with slightly higher estimates obtained under “fixed rates” and “constant rate”, as expected from our simulation study. As described above, a compensating effect is seen in the effective size of the MRCA population (θKHEXS). Estimates of demographic parameters further down the population phylogeny were fairly consistent with ones obtained in our original analysis.

We conclude that mutation rate variation across loci mainly influences our ability to accurately disentangle root population divergence from ancestral coalescence, with only a minor effect on more recent portions of the phylogeny. By using chimpanzee as an outgroup population, we appear to be able to obtain fairly robust estimates of demographic parameters in human population history, regardless of the assumed model for mutation rate variation.
S8.3 Intralocus Recombination

*G-PhoCS* assumes that each locus is described by a single (unknown) genealogy, implying no intralocus recombination throughout the history of the sample (or at least no “visible” recombination events, leading to changes in the genealogy along the sequence). In practice, of course, intralocus recombination cannot be completely eliminated; the best one can do is attempt to reduce it to negligible levels (see Section S7.1). To assess the potential effects of intralocus recombination on our parameter estimates, we conducted experiments on both simulated and real data, as follows.

**Simulation experiments**

We simulated several data sets with various levels of intralocus recombination, to observe the effect recombination might have on our parameter estimates. We used a demographic scenario similar to the one described in Section S6.3 (with $\tau_{KHEX} = 0.4 \times 10^{-4}$ and $\tau_{KHEXS} = 1.0 \times 10^{-4}$), no migration, and two recombination rates: 0.7 and 1.4 expected recombinations per site per unit time (where time is measured in expected mutations per site). These rates correspond to rates of $\{1, 2\} \times 10^{-8}$ recombinations per site per generation, assuming a mutation rate of $1.4 \times 10^{-8}$ substitutions per site per generation, in agreement with estimates by Myers et al. (2005). With additional filtering, described later this section under “additional recombination filters”, we make sure our loci conform with these average rates. Six data sets were generated under each of the two recombination rates, with various locus lengths: 500 bp, 1 kbp, 2 kbp, 5 kbp, 10 kbp, and 20 kbp. To reduce differences in information content between the different sets, we generated 20,000 loci for the three smallest lengths, and 2,000 loci for the three largest lengths. *G-PhoCS* was run on these 12 data sets using our standard settings (Section S6 and Table S6). The results of these runs are presented in Figure L. These results indicate that intralocus recombination does produce a clear bias in many parameter estimates, but this bias appears to be small for locus lengths of $<2$ kbp. Estimates of $\tau_{KHEXS}$ are elevated by $<5\%$ when moving from a locus length of 500 bp to our chosen length of 1 kbp.

![Figure L: Estimates based on data simulated with intralocus recombination. Each bar describes the mean and 95% CI for a specific parameter in a particular run. All estimates are scaled by a factor of $10^4$. The twelve runs are shown in two sets, corresponding to recombination rates of 0.7 (left) and 1.4 (right), and in increasing order of locus length within each set. The values assigned to the parameters in the simulations determine the baseline values (red labels, solid horizontal lines), and dashed lines indicate $\pm 10\%$ from the baseline.](image)
Validation with genomic data – perturbing locus length

Following the approach taken by Burgess and Yang (2008), we repeated our analysis with a series of alternative neutral loci, to see whether our estimates would change as loci were defined such that the effects of intralocus recombination should be increased or diminished. We defined loci similar to our original ones (Section S7.1) but with lengths of 500 bp, 1 kbp, 2 kbp, 5 kbp, 10 kbp, and 20 kbp. In order to have enough data to define longer loci, we had to relax our filters somewhat; in particular, contiguous genomic segments of length $\leq 500$ excluded by “hard” filters (see Section S1.2) were converted to soft filters. This strategy allows for larger numbers of loci, but these loci contain more missing data on average, due to masking by the new soft filters. For consistency, we recomputed the set of 1 kbp loci rather than reusing the previous ones. The number of loci obtained in each set were: 46,559 (1 kbp), 43,483 (2 kbp), 30,496 (5 kbp), 10,101 (10 kbp), and 501 (20 kbp). The 500 bp loci were defined by simply truncating the 46,559 1 kbp loci.

We ran G-PhoCS on each of these six sets of loci, using the Yoruban sample as the African ingroup $X$ and including a migration band $S\rightarrow X$. The resulting estimates are shown in Figure M. Most demographic parameters of interest, including $\tau_{\text{KHEX}}$, appear to be quite robust to choices of locus length between 500 bp and 2 kbp. The African-Eurasian divergence time $\tau_{\text{KHEX}}$ is somewhat more sensitive in this region, changing by $\sim 15\%$ when the locus length increases from 500 bp to 1 kbp. With loci of length $\geq 5$kb, we see a clear effect of intralocus recombination on all estimates, similar to that observed in our simulation experiments.

We conclude that our initial choice of 1 kb loci appears reasonable, particularly with respect to our estimates of the San divergence time, but our parameter estimates may reflect some minor biases from intralocus recombination.

![Figure M: Estimates for selected parameters with different choices of locus length (left to right): 500 bp, 1 kbp, 2 kbp, 5 kbp, 10 kbp, and 20 kbp. Each bar describes the mean and 95% CI for a specific parameter in a certain run. The y-axes show both the raw (×$10^{-4}$; left) and calibrated (right) versions, with $T_{\text{div}} = 6.5$ Mya. The calibrated effective population sizes are given in thousands and assume an average generation time of 25 years. The calibrated population divergence times are given in kya. Note that each run requires a separate calibration, due to slightly different estimates of $\tau_{\text{div}}$. The estimates from the run on 1 kb loci determine the baseline values (red/dark blue labels, solid horizontal line), and the dashed lines indicate $\pm 10\%$ from the baseline.](image-url)
Validation with genomic data – filtering loci suspected of high recombination rates

Finally, we repeated our analysis on a data set that excluded loci likely to be enriched for intralocus recombination, as follows. First, we estimated average recombination rates at all loci based on a recently published fine-scale recombination map (The 1000 Genomes Project Consortium, 2010). We simply averaged across the sampled populations (European, East-Asian and Yoruban) for this analysis. We produced three data sets by eliminating loci that exceeded three different thresholds for the average recombination rate (1.5, 2.0, and 3.0 cM/Mb). We also identified 597 loci (out of the original 37,574) that violate the four-gamete test under all possible phasings of heterozygotes. Assuming an infinite sites model for human populations (we do allow recurrent mutations on the long branch to the chimpanzee) and no genotyping error, these alignments are incompatible with a single genealogy and therefore must have experienced intralocus recombination. Of course, with real data, violations of the four-gamete test may also arise from homoplasy events or genotyping errors. For each of our three recombination thresholds, we generated a data set that additionally excluded these 597 loci. The strictest filtering scheme, using recombination threshold of 1.5 cM/Mb and removing loci which violate the four-gamete test, removed 5,034 loci (∼13%). As a control, we used a seventh data set that consisted of a random subset of 30,000 loci.

We ran G-PhoCS on each of these seven sets of loci, using the Yoruban sample as the African ingroup.

Figure N: Estimates for selected parameters with different filtering thresholds of pre-estimated recombination rates. The seven data sets considered are (left to right): a control set of randomly chosen 30,000 loci, three sets corresponding to recombination rate thresholds of 1.5, 2.0, and 3.0 cM/Mb, and the same three sets without loci violating the four-gamete test. Each bar describes the mean and 95% CI for a specific parameter in a certain run. The y-axes show both the raw ($\times 10^{-4}$; left) and calibrated (right) versions, with $T_{\text{div}} = 6.5$ Mya. The calibrated effective population sizes are given in thousands and assume an average generation time of 25 years. The calibrated population divergence times are given in kya. Note that each run requires a separate calibration, due to slightly different estimates of $T_{\text{div}}$. The estimates from the run on the control set determine the baseline values (red/dark blue labels, solid horizontal line), and the dashed lines indicate ±10% from the baseline.
X and including an $S\rightarrow X$ migration band. We used NA12891 as a European sample in this study (see also Section S8.5). The resulting estimates are shown in Figure N. The estimates obtained for all seven sets are generally highly concordant with our original estimates. The estimate of the San divergence time $\tau_{KHEXS}$ appears to be stable across data sets. Some effective population sizes ($\theta_{KHEXS}$ and $\theta_S$) show a small ($<5\%$) but clear reduction in their estimates as more loci showing evidence of possible recombination are filtered out of the analysis, however, overall we see no evidence of a strong bias from intralocus recombination.

### S8.4 Complex Changes in Effective Population Size

Our demographic model assumes that effective population sizes are constant along branches of the population phylogeny, and can change only when ancestral populations diverge. While this model can accommodate changes in population size to a degree, it has limited ability to capture phenomena such as bottlenecks or population expansions on individual branches. With small numbers of samples in the analysis, the specific demographic histories of individual lineages (particularly those near to the present day) are unlikely to impact inferences near the root of the population phylogeny. Nevertheless, human populations are believed to have experienced severe bottlenecks and dramatic growth (Schaffner et al., 2005; Keinan et al., 2007; Wall et al., 2009), so it is important to verify that this simplified model does not lead to biased estimates in the presence of complex (but plausible) demographic scenarios.

We executed a series of simulation tests to test the effect of complex demographic scenarios which might have taken place in the time since the San divergence on three of the central demographic parameters in this study: $\tau_{KHEX}$, $\tau_{KHEXS}$, and $\theta_{KHEX}$. We used results reported by Wall et al. (2009), Schaffner et al. (2005) and Keinan et al. (2007) as guidelines in designing these scenarios. In particular, we generated simulated data sets with the following features ($MS$ commands detailed below):

- **Bottlenecks.** An East Asian bottleneck in the KH population, a European bottleneck, an out-of-Africa bottleneck in the KHE population, and an ancient African bottleneck in the KHEX population.

- **Gene flow.** Gene flow between population X (designating the Yoruban population) and the three non-African populations. Scenarios 1 and 3 assume low rate of migration and scenarios 2 and 4 assume a high rate of migration.

- **Recent population expansion.** Two different scenarios of recent population expansion: instant expansion of populations E,K,H, and X, and exponential expansion of population X. Scenarios 1 and 2 assume instant population expansion, and scenarios 3 and 4 assume exponential population expansion.

#### Scenario 1:

```
MS 11 20000 -T -I 6 2 2 2 2 2 1
 -n 1 0.0077 -n 2 0.0077 -n 3 0.0077 -n 4 0.0077 -n 5 0.0015 -n 6 0.0010
 -m 1 4 2000 -m 2 4 1000 -m 3 4 1000
 -en 0.000004 4 0.00154 -en 0.000007 1 0.00077 -en 0.000008 2 0.00077
 -en 0.000008 3 0.00077 -em 0.000001 2 4 1000 -em 0.000001 3 4 0.0
 -ej 0.00001 3 2 -en 0.000001 2 0.00077 -en 0.0000012 2 0.0000230
 -en 0.00001277 2 0.00077 -en 0.0000014 1 0.000077 -en 0.00001477 1 0.00077
 -em 0.000002 1 4 0.0 -em 0.000002 2 4 0.0 -ej 0.000002 2 1
 -en 0.00002 1 0.00077 -en 0.000030 1 0.0000181 -en 0.000003077 1 0.00077
 -ej 0.00004 4 1 -en 0.00004 1 0.000077 -en 0.0000080 1 0.0000057
 -en 0.000008077 1 0.00077 -ej 0.000010 5 1 -en 0.000010 1 0.00077
 -ej 0.00410 6 1 -en 0.00410 1 0.0010
```

#### Scenario 2:
Figure O: Estimates obtained under complex demographic scenarios. Four simulated data sets are considered with different complex demographic scenarios, including bottlenecks, population expansions, and migration between Africans and non-Africans. Each bar describes the mean and 95% CI observed for a specific parameter in a certain run. The values assigned to the parameters in the simulations determine the baseline values (red labels, solid horizontal lines), and dashed lines indicate ±10% from the baseline.

The results of the *G-PhoCS* runs on these four data sets are summarized in Figure O. These runs indicate that *G-PhoCS* estimates both the San divergence time (τ_{KHEXS}) and the effective size of the MRCA population (θ_{KHEXS}) fairly accurately, despite the fact that the model can only very crudely describe these complex
changes in effective population size. The estimate of the African-Eurasian divergence time is biased upward by \(\sim 25\%\), mostly due to the ancient African bottleneck in the KHEX population.

**S8.5 Filter Settings**

Another factor that could have some influence on our parameter estimates is the specific definitions used for our data-quality and comparative filters (Section S1.2). These filters were designed based on our prior knowledge and related studies by other groups, but they inevitably depend on certain thresholds, many of which are difficult to set in a rigorous manner. Therefore, we selected four filters of particular importance, and examined the sensitivity of our parameter estimates to changes in the thresholds used to define these filters. These filters were: “clustered SNPs”, “recent transposable elements”, “coding regions”, and “effective coverage”. We used the thresholds described in Section S1.2 as a baseline, and perturbed each one separately, making use of the baseline value and two alternative values. This led to a total of eight alternative data sets in addition to the baseline. In order to test for combined effects of filter perturbation, we also considered a strict filter setting using the highest of the three thresholds defined for each of the four filters. We ran \(G\text{-PhoCS}\) on all ten data sets, using the Yoruban sample as the African ingroup population \(X\) and modeling migration along the migration band \(S \rightarrow X\). We used NA12891 as a European sample in this study, since its higher coverage allowed us to test the effect of the coverage filter.

Estimates of the main demographic parameters of interest in the baseline set show very high concordance with our original estimates, despite the use of NA12891 in place of the Venter genome. This suggests that reduced power for heterozygotes due to reduced coverage in the Venter genome (Section S3.1) does not have an appreciable effect on our main parameters of interest. Estimates obtained from the nine alternative data sets are compared to the baseline estimates in Figure P. The observed fluctuations are typically quite small, implying robustness of the demographic inference to the choice of filter thresholds. The threshold for the coverage filter (second and third bars from the right in Figure P) has a small, but noticeable, effect on the San divergence time \(\tau_{KHEXS}\). However, increasing the stringency of the coverage filter causes a large fraction of the data to be discarded, especially for the San genome which has average coverage of \(\sim 23\) (Table 1, main text), and the observed effects may in part reflect reduced signal. Another observation is that increasing the distance from coding regions in our filter for flanking sites \((1 \rightarrow 5 \rightarrow 10 \text{ kb})\) has little effect on our estimates. This increase causes a slight but steady increase in the estimate of \(\tau_{\text{div}}\), indicating an increase in overall mutation rate (which is expected since the effects of background selection are reduced). However, once the remaining parameters are calibrated according to \(\tau_{\text{div}}\), we see no significant changes in the estimates of population divergence times or effective sizes. The other two filters (clustered SNPs and recent transposable elements) appear to have minor effects on some of effective population size estimates, but all within the range of uncertainty of these parameters. Overall, we conclude that our analysis is robust to the setting of the various filter thresholds.
Figure P: Estimates for selected parameters with alternative definitions of data-quality and comparative filters. Each bar describes the mean and 95% CI observed for a specific parameter in a particular run of G-PhoCS. Different runs correspond to different data sets obtained using alternative filtering thresholds. The y-axes show both the raw ($\times 10^{-4}$; left) and calibrated (right) parameter estimates, with $T_{\text{div}} = 6.5$ Mya. The calibrated effective population sizes are given in thousands and assume an average generation time of 25 years. The calibrated population divergence times are given in kya. Note that each run requires a separate calibration, due to slightly different estimates of $\tau_{\text{div}}$. The baseline data set (leftmost estimate) was obtained using the thresholds described in Section S1.2. Each of the 8 alternative data sets was obtained by perturbing one threshold of interest. From left to right: clustered SNPs within a segment of 20 and 30 bp, recent transposable elements track score threshold of 27 and $\infty$, distance from coding regions of 5 kbp and 10 kbp, and effective coverage threshold of 10 and 15. The rightmost estimate corresponds to a combination of the strictest settings of all four filters. The baseline estimates are marked (red/dark blue labels, solid horizontal line), and the dashed lines indicate $\pm 10\%$ from the baseline.
References

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E., 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, 11:1005–1017.

Beerli, P. and Felsenstein, J., 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, 152:763–773.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, 27(2):573–580.

Burgess, R. and Yang, Z., 2008. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.*, 25:1979–1994.

Caswell, J. L., Mallick, S., Richter, D. J., Neubauer, J., Schirmer, C., Gnerre, S., and Reich, D., 2008. Analysis of chimpanzee history based on genome sequence alignments. *PLoS Genet.*, 4:e1000057.

Chen, F.-C. and Li, W.-H., 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet*, 68(2):444–456.

Fagundes, N. J., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., and Excoffier, L., 2007. Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 104:17614–17619.

Felsenstein, J., 1973. Maximum-likelihood and minimum-step methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.*, 22:240–249.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17:368–376.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., *et al.*, 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861.

Garrigan, D., Kingan, S. B., Pilkington, M. M., Wilder, J. A., Cox, M. P., Soodyall, H., Strassmann, B., Destro-Bisol, G., de Knijff, P., Novelletto, A., *et al.*, 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. *Genetics*, 177:2195–2207.

Goodman, M., 1999. The genomic record of Humankind’s evolutionary roots. *Am. J. Hum. Genet.*, 64:31–39.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., *et al.*, 2010. A draft sequence of the Neandertal genome. *Science*, 328:710–722.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.*, 5:e1000695.

Hastings, W. K., 1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.
Hoberman, R., Dias, J., Ge, B., Harmsen, E., Mayhew, M., Verlaan, D. J., Kwan, T., Dewar, K., Blanchette, M., and Pastinen, T., et al., 2009. A probabilistic approach for SNP discovery in high-throughput human resequencing data. *Genome Res.*, 19:1542–1552.

Hobolth, A., Christensen, O. F., Mailund, T., and Schierup, M. H., 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.*, 3:e7.

Huang, X. and Madan, A., 1999. CAP3: A DNA sequence assembly program. *Genome Res.*, 9:868–877.

Hubisz, M. J., Pollard, K. S., and Siepel, A., 2011. PHAST and RPHAST: Phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, 12:41–51.

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

International Hapmap Consortium, 2007. A second generation human haplotype map of over 3.1 million snps. *Nature*, 339(7164):851–861.

Jukes, T. H. and Cantor, C. R., 1969. Evolution of protein molecules. In Munro, H., editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York.

Keinan, A., Mullikin, J. C., Patterson, N., and Reich, D., 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.*, 39:1251–1255.

Kondrashov, A. S., 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.*, 21:12–27.

Laval, G., Patin, E., Barreiro, L. B., and Quintana-Murci, L., 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS ONE*, 5:e10284.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., et al., 2007. The diploid genome sequence of an individual human. *PLoS Biol.*, 5:e254.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R., 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079.

Li, H., Ruan, J., and Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18:1851–1858.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J., 2009b. SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, 19:1124–1132.

Liu, H., Prugnolle, F., Manica, A., and Balloux, F., 2006. A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.*, 79:230–237.

Lynch, M., 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U.S.A.*, 107:961–968.

Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitzel, N. O., Hillier, L., Kwok, P. Y., and Gish, W. R., et al., 1999. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, 23:452–456.
McVicker, G., Gordon, D., Davis, C., and Green, P., 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.*, 5:e1000471.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A., and Teller, E., 1953. Equation of state calculations by fast computing machines. *J Chem Phys*, 21:1087–1092.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P., 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310:321–324.

Nachman, M. W. and Crowell, S. L., 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156:297–304.

Nei, M. and Takezaki, N., 1996. The root of the phylogenetic tree of human populations. *Mol. Biol. Evol.*, 13:170–177.

Nielsen, R. and Wakeley, J., 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, 158:885–896.

Orangutan Genome Sequencing Consortium, 2010. Comparative and demographic analysis of orangutan genomes. *Nature*, In press.

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D., 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441:1103–1108.

Plagnol, V. and Wall, J. D., 2006. Possible ancestral structure in human populations. *PLoS Genet.*, 2:e105.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A., 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20:110–121.

Rambaut, A. and Drummond, A. J., 2007. Tracer v1.5. Available from http://beast.bio.ed.ac.uk/Tracer.

Rambaut, A. and Grassly, N. C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13:235–238.

Rannala, B. and Yang, Z., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164:1645–1656.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., et al., 2001. Linkage disequilibrium in the human genome. *Nature*, 411:199–204.

Relethford, J. H. and Jorde, L. B., 1999. Genetic evidence for larger African population size during recent human evolution. *Am. J. Phys. Anthropol.*, 108:251–260.

Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., et al., 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328:636–639.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mulikin, J. C., Mortimore, B. J., Willey, D. L., et al., 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:928–933.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D., 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, 15:1576–1583.
Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., et al., 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature*, **463**:943–947.

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., Liu, Y., Weinstock, G. M., Wheeler, D. A., Gibbs, R. A., et al., 2010. A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res.*, **20**:273–280.

Storz, J. F., Payseur, B. A., and Nachman, M. W., 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Mol. Biol. Evol.*, **21**:1800–1811.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, **467**:1061–1073.

Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y., Hudson, R. R., and Di Rienzo, A., 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**:18508–18513.

Wall, J. D., Lohmueller, K. E., and Plagnol, V., 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.*, **26**:1823–1827.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., et al., 2008. The diploid genome sequence of an Asian individual. *Nature*, **456**:60–65.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., et al., 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**:872–876.

Yang, Z. and Rannala, B., 2010. Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. U.S.A.*, **107**:9264–9269.

Zhivotovsky, L. A., Rosenberg, N. A., and Feldman, M. W., 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.*, **72**:1171–1186.