nature genetics

Robust and scalable inference of population history from hundreds of unphased whole genomes

Jonathan Terhorst[1], John A Kamm[1,2] & Yun S Song[1–4]

It has recently been demonstrated that inference methods based on genealogical processes with recombination can uncover past population history in unprecedented detail. However, these methods scale poorly with sample size, limiting resolution in the recent past, and they require phased genomes, which contain switch errors that can catastrophically distort the inferred history. Here we present SMC++, a new statistical tool capable of analyzing orders of magnitude more samples than existing methods while requiring only unphased genomes (its results are independent of phasing). SMC++ can jointly infer population size histories and split times in diverged populations, and it employs a novel spline regularization scheme that greatly reduces estimation error. We apply SMC++ to analyze sequence data from over a thousand human genomes in Africa and Eurasia, hundreds of genomes from a *Drosophila melanogaster* population in Africa, and tens of genomes from zebra finch and long-tailed finch populations in Australia.

Encoded in the genome of every living organism is a wealth of information about its antecedents. Unlocking this information promises to provide exciting new insights into the history of humans and many other species. Apart from its intrinsic interest, such knowledge is useful for understanding patterns of historical migration[1–4], natural selection[1,5,6], the relationship between humans and other hominids[7–10], and even distantly related phenomena like global climate change[11,12].

Over the past few years, there has been a surge of interest in using whole-genome sequence data from multiple individuals to learn about population demographic histories. There are two popular approaches to this problem: one based on the Poisson random field (PRF) model[13] using the sample frequency spectrum (SFS)[14] and the other based on the sequentially Markov coalescent[15–17]. SFS/PRF methods[18,19] employ results from diffusion or coalescent theory to characterize the sampling distribution of unlinked segregating sites in a random sample of DNA sequences. This distribution can be efficiently computed for a very large sample size[20] and for complex demographic models with multiple populations[21], but the model is incorrect when applied to modern whole-genome sequence data owing to correlation among neighboring sites. Also, the number of parameters that can be estimated using SFS-based methods is bounded by sample size.

With the increasing prevalence of whole-genome sequence data, attention has shifted to more complex models that incorporate recombination and linkage disequilibrium (LD) information. It has been shown that this approach can uncover past population history in unprecedented detail[22]. Because of the linear structure of DNA, these models commonly employ some form of hidden Markov model (HMM)[22–28]. Exploiting LD gives these methods more power to reconstruct past demographic events. At the same time, it entails an added level of mathematical and computational sophistication, such that state-of-the-art methods are limited to analyzing whole-genome samples from only a few genomes at a time. Also, most of these methods require as input computationally phased haplotypes, entailing a costly and potentially error-prone preprocessing step. As we show later, switch errors in phasing can catastrophically distort the inferred history.

In this paper, we present a new inference framework called SMC++ that combines the computational efficiency of the SFS method and the advantage of using LD information in coalescent HMMs. Our method is designed to take advantage of modern data sets consisting of hundreds of unphased whole genomes. It can also analyze pairs of diverged populations; in this scenario, it can pool information from both populations, as well as directly estimate the time of divergence. To our knowledge, this is the first demographic inference method capable of analyzing unphased whole-genome data from a large number of individuals in a computationally efficient and stable manner while taking linkage information into account.

RESULTS

We first demonstrate the accuracy and efficiency of our method, SMC++, on simulated data. Then, we apply the method to analyze real data from several large-scale whole-genome sequencing projects for various species: over a thousand genomes from eight human populations in Africa and Eurasia, hundreds of genomes from a *Drosophila* population in Africa, and tens of genomes from zebra finch and long-tailed finch populations in Australia. Throughout the rest of the paper, we denote the haploid sample size of a data set by n.

[1]Department of Statistics, University of California, Berkeley, Berkeley, California, USA. [2]Computer Science Division, University of California, Berkeley, Berkeley, California, USA. [3]Department of Integrative Biology, University of California, Berkeley, Berkeley, California, USA. [4]Departments of Biology and Mathematics, University of Pennsylvania, Philadelphia, Pennsylvania, USA. Correspondence should be addressed to Y.S.S. (yss@eecs.berkeley.edu).

Received 16 September; accepted 23 November; published online 26 December 2016; doi:10.1038/ng.3748

Accuracy and computational performance on simulated data

The exact details of our simulation procedure are provided in the Online Methods. Briefly, each scenario consisted of ten replicates of 3 Gb of data simulated under either a 'sawtooth' demography[24] featuring repeated exponential expansions and crashes over the last 1 million years or 'recent expansion', a stylized model of European population growth involving a bottleneck 200,000 years ago and ten-fold expansion over the last 10,000 years.

Effect of phasing error

Many inference procedures in population genetics require phased sequence data, leading to an unavoidable and potentially major source of computational expense and error. Because current phasing methods work best when large reference panels are available[29], these problems are especially acute when studying new populations for which a suitable reference panel is not available.

Phasing errors confound demographic inference by breaking up identity-by-state tracts in closely related haplotypes, biasing downward the number of inferred coalescences in the recent past and causing haplotype-based inference methods to infer large recent effective population sizes. To quantify this effect, we simulated $n = 4$ genomes under the sawtooth demography to generate three input data sets for MSMC and SMC++. One version of the data set contained the exact haplotypes generated by the simulation. The other two contained artificially induced switch errors at rates of 1% and 5%, which are approximate upper and lower bounds on the accuracy of existing phasing algorithms[29,30].

Figure 1 shows the result of running SMC++ and MSMC on these data sets. All of the fits exhibited some bias, but the methods showed fairly good agreement and low dispersion in the distant past. In the recent past, large differences emerged in the performance of MSMC as the error rate was increased. With 1% switch error, MSMC had difficulty accurately inferring demography more recent than 20,000 years ago, with highly divergent estimates (off by four orders of magnitude) over the past 3,000 years. With 5% phasing error, performance further deteriorated, with estimates diverging substantially from the truth beginning 20,000 years ago. With no phasing error, the accuracy of MSMC was similar to that of SMC++, with SMC++ giving higher resolution in the recent past. We have used MSMC for illustrative purposes, but we expect any demographic inference procedure that relies on phased data to exhibit similar inaccuracies in the presence of phasing error.

SMC++ in comparison to other methods

Next, we compared SMC++ with MSMC and the SFS-based inference method $\partial a \partial i$[18]. On the basis of the findings in the preceding section, we introduced phasing error into the data at a switch error rate of 1%. (Note that $\partial a \partial i$ is also invariant to phasing.)

For both the sawtooth and recent expansion simulations, SMC++ estimates were much more tightly concentrated around the true demography than either of the other two methods. The spline-based regularization scheme (Online Methods) employed by SMC++ effectively trades a small amount of bias for greatly reduced variance. Estimates obtained from $\partial a \partial i$ appeared to be biased only slightly but varied from run to run, and also between different epochs within the same run. MSMC had difficulty accurately inferring the true history, with estimates diverging quite severely for the very recent past. We note that these findings are somewhat at odds with the simulation results reported by Schiffels and Durbin[24], which seem to have been generated with error-free data.

The estimates for the sawtooth demography (**Fig. 2a**) seemed to deviate more from the truth than those for the recent expansion demography (**Fig. 2b**), and, in general, the sawtooth demography seemed to be harder to accurately infer than the recent expansion demography. This may in part be due to the somewhat pathological nature of the sawtooth demography, which experiences several rapid crashes over a relatively short period of time. It is known, for example, that a strong bottleneck complicates efforts to accurately estimate size change events that occurred before the bottleneck[31]. For the recent expansion demography (**Fig. 2b**), SMC++ tended to smooth over abrupt changes in the historical effective population size, a phenomenon that has been observed previously in related methods[22].

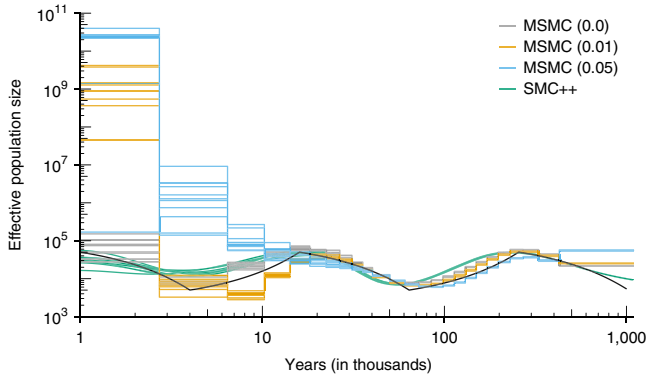

Figure 1 The effect of phasing error. The true population size history is represented by the bold black line, while colored lines correspond to inferred histories for ten simulations each with sample size $n = 4$. For MSMC, switch error was introduced at the rate of 0%, 1% or 5%, as indicated in parentheses. SMC++ does not require phased data, and its results are insensitive to phasing error. With phasing error, MSMC estimates can be off by orders of magnitude in the recent past. In the absence of phasing error, the accuracy of MSMC is comparable to that of SMC++, with SMC++ providing higher resolution in the recent past.

Estimating the rate of recombination

In the preceding simulations, we assumed that the recombination rate (denoted by $\rho/2$) was known ahead of time. When ρ is unknown, SMC++ can estimate this quantity simultaneously with the demography. To test this ability, we generated additional data sets with $n = 50$ under the recent expansion demography, varying ρ within the range $[0.1\theta, 10\theta]$, where $\theta/2$ is the mutation rate. We then used SMC++ to jointly infer both the demography and the recombination rate. Results for recombination rate estimation are shown in **Supplementary Figure 1**. SMC++ was able to estimate the rate of recombination fairly accurately over two orders of magnitude. It was most accurate when the ratio of recombination rate to mutation rate (ρ/μ) was below one. As ρ approaches 10θ, estimates of the ρ/θ ratio exhibit some downward bias, which caused the inferred demographies to become too large. In this regime, recombinations are roughly as common as mutations along the genome, complicating efforts by the HMM to establish the marginal time to the most recent common ancestor (TMRCA). The resulting demographic estimates exhibited additional variation but were qualitatively similar to those obtained when the recombination rate was known (**Fig. 2b** and **Supplementary Fig. 1**).

Inference of split times

SMC++ can analyze pairs of populations simultaneously to infer divergence times jointly with population size histories. The current version of our implementation assumes a 'clean split' model, in which no gene flow occurs after the populations split (**Supplementary Note**). As in the one-population case, divergence time estimates and the jointly inferred demographies do not depend on phasing.

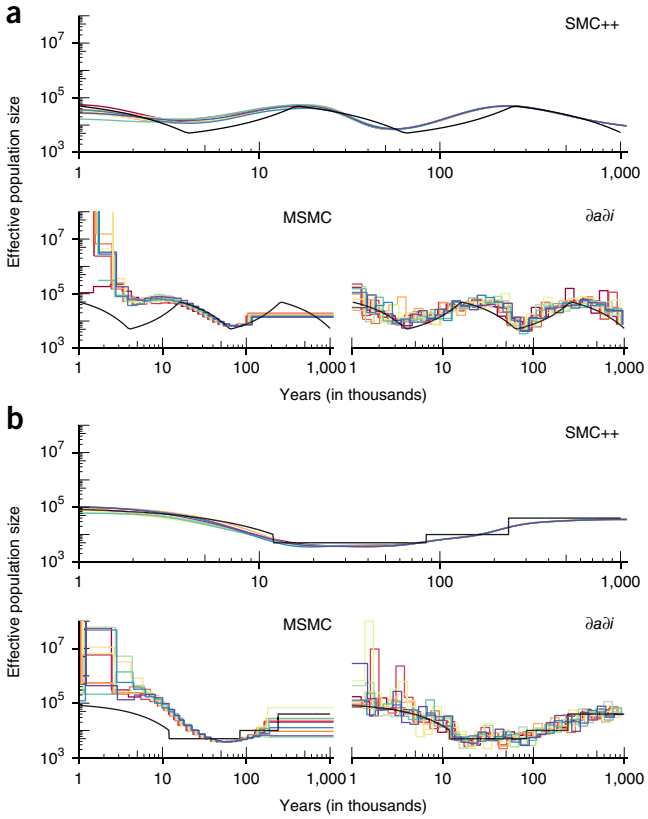

Figure 2 Performance of SMC++ in comparison to MSMC and ∂a∂i. (**a**) The sawtooth demography. (**b**) The recent expansion demography. Each method was used to analyze ten simulated data sets generated according to the demography shown in black. SMC++ was given sequence data from $n = 100$ lineages, and ∂a∂i analyzed the SFS from this data set. MSMC analyzed $n = 8$ of these lineages, the largest sample size for which it successfully ran. For this simulation, we introduced switch error at a rate of 1% at segregating sites. All plots use the same axis scales to aid in comparing the methods, but note that the MSMC fits again diverged to very large values (as high as $O(10^{10})$) in the recent past. (MSMC and ∂a∂i use the same breakpoints from run to run; we jittered the x values of the fits slightly to prevent overlaying.)

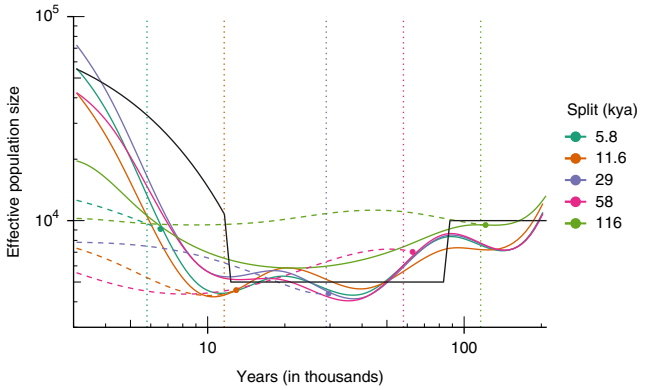

Figure 3 SMC++ results for jointly inferring population size histories and divergence times. Two populations were simulated under the recent expansion demography. Each population consisted of $n = 10$ lineages. Different colors correspond to different divergence times. From the point of divergence until the present, population 2 maintains a constant effective population size equal to the one it had at the time of the split. The solid colored lines represent the inferred demographies for population 1, which should follow the solid black line corresponding to the simulated demography. The dashed colored lines represent the inferred demographies for population 2, which should be flat from the time of the split onward. The vertical dotted lines represent the true values of the splits, whereas solid dots in corresponding colors correspond to the values of the inferred split times. These results show that our method is able to infer divergence times with low error over a wide range of split times, spanning approximately 6,000 to 120,000 years. kya, thousand years ago.

We verified by additional simulations that SMC++ could accurately determine divergence time and demography in the clean split case. For these simulations, we focused on somewhat smaller sample sizes of $n = 10$ lineages in each population. This analysis illustrates the performance of our method when there are limited data; additionally, the computations needed to jointly analyze two populations have higher computational complexity, so it is expedient to fit the joint model to somewhat smaller data sets.

The two populations were simulated according to the recent expansion demography described above. At a time that varied from simulation to simulation, population 2 split from population 1 (moving forward in time) and maintained a constant effective population size up until the present, while population 1 continued to follow the recent expansion demography. No gene flow occurred between the two populations after the split.

Simulation results are shown in **Figure 3**. Over the range of approximately 6,000 to 120,000 years ago, SMC++ was able to infer divergence times with low error. Additionally, the inferred demographies exhibit an acceptable level of accuracy. In situations where additional data are available, SMC++ can also estimate the demography of each population separately and then combine this information to infer the joint demography and divergence time.

Computational performance

We recorded the run time and peak memory consumption for SMC++, ∂a∂i and MSMC while analyzing the simulated data sets described above. To allow for fair comparison, we restricted SMC++ and MSMC to six threads, although we note that SMC++ is capable of exploiting all cores when possible. Results are shown in **Figure 4**. With $n = 8$, the largest setting for which we were able to successfully run MSMC, SMC++ required roughly an order of magnitude less memory and time. ∂a∂i, which operates on the SFS rather than sequence data, required extremely little memory (roughly 100 MB per run) and ran in roughly the same amount of time as SMC++.

For larger sample sizes, memory consumption with SMC++ is approximately constant, and run time also scales quite favorably. It may seem odd that the run time for SMC++ actually decreases slightly as n grows. This decrease is due to the fact that our method's computational performance improves as the density of segregating sites decreases (Online Methods), which occurs when we 'thin' the data more aggressively at larger sample sizes to break up correlations in the underlying ancestral recombination graph (Online Methods). All told, even though there are fewer segregating sites in the thinned data for large values of n, these sites are more demographically informative.

Improvement in posterior decoding

In addition to demographic inference, SMC++ can also be used to locally infer the TMRCA of pairs of lineages (Online Methods). We hypothesized that the 'conditioned SFS' (CSFS), by establishing a link between the coalescence time of a distinguished pair and the allelic status of the rest of the sample, could help to obtain an improved posterior distribution on the former quantity. To examine this hypothesis, we simulated additional 10-Mb stretches of sequence data and compared

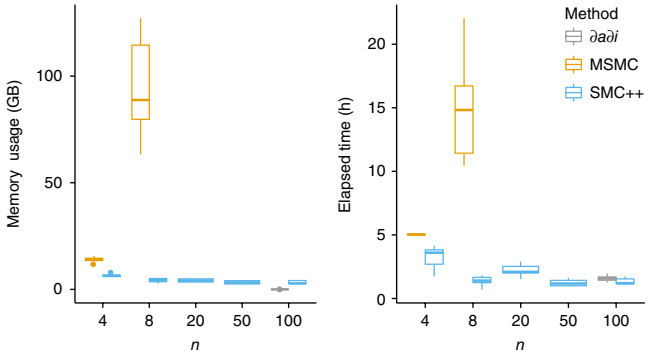

Figure 4 Computational performance of SMC++, MSMC and ∂a∂i. The plots show median memory usage and run time; error bars denote the interquartile range (IQR). Each datum comprises ten repetitions on 3 GB of simulated data. The largest sample size for which we were able to successfully run MSMC was $n = 8$. For large sample sizes ($n \geq 8$), SMC++ requires orders of magnitude less memory and run time than MSMC. For each plot, the lower and upper hinges represent the 25th and 75th percentiles and the middle line is the median; whiskers extend to the nearest observation less than 1.5 times the IQR beyond the corresponding hinge.

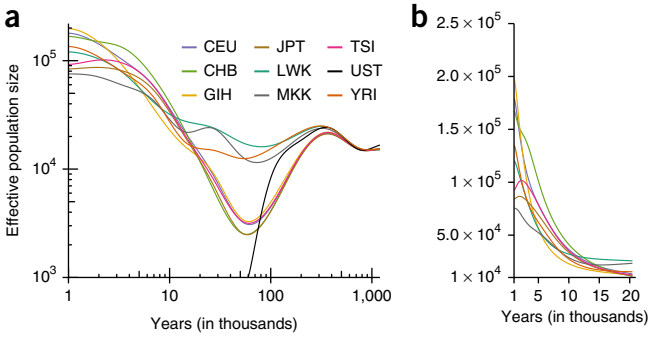

Figure 5 Results of effective population size inference across eight extant human populations and an ancient Ust'-Ishim individual. A generation time of 29 years was used to convert coalescent scaling to calendar time. (**a**) Results for all populations on a log–log scale. The plot assumes that the Ust'-Ishim individual was extant until 45,000 years ago. (**b**) Results for present-day populations on a linear scale over the past 20,000 years. See **Supplementary Tables 2** and **3** for a description of the populations and sample sizes.

the true TMRCA at each position for a distinguished pair to the posterior obtained by running our method. To quantify changes in accuracy, we calculated the root-mean-squared error (RMSE) between the inferred and true TMRCA estimates. (See the Online Methods for a precise definition.)

Supplementary Table 1 shows the results of these simulations. For each scenario, the average RMSE is given relative to the PSMC ($n = 2$) baseline decoding. We consider three strata, resulting from (i) varying the ratio of recombination rate to mutation rate, (ii) introducing genotype error and (iii) introducing missingness into the simulations. (See the Online Methods for a description of how we simulated these errors.)

In general, we confirmed that incorporating additional information from the CSFS improved posterior decoding (**Supplementary Table 1**). For large values of the ratio of recombination rate to mutation rate, the effect was on the order of 1–2% and was statistically significant in all scenarios. In contrast, for very low values of recombination ($\rho/\mu = 0.1$), the results exhibited considerable variance, with certain combinations of simulation parameters leading to substantial improvements and others resulting in no significant improvement. Manually comparing posterior decoding and true TMRCA in these cases identified long spans with very recent coalescence times, which did not accumulate enough mutations to make reliable inference on the TMRCA.

We also examined the case where mutation and recombination occur at comparable rates (**Supplementary Table 1**), which is the most relevant for analyzing human data. For high levels of genotype error (0.1% of sequenced bases), we saw a very substantial decrease in RMSE as sample size increased. This approach could potentially be useful in low-coverage sequencing applications. Lower genotype error rates saw improvements of 1–2% when there was 10% missingness in the data and 5–6% with 20% missingness. Finally, we found that increasing the sample size from 5 to 10, and then to 25, generally decreased the error, but we did not observe much of an improvement beyond this. We suspect that this is because the amount of correlation between the CSFS and the TMRCA of the distinguished pair declines marginally as additional samples are added.

Inference of human demography

We used SMC++ to infer the historical effective population size of eight human populations from Africa and Eurasia. We obtained whole-genome sequences from Complete Genomics[32] and the 1000 Genomes Project[2]. The Complete Genomics data are derived from 54 unrelated individuals obtained from various populations and have average coverage in excess of 45×. The 1000 Genomes Project data have a larger sample size (typically 50–100 individuals per population) but lower coverage. We combined these data sets by treating the genome of each individual from the Complete Genomics data as a 'distinguished' lineage and computing the full CSFS using all available individuals (see the **Supplementary Note** for details on these concepts). Within each population, we fit a model using composite likelihood, whereby SMC++ was instantiated once for each available pair of distinguished lineages and the sum of the log likelihoods was maximized using the expectation–maximization (EM) algorithm. Additionally, we analyzed a single ancient individual (Ust'-Ishim) who lived ~45,000 years ago in western Siberia[33]. The data used in this portion of the analysis are summarized in **Supplementary Tables 2** and **3**.

Results across all populations are shown in **Figure 5**, with estimates broken out by continent in **Supplementary Figures 2–4**. A constant generation time of 29 years[34] was used to convert the scale from coalescent time to calendar time, although we caution that the generation time(s) of ancient human populations are not known precisely, potentially distorting the time scale in the distant past. The per-generation mutation rate was fixed at 1.25×10^{-8} mutations per base pair for all extant populations and 1.45×10^{-8} mutations per base pair for the Ust'-Ishim individual[33].

In **Figure 5** and **Supplementary Figures 2–4**, we show the inferred size history over the period from 1 million to 1,000 years ago, as well as the size history for the past 20,000 years on a linear scale. Several interesting features were apparent. In the period from 1 million to 300,000 years ago, all population size histories experienced similar dynamics. Around 300,000 years ago, the African size histories began to deviate from all other populations, potentially reflecting the existence of population structure within Africa. In the period from 300,000 to 100,000 years ago, the African and non-African populations started to substantially diverge. During this period, the non-African populations experienced a steep decline in effective population size, while the African populations experienced a more moderate decline.

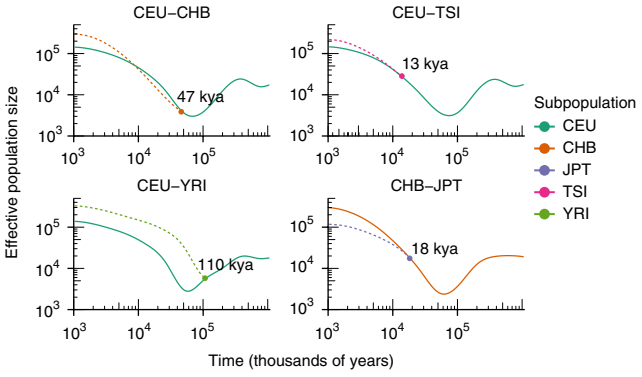

Figure 6 Inference of split times in modern humans. Results of jointly estimating population size histories and split times in a two-population model. The same data and generation times as in **Figure 5** were used to generate the plot.

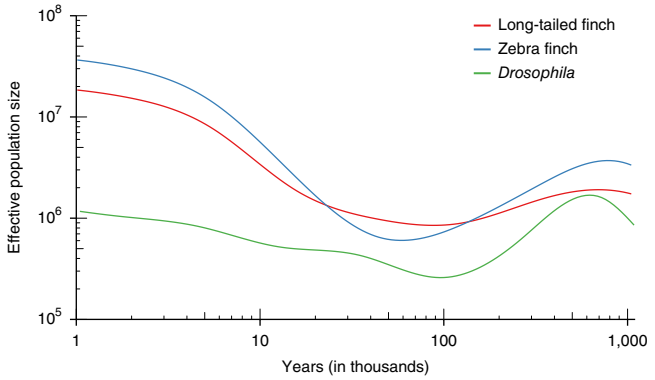

Figure 7 Results of effective population size inference for two finch species and *Drosophila*. Generation times of 3 months (finch) and 1 month (*Drosophila*) were used to convert coalescent scaling to calendar time. See **Supplementary Table 3** for a description of the populations and sample sizes.

There was strong concordance in the population size histories of the European and Asian populations over this time interval, consistent with the notion that they split from a common ancestral population more recently than 100,000 years ago.

In the period more recent than 100,000 years ago, all of the various populations became more distinct. Part of the observed short-term variability is due to the increased variance in our estimates as we approach the present (**Supplementary Fig. 5**), but differences in the global pattern should reflect real differences between population size histories. The African populations (**Supplementary Fig. 2**) all experienced a relatively mild bottleneck followed by growth in the period from 100,000 to 10,000 years ago. The Luhya (LWK) and Maasai (MKK) reached a nadir around 70,000 to 80,000 years ago, and the Yoruba (YRI) reach a nadir somewhat later, around 50,000 years ago. All three populations seemed to experience growth from at least 50,000 years ago up to the present, with a noticeable uptick starting approximately 15,000 years ago. Note that additional variability and lower recent effective population size seen for Maasai are likely due in part to the much smaller sample size of this population (**Supplementary Table 2**).

Among the Asian populations (**Supplementary Fig. 3**), a somewhat different pattern emerged. All three populations experienced a sharp bottleneck, reaching a nadir around 55,000 years ago, followed by growth up to the present. Sudden, rapid growth was experienced by the Gujarati (GIH) population around 5,000 to 10,000 years ago, whereas in the other two populations (Han Chinese (CHB) and Japanese (JPT)) growth began earlier, around 15,000 years ago. It is interesting to note that the Gujarati population appeared to be distinct from the East Asian populations starting around 100,000 years ago.

The European populations (**Supplementary Fig. 4**) experienced a similar bottleneck as Asians, ending around 50,000 years ago, followed by a period of rapid growth starting 10,000 to 15,000 years ago. The size histories of the Tuscan (TSI) and northern European (CEU) populations were nearly identical until around 5,000 years ago.

Several of the populations in the sample experienced similar size histories before diverging at some point in the recent past (**Supplementary Figs. 2–4**). To study this further, we jointly estimated their size histories and split times using the two-population model described above. Results are shown in **Figure 6**. Our model estimated a divergence time of 13,000 years ago for the northern European and Tuscan populations; 18,000 years ago for the Han Chinese and Japanese populations; 47,000 years ago for the northern European and Han Chinese populations; and 110,000 years ago for the northern European and Yoruban populations. The joint estimates of population size shown in **Figure 6** differ slightly from the marginal estimates shown in **Figure 5** because the assumption of continuity between the two size histories at the time of divergence places additional constraint on the estimation problem. Overall, however, the estimates are quite consistent.

Finally, we note the close agreement between the Ust'-Ishim individual and modern populations in **Figure 5** over the period from 1 million to 45,000 years ago. The highly disparate origins of these samples—the former consisting of ancient DNA and the latter obtained from present-day individuals—gives us confidence that the features emerging in correspond to actual historical events and not data or modeling artifacts.

Inference of demography in other species

To verify that SMC++ has applications beyond human demographic inference, we also inferred population size histories for two species of finch (the zebra finch, *Taeniopygia guttata*, and the long-tailed finch, *Poephila acuticauda*) from Australia[35], as well as a wild African population of *Drosophila* from Zambia[36]. These data sets are described in **Supplementary Table 3**. For finches, we assumed a generation time of 3 months and a per-generation mutation rate of 7×10^{-10} mutations per base pair[35]. For *Drosophila*, we assumed a generation time of 1 month and a per-generation mutation rate of 3×10^{-9} mutations per base pair[37]. Because of uncertainty surrounding the generation times of these species, we also plotted the results in generations (**Supplementary Fig. 6**). For all three species, we estimated the per-generation recombination rate from data. The genome-wide average ratios of recombination rate to mutation rate were estimated to be 1.6:1 (*Drosophila*), 1.15:1 (*P. acuticauda*) and 1.1:1 (*T. guttata*).

Results of demographic inference are shown in **Figure 7**. The two finch species had similar, stable size histories between 1 million and 500,000 years ago. Starting around 500,000 years ago, the populations experienced a decline that persisted until about 60,000 to 80,000 years ago, depending on the species. This was followed by a period of expansion, which accelerated around 15,000 years ago and led to more than a tenfold increase in effective population size by 1,000 years ago.

Estimates for *Drosophila* were overall lower, although we caution that the short generation time of *Drosophila*, in combination with their smaller genome size, complicates efforts to peer this far back into its past using genetic data. The *Drosophila* population seemed to

decline from 600,000 to 100,000 years ago, at which point it experienced steady growth leading to the present. Unlike the other species analyzed in this paper, for *Drosophila* we did not see as much evidence of a sudden increase in effective population size in the recent past.

DISCUSSION

In this paper, we have presented SMC++, a new demographic inference method capable of analyzing hundreds of unphased whole-genome sequences at a time while being fast, robust and easy to use. The ability to analyze much larger sample sizes makes our method's estimates substantially more accurate than those of previous methods, especially in the recent past. On simulated data, we obtain accurate estimates of the true effective population size history across a time span of three orders of magnitude. Furthermore, our method is able to estimate population split times with low error over a wide range of time. In real data, we obtain convincing estimates of the effective population size history of a number of different populations. In most cases, our estimates agree with previous findings on divergence times and historical migration patterns while also bringing to light some intriguing new features that could merit further study.

Two aspects of our method in particular are worth re-emphasizing. We have shown using simulations that introducing a modest amount of phasing error can severely corrupt the estimates of an existing demographic inference method. We conjecture that any method that is not invariant to phasing suffers from similar issues. Hence, the phase invariance of our method makes it more robust. Additionally, it eliminates a burdensome preprocessing step when analyzing real data.

We have also demonstrated that SMC++ requires an order of magnitude less memory and processing time than existing methods. We have found in practice that this is extremely useful for exploring and testing hypotheses in real data. The results of any demographic analysis depend on a number of a priori modeling assumptions, such as the functional form of the demography and various tuning parameters. At present, we lack a theoretical understanding of how to optimally choose these parameters. This is an important area for future research, and, in the meantime, the ability to explore the model space and receive rapid feedback from the algorithm is essential.

Our general theoretical framework, which couples the genealogical process for a given diploid individual with the allele frequency information in a large collection of other individuals, can be extended to more complex demographic models. In particular, we plan to extend our method to incorporate gene flow between populations. We believe that this approach opens up a new window of opportunity to utilize the information contained in a large collection of whole genomes to infer population demographic history in finer detail and with higher accuracy than has previously been possible.

URLs. PopGenMethods repository, https://github.com/popgenmethods/smcpp; Complete Genomics diversity panel, http://www.completegenomics.com/public-data/69-genomes/; 1000 Genomes Project Phase 3, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/; *Drosophila* Genome Nexus, http://www.johnpool.net/genomes.html.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Pool and C. Langley for helpful comments on our inferred *Drosophila* demography. We also thank H. Li for providing us with the Ust'-Ishim genome sequence. This research is supported in part by NIH grants R01 GM094402 and R01 GM108805 and by a Packard Fellowship for Science and Engineering (Y.S.S.).

AUTHOR CONTRIBUTIONS

J.T., J.A.K. and Y.S.S. conceived the study, developed the theoretical model and wrote the manuscript. J.T. developed software implementing the method and performed data analysis. J.A.K. contributed benchmarks of $\partial a \partial i$.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
2. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
3. Skoglund, P. *et al.* Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015).
4. Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
5. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
6. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16**, 359–371 (2015).
7. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
8. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
9. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
10. Vernot, B. & Akey, J.M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014).
11. Miller, W. *et al.* Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc. Natl. Acad. Sci. USA* **109**, E2382–E2390 (2012).
12. Stewart, J.R. & Stringer, C.B. Human evolution out of Africa: the role of refugia and climate change. *Science* **335**, 1317–1321 (2012).
13. Sawyer, S.A. & Hartl, D.L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).
14. Griffiths, R.C. & Tavaré, S. Sampling theory for neutral alleles in a varying environment. *Proc. R. Soc. Lond. B* **344**, 403–410 (1994).
15. Wiuf, C. & Hein, J. Recombination as a point process along sequences. *Theor. Popul. Biol.* **55**, 248–259 (1999).
16. McVean, G.A. & Cardin, N.J. Approximating the coalescent with recombination. *Phil. Trans. R. Soc. Lond. B* **360**, 1387–1393 (2005).
17. Marjoram, P. & Wall, J.D. Fast "coalescent" simulation. *BMC Genet.* **7**, 16 (2006).
18. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
19. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
20. Bhaskar, A., Wang, Y.X.R. & Song, Y.S. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Res.* **25**, 268–279 (2015).
21. Kamm, J.A., Terhorst, J. & Song, Y.S. Efficient computation of the joint sample frequency spectra for multiple populations. *J. Comput. Graph. Stat.* (in the press).
22. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
23. Dutheil, J.Y. *et al.* Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* **183**, 259–274 (2009).
24. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
25. Paul, J.S., Steinrücken, M. & Song, Y.S. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics* **187**, 1115–1128 (2011).
26. Steinrücken, M., Paul, J.S. & Song, Y.S. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theor. Popul. Biol.* **87**, 51–61 (2013).
27. Sheehan, S., Harris, K. & Song, Y.S. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* **194**, 647–662 (2013).
28. Steinrücken, M., Kamm, J.A. & Song, Y.S. Inference of complex population histories using whole-genome sequences from multiple populations. Preprint at. *bioRxiv* http://dx.doi.org/10.1101/026591 (2015).

29. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
30. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
31. Terhorst, J. & Song, Y.S. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc. Natl. Acad. Sci. USA* **112**, 7677–7682 (2015).
32. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
33. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
34. Langergraber, K.E. *et al.* Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci. USA* **109**, 15716–15721 (2012).
35. Singhal, S. *et al.* Stable recombination hotspots in birds. *Science* **350**, 928–932 (2015).
36. Lack, J.B. *et al.* The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics* **199**, 1229–1241 (2015).
37. Keightley, P.D., Ness, R.W., Halligan, D.L. & Haddrill, P.R. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics* **196**, 313–320 (2014).

ONLINE METHODS

SMC++ stands for 'sequential Markov coalescent + plenty of unlabeled samples'. SMC++ unites the PRF and coalescent HMM approaches, combining the strengths of each while overcoming several of their limitations. The inclusion of 'unlabeled samples' in the standard coalescent HMM is achieved via novel theoretical results on what we term CSFS, the sample frequency spectrum conditioned on the coalescence time and allelic state of a distinguished diploid individual. In comparison to existing methods, the main advantages of SMC++ are as follows.

1. Scalability. SMC++ can analyze hundreds of individuals at a time while requiring a modest amount of memory and processing time. Analyzing 100 human genomes takes roughly 1 h on a laptop.
2. Accuracy. By accommodating larger sample sizes, SMC++ has greatly improved power to infer demographic events, particularly in the recent past.
3. Phase invariance. SMC++ only requires unphased sequence data as input (that is, results do not depend on phasing).
4. Regularity. SMC++ uses cubic splines to enforce a smoothness constraint on the inferred demographies. In comparison with existing methods, the resulting estimates exhibit far less variance, with only a minimal increase in bias.

SMC++ extends a line of work that approximates the intractable sampling distribution of a contiguous segment of DNA (that of the so-called ancestral recombination graph[38]) by a tractable Markov process. We first give a brief overview of these related methods, followed by a description of the novel aspects of SMC++.

Related work. The starting point for our model is the well-known pairwise sequentially Markov coalescent[22]. The corresponding paper applied a simple likelihood model for the pattern of genetic mutations observed in a single diploid individual and showed that it had surprising power to infer the history of the individual's population. PSMC works as follows: each pair of chromosomes is divided into blocks of 100 bp and then mapped to a binary string according to whether each block contained one or more heterozygous sites. These 'emitted' symbols are modeled as an HMM whose hidden state is the marginal TMRCA, assumed to be constant within each block. Conditional on the TMRCA being equal to T, the probability of observing a heterozygous block is $1 - e^{-100\theta T}$ where $\theta/2$ is the per-base mutation rate. The distribution of the TMRCA is discretized into a number of disjoint intervals, and its stationary distribution and (discrete) transition function are then computed with respect to a piecewise-constant population size history. The transition function, which describes the conditional distribution of TMRCA between neighboring sites, is intractable in its exact form[15]. Instead, a simple Markov approximation is used[16]. The stationary, emission and transition probabilities together define a map from population size history to sampling distribution via the HMM. The likelihood function is then maximized iteratively using the EM algorithm.

Despite its apparent simplicity, PSMC has proved very effective in practice and has been used in hundreds of subsequent analyses of many different species. Nevertheless, there are a few clear areas where PSMC could be improved:

- The so-called SMC approximation[16] used to calculate transition probabilities in PSMC is inaccurate. A slight modification known as SMC' (ref. 17) results in a much more accurate approximation of the transition matrix, with negligible added computational burden[39,40].
- The computational performance of PSMC, although adequate, can be substantially improved via parallelism on modern, many-core workstations. Doing so renders the blocking approximation unnecessary.
- More fundamentally, PSMC is limited to analyzing a single diploid individual. This strongly limits the power of PSMC, particularly in the recent past where there will be few coalescent events in a sample of size two.

A follow-on method called MSMC[24] extends PSMC to larger sample sizes and also allows for the analysis of multiple populations. MSMC does so by redefining the HMM hidden variable to be the first coalescent event among the n haplotypes in the sample, with transition probabilities based on SMC'. For the emission probabilities, MSMC mainly uses the singleton mutations, only using the higher-frequency mutations to disallow certain coalescence events (if the two coalesced haplotypes have different alleles, the emission probability of the non-singleton mutation is 0; otherwise, the mutation is ignored). For the case of $n = 2$, MSMC essentially reduces to PSMC. Since the time to first coalescence depends more strongly on recent demography for larger sample sizes, MSMC has greater power to infer size changes in the recent past. On the other hand, the computational burden of MSMC also grows substantially with sample size, such that it is limited, in practice, to analyzing no more than four diploid whole-genome samples.

A different approach to performing demographic inference via a coalescent HMM may be found in diCal[27] and related methods[28,41]. These rely on the so-called conditional sampling distribution (CSD), which characterizes the probability of observing an kth sampled chromosome conditional on the previously observed $k - 1$ samples. Computing the CSD becomes substantially more difficult for larger sample sizes, so that even the most recent version of diCal is limited to analyzing no more than ten samples at a time at whole-genome scale. On the other hand, diCal is capable of analyzing complex demographic scenarios involving size changes, admixture, migration and population splits.

Theoretical model for SMC++. Above, we described how MSMC generalizes PSMC in two different directions, by altering both the hidden state space as well as the space of emitted symbols used in the HMM. We also demonstrated in simulation that MSMC has difficulty scaling to large values of n. This is partly due to the fact that there are $O(n^2)$ pairs of haplotypes that may be involved in the first coalescence; thus, the HMM must integrate over $O(n^2)$ hidden states per time interval per locus.

To scale to larger sample sizes, SMC++ generalizes PSMC in a different direction. The hidden state of SMC++ is exactly the same as in PSMC; specifically, the hidden state is the TMRCA between the two haplotypes of a single individual. Where SMC++ and PSMC differ is in the observed state. Whereas PSMC emits a binary symbol indicating whether the individual is heterozygous, SMC++ additionally emits the allele frequency of an extra $n - 2$ haplotypes. Specifically, the emission probability of SMC++ is based on the site frequency spectrum[14,42], conditioned on the TMRCA of a single 'distinguished' individual.

The differences between these three approaches are depicted in **Supplementary Figure 7**. PSMC and SMC++ track the same hidden information, but their emissions are different. Whereas the PSMC emissions are binary symbols, SMC++ emissions are 2-tuples $(a,b) \in \{0,1,2\} \times \{0,1,\ldots,n-2\}$. Here a and b denote the number of derived alleles present in the distinguished individual and the additional undistinguished n − 2 haplotypes, respectively. We refer to the distribution of these tuples as the CSFS: the site frequency spectrum conditional on the event that the TMRCA of the distinguished member falls in a given interval. The CSFS is a 2D generalization of the site frequency spectrum; indeed, summing the CSFS across rows exactly recovers the standard site frequency spectrum on n haploids. We explicitly define the SMC++ emission probability and rigorously derive the CSFS using coalescent theory, in the **Supplementary Note**.

To calculate the CSFS we assume that the mutation rate is known. Uncertainty in the mutation rate can potentially alter the estimates of SMC++, although in practice this effect will be attenuated by also using linkage information to infer the demography.

Thinning. An important caveat to our approach concerns correlation between observations of the CSFS. The HMM formally assumes that neighboring observations are independent conditional on knowing the underlying hidden state. If we just use the distinguished individual, without the additional undistinguished haplotypes, this assumption is correct (up to the SMC' approximation and time discretization). However, the conditional independence assumption is violated when we add the additional undistinguished lineages: correlations in branch lengths of the sample's underlying genealogy will remain even after conditioning on the TMRCA within the distinguished individual. Emitting the 'full' CSFS at every site therefore leads to model misspecification. To prevent this, we adopt a thinning strategy, in which the full CSFS is emitted only at every kth site for some prespecified constant k.

In this paper, we adopted the heuristic $k = Cn$ for a large constant C, so that the density of full SFS emissions decreases linearly in sample size. We found

doi:10.1038/ng.3748

that $C = 400$ worked well in practice, and all analyses reported in this paper are the result of using that choice of parameter.

Transition function. As discussed earlier, the SMC' transition function more accurately approximates the sequential coalescent. In SMC++, we employ a continuous time, conditioned Markov chain approximation to the sequential coalescent[39]. This approximation, which derives from the exact continuous time description of the two-locus coalescent with recombination[43], is actually slightly more accurate than SMC' as originally formulated[17] because it integrates over any number of recombinations and back-coalescences occurring between neighboring sites.

Expectation–maximization algorithm. For an HMM with M hidden states and L observations, the complexity of the forward–backward algorithm is $O(M^2L)$. Genetic data usually contains long stretches of monomorphic sites, a fact which can be used to decrease the running time of this algorithm to $O(M^2L_p)$, where p is the number of polymorphic loci in the data set[44].

However, extending this speedup to the EM algorithm is nontrivial: in particular, it was previously unknown how to compute the posterior expected number of transitions and emissions from each hidden state, which are needed to execute the 'M' step of the EM algorithm during model fitting. In the **Supplementary Note**, we present new results to compute these posterior expectations by taking advantage of the sparsity of polymorphic sites. The complexity of our modified algorithm is $O(M3L_p)$ that improves the running time $O(M^2L)$ of the non-sparse algorithm when $ML_p < L$. For large sample sizes $L = O(10^2L_p)$ so this requires that $M = O(10)$, a setting which we found produces acceptable results in practice.

Regularization. As mentioned above, we found that regularization dramatically improved convergence of our algorithm and the quality of the resulting estimates. SMC++ enforces a mild smoothness constraint by fitting a cubic spline to the data. Hence the space of models considered by our algorithm consists of continuously twice-differentiable (C^2) curves. (A weaker constraint requiring only C^1 smoothness is also available to the user.)

In addition to this 'implicit' regularization scheme, we added an explicit 'roughness' penalty to the optimization. Here roughness is defined as

$$P(f) \overset{\text{def}}{=} \int_R (f^n)^2$$

where R is the (compact) support of the spline f. The penalized likelihood used for fitting is then $Q(f) - \lambda P(f)$, where $Q(f)$ is the so-called complete log likelihood used in the EM algorithm[45] and $\lambda > 0$ is the regularization parameter.

By penalizing curvature, this type of regularizer shrinks to a linear fit and encodes our prior belief that real populations are unlikely to experience repeated crashes and expansions over short time intervals. For our simulated results, we found that $\lambda = 0.01$ worked well. On real data, which contain noise, missingness and other model violations, we found that slightly larger values, ranging from $\lambda = 1.0$ to $\lambda = 10.0$ depending on the amount and quality of available data, were necessary to obtain smooth fits.

Simulation procedure. We used the coalescent simulation program scrm[46] to generate simulated data sets. Each simulated data set consisted of 30 independently generated chromosomes of length 100 Mb. Ten replications were performed for each combination of demography, inference procedure and sample size. When running MSMC, we fixed the value of the recombination rate to its true value, while for SMC++ the recombination rate was estimated from the simulated data.

The command line used to simulate from the sawtooth demography was

```
scrm <n> 1 -p 10 -t 71560.0
-r 17889.9998211 100000000
-oSFS -seeds <s1> <s2> <s3> -eN 0.0 5.0
-eG 0.0 0.0 -eN 0.000582262 5.0
-eG 0.000582262 1318.18 -eN 0.00232905 0.500002043581
-eG 0.00232905 -329.546 -eN 0.00931919 5.00496081234
-eG 0.00931919 82.3865 -eN 0.0372648 0.500618264601
-eG 0.0372648 -20.5966 -eN 0.149059 5.0061592508
-eG 0.149059 5.14916 -eN 0.596236 0.500615510407
-eG 0.596236 0.0
```

The command line used to simulate from the human demography was

```
scrm <n> 1 -p 10 -t 50000.0
-r 12499.999875 100000000
-oSFS -seeds <s1> <s2> <s3> -eN 0.0 10.0
-eG 0.0 230.258509299 -eN 0.01 0.5 -eG 0.01 0.0
-eN 0.07 1.0 -eG 0.07 0.0 -eN 0.2 4.0 -eG 0.2 0.0
```

In these commands, `<n>` is the (haploid) sample size, which varied from experiment to experiment, and `<s{1,2,3}>` are the random number generator seeds, which varied from experiment to experiment but were retained to enable reproducibility.

To model the effect of computational phasing, we introduced switch error into the data sets using the following procedure. Let $\boldsymbol{X} \in \{0,1\}^{2\times S}$ be a binary matrix representing a pair of dimorphic haplotypes at S segregating sites. The ith column of $\boldsymbol{X}$ is denoted $\boldsymbol{X}_i = (X_{i,1}, X_{i,2})$. Also let U_i, $i = 1, \ldots, S$ denote a sequence of i.i.d. Uniform(0,1) random variables and let α denote the switch error rate. Execute the following random algorithm:

1. Set `switch_error = False`.
2. For $i = 1, \ldots, S$:
   a. If $U_i < \alpha$, then `switch_error = ¬ switch_error`.
   b. If `switch_error = True`, then $\boldsymbol{Y}_{i,1} = \boldsymbol{X}_{i,2}$ and $\boldsymbol{Y}_{i,2} = \boldsymbol{X}_{i,1}$; otherwise, $\mathbf{Y}_i = \mathbf{X}_i$.
3. Return $\boldsymbol{Y} \in \{0,1\}^{2\times S}$.

Posterior decoding. The RMSE between the true TMRCA and inferred posterior is defined as

$$\text{RMSE} = \left[\sum_{\ell=1}^{L}\sum_{m=1}^{M} \gamma_{\ell m}\,(h_m - t_\ell)^2\right]^{1/2}.$$

Here γ_{lm} is the posterior probability of coalescence in the mth hidden state for position l, t_1 is the true TMRCA at position l and h_m is the expected coalescence time conditional on coalescence occurring in hidden state m.

To simulate genotype error, we flipped a Poisson-distributed number of randomly chosen bases in each simulation. To simulate missingness, we replaced a Poisson-distributed number of segregating sites in each sample with missing values. The means of the Poisson distributions were chosen such that, in expectation, the fraction of bases affected by each type of error equaled the corresponding value in **Supplementary Table 1**.

Code availability. Our software, including source code and a Python-based command line interface, is publicly available at the PopGenMethods repository.

Data availability. All data analyzed in this study are publicly available. Whole-genome sequencing data were obtained from the Complete Genomics diversity panel. Frequency spectrum data were obtained from the 1000 Genomes Project Phase 3 release. Sequence data for the two species of finch were obtained from the European Nucleotide Archive, accession PRJEB10586. *Drosophila* data were obtained from the *Drosophila* Genome Nexus. Scripts used to generate the simulated data are available in the code repository.

38. Griffiths, R.C. & Marjoram, P. in *Progress in Population Genetics and Human Evolution* (eds. Donnelly, P. and Tavaré, S.) **87**, 257–270 (Springer-Verlag, 1997).
39. Hobolth, A. & Jensen, J.L. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theor. Popul. Biol.* **98**, 48–58 (2014).

40. Wilton, P.R., Carmi, S. & Hobolth, A. The SMC is a highly accurate approximation to the ancestral recombination graph. *Genetics* **200**, 343–355 (2015).
41. Tataru, P., Nirody, J.A. & Song, Y.S. diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics* **30**, 3430–3431 (2014).
42. Polanski, A. & Kimmel, M. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165**, 427–436 (2003).
43. Simonsen, K.L. & Churchill, G.A. A Markov chain model of coalescence with recombination. *Theor. Popul. Biol.* **52**, 43–59 (1997).
44. Paul, J.S. & Song, Y.S. Blockwise HMM computation for large-scale population genomic inference. *Bioinformatics* **28**, 2008–2015 (2012).
45. Bishop, C.M. *Pattern Recognition and Machine Learning* (Springer, 2006).
46. Staab, P.R., Zhu, S., Metzler, D. & Lunter, G. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* **31**, 1680–1682 (2015).

doi:10.1038/ng.3748