Inference of complex population histories using whole-genome sequences from multiple populations

Matthias Steinrücken[a,b], Jack Kamm[c,d], Jeffrey P. Spence[e], and Yun S. Song[c,d,f,1]

[a]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637; [b]Department of Human Genetics, University of Chicago, Chicago, IL 60637; [c]Department of Statistics, University of California, Berkeley, CA 94720; [d]Chan Zuckerberg Biohub, San Francisco, CA 94158; [e]Computational Biology Graduate Group, University of California, Berkeley, CA 94720; and [f]Computer Science Division, University of California, Berkeley, CA 94720

Edited by Elizabeth A. Thompson, University of Washington, Seattle, WA, and approved July 10, 2019 (received for review March 26, 2019)

There has been much interest in analyzing genome-scale DNA sequence data to infer population histories, but inference methods developed hitherto are limited in model complexity and computational scalability. Here we present an efficient, flexible statistical method, diCal2, that can use whole-genome sequence data from multiple populations to infer complex demographic models involving population size changes, population splits, admixture, and migration. Applying our method to data from Australian, East Asian, European, and Papuan populations, we find that the population ancestral to Australians and Papuans started separating from East Asians and Europeans about 100,000 y ago, and that the separation of East Asians and Europeans started about 50,000 y ago, with pervasive gene flow between all pairs of populations.

coalescent | population genetics | demography | statistical inference

Whole-genome sequences are now routinely available for population genetic analyses, and inference methods that can take better advantage of genome-scale data have received considerable attention in recent years. In particular, there has been much interest in methods that can use the genomic data of individuals from multiple populations to infer complex models of population history. In addition to being of historical interest, population demography is important to study because it influences patterns of genetic variation, and understanding the intricate interplay between demography and other evolutionary forces such as natural selection is a major aim in population genetics.

Inferring these demographic histories is computationally and statistically challenging. One class of methods (1–8) based on the sample frequency spectrum (SFS) is computationally efficient but ignores linkage information, and the minimax rate of convergence for such estimators is poor (9). Also, their utility is limited by the fact that the number of model parameters that can theoretically be identified using the SFS alone is bounded by the sample size (10). Although no identifiability results currently exist for the general case, methods (13–21) that take linkage structure into account are empirically more statistically efficient and can be used to infer models with many parameters even when the sample size is small.* This is of practical importance, since an increasing number of studies now seek to infer complex demographic models involving multiple populations using only a small number of individuals sampled from each population (e.g., refs. 22–25). A popular demographic inference method of this kind is PSMC (pairwise sequentially Markovian coalescent) (13), which uses a pair of sequences to infer piecewise-constant population size histories. Its extension, MSMC (multiple sequentially Markovian coalescent) (18, 19), can use sequences sampled from a pair of populations to infer a genetic separation history in addition to population size changes. A more recent method called SMC++ (20) can scale to hundreds of individuals, but it is able to analyze individuals from only a pair of populations that have diverged without subsequent gene flow.

Parallel to these developments, an inference method called diCal (Demographic Inference using Composite Approximate Likelihood) (16) was introduced to infer piecewise-constant effective population size histories using multiple sequences, thereby providing improved inference about the recent past. The key mathematical component of diCal is the conditional sampling distribution (CSD) π_Θ, which describes the conditional probability of observing a new sequence or haplotype given a collection of already observed haplotypes, under a given population genetic model with parameters Θ. The corresponding genealogical process can be formulated as a hidden Markov model (HMM), enabling efficient inference.

Here, we present our method diCal2, a scalable inference tool for population genomic analysis under general demographic models, which extends diCal in several ways. diCal2 has been successfully applied in recent empirical studies of human demographic history using whole-genome sequencing data (22, 24, 25). In the present paper, we provide a detailed description of the method, carry out a simulation study to benchmark its performance, and discuss its strengths and weaknesses.

To handle gene flow between populations, diCal2 builds on previous theoretical work (26) which introduced a CSD for subdivided populations with unchanging continuous migration; that earlier work did not address parameter estimation, which is the focus of this article. In contrast to MSMC, which does not explicitly model population structure, we consider fully parametric demographic models, including subdivided population structure with migration, that are easier to interpret. Our method also

Significance

An increasing number of population genomic studies now try to infer complex models of population history using a number of whole-genome sequences sampled from multiple populations. A key technical challenge to this effort is to compute model likelihoods, which involves integrating out latent variables (genealogical histories) that live in extremely high dimensions. This is a notoriously difficult computational problem, especially when the sample size is greater than a handful and the underlying population genetic model is complex. Here, we present an efficient, flexible statistical method that can scale to larger sample sizes and more populations than previously possible. Aside from demographic inference, our method can be used in other statistical inference problems in evolutionary biology and human genetics.

Author contributions: M.S. and Y.S.S. designed research; M.S. and J.K. developed algorithms and software; M.S., J.K., J.P.S., and Y.S.S. performed research; M.S. and J.P.S. analyzed data; and M.S., J.K., J.P.S., and Y.S.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: yss@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1905060116/-/DCSupplemental.

Published online August 6, 2019.

*Whether the distribution of pairwise coalescence times uniquely determines the demographic model has been answered recently (11, 12).

enables inference under demographic models more general than the 2-population clean-split model currently implemented in SMC++. Specifically, our method is flexible enough to model 1) an arbitrary number of populations specified by the user, 2) an arbitrary pattern of population splits and mergers, 3) more general population size changes (e.g., piecewise-exponential), 4) arbitrary migration patterns with time-varying continuous migration rates or pulse admixture events, and 5) an arbitrary poly-allelic mutation model at each site (including diallelic or tetraallelic). These are significant improvements on the previous version of diCal (16), which could only be used to infer piecewise-constant population size changes in a single population.

In addition to these features, we introduce major computational improvements which enable the use of whole-genome data. The mathematical details of our method and the computational extensions are provided in *Materials and Methods* and *SI Appendix*, *SI Text*. Below, we briefly highlight the key technical advances.

In PSMC and the earlier version of diCal, the demographic epochs and HMM discretization intervals are both fixed, and the latter forms a strict refinement of the former. In contrast, discretization intervals and demographic epochs are decoupled in our improved version of diCal. For example, population size change points or population split times can vary freely and do not need to coincide with discretization interval boundaries. This flexibility allows for more accurate parameter estimation, especially for population split times.

Moreover, the CSDs for different haplotypes can be combined in various ways to devise a composite likelihood that can be used in a maximum likelihood framework for parameter estimation. Our implementation of the expectation–maximization (EM) algorithm allows any composite likelihood that is composed of sums and products of CSDs, which includes the product of approximate conditionals (PAC) used by Li and Stephens (27) to detect recombination hotspots.

For substantial computational speedup, we implement a previously described "locus-skipping" algorithm (28), which analytically and exactly integrates over contiguous stretches of non-segregating loci. However, locus skipping is less computationally efficient with missing data, and thus, to efficiently incorporate missing alleles, we also implement an alternative speedup by grouping loci together into larger blocks. A similar speedup was used in PSMC, but it treats the whole block as a single diallelic site. In contrast, the blocks in our method are viewed as full nonrecombining haplotypes.

For complex demographic models, the likelihood function may have local optima. To address this issue, we implement a flexible genetic algorithm and combine it with the EM procedure to enable more efficient navigation of high-dimensional parameter space.

Lastly, we also present an application of our method to data from the Simons Genome Diversity Project (SGDP) (23) to investigate the population history of Australians, East Asians, Europeans, and Papuans. There has been some debate whether the population ancestral to Australians and Papuans (which we call Australo-Papuans, following ref. 29) split off prior to the divergence of East Asians and Europeans (e.g., ref. 29), or whether East Asians and Australo-Papuans first split from Europeans (e.g., ref. 30). We find substantial evidence in favor of the former hypothesis, but find that there has been pervasive gene flow between all of these populations since their divergence.

Results

To demonstrate the flexibility, accuracy, and efficiency of our method, we performed an extensive simulation study under a variety of biologically relevant demographic scenarios. DNA sequence data were simulated using the software scrm (31). We simulated 100 datasets for each demographic scenario and set the haplotype length to 250 Mbp for each dataset. We used 1.25×10^{-8} per generation for the per-site mutation and recombination rates and used a value of 100 kbp for the value of the -l parameter in scrm, which determines the length of the recombination history to be used during the simulation. Generation time was assumed to be 30 y. In each scenario, we used our method to estimate all demographic parameters of the underlying model.

Recent Exponential Growth. The first model we investigated involves recent exponential population growth. To investigate the performance of our method, we simulated data consisting of 10 haplotypes under the demographic model depicted in Fig. 1*A*. We fixed $T_B = 65$ ka, $T_G = 15$ ka, $N_A = 15,000$, and $N_B = 1,800$, and used 3 different values for the growth rate, r: $0.25\%, 0.5\%$, and 1.0% per generation.

We used the leave-one-out composite likelihood (LCL) in our EM procedure combined with a genetic algorithm to estimate all 5 parameters of the demographic model. For the genetic algorithm, we chose 50 random starting points that were each optimized for 5 EM iterations. Then we chose the 5 best parameter values ("parents") and replaced each of them with an average of 3 "offspring" parameter sets to obtain the next "generation." These were then optimized for 5 EM iterations. We repeated this procedure for 4 more "generations," and reported the parameters that achieved the overall maximal likelihood value. We found that the results are robust to the choice of composite likelihood scheme.

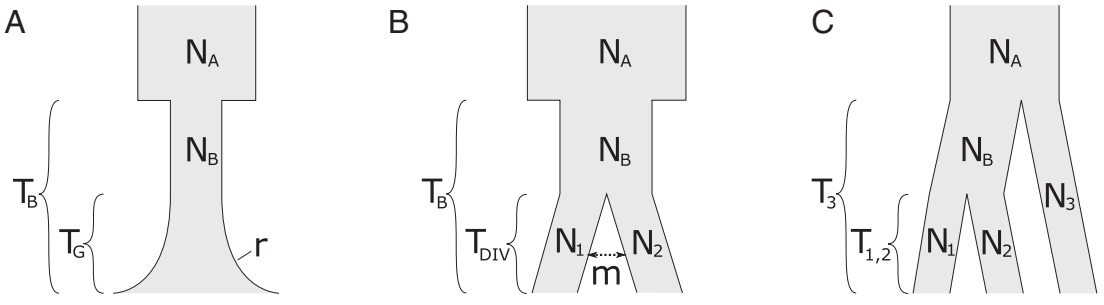

Fig. 1. Demographic models used in our simulation study. (*A*) Recent exponential population growth. An ancestral population of size N_A undergoes a bottleneck at time T_B, where its size is reduced to N_B. Growth starts at time T_G at an exponential rate r. (*B*) Demographic model of a population split. An ancestral population of size N_A undergoes a strong bottleneck that starts at time T_B in the past, and reduces the population size to N_B. At time T_{DIV}, this population then splits into 2 populations of size N_1 and N_2. Following the population split, migrants are exchanged at a rate m. (*C*) Demographic model of 3 populations with pure splits. An ancestral population of size N_A splits into 2 populations of size N_B and N_3 at time T_3. The former then again splits into 2 populations of size N_1 and N_2 at time $T_{1,2}$.

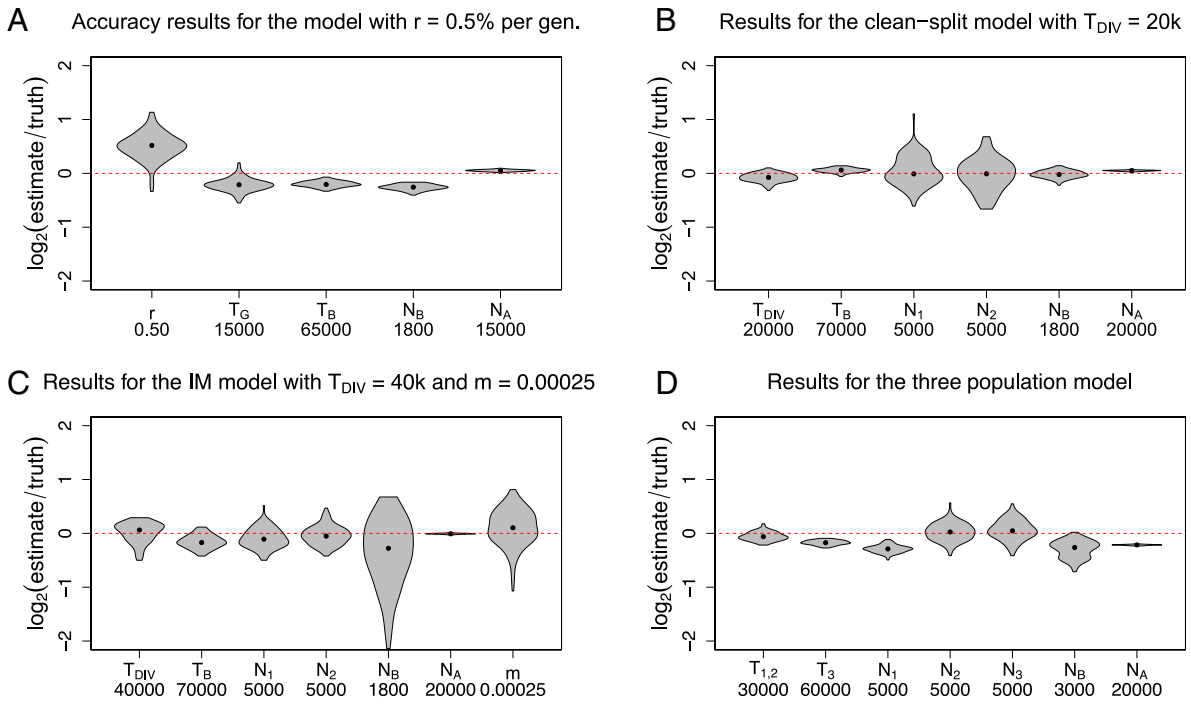

Fig. 2. Accuracy results of our method, diCal2. Each violin plot shows the base-2 logarithm of the relative error (estimate/truth) for the analysis of 100 simulated datasets. Thus, a value of 0 corresponds to an exact estimate, whereas +1 is a 2-fold overestimate and −1 is a 2-fold underestimate. True parameter values are shown on the *x* axis. (*A*) The recent exponential growth model shown in Fig. 1*A* with expansion rate $r = 0.5\%$ per generation. Parameter estimates were obtained using only 10 haplotypes, which is much less than the sample size (thousands to tens of thousands) required by SFS-based methods to get good estimates. (*B*) Accuracy results for the clean-split model (no gene flow, $m = 0$) shown in Fig. 1*B* with divergence time $T_{\mathrm{DIV}} = 20$ ka. Using only 2 haplotypes in each extant population, the parameters of this clean-split model could be estimated very accurately. (*C*) Accuracy results for the isolation with migration (IM) model shown in Fig. 1*B* with divergence time $T_{\mathrm{DIV}} = 40$ ka, and migration probability $m = 0.00025$. As in the clean-split case, only 2 haplotypes in each extant population were used. Most parameter estimates show little bias or variability. See the text for further discussion. (*D*) Accuracy results for the 3-population model shown in Fig. 1*C*, using 2 haplotypes in each extant population.

Violin plots representing the accuracy of the inferences are shown in Fig. 2*A* and *SI Appendix*, Fig. S1. Analysis of the simulated data shows that, in these scenarios, all parameters are estimated with little variability. However, the results indicate that the estimate of the exponential growth rate is biased upward. This bias is somewhat counterbalanced by a slight downward bias of the time when growth starts, and the population size before growth starts. In fact, the estimates lead to very accurate contemporary population sizes. We note that it is possible to empirically correct for biases in applications via simulation. Furthermore, using more sequence data for each individual reduces the variability of the estimates. We stress that our method accurately estimates recent exponential growth rates using only 10 haplotypes. This is far less than the sample size (thousands to tens of thousands) required by SFS-based methods to get reasonable estimates; see Bhaskar et al. (5) and references therein.

Note that, in these and the following simulations, the ancestral population size N_A is estimated with less variability than the other parameters. This is likely due to the fact that, in these scenarios, a sizable fraction of the informative genealogical events happen during the last epoch, and thus the power to estimate the single parameter for the population size during this last epoch is high. This power is not affected strongly by the different sample sizes used in the different scenarios.

Population Split. We also investigated a model of a past population split, depicted in Fig. 1*B*. This model allows for a bottleneck before the populations split, and subsequent gene flow following the split. We first focused on the case with no gene flow, i.e., with migration probability $m = 0$.

We simulated datasets with 2 haplotypes in each of the extant populations. We simulated 100 datasets each for $T_{\mathrm{DIV}} = 10$ ka and 20 ka, with the remaining parameters set to $T_{\mathrm{B}} = 70$ ka, $N_{\mathrm{A}} = 20,000$, $N_{\mathrm{B}} = 1,800$, and $N_1 = N_2 = 5,000$. This scenario has recently been used in a study of the demographic history of Native Americans (22). In addition, we simulated 100 datasets with $T_{\mathrm{DIV}} = 70$ ka, setting $N_{\mathrm{B}} = N_{\mathrm{A}} = 20,000$, thereby also removing the need for T_{B}. For the genetic algorithm, we used 60 random starting points, and 6 "parents" for each of the following 4 "generations" for the cases $T_{\mathrm{DIV}} = 10$ ka and 20 ka, and 40 starting points and 5 "parents" for the case $T_{\mathrm{DIV}} = 70$ ka. We used the LCL.

Fig. 2*B* and *SI Appendix*, Fig. S2 show the accuracy of the estimator. These empirical results demonstrate that our method is able to estimate the parameters in this clean-split model with high accuracy. Most parameters show little bias, and the empirical distributions are very narrow. Only the estimates of the extant population sizes N_1 and N_2 for $T_{\mathrm{DIV}} = 10$ ka and $T_{\mathrm{DIV}} = 20$ ka show a somewhat higher variability. Since this time frame is very recent on an evolutionary timescale, either more sampled haplotypes or more sequence data are required to better estimate these parameters.

Isolation with Migration. We also investigated the demographic model shown in Fig. 1*B* allowing for positive gene flow after the ancestral population splits into 2. We set the migration probability to $m = 0.00025$; i.e., an individual from population 1 can have a parent from population 2, and vice versa, with a probability of 0.00025 per generation. Using this migration probability, we simulated 100 datasets each consisting of 2

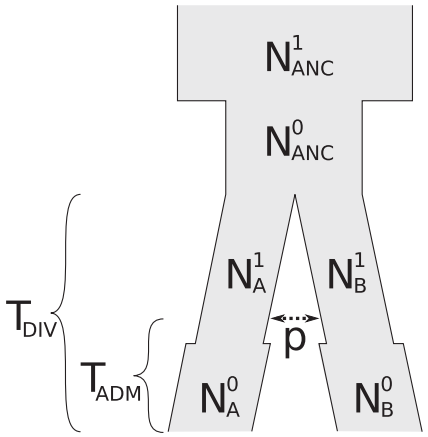

Fig. 3. The demographic model used for the analysis of the French, Han, Papuan, and Australian population from the SGDP dataset. The ancestral population has 2 periods of constant size, then splits into 2, and each of the extant populations has again 2 periods of constant size. Additionally, there is a symmetric pulse admixture event at T_{ADM}, replacing $p\%$ of the ancestors in each population.

haplotypes in each extant population, using $T_{\mathrm{DIV}} = 40$ ka, $T_{\mathrm{B}} = 70$ ka, $N_{\mathrm{A}} = 20,000$, $N_{\mathrm{B}} = 1,800$, and $N_1 = N_2 = 5,000$. We also simulated 100 datasets using $T_{\mathrm{DIV}} = 70$ ka, $N_{\mathrm{B}} = N_{\mathrm{A}} = 20,000$, and $N_1 = N_2 = 5,000$. In the former case, we used 70 starting points and 6 "parents" for each "generation" in the genetic algorithm, whereas, for the latter, we used 50 and 5, respectively.

Fig. 2*C* and *SI Appendix*, Fig. S3 show the accuracy of the estimator. In both scenarios, we used the pairwise composite likelihood. Again, most parameter estimates show little bias or variability, the exceptions being N_B and m in the first scenario. However, we note that the evolutionary timescales involved are, again, rather short, and thus the number of events informative about these parameters is small. In practice, the variability could be reduced by using additional chromosomes.

Three-Population Model. Our method can handle models with more than 2 extant populations each with several haplotypes. To test the accuracy of our method in this setting, we simulated data under the model depicted in Fig. 1*C* relating 3 extant populations. Under this model, an ancestral population of size N_A splits into 2 populations of size N_B and N_3 at time T_3. The one of size N_B then splits into 2 populations of size N_1 and N_2 at time $T_{1,2}$. We simulated 100 datasets with 2 haplotypes in each of the extant populations. We set the parameters to $T_{1,2} = 30$ ka, $T_3 = 60$ ka, $N_A = 20,000$, $N_B = 3,000$, and $N_1 = N_2 = N_3 = 5,000$. For the genetic algorithm, we chose 70 starting points, and 6 "parents," and used the LCL. Fig. 2*D* shows the accuracy of our method. Again, the empirical distribution of the estimates shows little bias or variability.

Application to SGDP Data. We used our method to investigate the pattern of population splits between Australians, East Asians, Europeans, and Papuans. There has been some debate about the relative ordering of population splits; specifically, there has been competing evidence about whether East Asians and Europeans split most recently (e.g., ref. 29) or whether Australo-Papuans and East Asians split most recently (e.g., ref. 30). To date these splits, we used Australian, French, Han, and Papuan individuals from the SGDP (23) and fit models for each of the 6 possible pairs of these populations, allowing for recent population size changes and pulse admixture. The model is depicted in Fig. 3, and additional details are given in *Materials and Methods*. The estimates of the divergence time T_{DIV} and admixture fraction p together with confidence intervals obtained using a parametric bootstrapping approach are presented in Table 1. We found compelling evidence that Australo-Papuans and Eurasians diverged first, about 100 ka, with subsequent French–Han divergence at 53.6 ka, and the Australian–Papuan divergence at 33.9 ka. Note that, while these estimates of divergence times are largely consistent with a tree, some estimates appear to imply slightly different split times (for example, the Australian–Han divergence time is about 15 ka earlier than the Australian–French divergence time); this is likely due to model misspecification resulting from an overly simplistic model.

We also found evidence of pervasive recent gene flow. In particular, we found pulse admixture proportions of 15 to 26% between each pair of populations, all occurring 5 to 20 ka. We note that our model cannot capture all of the intricacies of human demographic history: There has likely been continuous gene flow between all populations punctuated by a few mass migrations. Our gene flow estimates likely attempt to capture both of these modes simultaneously along with indirect gene flow through intervening populations. While it is unlikely that about a quarter of any population was replaced by a geographically distant population, our results suggest that, since their divergences tens of thousands of years ago, these populations have exchanged a considerable number of migrants.

Additional details about the data, data processing, and parameter settings for our method are presented in *Materials and Methods*. All parameter estimates, bootstrap results, and measures of goodness-of-fit evaluated using cross-coalescence rate curves (CCRs) may be found in *SI Appendix*, Fig. S4 and Table S1.

There is also a separate debate about the number of out-of-Africa events (23, 29, 32), with some studies suggesting that a second, earlier wave left traces of ancestry specifically in Australo-Papuans. We caution against interpreting our results in this context, since directly testing this hypothesis would require explicitly including African populations in the analysis. Moreover, our estimates of T_{DIV} are not directly comparable to divergence time estimates given in refs. 23, 29, and 32; as the CCRs in *SI Appendix*, Fig. S4 and similar curves in refs. 23 and 29 show, the populations involved already exhibited a substantial degree of structure 100 ka. Representing this complex population history by a simple model with a single pulse admixture event after splitting as in our analysis, or by a single estimate for the divergence time as in refs. 23 and 29, is certainly an oversimplification that omits relevant details. Lastly, we do not explicitly model the excess traces of Denisovan ancestry that are found in Papuans (33), which may cause differences in the estimated divergence times.

Discussion

The results described above demonstrate that our method can efficiently and accurately estimate demographic parameters in biologically relevant scenarios. Our method has recently been used to study the history of Native American peoples (22, 24, 25),

Table 1. Estimates of the divergence time T_{DIV} (in kiloyears before present) and the admixture percentage p (in percent) for the respective pair of populations from the SGDP dataset

	Han	Papuan	Australian
French	54 ka [52,55] 14.8% [14.4,15.2]	106 ka [104,108] 23.5% [23.0,24.0]	106 ka [105,108] 25.0% [24.3,25.7]
Han		113 ka [110,115] 26.3% [25.0,27.7]	91 ka [91,91] 24.5% [23.7,25.4]
Papuan			34 ka [33,35] 15.4% [14.6,16.1]

The 95% confidence intervals obtained from the parametric bootstrap procedure are shown in square brackets.

due to the flexible framework underlying the method, enabling the consideration of a wide range of population histories.

A limitation of our method is that it relies on the haplotype structure of the sample, and thus requires phased data. Phasing errors can lead to biased parameter estimation; see supplementary information, section S7, specifically table S13, of ref. 22 for a simulation study that explores the bias when estimating divergence times. Moreover, note that the simulations presented above were performed using homogeneous recombination and mutation rates along the genome. The EM procedure underlying our method could be modified to accommodate heterogeneous rates without severely impacting the runtime. If the correct rates are specified, we do not expect the accuracy of inference to be adversely affected. If the correct rates are not known, it would also be possible to adjust the method for joint inference, or to use rates obtained from alternative approaches (34).

Aside from demographic inference, we note that our method can be used in other population genetic problems of interest, such as model selection (see, for example, supplementary information 18.4 of ref. 24). Furthermore, the posterior decoding of latent variables in our CSD can be used in detecting admixture tracts (35), estimating fine-scale recombination rates in admixed individuals, distinguishing ancestral and introgressed polymorphism, and detecting incomplete lineage sorting. Also, applying our CSD in methods for phasing genotypes, imputing missing sequence data, and detecting identity-by-descent tracts (36) would make it possible to properly account for demography or infer it simultaneously, thus potentially improving accuracy. Lastly, it is straightforward to incorporate temporal samples (ancient DNA sequences) into our method (24), which leads to further interesting applications.

Materials and Methods

Here, we briefly describe our method, a composite likelihood framework to estimate demographic parameters using EM. Further details are provided in *SI Appendix*, *SI Text*. We also describe our analysis of the SGDP data.

Demographic Inference Using diCal2. A central building block of our method diCal2 for demographic inference is the CSD $\pi_\Theta(h|\alpha, \mathbf{n})$. It denotes the probability of observing the sequence or haplotype h in subpopulation α, given that the haplotypes $\mathbf{n}$ have already been observed in their respective subpopulations and the underlying demography is described by the parameters Θ. The CSD can be described using a sequentially Markovian genealogical process (37, 38) that approximates the true conditional genealogical process. Subsequent approximations to this sequential process lead to an HMM with finite hidden state space that can be used to efficiently compute approximate CSDs. We provide the details of the HMM approximations in *SI Appendix*, section 1. The CSDs presented in Steinrücken et al. (26) and Sheehan et al. (16) can be obtained as special cases of the model presented here.

The CSD can then be used to define composite likelihood functions, which, in turn, enable us to perform maximum composite likelihood inference of the demographic parameters, Θ. We can use any such composite likelihood function that is composed of sums and products of CSDs, for example, the PAC framework which has been used successfully by Li and Stephens (27) to infer recombination hotspots.

To find the parameter values that maximize this composite likelihood, we employ the composite likelihood in the standard EM framework (39). While, in principle, all parameters of the model can be inferred, we focus on the demographic parameters, Θ. Since it is not possible to derive a closed form solution for the maximum in the maximization step in general, we employ numerical optimization schemes, like the Nelder–Mead simplex algorithm (40), to efficiently determine the requisite maximum. We provide mathematical details for the implementation of the EM algorithm in *SI Appendix*, section 3.

In *SI Appendix*, section 4, we provide details on the implementation of the "locus-skipping" algorithm, and the alternative speedup that groups loci into larger blocks. Furthermore, in *SI Appendix*, section 5, we provide mathematical details of the modifications to the trunk genealogy to increase accuracy. Finally, we describe, in *SI Appendix*, section 6, how to employ a discretization for the HMM computations that differs from the partition induced by the demography and remains fixed throughout the optimization procedure.

Runtime. The runtime of our implementation of the EM algorithm is linear in the number of haplotypes times the number of CSDs in the composite likelihood and quadratic in the number of populations involved. The E step depends linearly on the length of the haplotypes, whereas the M step is independent of this quantity. The exact complexity and runtime of parameter estimation depends on the composite likelihood used, the details of the genetic algorithm, and the number of parameters to estimate. The analyses of the simulated data presented in this section were performed on a cluster of AMD Opteron processors. The raw sequential runtime of analyzing a single dataset averaged 100 to 120 CPU hours, but, by taking advantage of the independence structure of the composite likelihood and the genetic algorithm, we were able to decrease the runtime to an average of 15 to 20 wall clock hours, using up to 16 cores in parallel. The one exception was the 3-population model, where the parallelized version took, on average, 70 wall clock hours, due to the more complex demographic model and the increased number of haplotypes.

SGDP Analysis. For the analysis of the SGDP data, we used the following individuals: B_Australian-3, B_Australian-4, S_French-1, S_French-2, B_French-3, S_Han-1, S_Han-2, B_Han-3, S_Papuan-1, S_Papuan-3, and B_Papuan-15. The data were phased using Shapeit (41) with read-based phasing (phased data provided by I. Mathieson, Department of Genetics, University of Pennsylvania, Philadelphia, PA), and all sites in Heng Li's 75-bp universal mask (23) were treated as missing. As in our simulations, we used a mutation rate and recombination rate of 1.25×10^{-8} per base per generation, and assumed a generation time of 30 y. For each pair of populations, we used all of the individuals from those populations and used all of the autosomes to fit a model where each population has a constant size from present to 5 ka, and another constant size from 5 ka until the divergence time of the populations. The ancestral population is assumed to be a constant size from the divergence time until 100 ka, beyond which we infer a separate constant size. We allow a symmetric pulse migration between the 2 populations. To fit this model, we performed 4 iterations of the genetic algorithm, starting from 15 arbitrary points, keeping the 3 best particles at each iteration, and then spawning a total of 10 particles. Each genetic algorithm iteration consisted of 6 EM iterations for each particle, using the LCL. For computational efficiency, we grouped loci into 2.5-kbp bins (*SI Appendix*, section 4), and discretized time with 8 log-uniformly spaced break points between 1.5 ka and 5 Ma. Each genome-wide analysis of a pair of populations with combined sample size 10 to 12 took $\sim$90 to 145 wall clock hours on AMD Opteron processors, using up to 10 cores in parallel and less then 30 GB of memory.

As seen in the simulations, the raw inferred parameters may be biased. To address this issue and infer confidence intervals, we performed a parametric bootstrap using msprime (42). For each pair of populations, we simulated 10 full genome datasets, and reran our method on each of these datasets. Our reported estimates are "debiased" estimates, obtained by subtracting the estimated bias from our raw estimates. We then used the bootstraps to estimate a SD for each parameter, and reported confidence intervals based on a normal approximation (i.e., the debiased estimate ± 1.96 SD). To avoid having these debiased estimates or confidence intervals fall outside of the domain of the parameters (e.g., negative population sizes, times, or pulse proportions or pulse proportions $> 100\%$), all debiasing and confidence intervals were computed in log space for population sizes and times and in logit space for pulse proportions. The resulting estimates and confidence intervals were then transformed back to their natural space using the exponential map and logistic map, respectively. We note that this procedure means that our estimates are unbiased in log space or logit space, and may be slightly biased in their natural scale. All parameter estimates and bootstrap results are presented in *SI Appendix*, Table S1. We also note that the bootstrap results were obtained using the same grouping of loci (2.5-kbp bins), and thus show that this procedure does not severely impact accuracy.

To assess goodness of fit, we used MSMC to infer CCRs on the real data, and then on data simulated under our debiased estimates, presented in *SI Appendix*, Fig. S4. For each pair of populations, we used a single diploid from each population (B_Australian-3, S_French-1, S_Han-1, and S_Papuan-1), using all of the autosomes and again treating all sites in Heng Li's 75-bp universal mask as missing. We simulated 5 replicates of each pair of populations to assess the variability in the `MSMC` CCRs. The CCRs are qualitatively similar between the real and simulated data, and the fit is quite good for

a model with only 9 parameters. As discussed above, introducing additional size changes and migration rates would likely improve the fit.

Software Availability. The algorithms described here are implemented in a new version of the software package diCal2, which is available for download at https://sourceforge.net/projects/dical2.

ACKNOWLEDGMENTS. We thank Sara Mathieson and Geno Guerra for helpful discussions and for testing our software. Furthermore, we thank Iain Mathieson for helpful discussions and providing the phased SGDP data. This research is supported, in part, by National Institutes of Health Grant R01-GM094402 and a Packard Fellowship for Science and Engineering. Y.S.S. is a Chan Zuckerberg Biohub Investigator.

1. R. Nielsen, Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942 (2000).
2. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
3. S. Lukić, J. Hey, Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* **192**, 619–639 (2012).
4. L. Excoffier, I. Dupanloup, E. Huerta-Sanchez, V. Sousa, M. Foll, Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
5. A. Bhaskar, Y. R. Wang, Y. S. Song, Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Res.* **25**, 268–279 (2015).
6. J. A. Kamm, J. Terhorst, Y. S. Song, Efficient computation of the joint sample frequency spectra for multiple populations. *J. Comput. Graph. Stat.* **26**, 182–194 (2017).
7. J. Jouganous, W. Long, A. P. Ragsdale, S. Gravel, Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics* **206**, 1549–1567 (2017).
8. J. Kamm, J. Terhorst, R. Durbin, Y. S. Song, Efficiently inferring the demographic history of many populations with allele count data. *J. Am. Stat. Assoc.*, 10.1080/01621459.2019.1635482 (2019).
9. J. Terhorst, Y. S. Song, Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7677–7682 (2015).
10. Bhaskar A, Song YS, Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann. Stat.* **42**, 2469–2493 (2014).
11. J. Kim, E. Mossel, M. Z. Rácz, N. Ross, Can one hear the shape of a population history? *Theor. Popul. Biol.* **100**, 26–38 (2015).
12. Y. Kim, F. Koehler, A. Moitra, E. Mossel, G. Ramnarayan, "How many subpopulations is too many? Exponential lower bounds for inferring population histories" in *Research in Computational Molecular Biology. RECOMB 2019 Research in Computational Molecular Biology. RECOMB 2019*, L. Cowen, Ed. (Lecture Notes in Computer Science, Springer, 2019), vol. **11467**, pp. 136–157.
13. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
14. T. Mailund, J. Y. Dutheil, A. Hobolth, G. Lunter, M. H. Schierup, Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genet.* **7**, e1001319 (2011).
15. P. F. Palamara, T. Lencz, A. Darvasi, I. Pe'er, Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91**, 809–822 (2012).
16. S. Sheehan, K. Harris, Y. S. Song, Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics* **194**, 647–662 (2013).
17. K. Harris, R. Nielsen, Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.* **9**, e1003521 (2013).
18. S. Schiffels, R. Durbin, Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
19. K. Wang, I. Mathieson, J. O'Connell, S. Schiffels, Tracking human population structure through time from whole genome sequences. bioRxiv:10.1101/585265 (21 March 2019).
20. J. Terhorst, J. A. Kamm, Y. S. Song, Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
21. J. P. Spence, M. Steinrücken, J. Terhorst, Y. S. Song, Inference of population history using coalescent HMMs: Review and outlook. *Curr. Opin. Genet. Dev.* **53**, 70–76 (2018).
22. M. Raghavan *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
23. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
24. J. V. Moreno-Mayar *et al.*, Terminal Pleistocene Alaskan genome reveals first founding population of native Americans. *Nature* **553**, 203–207 (2018).
25. J. V. Moreno-Mayar *et al.*, Early human dispersals within the Americas. *Science* **362**, aav2621 (2018).
26. M. Steinrücken, J. S. Paul, Y. S. Song, A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theor. Popul. Biol.* **87**, 51–61 (2013).
27. N. Li, M. Stephens, Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* **165**, 2213–2233 (2003).
28. J. S. Paul, Y. S. Song, Blockwise HMM computation for large-scale population genomic inference. *Bioinformatics* **28**, 2008–2015 (2012).
29. A. S. Malaspinas *et al.*, A genomic history of Aboriginal Australia. *Nature* **538**, 207–214 (2016).
30. J. D. Wall, Inferring human demographic histories of non-African populations from patterns of allele sharing. *Am. J. Hum. Genet.* **100**, 766–772 (2017).
31. P. R. Staab, S. Zhu, D. Metzler, G. Lunter, scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* **31**, 1680–1682 (2015).
32. L. Pagani *et al.*, Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**, 238–242 (2016).
33. S. Sankararaman *et al.*, The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr. Biol.* **26**, 1241–1247 (2016).
34. J. P. Spence, Y. S. Song, Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations, bioRxiv:10.1101/532168 (28 January 2019).
35. M. Steinrücken, J. P. Spence, J. A. Kamm, E. Wieczorek, Y. S. Song, Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Mol. Ecol.* **27**, 3873–3888 (2018).
36. P. Tataru, J. Nirody, Y. S. Song, diCal-IBD: Demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics* **30**, 3430–3431 (2014).
37. C. Wiuf, J. Hein, Recombination as a point process along sequences. *Theor. Pop. Biol.* **55**, 248–259 (1999).
38. G. A. McVean, N. J. Cardin, Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 1387–1393 (2005).
39. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B Met.* **39**, 1–38 (1977).
40. J. A. Nelder, R. Mead, A simplex method for function minimization. *Comput. J.* **7**, 308–313 (1965).
41. O. Delaneau, B. Howie, A. Cox, J. F. Zagury, J. Marchini, Haplotype estimation using sequence reads. *Am. J. Hum. Genet.* **93**, 787–696 (2013).
42. J. Kelleher, A. M. Etheridge, G. McVean, Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, 1–22 (2016).