Inferring the ancestry of parents and grandparents from genetic data

Jingwen Pei[1], Rasmus Nielsen[2,3,*] and Yufeng Wu[1,*]

[1]Department of Computer Science and Engineering

University of Connecticut

Storrs, CT 06269, U.S.A.

[2]Departments of Integrative Biology and Statistics

University of California, Berkeley

Berkeley, CA 94720, U.S.A.

[3]Museum of Natural History

University of Copenhagen

Copenhagen K, 1350, Denmark.

Running title: Inferring ancestry of parents and grandparents

Key words: population genetics, population admixture, ancestry inference, pedigree, algorithm.

*Corresponding author: `rasmus_nielsen@berkeley.edu` and `yufeng.wu@uconn.edu`.

ABSTRACT

Inference of admixture proportions is a classical statistical problem in population genetics. Standard methods implicitly assume that both parents of an individual have the same admixture fraction. However, this is rarely the case in real data. In this paper, we show that the distribution of admixture tract lengths in a genome contains information about the admixture proportions of the ancestors of an individual. We develop a Hidden Markov Model (HMM) framework for estimating the admixture proportions of the immediate ancestors of an individual, i.e., a type of splitting of an individual's admixture proportions into further subsets of ancestral proportions in the ancestors. Based on a genealogical model for admixture tracts, we develop an efficient algorithm for computing the sampling probability of the genome from a single individual as a function of the admixture proportions of the ancestors of this individual. This allows us to perform probabilistic inference of admixture proportions of ancestors, using only the genome of an extant individual. We perform extensive simulations to quantify the error in the estimation of ancestral admixture proportions under various conditions. As an illustration, we also apply the method on real data from the 1000 Genomes Project.

INTRODUCTION

Ancestry inference is one of the most commonly used tools in human genetics. It arguably provides the most popular information from commercial genotyping companies such as *Ancestry.com* and *23andMe* to millions of customers. It also forms the basis of many standard population genetic analyses with most population genomic publications including ancestry inference analyses in one form or another (e.g., Rosenberg, et al. (2002); Li, et al. (2008)). Modern ancestry inference has roots in the seminal paper on STRUCTURE (Pritchard, et al., 2000). The model introduced in that paper assumes each individual can trace its ancestry, fractionally, to a number of discrete populations. For each individual, independence is assumed between the two alleles at a locus. The ancestry for each allele is then described by a mixture model in which the allele is assumed to be sampled from each of the ancestral populations with probability equal to the *admixture proportion* of this ancestral population. Many subsequent methods are based on the same model including FRAPPE (Tang, et al., 2005) and ADMIXTURE (Alexander, et al., 2009). Notice that this model implicitly assumes that the admixture proportions for each parent of an individual are the same. This assumption is arguably unrealistic for many human populations. In fact, for recently admixed populations, we would expect the admixture proportions to differ between the parents. However, the commonly used methods for admixture inference do not allow estimation of ancestry components separately for the two parents. We note that there is substantial information in genotypic data on parental admixture proportions. Even without linkage information, the genotypes can be used to infer parental ancestry. For example, consider the extreme case of a locus with two alleles, T and t at a frequency of 1 and 0, respectively, in ancestral population A, and a frequency of 0 and 1, respectively, in ancestral population B (i.e., a fixed difference between two populations). Then, the sampling probability of an offspring of genotype Tt, resulting from matings between individuals from populations A and B, is equal to one. However, if the two parents are both 50:50 (%) admixed between populations A and B, the probability, in the offspring, of genotype Tt is 0.5. In both cases, the average admixture proportion of the offspring individual is 0.5. This is an extreme example, but it clearly illustrates that the offspring genotype distributions contain information regarding the parental genotypes that can be used to infer admixture proportions in the parents.

Recently, a method was developed for inferring admixture proportions, and admixture tracts, in

the two parents separately from phased offspring genotype data (Zou, et al., 2015). As the length distribution of admixture tracts is well-known not to follow the exponential prediction of a Markov process (Gravel, 2012), this method models the ancestry process along each of the chromosomes as a semi-Markov process. It uses inference methods based on Markov Chain Monte Carlo (MCMC), Stochastic Expectation Maximization (EM), and a faster non-stochastic method for the case of a Markovian approximation to the ancestry process, and shows that parental ancestry can be estimated with reasonable accuracy.

The objective of this paper is to explore the possibility of not only estimating admixture proportions in parents, but in grandparents, or even great grandparents. A common assumption is that each genetic variation is assumed to be independent. In this case, the marginal genotype probabilities provide no information that would allow us to distinguish between different admixture proportions in the grandparents compatible with the same parental genotype distribution. However, the distribution of tract lengths does provide such information. By modeling the segregation of admixture tracts inside a pedigree, we obtain a likelihood function that can be used to estimate admixture proportions in grandparents and great grandparents, in addition to parents.

RESULTS

We now evaluate the performance of our method, implemented in a program called PedMix. We show results on simulated, semi-simulated, and real data.

Results on simulated data

Simulation settings and evaluation

We conduct extensive simulations to evaluate the performance of our method. We first simulate a number of haplotypes using macs (Chen, et al., 2009) from two ancestral populations which diverged from one ancestral population at $4N_e t$ generations in the past. Here N_e is the effective population size. An admixed population is then formed by merging the two ancestral populations and simulating the process of random mating, genetic drift, and recombination using a diploid Wright-Fisher model for g additional generations. We model recombination rate variation using the local recombination estimates from the 1000 Genomes Project (1000 Genomes Project Consortium,

2015). The hotspot maps of the 22 human autosomal chromosomes are concatenated for a single string of $3 \times 10^9 bp$ and subsequently simulated genomes are divided into 22 chromosomes of equal length to facilitate clearly interpretable explorations of the relationship between accuracy and the amount of data. Haplotypes are paired into genotypes and phasing errors are then added stochastically by placing them on the chromosome according to a Poisson process with rate p_p. By default, no phasing error is included in the simulations. The parameters we use in the simulations are listed and explained in Table 1 together with their default values. For the default setting, the approximate total number of single nucleotide polymorphisms (SNPs) simulated by macs is $\sim 14.7M$. Here we apply frequency-based pruning to trim data (see the Methods Section). Frequency-based pruning removes SNPs with a minor allele frequency difference in two ancestral populations less than the pruning threshold d_f. After pruning with the default d_f, each of the 22 chromosomes contains $\sim 26,000$ SNPs. In some cases, the default simulated length, $L = 3 \times 10^9 bp$, can result in a high computational workload. Therefore, in some simulations, we also use a shorter length of $L = 5 \times 10^8 bp$, divided into 3 chromosomes. If not otherwise stated, we use $L = 3 \times 10^9 bp$ to be the default setting.

PedMix is applied to the simulated genotype data from the admixed population for inference of admixture proportions of ancestors in the 1st generation (parents), the 2nd generation (grandparents), and so on. To evaluate accuracy, we use the mean absolute error (MAE) between the estimated admixture proportion, $\widehat{m^i}$, and the true admixture proportion, m^i, for the ith ancestor in the Kth generation, as the metric of estimation error. If there are multiple individuals, we further take the average over all individuals, i.e., the mean error for n individuals as defined by Equation 1. Without loss of generality, we only consider the estimate of the proportions of ancestral population A. Since we assume two ancestral populations, the expected mean errors of admixture proportions for two ancestral populations are identical.

$$\text{Mean error } = \frac{1}{n \cdot 2^K}\Big(\sum_{1 \le j \le n} \sum_{1 \le i \le 2^K} |\widehat{m_j^i} - m_j^i|\Big) \tag{1}$$

As the admixture proportions inferred by the method are unlabeled with respect to individuals, this leads to ambiguity on how to match the inferred proportions to the true proportions. We address this problem using a "best-match" procedure by rotating the parents for each internal

node in a pedigree to find the best match between the inferred and the simulated admixture proportions. For example, for inference in parents, denote the true admixture proportions for two parents as (m^1, m^2) and the estimated admixture proportions as $(\widehat{m^1}, \widehat{m^2})$. Both $(\widehat{m^1}, \widehat{m^2})$ and $(\widehat{m^2}, \widehat{m^1})$ are then matched with (m^1, m^2), and the one with smaller mean error is chosen. For the case of grandparent inference, there are eight possible matchings and we explore all eight to obtain the best match.

Evaluation of ancestral inference accuracy

Figure 1 shows the mean error when inferring admixture proportions of parents, grandparents, and great grandparents under the default simulation settings (as shown in Table 1). The performance of PedMix is compared to what is expected from a random guess based on a Bayesian model. The random guess is described in Supplemental Methods. Ten genotypes are sampled from simulated admixed population for g (with $g \geq 3$) generations.

Inference of great grandparents admixture proportions is computationally demanding in the current framework. Therefore, we use a more extreme trimming threshold, $d_f = 0.9$, when inferring great grandparent admixture, resulting in only 26,638 SNPs.

As expected, it is easier to estimate admixture proportions of more recent ancestors. This is because, as we trace the ancestry of a single individual back in time, the genome of the extant individual contains progressively less information about an ancestor.

Comparison of PedMix to existing methods

Although there are no existing methods for inferring the admixture proportions of grandparents and great grandparents that we can compare PedMix to, there is a method called ANCESTOR (Zou, et al., 2015) that infers admixture proportions of parental genomic ancestries given ancestry of a focal individual. There are many methods (e.g., ADMIXTURE (Alexander, et al., 2009) and RFmix (Maples, et al., 2013)) for inferring admixture proportions of individuals of the current generation. In this section, we first compare estimates of the admixture proportions of a focal individual obtained from ADMIXTURE and RFmix, arguably the state-of-the-art methods for ancestry inference, to the average of parental or grandparental admixture proportions inferred using PedMix. Here, we use the average of the estimated admixture proportions from ancestors as the

proxy for the admixture proportion of the focal individual. If the admixture proportions of ancestors inferred by PedMix are accurate, the average of these admixture proportions of ancestors should be able to serve as a good approximation for the focal individual. Furthermore this average is expected to be approximately as accurate as the admixture proportions inferred by RFmix and ADMIXTURE. This is verified with simulation data (Supplemental Methods and Supplemental Table S1). It should be noted that the high accuracy of PedMix in inferring the admixture proportion of a focal individual from the average of parental or grandparental proportions does not necessarily imply that the parental and grandparental admixture proportions themselves are accurately inferred. However, if the admixture proportion of a focal individual is poorly estimated from the inferred admixture proportions of the ancestors, this may suggest that the admixture proportion estimates for the ancestors also are not accurate.

20 individuals are randomly sampled from an admixed population and run ADMIXTURE, RFmix, and PedMix on the same datasets. The genotypes are preprocessed with LD pruning (Supplemental Methods) and contain phasing errors simulated with rate $p_p = 0.00002$ per bp. The ancestry of each individual is deduced using PedMix from the inferred admixture proportions of either parents or grandparents, by using the average of the inferred admixture proportions of the ancestors. ADMIXTURE and RFmix infer the admixture proportions of extant individuals directly. Table 2 shows the mean error as defined in Equation 1 and the error's standard deviation. Our results show that the admixture proportions inferred from the average of ancestral admixture proportions in PedMix are comparable to those of RFmix and ADMIXTURE. The estimate of parents matches the results of RFmix and ADMIXTURE.

We further compare the estimates of admixture proportions of parents from PedMix to those from ANCESTOR. Note that ANCESTOR requires the ancestry states with tract length. That is, ANCESTOR needs to know the ancestral population for each position. Here we use the inferred ancestry by RFmix when running ANCESTOR. Mean error is computed between the true admixture proportions of parents and the estimates from ANCESTOR or PedMix (Table 3). Estimates by PedMix are more accurate than ANCESTOR by about 3% on average.

Details regarding the application of ADMIXTURE, RFmix, and ANCESTOR are given in Supplemental Methods.

Impact of simulation parameters

PedMix is run on different amounts of data, by sub-sampling 5, 10, 15, and 20 chromosomes, to evaluate the effect of data amount on inference accuracy. The mean error is estimated over 10 samples. As shown in Figure 2, there is a clear linear decrease of mean error for parents, grandparents, and great grandparents inference, as more data is added. The highest mean error for five chromosomes is 18.07%, which is still much lower than the random guess (about 35%, Figure 1).

We perform additional simulations to investigate the impact of various simulation parameters on the accuracy of our method. To investigate the effect of mutation rates and recombination rates, the default setting is used with a shorter genome of length $L = 5 \times 10^8$ (Table 1) to reduce the computational time (Figure 3). The expected number of SNPs simulated in a region increases linearly with the mutation rate. This leads to a reduction in the mean error with increased mutation rates, as more informative markers are available for analysis (Figure 3 (A)). However, the reduction is modest because the statistical accuracy is mostly limited by the number of admixture tracts and not by the number of markers. In contrast, recombination rate has a much stronger effect on the accuracy than mutation rate because increased recombination rates introduce more admixture tracts (Figure 3 (B) and Supplemental Fig S2). The mean error for both parental and grandparental inferences decreases and then levels off as recombination rate increases further. When the recombination rate increases to more than 5×10^{-8}, the improvement in accuracy narrows, especially for parental inference. As the length of each tract decreases, the information regarding the ancestry for each tract also decreases. Even with very high recombination rates, there may still be some error determined by the degree of genetic divergence between populations and the number of generations since admixture.

The simulations assume a model of two ancestral populations that diverged $4N_e t$ generations ago and then admixed g generations ago. The performance of the method clearly depends on these parameters. If g is small, the number of admixture tracts is also small, complicating inferences, particularly in the grandparental generation. As g increases from 4 to 10, the mean error reduces from 7.13% to 4.47% for parent inference and from 15.34% to 7.36% for grandparents respectively (Figure 3 (C)). There is also a strong effect of t on the accuracy. As t increases, allele frequency dif-

ferences between the admixing populations increase and it becomes easier to distinguish admixture tracts from two ancestral populations (Figure 3 (D)). When $t > 0.5$ the mean error for parental inferences drops to below 1%.

Data Trimming

To investigate the effect of frequency-based pruning and LD pruning, a small investigation of the relative effect of LD-pruning and frequency-based pruning on the same simulated dataset is performed. To efficiently compare the two trimming strategies, a shorter length genome $L = 5 \times 10^8$ is simulated. The simulation settings are chosen in order to better compare the two ways of trimming and also to ease the computational burden.

There are $\sim 2.44M$ SNPs simulated for the whole genome. Here two cases of data trimming are examined. In both cases, the window size of $W = 10Kbp$ and $c = 0.1$ are used for LD pruning following the procedure described in Supplemental Methods. The frequency-based pruning is conducted following the procedure described in Supplemental Methods. In the first case of LD-pruning, $f = 0.05$ is used to remove rare variants (i.e., SNPs with combined frequency in the two populations being smaller than f), this results in 284K SNPs left. As a comparison, frequency-based pruning is performed with the threshold $d_f = 0.27$ that leads to the similar amount of SNPs (283K) in this case. In the other case, rare variants are removed with $f = 0.2$, resulting in 97K SNPs. As a comparison, frequency-based pruning with $d_f = 0.5$ results in 94K SNPs (Table 4). In both cases, inferences improve as more SNPs are removed. Overall, the frequency-based pruning approach is slightly better than the LD pruning approach, at least in this simulation. Therefore, frequency based trimming is used as the default strategy for data trimming. A more extreme case is examined and results are shown in Supplemental Table S3.

To further investigate the effect of frequency-based pruning, haplotypes from a small region are simulated. The detailed results are shown in Supplemental Methods and Supplemental Fig S1. We investigate different values of the allele frequency threshold d_f. More than 75% of simulated SNPs are being removed with a fairly small $d_f = 0.1$. As we trim with increasing d_f value, it results in a substantial reduction in mean inference error, especially for the inference of admixture proportions in grandparents. However, when the trimming threshold is too large (e.g., $d_f = 0.7$), an increase in the mean error is observed as too much information is now being lost by removing SNPs. The

optimal trimming threshold depends on many parameters, such as mutation rate and ancestral populations' split time. It is difficult to decide the optimal value for d_f when analyzing a specific dataset. As a rule of thumb, it is desirable to have at least 100 SNPs per tract (from one ancestral population) after frequency-based pruning.

Clearly, more work is needed to identify optimal pruning strategies for real data analyses, for the current method, and for other methods that use HMMs for population genetic inferences. However, such investigations are not the main subject of the current paper.

Phasing error

Real data may contain phasing error (i.e., errors of haplotypes at heterozygous sites). We have implemented a preprocessing approach for reducing the phasing error. See the Methods Section for details. To evaluate the effect of phasing error, haplotypes with a phasing error rate of 2×10^{-5} per bp are simulated. We then compare the mean error by PedMix using genotype data without phasing errors, genotype data with phasing errors, and genotype data with phasing errors preprocessed to remove some phasing errors. As for the data generated with a phasing error rate of 2×10^{-5}, PedMix is directly run without preprocessing. Then the technique described in Supplemental Methods is used to preprocess the data.

As shown in Figure 4 (A), phasing error reduction by preprocessing increases the inference accuracy. Note that at the default setting, the phasing error rate is nearly 2,000 times larger than the recombination rate, which can affect the accuracy of the method significantly. Thus, in real data analyses, it is important to reduce phasing error in some way.

We also consider the case of unadmixed individuals. In that case, phasing errors may be interpreted as recombination events during inference, but the overall admixture proportion estimates should be relatively unaffected. To illustrate this point, 10 individuals are sampled from the same population and PedMix is run to infer their admixture proportions. The performance of PedMix is very stable as shown in Figure 4 (B), with an inference error of approximately 3% for parents and grandparents.

Phasing error adds noise to the model, especially in the region where the two haplotypes have different ancestral states. As the effect of the phasing error in the data using preprocessing is reduced, the inference error decreases significantly.

Results on real data

PedMix is run on the data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015). The 1000 Genomes Project recently released phased haplotypes on 22 chromosomes for 1,092 individuals. The haplotypes analyzed in this manuscript are downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/shapeit2_phased_haplotypes/`. Data from the CEU (Utah Residents with Northern and Western European Ancestry), YRI (Yoruba in Ibadan, Nigeria) and ASW (Americans of African Ancestry in SW USA) populations are analyzed. The African-American population tends to have admixed European and African (primarily West-African) ancestry. In this work CEU and YRI are regarded as the two source populations for ASW and infer the admixture proportions of parents and grandparents of ASW individuals. For these three populations, there are 85 CEU individuals (170 haplotypes), 88 YRI individuals (176 haplotypes) and 61 ASW individuals (122 haplotypes) in total in the 1000 Genomes Project data.

We approximate the allele frequencies in the two hypothetical source populations using the average allele frequencies in the CEU and YRI populations. The original data has 1,060,387 SNPs in total. After applying the frequency-based pruning with $d_f = 0.5$, there are 256,122 SNPs (about 24%) left. The recombination fractions are calculated based on the recombination hotspot map of the 1000 Genomes Project (1000 Genomes Project Consortium, 2015). Seven individuals from the CEU, YRI, and ASW populations are sampled and PedMix is applied to infer admixture proportions in their parents and grandparents.

The inference results of admixture proportions for parents and grandparents for the seven 1000 Genomes individuals are shown in Figure 5. The admixture proportions of the CEU individual ancestors are estimated to be 98% of CEU origin on average. Similarly, the admixture proportions of YRI individuals ancestors are 98% of YRI origin on average. The admixture proportions in the African-American ASW population vary considerably among individuals (Figure 5). Note that some ancestors for the CEU and YRI individuals have small (but non-zero) inferred admixture proportions. Since the proportions are very small (within the error margin of our method), whether these ancestors are admixed or not cannot be determined.

To further validate our results, 61 individuals from ASW population are analyzed using ADMIXTURE and RFmix. Genotypes are pruned with the default LD pruning setting (Supplemental

Methods). Meanwhile, genotypes from CEU and YRI populations are provided as the two ancestral populations in both tools. Using PedMix, the average admixture proportions from parents and grandparents are computed to see if the results are consistent with those from ADMIXTURE and RFmix. The percentages of the CEU origin of the five ASW individuals are shown in Table 5. Five individuals from ASW with different level of admixture proportions (from 20% to 90%) are listed. Although the true admixture proportions of these individuals are unknown, the results of PedMix are consistent with those from RFmix and ADMIXTURE. Moreover, the estimates by PedMix are highly correlated with those by ADMIXTURE and RFmix. Pearson correlation coefficients are 0.9954 and 0.9945 respectively (Supplemental Table S2).

Results on semi-simulated data

Results on semi-simulated data are presented next. Here, genotypes of CEU/YRI/ASW populations from the 1000 Genomes Project are used as the founders of a fixed pedigree topology as shown in Figure 6. This way, the genotypes are closer to the real data and the origin of the founders is known. For this pedigree of two generations, four genotypes are selected from one or more populations among CEU, YRI, and ASW populations as grandparents. It is assumed that there is no phasing error along these grandparental genomes. Then we simulate two genotypes as parents and one genotype as the focal individual forward in time based on the standard genetics law. Recombination rate is modeled from the hotspot maps of the 1000 Genomes Project as in the other simulations. To assess the impact of phasing errors, we also create data with phasing errors by adding phasing errors stochastically with the rate $p_p = 0.00002$ per bp for the focal individuals. PedMix is run on the genotypes of the focal individual both with and without phasing errors to infer the admixture proportions of parents and grandparents. RFmix is run to estimate the admixture proportions of parents (respectively grandparents) using the genotypes of parents (respectively grandparents). The estimates from RFmix with the ancestral genotypes are used as the ground truth on the admixture proportions of ancestors. Here six cases with different ancestral origins of the grandparents: CCCY, CCYY, CYCY, AAAA, AAAC, and AACY (where C is for CEU, Y is for YRI, and A is for ASW) are examined. As an example, CCCY stands for the four grandparents from CEU, CEU, CEU, and YRI respectively. Figure 6 shows the estimates by RFmix and PedMix. Mean error is computed from the six inferred admixture proportions (two parents and four grandparents) in the pedigree

and their estimates by RFmix. Although the focal individuals in the pedigrees CCYY and CYCY both have around 50% admixture proportion, PedMix is able to tell the difference in the parents by estimating the parental admixture proportions being 82.41% and 16.43% for CCYY and 47.05% and 43.78% for CYCY. This largely agrees with the true admixture proportions of the parents, which are 100% and 0% for CCYY and 50% and 50% for CYCY. Note that the true admixture proportions for the two parents are known for a pedigree with the known grandparental origin. For example, in the CCYY case the true parental admixture proportions are 100% and 0%. This is because one parent has two CEU grandparents and thus this parent is 100% CEU. Similarly, the other parent is 100% YRI. This indicates that PedMix is able to collect useful information from the admixture tract lengths in the focal individual. Results on genotypes without phasing errors tend to be more accurate than those with phasing errors. These results indicate that phasing errors can indeed lead to larger inference error for some cases. Thus, it is useful to use haplotypes with fewer phasing errors. Estimates for parents are more accurate than those of grandparents.

DISCUSSION

In this paper we developed a method for inference of admixture proportions of recent ancestors such as parents and grandparents. To the best of our knowledge, there are no other methods for inferring admixture proportions for grandparents or great grandparents. The key idea is using the distribution of admixture tracts which is influenced by the ancestral admixture proportions. Admixture tracts capture the important linkage disequilibrium information. Treating SNPs as independent sites is insufficient for inferring ancestral admixture proportions in general (Supplemental Methods). Our method uses a pedigree model, which is a reasonable model for recent genealogical history of a single individual. There exists earlier approaches for inferring ancestry that use additional information. For example, maps or other localization information are used in Yang, et al. (2014); Margalit,et al. (2015). Our approach only uses the genetic data from a single individual. See Supplemental Methods for a brief introduction on how to use PedMix.

A natural question is whether our method can be extended to more distant ancestors. In theory it could, but as the number of generations increases, the amount of information for each ancestor decreases and the computational time increases. This can be seen from Figure 1, where the inference

error for great grandparents is significantly higher than those for parents or grandparents. On the other hand, Figure 1 shows that there is still information obtained from the inference even in the more difficult great grandparent case.

In Table 2, we show that PedMix can be used to infer the admixture proportions of an extant individual by averaging the inferred admixture proportions of ancestors. In comparison between ADMIXTURE and RFmix, we find that the admixture proportions inferred from the average of ancestral admixture proportions in PedMix is comparable to that of RFmix and ADMIXTURE. The key difference between PedMix and RFmix/ADMIXTURE is that PedMix infers the admixture proportions of ancestors while the other methods infer the admixture proportions of the focal individuals. Also note that when we compare PedMix with RFmix and ADMIXTURE, PedMix uses recombination fractions in the founding populations, which are not used by RFmix and ADMIXTURE.

Inference with PedMix is affected by the parameter settings of the underlying population genetic process (Figure 3). The inference error of PedMix can be significantly reduced if the recombination rate is high, admixture is more ancient, or the divergence time between the two source populations is larger. Increasing the chromosome length has a similar effect on inference as increasing the recombination rate.

In line with other studies (e.g., Anderson, et al. (2010)), we find that pruning of SNPs and preprocessing to remove potential phasing errors is critical for obtaining reasonably accurate results. In the Results Section, we compare two trimming strategies, LD pruning and frequency-based pruning. LD pruning is a common strategy used in HMM-based applications for removing background LD that is not modeled by the HMM. However, as low-frequency SNPs are more likely to have small values of r^2, but are less informative for inference, strategies for removing SNPs based solely on measures of LD such as r^2 might not be optimal. In fact, in the limited simulations performed here (Table 4 and Supplemental Fig S1), we find, perhaps surprisingly, that pruning strategies based on removing low-frequency SNPs, rather than SNPs in high LD, lead to the best performance. Based on our experience, we use the frequency-based trimming as our default data trimming approach. The objective of this paper is not to explore SNP pruning strategies for HMMs, but our results suggest that existing methods could be improved by devising better methods for SNP pruning.

PedMix works with haplotypes. At present, haplotypes are mainly inferred from genotype data

and thus usually contain errors. Figure 4 shows that if untreated, phasing error can indeed greatly increase the inference error of PedMix. On the other hand, when preprocessing is applied to remove the obvious phasing errors, inference errors can be significantly reduced. Nonetheless, phasing error can still reduce inference accuracy. Note that phasing methods are constantly improving and the problem of phasing errors may be greatly reduced in the near future.

As shown in Figure 2, it is desirable to use larger (e.g., whole genome) genetic data for the genetic settings that are similar to those of humans. Simulations show that PedMix can scale to whole genome data, when proper data preprocessing is performed. See Supplemental Methods and Supplemental Fig S3 on the running time of PedMix. The current implementation of PedMix assumes two ancestral populations. In principle, PedMix can be extended to allow more than two ancestral populations, although this may lead to increased computational time.

METHODS

Inferring admixture proportions from genetic data

Consider a single diploid individual from an admixed population. We assume two haplotypes H_1 and H_2 for this individual are given. Here, a haplotype is a binary vector of length n, where n is the number of single nucleotide polymorphisms (SNPs) within the haplotype. Note that in real data, H_1 and H_2 are usually inferred from the genotypes G and may have phasing errors. For the ease of exposition, we initially assume the absence of phasing errors in the haplotypes, and then extend the inference framework to allow phasing errors. The admixed population is assumed to be formed by an admixture of two ancestral populations (denoted as populations A and B) g generations ago. g is assumed to be known. For simplicity it is assumed that there are two ancestral populations, although the method can be extended to allow more than two ancestral populations. We further assume allele frequencies in the two ancestral populations are known for all SNPs. Note that allele frequencies from extant populations that are closely related to the ancestral populations are typically available. For example, suppose the admixed individual has genetic ancestry in West Africa and Northern Europe. So the allele frequencies from the extant YRI and CEU populations, available from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015), are used as approximations of the real ancestral allele frequencies. It is also assumed that recombination

fractions between every two consecutive SNPs are known. For human populations, recombination fractions are readily available (e.g., 1000 Genomes Project Consortium (2015)).

Likelihood computation on the perfect pedigree model

The *perfect pedigree* model in Liang and Nielsen (2014) can be used to describe the segregation of admixture tracts. Here, an admixture tract is a segment of the genome which originates from a single ancestral population. This model differs from many of the models typically used for inferring admixture tracts of an extant individual (e.g., Tang, et al. (2006); Price, et al. (2009); Sankararaman, et al. (2008); Paşaniuc, et al. (2009)). This model directly models the segregation of admixture tracts within a pedigree. Most current models assume that the ancestry process follows a Markov chain along the chromosome. However, because of recombination between tracts from multiple ancestors, the exact process does not follow a first-order Markov process (Liang and Nielsen, 2014; Gravel, 2012). The perfect pedigree model establishes a more accurate, but also much more computationally demanding, model that does not assume a Markov process for the ancestral process, especially for recent admixture events.

Figure 7 illustrates the perfect pedigree model for an extant observed haplotype H at a single site. A perfect pedigree is a perfect binary tree where each node represents a haplotype. All internal nodes in the pedigree are ancestors of H (the single leaf in the pedigree). We trace the ancestry of H backwards in time until reaching the time of admixture, T_m generations ago. The 2^{T_m} haplotypes at this time are called "founder" haplotypes (which themselves are unadmixed but may be from different ancestral populations). Under the assumption of no inbreeding, all ancestors are distinct. Notice that there is an assumption of a single admixture event. However, the model can easily be generalized to multiple admixture events.

There are two main aspects of the perfect pedigree: the ancestry vector C and the recombination vector R. C specifies which ancestral population each particular founder haplotype is from. For example, in Figure 7, C is a vector $(ABBAAABB)$, of length 8. It indicates that the leftmost founder is from the ancestral population A while the rightmost founder is from ancestral population B. As founders are unadmixed, C does not change along the genome. R specifies from which of the two parental haplotypes each descendant haplotype inherits its DNA at a particular genomic position. R is the key component in the well-known Lander-Green algorithm (Lander and Green,

1987). As shown in Figure 7, one can visualize R as a set of arrows, one for each meiosis, pointing to the left or right. There is a list of recombination vectors for n sites $(R_1, R_2, \ldots, R_n)$, where R_i is the recombination vector for site i.

The most obvious method for computing the likelihood $P(H|M)$ of the given haplotype H on the perfect pedigree model is using the Lander-Green algorithm (Lander and Green, 1987) to compute the probability of H for a given C. Then the sum of these probabilities over all possible C is equal to $P(H|M)$. Here M is a vector of admixture proportions for ancestors of interests in the pedigree. However, computation of $P(H|M)$ directly using the Lander-Green algorithm is not practical for most datasets. This is because first we need to determine the ancestral setting, C, which specifies the ancestral population for each founder. Moreover, the number of possible R grows very fast relative to the number of generations in the pedigree. Note that the Lander-Green algorithm needs to enumerate all possible R values. Even considering just a single site i, there are $2^{2^{T_m}}$ possible values of C and $2^{2^{T_m}-1}$ possible values of R_i $(1 \leq i \leq n)$. These numbers are prohibitively large for e.g., $T_m = 10$. To circumvent this problem, a two-stage model is adopted as described below.

A two-stage Markovian pedigree model for genotypes

Our objective is to infer the admixture proportions of ancestors in the perfect pedigree at the Kth generation in the past. Here, K is usually much smaller than the number of generations since admixture. For example, 1st generation inference $(K = 1)$ is for parents and 2nd generation inference $(K = 2)$ is for grandparents. The first phase of the two-stage model involves modeling the first K generations in the past using the perfect pedigree model. In the second phase, starting at the Kth generation in the past, there are 2^K ancestors, which are assumed to have ancestry distributions following the standard Markovian model. The ancestry of these 2^K founders can change along the genome following the standard Markovian process. This allows the ability to model the admixture of recent ancestors (e.g., parents and grandparents) without explicitly considering the entire pedigree.

The model defined so far concerns haploid genomes/chromosomes. However, most real data are from diploid individuals, possibly with unknown or relatively poorly estimated haplotype phasing. We extend the two-stage pedigree model by assuming that each of the two haplotype from the extant individual has been estimated, but with phasing errors that occur at a constant switch error rate. This leads to a genotype-based perfect pedigree model.

Figure 8 (A) illustrates the genotype-based perfect pedigree at a single position. It consists of two perfect pedigrees, one for each of the two haplotypes H_1 and H_2. Each node in the outline tree denotes an ancestral genotype of the extant genotype G. The two haplotypes H_1 and H_2 of G follow different pedigrees independently. For simplicity, we use a single haplotype with "average" admixture tracts to represent a diploid founder, which works well in practice. Note that the estimated admixture proportion of a founder is the average of the admixture proportions of its two haplotypes. One can view this "average" haplotype has the admixture proportion equal to the diploid founder.

To allow phasing errors between H_1 and H_2, we introduce the phase-switching indicator P, which indicates whether at this position the two haplotypes switch or not. One can visualize P as the arrow labeled by P in Figure 8. A P-arrow pointing to the left indicates that H_1 traces to the left half of the pedigree and H_2 traces to the right half of the pedigree. A P-arrow pointing to the right indicates the opposite. When moving along the diploid sequence (genotype), the direction of P changes when a phasing error occurs. Thus we can combine the two pedigrees for H_1 and H_2 and let the two haplotypes from a single individual collapse into one node, as illustrated in Figure 8 (B).

The full information regarding the ancestry of a genotype, $G = \{H_1, H_2\}$, in a fixed pedigree is then given by the ancestral configuration $AC = (P, C, R)$. The sampling probability of G can be computed naively by summing over all possible ACs. The ancestral configuration AC naturally leads to a Hidden Markov Model (HMM) that can be used for efficient calculation of the likelihood.

Let $\mathcal{AC}_i$ denote a set containing all possible ancestral configurations at site i and AC_i denote an element that belongs to $\mathcal{AC}_i$ ($AC_i \in \mathcal{AC}_i$). In a perfect pedigree of K generations, $AC_i = (P_i, C_i, R_i)$ is a binary vector of $2^{K+1} - 1$ bits and represents a state at site i. For each state, P_i has exactly one bit where a "0" (respectively "1") represents the phasing arrow pointing to the left (respectively right) and "1" represents the phasing arrow pointing to the right. R_i is a binary vector of $2^K - 2$ bits indicating the recombination states associated with all $2^K - 2$ meiosis in the pedigree, where "0" (respectively "1") represents a recombination arrow pointing to the left (respectively right). C_i is a binary vector that indicates the ancestry of each of the 2^K ancestors and contains 2^K bits when there are two ancestral populations. Also, if $C[j] = 0$ (respectively $C[j] = 1$) the j-th founder is from the population A (respectively B) at the current site. In the example in Figure 8 (B),

$AC = (P, C, R)$ at this site can be expressed as the binary vector $(1, 0101, 10)$.

We define $h(AC_i)$ as the joint probability of the length-i prefix of G (i.e., $G[1..i]$) and the ancestral configuration AC_i at site i. Given a genotype G with n sites, the likelihood $p(G|M) = \sum_{AC_n \in \mathcal{AC}_n} h(AC_n)$. The critical step is the computation of $h(AC_i)$ for each configuration AC_i at site i. This can be carried out in a recurrence for $i \geq 2$:

$$h(AC_i) = [\sum_{AC_{i-1} \in \mathcal{AC}_{i-1}} p_t(AC_i|AC_{i-1}) h(AC_{i-1})] \cdot p_e(AC_i) \tag{2}$$

where $p_t(AC_i|AC_{i-1})$ is the transition probability from AC_{i-1} at site $i-1$ to AC_i at site i and $p_e(AC_i)$ is the emission probability of an allele given the ancestral configuration AC_i at site i. This is the standard forward algorithm for HMMs. Transitions in the HMM may occur between adjacent sites and we assume, for generality, that the configurations at sites $i-1$ and i are fully connected as illustrated in Figure 9.

Transition and Emission probabilities of the HMM

Consider a founder j and two sites that are separated by d nucleotides. We first define the one-step ancestry transition probabilities $P^j_{A \to B}$ and $P^j_{B \to A}$. $P^j_{A \to B}$ (respectively $P^j_{B \to A}$) is the probability that ancestral population A (respectively B) changes to ancestral population B (respectively A) along the genome for the founder j when $d = 1$. Recall that the ancestry process of an ancestor follows a standard Markovian model. Suppose a haplotype of the individual j has ancestral population A at site $i-1$. The probability that site i has the ancestral population B is approximately $d \cdot P^j_{A \to B}$, while the probability that site i has ancestral population A is approximately $1 - d \cdot P^j_{B \to A}$, assuming that d (the number of bp between sites i and $i-1$) is small. Multiple transitions in the interval are ignored. Then T_j is defined to be the d-step transition probability of the ancestral settings for ancestor j:

$$T_j = \begin{cases} d \cdot P^j_{A\to B} & C_{i-1}[j] = 0, C_i[j] = 1 \\ d \cdot P^j_{B\to A} & C_{i-1}[j] = 1, C_i[j] = 0 \\ 1 - d \cdot P^j_{A\to B} & C_{i-1}[j] = 0, C_i[j] = 0 \\ 1 - d \cdot P^j_{B\to A} & C_{i-1}[j] = 1, C_i[j] = 1 \end{cases} \tag{3}$$

Notice that this is a function of d, and that i and $i-1$ are suppressed in the notation. Using similarly simplified notation, the phasing transition probabilities I are defined as:

$$I = \begin{cases} d \cdot p_p & P_{i-1} \neq P_i \\ 1 - d \cdot p_p & P_{i-1} = P_i \end{cases} \tag{4}$$

where p_p is the probability of a phasing error per unit length (assumed to be known and small enough that double or more phasing errors can be ignored).

We also define B_k as the transition probability of the recombination vector for the k'th bit. Given the recombination map of G, the recombination probability B_k between the two sites is computable. Let $p^r_{i,i-1}$ denote the probability of one recombination event between sites i and $i-1$, then

$$B_k = \begin{cases} p^r_{i,i-1} & R_{i-1}[k] \neq R_i[k] \\ 1 - p^r_{i,i-1} & R_{i-1}[k] = R_i[k] \end{cases} \tag{5}$$

Using this simplified notation, and assuming independence among transitions associated with recombination, phasing errors, and the ancestral population setting, the transition probabilities of the Markov chain are then given by:

$$p_t(AC_i|AC_{i-1}) = I \cdot \prod_k B_k \cdot \prod_j T_j \tag{6}$$

As mentioned above, the emission probability at site i is a function of the ancestral population assignment, C_i, and the alleles of the focal individual. At site i of the genotype G, there are two haplotypes (h_1, h_2). Let $f_{h_j}(AC_i)$ be the allele frequency in the population specified by AC_i for the allele observed at the position i of h_j $(j = 1, 2)$. The emission probability is then

$$p_e(AC_i) = \prod_{j=1,2} f_{h_j}(AC_i) \tag{7}$$

as in the standard definitions in genetic ancestry models (e.g., Pritchard, et al. (2000)).

Fast computation of sampling probability in PedMix

The main computational work in the evaluation of Equation 2 is that the calculation of $h(AC_i)$ requires a multiplication of the transition probability matrix and the vector $h(AC_{i-1})$. This leads to a computational complexity of $O(N_K^2)$, where N_K is the number of possible states in $\mathcal{AC}_i$. This is a significant burden on computation: for example if $K = 3$, the run time is on the order of 2^{30}. To address this problem, we develop a divide and conquer algorithm for computing the probability of ACs, which runs in $O(N_K log(N_K))$ time.

Let P_i denote the probability vector that contains all $h(AC_i)$ for $AC_i \in \mathcal{AC}_i$ at site i. Let $T_{i-1,i}$ denote the transition probability matrix containing the transition probabilities $p_t(AC_i|AC_{i-1})$ that one AC at site $i-1$ transits to another AC at site i. To obtain P_i, we need to compute $T_{i-1,i}P_{i-1}$. Direct computation leads to quadratic complexity.

For simplicity, we omit the site index notation i or $i-1$ in $T_{i-1,i}$ and P_{i-1}. Let T^b denote the transition probability matrix for AC that has b bits. The AC is represented as a binary vector of length b. Let P^b denote the probability vector for the previous site $(i-1)$ of length b. A bipartition of a matrix is a bipartition of each dimension, which divides a matrix into four sub-matrices with equal size. A bipartition of a vector is a division that equally cuts the vector into two sub-vectors. Figure 10 shows an example of a transition probability matrix T^3 and a probability vector P^3 for AC with 3 bits. For example, the $(2,3)$ element in T^3 is the transition probability $p_t((001)|(010))$. The bipartition for T^3 and P^3 is shown as red lines.

We observe that each bit in an AC_{i-1} transits to a bit in AC_i independently (i.e., the transition probability of each bit in AC_i doesn't depend on other bits). t^b_{xy} is used to denote the transition probability of the bth bit from x to y $(x, y \in \{0, 1\})$. The divide-and-conquer approach in Idury and Elston (1997) is adapted to our problem as follows. With bipartition, T^b can be viewed as four sub-matrices, and P^b can be divided into two sub-vectors. The key of the divide and conquer approach is given in the Equation 8.

$$T^b P^b = \begin{pmatrix} t^b_{00} T^{b-1} & t^b_{10} T^{b-1} \\ t^b_{01} T^{b-1} & t^b_{11} T^{b-1} \end{pmatrix} \begin{pmatrix} P^{b,0} \\ P^{b,1} \end{pmatrix} = \begin{pmatrix} t^b_{00} T^{b-1} P^{b,0} + t^b_{10} T^{b-1} P^{b,1} \\ t^b_{01} T^{b-1} P^{b,0} + t^b_{11} T^{b-1} P^{b,1} \end{pmatrix} \tag{8}$$

Each sub-matrix of T^b is equal to T^{b-1} multiplied by t^b_{xy}. Here, T^{b-1} is a transition probability matrix where the bth bit of T^b is masked off. For example, the top left sub-matrix of T^3 (Figure 10) is equal to $t^3_{00} T^2$ and the top right sub-matrix is equal to $t^3_{10} T^2$. Let $P^b = (P^{b,0}, P^{b,1})$ denote the bipartition of the probability vector. Then $T^b P^b$ can be computed by computing $T^{b-1} P^{b,0}$ and $T^{b-1} P^{b,1}$. In general, $T^{b-1} P^{b,0}$ and $T^{b-1} P^{b,1}$ can then be divided in a similar way until we reach T^1 (masking off $b-1$ bits in AC). For the Kth generation inference, each AC has $b = 2^{K+1} - 1$ bits, which leads to $N_K = 2^B = 2^{2^{K+1}-1}$ possible states at each site. The divide and conquer scheme reduces computational complexity from $O(N_K^2)$ to $O(N_K log(N_K))$.

Probabilistic inference

Maximum Likelihood (ML) inference of admixture proportions can be obtained by maximizing the sampling probability $p(G|M)$ of the $\mathcal{AC}$-based HMM model:

$$M^* = \arg\max_M p(G|M) \tag{9}$$

Let m^j_0 and m^j_1 denote the admixture proportions of the populations A and B respectively for the ancestor j. These admixture proportions are then given by the stationary frequencies of the Markov chain, which according to standard theory, are given by

$$m^j_0 = \frac{P^j_{B\to A}}{P^j_{A\to B} + P^j_{B\to A}} \tag{10}$$

and

$$m^j_1 = \frac{P^j_{A\to B}}{P^j_{A\to B} + P^j_{B\to A}} = 1 - m^j_0 \tag{11}$$

respectively. From the invariance principle of ML, it follows that if $P^j_{A\to B}$ and $P^j_{B\to A}$ are estimated by ML, the resulting estimates of m^j_0 and m^j_1 are also ML estimates.

To obtain ML estimates of $P^j_{A\to B}$ and $P^j_{B\to A}$, the Boyden-Fletcher-Goldfarb-Shanno (BFGS)

method of optimization is used. We use an implementation of the limited-memory version of the algorithm (L-BFGS) (`http://www.chokkan.org/software/liblbfgs`) and the finite difference method for estimating derivatives. We transform bounded parameters to unbounded parameters using the logit function to accommodate bound constraints.

Preprocessing

There are several aspects of real data that are not considered by our models and may affect the inference accuracy, in particular background Linkage Disequilibrium (LD) and phasing errors. Background LD refers to non-random association between alleles not caused by admixture. Background LD may mislead HMM methods which assume conditional independence among SNPs. As a consequence it may confuse the background LD with the admixture LD. Phasing errors may also introduce an extra layer of noise.

The traditional approach for addressing the problem of background LD is to trim the data sets by removing SNPs. We compare two possible strategies for doing this:

1. Data trimming based on allele frequency differences (frequency-based pruning).

2. Data trimming based on LD patterns (LD pruning).

Frequency-based pruning relies on a trimming threshold d_f, which specifies the minimum allele frequency difference in the two source populations. A SNP site is trimmed if the absolute difference between the allele frequencies in the source populations is smaller than d_f (Supplemental Methods).

In LD pruning, SNPs are removed in order to minimize the LD among SNPs located in the same region. This is the more commonly used strategy implemented in programs such as PLINK (Purcell, et al., 2007). See Supplemental Methods for more details on LD pruning.

The advantage of the second approach is that it more directly reduces LD in the data. The advantage of the first approach is that it keeps the most ancestry informative SNPs in the data set. Both approaches improve inference accuracy and reduce computational time. However, our implementation of frequency-based pruning leads to slightly better performance and therefore, this method is used as the default unless otherwise stated.

Phasing error

In real haplotype datasets, phasing error usually cannot be eradicated when haplotypes are inferred from genotypes. In some sense, phasing errors and recombination have similar effects on the genomes of the extant individual. We develop a technique for removing some phasing errors during preprocessing. Briefly, the admixture tracts for the current haplotypes are first estimated. Note that admixture tracts are expected to be relatively long, however phasing errors may shorten admixture tracts. So the potential phasing errors can be inferred by examining the unexpected short admixture tracts (Supplemental Methods and Supplemental Fig S4).

Software availability

The program PedMix is available from the Supplemental Materials and also can be downloaded from `https://github.com/pjweggy/PedMix`.

DISCLOSURE DECLARATION

The authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

This work is partly supported by U.S. National Science Foundation grants IIS-1526415 and CCF-1718093. We thank three anonymous reviewers for helping to improve the manuscript. We also thank Walter Krawec and Aaron Palmer for reading and helping to improve the manuscript.

REFERENCES

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*, 526(7571):68.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664.

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. 2010. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564.

Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome research*, 19(1):136–142.

Ewing G, Hermisson J. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065.

Gravel S. 2012. Population genetics models of local ancestry. *Genetics*, 191(2):607–619.

Hellenthal G, Stephens M. 2007. msHOT: modifying hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, 23(4).

Hudson RR. 2002. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

Idury RM, Elston RC. 1997. A faster and more general hidden markov model algorithm for multipoint likelihood calculations. *Human heredity*, 47(4):197–202.

Kruglyak L, Lander ES. 1998. Faster multipoint linkage analysis using Fourier transforms. *Journal of Computational Biology*, 5(1):1–7.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8):2363–2367.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *science*, 319(5866):1100–1104.

Liang M, Nielsen R. 2014. The lengths of admixture tracts. *Genetics*, 197(3):953–967.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFmix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288.

Margalit Y, Baran Y, Halperin E. 2015. Multiple-Ancestor Localization for Recently Admixed Individuals. In: Pop M., Touzet H. (eds) Algorithms in Bioinformatics. WABI 2015. Lecture Notes in Computer Science, vol 9289. Springer, Berlin, Heidelberg.

Paşaniuc B, Sankararaman S, Kimmel G, Halperin E. 2009. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):i213–i221.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6):e1000519.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *science*, 298(5602):2381–2385.

Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, Clark AG. 2010. Inferring genetic ancestry: opportunities, challenges, and implications. *The American Journal of Human Genetics*, 86(5):661–673.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303.

Tang H, Coram M, Wang P, Zhu X, Risch N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1):1–12.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology*, 28(4):289–301.

Wakeley J, King L, Low BS, Ramachandran S. 2012. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics*, 190(4):1433–1445.

Yang W-Y, Platt A, Chiang CW-K, Eskin E, Novembre J, Pasaniuc B. 2014. Spatial Localization of Recent Ancestors for Admixed Individuals. G3: Genes—Genomes—Genetics. 2014;4(12):2505-2518. doi:10.1534/g3.114.014274.

Zou JY, Halperin E, Burchard E, Sankararaman S. 2015. Inferring parental genomic ancestries using pooled semi-Markov processes. *Bioinformatics*, 31(12):i190-6.

Figures Legend

***Figure 1:* Comparison between the accuracy of PedMix and random guess.**

Mean errors for parents, grandparents, and great grandparents. Smaller number of SNPs are used for great grandparent simulations.

***Figure 2:* Impact of data amount.**

5, 10, 15, and 20 chromosomes are used respectively. For great grandparents inference, $d_f = 0.9$ is used due to the computational burden.

***Figure 3:* Mean error for different simulation parameter settings.**

(A) Varying mutation rates. $L = 5 \times 10^8$.

(B) Varying recombination rates. $L = 5 \times 10^8$.

(C) Varying the time since admixture.

(D) Ancestral population split time. For $t = 0.01$, there are no SNPs left after using the $d_f = 0.5$ cut-off. As a result, $d_f = 0.2$ is used, leaving 74,227 SNPs for analysis.

Default parameters are used except for the variable indicated by the X axis of each plot.

***Figure 4:* Inference error vs. phasing error.**

Comparison among three different datasets: dataset without phasing errors, dataset with phasing errors, and preprocessed datasets.

(A) Individuals sampled from an admixed population.

(B) Individuals sampled from an unadmixed population.

***Figure 5:* Example pie charts of admixture proportions for six ancestors.**

For each individual, six pie charts are displayed which represent two parents and four grandparents respectively. The blue part shows the percentage of CEU population and the orange part shows the percentage of YRI population.

***Figure 6:* Six pedigrees for semi-simulated data analysis.**

The percentage of CEU origin is shown in pedigrees. Admixture proportions in black are estimates by RFmix using ancestors' genotypes and are assumed to be the true proportions. Admixture proportions in blue and red are estimates by PedMix using genotypes of the focal individual with (red) or without (blue) phasing error. Mean error is the average over the two parents and four grandparents.

Figure 7: The perfect pedigree model for $T_m = 3$ for a single site.

H: extant haplotype. Ancestry origins of 8 founders are listed above the perfect pedigree. At one site, A and B indicate which of the two ancestral populations of each founder's haplotype. The combined vector of these values is C. Arrows: the recombination vector R. The population A shown in red: the ancestry of H as traced back by the recombination setting. Arrows can change direction at the next site.

Figure 8: The perfect pedigree model for genotype.

$G = (H_1, H_2)$ at a site, and $K = 2$.

(A) Outline pedigree in black: the perfect pedigree for genotype G. Two pedigrees embedded in red are for haplotype H_1 and H_2 respectively. Ancestral settings and recombination settings with the same label have the same meaning.

(B) The simplified perfect pedigree for genotype G. Ancestral vector C: $(ABAB)$. The arrows without label define recombination vector R. P: the phasing error setting.

Figure 9: An example of HMM.

An example of $\mathcal{AC}$-based HMM with 128 ACs as states for $K = 2$ generations (i.e., grandparents) with arrows indicating possible transitions along the Markov chain from site $i - 1$ to i. The vectors under each pedigree provide the binary representations of P, C, and R, respectively, for the pedigree. The two top thick arrows and the lower thick arrow indicate the settings of R and P, respectively.

Figure 10: Faster calculation of the probabilities of ACs.

Red lines break the transition probability matrix into four smaller pieces (for ACs with length 2).

The probability vector at the previous site is broken into two pieces. Multiplication of the matrix and the vector is faster due to shared parts between these pieces.

Table 1: *A list of parameters and their default values used in the simulation.*

Description	Symbol	Default
The number of haplotypes	n_h	1000
The number of chromosomes	n_c	22
Effective population size	N_e	10000
Region length (bp)	L	$3 \times 10^9 / 5 \times 10^8$
Mutation rate (per generation per bp)	μ	1×10^{-8}
Recombination rate (per generation per bp)	ρ	1×10^{-8}
Ancestral populations splitting time (scaled with N_e)	t	0.2
The number of generations for admixture	g	10
The number of individuals to infer	n_i	10
Frequency-based pruning threshold	d_f	0.5

Table 2: *Mean and standard deviation of the error in the estimate of admixture proportions for ADMIXTURE, RFmix, and PedMix (in unit of %). Note the PedMix results are the average proportions from the estimated admixture proportions of parents or grandparents .*

Error (in %)	ADMIXTURE	RFmix	PedMix (ave. of parents)	PedMix (ave. of grandparents)
mean	1.16	1.77	2.01	3.53
standard deviation	1.43	1.52	2.5	2.53

Table 3: *Mean and standard deviation of the error in the estimate of admixture proportions of parents from ANCESTOR and PedMix (in unit of %).*

Error (in %)	ANCESTOR (parents)	PedMix (parents)
mean	9.63	6.54
standard deviation	6.13	2.14

Table 4: *Compare the effects of LD pruning and frequency-based pruning on inference accuracy. LD: results with LD pruning. F-prune: results with frequency-based pruning.*

Inference error (%)	LD (284K)	F-prune (283K)	LD (97K)	F-prune (94K)
parents	14.99	14.38	13.03	7.93
grandparents	18.02	16.48	16.05	14.79

Table 5: *Comparison of admixture proportion estimate among ADMIXTURE, RFmix, and PedMix for five ASW individuals. The PedMix results are the averages of the admixture proportions of parents and grandparents inferred by PedMix. These averages are used as proxies for focal individuals, and are then compared with estimates of ADMIXTURE and RFmix. The percentage of CEU origin is shown in table.*

individual ID	ADMIXTURE	RFmix	PedMix (ave. of parents)	PedMix (ave. of grandparents)
NA20296	22.06%	21.78%	20.46%	17.37%
NA19625	32.02%	34.35%	31.96%	31.13%
NA19921	41.22%	41.20%	42.61%	39.84%
NA20414	63.84%	68.76%	63.88%	67.00%
NA20314	92.38%	97.93%	88.76%	88.13%

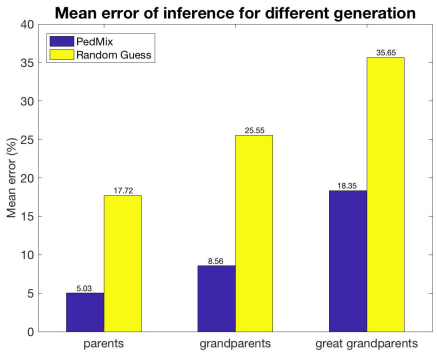

Figure 1: *Comparison between the accuracy of PedMix and random guess for parents, grandparents and great grandparents. Smaller number of SNPs are used for great grandparent simulations.*

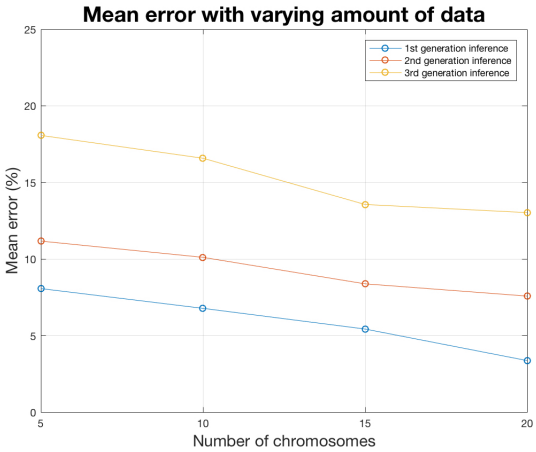

Figure 2: *Impact of data amount. We use 5, 10, 15, and 20 chromosomes respectively. For great grandparents inference, $d_f = 0.9$ is used due to computational burden.*

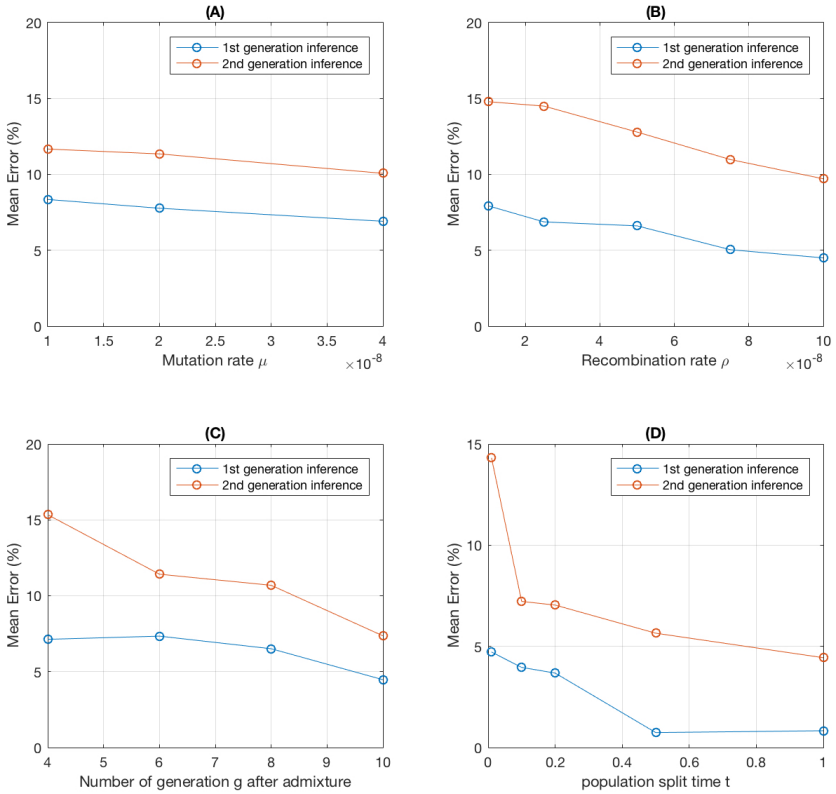

Figure 3: *Mean error for different simulation parameter settings. (A) Varying mutation rates.* $L = 5 \times 10^8$. *(B) Varying recombination rates.* $L = 5 \times 10^8$. *(C) Varying the time since admixture. (D) Ancestral population split time. For* $t = 0.01$, *there are no SNPs left after using the* $d_f = 0.5$ *cut-off. As a result, we use* $d_f = 0.2$, *leaving 74,227 SNPs for analysis. Default parameters are used except for the variable indicated by the X axis of each plot.*

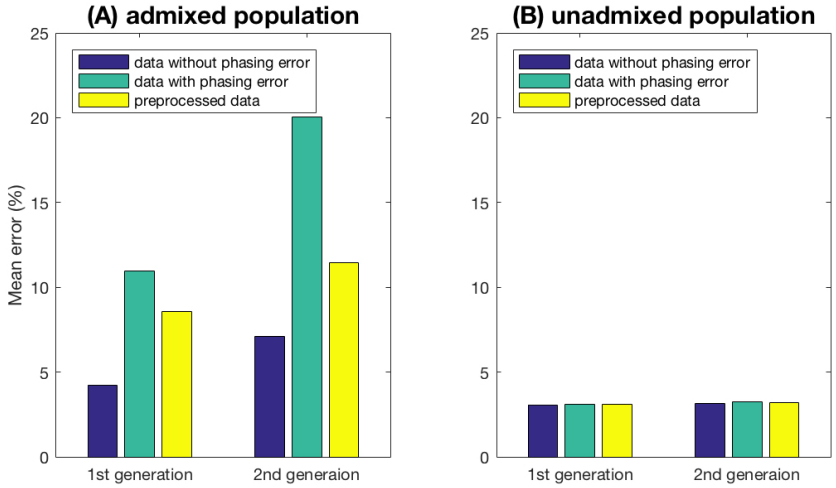

Figure 4: *Inference error vs phasing error: comparison among three different datasets: dataset without phasing errors, dataset with phasing errors and preprocessed datasets. (A) Samples from an admixed population. (B) Samples from an unadmixed population.*

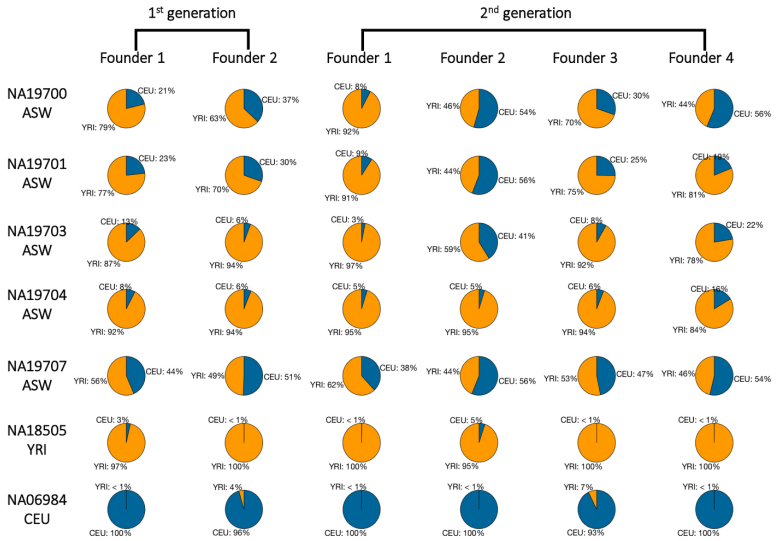

Figure 5: *Example pie charts of admixture proportions for six ancestors. For each individual, we display six pie charts which represent two parents and four grandparents respectively. The blue part shows the percentage of CEU population and the orange part shows the percentage of YRI population.*

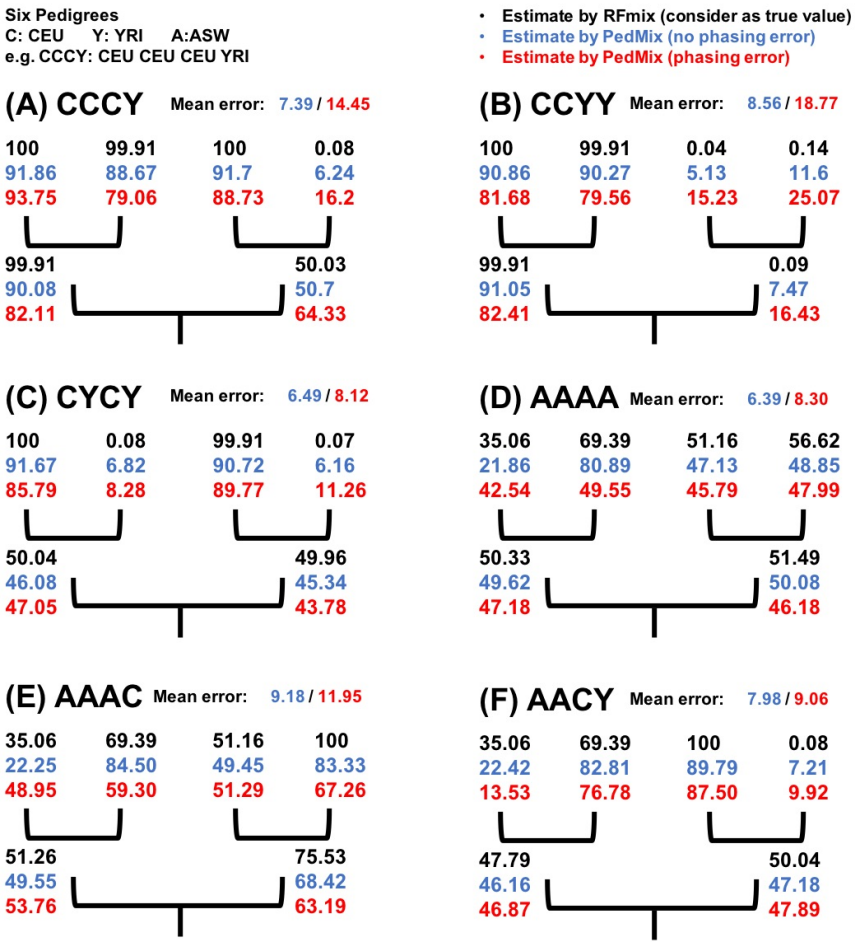

Figure 6: *Six pedigrees for semi-simulated data analysis. The percentage of CEU origin is shown in pedigrees. Admixture proportions in black are estimates by RFmix using ancestors' genotypes and are assumed to be the true proportions. Admixture proportions in blue and red are estimates by PedMix using genotypes of the focal individual with (red) or without (blue) phasing error. Mean error is the average over the two parents and four grandparents.*

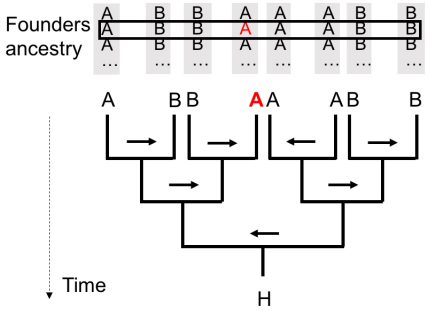

Figure 7: *The perfect pedigree model for $T_m = 3$ for a single site. H: extant haplotype. Ancestry origins of 8 founders are listed above the perfect pedigree. At one site, A and B indicate which of the two ancestral populations of each founder's haplotype. The combined vector of these values is C. Arrows: the recombination vector R. The population A shown in red: the ancestry of H as traced back by the recombination setting. Arrows can change direction at the next site.*

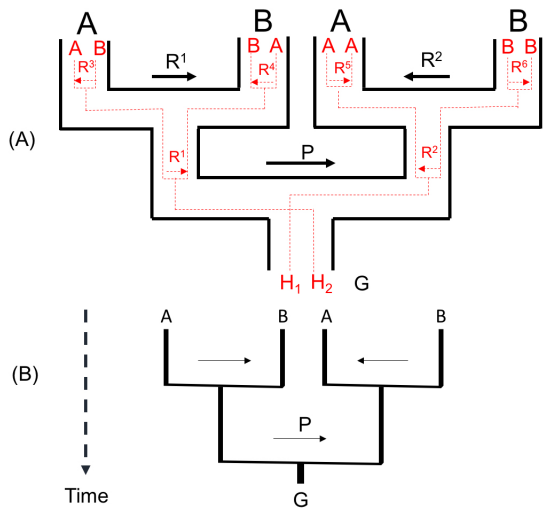

Figure 8: *The perfect pedigree model for genotype $G = (H_1, H_2)$ at a site, and $K = 2$. (A) Outline pedigree in black: the perfect pedigree for genotype G. Two pedigrees embedded in red are for haplotype H_1 and H_2 respectively. Ancestral settings and recombination settings with the same label have the same meaning. (B) The simplified perfect pedigree for genotype G. Ancestral vector C: $(ABAB)$. The arrows without label define recombination vector R. P: the phasing error setting.*

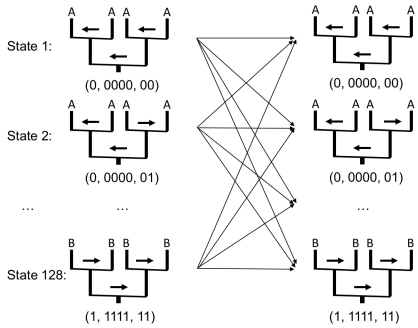

Figure 9: *An example of $\mathcal{AC}$-based HMM with 128 ACs as states for $K = 2$ generations (i.e. grandparents) with arrows indicating possible transitions along the Markov chain from site $i-1$ to i. The vectors under each pedigree provide the binary representations of P, C, and R, respectively, for the pedigree. The two top thick arrows and the lower thick arrow indicate the settings of R and P, respectively.*

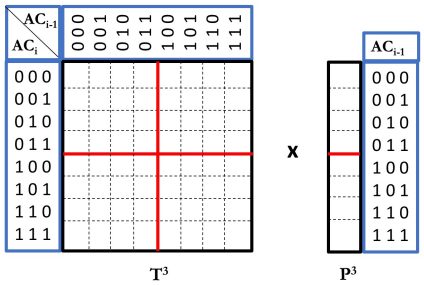

Figure 10: *Faster calculation of the probabilities of ACs. Red lines break the transition probability matrix into four smaller pieces (for ACs with length 2). The probability vector at the previous site is broken into two pieces. Multiplication of the matrix and the vector is faster due to shared parts between these pieces.*