RESEARCH ARTICLE

Inferring the ancestry of parents and grandparents from genetic data

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

1 Department of Computer Science and Engineering, University of Connecticut, Storrs, Connecticut, United States of America, **2** Departments of Integrative Biology and Statistics, University of California, Berkeley, Berkeley, California, United States of America, **3** Museum of Natural History, University of Copenhagen, Copenhagen, Denmark

* rasmus_nielsen@berkeley.edu (RN); yufeng.wu@uconn.edu (YW)

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 decomposition 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 only using the genome of an extant individual. We perform extensive simulations to quantify the error in the estimation of ancestral admixture proportions under various conditions. To illustrate the utility of the method, we apply it to real genetic data.

Author summary

Ancestry inference is an important problem in genetics and is used commercially by a number of companies affecting millions of consumers of genetic ancestry tests. In this paper, we show that it is possible, not only to estimate the ancestry fractions of an individual, but also, with some uncertainty, to estimate the ancestry fractions of an individual's recent ancestors. For example, if an individual traces his/her ancestry 50% to Asia and 50% to Europe, it is possible to distinguish between the individual having two parents that each are 50:50 composites of Asian and European ancestry, or one parent from Asia and one from Europe. It is likewise also possible to make inferences about grandparents. We present a computationally efficient method for making such inferences called PedMix. PedMix is based on a probabilistic model for the descendant and the recent ancestors. PedMix infers admixture proportions of recent ancestors (parents, grandparents or even great grandparents) using whole-genome genetic variation data from a focal individual.

OPEN ACCESS

Citation: Pei J, Zhang Y, Nielsen R, Wu Y (2020) Inferring the ancestry of parents and grandparents from genetic data. PLoS Comput Biol 16(8): e1008065. https://doi.org/10.1371/journal.pcbi.1008065

Editor: Sergei L. Kosakovsky Pond, Temple University, UNITED STATES

Received: October 6, 2019

Accepted: June 17, 2020

Published: August 14, 2020

Data Availability Statement: The PedMix software is free to download from: https://github.com/yufengwudcs/PedMix.

Funding: This work is partly supported by U.S. National Science Foundation grants IIS-1526415, CCF-1718093 and IIS-1909425. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Results on both simulated and real data show that PedMix performs reasonably well in most scenarios.

This is a *PLOS Computational Biology* Methods paper.

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. There are now at least 26 million participants in commercial genetic databases in the US alone. A primary objective for participants to join is to learn more about their genetic ancestry. Consumers consider such information relevant for their identity and for understanding their family history, and results are often compared to knowledge handed down orally. In some cases, such as for adopted children, ancestry tests can be particularly important. Despite many warnings not to allow genetics dictate issues of identity [1], ancestry tests are nonetheless becoming an important, and often identity defining, part of American consumer culture. Ancestry inference also forms the basis of many standard population genetic analyses and most population genomic publications include ancestry inference analyses in one form or another (e.g., [2, 3]). Modern ancestry inference has roots in the seminal paper on STRUCTURE [4]. The model introduced in that paper assumes that 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, and the ancestry for each allele is then described as 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 [5] and ADMIXTURE [6]. 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 two parents.

Nonetheless, this assumption can be relaxed opening up the possibility of direct inference of the admixture proportions in parents, grandparents, or great-grandparents. Such inference will be of interest to individuals who are trying to understand their family history and the origins of their immediate family. For example, it will provide adopted children of admixed genetic ancestry an opportunity to understand the origins of their different parents and grandparents. It will also open up the possibility in ecological studies to investigate a variety of questions related to parentage without actually sampling the parents. For example, questions regarding assortative mating can be explored in recently admixed populations without direct access to the parents of the genotypes/sequenced individuals. This approach of studying parentage without pedigrees will be particularly important in organisms where it is difficult to directly observe mating pairs.

We note that there is substantial information in genotypic data on parental admixture proportions. Even without linkage information, 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 the ancestral population A, and a frequency of 0 and 1, respectively, in the 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 the populations A and B, is equal to one. However, if the two parents are both 50:50 (%) admixed between the 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 [7]. This method models the ancestry process along each of the chromosomes as a semi-Markov process, as the length distribution of admixture tracts is well-known not to follow the exponential prediction of a Markov process [8]. 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 show that parental ancestry can be estimated with reasonable accuracy. However, to our knowledge, no software program for inferring parental ancestry was distributed with the publication of [7]. The major difference between our method and their method is that their method first estimates ancestry segments, while ancestry segment inference is directly incorporated into our method allowing uncertainty in ancestry inference to be modeled directly. Note that the method in [7] is only applicable to parents, and not grand-parents or great-grandparents. Moreover, our results show that their method is comparatively slow and could be difficult to apply to large data sets.

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. We show that the distribution of tract lengths provides information that can be used for such inference. 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. While these estimates are associated with some variance, we show that they nonetheless can be used to distinguish between various hypotheses regarding the admixture proportions of parents, grandparents and great-grandparents. Our method has been implemented in a computer program called PedMix (available for download at https://github.com/yufengwudcs/PedMix).

Materials and methods

Inferring admixture proportions from genetic data

We 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*. *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. For simplicity we assume 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, especially for recent admixed populations. For example, suppose the admixed individual has genetic ancestry in West Africa and Northern Europe. Then we may use the allele frequencies from the extant YRI and CEU populations, available from the

1000 Genomes Project [9], as approximations of the real ancestral allele frequencies. Here, CEU refers to Utah Residents with Northern and Western European Ancestry, and YRI refers to Yoruba in Ibadan, Nigeria. We also assume recombination fractions between every two consecutive SNPs are known. For human populations, recombination fractions are readily available (e.g. [9]).

Likelihood computation on the perfect pedigree model

The *perfect pedigree* model in [10] 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. [11, 12, 13, 14]). 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 [10, 8]. 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.

Fig 1 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, g generations ago. The 2^g 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.

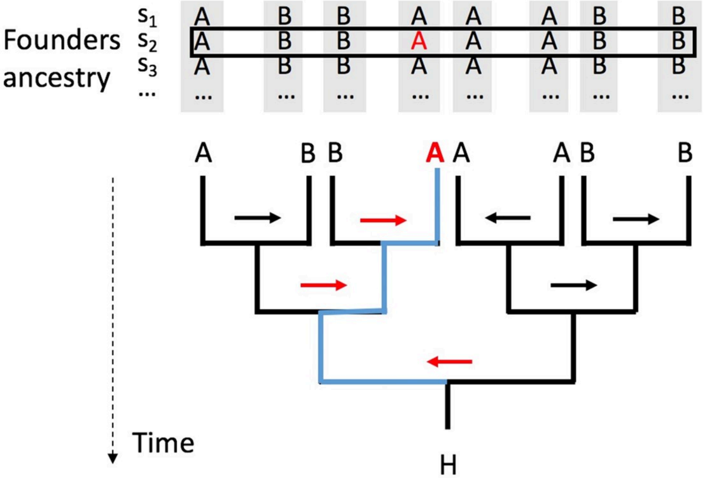

Fig 1. The perfect pedigree model for $g = 3$ for a single site. H: extant haplotype. Pedigree is haplotype-based, which models ancestry changes along the genome. Ancestry origins of 8 founders are listed above the perfect pedigree. At one site, A and B indicate which of the two ancestral populations of each founder’s haplotype. The combined vector of these values is C. Here, $C = (ABBAAABB)$. Arrows: the recombination vector R. Here, $R = 0111101$, where meiosis is ordered the in reverse time order and also from left to right. 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. Founder ancestry is at a specific sites (say $s_1, s_2, s_3, \ldots$). Note that founder ancestry at the founders of the pedigree remains the same at different genomic position: these founders are the founding members of the admixed population and they are not admixed themselves.

https://doi.org/10.1371/journal.pcbi.1008065.g001

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 Fig 1, 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 [15]. As shown in Fig 1, one can visualize R as a set of arrows, one for each meiosis, pointing to the left or right. There are seven such arrows in R. We have 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 [15] to compute the probability of H for a given C. Then we sum these probabilities over all possible C to obtain $P(H|M)$, assuming each C is equally likely. Here M is a vector of admixture proportions for ancestors of interest 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 with 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^g} possible values of C and 2^{2^g-1} possible values of R_i ($1 \leq i \leq n$). These numbers are prohibitively large for e.g. $g = 10$. To circumvent this problem, we adopt a two-stage model 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 K^{th} generation in the past. Here, K is usually much smaller than the number of generations since admixture. For example, 1^{st} generation inference ($K = 1$) is for parents and 2^{nd} 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 K^{th} 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 us 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 haplotypes 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.

Fig 2A 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 as one with the admixture proportion equal to the diploid founder.

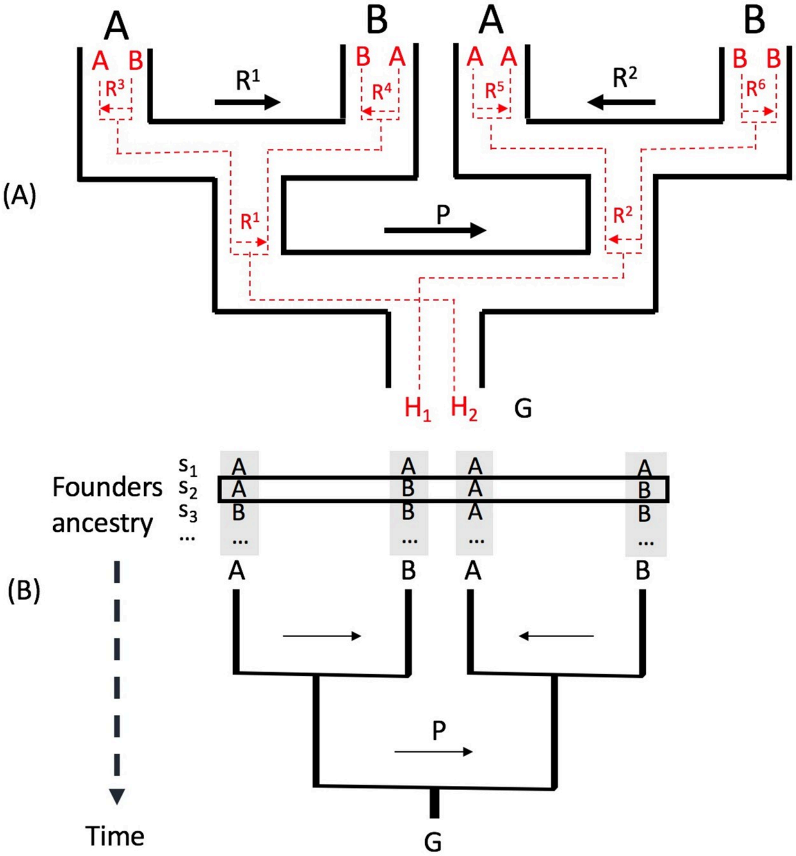

Fig 2. The perfect pedigree model for genotype $G = (H_1, H_2)$ at a site, and $K = 2$. Two sources populations: A and B. (A) Outline pedigree in black: the perfect pedigree for genotype G. Two pedigrees embedded in red are for haplotype H_1 and H_2 respectively. Ancestral settings and recombination 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. Each R^i is for a meiosis in the pedigree. P: the phasing error setting. Note that different from Fig 1, the founders in this pedigree can be admixed themselves (i.e., ancestry of these founders can change along the genome) based on the two-stage Markovian pedigree model.

https://doi.org/10.1371/journal.pcbi.1008065.g002

To allow phasing errors between H_1 and H_2, we introduce the phase-switching indicator P. It indicates whether at this position the two haplotypes switch or not. One can visualize P as the arrow labeled by P in Fig 2. 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 Fig 2B.

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 an 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 the 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). 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 Fig 2B, $AC = (P, C, R)$ 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 the 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 the site i. This can be carried out in a recurrence for $i \geq 2$ (which resembles the Lander-Green algorithm on the high-level):

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

where $p_t(AC_i|AC_{i-1})$ is the transition probability from AC_{i-1} at the site $i - 1$ to AC_i at the site i and $p_e(AC_i)$ is the emission probability of an allele given the ancestral configuration AC_i at the site i. This is the standard forward algorithm for HMMs. Details are given below. 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 Fig 3.

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 the ancestral population A

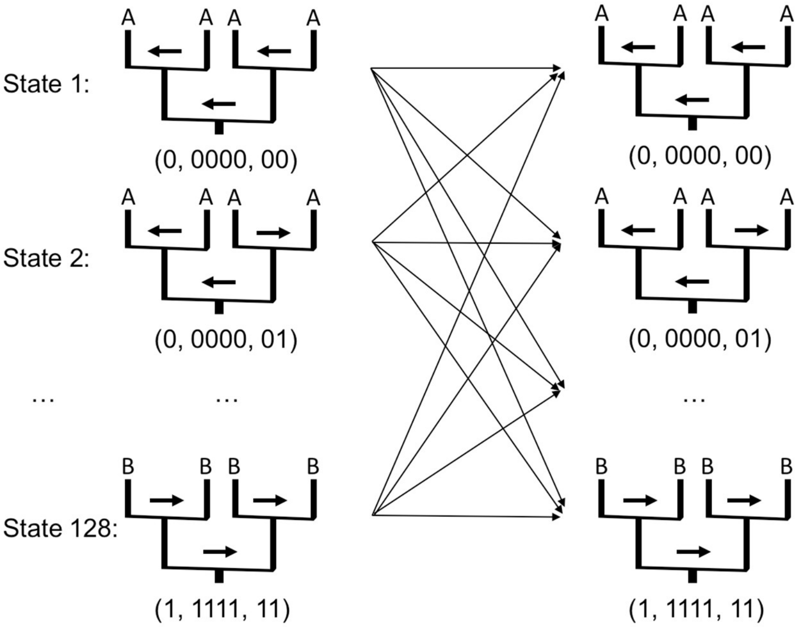

Fig 3. An example of $\mathcal{AC}$-based HMM with 128 ACs as states for $K = 2$ generations (i.e. grandparents). Arrows: possible transitions along the Markov chain from site $i - 1$ to i. The vectors under each pedigree provide the binary representations of P, C, and R, respectively, for the pedigree. The two top thick arrows and the lower thick arrow indicate the settings of R and P, respectively.

https://doi.org/10.1371/journal.pcbi.1008065.g003

(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 the ancestral population A at the site $i - 1$. The probability that the site i has the ancestral population B is approximately $d \cdot P^j_{A\to B}$ and the probability that the site i has the ancestral population A is approximately $1 - d \cdot P^j_{A\to B}$, assuming that d (the number of bp between the sites i and $i - 1$) is small. Multiple transitions in the interval are ignored. We then define T_j to be the d-step transition probability of the ancestral settings for ancestor j:

$$T_j = \begin{cases} d \cdot P^j_{A\to B} & C_{i-1}[j] = 0, C_i[j] = 1 \\ d \cdot P^j_{B\to A} & C_{i-1}[j] = 1, C_i[j] = 0 \\ 1 - d \cdot P^j_{A\to B} & C_{i-1}[j] = 0, C_i[j] = 0 \\ 1 - d \cdot P^j_{B\to A} & C_{i-1}[j] = 1, C_i[j] = 1 \end{cases} \tag{2}$$

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

$$I = \begin{cases} d \cdot p_p & P_{i-1} \neq P_i \\ 1 - d \cdot p_p & P_{i-1} = P_i \end{cases} \tag{3}$$

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 kth 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

$$B_k = \begin{cases} p^r_{i,i-1} & R_{i-1}[k] \neq R_i[k] \\ 1 - p^r_{i,i-1} & R_{i-1}[k] = R_i[k] \end{cases} \tag{4}$$

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:

$$p_t(AC_i|AC_{i-1}) = I \cdot \prod_k B_k \cdot \prod_j T_j \tag{5}$$

As mentioned above, the emission probability at the site i is a function of the ancestral population assignment, C_i, and the alleles of the focal individual. At the 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

$$p_e(AC_i) = \prod_{j=1,2} f_{h_j}(AC_i) \tag{6}$$

as in the standard definitions in genetic ancestry models (e.g., [4]).

Note that, in contrast to the original perfect pedigree model, the number of generations since admixture, g, is not a parameter in the two-stage model. This is because the likelihood is only dependent on the ancestry change within each founder in the two-stage model, not on g.

Fast computation of sampling probability in PedMix. The main computational burden in the evaluation of Eq 1 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 have developed 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 the 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 the site $i - 1$ transits to another AC at the 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 $(i - 1)$. 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. Fig 4 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). We use t^b_{xy} to denote the transition probability of the bth bit from x to y $(x, y \in \{0, 1\})$. We can adapt the divide-and-conquer approach in [16] 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 Eq 7.

$$T^b P^b = \begin{pmatrix} t^b_{00} T^{b-1} & t^b_{10} T^{b-1} \\ t^b_{01} T^{b-1} & t^b_{11} T^{b-1} \end{pmatrix} \begin{pmatrix} P^{b,0} \\ P^{b,1} \end{pmatrix} = \begin{pmatrix} t^b_{00} T^{b-1} P^{b,0} + t^b_{10} T^{b-1} P^{b,1} \\ t^b_{01} T^{b-1} P^{b,0} + t^b_{11} T^{b-1} P^{b,1} \end{pmatrix} \qquad (7)$$

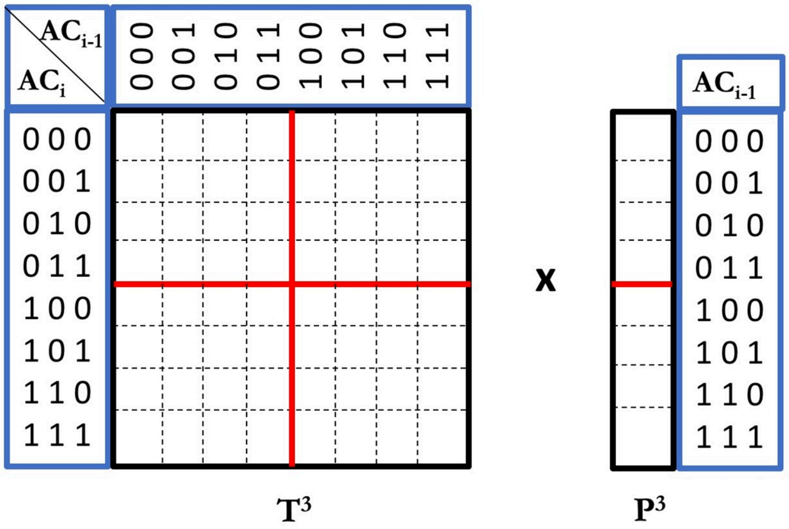

Fig 4. Faster calculation of the probabilities of ACs. Red lines break the transition probability matrix into four smaller pieces (for ACs with length 2). The probability vector at the previous site is broken into two pieces. Multiplication of the matrix and the vector is faster due to shared parts between these pieces.

https://doi.org/10.1371/journal.pcbi.1008065.g004

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 (Fig 4) 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 K^{th} 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 $O(N_K\, log(N_K))$. For more detailed explanation of this divide-and-conquer approach, we refer the readers to a related approach in [16].

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:

$$M^* = \arg\, \max_M P(G|M) \tag{8}$$

Let m_0^j and m_1^j 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

$$m_0^j = \frac{P^j_{B\to A}}{P^j_{A\to B} + P^j_{B\to A}} \tag{9}$$

and

$$m_1^j = \frac{P^j_{A\to B}}{P^j_{A\to B} + P^j_{B\to A}} = 1 - m_0^j \tag{10}$$

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_0^j and m_1^j are also ML estimates.

To obtain ML estimates of $P^j_{A\to B}$ and $P^j_{B\to A}$ we apply the Boyden-Fletcher-Goldfarb-Shanno (BFGS) method of optimization. We use an implementation of the limited-memory version of the algorithm, L-BFGS [17], from http://www.chokkan.org/software/liblbfgs. We use the finite difference method for estimating derivatives. We transform bounded 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. See Section S1.1.1 in S1 Text for more details on this approach.

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 [18]. See Section S1.1.2 in S1 Text for more details on LD pruning. The advantage of the first approach is that it keeps the most ancestry informative SNPs in the data set. The advantage of the second approach is that it more directly reduces LD in the data. Both approaches improve inference accuracy and reduce computational time. For frequency-based trimming, for example, the parental inference error without trimming is about 6.5%, while with trimming, the error rate can be about 5% (with 50% threshold). However, our implementation of frequency-based pruning leads to slightly better performance (see Table C in S1 Text), and we, therefore, use this method 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 have developed a technique for removing some phasing errors during preprocessing. Briefly, we first estimate the admixture tracts for the current haplotypes. We expect admixture tracts to be relatively long, but may be shortened by phasing errors. Phasing errors can, therefore, be removed to some extent by removing unexpectedly short admixture tracts. See Section S1.2 for details.

Results

This section contains results on simulated, semi-simulated and real data. Some results are given in Section S1.6 in S1 Text.

Results on simulated data

Simulation settings and evaluation. We perform extensive simulations to evaluate the performance of our method. The parameters we use in the simulations are listed and explained in Table 1 together with their default values. We first simulate a number of haplotypes using macs [19] from two ancestral populations which diverged from one ancestral population at $4N_et$ 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

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 since admixture	g	10
The number of individuals to infer	n_i	10
Frequency-based pruning threshold	d_f	0.5

https://doi.org/10.1371/journal.pcbi.1008065.t001

generations. We model recombination rate variation using the local recombination estimates from the 1000 Genomes Project [9]. The hotspot maps of the 22 human autosomal chromosomes are concatenated for a single string of 3 × 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. For the default setting, the approximate total number of SNPs simulated by macs is $\sim 14.7M$. Here we apply frequency-based pruning to trim data. 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 ∼26, 000 SNPs. In some cases, the default simulated length $L = 3 \times 10^9$ bp results in high computational burden. 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.

Default settings are chosen to be some reasonable values. For example, for g (the number of generations since admixture), ten generations roughly correspond to 200 to 300 years in humans, where there are some significant events related to admixture in human history within this time frame. We acknowledge that the choice of parameters can affect the accuracy. To explore this issue further, we have made additional simulations, where we show that simulation parameters may indeed affect accuracy. For example, when g increases, accuracy gets somewhat better.

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 K^{th} 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 is defined by Eq 11. Without loss of generality, we only consider the estimate of the proportions of the ancestral population A. Because we assume two ancestral populations, the expected mean errors of admixture proportions for two ancestral populations are identical.

$$\text{Mean error} \;= \frac{1}{n \cdot 2^K}\left(\sum_{1 \le j \le n} \sum_{1 \le i \le 2^K} |\widehat{m^i_j} - m^i_j|\right) \tag{11}$$

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 for ancestors. We address this problem using a "best-match" procedure by rotating the parents for each internal node in pedigree to find the best match between the inferred and the simulated admixture proportions. For example, for inference in parents, we have true admixture proportions for two parents (m^1, m^2) and estimated admixture proportions $(\widehat{m^1}, \widehat{m^2})$. We match both $(\widehat{m^1}, \widehat{m^2})$ and $(\widehat{m^2}, \widehat{m^1})$ to (m^1, m^2) and choose the one with smaller mean error. 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. Fig 5 shows the mean error when inferring admixture proportions of parents, grandparents, and great grandparents under the default simulation settings (as shown in Table 1). We compare the performance of PedMix to what is expected from random guess based on a Bayesian model. The random guess is described in

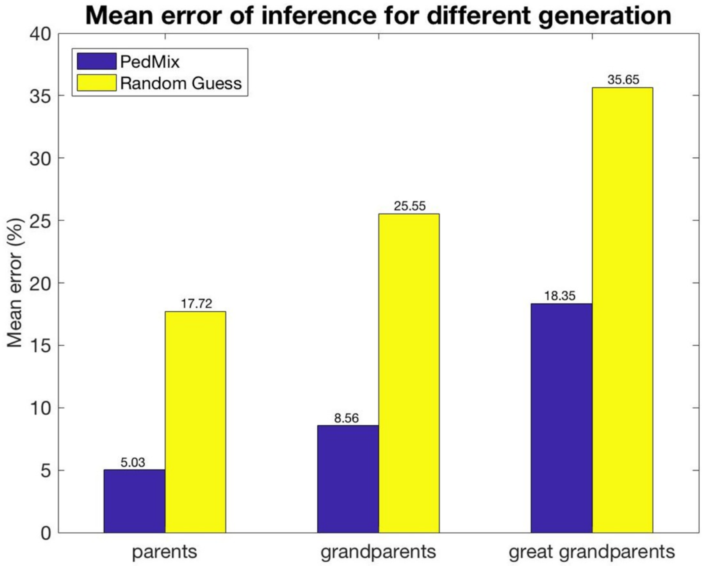

Fig 5. Comparison between the accuracy of PedMix and a random guess for parents, grandparents and great grandparents. About 570,000 SNPs are used for parent and grandparent simulations. About 26,000 SNPs are used for great grandparent simulations since this case needs more computing resources.

https://doi.org/10.1371/journal.pcbi.1008065.g005

Section S1.3 of S1 Text. We sample 10 genotypes from simulated admixed population for g ($g \geq 3$) generations.

Note that 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 [7] that infers admixture proportions of parental genomic ancestries given ancestry of a focal individual. And there are many methods (e.g., ADMIXTURE [6] and RFMix [20]) 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, we would expect the average of these admixture proportions of ancestors can serve as a good approximation for the focal individual. And this average should be approximately as accurate as the admixture proportions inferred by RFMix and ADMIXTURE. This is verified with simulation data in Section S1.6.1.1 of S1 Text. We note 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

Table 2. Mean and standard deviation of the error in the estimate of admixture proportions for ADMIXTURE, RFMix and PedMix (in units of %). Note the PedMix results are the average proportions from the estimated admixture proportions of parents (denoted as par.) or grandparents (denoted as grandpar.).

Error (in %)	ADMIXTURE	RFMix	PedMix (par.)	PedMix (grandpar.)
mean	1.16	1.77	2.01	3.53
standard deviation	1.43	1.52	2.5	2.53

https://doi.org/10.1371/journal.pcbi.1008065.t002

the inferred admixture proportions of the ancestors, this may suggest that the admixture proportion estimates for the ancestors also are not accurate.

We randomly sample 20 individuals from an admixed population and run ADMIXTURE, RFMix and PedMix on the same datasets. The genotypes are preprocessed with LD pruning (see Section S1.1.2 of S1 Text) and contain phasing errors simulated with rate $p_p = 0.00002$ per *bp*. We deduce the ancestry of each individual 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. More details on how ADMIXTURE and RFMix are applied are given in Section S1.6.1.2 of S1 Text. Table 2 shows the mean error as defined in Eq 11 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 by parents matches (within one standard deviation) the results of RFMix and ADMIXTURE. The estimate by RFMix and ADMIXTURE is slightly better than the estimate by grandparents. Note that the simulation setting tested here is only one out of many possible simulation settings.

We further compare the estimates of admixture proportions of parents from PedMix to those from ANCESTOR. ANCESTOR requires that the ancestry states and the tract lengths are provided as input for the algorithm. To accomplish this, we use the ancestry tracts inferred by RFMix when running ANCESTOR. More details on how ANCESTOR is applied are given in Section S1.6.1.2 of S1 Text. 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 (approx. 6.5% error versus 9.6% error).

Impact of simulation parameters. 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, we use the default setting with a shorter genome of length $L = 5 \times 10^8$ (Table 1) to reduce the computational time (Fig 6). 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 (Fig 6A). 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 (Fig 6B, and also Fig D of S1 Text). The mean error for both parental and grandparental inferences decreases as

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

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

https://doi.org/10.1371/journal.pcbi.1008065.t003

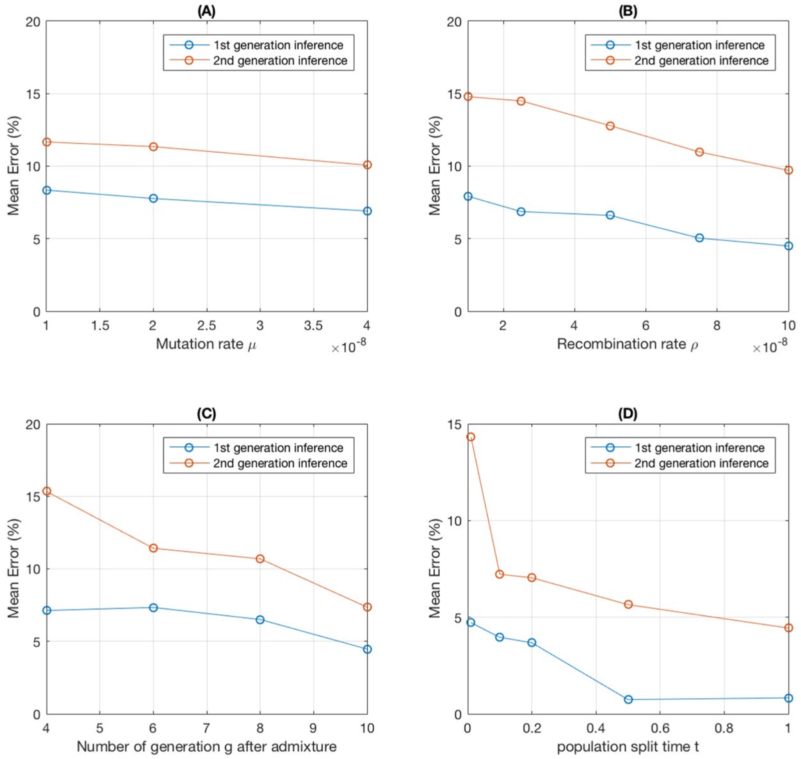

Fig 6. Mean error for different simulation parameter settings. (A) Varying mutation rates. $L = 5 \times 10^8$. (B) Varying recombination rates. $L = 5 \times 10^8$. (C) Varying the time since admixture. (D) Ancestral population split time. For $t = 0.01$, there are no SNPs left after using the $d_f = 0.5$ 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.

https://doi.org/10.1371/journal.pcbi.1008065.g006

recombination rate increases. When the recombination rate increases to more than 5×10^{-8}, the improvement in accuracy becomes smaller, especially for parental inference (also see Section S1.6.3 of S1 Text). 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 (Fig 6C). There is also a strong effect of t on the accuracy. As t increases, allele frequency differences between the admixing populations increase and it becomes easier to distinguish admixture tracts from two ancestral populations (Fig 6D). When $t > 0.5$ the mean error for parental inferences drops to below 1%.

Phasing error. Real data may contain phasing error. We have implemented a preprocessing approach for reducing the phasing error. See Section S1.2 of S1 Text for details. To evaluate the effect of phasing error, we simulate data with the phasing error rate 2×10^{-5} per bp. 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 phasing error rate 2×10^{-5}, we run

PedMix directly without preprocessing. Then we use the technique described in Section S1.2 of S1 Text to preprocess the data.

As shown in Fig 7A, 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 the 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, we sample 10 individuals from the same population and run PedMix to infer their admixture proportions. The performance of PedMix is very stable as shown in Fig 7B, 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 we decrease the effect of the phasing error in the data using preprocessing, the inference error decreases significantly.

Results on semi-simulated data

We now show results on semi-simulated data. Here, we use genotypes of CEU/YRI/ASW populations from the 1000 Genomes Project as the founders of a fixed pedigree topology as shown in Fig 8. Here, ASW refers to Americans of African Ancestry in SW USA. This way, the genotypes are closer to the real data and we know the origin of the founders. For this pedigree of two generations, we select four genotypes from one or more populations among CEU, YRI and ASW populations as grandparents. We assume there is no phasing error along these grandparental genomes. Then we simulate two genotypes as parents and one genotype as the focal individual based on the pedigree with recombination. 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. We run PedMix on the genotype of the focal individual genotype with or without phasing error to infer the admixture proportions of parents and grandparents. RFMix is run to estimate the admixture proportions

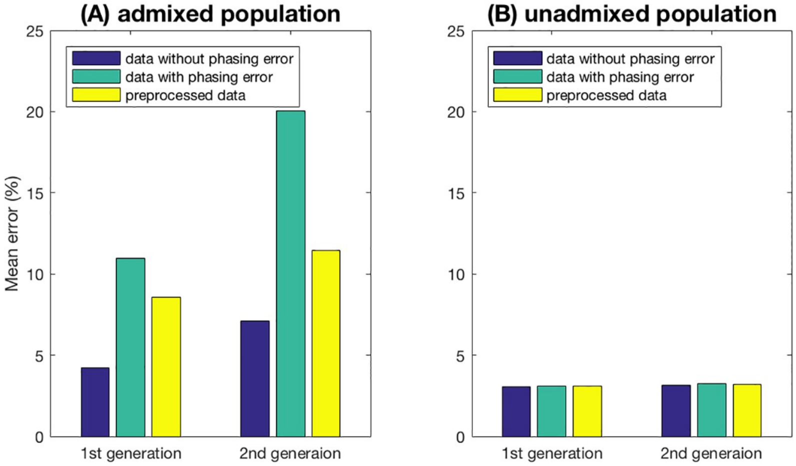

Fig 7. Inference error vs phasing error: Comparison among three different datasets. These include: dataset without phasing errors, dataset with phasing errors and preprocessed datasets. (A) Samples from an admixed population. (B) Samples from an unadmixed population.

https://doi.org/10.1371/journal.pcbi.1008065.g007

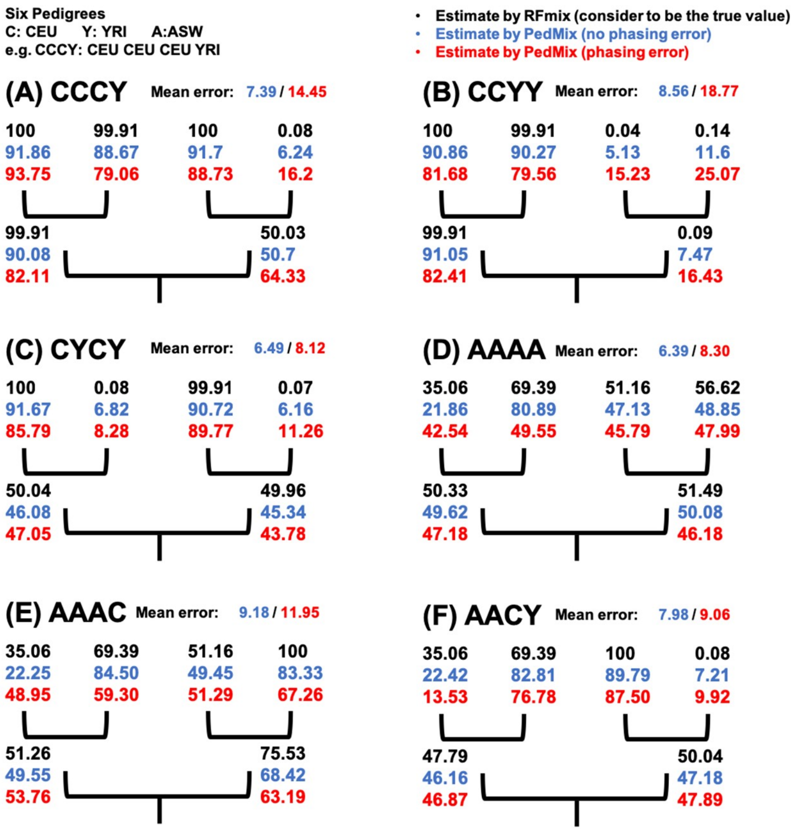

Fig 8. Six pedigrees used for the simulations scheme of semi-simulated data. The percentage of CEU origin is shown in the pedigrees. Admixture proportions in black are estimates by RFMix using the ancestors' genotypes directly and are assumed to be the true proportions. Admixture proportions in blue and red are estimates by PedMix using genotypes of the focal descendant individual with (red) or without (blue) phasing error. Mean error is the average over two parents and four grandparents.

https://doi.org/10.1371/journal.pcbi.1008065.g008

of parents (respectively grandparents) using the genotypes of parents (respectively grandparents). We use the estimates from RFMix with the ancestral genotypes as the ground truth on the admixture proportions of ancestors. This is because these ancestors are from real data and their true admixture proportions are not known. Here we examine 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). As an example, CCCY stands for the four grandparents from CEU, CEU, CEU and YRI respectively. Fig 8 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 99.91% and 0.09% for CCYY and 50.04% and 49.96% 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 99.91% and 0.09%. This is because one parent has two CEU grandparents and thus this parent is about 100% CEU. Similarly, the other parent is around 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. Our results indicate that phasing errors can indeed lead to larger inference error for some cases. Thus, it is useful to use haplotypes with less phasing errors. Estimates for parents are more accurate than those of grandparents.

Results on real data

To evaluate the performance of PedMix on real genetic data, we run PedMix on the haplotypes of ten trios from the ASW population from the HapMap data [21]. Phased genotypes of the parents are available from parents, while children's genotypes are unphased. We use the program Beagle [22] to phase children's genotypes from the phased haplotypes of the parents. Note that the exact ancestry of parents and grandparents in these trios are not known. Here, for parents, we use the inferred admixture proportions by running RFMix on the given haplotypes of parents as the true admixture proportions of the parents. The case of grandparents is more difficult because genetic data of grandparents is not available from the HapMap project. In order to evaluate the performance on grandparents, we adopt the following indirect approach: for each trio, we run PedMix on the child's haplotypes to infer grandparents' ancestry; we also run PedMix on the parents' haplotypes to infer grandparents' ancestry; we then examine whether the two inference results are consistent. That is, we view the inferred grandparents' admixture proportions from parents as proxy of the true grandparental proportions. While this is not a direct evaluation, we believe it at least offers some hints on how well grandparental inference is likely to perform on real data.

We use the YRI and CEU populations as the two ancestral populations. Allele frequencies of these two ancestral populations and genetic maps from the HapMap project are used. We preprocess the data using the frequency-based trimming (default threshold is 0.4). We apply the phasing error removal procedure as described in the Methods Section. We use the reduced phasing error rate (see Section S1.2 of S1 Text for details) when running PedMix. Fig 9 shows the average admixture proportion inference error for parents and grandparents over ten trios. We show results for individual trios and also the average error. The average inference error for

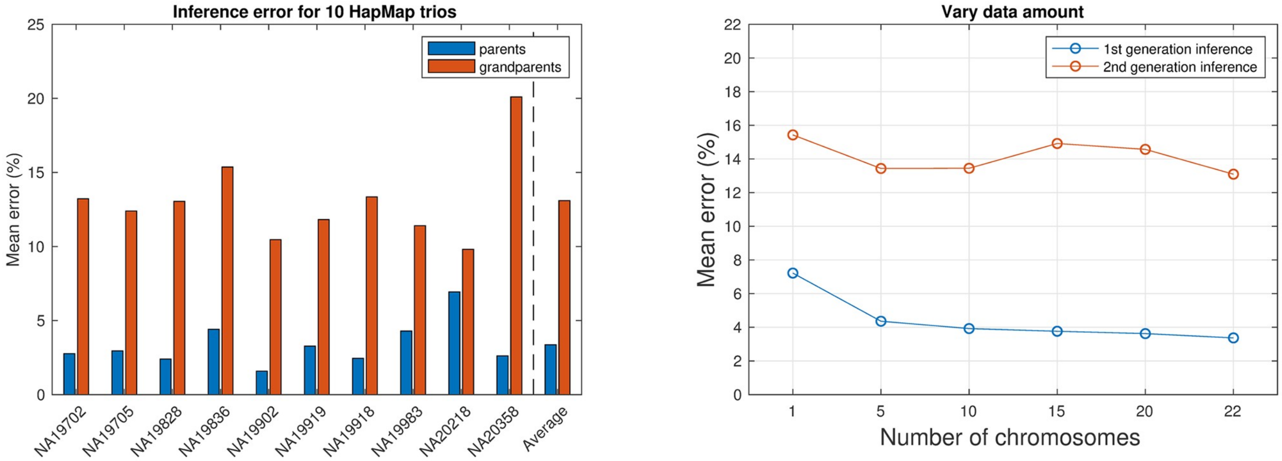

Fig 9. Admixture proportion inference error for parents and grandparents for ten HapMap trios from the ASW population. Left: inference error for ten HapMap trios individually or on average. Right: inference error with different amounts of data. X-axis: the number of chromosomes used. Y-axis: parental/grandparental inference error. Parental: the difference between PedMix's parental inference for the child and RFMix for the parents' haplotypes. Grandparental: the difference between PedMix's grandparental inference for the child and PedMix's parental inference for the parents' haplotypes in the trio.

https://doi.org/10.1371/journal.pcbi.1008065.g009

parents is about 3.4%, while the difference between the two inference results for grandparents is about 13%. Accuracy for most trios is relatively small, especially for parental inference. But for grandparent inference, the difference between the two estimates is relatively large for some trios. Fig 9 shows the parental/grandparental inference error with different amount of data (in terms of the number of chromosomes used). Our results show that inference error decreases when the amount of data increases, although sometimes the decrease in inference error is not very large. Overall, our results show that PedMix performs reasonably well on the real data.

We also compare PedMix with ANCESTOR on real data. The results are given in Section S1.6.7 of S1 Text.

Discussion

In this paper we develop a method for the 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 to use the distribution of admixture tracts which is influenced by the ancestral admixture proportions. Our theoretical analysis shows that treating SNPs as independent sites is insufficient for inferring ancestral admixture proportions in general. In fact, it is unidentifiable to use single SNPs to infer admixture proportions of ancestors. See Section S1.4 of S1 Text for details. Our method uses a pedigree model, which is a reasonable model for recent genealogical history of a single individual. Some previous approaches for inferring ancestry use additional information, for example geospatial information [23, 24]. Our approach only uses the genetic data from a single individual (so does [7]).

Existing admixture inference methods such as ADMIXTURE and RFMix (also see [25]) infer local ancestry for the focal individuals. However, these local ancestry estimates cannot be directly applied for the inference of recent ancestors. Note that there are 22 unlinked human chromosomes, each of which will on average experience one or two cross-overs in each meiosis. Analyses using methods such as RFMix will allow ancestry inferences of each of these chromosomal segments in the genome. However, it is not clear which parent each of the two haplotypes comes from; segments are inferred with respect to population, not with respect to parent. Therefore, ancestry paintings do not directly provide admixture proportions for parents. The joint transmission probabilities over all the segments in the pedigree need to be taken into account modeling the possibility of back-recombination over multiple meiosis, as well as phasing error. This is exactly what PedMix achieves.

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 burden increases. This can be seen from Fig 5, where the inference error for great grandparents is significantly higher than those for parents or grandparents. On the other hand, Fig 5 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 comparisons with 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 only infer the admixture proportions of the focal individuals.

Inferences by PedMix are affected by the assumptions of the underlying population genetic processes (Fig 6). 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 large. Increasing the chromosome length has similar effect on inference accuracy as increasing the recombination rate. Moreover, noises in population genetic data may also affect inference accuracy. In Section S1.6.6 of S1 Text, significant noise in genetic maps, allele frequencies of reference populations, or inappropriate reference populations can indeed reduce inference accuracy.

In line with other studies (e.g. [26]), we find that pruning of SNPs and preprocessing to remove potential phasing errors is critical for obtaining reasonably accurate results. In Section S1.6.2 of S1 Text, we compared 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, 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. Fig 7 shows that if untreated, phasing error can indeed greatly increase the inference error of PedMix. On the other hand, when we apply preprocessing to remove the obvious phasing errors, inference errors can be significantly reduced. Nonetheless, phasing error can still reduce inference accuracy. We note that phasing methods are constantly improving and the problem of phasing errors may be greatly reduced in the near future.

Simulation shows that PedMix can scale to whole genome data, when proper data preprocessing is performed. The current implementation of PedMix assumes two ancestral populations. In principle, PedMix can be extended to allow more than two ancestral populations. However, this may lead to increased computational time. Inference with more than two ancestral populations may also lower inference accuracy because the underlying model becomes more complex.

Supporting information

S1 Text. Due to the space limit, we place parts of our results in the supplemental materials. These include additional methods, additional results and also a theoretical result which show that the ancestry inference problem studied in this paper will be infeasible without using linkage disequilibrium (LD).
(PDF)

Author Contributions

Conceptualization: Rasmus Nielsen, Yufeng Wu.

Data curation: Jingwen Pei.

Formal analysis: Jingwen Pei, Yiming Zhang.

Funding acquisition: Yufeng Wu.

Investigation: Jingwen Pei, Rasmus Nielsen, Yufeng Wu.

Methodology: Jingwen Pei, Yiming Zhang, Rasmus Nielsen, Yufeng Wu.

Project administration: Rasmus Nielsen, Yufeng Wu.

Software: Jingwen Pei, Yiming Zhang, Yufeng Wu.

Validation: Jingwen Pei, Yiming Zhang.

Writing – original draft: Jingwen Pei, Rasmus Nielsen, Yufeng Wu.

Writing – review & editing: Jingwen Pei, Yiming Zhang, Rasmus Nielsen, Yufeng Wu.

References

1. Nordgren A. Genetics and Identity. Community Genet. 2008; 11:252–266. https://doi.org/10.1159/000121396
2. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science. 2002; 298:2381–2385. https://doi.org/10.1126/science.1078311 PMID: 12493913
3. Li J, Absher D, Tang H, Southwick A, Casto A, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008; 319:1100–1104. https://doi.org/10.1126/science.1153717 PMID: 18292342
4. Pritchard JP, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 2000; 155:945–959.
5. Tang H, Peng J, Wang P, Risch N. Estimation of individual admixture: analytical and study design considerations. Genetic epidemiology. 2005; 28:289–301. https://doi.org/10.1002/gepi.20064
6. Alexander D, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009; 19:1655–1664. https://doi.org/10.1101/gr.094052.109
7. Zou JY, Halperin E, Burchard E, Sankararaman S. Inferring parental genomic ancestries using pooled semi-Markov processes. Bioinformatics. 2015; 31:i190–6. https://doi.org/10.1093/bioinformatics/btv239
8. Gravel S. Population genetics models of local ancestry. Genome research. 2012; 191:607–619.
9. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526:64–74.
10. Liang M, Nielsen R. The Lengths of Admixture Tracts. Genetics. 2014; 197:953–967. https://doi.org/10.1534/genetics.114.162362
11. Tang H, Coram M, Wang P, Zhu X, Risch N. Reconstructing Genetic Ancestry Blocks in Admixed Individuals. American Journal of Human Genetics. 2006; 79:1–12. https://doi.org/10.1086/504302
12. Price A, Tandon A, Patterson N, Barnes K, Rafaels N, Ruczinski I, et al. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS Genetics. 2009; 5:e1000519. https://doi.org/10.1371/journal.pgen.1000519 PMID: 19543370
13. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. American Journal of Human Genetics. 2008; 82:290–303. https://doi.org/10.1016/j.ajhg.2007.09.022
14. Paşaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. Bioinformatics. 2009; 25:i213–i221. https://doi.org/10.1093/bioinformatics/btp197
15. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc of the Nat Academy of Science. 1987; 84:2363–2367. https://doi.org/10.1073/pnas.84.8.2363
16. Idury R, Elston R. A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. Hum Hered. 1997; 47:197–202. https://doi.org/10.1159/000154413
17. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical Programming. 1989; 45:503. https://doi.org/10.1007/BF01589116
18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81:559–575. https://doi.org/10.1086/519795 PMID: 17701901
19. Chen G, Marjoram P, Wall J. Fast and flexible simulation of DNA sequence data. Genome research. 2009; 19:136–142. https://doi.org/10.1101/gr.083634.108
20. Maples B, S G, Kenny E, Bustamante C. RFmix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics. 2013; 93:278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

21. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. https://doi.org/10.1038/nature09298

22. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. American Journal of Human Genetics. 2007; 81:1084–1097. https://doi.org/10.1086/521987

23. Yang WY, Platt A, Chiang CK, Eskin E, Novembre J, Pasaniuc B. Spatial Localization of Recent Ancestors for Admixed Individuals. G3: Genes, Genomes, Genetics. 2014; 4:2505–2518. https://doi.org/10.1534/g3.114.014274

24. Margalit Y, Baran Y, Halperin E. Multiple-Ancestor Localization for Recently Admixed Individuals. In: Algorithms in Bioinformatics. WABI 2015; 2015. p. 121–135.

25. Salter-Townshend M, Myers S. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. Genetics. 2019; 212:869–889. https://doi.org/10.1534/genetics.119.302139

26. Anderson C, Pettersson F, Clarke G, Cardon L, Morris A, Zondervan K. Data quality control in genetic case-control association studies. Nature protocols. 2010; 5:1564. https://doi.org/10.1038/nprot.2010.116