PNAS RESEARCH ARTICLE | EVOLUTION

A general approach for inferring the ancestry of recent ancestors of an admixed individual

Yiming Zhang[a], Haotian Zhang[a], and Yufeng Wu[a,1]

Edited by Marcus Feldman, Stanford University, Stanford, CA; received September 18, 2023; accepted November 27, 2023

The genome of an individual from an admixed population consists of segments originated from different ancestral populations. Most existing ancestry inference approaches focus on calling these segments for the extant individual. In this paper, we present a general ancestry inference approach for inferring recent ancestors from an extant genome. Given the genome of an individual from a recently admixed population, our method can estimate the proportions of the genomes of the recent ancestors of this individual that originated from some ancestral populations. The key step of our method is the inference of ancestors (called founders) right after the formation of an admixed population. The inferred founders can then be used to infer the ancestry of recent ancestors of an extant individual. Our method is implemented in a computer program called PedMix2. To the best of our knowledge, there is no existing method that can practically infer ancestors beyond grandparents from an extant individual's genome. Results on both simulated and real data show that PedMix2 performs well in ancestry inference.

genetics | genetic tests | ancestry inference | population admixture | recombination

Significance

Ancestry inference has been performed for millions of customers in commercial genetic tests. A popular ancestry test offered by these tests provides the percentages of an extant genome that are from different ancestral populations. Most existing approaches focus on inferring the ancestry of an extant individual. Now, what if someone wants to know more about his/her ancestry by asking: what are the ancestries of my parents, grandparents, great-grandparents, etc.? In this paper, we present an inference approach that can estimate the ancestries of recent ancestors of an extant individual from this individual's genome. Our method can provide estimates of ancestry for much more distantly related ancestors than existing methods. Therefore, our approach is more general and applicable than existing methods.

Ancestry inference from individual genomes has become a major component of commercial genetic tests offered by companies including 23andMe and Ancestry.com. These tests often produce reports about the ancestral origin of the genome under test. One of the most popular reports is about admixture inference, which is usually in the form of the so-called "chromosome painting." The concept of chromosome painting can be illustrated by considering an individual from an admixed population formed by two or more ancestral populations. The genome of this individual can be divided into segments, where each segment originated from an ancestral population. If we assign a distinct color to each ancestral population, the genome can be colored (i.e., painted) based on these segments. The painted chromosome can be used to calculate quantities about ancestry, e.g., admixture proportion: the percentage of the genome that originated from a specific ancestral population (i.e., painted in a specific color). In practice, segments are not directly observable and need to be inferred. Therefore, chromosome painting is a computational problem that aims at inferring the segments given the extant genome, along with other population genetic data (e.g., allele frequencies) about the relevant ancestral populations. The basic concept of chromosome painting is behind the seminal STRUCTURE paper (1). Chromosome painting has been actively studied recently. There exist several computational methods (e.g., refs. 2–7) for performing chromosome painting of extant genomes.

Chromosome painting has been performed extensively in consumer genetics. A natural question is whether more information can be obtained from an extant genome. Most existing approaches for admixture inference from extant genomes focus on extant individuals and do not provide much information about the recent ancestors of extant individuals. An emerging research problem on admixture inference is the inference of recent ancestors of an extant individual from the genome of this individual. One of the first approaches for this kind of ancestry inference is PedMix (8) (see also refs. 9 and 10). Unlike existing admixture inference approaches, PedMix aims at inferring the ancestry of parents or grandparents of an extant individual (not the extant individual him/herself) from an extant genome. For example, suppose we are given the genome of an admixed individual with ancestry from two ancestral populations A and B. PedMix can be used to answer questions such as "Are the two parents of this individual both 50 to 50 admixed, or is one 100% A and the other 100% B?". Simulation shows PedMix can provide a reasonably accurate estimate (with some variance) of admixture proportions of parents and grandparents (8). One major downside of PedMix is that it can only infer the ancestries of parents and grandparents.

Author affiliations: [a]School of Computing, College of Engineering, University of Connecticut, Storrs, CT 06269

Author contributions: Y.Z. and Y.W. designed research; Y.Z., H.Z., and Y.W. performed research; Y.Z. and H.Z. analyzed data; and Y.Z., H.Z., and Y.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: yufeng.wu@uconn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2316242120/-/DCSupplemental.

Published January 2, 2024.

Another approach called PAPI (10) also aims at inferring parental admixture proportions. One feature of PAPI is that it can infer the time (the number of generations) since admixture. However, PAPI can only infer parental admixture. More recently, we developed another ancestry inference approach called parMix (11), which can perform chromosome painting for parents given genomes from multiple children. A main limit of parMix is that parMix needs to have two or more children of the same parents. Also, parMix can only work for parental inference.

Since PedMix can only estimate admixture proportions of parents and grandparents of an extant individual, a natural research question is inferring aspects of admixture for more distantly related ancestors. Simple genetics principles indicate that when only an extant genome is given, admixture inference of distant ancestors of this individual does not seem easy. Suppose we want to know the admixture proportions of ancestors ten generations ago, where there are up to 1,024 distinct ancestors. Recall that we only have one extant genome. The extant individual inherits $\frac{1}{1024}$ genome of each of these 1,024 ancestors on average. That is, only a tiny fraction of the genomes of the 1,024 distantly related ancestors can be found in the extant genome. Moreover, these ancestors may be admixed themselves, so different regions of an ancestor may originate from different ancestral populations. When only one extant genome is given, it seems unlikely that we can obtain any meaningful estimates of the ancestries of distantly related ancestors. Therefore, it is unsurprising that no methods exist in the literature (to the best of our knowledge) that can perform admixture inference of distant ancestors from a single extant genome. The existing method that is closest to performing such inference is PedMix. PedMix is based on a hidden Markov model built on the genetic process's parameters, including states of ancestry and recombination in recent ancestors (e.g., parents and grandparents) at a specific site. But PedMix is not applicable to more distantly related ancestors due to computational complexity.

As it turns out, it is a little surprising that estimating admixture proportions of recent ancestors (e.g., parents) of an extant individual can be obtained by considering ancestors that are much more distantly related to this extant individual than parents. We consider a particular type of ancestor: founding ancestors. Founding ancestors (or simply founders) are the ancestors of an extant individual at the time of population admixture. That is, founders are individuals who originated from an ancestral population. A simple but important fact is that founders are not admixed. Different from ancestors who are not founders, each genomic region of a founder originated from the same ancestral population. Now suppose we know which founder originated from which ancestral population. Why are these founders helpful in knowing the ancestry of recent ancestors? We denote the fraction of all founders who are from a specific ancestral population A and are ancestral to an individual as the founder ratio of this individual for A. The founder ratio is related (but not identical) to admixture proportion. A crucial empirical observation is that as the number of single nucleotide polymorphism (SNP) sites increases, the founder ratio and the admixture proportion of an individual converge. That is, founders can provide an estimate of admixture proportions of an individual if the ancestry of the founders for this individual is known.

Founders are not directly observable and need to be inferred from an extant genome. Assume that an admixed population was founded g generations ago; then, each ancestor of an extant individual from this population at g generations ago is assumed to be a founder. This is due to the standard Wright–Fisher model. We denote the ancestry of founders (i.e., a list of ancestral populations, one for each founder, from which a founder originated) as founder configuration (or simply configuration). Founder configuration inference from an extant genome is more straightforward than the inference of non-founders. Intuitively, one may be able to infer the ancestry origin of a founder from a small segment of the extant genome, assuming it is known this segment is from this founder. Recall that different regions of founders have the same ancestry. An important aspect here is meiotic recombination. Due to recombination, an extant genome consists of segments from multiple founders. Since an admixed population is usually formed relatively recently (i.e., g is relatively small), we expect a relatively long segment of a founder's genome to be passed to an extant individual. Therefore, an extant genome, in principle, allows ancestry inference of multiple (possibly all) founders. However, due to the stochastic nature of recombination, a rigorous method is needed for ancestry inference.

In this paper, we present a method for ancestry inference of recent ancestors of an extant genome. A key step of this method is inferring the founder configuration from an extant genome. The inference of founder configuration is non-trivial. First, it is unknown which part of an extant genome is from which founder since recombination is stochastic. Moreover, the number of possible founder configurations can be vast. Suppose there are two ancestral populations. Then there are $2^{1,024}$ (which is astronomically large) possible founder configurations ten generations ago. A main contribution of this paper is that we show that our method can infer founder configuration reasonably accurately (with some variance). Another significant contribution of our method is that it can, in principle, provide estimates of admixture proportions for all recent ancestors. Thus, it is a more general ancestry inference approach than existing methods (e.g., PedMix). Also, the method uses a computational approach that is fundamentally different from that of PedMix. We have implemented this method in the program PedMix2. Given an extant genome and also relevant population genetic information about ancestral populations (including allele frequencies and recombination fractions), PedMix2 can perform the following inference.

1. Founder configuration inference: infer from which ancestral population each founder originated. Our experience shows that PedMix2 can infer founder configuration with reasonable accuracy (with some variance).
2. Estimating ancestral proportions of recent ancestors: estimate the percentage of alleles of all ancestors (at certain generations ago) of an extant individual that is from a specific ancestral population. While existing tools (e.g., PedMix) can perform inference for parents or grandparents, our simulation shows that PedMix2 can give more accurate estimates of admixture proportions for parents and grandparents than PedMix. Moreover, PedMix2 can, in principle, provide admixture proportion estimate for all ancestors up to the founders. Note that accuracy may be lower for estimates of more distantly related ancestors.

PedMix2 is available for download at https://github.com/biotoolscoders/PedMix2.

Results

The main result of this paper is an ancestry inference approach. In the following, we first describe this approach. Some details of this method are provided in *SI Appendix, Text*. We then provide empirical results of our method on simulated and real data.

Inference of Founder Configuration: Concepts and Assumptions. We first give an introduction to the perfect pedigree model, which is the foundation of PedMix (8) and also our method. Suppose we have a haplotype from an extant individual who is from an admixed population formed by K ancestral populations. Throughout the paper, we assume haplotypes are phased properly without phasing errors (see the *Discussion* section for the issue of phasing errors). An allele of this haplotype is passed from one of the two parents, which in turn is passed from their parents. We trace this allele until the founding generation of this admixed population when the population was formed by the admixture of K ancestral populations. This genetic process leads to a pedigree, where an extant allele follows a path (called inheritance path) from the extant haplotype to one of the founders of the pedigree. Here, the founders of the pedigree are the haplotypes (called founding haplotypes) from founders of this extant individual at the time (g generations ago) of population admixture. Recall that founding haplotypes are not admixed. Due to meiotic recombination, different alleles of a haplotype may come from different parents. That is, different alleles of the extant haplotype can take different inheritance paths in the pedigree. By the standard Wright–Fisher model, all founding haplotypes of the pedigree are at g generations ago. That is, an inheritance path in the pedigree from the (single) extant haplotype to a founding haplotype always has the same length (g generations). This pedigree is called the perfect pedigree (8, 12). A founder configuration C can be represented as a length-2^g vector, where each value of C is a population label representing from which ancestral population one founder originated. See Fig. 1 for an illustration of the perfect pedigree model and founder configuration.

Founder configuration is an essential aspect of the genetic ancestry of an extant haplotype. A central computational problem addressed in this paper is that given a set of (un-linked) haplotypes from an extant admixed individual along with relevant genetic information about ancestral populations, infer the founder configuration of this individual. We assume the following information is given:

1. A set of haplotypes from an extant individual, which are assumed to be phased correctly. Haplotypes from different chromosomes are assumed to be un-linked.
2. Allele frequencies of ancestral populations and recombination fractions between single nucleotide polymorphisms (SNPs).
3. g: the number of generations since admixture. We assume population admixture is relatively recent: g is assumed to be less than 15.

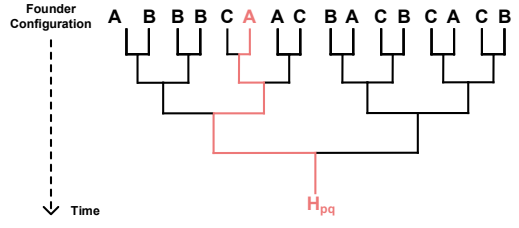

Fig. 1. Example of a perfect pedigree with $K = 3$ ancestral populations with four generations ($g = 4$). The founders are on the top. At each generation (viewing upward), the number of ancestors doubles. The red lines show an inheritance path for a site. A single extant haplotype H_{pq} is at the bottom. Each founder is labeled by one of three ancestral populations (A, B, and C) that form the admixed population. The list of these labels forms the founder configuration.

The High-Level Approach. In order to infer the founder configuration from the given extant haplotypes, we consider the inheritance paths that determine the relationships between founders and extant haplotypes. An inheritance path specifies the sequence of ancestors (from a founder down to the extant individual) along which ancestral alleles pass along at a specific site. See Fig. 1 for an illustration of the inheritance path. The inheritance path cannot be fully determined from an extant genome due to the stochastic nature of recombination. To infer the founder configuration C from a set of extant haplotypes $\mathcal{H}$, we want to compute the likelihood $P(\mathcal{H}|C)$ of the extant haplotypes $\mathcal{H}$. Recall that due to recombination, two nearby SNPs may follow different inheritance paths. We denote the set of inheritance paths at each SNP site as an inheritance path set. This likelihood can be viewed as a sum of probabilities over all possible inheritance path sets Π (which specifies an inheritance path $\pi \in \Pi$ for each SNP position):

$$P(\mathcal{H}|C) = \prod_{H\in\mathcal{H}} \sum_{\pi\in\Pi} P(H|\pi, C)P(\pi). \quad [1]$$

In principle, we can infer C by maximizing the likelihood computed by Eq. **1**. However, direct computation of the likelihood $P(\mathcal{H}|C)$ using Eq. **1** is infeasible even for moderate-sized data since the size of Π can be very large. We use several techniques to make the likelihood computation practical. First, we divide a haplotype into a relatively small number of blocks where recombinations are assumed to occur only between blocks. The key component of our method is an efficient algorithm that calculates $P(H|C)$ approximately for a single haplotype H by integrating over all inheritance paths of blocks of H. This algorithm is practical for use in inference when the number of blocks n_b and g are modest (say $n_b \leq 50$ and $g \leq 15$). See *SI Appendix* for details.

Maximum likelihood estimate of founder configuration. Now, given a set of haplotypes $\mathcal{H}$, we infer the most likely founder configuration C by maximum likelihood:

$$C_{opt} = argmax_C \prod_{H\in\mathcal{H}} P(H|C). \quad [2]$$

It is not practical to enumerate all possible founder configurations C when g and K (the number of ancestral populations) are large: There are K^{2g} possible founder configurations. To find C_{opt} practically, we use a local search heuristic to find C_{opt}. The details of the local search are given in *SI Appendix*.

Ancestry Inference from the Inferred Founder Configuration. Suppose we have inferred the optimal founder configuration C_{opt} for an extant individual. We denote the percentage of founders from a specific ancestral population as the founder ratio of this individual for this population. Under the perfect pedigree model, the notion of the founder ratio can also be generalized to any ancestor (starting from the extant individual up to the founders) of this individual. This is because the perfect pedigree of this ancestor is contained inside the perfect pedigree of the extant individual as a sub-graph (called sub-pedigree). For example, we can break C_{opt} (the configuration for the entire pedigree) into two halves. Each half is for one parent of the extant individual. Then, we can calculate the founder ratios of the two parents from the two halves.

The founder ratio concerns the composition of founders of an (extant or not) individual. The founder ratio can be computed

when the founder configuration is known for this individual. Recall that admixture proportion is of main interest in ancestry inference. Admixture proportion concerns the composition of alleles of an individual (who is often an extant individual but can be an ancestor in the pedigree). While conceptually different, these two quantities are correlated. It is obvious that the more founders from a specific ancestral population (say A), the higher the admixture proportion of the extant individual for A tends to be, and vice versa. We now investigate the relationship between founder ratio and admixture proportion in a more formal way.

Suppose we focus on the extant individual (the cases of other individuals in the pedigree are almost the same). We fix the founder configuration in the perfect pedigree (and so the founder ratio for this individual is also fixed). Under this setting, the genome of the extant individual is composed of segments from founders. The process of the extant genome formulation is a stochastic process, where the randomness is due to meiotic recombination and also genetic inheritance (i.e., from which side of parents an un-linked allele originates). Admixture proportion is determined by this stochastic process within the pedigree. That is, an admixture proportion can be viewed as a random variable, which depends on the founder configuration, recombination, and inheritance choices. We have the following simple observation.

Proposition 1. *Admixture proportion is an unbiased estimator of founder ratio. That is, the expectation of admixture proportion is equal to founder ratio.*

Proof: Let C be the fixed founder configuration. Let $f(C)$ be the founder ratio for an ancestral population A (i.e., the fraction of founders from A) as specified by C. We consider a single site s. The probability $P_s(A)$ that s's allele originates from A is equal to $f(C)$. This is because s has a random inheritance path, and so s's allele is from a random founder which is from A with probability $f(C)$. We define an indicator variable $I_s(A)$ for a site s ($1 \le s \le n$) that s's allele is from A (i.e., $I_s(A) = 1$ if the site s has the ancestry A). Note that $E(I_s(A)) = P_s(A)$. Let $n(A)$ be the number of sites that are from A among data with n sites. Then, the expectation of $n(A)$ is:

$$E(n(A)) = E(\sum_{s=1}^{n} I_s(A)) = \sum_{s=1}^{n} E(I_s(A))$$
$$= \sum_{s=1}^{n} P_s(A) = \sum_{s=1}^{n} f(C) = nf(C).$$

The above follows by the well-known fact of linearity of expectations. Therefore, the expected admixture proportion is equal to $\frac{E(n(A))}{n} = \frac{nf(C)}{n} = f(C)$. □

Proposition 1 implies that founder ratio can be approximated by admixture proportion. But note that we already know the founder ratio for the individual (from the inferred founder configuration) and we want to estimate the admixture proportion of this individual instead. To estimate admixture proportion, we apply Proposition 1 in the opposite direction. That is, we use the founder ratio computed from the inferred configuration as an estimate of the admixture proportion. In the *Results*, we use simulation to validate that founder ratios and admixture proportions converge when the data size reaches a level similar to the whole genome in humans.

Since we have the inferred founder configuration C_{opt} for the entire pedigree, this provides estimates for admixture proportions for all the individuals in the pedigree, from the extant individual to founders. That is, for each ancestor in the pedigree, we use the sub-pedigree rooted at this ancestor to obtain the inferred configuration for this ancestor from C_{opt}, which allows us to estimate the admixture proportion for this ancestor. Note, however, a sub-pedigree rooted at an ancestor that is closer to founders has a smaller size. This can lead to higher bias in admixture proportion estimates for such ancestors.

The algorithms for founder configuration inference and admixture proportion estimate are implemented in the program PedMix2. We now evaluate the performance of PedMix2 on simulated and real data.

Empirical Results on Simulated Data. We use simulated data to evaluate the performance of PedMix2. We first simulate n_h haplotypes using msprime (13) from two ancestral populations which diverged from one ancestral population at $4N_e t$ generations in the past. Then, an admixed population is formed by admixing these two ancestral populations. This admixed population has $\frac{n_h}{2}$ diploid individuals (n_h haplotypes). That is, the admixture ratio is 0.5. We then simulate forward in time the genetic process starting from the time of admixture for additional g generations. The process includes random mating, genetic drift, and recombination using a diploid Wright–Fisher model. The varying recombination rates from the 1000 Genomes Project (14) are used in the simulation. The length of each chromosome is based on human data.

We assume that haplotypes are properly phased. The number of SNPs for the first chromosome simulated by *msprime* is ~149,000 under the default settings. We divide each chromosome into n_b blocks. Table 1 shows the parameters (with explanations and their default values) used in the simulations. Here, the number of blocks is due to the "block assumption" for efficient likelihood computation. See *SI Appendix* for details. For accuracy evaluation, we benchmark PedMix2 in several aspects.

Admixture proportion accuracy. Recall that we use the founder ratios calculated from the inferred founder configuration as the estimates of admixture proportions of recent ancestors. The founder ratio can be calculated from the inferred founder configuration as follows. For example, if the inferred configuration is "ABCBABCC," the ancestry ratio of population "A," "B," and "C" will be 25%, 37.5%, and 37.5%, respectively. The accuracy of admixture proportion estimates (or simply proportion accuracy) rate can be defined as follows.

Table 1. List of parameters and their default values in the simulation data generation and testing

Symbol	Default	Description
n_h	1,000	Number of haplotypes
N_e	10,000	Effective population size
t	0.15	Ancestral populations splitting time
μ	1×10^{-9}	Mutation rate (per generation per bp)
ρ	1×10^{-8}	Recombination rate (per generation per bp)
n_c	22	Number of chromosomes
g	9	Number of generations since admixture
ad_p	0.5	Admixture ratio
n_f	10	Number of individuals in testing for each setting
n_b	20	Number of blocks per chromosome

The top part shows the values in simulation data generation, and the bottom part shows the values in testing.

$$R_a = 1 - \frac{\sum_p |R_p - R_p^{'}|}{K}, \quad [3]$$

where R_p is the inferred admixture proportion of reference population p, and $R_p^{'}$ is the true admixture proportion of an ancestral population p. K is the number of ancestral populations.

Configuration accuracy. We evaluate the accuracy of inferred founder configurations with the true simulated founder configuration. The metric we use is called "configuration accuracy." Configuration accuracy is the accuracy of the inferred founder configuration C_{opt} compared with the true founder configuration C_0. Founder configuration has the symmetry property: Two configurations can be equivalent by "rotating" the pedigree (see *SI Appendix* for details). To compare two configurations with symmetry allowed, we apply the so-called "maximum accuracy algorithm" or MAA. The MAA is given in *SI Appendix*.

Comparing with existing methods. There are no existing methods for founder inference. There are existing methods that can infer admixture proportions of extant individuals and their recent ancestors (especially parents or grandparents). To evaluate the proportion accuracy of PedMix2, we compare it with the following alternative approaches that can estimate admixture proportions.

1. RFMix. We can use RFMix to estimate the admixture proportion of an individual whose genome is given. Note that RFMix is not applicable to individuals (e.g., ancestors) whose genomes are not given. RFMix uses "Random Forest" and "Decision Tree" methods to perform ancestry inference for the extant individual.
2. PedMix. PedMix can estimate admixture proportions of recent ancestors (i.e., parents or grandparents) of an extant individual. However, PedMix becomes very slow for ancestors that are more distant than grandparents. Moreover, the current PedMix implementation only allows two ancestral populations.

Empirical validation of the convergence of admixture proportion and founder ratio. A key aspect of PedMix2 is using the founder ratio to estimate the admixture proportion for an ancestor. One justification is Proposition 1, which shows that the expectation of admixture proportion is equal to founder ratio. We designed a large-scale experiment about admixture proportions and founder ratios to provide empirical justification for this approach. First, 100 individuals from two simulated populations are used as the founders of an admixed population. The admixture ratio ad_p is 0.5. Then, for the first chromosome, we simulate 100 individuals based on the Wright–Fisher model for 10 generations. This leads to a founder configuration for the 100 extant individuals. We then use the same ancestors in the pedigree to generate 99 more chromosomes for these 100 individuals.

We calculate the accumulated admixture proportions of the first k chromosomes. The absolute mean differences between the ground truth of admixture proportion for the first k chromosomes and the ground truth of founder ratio (which does not depend on k) are shown in Fig. 2*A*. We can see that the absolute mean difference decreases when more chromosomes are used. The difference is less than 1% if there are 100 chromosomes and 2.3% if there are 22 chromosomes (the whole genome for humans). Therefore, founder ratio can provide a reasonable estimate of admixture proportions when the data are large.

To benchmark the proportion accuracy of PedMix2, we use the estimated founder ratio as the admixture proportion of the extant individual for PedMix2. We then compare PedMix2's estimates of founder ratios and admixture proportions with the ground truth of founder ratio and admixture proportion of extant individuals. Note that the two estimates by PedMix2 are the same; the two comparisons are against different ground truth. We also compare RFMix's estimates of admixture proportions with the ground truth of admixture proportion. We vary the number of chromosomes from one to twenty-two (i.e., increase the data size). The results are shown in Fig. 2*B*. When data are small, founder ratio estimate is not very accurate. However, the estimated founder ratio can still serve as a good estimate of admixture proportion even with small data. This may be due to the large variance with small data. When data size increases, founder ratio estimate becomes more accurate. Overall, the founder ratio calculated by PedMix2 provides a good estimate of admixture proportion when data are large.

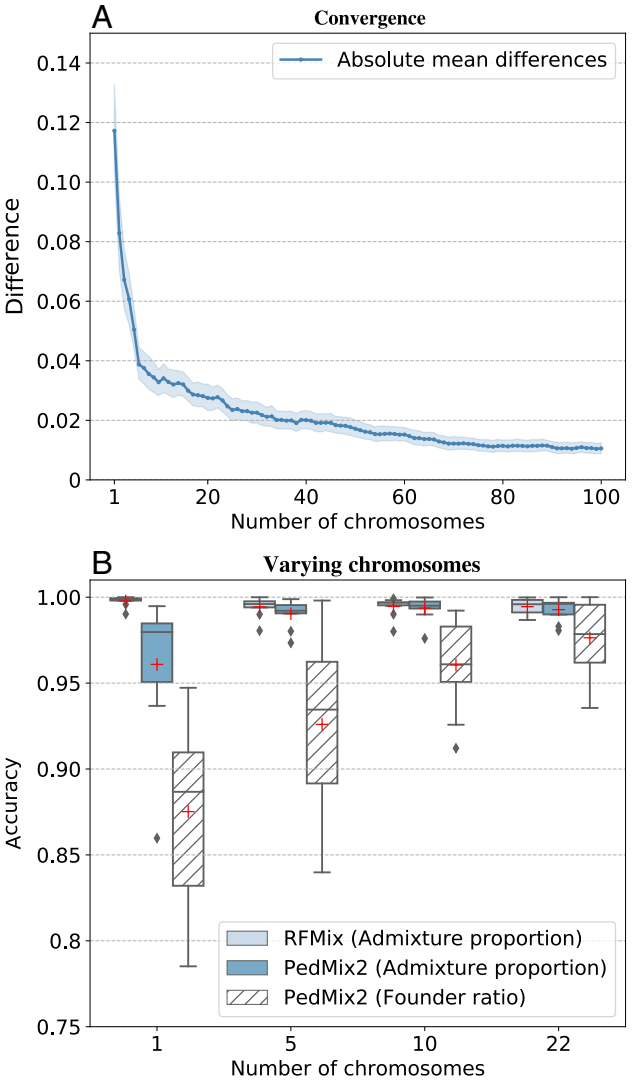

Fig. 2. Empirical investigation of the relationship between founder ratio and admixture proportion. Part 2 (*A*): absolute mean differences between the true admixture proportions and the true founder ratios with varying numbers of chromosomes (under 95% CI). The differences converge to 1% when there were 100 chromosomes, which provides empirical justification of Proposition 1. Part 2 (*B*): comparison of PedMix2 with RFMix in the accuracy of admixture proportion and founder ratio accuracy of PedMix2 with different numbers of chromosomes. The default settings in Table 1 are used. "red +": the mean value.

Proportion accuracy on simulated data. We compare PedMix2 with several existing methods on admixture proportion accuracy. RFMix can estimate admixture proportions for individuals with given genomes. PedMix can infer the admixture proportions of recent ancestors from an extant genome. PedMix2 can do both and also work for more distant ancestors (up to the founder generation). Therefore, we compare PedMix2 with RFMix for the extant individuals' inference. We compare PedMix2 with PedMix for both parents' and grandparents' inference. We then evaluate the accuracy for "all-generations" ancestors' inference. Note that symmetry in pedigrees should be accommodated when calculating the admixture proportion accuracy of ancestors. Admixture proportion accuracy is calculated from the best match between the inferred and the true proportions. This best-match algorithm is given in *SI Appendix*.

In this experiment, the data were simulated forward in time for 11 generations. We evaluate the accuracy for each of the 11 generation's inference. As shown in Fig. 3, the admixture proportion accuracy by PedMix2 for the extant individuals is about 99%, which is only slightly lower than the RFMix's result. The comparison between PedMix and PedMix2 shows that PedMix2 provides more accurate estimates of admixture proportions for parents and grandparents. For great-grandparents, the admixture proportion accuracy is 93%. Even for great-great-grandparents, the admixture proportion accuracy is still higher than 90%. Our results show that PedMix2 can provide reasonably accurate estimates of the admixture proportion of all ancestors in a pedigree. As expected, proportion accuracy for more distantly related ancestors is lower. Nonetheless, the accuracy for founders 10 generations ago is still above 75%.

Robustness tests for proportion accuracy: varying parameters. Our experiments so far are for simulation data under the default parameter settings. There are a number of parameters that may affect the inference accuracy. To evaluate the performance of PedMix2 under different settings, we have performed extensive experiments for the following settings:

1. Three important parameters in simulations may influence the performance of PedMix2: ρ (recombination rate), μ (mutation rate), and t (splitting time of ancestral populations). In addition, n_b, the number of blocks, may also influence the performance of PedMix2 in inference.
2. Varying the number of generations g since the formation of the admixed population.
3. Mis-specifying the number of generations g since admixture. So far, we assume g is known. In practice, the true value of g may not be known exactly. We evaluate the effect on inference if g is mis-specified.
4. Varying the number of ancestral populations. By default, there are two ancestral populations. We now test the case of more than two ancestral populations.
5. Phasing errors. We evaluate the accuracy of admixture proportion estimates on data with phasing errors.
6. Complex admixture where admixture occurs in more than one generations during the formation of the admixed population.
7. Unbalanced admixture where the admixture ratio is different from the default value 0.5.

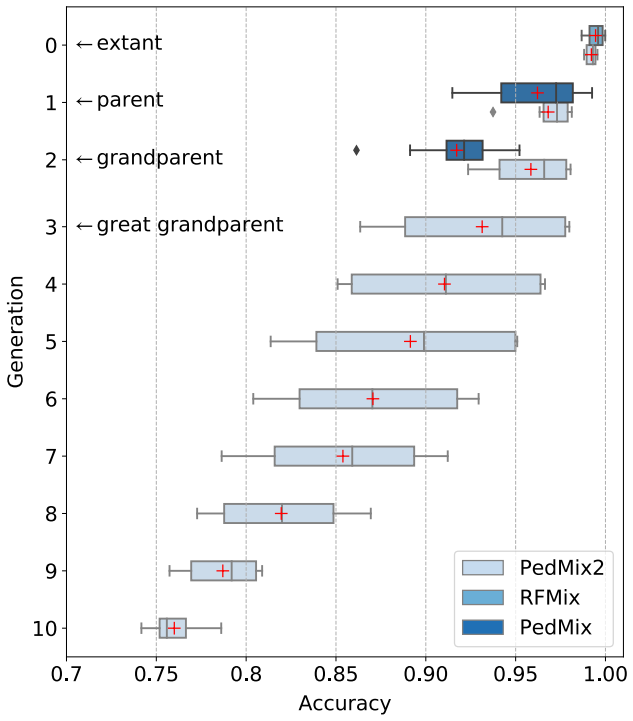

Fig. 3. Comparison of PedMix2 with RFMix and PedMix on admixture proportion accuracy for the extant individual and all 10 generations of ancestors. Data simulated under the default settings in Table 1. $g = 11$, "red +": the mean value.

Fig. 4 shows the results for several robustness tests. In Fig. 4*A*, we can see that very low ρ values may reduce proportion accuracy. However, the difference in proportion accuracy with different ρ values is not very significant. In Fig. 4*B*, we evaluate the effects of the splitting time of ancestral populations. There appears to be no very strong correlation between population splitting time and accuracy. Fig. 4*C* shows that the effect of μ values on proportion accuracy is also not very significant. This may be due to the fact that the number of SNPs is usually very large within each block. On the other hand, Fig. 4*D* shows that the value of n_b has larger impact on proportion accuracy than other parameters. The larger n_b is, the higher proportion accuracy is. The reason is that when the chromosome is divided into more blocks, recombination is more likely to be placed at the true position. A larger number of blocks inside a haplotype leads to a longer running time. To speed up computation, we use 20 as the default value for n_b in the experiments. Overall, proportion accuracy is not significantly affected by the parameters we tested. We note that there may be variance in simulation experiments. Nonetheless, the average proportion accuracy is in the range of 90% and above for the extant individuals under most settings. This shows that PedMix2 may be applicable for data with different settings. Other results are given in *SI Appendix*.

Configuration accuracy on simulated data. We test the accuracy of PedMix2 for inferring founder configuration, which is a main technical aspect of PedMix2. Under the default settings in Table 1, the configuration accuracy of PedMix2 is 74.9%. Note that with $g = 9$ generations, there are 2^{512} possible configurations. While PedMix2 may not infer the exact configuration, its inferred configuration is reasonably accurate, especially considering the vast space of possible configurations for relatively large g.

We also evaluate the performance of PedMix2 on data generated by varying several simulation parameters (similar to the proportion accuracy robustness test). The results are given in *SI Appendix*. Overall, configuration accuracy is acceptable under various settings of parameters.

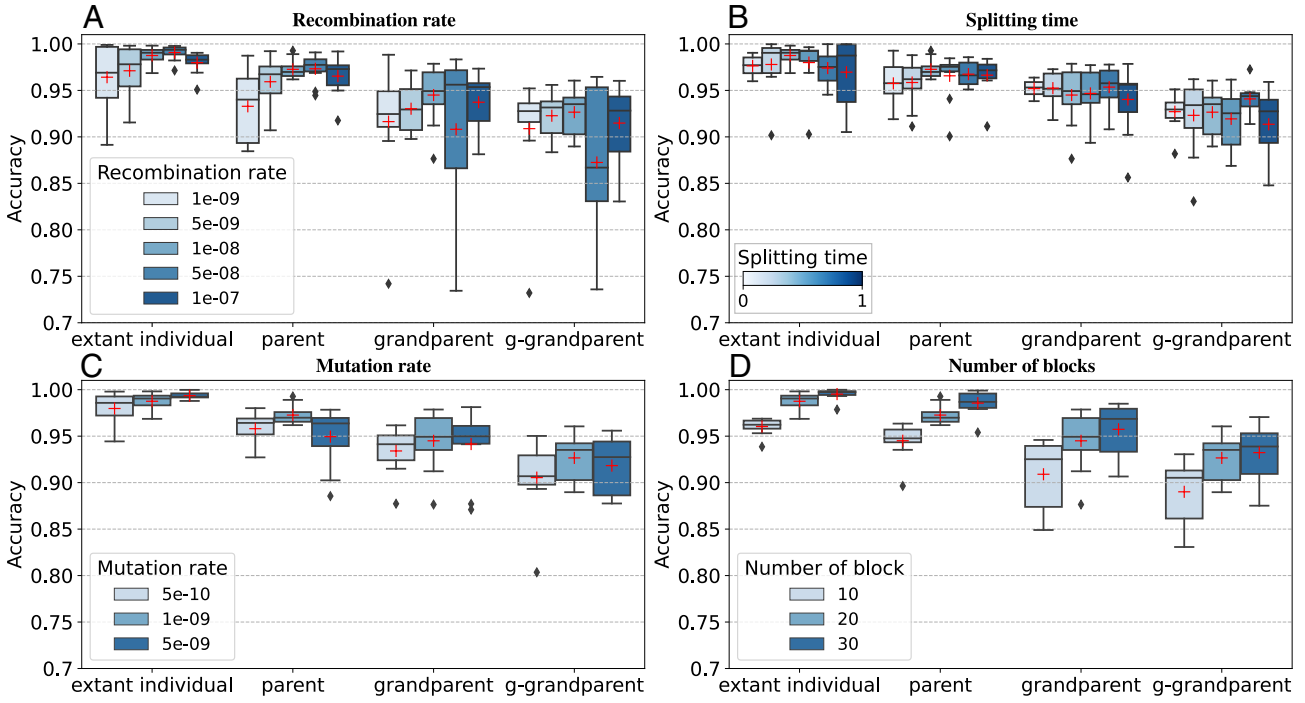

Fig. 4. Accuracy of inferred admixture proportions (of extant individuals and the ancestors within three generations) under various ρ, t, μ, and n_b values. (*A*) Different recombination rates ρ. (*B*) Different splitting time t. (*C*): different mutation rates μ. (*D*): different numbers of blocks n_b. t: six values (0.001, 0.01, 0.15, 0.3, 0.5, and 0.9 coalescent unit). "g-grandparent": great grandparent. "red +": the mean value.

Running time. We evaluate the running time of PedMix2 under different numbers of generations g and numbers of blocks n_b. Among all the simulation parameters, g and n_b greatly impact the running time. Our experiments are run on a machine with Linux and an Intel(R) Core(TM) i9-9900K CPU (3.60 GHz). Each data point in Fig. 5 represents the running time (in natural logs) for a single local search starting point. For example, the total running time of PedMix2 under the default settings is around 9 h (9 generations, 20 blocks per chromosome, 22 chromosomes, 2 ancestral populations, and 10 local search starting points).

In comparison, RFMix is very fast. PedMix is also faster than PedMix2 for parental and grandparental inference. While PedMix2 is slower than the two existing methods, one advantage of PedMix2 is that it can perform ancestry inference of all ancestors at once. The main time-consuming part of PedMix2 is the inference of the optimal configuration, but estimating the admixture proportion of each ancestor from an inferred configuration is trivial. In contrast, PedMix needs to run separately for parental and grandparental inference.

Results on Real Data. We now test PedMix2 on phased haplotypes of the trios from the 1000 Genomes Project. We use the genotypes from CEU (Utah residents with Northern and Western European ancestry) and YRI (Yoruba in Ibadan, Nigeria) as ancestral populations. The test samples are haplotypes of ten trios from the ASW (African Ancestry in Southwest US) population. In these ten trios, phased genotypes of the parents are available, while children's genotypes are unphased. We use Beagle (15) to phase children's genotypes from the phased haplotypes of parents. All parents' genotypes and reference populations are from the 1000 Genomes Project phase 3 data (https://www.internationalgenome.org/data-portal/population/).

We choose to use trio data so that the genomes of parents are given. We do not have the ground truth of admixture proportions and founder configurations. So we use RFMix to estimate the admixture proportions for the extant individuals and their parents. The estimated admixture proportions are then treated as the ground truth for the extant individuals and their parents.

Fig. 6 *A* and *B* show the results of PedMix2 for estimating the admixture proportions of extant individuals and their parents. Here, we use $g = 9$ and $n_b = 20$. The mean admixture proportion accuracy of PedMix2 is ∼97% for extant individuals (the 10 children in the trios) and ∼95.8% for their parents. Phasing errors may influence the accuracy of PedMix2, although our results show that PedMix2 still performs well on real data which may contain phasing errors.

For comparison, we run PedMix on the same data to estimate the admixture proportions of the parents and the grandparents. Note, however, the performance of PedMix is affected by its data-

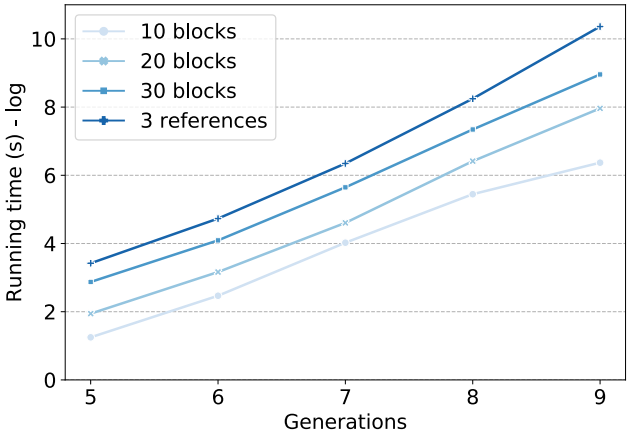

Fig. 5. Running time (in natural log) of PedMix2 under varying the number of generations in the pedigree and the number of blocks per chromosome. For the experiments with three reference populations, the number of blocks per chromosome (n_b) is set to 20.

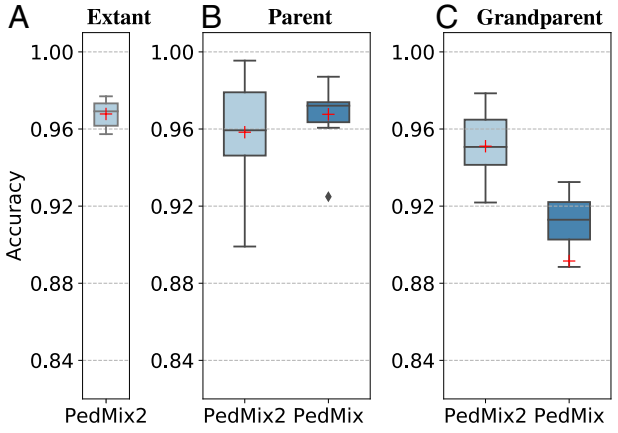

Fig. 6. Admixture proportion accuracy of extant individuals and their parents/grandparents for 10 trios from the ASW population of the 1000 Genomes Project. (*A*) Accuracy of PedMix2 for extant individual inference. (*B*) (parents) and (*C*) (grandparents): compare PedMix2 and PedMix with data that are trimmed for estimating parental and grandparental admixture proportion accuracy. "red +": the mean value.

trimming approach, which discards some SNPs from the data based on specific criteria such as linkage disequilibrium or allele frequencies. It is known that using certain trimming approaches can improve the accuracy of PedMix (8). The results by PedMix in Fig. 6*B* are the proportion accuracy of PedMix when data trimming is applied. Details on data trimming are provided in *SI Appendix*. If data are not trimmed, PedMix's proportion accuracy is significantly lower (e.g., ∼92% for parental inference). Nonetheless, it is often unclear how to apply data trimming in real data analysis. PedMix2 does not need to perform data trimming and still gets accurate inferences. For grandparents, we cannot use RFMix to estimate the true admixture proportions of grandparents since the genomes of grandparents are unavailable. So we estimate the parental admixture proportions inferred by PedMix2 from the genomes of parents and treat these as the ground truth for grandparents. Alternatively, we can use the parental results from PedMix. But this leads to higher grandparental inference errors for both methods. Fig. 6*C* shows that PedMix2 is more accurate than PedMix in grandparental inference. This is consistent with the results on simulated data.

Discussions

PedMix2 performs ancestral admixture inference from individual genomes. We have shown that PedMix2 can be applied to simulated genetic data to infer founders at 10 or more generations ago. Then, the inferred founders can be used to estimate admixture proportions of recent ancestors of an extant individual. A unique advantage of PedMix2 is that it can, in principle, provide admixture proportions of all recent ancestors whose genomes are not available. The main contribution of this paper is that we show it is feasible to infer useful aspects about the admixture of distantly related ancestors of an extant individual when only the genome of this extant individual is given. In contrast, most existing methods (e.g., RFMix) can only estimate admixture proportions for individuals with given genomes. Experiments on one real genetic data appear to suggest that PedMix2 can infer aspects of ancestry from real genetic data, although we acknowledge that a thorough validation on real data is not easy due to the lack of ground truth in these real data. Our results show that proportion estimates of extant individuals by PedMix2 and RFMix have similar accuracy. Note that RFMix runs faster than PedMix2 and can also perform chromosome painting. PedMix2 performs better than PedMix in almost every scenario. PedMix2 is more generally applicable than PedMix, which works for only parental and grandparental inference. PedMix2 allows more than two reference populations, while the current implementation of PedMix only allows two. However, PedMix allows genotypes with some phasing errors, while PedMix2 does not explicitly address phasing errors. Real data often have phasing errors. While phasing errors are common in current genetic data, we expect newer technologies in genotyping (e.g., long reads sequencing) can lead to data with very low (or even no) phasing errors. Moreover, our simulation shows that the effect of phasing errors on accuracy appears to be relatively modest (see the results in *SI Appendix*).

There are a number of parameters that may affect the performance of PedMix2.

1. Recombination rate. PedMix2 uses the recombination rate per base pair per generation in the inference. For humans, the recombination rate is around $1.0e^{-8}$. Our results show that variation in recombination rate appears to have only a modest effect on the performance of PedMix2.
2. Number of generations. The number of generations since admixture plays an important role in PedMix2 inference because it directly determines the number of founders of an extant individual. Empirical results show that the current implementation of PedMix2 runs reasonably fast for up to 12 generations. When using a larger number of generations, The number of blocks per chromosome should also increase. PedMix2 assumes the number of generations is known. In some cases, the number of generations is unknown and may need to be inferred, e.g., using the PAPI approach (10). Simulation shows that mis-specifying the number of generations does not significantly influence the proportion accuracy of PedMix2. We note that using a larger generation number than the true value does not necessarily lead to lower accuracy (but inference time would be longer). This is because, in this case, the founders of the perfect pedigree model are still from an ancestral population (i.e., un-admixed). It can be more problematic to use a number of generations that is smaller than the true value.
3. Number of blocks. Allowing each SNP to have its own inheritance path is the most accurate way to compute the likelihood. However, this leads to a long computation time. To obtain a practical method, PedMix2 divides a chromosome into several blocks. Increasing the number of blocks can significantly increase the running time of PedMix2.
4. Number of ancestral populations. PedMix2 allows multiple ancestral populations. Our results show that the admixture proportion accuracy rate does not decrease significantly with a larger number of ancestral populations. However, running time increases exponentially with the number of ancestral populations.
5. Number of initial configurations in local search. PedMix2 uses local search that starts from some initial configurations to search for the optimal founder configuration. To obtain a more accurate inference, one can run PedMix2 with a larger number of initial configurations. However, the running time of PedMix2 increases when the number of initial configurations increases.

There are several aspects of PedMix2 that need to be improved in the future. First, it is desirable to further speed up for a larger number of generations. Moreover, there may still be room for improving the inference accuracy. As shown in Fig. 2, the difference between admixture proportion and founder ratio may

not be small unless the data size is large. Therefore, for smaller data, PedMix2 may not give very accurate estimates of admixture proportions especially for more distant ancestors. To obtain more accurate inference, it is likely that more information (such as linkage disequilibrium in the ancestral populations) needs to be utilized in inference. However, more complex methods may also increase computational time and there may be a trade-off between accuracy and efficiency.

Overall, PedMix2 is currently the only general ancestry inference approach that can infer aspects of admixture for ancestors beyond grandparents from the genome of an extant individual. Our results show that it performs reasonably well in both accuracy and running time.

Data, Materials, and Software Availability. Computer programs with tutorial and examples for users data have been deposited in GitHub (https://github.com/biotoolscoders/pedmix2) (16).

ACKNOWLEDGMENTS. Research is partly supported by U.S. NSF grant IIS-1909425 (to Y.W.).

1. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
2. A. L. Price *et al*., A genomewide admixture map for latino populations. *Am. J. Hum. Genet.* **80**, 1024–1036 (2007).
3. B. Maples, G S, K EE, B CD., Rfmix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
4. A. Price *et al*., Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**, e1000519 (2009).
5. S. Sankararaman, S. Sridhar, G. Kimmel, E. Halperin, Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.* **82**, 290–303 (2008).
6. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
7. H. Tang, J. Peng, P. Wang, N. J. Risch, Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.: Off. Publ. Int. Genet. Epidemiol. Soc.* **28**, 289–301 (2005).
8. J. Pei, Y. Zhang, R. Nielsen, Y. Wu, Inferring the ancestry of parents and grandparents from genetic data. *PLoS Comput. Biol.* **16**, e1008065 (2020).
9. J. Y. Zou, E. Halperin, E. Burchard, S. Sankararaman, Inferring parental genomic ancestries using pooled semi-markov processes. *Bioinformatics* **31**, i190–6 (2015).
10. S. Avadhanam, A. L. Williams, Simultaneous inference of parental admixture proportions and admixture times from unphased local ancestry calls. *Am. J. Hum. Genet.* **109**, 1405–1420 (2022).
11. Y. Zhang, Y. Wu, Joint inference of ancestry and genotypes of parents from children. *iScience* **25**, 104768 (2022).
12. M. Liang, R. Nielsen, The lengths of admixture tracts. *Genetics* **197**, 953–967 (2014).
13. F. Baumdicker *et al*., Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* **220**, iyab229 (2022).
14. The 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 64–74 (2015).
15. B. L. Browning, X. Tian, Y. Zhou, S. R. Browning, Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
16. Y. Zhang, PedMix2. GitHub. https://github.com/biotoolscoders/pedmix2. Deposited 12 April 2023.