Simultaneous inference of parental admixture proportions and admixture times from unphased local ancestry calls

Authors

Siddharth Avadhanam, Amy L. Williams

Correspondence

alw289@cornell.edu

Avadhanam & Williams, 2022, The American Journal of Human Genetics *109*, 1405–1420
August 4, 2022 © 2022 American Society of Human Genetics.
https://doi.org/10.1016/j.ajhg.2022.06.016

CellPress

Simultaneous inference of parental admixture proportions and admixture times from unphased local ancestry calls

Siddharth Avadhanam[1] and Amy L. Williams[1,*]

Summary

Population genetic analyses of local ancestry tracts routinely assume that the ancestral admixture process is identical for both parents of an individual, an assumption that may be invalid when considering recent admixture. Here, we present Parental Admixture Proportion Inference (PAPI), a Bayesian tool for inferring the admixture proportions and admixture times for each parent of a single admixed individual. PAPI analyzes unphased local ancestry tracts and has two components: a binomial model that leverages genome-wide ancestry fractions to infer parental admixture proportions and a hidden Markov model (HMM) that infers admixture times from tract lengths. Crucially, the HMM accounts for unobserved within-ancestry recombination by approximating the pedigree crossover dynamics, enabling inference of parental admixture times. In simulations, we find that PAPI's admixture proportion estimates deviate from the truth by 0.047 on average, outperforming ANCESTOR and PedMix by 46.0% and 57.6%, respectively. Moreover, PAPI's admixture time estimates were strongly correlated with the truth ($\boldsymbol{R} = 0.76$) but have an average downward bias of 1.01 generations that is partly attributable to inaccuracies in local ancestry inference. As an illustration of its utility, we ran PAPI on African American genotypes from the PAGE study (N = 5,786) and found strong evidence of assortative mating by ancestry proportion: couples' ancestry proportions are highly correlated ($R = 0.87$) and are closer to each other than expected under random mating ($p < 10^{-6}$). We anticipate that PAPI will be useful in studying the population dynamics of admixture and will also be of interest to individuals seeking to learn about their personal genealogies.

Introduction

Widespread consumer interest in one's personal genetic ancestry and genealogy has fueled the rise of direct-to-consumer genetic testing companies,[1,2] enabling high resolution studies of population structure, admixture, and migration patterns.[3–5] One of the most popular features these companies provide is ancestry estimation: the geographic locations of the populations from which customers inherited their DNA. The techniques for such estimation, first developed by academic researchers,[6–8] have applications both to demographic inference[9–11] as well as to disease association mapping.[12,13]

Inferring an individual's genome-wide and locus-specific ancestry—i.e., local ancestry[6]—reveals not only that person's genetic heritage but also information about the genetic ancestry of his or her parents.[14,15] Indeed, each parent transmits half of his or her genome to a child, so in principle, one can recover half the DNA of each parent by analyzing data from a single individual. In practice however, phasing errors and uncertain interchromosomal phase pose substantial challenges to reconstructing the parents' genomes, although methods that analyze multiple siblings and/or use population allele frequencies can be effective.[16–18]

Many past studies of admixed individuals have focused on population-level ancestry questions, including migration times,[5,9,19] yet analyzing data from couples has the power to provide refined information about demographic patterns. For example, it makes possible studies of assortative mating by ancestry,[20–22] could help identify subgroups with different admixture histories, and may detect outlier individuals with patterns distinct from the bulk of the population.

The features derivable from local ancestry signals include population ancestry proportions and estimated times since admixture based on tract lengths.[9,23] Individual-level ancestry proportions are generally reliable to infer, depending on the ancestral populations,[6,24,25] yet computing admixture times for an individual (or his/her parents) from only one genome has not received careful attention and may be difficult. In particular, high variance in tract lengths[9] and the finite genome size can lead to noisy admixture times estimates.[26] Even so, noisy estimates may still be useful when analyzed in aggregate across many samples.

Although phase accuracy remains an issue in the context of local ancestry inference,[13] homozygous ancestry regions provide unambiguous information about the ancestry of the two parents at that locus. As such, even when analyzing unphased local ancestry tracts,[6,25] one can leverage these homozygous regions (as well as the lack thereof) to estimate the parental ancestry fractions. Analyzing unphased local ancestry tracts also helps prevent biases due to phase inaccuracy, both in estimating the parental ancestry proportions and the time since admixture (see subjects and methods).

[1]Department of Computational Biology, Cornell University, Ithaca, NY 14853, USA
*Correspondence: alw289@cornell.edu
https://doi.org/10.1016/j.ajhg.2022.06.016.

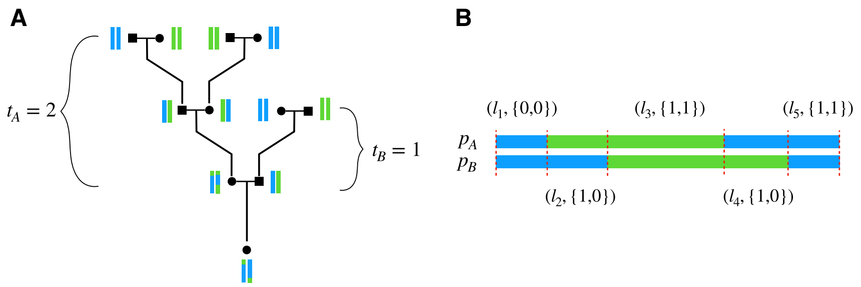

Figure 1. Example pedigree and local ancestry data

(A) Pedigree with vertical bars next to each individual representing two of their chromosomes, each colored to represent local ancestry from two populations, blue and green. t_A and t_B are the number of generations to the unadmixed ancestors for parents A and B, respectively; this is the generation in which at least one couple contains individuals of different ancestries.

(B) PAPI takes in unphased local ancestry tracts denoted by $x_i = (l_i, a_i)$. The image depicts a particular phasing of the data with labels that correspond to the observed unphased data.

We developed Parental Ancestry Proportion Inference (PAPI), a tool that provides estimates of both the admixture proportions and times since admixture of both parents of two-way admixed individuals. PAPI uses both a binomial model and a hidden Markov model (HMM) to infer these parental values, and, by default, combines these two models into a composite likelihood. We validated PAPI by using both simulated African Americans and trios from the HapMap African ancestry in Southwest USA (ASW) population.[27] PAPI's parental ancestry estimates are highly accurate, its time since admixture estimates are well correlated with the ground truth, and it can determine when the two parents have different times since admixture. We applied PAPI to African Americans from the PAGE study[28] and find strong evidence for ancestry-based assortative mating (results).

Subjects and methods

PAPI is designed to infer the ancestry proportions (p_A,p_B) and the average number of generations to unadmixed ancestors (t_A,t_B) for the unsexed parents A and B of a single admixed individual (Figure 1). It takes unphased diploid local ancestry tracts for focal admixed individuals as input and works on two-way admixed samples, but the approach can be extended to handle multi-way admixture. Throughout this paper, we label the two populations being analyzed arbitrarily as 0 and 1.

The advantage of using unphased local ancestry calls is that they are immune to switch errors that affect the phased input to (and therefore output of) haplotype-based local ancestry methods. PAPI makes inferences of times since admixture (t_A,t_B) based on local ancestry tract lengths, and switch errors can create artificially short or long ancestry tracts that would bias this approach if it relied on phased input. Additionally, switch errors in phased input can put an ancestry tract on an incorrect haplotype, which would lead to incorrect (p_A,p_B) estimates.

PAPI employs a composite likelihood built from two component models: a binomial model that infers (p_A,p_B) and a hidden Markov model (HMM) that jointly infers all parameters $\boldsymbol{\theta} = (p_A, p_B, t_A, t_B)$. The binomial model uses genome-wide diploid ancestry proportions to infer (p_A,p_B). In turn, the HMM estimates $\boldsymbol{\theta}$ by integrating over all possible phasings of the input local ancestry tracts using the forward algorithm. Our empirical findings indicate that forming a composite likelihood from these two models provides more accurate estimates of $\boldsymbol{\theta}$ than those of each component alone (results).

A key challenge in inferring (t_A,t_B) is that a subset of the crossovers the ancestors transmitted typically occurs between haplotypes of the same ancestry and, as a result, are not observable in local ancestry calls. The HMM accounts for this by approximating the rate of these hidden crossovers from the parental ancestry proportions (p_A,p_B) as described below. This approximation relies on a population genetic model of ancestry tract length distributions following a pulse migration,[9] which serves here as a simplified model of the complex dynamics of the meioses in the focal individual's pedigree.

A binomial model to infer parent ancestry proportions

The binomial model for estimating (p_A,p_B) operates on genome-wide diploid ancestry fractions. As a motivating example, consider a locus in the focal individual that is homozygous for ancestry. This locus carries unambiguous information about both parents' ancestries—it is certain that each parent carries at least one copy of the given ancestry in that region. Furthermore, low proportions of homozygous ancestry regions indicate that the parents most likely differ in their ancestry at most loci. For example, a child that is heterozygous for ancestry at every position most likely has two parents with very high levels of ancestry from distinct populations. An alternative but unlikely explanation of such data is that the parents have ancestries that alternate between the two populations across chromosomes. Thus, considering the genome-wide fraction of each type of unphased local ancestry call in the focal sample (i.e., homozygous for population 0 or 1 or heterozygous for ancestry) allows us to formulate a parsimonious model of the parents' ancestry fractions. These observations, which may not be fully captured in some other models of parental ancestry, are the key motivation behind PAPI's binomial model. We note that the approach described below is equivalent to that adopted in another recently developed method to model parental ancestry.[29]

PAPI's binomial model treats the DNA transmitted by a given parent $j \in \{A, B\}$ as a sequence of Bernoulli trials parameterized by the parent's ancestry proportion—i.e., by the target of inference p_j. We calculate p_j as the fraction of parent j's genome contained in tracts of ancestry 0, calculated with Morgan units. Note that PAPI only has access to the ancestry proportion of the haplotype the parent *transmitted*. This proportion will in general differ from p_j because of recombination and random assortment of alleles to gametes and is only equal to p_j in expectation. The difference between the transmitted and true parent proportions most likely accounts for some of the variance in estimation accuracy we observe in practice (results).

Formally, let $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be the input unphased diploid local ancestry tracts of a focal sample. Each of these tracts is a tuple $\boldsymbol{x}_i = (l_i, a_i)$ for $i \in \{1, \ldots, N\}$ where l_i represents the tract length in

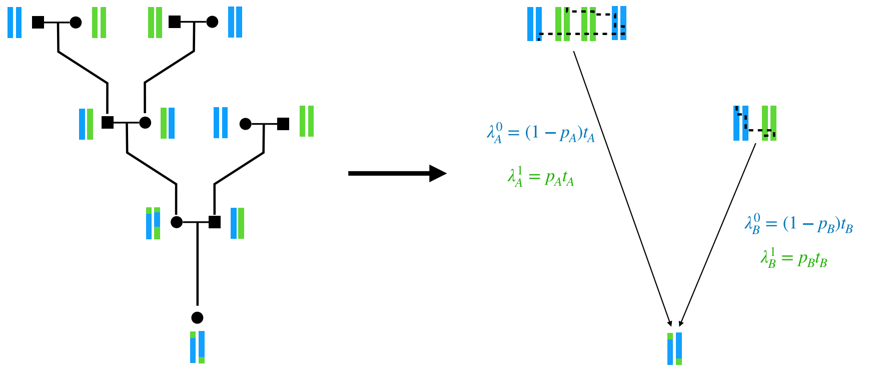

Figure 2. Markovian approximation to an individual's full pedigree
Example pedigree for an admixed individual (left). Viewing the unadmixed ancestors as a pool of founder haplotypes (right), one for each parent, and further viewing recombinations as a Markovian switching process in this pool motivates the use of a single-pulse migration model to capture the effect of t_j and p_j on the observed between-ancestry Poisson switch rates λ_j^0 and λ_j^1.

Morgans and $a_i \in \{\{0,0\},\{0,1\},\{1,1\}\}$ is its unphased local ancestry (Figure 1B). Curly braces in the ancestry component indicate that these values are unordered, since the data are unphased, so the $\{0,1\}$ class is consistent with the phased, ordered states (in square brackets) $[1,0]$ and $[0,1]$.

The likelihood of the binomial model is:

$$P(\boldsymbol{S}|p_A,p_B) = (p_A p_B)^{s_{\{0,0\}}} \times (p_A(1-p_B) + p_B(1-p_A))^{s_{\{0,1\}}} \times ((1-p_A)(1-p_B))^{s_{\{1,1\}}} \quad \text{(Equation 1)}$$

where $1-p_j$ is the fraction of parent j's genome contained in tracts of ancestry 1, $\boldsymbol{S} = (s_{\{0,0\}}, s_{\{0,1\}}, s_{\{1,1\}})$, and $s_a = \sum_i l_i I(a = a_i)\tau$, with I being the indicator function. That is, s_a is the total Morgan length of the genome contained in tracts of ancestry type a, scaled by a hyperparameter τ that accounts for the effective number of independent markers. This hyperparameter is necessary because the segment lengths are in Morgan units, and the effective number of independent loci in a tract of length 1 Morgan is the number of crossovers that have occurred in that interval since admixture began. In expectation, this number is exactly t_j for the haplotype parent j transmitted. Following this reasoning, we set τ to a rough population averaged number of generations since admixture. For all the analyses presented in results, we set $\tau = 7$, consistent with average estimates of the time since admixture from previous studies in African Americans,[6,30,31] and with PAPI's own estimates in the HapMap ASW population of African American samples[27] (results). PAPI is strongly robust to even large variations in τ: considering $\tau = 7$ versus $\tau = 20$, the accuracy of its admixture proportion estimates decreases by only 1.32% (in relative terms and measured with $d_{abs,p}$; see "computing measures of accuracy and bias") and that of its admixture time estimates by 0.621% (measured with Pearson correlation; Figures S2 and S3).

Modeling the relationship between admixture time and ancestry switch rates

Crossover placement along a haplotype that was subject to m meioses can be modeled as a Poisson process with rate m per Morgan. If $m \geq 4$, this model provides a good fit to intercrossover distances generated under a crossover interference model.[26] However, applying the Poisson model to local ancestry segments is complicated by the presence of within-population *hidden* crossovers. These are historical crossovers that occur at homozygous ancestry positions and, as a result, are not directly observed in local ancestry tracts. Their presence requires an adjustment to the ancestry switch rate parameter—i.e., this rate is not a direct measure of the number of meioses since admixture.

In this subsection we assume that for each parent $j \in \{A,B\}$, all the ancestors t_j generations ago are unadmixed and that at least one couple in that generation includes individuals of both ancestries, while the other individuals have the same ancestry (Figure 1A). We refer to this form of pedigree throughout as the "base case," and it corresponds to a single pulse of admixture, with migrants entering only t_j generations ago. This means that at least one of the ancestors $t_j - 1$ generations ago is entirely heterozygous for ancestry. We do this to defer consideration of the crossover rate parameter in more complex admixture scenarios—those with multiple pulses of admixture—to later in the text (see "simulating multiple pulses of admixture on a pedigree").

The model we adopt approximates pedigree-generated crossovers as Markovian transitions within a pool of haplotypes (Figure 2). The proportion of haplotypes in the pool from population 0 is p_j, the parental ancestry proportion. Gravel proposed a population genetic model of this form in the context of pulse admixture events,[9] and noted that the probability of transitioning to a haplotype of a different ancestry is simply the proportion of haplotypes of that ancestry in the population. In particular, the Poisson rates λ_j^0 for switching from ancestry 0 to 1 and λ_j^1 for switching from 1 to 0, are as follows:

$$\lambda_j^0 = \left(1 - p_j\right)t_j \text{ and } \lambda_j^1 = p_j t_j. \quad \text{(Equation 2)}$$

Thus, to obtain the switch rate out of a given ancestry, the adjustment to the crossover rate t_j is the proportion of haplotypes of the opposite ancestry. Intuitively, one can view these rate parameters (λ_j^0 and λ_j^1) as the expected count of crossovers per Morgan that result in an ancestry switch.

Note the nature of the approximation made here, which is 2-fold: first, crossovers in the model are no longer constrained to occur within an individual, which ignores long-range correlations caused by within-individual meioses.[32] Second, p_j is used to represent the proportion of founder haplotypes of ancestry 0 even though this value—parent j's ancestry proportion—may differ from the fraction of founders that have this ancestry. Despite these caveats, our empirical results suggest that this is an effective approximation (results).

An important limitation in applying the formulas in Equation 2 is that they incorrectly account for the crossovers transmitted by an individual that is entirely heterozygous for ancestry. To illustrate this, consider such a parent j, who will have $p_j = 0.5$. Equation 2 implies that $t_j = 2\lambda_j^0 = 2\lambda_j^1$, and because every locus in parent j's genome is heterozygous for ancestry, all of j's crossovers will produce ancestry switches. This makes the rate of switches $\lambda_j^0 = \lambda_j^1 = 1$ per Morgan, and so, despite including only one meiosis, the equations will suggest that $t_j = 2$. The issue is that the model implicitly assumes that the ancestry proportions of parent j's two chromosomes are both 0.5 when in fact one is entirely composed of ancestry 0 and the other of ancestry 1. This scenario arises for all ancestors with entirely heterozygous ancestry, and we have found that subtracting 1 from the raw estimates is an effective fix no matter how large t_j is[26] (data not shown). PAPI applies this correction internally by subtracting 1 from the t_j point estimates and from each Markov chain Monte Carlo (MCMC) sample followed by rounding up negative values to 0.

A hidden Markov model to jointly infer admixture times and ancestry proportions

PAPI computes the likelihood of the observed unphased local ancestry tracts $\boldsymbol{X}$ given assignments of the parameters $\boldsymbol{\theta}$, and uses either an MCMC or gradient descent (GD) method to estimate these parameters (below). The hidden states of the HMM encode phased local ancestry assignments $\mathbf{z} = \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$, where $\mathbf{z}_i \in \{[0,0],[0,1],[1,0],[1,1]\}\ \forall i$, and the two elements of $\mathbf{z}_i$ correspond to a transmitted chromosome by a given parent; as a matter of convention, we let the first element $z_{i,1}$ be the chromosome transmitted by parent A. Because the phase is unknown, the HMM integrates over all possible phase assignments to compute the likelihood of the unphased data. In order to model the relationship of p_j and t_j to the observed data while accounting for hidden recombinations (above), PAPI internally reparameterizes $\boldsymbol{\theta} = (p_A, p_B, t_A, t_B)$ as $\boldsymbol{\gamma} = (\lambda_A^0, \lambda_A^1, \lambda_B^0, \lambda_B^1)$ by applying Equation 2 for $j \in \{A, B\}$, and the method therefore calculates the likelihood $P(\boldsymbol{X}|\boldsymbol{\theta}) = P(\boldsymbol{X}|\boldsymbol{\gamma})$. Below, we treat the conditioning on parameters as implicit and omit them from the probabilities.

The HMM models the local ancestry tract lengths as exponentially distributed, as follows from a Poisson crossover model. More specifically, the initial probability is a product of the probabilities of the first observed tracts (for haplotypes A and B) in $\mathbf{z}_1$, both of which have length l_1: $P(\mathbf{z}_1) = P(l_1|\lambda_A^{z_{1,1}})P(l_1|\lambda_B^{z_{1,2}}) = \lambda_A^{z_{1,1}}\exp(-\lambda_A^{z_{1,1}}l_1)\lambda_B^{z_{1,2}}\exp(-\lambda_B^{z_{1,2}}l_1)$. We define the transition probabilities to ensure that tracts that extend beyond the interval they start in have a total probability consistent with their full length. That is, the transition probability $P(\mathbf{z}_i|\mathbf{z}_{i-1})$ depends on whether the ancestry of a haplotype changes between adjacent intervals and only includes the leading λ_j term in the first interval where haplotype j has switched ancestries. We express these as follows:

$$P(\mathbf{z}_i|\mathbf{z}_{i-1}) = \begin{cases} \lambda_A^{z_{i,1}}\exp\left(-\lambda_A^{z_{i,1}}l_i\right)\exp\left(-\lambda_B^{z_{i,2}}l_i\right) & \text{if } z_{i,1} \neq z_{i-1,1} \text{ and } z_{i,2} = z_{i-1,2} \\ \lambda_B^{z_{i,2}}\exp\left(-\lambda_B^{z_{i,2}}l_i\right)\exp\left(-\lambda_A^{z_{i,1}}l_i\right) & \text{if } z_{i,1} = z_{i-1,1} \text{ and } z_{i,2} \neq z_{i-1,2} \\ \exp\left(-\lambda_A^{z_{i,1}}l_i\right)\exp\left(-\lambda_B^{z_{i,2}}l_i\right) & \text{if } z_{i,1} = z_{i-1,1} \text{ and } z_{i,2} = z_{i-1,2} \\ 0 & \text{if } z_{i,1} \neq z_{i-1,1} \text{ and } z_{i,2} \neq z_{i-1,2} \end{cases}. \qquad \text{(Equation 3)}$$

For example, in the first case in Equation 3, parent B's transmitted chromosome does not change ancestry ($z_{i,2} = z_{i-1,2}$), and the transition probability incorporates the $\exp(-\lambda_B^{z_{i,2}}l_i)$ term to reflect this. The total probability of this tract from its origin in a given interval k to its termination at the end of some interval ℓ then is $\lambda_B^{z_{i,2}}\exp(-\lambda_B^{z_{i,2}}(l_k + l_{k+1} + \ldots + l_\ell))$—which is identical to the probability of observing a tract of length $(l_k + l_{k+1} + \ldots + l_\ell)$. By contrast, for this case to apply, haplotype A would have changed ancestries, and the formula includes a leading $\lambda_A^{z_{i,1}}$ term here where the tract begins. Other cases follow this same pattern, but the final case (in Equation 3) requires two crossovers to occur at the same position (one on each haplotype). This is prohibited by the model, and the corresponding transition probability is set to 0. If the observed data implies such a transition and PAPI is run without an error model (below), it will throw an error. Although in our analyses such transitions never occurred in the HAPMIX local ancestry tracts, a user can address any such cases by pre-processing them to introduce a small tract that switches the ancestry of only one haplotype between the two disallowed tracts.

The above formulation is somewhat non-standard in that the initial and transition probabilities depend on the tract lengths, which are part of the observed data. An alternative might put the exponential terms—which are calculated from the tract lengths—into the emission probabilities, leaving only the leading λ terms in the transition probabilities. However, doing this yields transition probabilities that are typically greater than one and obscures the conceptual underpinnings of the model—that of Poisson processes. The current formulation can be thought of as implicitly combining more standard transition and emission terms in the transition probabilities. We therefore set the emission probabilities in this formulation to 1 if the hidden state is a phased version of the observed local ancestry and 0 if not. We describe an optional error model that allows for mismatched hidden and observed ancestry states in the next subsection.

We treat the final tract of a chromosome as partially observed, as the position of the next crossover is unknown and would land somewhere beyond the chromosome endpoint. Such tracts should contribute only an exponential term to the likelihood, with the leading λ_j term omitted—corresponding to the probability of a tract of length greater than that observed. PAPI accomplishes this by setting the probability of transitioning from the final tract to a special end state Ω as follows: $P(\mathbf{z}_* = \Omega|\mathbf{z}_N) = \left(\lambda_A^{z_{N,1}}\right)^{-1}\left(\lambda_B^{z_{N,2}}\right)^{-1}$. These inverse terms act to cancel out the $\lambda_A^{z_{N,1}}$ and $\lambda_B^{z_{N,2}}$ terms that were introduced when the tracts with ancestry $z_{N,1}$ and $z_{N,2}$ were first transitioned to, leaving only exponential terms as desired.

With the HMM's initial, transition, and emission probabilities specified, PAPI computes the likelihood of the observed data $\boldsymbol{X}$ given $\boldsymbol{\theta}$ by marginalizing over all possible hidden state paths $P(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z})$ with the forward algorithm. The model treats each chromosome as independent given $\boldsymbol{\theta}$, so the total likelihood is the product of each chromosome's likelihood. Note that the fact that parent labels are arbitrary and interchangeable introduces non-identifiability into the model. In particular, $\boldsymbol{\theta} = (p_1, p_2, t_1, t_2)$ and $\boldsymbol{\theta}^* = (p_2, p_1, t_2, t_1)$ have the same likelihood and only produce swapped parent labels. When run in the MCMC inference mode (see below), PAPI handles this non-identifiability by restricting the parameter space to points such that $p_A > p_B$. When run in the GD inference mode (below), there is no need to restrict the space, as PAPI will find a mode of the posterior distribution, and the presence of more than one equivalent mode does not hamper the inference.

The above HMM is similar to that of ANCESTOR,[14] another approach for inferring parent ancestry proportions. Both methods apply to unphased local ancestry calls and model the phase of those calls in the hidden states. A point of departure is that PAPI parameterizes the model explicitly in terms of $\boldsymbol{\theta}$ (although it still performs the underlying calculations in terms of $\boldsymbol{\gamma}$), which affords it two advantages. First, by modeling hidden crossovers (above), PAPI can produce direct estimates of t_A and t_B rather than only λ_A and λ_B. Second, PAPI's HMM uses more of the available information for inferring (p_A, p_B) by incorporating these parameters into the transition probabilities (via λ_A and λ_B), thus producing more accurate estimates of (p_A, p_B) across a range of scenarios (results). While the (p_A, p_B) estimates from PAPI's HMM are in general more accurate than those of ANCESTOR, for the most accurate p_j estimates, PAPI's binomial component is important, and PAPI performs best when run with the HMM and binomial component together with the full model (results).

Handling erroneous tracts

Despite being immune to switch errors, unphased local ancestry segments are still subject to erroneous calls that have the potential to bias the inference of t_j. We address this by providing an optional error model that has a non-zero probability of emitting unphased local ancestry tracts with a different ancestry than that of the hidden state. Using the congruence operator, we say $\boldsymbol{z} \cong a$ if the ordered hidden state $\boldsymbol{z}$ has the same ancestry values as those in the unordered observed local ancestry a; e.g., $[1,0] \cong \{0,1\}$ and $[0,0] \ncong \{0,1\}$. Using this notation, the emission probabilities of the error model are as follows:

$$P(\boldsymbol{x}_i|\boldsymbol{z}_i) = \begin{cases} \exp(-\varphi l_i)/2 & \text{if } \boldsymbol{z}_i \ncong a_i \\ 1 - \exp(-\varphi l_i) & \text{if } \boldsymbol{z}_i \cong a_i \end{cases}. \quad \text{(Equation 4)}$$

Here, φ is a hyperparameter that we set heuristically so that when $\boldsymbol{z}_i \ncong a_i$, $P(\boldsymbol{x}_i|\boldsymbol{z}_i)$ decreases by a factor of 1/2 for every centiMorgan unit increase in l_i, so $\varphi = 100 \ln 2$. We divide the first case in Equation 4 by 1/2 because there are three unordered local ancestry calls and therefore two ways to select a_i such that $\boldsymbol{z}_i \ncong a_i$. The exponential form of these equations leads to higher emission probabilities for incongruous tracts that are small. Another way to think about this is that the model applies a penalty for treating larger tracts as erroneous and that the penalty decreases exponentially with decreasing tract size. So if the HMM encounters a sufficiently small tract where a_i differs from both a_{i-1} and a_{i+1}, it may favor the incongruous hidden state (i.e., where $\boldsymbol{z}_i \ncong a_i$ but $\boldsymbol{z}_i = \boldsymbol{z}_{i-1} = \boldsymbol{z}_{i+1}$) in order to avoid ancestry state switches over a small distance (see Equation 3) while incurring the incongruous tract penalty from Equation 4.

Combining component models and performing inference

As noted earlier, PAPI can combine the output signals from its two component models—the binomial model and the HMM—which improves its inference of $\boldsymbol{\theta}$ compared to either model alone (results). Notably, both models operate on the same data $\boldsymbol{X}$, which renders their likelihoods non-independent, but the HMM analyzes the unphased local ancestry calls $\boldsymbol{X}$ directly, while the binomial model considers the summary statistics in $\boldsymbol{S}$. Despite this non-independence, PAPI's default setting combines them to form a composite likelihood, treating $P(\boldsymbol{X}, \boldsymbol{S}|\boldsymbol{\theta}) \approx P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{S}|\boldsymbol{\theta})$. Furthermore, PAPI uses a Bayesian model, computing the following posterior based on the composite likelihood (which we refer to as the *full* model):

$$\begin{aligned} P(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{S}) &\propto P(\boldsymbol{X}, \boldsymbol{S}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \approx P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{S}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \\ &= P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{S}|p_A, p_B)P(p_A)P(p_B)P(t_A)P(t_B) \end{aligned} \quad \text{(Equation 5)}$$

Here, we make explicit the assumption that all the model parameters (in $\boldsymbol{\theta}$) are independent, and note that the binomial likelihood $P(\boldsymbol{S}|\boldsymbol{\theta}) = P(\boldsymbol{S}|p_A, p_B)$ since that model only considers (p_A, p_B). We use uniform priors on the parameters, with $p_A, p_B \sim U(0,1)$ and $t_A, t_B \sim U(0, 25)$. The full model can be thought of as treating the posterior distribution on (p_A, p_B) from the binomial model as a prior to the HMM. This may explain the improved accuracy (and speed) of the full model when compared to either the binomial or HMM alone (results); i.e., the genome-wide information from the binomial likelihood shapes the overall distribution, making it more strongly peaked near the true p_A and p_B.

PAPI uses one of two methods to perform inference, as specified by the user. By default, it performs gradient descent (GD) with the constrained optimization algorithm L-BFGS-B (implemented in the scipy-optimize[33] Python package). This yields a point estimate corresponding to a local maximum of the likelihood function. The empirical results suggest that these estimates, despite being local maxima, are reliable in practice. The second inference method uses MCMC in order to draw from the posterior distribution—specifically performing slice sampling (via the PyMC3 package[34]). In this setting, PAPI performs 500 burn-in iterations followed by 1,000 MCMC draws across four independent chains per individual (6,000 samples total) and outputs a 90% credible interval for each of the parameters.

The GD inference method is quite fast, whereas the MCMC approach requires larger compute times. In light of timing and performance (see results), we recommend running PAPI with the full model (Equation 5) in the GD inference mode. It is also possible to run the HMM or binomial model alone, i.e., to analyze $P(\boldsymbol{\theta}|\boldsymbol{X})$ or $P(\boldsymbol{\theta}|\boldsymbol{S})$, respectively. (In these settings, PAPI does not leverage a composite likelihood.) Note that all three models supply p_j estimates, and a user can run the binomial model in MCMC mode if posterior distributions of p_j alone are required (which we recommend, as the MCMC runs are much faster with the binomial model alone [results]).

Simulating realistic admixed genotypes

Perhaps the most realistic way to simulate genetic data for an admixed individual is to explicitly model their pedigree and simulate a meiosis for each of their ancestors. This would appropriately constrain crossovers to occur only between the two haplotypes of each ancestor and can incorporate crossover interference modeling.[26] However, doing this requires the use of data from 2^T founders, where T is the number of ancestral generations in the pedigree, thus consuming large data resources from unadmixed individuals to simulate only a single admixed individual. Furthermore, some unadmixed samples must be held out of the simulation for use in the panels required by the local ancestry inference software.

We adopted a three-step hybrid simulation scheme that explicitly accounts for pedigree-based transmissions while using far fewer unadmixed samples (Figure S1). First, we ran Ped-sim[26] to simulate individuals from a specified pedigree structure, similar to the above approach, but without providing input genetic data, and thus obtained no output genotypes at this stage. We used a sex-specific genetic map,[35] a Poisson crossover model, and ran Ped-sim in a mode (using the –bp option) that outputs the founder haplotype ID that the focal admixed individual carries at each position, with crossover break points indicated. Rather than using the resulting break point file as is, with its 2^T unique haplotypes, in step two we modified it by mapping each founder haplotype to one of the two ancestral population labels according to the desired value of $E[(p_A, p_B)]$ as well as (t_A, t_B) (see below). Finally, we used this population-only break point file as input to admix-simu,[36] which also takes in genetic data corresponding to unadmixed founder haplotypes from each population. Admix-simu treats the founder haplotypes as a pool and samples haplotype segments on the basis of the population labels and break points contained in the break point file. More specifically, at each break point, the method randomly samples a new haplotype (of the appropriate population) from the pool, creating non-overlapping sampling paths (Figure S1). This interrupts the linkage disequilibrium (LD) on either side of a break point as in a real crossover and approximates the full pedigree model. After sampling all segments of the target admixed individuals' haplotypes, admix-simu outputs the final haplotypes, which we then grouped in pairs to form individual diploid genotypes.

We simulated pedigrees of the base case form introduced earlier (see "modeling the relationship between admixture time and ancestry switch rates"), assigning the founders in generations (t_A, t_B) to different populations according to the proportions given by the expectation $E[(p_A, p_B)]$. This expectation will in general differ from the realized values (p_A, p_B) (in the parents of the focal individual) that we used to calculate deviation statistics (see "computing measures of accuracy and bias"). Further, the range of $E[(p_A, p_B)]$ values is constrained by the number of founder individuals in the pedigree, which depends on (t_A, t_B) (e.g., it is impossible to simulate $E[(p_A, p_B)] = (0.1, 0.1)$ when $(t_A, t_B) = (2, 2)$, as the number of founder haplotypes is too small). Therefore, we assigned the founder ancestries to approximate the expected proportions as closely as possible.

We used data from the HapMap CEU and YRI populations as the input unadmixed founder haplotypes for simulating. In order to provide large unadmixed panels to the local ancestry inference software and ensure these are unrelated to the target simulated individuals, we first excluded children from the HapMap CEU and YRI populations and split the remaining individuals (trio parents, duo parents, and unrelated samples) into four sets of 22 individuals for the two populations. We then used the pipeline described above to simulate admixed individuals in four separate runs, each using one of the four sets of simulation founders as input to admix-simu and with each assigned a different $E[(p_A, p_B)]$ value: $(0.25, 0.75), (0.4, 0.6), (0.25, 0.5)$ or $(0.5,\ 0.5)$. We repeated this procedure (with the same four founder datasets) for each of the desired settings of (t_A, t_B) (see results for these values). This produces four sets of simulated samples for every (t_A, t_B) that are unrelated to the other three sets of unadmixed samples. Doing this allows us to utilize the three sets of unadmixed samples (66 individuals in total) as panels in HAPMIX when analyzing the corresponding simulated individuals, thus leveraging all the available data for simulating while still maintaining sizeable panels for local ancestry inference.

To phase the HapMap samples, we first downloaded the Phase III genotypes (see data and code availability) and performed a lift over to GRCh37 coordinates.[37] We then removed all markers that were unmapped by the lift over and, for positions typed on both the Affymetrix and Illumina arrays, removed the Affymetrix genotypes. A few positions remained duplicated, and we retained one random genotype for each such site. Following this, we used the reported relationships to identify trios and duos and used PLINK 1.9[38] to set sites with Mendelian errors to missing (using the –set-me-missing option). Finally, we phased the data by running Beagle 3.3.2 (October 31, 2011)[39] on all the samples, supplying the labeled trios and duos as well as all unrelated individuals. (Notably this phasing run included all HapMap populations, not just the YRI and CEU samples.)

Simulating multiple pulses of admixture on a pedigree

The base case pedigree only includes a subset of the admixture histories that are possible, as unadmixed ancestors from distinct populations will in general appear at more than one generation in an admixed individual's pedigree. We sought to ask what value of t_j our pulse-migration model (Equation 2) will estimate in more general cases.

To characterize this, we simulated admixed haplotypes that have two admixture time parameters, t_j^e and t_j^l, which are, respectively, the earlier and later generations ($t_j^e > t_j^l$) in which couples containing unadmixed individuals of different ancestries occur (all other individuals are unadmixed of the majority ancestry). To perform these simulations, we used a base case pedigree for t_j^l with a desired setting of $E[p_j]$ and generated new pedigrees by replacing lineages—starting from the left-most lineage and working to the right—by lineages with a time since admixture of t_j^e and with the same $E[p_j]$. We use a parameter q to define the proportion of all the couples in the t_j^e generation composed of individuals of different ancestries. By this definition, the range of q depends on $E[p_j]$, as the maximum fraction of couples containing unadmixed individuals of different ancestries is $2E[p_j]$, so when $E[p_j] = 0.5$, q ranges between 0 and 1 and when $E[p_j] = 0.25$, q's maximum value is 0.5.

Note that for this analysis we used only the exact local ancestry tracts derived from the Ped-sim break point positions. We represent the tracts lengths from population 0 as l_k^0 for $k \in \{1, \ldots, N^0\}$ and from population 1 as l_ℓ^1 for $\ell \in \{1, \ldots, N^1\}$. From these, we estimated the Poisson rates by using the maximum likelihood estimator for the 22 finite autosomes[26] as $\hat{\lambda}_j^0 = (N^0 - 22)/\sum_k l_k^0$ and $\hat{\lambda}_j^1 = (N^1 - 22)/\sum_\ell l_\ell^1$. Finally, from these same data, we also estimated $\hat{p}_j = \sum l_k^0 / \left(\sum l_k^0 + \sum l_\ell^1\right)$.

In order to calculate the value of t_j the model in Equation 2 infers, we reasoned that $\hat{\lambda}_j^0 / \left(1 - \hat{p}_j\right)$ and $\hat{\lambda}_j^1 / \hat{p}_j$ each give a measure of t_j, and we compute a weighted average of these two estimators. Specifically, because our estimates of the total Morgan fraction of the genome that fall in tracts of ancestry 0 and 1 are $\hat{p}_j$ and $1 - \hat{p}_j$ respectively, we calculated this as a weighted average:

$$\hat{t}_j = \hat{p}_j \frac{\hat{\lambda}_j^0}{1 - \hat{p}_j} + \left(1 - \hat{p}_j\right) \frac{\hat{\lambda}_j^1}{\hat{p}_j} - 1. \qquad \text{(Equation 6)}$$

Here, as in PAPI, we subtract 1 for reasons described earlier (see "modeling the relationship between admixture time and ancestry switch rates").

Handling local ancestry calls as input to PAPI

PAPI reads local ancestry calls produced by either HAPMIX or LAMP-LD. (Note that methods such as RFMix[24] that require phased haplotypes and produce phased local ancestry calls are unsuitable for input to PAPI.) HAPMIX outputs posterior probabilities of the three possible unphased diploid local ancestry states at every marker, which PAPI converts into an internal representation of fixed (i.e., non-probabilistic) local ancestry calls. This conversion works by taking the maximally probable state at each marker. In contrast, LAMP-LD outputs fixed local ancestry calls and PAPI converts them directly into an internal representation. In simulated data, we found that LAMP-LD is prone to producing erroneous short segments, leading to an upward bias in admixture time inference. To address this, we filtered out LAMP-LD tracts shorter than 1 cM, which helps correct this bias (results not shown). Note that filtering out tracts can introduce state transitions prohibited by the HMM (see "a hidden Markov model to jointly infer admixture times and ancestry proportions"). In order to avoid this, we filtered out only those tracts that are flanked by tracts of identical ancestry state. The filtering process consists of deleting the target tract and merging the flanking tracts to fill in the missing space. We applied this filter in all the LAMP-LD-based analyses.

Data processing for ANCESTOR and PedMix

We formed the input to ANCESTOR by converting the non-probabilistic local ancestry calls that PAPI uses into the format expected by ANCESTOR. In order to run PedMix, we followed the authors recommendations and applied (1) a 0.05 minor allele frequency filter, (2) LD pruning (removing SNPS with $r^2 > 0.25$) by using overlapping blocks of 1,000 SNPs and a sliding window increment of 25 SNPs (with the PLINK option –indep-pairwise 1000 25 0.25), and (3) a switch-error correction tool provided by the authors. An additional utility for filtering SNPs on the basis of allele frequency differences between the ancestral populations reduced PedMix's performance, and we thus omitted this step.

Computing measures of accuracy and bias

The A and B parent labels in PAPI's $\hat{\boldsymbol{\theta}} = (\hat{p}_A, \hat{p}_B, \hat{t}_A, \hat{t}_B)$ estimates separate the inferred time and ancestry proportions of one parent from the other. However, PAPI does not infer the sexes of these parents, and in order to compute measures of accuracy and bias in our simulations, it is necessary to map these labels to specific parents. To do this, we choose the assignment that minimizes the sum of squared errors. Let p_{max} and p_{min} be the ancestry proportions of the parent with the larger and smaller estimates, respectively. Then, representing the estimated labels by an ordered vector $\boldsymbol{v} \in \{(A,B),(B,A)\}$, we minimize $f(\boldsymbol{v}) = (\hat{p}_{\boldsymbol{v}_1} - p_{max})^2 + (\hat{p}_{\boldsymbol{v}_2} - p_{min})^2$ and use $\boldsymbol{v}^* = arg\, min_{\boldsymbol{v}} f(\boldsymbol{v})$ as the mapping between estimated and true parent labels.

In results, we report the mean absolute deviation of the estimated parameters from the truth, which serves as a measure of accuracy, computing this as

$$d_{abs,p} = \frac{\left|\hat{p}_{\boldsymbol{v}_1^*} - p_{max}\right|}{2} + \frac{\left|\hat{p}_{\boldsymbol{v}_2^*} - p_{min}\right|}{2}. \qquad \text{(Equation 7)}$$

We also report a measure that uses exact rather than absolute differences, which serves as a measure of bias. This statistic is a vector of two numbers corresponding to the bias of each parent's ancestry proportion individually:

$$d_{bias,p} = \left(\hat{p}_{\boldsymbol{v}_1^*} - p_{max}, \hat{p}_{\boldsymbol{v}_2^*} - p_{min}\right). \qquad \text{(Equation 8)}$$

We also compute analogous metrics for the $\hat{t}_j$ estimates, denoted $d_{abs,t}$ and $d_{bias,t}$.

Filtering and processing the PAGE data

We used PAPI to analyze African American individuals from the BioMe Biobank subset of the Population Architecture using Genomics and Epidemiology (PAGE II; dbGaP: phs000925.v1.p1) study.[28] In order to identify admixed individuals with African and European ancestry, we merged all the individuals (N = 12,754) with the HapMap CEU and YRI trio parents—the latter providing known population labels—and ran ADMIXTURE v1.3[40] with $K = 5$. We analyzed the estimated ancestry proportions for the HapMap individuals to determine which of the K populations corresponded to African and European ancestry. We then selected PAGE individuals that met each of the following three criteria as subjects for our study: (1) has ≥5% African ancestry, (2) has ≥5% European ancestry, and (3) the sum of these two ancestries is ≥99.5%. This identified a total of 5,786 subjects for our analysis.

Because the number of SNPs that intersect between the PAGE and HapMap III datasets is small, we used 107 YRI and 95 CEU unrelated individuals from the 1,000 Genomes Project[41] (1000G) data as panels for performing local ancestry inference in the PAGE subjects. To get a common set of SNPs between 1000G and PAGE, we first filtered out multiallelic and indel variants from both datasets. Next, in order to prevent any strand inconsistencies in the allele encodings of the two datasets, we ran PLINK v2.00a2.3LM with the –ref-from-fa option to recode the PAGE alleles to the forward strand of the GRCh37 reference genome (the 1000G data are already encoded in this manner). This may not correctly assign alleles at A/T and C/G SNPs, and we filtered these SNPs later (below). Next, we applied the "composite filter" provided in the quality control report distributed with the PAGE data, which includes filters for Hardy-Weinberg equilibrium, discordant calls in duplicated samples, and Mendelian errors, among others. In order to identify any remaining incorrectly coded alleles, we applied a chi-squared test for allele frequency differences between two groups of subjects—the 5,786 PAGE individuals and a subset of 1000G composed of 80 YRI and 20 CEU individuals. Markers in the PAGE data and the 1000G subset are expected to have similar allele frequencies, and those with statistically significant differences indicate a potential mismatch in the strand encoding. We filtered all markers with $|z| > 4$ from both datasets. Finally, we removed all A/T and C/G SNPs, as allele-encoding mismatches in these markers (e.g., at high minor allele frequency variants) may not be detected by any of the analyses above. All of these steps left 409,095 markers that overlapped between PAGE and the 1000G data, and we used these to perform local ancestry inference via HAPMIX on each PAGE sample with the 1000G data as panels.

Results

We validated PAPI's approach to deconvolving parental transmissions within unphased diploid local ancestry tracts of a given individual by using simulated and real data. Our first analysis contrasts the accuracy of PAPI's estimates with each of its models and inference modes. Following this, we compare PAPI's parameter estimates of (p_A,p_B) with those of ANCESTOR and PedMix (which do not estimate (t_A,t_B)) under a number of simulated scenarios, finding that PAPI reliably infers these parent ancestry fractions. We then turn to an examination of the quality of PAPI's (t_A,t_B) estimates by using both inferred local ancestry tracts and exact tracts from our simulation pipeline. For these analyses, we considered a range of admixture histories, including the base case and more complex pedigrees (subjects and methods).

Following these efforts, we applied PAPI to a large dataset of African Americans from the PAGE study. By leveraging PAPI's power to accurately infer parental ancestry proportions, we examined the relationship of couples' ancestries to one another and find strong evidence of non-random mating, as described below.

Evaluating the performance of PAPI's component models

We simulated the genotypes of admixed individuals under three scenarios: (1) $t_A > t_B$, with $t_A = 9$ and $t_B \in [2, 8]$; (2) $t_A = t_B \in [2, 9]$; and (3) $(t_A, t_B) \in \{(0,0),(1,1)\}$ (Table S1; see subjects and methods for simulation details). For every (t_A, t_B) in scenarios 1 and 2, we simulated 22 admixed individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5),$ $(0.25, 0.5), (0.4, 0.6), (0.25, 0.75)\}$ for a total of 88 individuals per scenario. In scenario 3, the small numbers of generations correspond to limited ranges of parent ancestry proportions; as such, when $(t_A, t_B) = (0, 0)$, we simulated a total of 88 admixed individuals for $E[(p_A, p_B)] \in \{(0, 0),$ $(1, 0), (1, 1)\}$ (22 individuals for each of $E[(p_A, p_B)] \in$ $\{(0, 0), (1, 1)\}$ and 44 individuals for $E[(p_A, p_B)] = (1, 0)$) and when $(t_A, t_B) = (1, 1)$, we simulated 22 admixed individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5), (0, 0.5)\}$ for a total of 132 individuals. Note that here and elsewhere in results, p_j corresponds to African ancestry proportions. As described in subjects and methods, our simulation pipeline assigns population labels to unadmixed founders while matching as closely as possible the desired parent ancestry proportions. Because of random assortment and recombination, the realized (p_A,p_B) vary, and in general these founder ancestry assignments only give the expectation $E[(p_A, p_B)]$ (or more precisely, the closest possible values to this expectation). However, we used the realized values of (p_A,p_B) for computing deviation metrics $d_{abs,p}$ and $d_{bias,p}$.

PAPI can estimate (p_A,p_B) by using any of its three models: the binomial, the HMM, or the full model (a composite of the binomial and the HMM; subjects and methods). We ran all three models on the scenario 1 simulated individuals (Figure S4), and, averaged across all data points taken together, the GD estimates of (p_A,p_B) under the full model are quite reliable with $d_{abs,p} = 0.043$. This is the best performing GD-based model overall with an absolute deviance 6.52% and 14.0% smaller than those of the HMM ($d_{abs,p} = 0.046$) and binomial model ($d_{abs,p} = 0.050$), respectively. The same trend holds for the MCMC estimates, albeit with lower relative differences among the models: averaged across all scenario 1 data points, the full, HMM, and binomial models achieve $d_{abs,p}$ values of 0.040, 0.041, and 0.042, respectively. The differences between the models are most pronounced when $E[(p_A, p_B)] = (0.5, 0.5)$; in particular, using the GD inference mode for this setting, the $d_{abs,p}$ statistics for the binomial, HMM, and full models are, respectively, 0.061, 0.054, and 0.053; in turn, using MCMC inference, the corresponding $d_{abs,p}$ measures are 0.061, 0.054, and 0.049, respectively. Note that while the MCMC (p_A,p_B) estimates under the full model are more accurate than the corresponding GD estimates overall, this method's increased runtime and loss of accuracy in (t_A,t_B) estimation (below) make the MCMC estimates less effective when a user also seeks admixture time estimates.

Considering PAPI's (t_A,t_B) estimates, we measured the Pearson correlation R of the inferred t_B estimates with the true simulated t_B values in scenario 1 data ($t_A = 9$ is held fixed in scenario 1, and we ignore it in the correlation analysis). Here, PAPI's GD estimates under the full model yield $R = 0.761$, representing a modest improvement compared to the HMM ($R = 0.713$). By contrast, the MCMC results in this case show weaker performance, with the full model yielding $R = 0.678$ and the HMM model alone obtaining $R = 0.681$.

We also benchmarked PAPI's runtime for every combination of options (inference mode and model) on compute nodes with Intel Xeon E5 4620 processors, utilizing 4 GB of RAM (Table S2). We analyzed the 88 individuals simulated under scenario 2 with $t_A = t_B = 5$, one individual per PAPI run. The average wall clock time for the full model using the MCMC mode is 8.97 h per individual compared to 23.9 s using the GD mode, making the GD mode 1,346× faster than MCMC. The MCMC and GD runtimes for the full model are, somewhat counter-intuitively, slightly shorter than those for the HMM model (6.00% and 9.20% shorter, respectively), while runtimes for the HMM with the error model option turned on are slightly longer (3.29% and 8.75% longer for the MCMC and GD options, respectively). Finally, the binomial model runtimes are quite fast, with the MCMC mode completing in 5.38 min and GD mode requiring only 7.3 s.

In light of these results, we recommend running PAPI with the full model and using the GD inference option in most cases. These are PAPI's default settings, and unless otherwise stated, we used these options to produce the rest of the results. A notable exception to this recommendation is when a user seeks only (p_A, p_B) estimates; in this case the MCMC runtimes for the binomial model are relatively fast, and while this mode provides the least accurate (p_A, p_B) estimates of all MCMC-based models, they are more accurate than the best GD estimates, making the speed and accuracy trade-off worthwhile.

Accuracy of parent ancestry estimates across admixture scenarios

To evaluate PAPI's performance in the context of other methods, we compared its inferred parental ancestry proportions with those of ANCESTOR and PedMix across all three simulation scenarios described above. Figure 3A shows $d_{abs,p}$ for the three methods considering all scenario 1 and 2 data points, and Figures 3B and 3C show

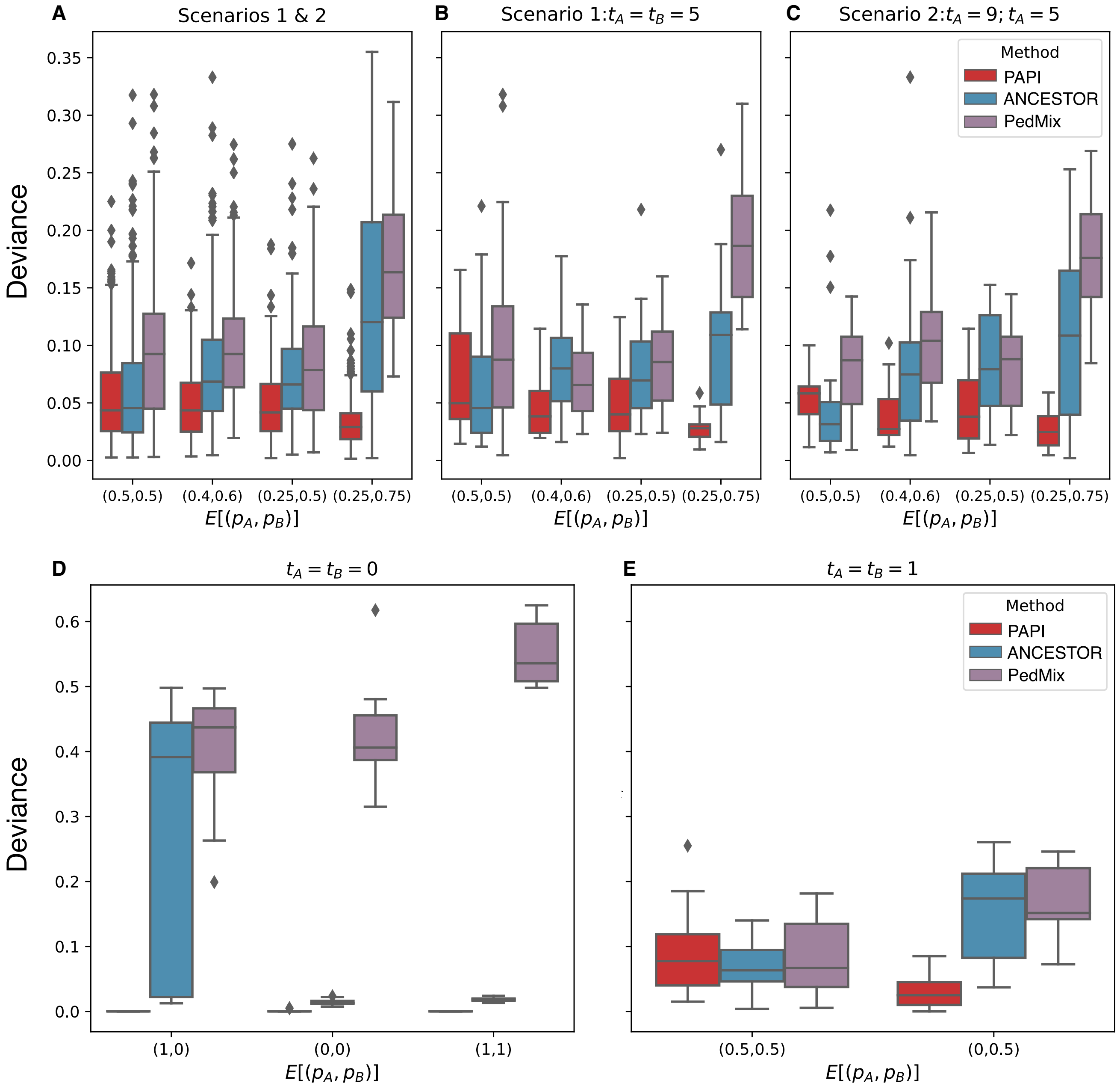

Figure 3. Deviance statistics in simulated data for inferred parent ancestry proportions (p_A, p_B) from PedMix, ANCESTOR, and PAPI
(A) Average absolute deviances ($d_{abs,p}$) on all scenario 1 and 2 data.
(B and C) Absolute deviances for data points where (B) $t_A = t_B = 5$ and (C) $t_A = 9$ and $t_B = 5$.
(D and E) Average absolute deviance on all scenario 3 data.
(D) PAPI estimates the ground truth near perfectly when $E[(p_A, p_B)] \in \{(1,0), (0,0), (1,1)\}$ and (E) ANCESTOR and PedMix outperform PAPI when $E[(p_A, p_B)] = (0.5, 0.5)$ but PAPI's deviance is low when $E[(p_A, p_B)] = (0, 0.5)$.

representative cases from each scenario. Overall, PAPI achieves $d_{abs,p} = 0.047$ on average for all scenario 1 and 2 cases compared to $d_{abs,p} = 0.087$ and $d_{abs,p} = 0.111$ from ANCESTOR and PedMix (representing improvements of 45.9% and 57.6%), respectively. PAPI's performance improves with increasing $|p_A - p_B|$; for example, when $E[(p_A, p_B)] = (0.25, 0.75)$, PAPI's absolute deviance (averaged over admixture times) is 0.032, which is 75.6% and 81.1% lower than ANCESTOR and PedMix's corresponding values of $d_{abs,p} = 0.131$ and $d_{abs,p} = 0.169$. In fact, this $E[(p_A, p_B)]$ setting yields PAPI's lowest mean deviance and ANCESTOR and PedMix's highest. This trend is recapitulated when using PAPI's HMM alone compared to ANCESTOR (Figure S5). While the exact explanation of these results is likely multi-factorial, it seems that the PAPI HMM is better equipped to exploit the strictly less informative heterozygous ancestry regions as well as those with homozygous ancestry (see discussion). We examined the relative performance of PAPI, ANCESTOR, and PedMix in every setting of (t_A,t_B) from scenarios 1 and 2 and find

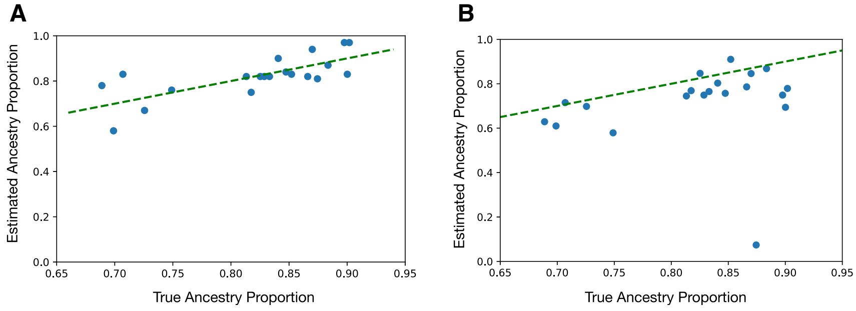

Figure 4. PAPI and ANCESTOR's parent ancestry proportion estimates for the HapMap ASW trio children
Scatter plots of estimated versus true parent ancestry proportions for (A) PAPI's full model and (B) ANCESTOR.

that PAPI outperforms both ANCESTOR and PedMix in the majority of cases, with a handful of exceptions when $E[(p_A, p_B)] = (0.5, 0.5)$ (see Figures S6 and S7).

PAPI's parent ancestry estimates have little bias, with average $d_{bias,p} = (0.0085, -0.0078)$ in scenario 1 samples and $d_{bias,p} = (0.0096, -0.010)$ in scenario 2 samples. The corresponding biases for ANCESTOR are higher at $d_{bias,p} = (-0.056, 0.022)$ and $d_{bias,p} = (-0.047, 0.035)$ in scenarios 1 and 2, respectively. Those of PedMix are also higher than PAPI with $d_{bias,p} = (-0.069, -0.013)$ in scenario 1 and $d_{bias,p} = (-0.069, -0.009)$ in scenario 2. Furthermore, the standard deviations of PAPI's $d_{abs,p}$ estimates are low at 0.035 in scenario 1 and 0.029 in scenario 2 samples, with corresponding values for ANCESTOR being 0.064 and 0.065 and for PedMix being 0.059 and 0.062.

Turning to the scenario 3 simulated data, in cases where both parents of the focal sample are unadmixed and from different populations ($E[(p_A, p_B)] = (1,0)$, so the focal sample is heterozygous for ancestry genome-wide), PAPI infers the ground truth with near-perfect accuracy ($d_{abs,p} = 0.0018$). In turn, ANCESTOR and PedMix have fairly high deviances in this case with average $d_{abs,p} = 0.277$ and 0.416, respectively (Figure 3D). When all four grandparents are unadmixed with each couple containing individuals of each ancestry ($t_A = t_B = 1$ and $E[(p_A, p_B)] = (0.5, 0.5)$), PAPI is less accurate than both ANCESTOR and PedMix, with the three methods having $d_{abs,p}$ values of 0.089, 0.067, and 0.087, respectively (Figure 3E). Even so, when exactly one parent is unadmixed ($t_A = t_B = 1$ and $E[(p_A, p_B)] = (0, 0.5)$), PAPI achieves $d_{abs,p} = 0.030$, which is considerably better than ANCESTOR's $d_{abs,p} = 0.150$ and PedMix's 0.176 (Figure 3E).

Finally, we validated PAPI's parent ancestry estimates in real data by running it on the offspring of all ten African American trios in the HapMap ASW population. The availability of parental data provides a direct measure of the ground truth via those parents' local ancestries (which we inferred with HAPMIX). As in the simulated data, we assigned the estimated parent labels to the real parents by minimizing the sum of squared deviances (subjects and methods). Figure 4 shows the estimated and true parent ancestry proportions of the trio children for PAPI and ANCESTOR. In these real samples, PAPI achieves $d_{abs,p} = 0.0497$ and $d_{bias,p} = (0.0272, -0.0263)$, which is more accurate and less biased compared to ANCESTOR's values of $d_{abs,p} = 0.111$ and $d_{bias,p} = (-0.150, -0.0547)$. PAPI's parental ancestry proportion estimates with the binomial and HMM models alone achieve similar but slightly reduced performance compared to these full model results (binomial model mean $d_{abs,p} = 0.0542$ and $d_{bias,p} = (-0.0264, 0.0253)$; HMM mean $d_{abs,p} = 0.0597$ and $d_{bias,p} = (-0.0184, 0.0263)$; Figure S8).

Accuracy of time since admixture estimates

We used PAPI to estimate (t_A, t_B) in data simulated under scenarios 1 and 2 (above) and found that it reliably recovers the truth under both scenarios (Figures 5A and 5B). In scenario 1 data, the PAPI estimates for the two parents are well separated, with the difference between the two estimates proportional to the true difference. More specifically, there is a high correlation between t_B and the corresponding PAPI estimate ($R = 0.76$), and a more moderate one between the inferred branch differences $\widehat{t}_A - \widehat{t}_B$ and the truth ($R = 0.53$). PAPI achieves similar quantitative results in the scenario 2 data, where the correlation between its estimates and the truth is $R = 0.79$ (calculated with both the t_A and t_B data points).

PAPI's estimates in these simulated data have an overall downward bias, with mean $d_{bias,t} = (-1.03, -0.986)$ for scenario 1 and $d_{bias,t} = (-1.42, 0.075)$ for scenario 2 data points. In order to understand the source of these biases, we ran PAPI on the exact ancestry tracts output by the simulator for data from both scenarios. This analysis yields reduced bias measures of mean $d_{bias,t} = (0.041, -0.701)$ for scenario 1 and $d_{bias,t} = (-1.06, 0.394)$ for scenario 2 samples (Figures 5C and 5D), suggesting that some of the downward bias can be explained by errors in local ancestry inference. The larger number of outliers in Figures 5C and 5D may be due to the presence of small tracts in the exact simulated data, which may go undetected by HAPMIX.

We contrasted these HAPMIX-based results with those obtained with tracts inferred by LAMP-LD (after filtering out tracts smaller than 1 cM; subjects and methods). As shown in Figure S9, the LAMP-LD-based estimates are biased upwards on average for both scenarios, with mean $d_{bias,t} = (1.07, -0.089)$ for scenario 1 and $d_{bias,t} = (-0.576, 1.66)$ for scenario 2 data points. The

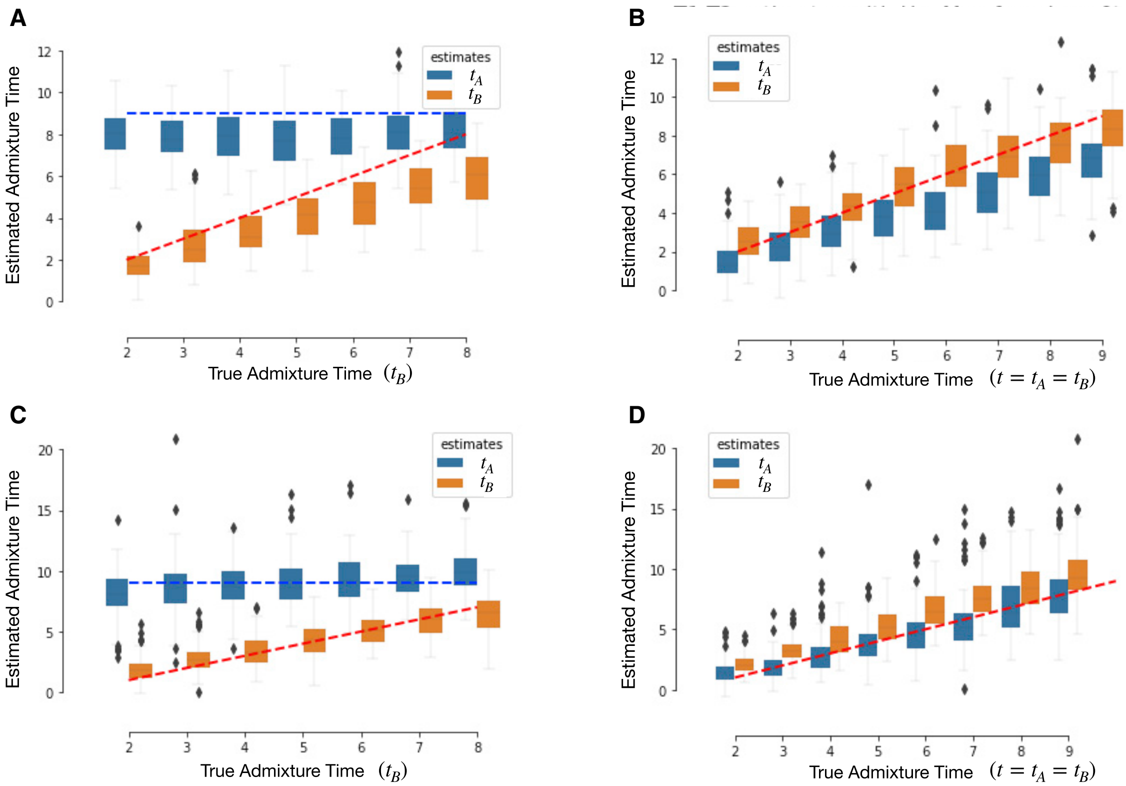

Figure 5. PAPI's estimated time since admixture in simulated individuals

(A and B) Box plots of estimated times based on HAPMIX local ancestry tracts for (A) scenario 1 samples (where $t_A > t_B$) and (B) scenario 2 data (where $t_A = t_B$).

(C and D) Estimated times based on exact local ancestry tracts for (C) scenario 1 and (D) scenario 2 individuals. In (A) and (C), the x-axis gives the true t_B since t_A is held fixed in these cases. Blue and orange dashed lines represent the simulation ground truths for t_A and t_B, respectively in (A) and (C), while the red dashed lines in (B) and (D) represent the simulation ground truths for t_A and t_B.

LAMP-LD-based estimates also show weaker correlations with the truth compared to the HAPMIX-based estimates with $R = 0.59$ for scenario 1 (considering only t_B values) and $R = 0.69$ for scenario 2 data points. Given these results, and the fact that HAPMIX further classifies $\geq 98.5\%$ of SNPs correctly for all three diploid ancestry states (see Figure S10), we conducted all further analyses with HAPMIX local ancestry tracts.

The above results consider the $(\widehat{t}_A, \widehat{t}_B)$ estimates in aggregate across simulated individuals, but PAPI is designed to be applied to individual samples as well. We plotted sample-specific point estimates of (t_A, t_B) (obtained with the MCMC inference mode) in the scenario 2 data along with their 90% Bayesian credible intervals (CIs) in Figure S11. The 90% CIs overlap true values in 86% of data points and have a mean range of 4.99. Notably, when the true $t_A = t_B = 0$, the 90% CIs of the model (in this case run with the error model option) have very large mean range of 16.5 and are strongly biased ($d_{bias,t} = 7.06$; Figure S12). This is due to the fact that when $t_A = t_B = 0$, both p_A and p_B must be either 0 or 1, and hence the inferred values in $(\widehat{p}_A, \widehat{p}_B)$ are each close to one of these extremes. Equation 2 implies that when this happens, either $\lambda_j^0 \approx 0 t_j$ or $\lambda_j^1 \approx 0 t_j$, and therefore relatively small changes in λ_j can lead to large changes in t_j; thus small erroneous tracts can readily bias the results. Note that PAPI performs better in GD inference mode when using the error model (see "effectiveness of PAPI's error model"); this may be because GD identifies a local maximum while the MCMC calculation integrates error and non-error state assignments for short, erroneous tracts. Still, as noted below in the discussion of the error model, we advise caution in interpreting the t_j estimates when $p_j > 0.95$ or $p_j < 0.05$.

We further validated PAPI's t_j estimates in real data by again analyzing the ten ASW trio offspring. Here, the parents' true time since admixture cannot be directly ascertained, so we first ran PAPI on each trio parent and computed the truth as an average of that parent's time since admixture estimates, adding one for the meiosis to the trio offspring: $t_P = (\widehat{t}_A^P + \widehat{t}_B^P)/2 + 1$ and $t_M = (\widehat{t}_A^M + \widehat{t}_B^M)/2 + 1$, where for $j \in \{A, B\}$, $\widehat{t}_j^P$ and $\widehat{t}_j^M$ represent PAPI's estimates on the trio father and trio mother, respectively. We then assigned the trio children's $(\widehat{t}_A, \widehat{t}_B)$ estimates to parent labels by minimizing the sum of squared differences (subjects and methods). PAPI's estimates in these data have a strong correlation with the truth of $R = 0.77$, in line with that observed in the simulated data (Figure S13). These estimates are somewhat biased with $d_{bias,t} = (2.061, -0.332)$.

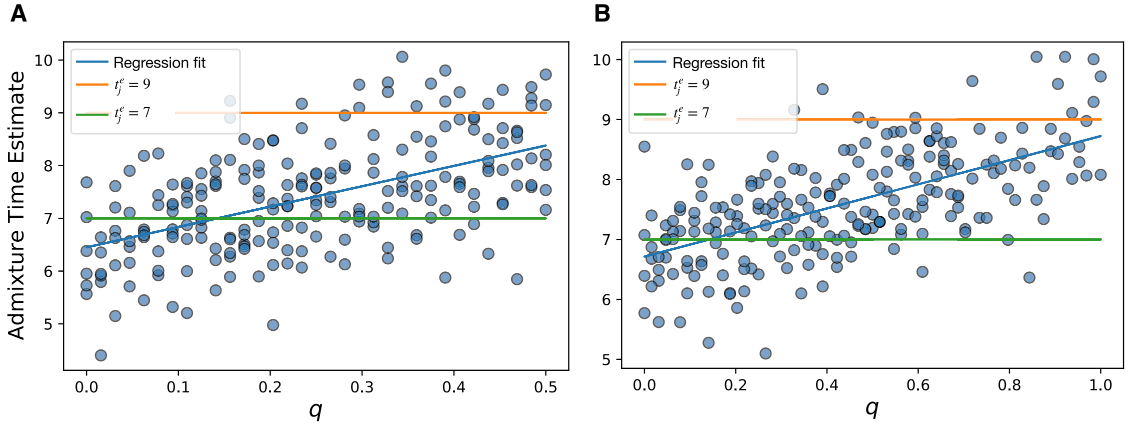

Figure 6. PAPI's estimated time since admixture with two migrant pulses
Plots show estimated time since admixture versus q (the proportion of couples in the t_j^e generation with unadmixed individuals of different ancestries) when $t_j^e = 9$ and $t_j^l = 7$ for (A) $E[p_j] = 0.25$ and (B) $E[p_j] = 0.5$. The blue lines depict the least-squares regressions on the plotted data points.

Interpreting time since admixture estimates

As described in subjects and methods, we simulated single haplotypes under a larger class of pedigree structures defined by four parameters—t_j^e, t_j^l, $E[p_j]$, and q. We carried out these simulations with $(t_j^e, t_j^l) = (9,7)$ for each of $E[p_j] \in \{0.25, 0.5\}$, where, when $E[p_j] = 0.5$, q varied between $[0,1]$ in increments of 2^{-6}, and when $E[p_j] = 0.25$, q varied between $[0, 0.5]$ in the same increments (q's maximum value is 0.5 for $E[p_j] = 0.25$). We further simulated structures with $(t_j^e, t_j^l) = (9,5)$ for $E[p_j] = 0.5$, varying q between $[0,1]$ in increments of 2^{-4}.

Given the exact local ancestry segments output by the simulation pipeline, we applied Equation 6 to estimate $\widehat{t}_j$ and then empirically examined the relationship between $\widehat{t}_j$ and the simulation parameters t_j^e, t_j^l, $E[p_j]$, and q. Our results suggest that $\widehat{t}_j$ linearly increases from t_j^l to t_j^e as q increases from 0 to its maximum value (Figures 6 and S14), but notably, the $\widehat{t}_j$ estimates are downwardly biased from the weighted average over all N simulated datapoints. In particular, when $E[p_j] = 0.5$, this weighted average is $t_j^{avg} = \frac{1}{N}\sum_i (q_i t_j^e + (1 - q_i) t_j^l)$, where i indexes individual datapoints. When $E[p_j] = 0.25$, we use $2q_i$ and $(1 - 2q_i)$ as the weights on t_j^e and t_j^l, respectively, with the factor of 2 accounting for q's smaller range in this case. We measured the bias as $\left(\frac{1}{N}\sum_i \widehat{t}_{j,i}\right) - t_j^{avg} = -0.584$ when $E[p_j] = 0.25$ and -0.282 when $E[p_j] = 0.5$. Similar biases arise in PAPI's t_j estimates when using both simulated exact ancestry tracts and tracts inferred by HAPMIX (see the previous subsection).

While these results indicate that the estimated times since admixture may generally be a weighted average of the times in which individuals of a given ancestry entered an individual's pedigree, a full understanding of the quantitative relationship requires more detailed analyses that are beyond the scope of this work. Instead we note simply that PAPI's admixture time estimates potentially represent inputs from several generations and should not be interpreted as *the* generation in which admixture occurred.

Effectiveness of PAPI's error model

While generally reliable, local ancestry inference methods do produce erroneous tracts, which can lead to biased (t_A, t_B) estimates (see "accuracy of time since admixture estimates"). These biases are most pronounced when the true p_A and p_B values are close to 0 or 1, such as in a subset of the scenario 3 simulated data (where $E[(p_A, p_B)] \in \{(0,0), (1,0), (1,1)\}$ and $(t_A, t_B) = (0,0)$). In these cases, PAPI's error model helps account for any erroneous tracts, leading to improved estimates of (p_A, p_B) ($d_{abs,p} = 0.0078$ with the error model as opposed to $d_{abs,p} = 0.0148$ without) and (t_A, t_B) ($d_{bias,t} = 0.957$ with the error model and $d_{bias,t} = 9.934$ without). Despite these improvements when the true $t_A = t_B = 0$, note that in practice a few small tracts in real data that this model considers to be erroneous may in fact be real and reflect old admixture. As such, by default, PAPI runs without the error model, and we advise caution in interpreting its t_j estimates when $p_j > 0.95$ or $p_j < 0.05$. If a user is confident that the population under study is only recently admixed, using the error model can improve the results. In either case, it may be prudent to re-run PAPI in MCMC mode and examine the posterior distribution of (t_A, t_B) for samples with $p_j > 0.95$ or $p_j < 0.05$ or to simply avoid consideration of results from such parents.

Ancestry proportions of parental couples are strongly correlated with each other in PAGE African Americans

We ran PAPI on detected African Americans from the BioMe Biobank subset of the PAGE study,[28] considering a total of 5,786 individuals that met our ancestry-based

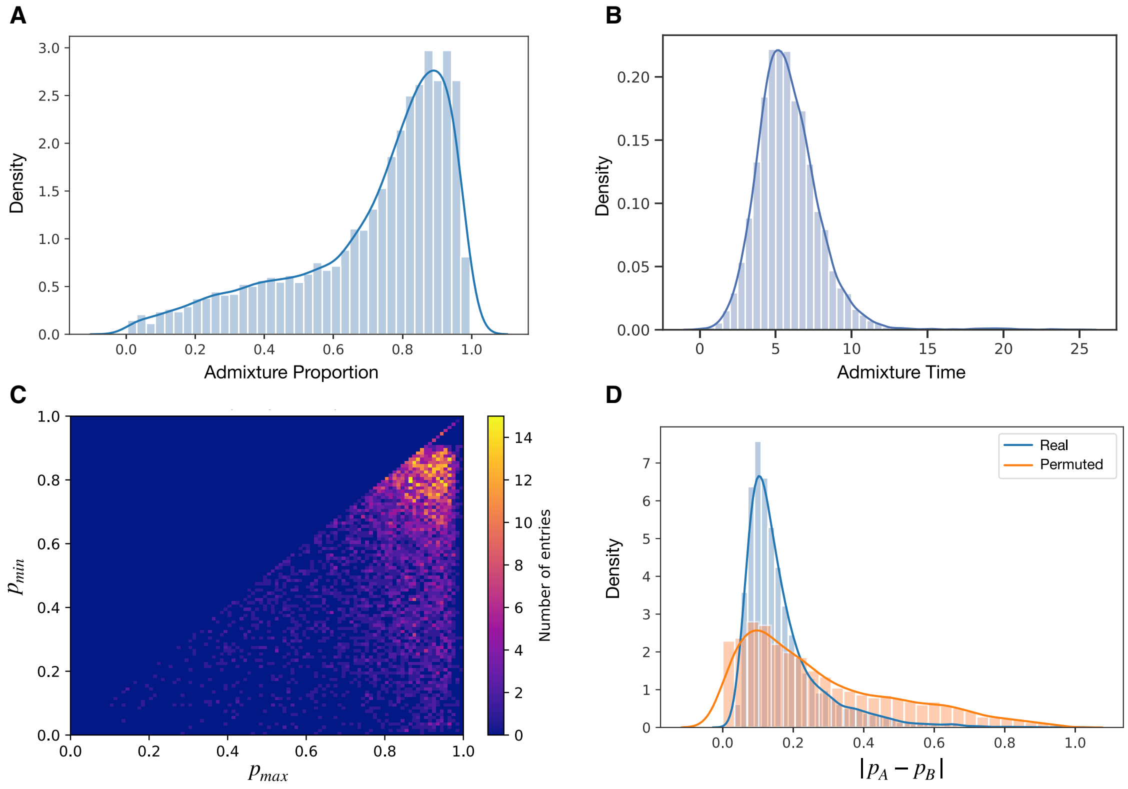

Figure 7. PAPI analyses from African Americans in the BioMe Biobank subset of PAGE

(A) Distribution of the parent ancestry estimates (both p_A and p_B plotted separately) in these African Americans.
(B) Distribution of time since admixture estimates (t_A and t_B plotted separately).
(C) Heatmap of the ancestry proportions of couples (p_A, p_B) ordered as (p_{max}, p_{min}).
(D) Distribution of $|p_A - p_B|$ estimates from the real couples along with a permuted distribution.

criteria for inclusion (subjects and methods). Here, we initially examined the parental ancestry proportion estimates from PAPI's binomial model run in MCMC inference mode—the recommended mode when only p_j estimates are required (see "evaluating the performance of PAPI's component models"). The average African ancestry proportion in the parents is $p_j = 0.71$, which is roughly consistent with previous estimates in African Americans.[6,30,31] The majority of parents have proportions between 0.8 and 1.0, corresponding to high African ancestry, but a heavy left tail in the distribution drives down the mean (Figure 7A). Figure 7B plots the estimated time since admixture—obtained by running PAPI with the default options—considering all parents that have $p_j \in [0.05, 0.95]$ for $j \in \{A, B\}$. This distribution has an overall Gaussian-like shape, which may suggest random mating with respect to the time since admixture among the ancestors of these African Americans. The population mean of t_j estimates is 6.10, again roughly consistent with previous estimates in African Americans.[6,30,31] We infer a small set of individuals (0.04% of the dataset) with $t_j > 15$, but since the simulated data also contain some overestimated extreme values, we assumed that these estimates are inflated.

We analyzed PAPI's parental admixture proportion estimates within all parent pairs—all couples—from these African Americans. Figure 7C depicts a heatmap of $(p_{max}, p_{min}) = (\max(p_A, p_B), \min(p_A, p_B))$ (since the A and B labels are arbitrary) where one salient feature is a strong cluster of points centered roughly at $(0.95, 0.85)$ that spreads diffusely outward. This cluster is consistent with a pattern of assortative mating between individuals with high African ancestry, and includes a meaningful fraction of the couples (25.2% of individuals in the dataset have $p_{max} \geq 0.90$ and $p_{min} \geq 0.80$). Furthermore, the Pearson correlation coefficient between p_{max} and p_{min} is $R = 0.871$, also indicating assortative mating by ancestry.

To more precisely characterize the strength of the signal for non-random mating in these couples, we examined the per-individual distribution of $|p_A - p_B|$ (Figure 7D). This density is roughly Gaussian with a heavy right tail and has a mean of 0.169, largely driven by the high density of couples noted above. This real data distribution is visibly distinct from what would arise under random mating, as randomly shuffling the members of the couples creates a much more dispersed, thick-tailed density (Figure 7D). Indeed, the mean $|p_A - p_B|$ across 10^6 permutations is 0.272, and no permutation mean reached as small a value as in the real data (minimum permutation mean 0.264; $p < 10^{-6}$). Notably the real data distribution has a depletion of mass near $|p_A - p_B| = 0$, which could be due to PAPI being biased when $p_A \approx p_B$. Consistent with this, PAPI shows slightly reduced performance in simulated data when $E[(p_A, p_B)] = (0.5, 0.5)$ (see Figure 3).

In order to ensure that the assortative mating signal is not being driven by this bias (or other biases in PAPI's inference), we simulated genotypes for 176 admixed African Americans in the same manner as before (subjects and methods) but with $t_A = t_B = 6$ and with the (p_A, p_B) pairs randomly sampled from the parental admixture proportion distribution depicted in Figure 7A. As we sampled p_A and p_B independently, this produces a null (random mating) distribution akin to the shuffled distribution above. The results from running PAPI on these simulated individuals (see Figure S15) are remarkably similar to the permuted distribution in Figure 7D, and the mean $|p_A - p_B|$ in these simulated individuals is 0.261. This indicates that potential bias in PAPI's inference is not driving the assortative mating signal. Moreover, the simulated data distribution also lacks mass near $|p_A - p_B| = 0$, suggesting that, while PAPI is somewhat biased in this region of the parameter space, this bias alone does not drive the assortative mating signal. Taken together, these observations provide strong evidence for assortative mating by ancestry proportion in the parents of these African American individuals.

Discussion

In this paper, we described and evaluated PAPI, a novel approach to inferring the admixture proportions and times since admixture for each parent of an extant individual that combines two models. The first is a binomial model that takes a global view of the ancestry of an individual, considering the overall proportions of homozygous and heterozygous ancestry in their genome. As noted earlier, this model is equivalent to that described in a recent paper.[29] PAPI employs a Bayesian framework and combines this binomial model with a second HMM that leverages the information present in the length distribution of local ancestry tracts. The HMM incorporates a population genetic model[9]—but uses parameters specific to the individual's pedigree—to capture the hidden recombinations within these tracts (i.e., those that do not switch the ancestry of the haplotype).

To characterize PAPI's accuracy, we simulated African Americans under various pedigree topologies and distributions of parent ancestry proportions and also used the real HapMap ASW trios. Our results show that PAPI is more accurate than ANCESTOR and PedMix at inferring parent ancestry proportions in many settings (Figure 4) and can detect when $t_A \neq t_B$ (Figure 5). In simulations that rely on inferred local ancestry tracts, PAPI is able to recover t_j estimates highly correlated with the truth, albeit with a downward bias when using HAPMIX local ancestry tracts and an upward bias when using tracts from LAMP-LD (Figure 5).

PAPI's approach models historical crossovers with a Markovian approximation to the more elaborate dynamics of crossover transmissions in a pedigree. The transition probabilities of PAPI's HMM are otherwise similar to those of ANCESTOR, although ANCESTOR also supports non-exponential tract length distributions by using a stochastic expectation maximization technique for inference. In contrast, PedMix[15] parameterizes the full pedigree up to the grandparental generation by using an ancestral configuration vector (which captures grandparental ancestry states, phase error, and all transmitted recombinations) while treating older recombinations as Markovian. Despite only approximating pedigree-based crossovers, and through its two component models, PAPI produces highly accurate estimates of parental ancestry proportions without relying on phased input. Note that neither PedMix nor ANCESTOR estimate parental admixture times (ANCESTOR instead provides estimates of average ancestry tract lengths for each parent).

As noted earlier, ANCESTOR and PedMix's performance decline as the difference $|p_A - p_B|$ increases, while PAPI maintains strong performance throughout the (p_A, p_B) settings we considered. By using a pulse-migration model for local ancestry, PAPI's HMM embeds the parent ancestry proportions p_j in its transition probabilities, enabling it to derive signal from all tracts, including those of heterozygous ancestry that are phase-uncertain. Compared to ANCESTOR, this change in the likelihood function shifts probability mass to regions of parameter space with divergent p_A and p_B when the focal individual has high heterozygous ancestry. ANCESTOR's model, on the other hand, seems to derive signal for (p_A, p_B) by using the HMM initial probabilities, i.e., from the first tract on each chromosome, rather than from all tracts (see section 2.3 in the ANCESTOR paper for its transition probabilities[14]). In turn, PedMix's performance is affected by phase errors that its model partially addresses[15]—an issue that both PAPI and ANCESTOR avoid. Another factor in PAPI's improved performance when $|p_A - p_B|$ is large may be that the difference between the true parent ancestry p_j and that transmitted to the focal individual has lower variance when p_A and/or p_B are near either end of the $[0, 1]$ range.

PAPI's ancestry estimates in the PAGE dataset add to a growing body of evidence for assortative mating by ancestry proportion in African Americans.[22] Others have noted these signals in Latinos[21] and found that they frequently cannot be explained by socioeconomic factors alone and are important to take into account when designing genetic studies. For example, using Wright-Fisher simulations that allow for ancestry-based assortative mating, Zaitlen et al.[22] showed that assuming random-mating leads to significant underestimates of population-scale admixture times.

Overall, PAPI's applications include both the study of the population genetics of mating in African Americans and the potential to provide individual-level information to study subjects. Given this ability to infer parental ancestry proportions for individual samples, we expect PAPI to be of interest to direct-to-consumer genetic testing companies and to consumers themselves who wish to better understand their genetic heritage.

Data and code availability

The PAPI code is available at https://github.com/williamslab/papi. Genotype data for BioMe Biobank subset of the PAGE II dataset are available (dbGaP: phs000925.v1.p1), and the phased 1000 Genomes data used for local ancestry inference are publicly available at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/. The simulated genotype data supporting the current study have not been deposited in a public repository but can be reproduced using Ped-sim (https://github.com/williamslab/ped-sim), admix-simu (https://github.com/williamslab/admix-simu), scripts distributed with PAPI, and the publicly available HapMap CEU and YRI data (https://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2009-01_phaseIII/plink_format/).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.06.016.

Acknowledgments

We thank the two anonymous reviewers for their valuable feedback that improved this paper. Funding for this work was provided by NIH grant R35 GM133805. Computing was performed on a cluster administered by the Biotechnology Resource Center at Cornell University.

Declaration of interests

A.L.W. is an employee of and holds stock in 23andMe and is the owner of HAPI-DNA LLC.

Received: January 5, 2022
Accepted: June 24, 2022
Published: July 26, 2022

References

1. Ruhl, G.L., Hazel, J.W., Wright Clayton, E., and Malin, B.A. (2019). Public attitudes toward direct to consumer genetic testing. AMIA Annu. Symp. Proc. *2019*, 774.
2. Regalado, A. (2019). More than 26 million people have taken an at-home ancestry test. MIT Tech. Rev. *11*, 2019.
3. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., and Mountain, J.L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet. *96*, 37–53.
4. Han, E., Carbonetto, P., Curtis, R.E., Wang, Y., Granka, J.M., Byrnes, J., Noto, K., Kermany, A.R., Myres, N.M., Barber, M.J., et al. (2017). Clustering of 770, 000 genomes reveals post-colonial population structure of North America. Nat. Commun. *8*, 14238.
5. Micheletti, S.J., Bryc, K., Ancona Esselmann, S.G., Freyman, W.A., Moreno, M.E., Poznik, G.D., Shastri, A.J., 23andMe Research Team, and Mountain, J.L., et al. (2020). Genetic consequences of the transatlantic slave trade in the Americas. Am. J. Hum. Genet. *107*, 265–277.
6. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
7. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
8. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. *5*, e1000519.
9. Gravel, S. (2012). Population genetics models of local ancestry. Genetics *191*, 607–619.
10. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.
11. Waples, R.K., Hauptmann, A.L., Seiding, I., Jørsboe, E., Jørgensen, M.E., Grarup, N., Andersen, M.K., Larsen, C.V.L., Bjerregaard, P., Hellenthal, G., et al. (2021). The genetic history of Greenlandic-European contact. Curr. Biol. *31*, 2214–2219.e4.
12. Seldin, M.F., Pasaniuc, B., and Price, A.L. (2011). New approaches to disease mapping in admixed populations. Nat. Rev. Genet. *12*, 523–528.
13. Atkinson, E.G., Maihofer, A.X., Kanai, M., Martin, A.R., Karczewski, K.J., Santoro, M.L., Ulirsch, J.C., Kamatani, Y., Okada, Y., Finucane, H.K., et al. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. Nat. Genet. *53*, 195–204.
14. Zou, J.Y., Halperin, E., Burchard, E., and Sankararaman, S. (2015). Inferring parental genomic ancestries using pooled semi-Markov processes. Bioinformatics *31*, i190–i196.
15. Pei, J., Zhang, Y., Nielsen, R., and Wu, Y. (2020). Inferring the ancestry of parents and grandparents from genetic data. PLoS Comput. Biol. *16*, e1008065.
16. Nettelblad, C. (2012). Inferring haplotypes and parental genotypes in larger full sib-ships and other pedigrees with missing or erroneous genotype data. BMC Genet. *13*, 85.
17. Young, A.I., Moeen Nehzati, S., Lee, C., Benonisdottir, S., Cesarini, D., Benjamin, D.J., Turley, P., and Kong, A. (2020). Mendelian Imputation of Parental Genotypes for Genome-wide Estimation of Direct and Indirect Genetic Effects. Preprint at bioRxiv. https://doi.org/10.1101/2020.07.02.185199.
18. Hwang, L.-D., Tubbs, J.D., Luong, J., Lundberg, M., Moen, G.-H., Wang, G., Warrington, N.M., Sham, P.C., Cuellar-Partida, G., and Evans, D.M. (2020). Estimating indirect parental genetic effects on offspring phenotypes using virtual parental genotypes derived from sibling and half sibling pairs. PLoS Genet. *16*, e1009154.
19. Johnson, N.A., Coram, M.A., Shriver, M.D., Romieu, I., Barsh, G.S., London, S.J., and Tang, H. (2011). Ancestral components of admixed genomes in a Mexican cohort. PLoS Genet. *7*, e1002410.
20. Howe, L.J., Battram, T., Morris, T.T., Hartwig, F.P., Hemani, G., Davies, N.M., and Smith, G.D. (2021). Assortative mating and within-spouse pair comparisons. PLoS Genet. *17*, e1009883.
21. Zou, J.Y., Park, D.S., Burchard, E.G., Torgerson, D.G., Pino-Yanes, M., Song, Y.S., Sankararaman, S., Halperin, E., and Zaitlen, N. (2015). Genetic and socioeconomic study of mate choice in Latinos reveals novel assortment patterns. Proc. Natl. Acad. Sci. USA. *112*, 13621–13626.
22. Zaitlen, N., Huntsman, S., Hu, D., Spear, M., Eng, C., Oh, S.S., White, M.J., Mak, A., Davis, A., Meade, K., et al. (2017). The

effects of migration and assortative mating on admixture linkage disequilibrium. Genetics *205*, 375–383.

23. Liang, M., and Nielsen, R. (2014). The lengths of admixture tracts. Genetics *197*, 953–967.
24. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). A discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
25. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. Bioinformatics *28*, 1359–1367.
26. Caballero, M., Seidman, D.N., Qiao, Y., Sannerud, J., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., Carmi, S., and Williams, A.L. (2019). Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. PLoS Genet. *15*, e1007979.
27. International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
28. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
29. Pfaffelhuber, P., Sester-Huss, E., Baumdicker, F., Naue, J., Lutz-Bonengel, S., and Staubach, F. (2022). Inference of recent admixture using genotype data. Forensic Sci. Int. Genet. *56*, 102593.
30. Smith, M.W., Patterson, N., Lautenberger, J.A., Truelove, A.L., McDonald, G.J., Waliszewska, A., Kessing, B.D., Malasky, M.J., Scafe, C., Le, E., et al. (2004). A high-density admixture map for disease gene discovery in African Americans. Am. J. Hum. Genet. *74*, 1001–1013.
31. Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., and Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet. *7*, e1001373.
32. Nelson, D., Kelleher, J., Ragsdale, A.P., Moreau, C., McVean, G., and Gravel, S. (2020). Accounting for long-range correlations in genome-wide simulations of large cohorts. PLoS Genet. *16*, e1008619.
33. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.; SciPy 1.0 Contributors (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods *17*, 261–272.
34. Salvatier, J., Wiecki, T.V., and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. PeerJ Comp. Sci., *2*, e55.
35. Bhérer, C., Campbell, C.L., and Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. Nat. Commun. *8*, 14994.
36. Williams, A. (2016). Admix-Simu: Program To Simulate Admixture Between Multiple Populations. https://github.com/williamslab/admix-simu.git.
37. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598. Database issue):D590–D598.
38. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.
39. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.
40. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
41. 1000 Genomes Project Consortium, Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

The American Journal of Human Genetics, Volume *109*

Supplemental information

Simultaneous inference of parental admixture proportions and admixture times from unphased local ancestry calls

Siddharth Avadhanam and Amy L. Williams

Supplemental Data for Simultaneous inference of parental admixture proportions and admixture times from unphased local ancestry calls

Siddharth Avadhanam[1] and Amy L. Williams[1,*]

[1]Department of Computational Biology, Cornell University, Ithaca, NY 14853, USA

Scenarios Settings	**Scenario (1)**	**Scenario (2)**	**Scenario (3)**
Admixture time	$t_A = t_B = t$; $t \in \{2, 9\}$	$t_A > t_B$; $t_A = 9, t_B \in \{2, 8\}$	$(t_A, t_B) \in$ $\{(0, 0), (1, 1)\}$
Admixture proportions and number of individuals	22 individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5), (0.25, 0.5), (0.4, 0.6), (0.25, 0.75)\}$	22 individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5), (0.25, 0.5), (0.4, 0.6), (0.25, 0.75)\}$	22 individuals for each of $E[(p_A, p_B)] \in \{(0, 0), (1, 1)\}$; 44 individuals for $E[(p_A, p_B)] = (1, 0)$; 22 individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5), (0, 0.5)\}$

Table S1: **Simulation scenarios.** Summary of all the admixture scenarios simulated in this paper. Columns are the scenario labels, the first row gives the admixture time settings, and the second row provides the number of individuals and admixture proportion setting simulated *for each* setting of (t_A, t_B).

Model Setting Inference Mode	**Binomial**	**Binomial w/Error Model**	**HMM**	**HMM w/Error Model**	**Full**	**Full w/Error Model**
GD	7.3s	–	21.7s	23.6s	23.9s	27.0s
MCMC	5m 23s	–	8h 36m 49s	8hr 58m 31s	8hr 5m 58s	8hr 55m 6s

Table S2: **PAPI runtimes for each model and inference mode.** Times are the per-individual average from analyzing simulated samples with $t_A = t_B = 5$, including 22 individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5), (0.25, 0.5), (0.4, 0.6), (0.25, 0.75)\}$, for a total of 88 individuals per run. Table gives wall clock times from compute nodes with Intel Xeon E5 4620 processors utilizing 4 GB of RAM.

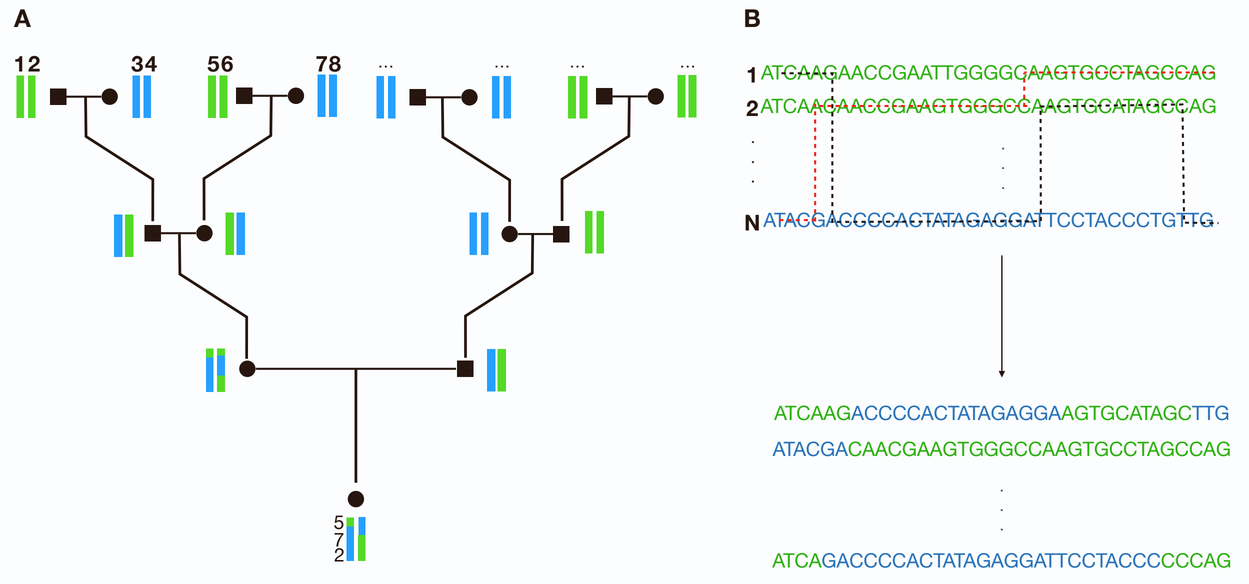
A
12
34
56
78
...
...
...
...
5
7
2
B
1 ATCAAGAACCGAATTGGGGCAAGTGCCTAGCCAG
2 ATCAAGAACCGAAGTGGGCCAAGTGCATAGCCAG
N ATACGACCCCACTATAGAGGATTCCTACCCTGTTG
ATCAAGACCCCACTATAGAGGAAGTGCATAGCTTG
ATACGACAACGAAGTGGGCCAAGTGCCTAGCCAG
ATCAGACCCCACTATAGAGGATTCCTACCCCCCAG

Figure S1: **Overview of simulation procedure.** (A) Given a pedigree topology (specified by (t_A, t_B)), Ped-sim generates a breakpoint file for the simulated samples containing recombination break points between founder haplotypes, each of which have a unique id number. Subsequently, we assign these haplotype ids to population labels according to the desired values of $E[(p_A, p_B)]$ and (t_A, t_B). In the example depicted here, $(t_A, t_B) = (2, 1)$ and $E[(p_A, p_B)] = (0.5, 0.5)$, where parent A is on the left and B is on the right. (B) Given the adapted breakpoint file containing population labels, admix-simu generates offspring haplotypes by sampling non-overlapping paths (black and red dashed lines representing the sampling paths of two haplotypes) from the input unadmixed haplotypes.

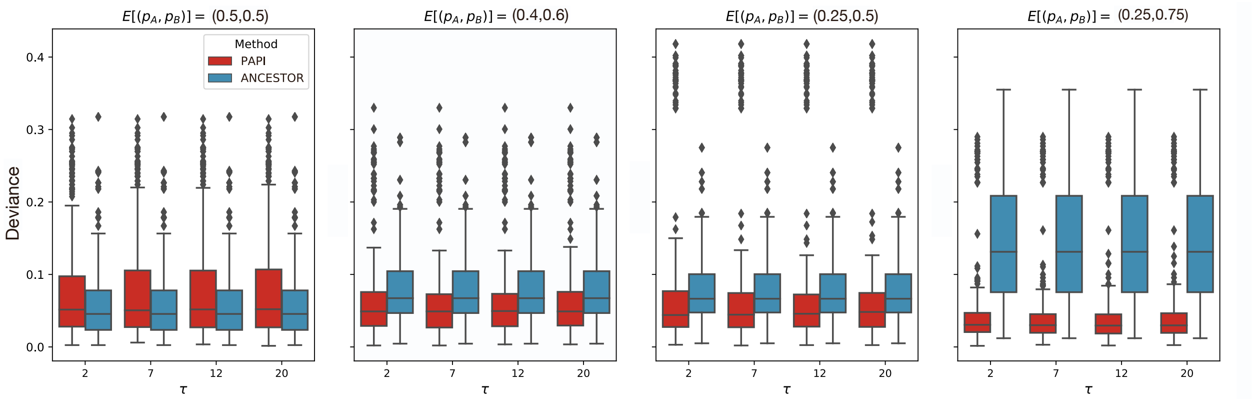

Figure S2: **Comparison of PAPI's admixture proportion estimates using the full model under different settings of τ.** Each panel displays box plots of $d_{abs,p}$ from PAPI and ANCESTOR versus τ for one of the settings of $E[(p_A, p_B)]$ for all scenario (1) and (2) data. PAPI is strongly robust to variable specification of τ. Note that ANCESTOR results are the same for each x-axis tick, as ANCESTOR does not utilize τ.

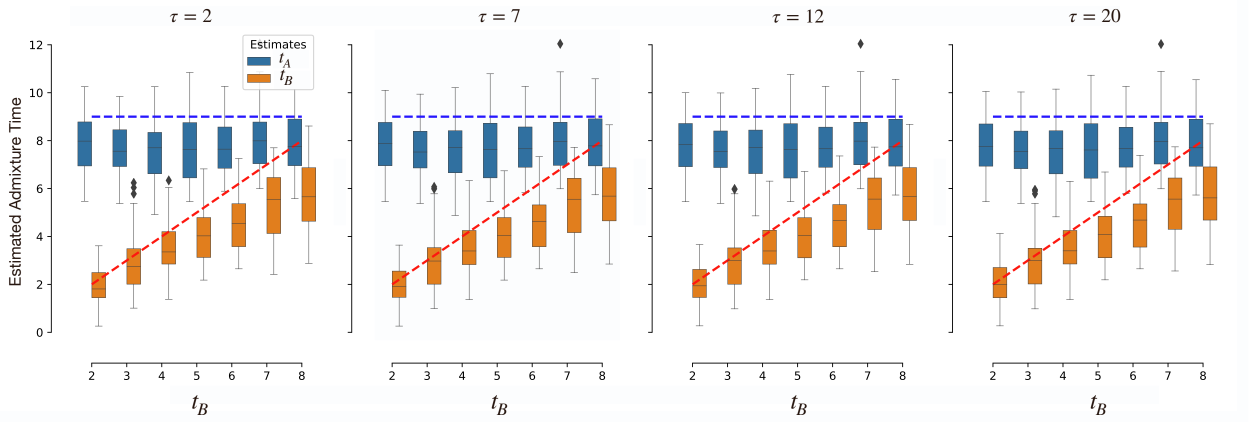

Figure S3: **Comparison of PAPI's admixture time estimates using the full model under different settings of τ.** Each panel displays box plots of PAPI's time estimates versus t_B for a different setting of τ for all scenario (1) datapoints. PAPI's results are largely unchanged regardless of the specification of τ.

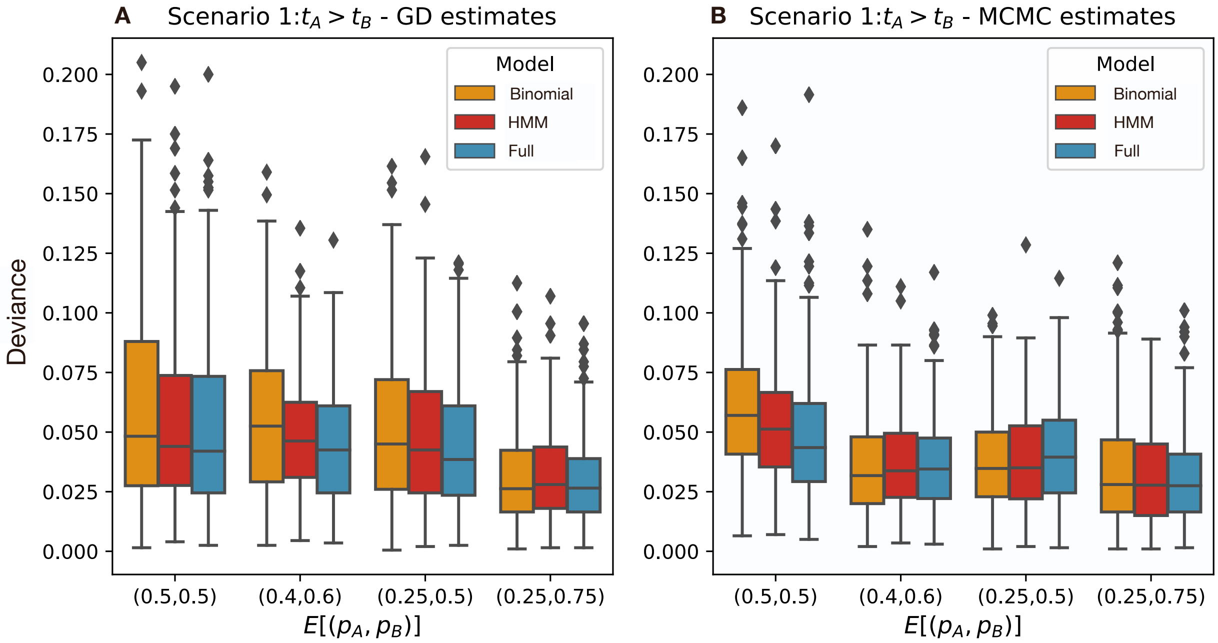

Figure S4: **Deviance statistics for inferred parent ancestry proportions** (p_A, p_B) **for each of PAPI's component models.** Average absolute deviances ($d_{abs,p}$) on all scenario (1) data under the (A) GD and (B) MCMC inference modes.

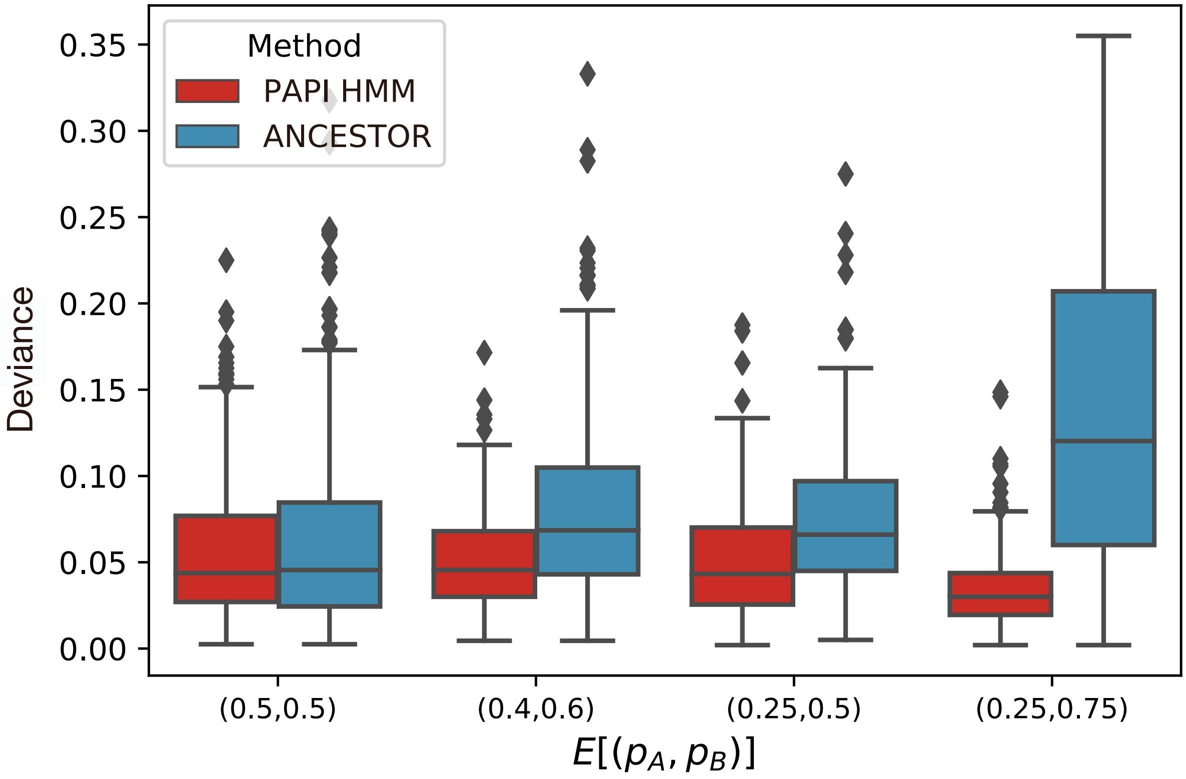

Figure S5: **Deviance statistics for inferred parent ancestry proportions** (p_A, p_B) **from ANCESTOR and PAPI's HMM alone for all scenario (1) and (2) data points.**

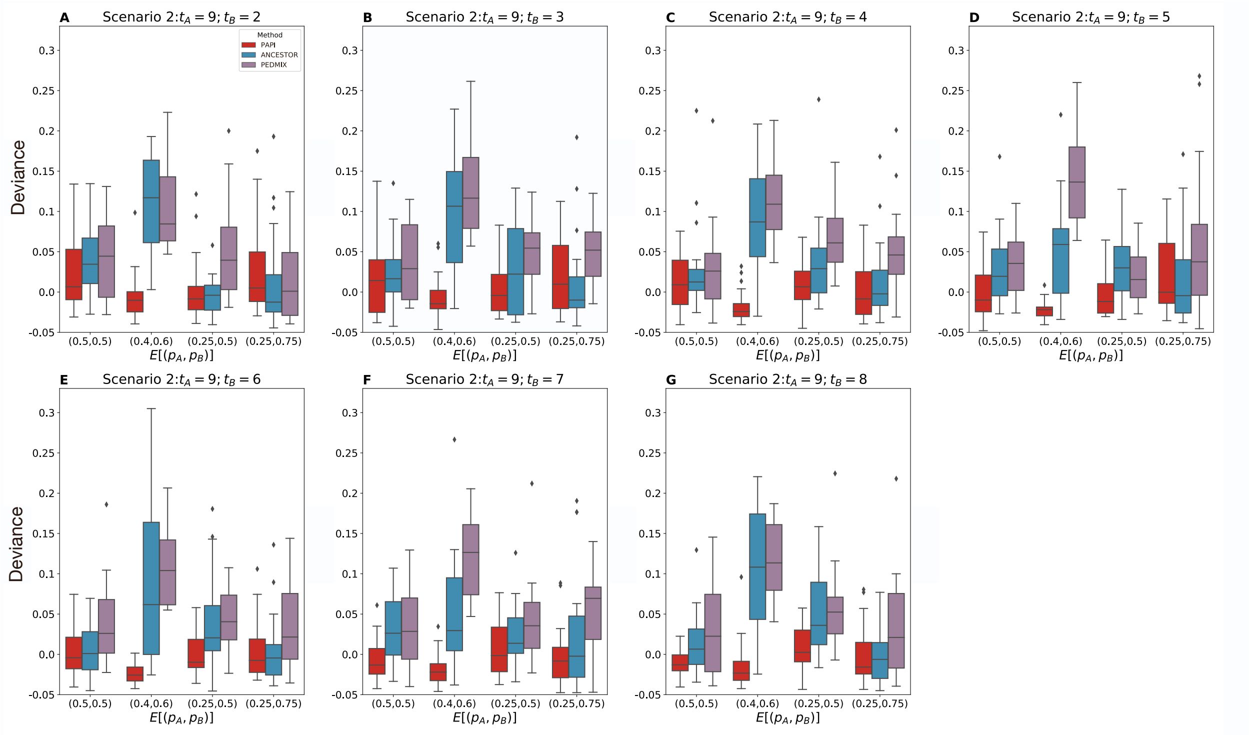

Figure S6: **Deviance statistics for inferred parent ancestry proportions** (p_A, p_B) **from PAPI, ANCESTOR and PedMix for each scenario (1) parameter setting.** (A-G) Average absolute deviances $(d_{abs,p})$ for each value of (t_A, t_B) as indicated in the panel title.

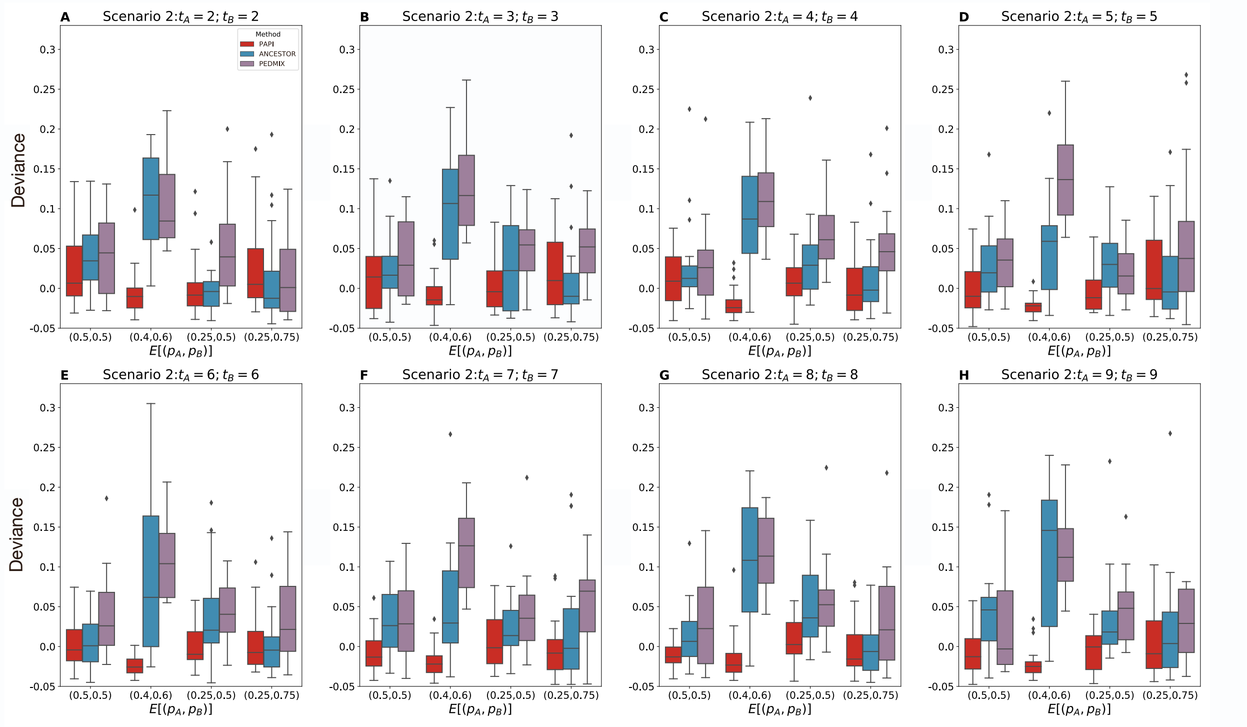

Figure S7: **Deviance statistics for inferred parent ancestry proportions (p_A, p_B) from PAPI, ANCESTOR and PedMix for each scenario (2) parameter setting.** (A-H) Average absolute deviances $(d_{abs,p})$ for each value of (t_A, t_B) as indicated in the panel title.

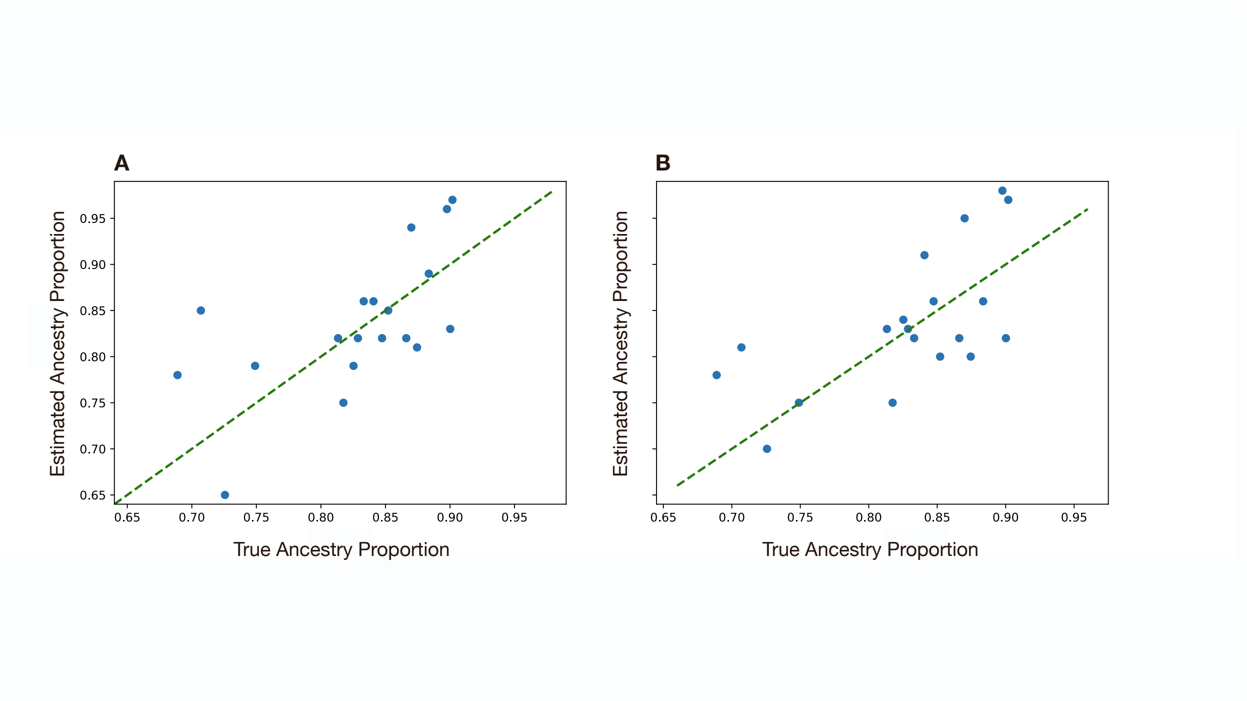

Figure S8: **Accuracy of admixture proportion estimates for PAPI's binomial model versus PAPI's HMM in the HapMap ASW trio children.** Scatter plots of estimated versus true parent ancestry proportions p_j for (A) estimates from PAPI's binomial model (B) estimates from PAPI's HMM.

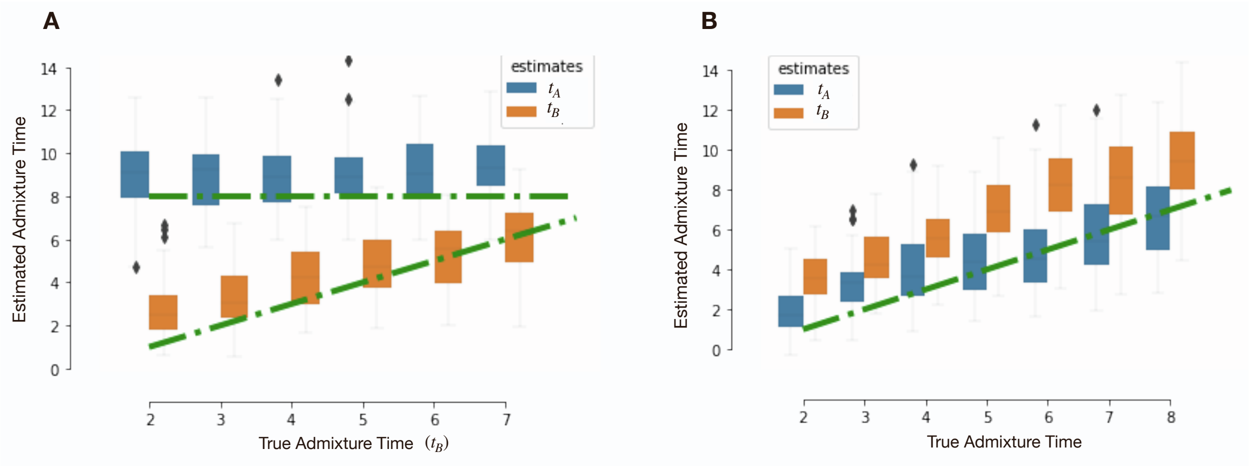

Figure S9: **PAPI's estimated time since admixture in simulated individuals using local ancestry tracts inferred by LAMP-LD.** Box plots of estimated times for (A) scenario (1) samples (where $t_A > t_B$) and (B) scenario (2) data (where $t_A = t_B$).

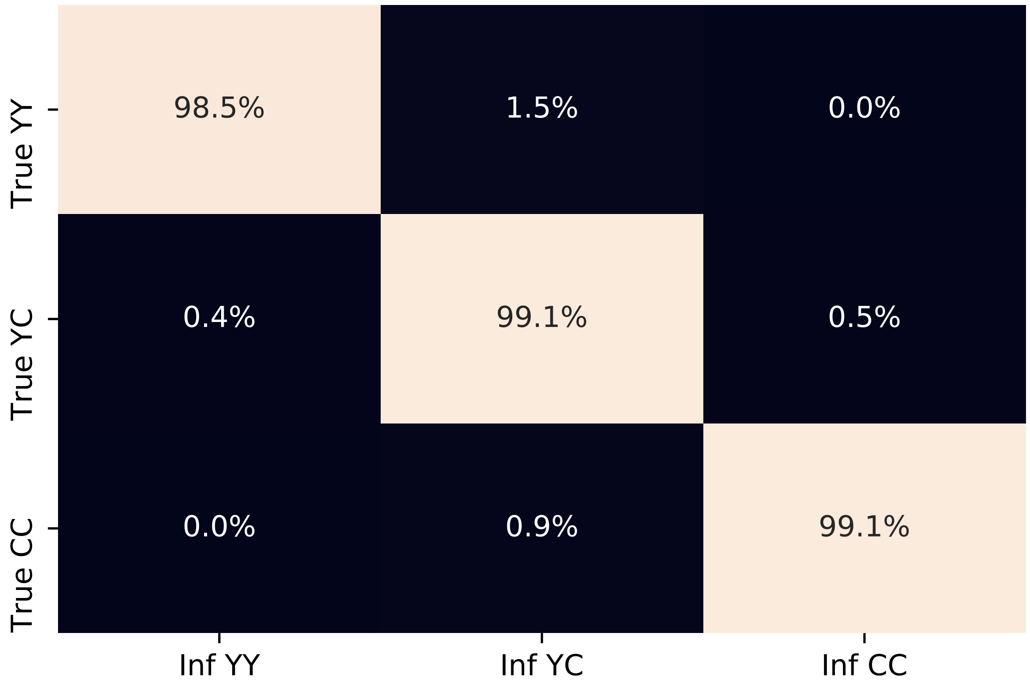

Figure S10: **Per-marker confusion matrix of true diploid ancestries versus HAPMIX-inferred diploid ancestries.** HAPMIX shows a high true positive rate (> 98.5%). Matrix includes across all scenario (1) and (2) simulated data points. The rows are labeled true ancestry states, and the columns are inferred ancestry states, with Y and C representing HapMap YRI (African) and CEU (European) ancestry, respectively. For example, the cell at row 1, column 1 contains the percentage of true YY calls (diploid African) that were inferred to be YY.

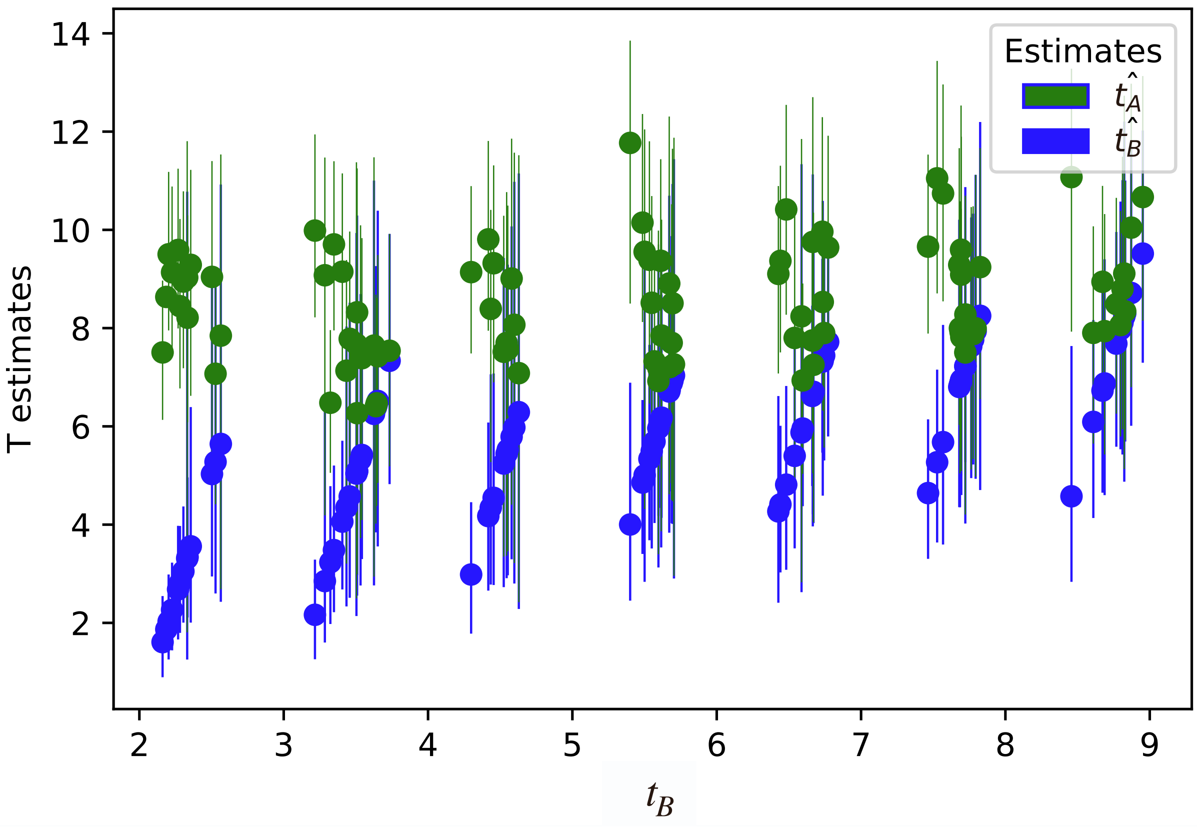

Figure S11: **Admixture time estimates with credible intervals for scenario (1) data points.** Plot shows point estimates along with their 90% credible intervals from PAPI's MCMC inference mode run under the full model. The x-axis coordinate for each point is shifted slightly from the truth to aid visualization: the estimate pairs $(\hat{t}_A, \hat{t}_B)$ are sorted by the value of $\hat{t}_B$.

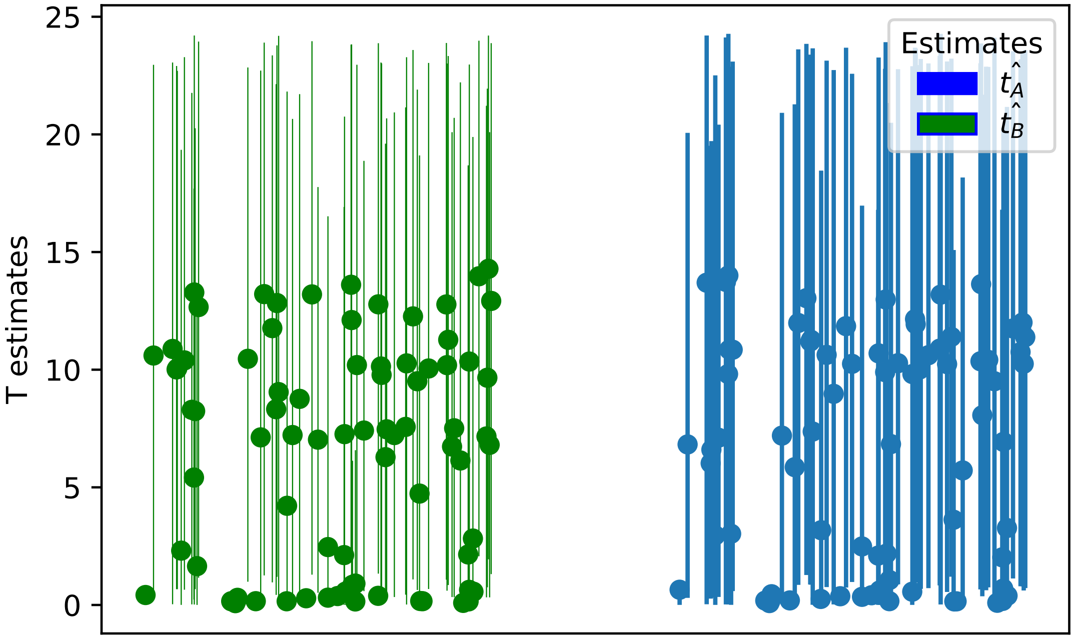

Figure S12: **Admixture time estimates with credible intervals for the subset of scenario (3) data points with $t_A = t_B = 0$.** Plot shows point estimates along with their 90% credible intervals from PAPI's MCMC inference mode run under the full model. Many estimates are biased and have very large 90% credible intervals because of a small number of erroneous tracts.

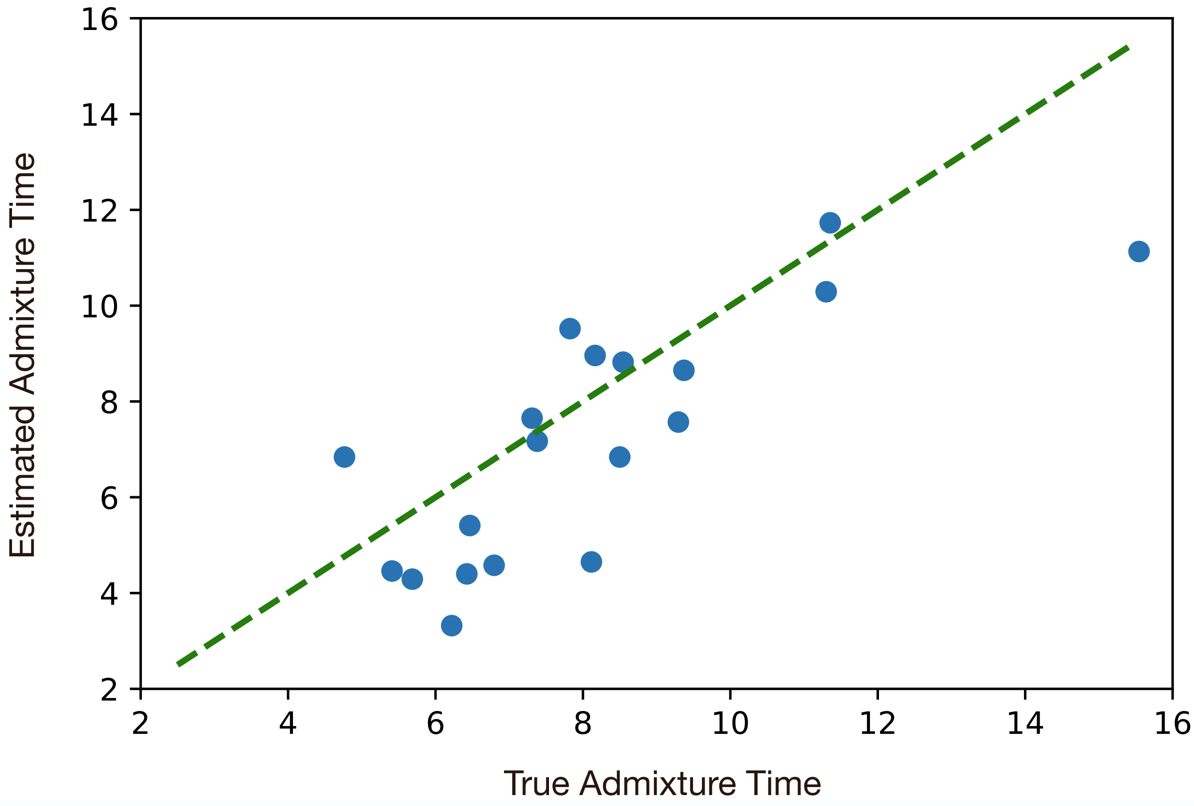

Figure S13: **Accuracy of admixture time estimates for PAPI's full model in the HapMap ASW trio children.** Scatter plots of estimated versus true parent admixture proportions t_j for PAPI's full model.

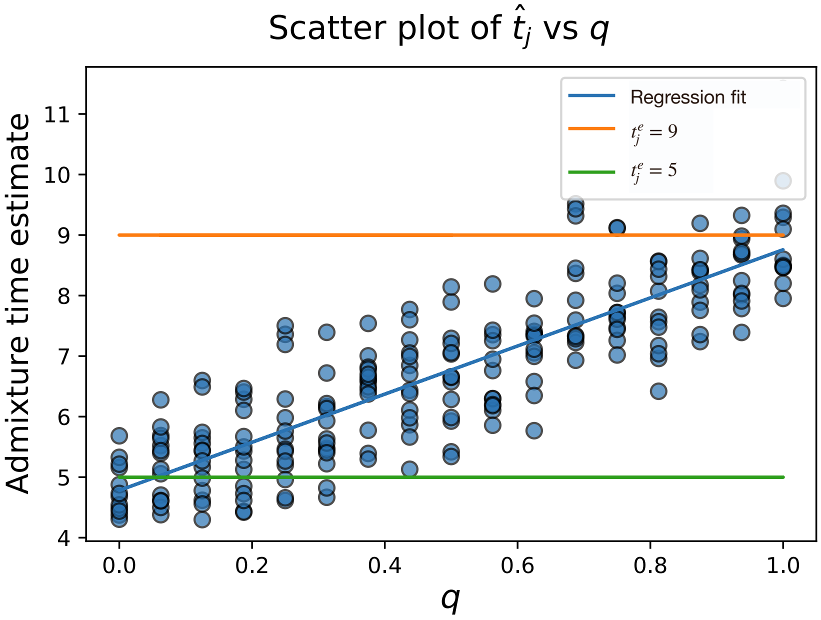

Figure S14: **PAPI's estimated time since admixture with two migrant pulses where $t_j^e = 9$ and $t_j^l = 5$.** Plot of estimated time since admixture versus q (the proportion of unadmixed couples in the t_j^e generation of different ancestries) when $E[p_j] = 0.5$. The blue line depicts the least-squares regressions on the plotted data points.

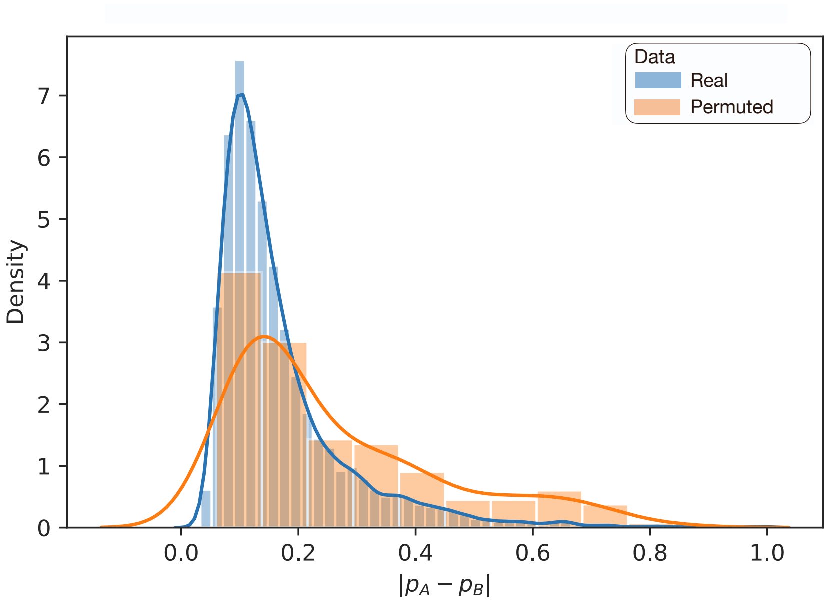

Figure S15: **Real versus simulated random mating distribution of $|p_A - p_B|$ in the PAGE dataset.** The orange distribution uses parent ancestry proportions randomly sampled from the inferred $\hat{p}_j$ values from the PAGE African Americans.