The American Journal of Human Genetics, Volume *110*

## Supplemental information

# Fast, accurate local ancestry inference with FLARE

**Sharon R. Browning, Ryan K. Waples, and Brian L. Browning**

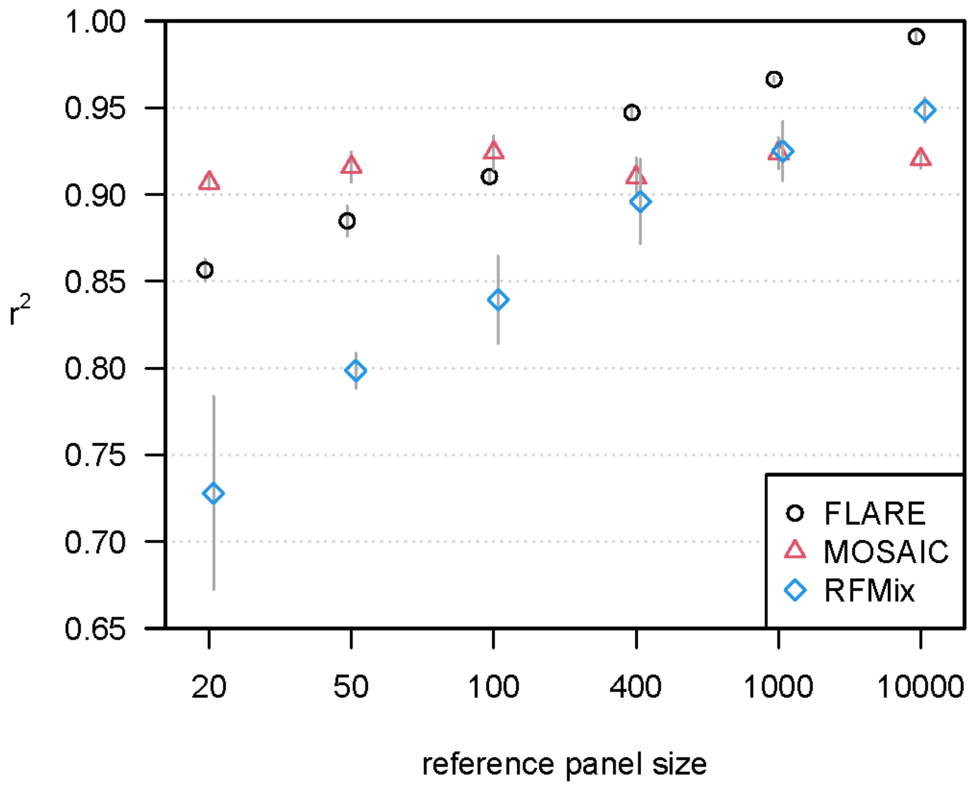

**Figure S1: Accuracy when increasing reference panel size for simulated array data with three-way admixture.** The y-axis shows squared correlation between true and inferred local ancestry dose averaged over ancestries and across four replicate simulations (details in Methods). Error bars (+/- 2 standard errors) are shown as gray lines. Each of the three ancestries is represented by a reference panel of size shown on the x-axis. Each analysis includes 100 admixed individuals.

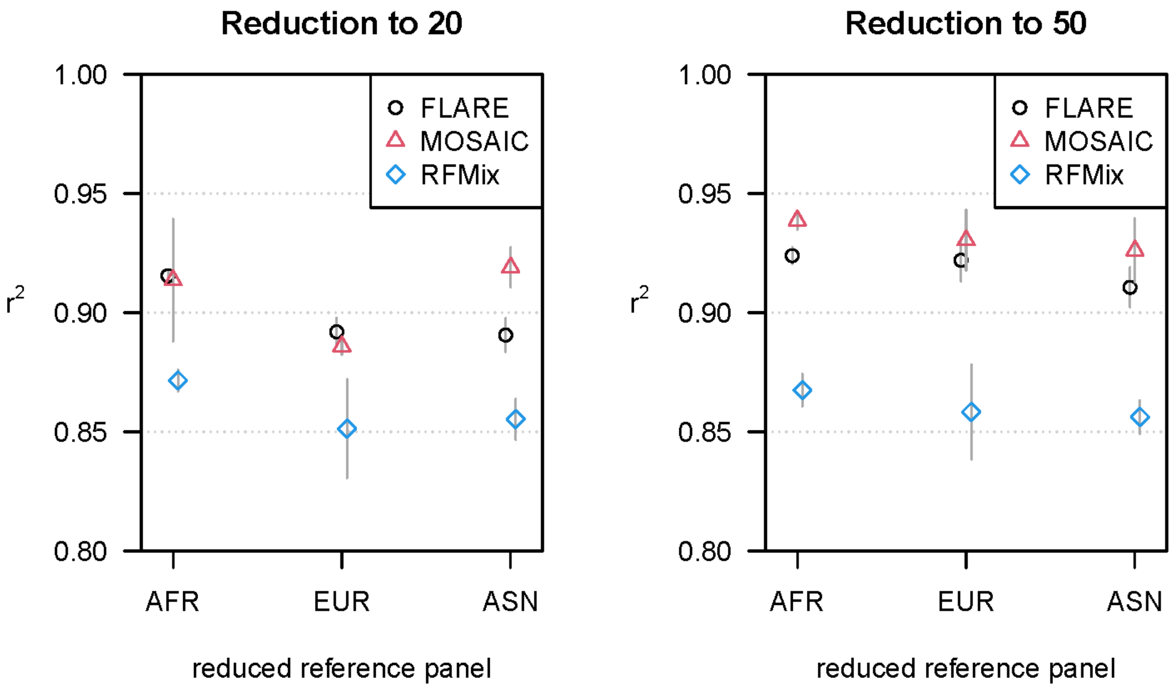

**Figure S2: Accuracy when reducing the size of one of three reference panels.** The data are simulated sequence data for three-way admixture. Each reference panel has size 400, except for the reference panel that is denoted on the x-axis (AFR is West African, EUR is European, ASN is East Asian) which has size 20 (left plot) or 50 (right plot). The y-axis shows squared correlation between true and inferred local ancestry dose averaged over ancestries and across four replicate simulations (details in Methods). Error bars (+/- 2 standard errors) are shown as gray lines. Each analysis includes 100 admixed individuals.

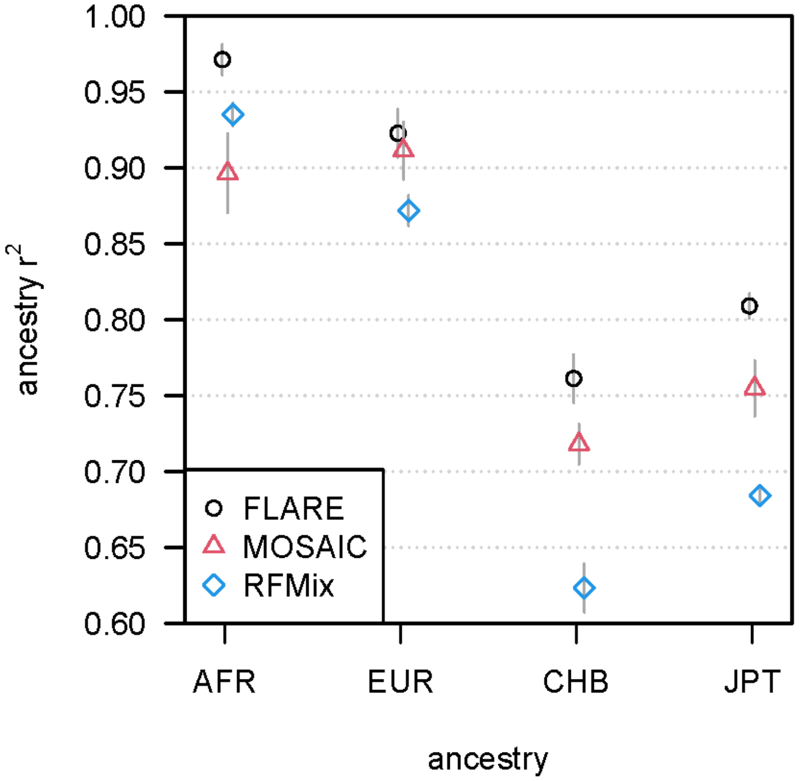

**Figure S3: Accuracy by ancestry for the simulated array data with four-way admixture**. The y-axis is the squared correlation between the true and inferred ancestry dose for a single ancestry averaged across four replicate simulations. Error bars (+/- 2 standard errors) are shown as gray lines. The ancestry is shown on the x-axis (AFR is simulated West African, EUR is simulated European, CHB is simulated Han Chinese, JPT is simulated Japanese). The simulated array data have 100 admixed individuals and 400 individuals in each of the four reference panels.

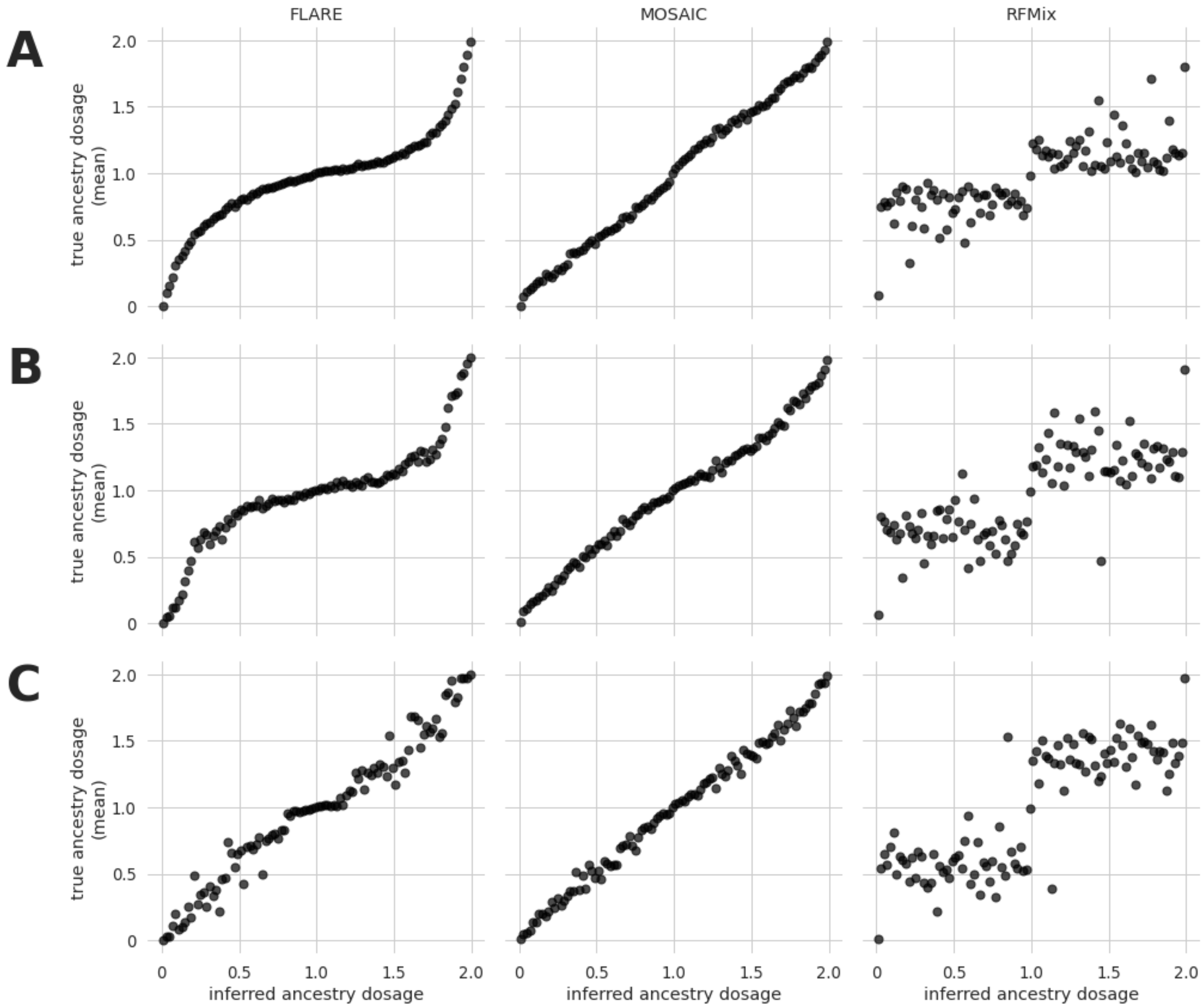

**Figure S4: Calibration of estimated diploid ancestry dose on simulated three-way admixture data.** Estimated diploid ancestry dose is binned into bins of width 0.02 along the x-axis. The y-axis is the average true diploid ancestry dose for each bin. Results for FLARE, MOSAIC, and RFMix are shown in the left, middle, and right panels respectively. Sequence and array data are combined in these plots and reference panel sizes are combined in three size groups to reduce noise. A) Reference panels of sizes 20, 50, and 100. B) Reference panels of sizes 400 and 1000. C) Reference panels of size 10,000.

# Supplementary Methods 1: Transition probabilities

The transition probabilities described in the main text can be expressed as:

$$P\big(S_m = (i', h')\big|S_{m-1} = (i, h)\big) \\ = \begin{cases} \big(1 - e^{-d_m T}\big)\mu_{i'} q_{i'h'} + e^{-d_m T}\big(1 - e^{-d_m \rho_i}\big)q_{i'h'} + e^{-d_m T} e^{-d_m \rho_i} & i = i', h = h' \\ \big(1 - e^{-d_m T}\big)\mu_{i'} q_{i'h'} + e^{-d_m T}\big(1 - e^{-d_m \rho_i}\big)q_{i'h'} & i = i', h \neq h' \\ \big(1 - e^{-d_m T}\big)\mu_{i'} q_{i'h'} & i \neq i' \end{cases}$$

# Supplementary Methods 2: Algorithm for posterior probabilities of ancestry

We estimate the posterior ancestry probabilities using the hidden Markov model forward-backward algorithm.[1]

Consider an admixed haplotype, $\boldsymbol{Y}$. Let $Y_m$ be the allele at marker $m$, with markers indexed $1, \ldots, M$. The forward probabilities are

$$\alpha_m(i, h) = P(Y_1, \ldots, Y_m, S_m = (i, h)) \qquad \text{(S1)}$$

where $S_m$ represents the (ancestry, haplotype) state at the $m$th marker. The backward probabilities are

$$\beta_m(i, h) = P\big(Y_{m+1}, \ldots, Y_M \big| S_m = (i, h)\big). \qquad \text{(S2)}$$

*Forward probabilities at first marker*: For each ancestry $i$ and haplotype $h$,

$$\alpha_1(i, h) = \pi(i, h) e_1(i, h).$$

where $\pi(i, h)$ is the prior probability that the state is $(i, h)$, and the emission probability $e_m(i, h)$ is the probability of observing the allele $Y_m$ at marker $m$ on the admixed haplotype when the hidden state at this marker is $(i, h)$.

*Forward probabilities*: Suppose we have already calculated $\alpha_{m-1}(i, h)$ for all $(i, h)$, and we want to calculate $\alpha_m(i', h')$. Let $d_m$ be the distance in Morgans between markers $m - 1$ and $m$. Pre-calculate

$$f_i = \sum_h \alpha_{m-1}(i, h)$$

for each $i$, and

$$s_f = \sum_i f_i.$$

The values $f_i$ and $s_f$ are temporary variables that are over-written for each successive marker. Their purpose is to avoid duplicate calculation.

Then for each $i'$ and $h'$ calculate (using equation S1)

$$\begin{aligned}\alpha_m(i',h') &= e_m(i',h')\sum_{i,h} P\big(S_m = (i',h')\big|S_{m-1} = (i,h)\big)\alpha_{m-1}(i,h) \\ &= e_m(i',h')\big[\big(1-e^{-d_mT}\big)\mu_{i'}q_{i'h'}s_f + e^{-d_mT}\big(1-e^{-d_m\rho_{i'}}\big)q_{i'h'}f_{i'} \\ &+ e^{-d_mT}e^{-d_m\rho_{i'}}\alpha_{m-1}(i',h')\big].\end{aligned}$$

In the computation, we normalize the $\alpha_m(i',h')$ to sum to one and store the normalization factors in order to avoid numerical underflow.

*Backwards probabilities*: Let $\beta_M(i,h) = 1$ for all ancestries $i$ and reference haplotypes $h$.

Suppose the $\beta_{m+1}(i,h)$ values have been calculated for all ancestries $i$ and reference haplotypes $h$. Let $d_{m+1}$ be the distance in Morgans between markers $m$ and $m+1$. Pre-calculate

$$b_i = \sum_h \beta_{m+1}(i,h)q_{ih}\,e_{m+1}(i,h)$$

for each $i$, and $s_b = \sum_i b_i\mu_i$

The values $b_i$ and $s_b$ are temporary variables that are over-written for each successive marker. Their purpose is to avoid duplicate calculation.

Then for each $i$ and $h$, calculate (using equation S2)

$$\begin{aligned}\beta_m(i,h) &= \sum_{i',h'} e_{m+1}(i',h')P\big(S_m = (i',h')\big|S_{m-1} = (i,h)\big)\beta_{m+1}(i',h') \\ &= \big(1-e^{-d_{m+1}T}\big)s_b + e^{-d_{m+1}T}\big(1-e^{-d_{m+1}\rho_i}\big)b_i + e^{-d_{m+1}T}e^{-d_{m+1}\rho_i}\beta_{m+1}(i,h)e_{m+1}(i,h)\end{aligned}$$

In the computation, we normalize the values of $\beta_m(i,h)$ to sum to one and store the normalization factors to avoid numerical underflow.

*Posterior probability of ancestry*:

Let

$$v_m(i) = \sum_h \alpha_m(i,h)\beta_m(i,h)$$

The posterior probability of ancestry $i$ at marker $m$ is $w_m(i) = v_m(i)/\sum_{i'} v_m(i')$.

## Supplementary Methods 3: Initialization and updating parameter values

The initial values of the parameters are set as described below, or as specified by the user. If the EM updating option is turned on (which it is by default), we update parameters using a variant of the Baum-Welch algorithm.[1]Each EM iteration estimates local ancestry for 100 randomly selected admixed haplotypes (using a separate random selection for each EM iteration) and the ancestry proportions and admixture time are updated as described below. Twenty EM iterations are performed unless the EM updating converges sooner. Convergence is defined as a relative change less than 5% in each ancestry proportion $\mu_i$ from the value in the preceding iteration, excluding those ancestries for which $\mu_i < 0.001$. A 5% relative change in a $\mu_i$ taking value of 0.1 in the previous iteration would be 0.005.

**Mismatch probabilities $\theta_{i,j}$:**

The default mismatch probabilities are the same for each ancestry and panel, and are defined as: $\theta_{i,j} = \lambda/(2\lambda + 2N)$ where $\lambda = 1/(\log N + 0.5)$ and $N$ is the total number of reference haplotypes.[2] We do not update this parameter.

**Panel probabilities $p_{ij}$ and switch rates $\rho_i$:**

The panel probabilities are obtained via a single iteration of training on the reference panel. Considering ancestry $i^*$, which is represented by one reference panel, we take one haplotype at a time out of that reference panel and run the forwards-backwards algorithm using all other reference haplotypes. For this analysis we set $\mu_{i^*} = 1$, $\mu_i = 0$ for $i \neq i^*$, $T = 0$, and $p_{i^*j} = n_j/N$ where $n_j$ is the number of reference haplotypes in panel $j$. We use the default mismatch probabilities $\theta_{ij}$ defined in the preceding section, and we set $\rho_i = 4N_e/N$ where $N_e = 50{,}000$.[2; 3] We perform the analysis for 100 haplotypes selected at random from the reference panel.

The updated panel probability is the average posterior probability that the copied haplotype is from panel $j$, given that the ancestry is $i$. The posterior probability for state $(i,h)$ at marker $m$ for selected reference haplotype $k$ is proportional to $\alpha_{m,k}(i,h)\beta_{m,k}(i,h)$. That is, the posterior probability for state $(i,h)$ is

$$\sum_{h \text{ in panel } j} \alpha_{m,k}(i,h)\beta_{m,k}(i,h) \Big/ \sum_h \alpha_{m,k}(i,h)\beta_{m,k}(i,h)$$

and we average this over markers $m$ and selected reference haplotypes indexed by $k$ to obtain the estimated panel probability

$$\hat{p}_{ij} = \sum_{m,k}\left(\sum_{h \text{ in panel } j} \alpha_{m,k}(i,h)\beta_{m,k}(i,h) \Big/ \sum_h \alpha_{m,k}(i,h)\beta_{m,k}(i,h)\right) \Bigg/ \sum_{m,k} 1$$

The updated switch rate $\rho_i$ is determined from the posterior probabilities of a change of haplotype state, as follows:

The probability of transitioning to the same state is:

$$\begin{aligned} P\big(S_m = (i,h) \big| S_{m-1} = (i,h)\big) &= \big(1 - e^{-d_m T}\big)\mu_i q_{ih} + e^{-d_m T}\big(1 - e^{-d_m \rho_i}\big) q_{ih} + e^{-d_m T} e^{-d_m \rho_i} \\ &= \big(1 - e^{-d_m T}\big)\mu_i q_{ih} + e^{-d_m T} - e^{-d_m T}\big(1 - e^{-d_m \rho_i}\big)(1 - q_{ih}) \end{aligned}$$

Solving for $\big(1 - e^{-d_m \rho_i}\big)$ gives:

$$1 - e^{-d_m \rho_i} = \frac{\big(1 - e^{-d_m T}\big)\mu_i q_{ih} + e^{-d_m T} - P\big(S_m = (i,h) \big| S_{m-1} = (i,h)\big)}{e^{-d_m T}(1 - q_{ih})} \qquad \text{(S3)}$$

We write $\tau_{m,i} = 1 - e^{-d_m \rho_i}$. We estimate $\tau_{m,i}$ using the observed transition probabilities in place of the prior transition probabilities $P\big(S_m = (i,h)\big|S_{m-1} = (i,h)\big)$:

$$P(S_m = (i,h)|S_{m-1} = (i,h), \boldsymbol{Y}) = \frac{P(S_m = (i,h), S_{m-1} = (i,h), \boldsymbol{Y})}{P(S_{m-1} = (i,h), \boldsymbol{Y})}$$

We average over haplotype state $h$, weighting by the observed state probabilities conditional on ancestry $i$,

$$\frac{P(S_{m-1} = (i,h)|\boldsymbol{Y})}{\sum_{h'} P(S_{m-1} = (i,h')|\boldsymbol{Y})} = \frac{P(S_{m-1} = (i,h), \boldsymbol{Y})}{\sum_{h'} P(S_{m-1} = (i,h'), \boldsymbol{Y})},$$

in the right-hand side of equation S3 to obtain:

$$\hat{\tau}_{m,i} = \sum_{h=1}^{H} \frac{\left(1 - e^{-d_m T}\right)\mu_i q_{ih} + e^{-d_m T} - P(S_m = (i,h)|S_{m-1} = (i,h), \boldsymbol{Y})}{e^{-d_m T}(1 - q_{ih})} \frac{P(S_{m-1} = (i,h), \boldsymbol{Y})}{\sum_{h'} P(S_{m-1} = (i,h'), \boldsymbol{Y})}$$
$$= \sum_{h=1}^{H} \frac{\left(1 - e^{-d_m T}\right)\mu_i q_{ih} + e^{-d_m T} - P(S_m = (i,h)|S_{m-1} = (i,h), \boldsymbol{Y})}{e^{-d_m T}(1 - q_{ih}) \sum_{h'} P(S_{m-1} = (i,h'), \boldsymbol{Y})} P(S_{m-1} = (i,h), \boldsymbol{Y})$$
$$= \sum_{h=1}^{H} \frac{\left(\left(1 - e^{-d_m T}\right)\mu_i q_{ih} + e^{-d_m T}\right) P(S_{m-1} = (i,h), \boldsymbol{Y}) - P(S_m = (i,h), S_{m-1} = (i,h), \boldsymbol{Y})}{e^{-d_m T}(1 - q_{ih}) \sum_{h'} P(S_{m-1} = (i,h'), \boldsymbol{Y})}.$$

At each marker $m > 1$,

$$P(S_m = (i,h), S_{m-1} = (i,h), \boldsymbol{Y}) = \beta_m(i,h) e_m(i,h) P\big(S_m = (i,h)\big|S_{m-1} = (i,h)\big)\alpha_{m-1}(i,h)$$

and

$$P(S_{m-1} = (i,h), \boldsymbol{Y}) = \alpha_{m-1}(i,h)\beta_{m-1}(i,h)$$

We use the linear approximation $\tau_{m,i} = 1 - e^{-d_m \rho_i} \approx \rho_i d_m$ to estimate $\rho_i$. After we have estimated the $\hat{\tau}_{m,i,k}$ for each marker $m$ and each target haplotype $k$, we estimate $\rho_i$ with a slope estimator weighted by the conditional probability of ancestry $i$ given the data, $\sum_h P(S_{m-1} = (i,h)|\boldsymbol{Y})$:

$$\hat{\rho}_i = \frac{\sum_{m,k} \sum_h P(S_{m-1} = (i,h)|\boldsymbol{Y})\, \hat{\tau}_{m,i,k}}{\sum_{m,k} \sum_h P(S_{m-1} = (i,h)|\boldsymbol{Y})\, d_m}$$

Note that

$$\sum_h P(S_{m-1} = (i,h)|\boldsymbol{Y}) = \frac{\sum_h \alpha_{m-1}(i,h)\beta_{m-1}(i,h)}{\sum_{i'} \sum_h \alpha_{m-1}(i',h)\beta_{m-1}(i',h)}$$

After initializing the $p_{ij}$ and $\rho_i$, these parameters are fixed for the remainder of the analysis.

**Ancestry proportions, $\boldsymbol{\mu_i}$:** The default initial value is $1/A$, where $A$ is the number of ancestries. The updated value following each EM iteration is a weighted average of the posterior probability $w_m(i)$ for ancestry $i$. We include only positions for which the posterior probability of the ancestry is at least 0.9 in order to speed convergence. The selected haplotypes are indexed by $k$.

$$\hat{\mu}_i = \frac{\sum_{m,k} w_{m,k}(i) 1\{w_{m,k}(i) \geq 0.9\}}{\sum_{i'} \sum_{m,k} w_{m,k}(i') 1\{w_{m,k}(i') \geq 0.9\}}$$

**Admixture time $\boldsymbol{T}$:**

The default initial value of $T$ is 10 generations.

The updated admixture time is determined from the posterior probabilities of a change of ancestry state, as follows:

The probability of transitioning to the same ancestry state is

$$\sum_{h'} P\big(S_m = (i, h') \big| S_{m-1} = (i, h)\big) = \big(1 - e^{-d_m T}\big)\mu_i + e^{-d_m T}$$

Solving for $\big(1 - e^{-d_m T}\big)$ we obtain

$$\big(1 - e^{-d_m T}\big) = \frac{1 - \sum_{h'} P\big(S_m = (i, h') \big| S_{m-1} = (i, h)\big)}{1 - \mu_i}$$

We write $\gamma_m = 1 - e^{-d_m T}$. We estimate $\gamma_m$ using the observed transition probabilities in place of the prior transition probabilities $P\big(S_m = (i, h) \big| S_{m-1} = (i, h)\big)$:

$$P(S_m = (i, h') | S_{m-1} = (i, h), \boldsymbol{Y}) = \frac{P(S_m = (i, h'), S_{m-1} = (i, h), \boldsymbol{Y})}{P(S_{m-1} = (i, h'), \boldsymbol{Y})}$$

We average over haplotype state $h$ and ancestry $i$ at marker $m - 1$, weighting by the observed state probabilities:

$$\frac{P(S_{m-1} = (i, h) | \boldsymbol{Y})}{\sum_{i^*} \sum_{h^*} P(S_{m-1} = (i^*, h^*) | \boldsymbol{Y})} = \frac{P(S_{m-1} = (i, h), \boldsymbol{Y})}{\sum_{i^*} \sum_{h^*} P(S_{m-1} = (i^*, h^*), \boldsymbol{Y})}$$

to obtain

$$\begin{aligned} \hat{\gamma}_m &= \sum_i \sum_h \frac{1 - \sum_{h'} P(S_m = (i, h') | S_{m-1} = (i, h), \boldsymbol{Y})}{1 - \mu_i} \frac{P(S_{m-1} = (i, h), \boldsymbol{Y})}{\sum_{i^*} \sum_{h^*} P(S_{m-1} = (i^*, h^*), \boldsymbol{Y})} \\ &= \sum_i \sum_h \frac{1 - \sum_{h'} P(S_m = (i, h'), S_{m-1} = (i, h), \boldsymbol{Y}) / P(S_{m-1} = (i, h), \boldsymbol{Y})}{1 - \mu_i} \frac{P(S_{m-1} = (i, h), \boldsymbol{Y})}{\sum_{i^*} \sum_{h^*} P(S_{m-1} = (i^*, h^*), \boldsymbol{Y})} \\ &= \sum_i \sum_h \frac{P(S_{m-1} = (i, h), \boldsymbol{Y}) - \sum_{h'} P(S_m = (i, h'), S_{m-1} = (i, h), \boldsymbol{Y})}{(1 - \mu_i) \sum_{i^*} \sum_{h^*} P(S_{m-1} = (i^*, h^*), \boldsymbol{Y})}. \end{aligned}$$

At each marker $m > 1$,

$$\begin{aligned} &P(S_m = (i, h'), S_{m-1} = (i, h), \boldsymbol{Y}) \\ &\qquad = \beta_m(i, h') e_m(i, h') P\big(S_m = (i, h') \big| S_{m-1} = (i, h)\big) \alpha_{m-1}(i, h) \end{aligned}$$

and

$$P(S_{m-1} = (i,h), \boldsymbol{Y}) = \alpha_{m-1}(i,h)\beta_{m-1}(i,h)$$

After we have estimated the $\hat{\gamma}_{m,k}$ for each marker $m$ and each target haplotype $k$, we estimate the constant of proportionality $T$ in the relationship $\hat{\gamma}_{m,k} \approx Td_m$ as:

$$\hat{T} = \frac{\sum_{m,k} \hat{\gamma}_{m,k}}{\sum_{m,k} d_m}$$

## Supplementary References

1. Rabiner, L.R. (1989). A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77, 257-286.
2. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39, 906-913.
3. Browning, B.L., and Browning, S.R. (2016). Genotype imputation with millions of reference samples. Am J Hum Genet 98, 116-126.