Fast, accurate local ancestry inference with FLARE

Authors

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

Correspondence

sguy@uw.edu (S.R.B.), browning@uw.edu (B.L.B.)

Local ancestry is the ancestral origin at each point in the genome of an admixed individual. FLARE (fast local ancestry estimation) is a program for local ancestry inference that is highly accurate and computationally efficient. FLARE can analyze array and sequence data and can utilize large reference panels.

Browning et al., 2023, The American Journal of Human Genetics *110*, 326–335
February 2, 2023
https://doi.org/10.1016/j.ajhg.2022.12.010

CellPress

Fast, accurate local ancestry inference with FLARE

Sharon R. Browning,[1,*] Ryan K. Waples,[1] and Brian L. Browning[1,2,*]

Summary

Local ancestry is the source ancestry at each point in the genome of an admixed individual. Inferred local ancestry is used for admixture mapping and population genetic analyses. We present FLARE (fast local ancestry estimation), a method for local ancestry inference. FLARE achieves high accuracy through the use of an extended Li and Stephens model, and it achieves exceptional computational performance through incorporation of computational techniques developed for genotype imputation. Memory requirements are reduced through on-the-fly compression of reference haplotypes and stored checkpoints. Computation time is reduced through the use of composite reference haplotypes. These techniques allow FLARE to scale to datasets with hundreds of thousands of sequenced individuals and to provide superior accuracy on large-scale data. FLARE is open source and available at https://github.com/browning-lab/flare.

Introduction

All humans are admixtures of various historical source populations.[1] This admixture has occurred across a range of timescales, from the recent intercontinental admixture in African Americans and Hispanics[2–4] to the ancient admixture with Neanderthals that occurred when modern humans migrated out of Africa around 50,000 years ago.[5,6]

Local ancestry is the source ancestry of an individual's chromosomes at each point in the genome. Local ancestry can be inferred on cross-continental admixtures for recently admixed groups, such as admixed populations in the Americas which have admixed ancestry deriving from indigenous Americans, West Africans, and Western Europeans. With sensitive methods and appropriate reference panels, local ancestry of recent within-continental admixtures with less genetic divergence can also be inferred.[7] Indeed, some consumer genetics companies now report sub-continental level ancestry.[8]

Inferred local ancestry is required for admixture mapping. Admixture mapping tests for association between local ancestry and phenotype and provides a complementary approach to genome-wide association testing in admixed populations.[9,10] Local ancestry can act as a proxy for a genetic variant that is not well captured by the available SNP-array or sequencing data, such as a structural variant that is difficult to genotype accurately.

Once a variant is found to be associated with a phenotype, local ancestry can be used to investigate the ancestral origin of an allele. For example, an Amerindian-specific variant of *ACTN1* is associated with platelet count in US-based Hispanics,[11] and an Amerindian-specific variant of *BCL2L11* is associated with urine albumin-to-creatinine ratio.[12] In African Americans, two African-specific variants in *APOL1* are associated with kidney disease.[13] Identification of the ancestry of disease-associated variants is helpful for understanding and addressing disparities in disease rates.[14]

Local ancestry is also useful for population genetics analyses. Local ancestry segments are used to infer demographic history, including the timing of admixture,[15] the identity of source populations,[15,16] and the effective size of ancestral populations.[17,18] Local ancestry can be used for recombination rate inference[19,20] because changes in ancestry along an individual's genome represent crossovers that have occurred since admixture. Genomic regions with local ancestry proportions that deviate from the genome-wide average can signal post-admixture selection.[21,22]

Increasing amounts of high-coverage whole-genome sequence data are available from diverse and admixed populations.[23–25] This presents opportunities, because substantially increasing the number of reference individuals increases the accuracy of local ancestry inference, particularly in resolving ancestries that are less genetically diverged. However, larger reference panels also increase the computational burden.

Given these opportunities and challenges, we developed FLARE, which is based on the Li and Stephens model for haplotype frequencies[26] and follows in the footsteps of the HAPMIX and MOSAIC local ancestry inference methods.[7,27] The Li and Stephens model has been widely used for genotype phasing and imputation[28–32] because it provides high accuracy and it can be combined with powerful computational optimizations.[28–34] Its computation time is linear in the number of genetic markers, and after optimization its computation time is approximately linear in the number of individuals.[29,32] Extending this model to incorporate ancestry extends these advantages to local ancestry inference.

HAPMIX pioneered the application of the Li and Stephens model to local ancestry inference. However, we do not compare FLARE with HAPMIX because HAPMIX is limited to two ancestries. Instead, we compare FLARE with MOSAIC, which is a recent method based on the Li and Stephens model that allows for an arbitrary number

[1]Department of Biostatistics, University of Washington, Seattle, WA, USA; [2]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA
*Correspondence: sguy@uw.edu (S.R.B.), browning@uw.edu (B.L.B.)
https://doi.org/10.1016/j.ajhg.2022.12.010.

of ancestries and unknown relationships between reference panels and ancestry. We show that FLARE has better computational performance than MOSAIC and similar accuracy in our simulation scenarios.

Other frameworks for inferring local ancestry are possible. One of the most popular alternative methods for local ancestry inference is RFMix. Rather than utilizing a generative model of haplotype frequencies, RFMix is discriminative and employs a conditional random field.[35] We compare our method to RFMix and show that FLARE has better computational performance than RFMix and has superior accuracy in our simulation scenarios.

FLARE incorporates several computational techniques which allow it to scale to enormous datasets while maintaining high accuracy. FLARE performs on-the-fly compression of reference haplotypes and stores checkpoints when calculating probabilities to reduce memory requirements. FLARE constructs composite reference haplotypes to reduce computation time. These techniques are described in subjects and methods.

FLARE is designed to analyze both SNP-array and whole-genome sequence data. Most existing local ancestry inference methods were designed only for SNP-array data. For example, the MOSAIC method was tested using only SNP-array data, and we found that modifications to the program parameters were necessary when analyzing sequence data. FLARE can perform local ancestry inference on sequence data without the information loss that would result from substantial marker thinning.

Subjects and methods

Hidden markov model

FLARE uses a hidden Markov model (HMM). The input data are phased reference haplotypes and phased admixed haplotypes. We use the reference haplotypes to infer local ancestry in one admixed haplotype at a time. Each target admixed haplotype is modeled as an imperfect mosaic of reference haplotypes.[26,27] For a target haplotype, the unobserved state $S_m = (i, h)$ at marker m is comprised of the target haplotype's ancestry, i, at that position and the donor reference haplotype, h, whose allele is being copied at that position.

We assume that there are A ancestries contributing to the admixed genomes and that these ancestries are represented by the reference haplotypes. The reference data consist of J panels. In our analyses each reference panel is associated with a single ancestry and $A = J$. Our algorithm for estimating model parameters (see Method S3) assumes this one-to-one matching, but the remainder of the methodology does not require a one-to-one matching of reference panels and ancestries, and one could have $J < A$, $J = A$, or $J > A$. As an example of a case where $J < A$, one might have two (or more) reference panels representing a single ancestry, such as British and Italian reference panels for European ancestry; keeping these reference panels separate rather than combining them into one allows for different parameters for each of them. The case of $J < A$ could occur if there is no reference panel available for one of the ancestries; in this case it may be possible to choose parameters that would encourage the algorithm to infer that missing ancestry in regions where there is no good match to any of the reference panels, although accuracy may not be overly high in such a scenario.

The number of haplotypes in the j^{th} reference panel is denoted n_j. The total number of reference haplotypes is $N = \sum_j n_j$. We write p_{ij} for the probability that the donor haplotype is from reference panel j when the target haplotype is from ancestry i.

State transitions between two adjacent marker positions can occur due to crossover events. Crossover events that occur after admixture can change both the ancestry state, i, and the reference haplotype h. Crossover events that occur prior to admixture do not change the ancestry state but can change the reference haplotype h. We model this second class of crossover events using an ancestry-specific switch rate ρ_i. Ancestry-specific switch rates allow each ancestry to have a different effective population size and a different number of reference haplotypes.

The parameter $\boldsymbol{\mu}$ is a vector of length A giving the overall ancestry proportions of the admixed samples. The component μ_i is the prior probability that an arbitrary position in the genome is derived from ancestry i. Ancestry probabilities sum to one, i.e. $\sum_i \mu_i = 1$. It is assumed that all admixed samples included in the same analysis have similar ancestry proportions. If there are subgroups of admixed individuals with differing demographic histories, each subgroup can be analyzed separately.

Given a target haplotype, the prior probability for the state at any position is defined as follows. First the ancestry i is selected according to the probabilities μ_i. Then the reference panel j is chosen according to the probabilities p_{ij}. Finally, the donor haplotype is chosen randomly from the reference haplotypes for panel j. If h is from panel j, we write $q_{ih} = p_{ij}/n_j$ for the probability that the reference haplotype is h when the ancestry is i. Thus, the prior probability that the state is (i,h) is $\pi(i,h) = \mu_i q_{ih}$.

The parameter T is the number of generations since admixture. The distances between consecutive pairs of crossovers arising in the last T generations are exponentially distributed with mean $1/T$ Morgans ($100/T$ centiMorgans [cM]).

Any crossover in the past T generations may change the ancestry state. Consider two markers indexed by $m-1$ and m and separated by an interval of d_m Morgans. The probability of at least one such crossover occurring in this interval is $1 - e^{-d_m T}$. When a crossover occurs, a new ancestry is chosen according to the global ancestry probability vector $\boldsymbol{\mu}$. The probability of a transition from ancestry state i to ancestry state i' is thus $\mu_{i'}(1 - e^{-d_m T})$ for $i \neq i'$.

Changes in the donor reference haplotype h can occur regardless of whether there is a change in the ancestry. If there is no change in ancestry between the two markers, selection of a new donor reference haplotype occurs with probability $1 - e^{-d_m \rho_i}$, where ρ_i is a population-specific switch rate and i is the ancestry state at both markers. If there is a change to ancestry i' between the two positions, a new donor reference haplotype h is always selected. In either case, the donor reference haplotype in the new state (i', h') is selected according to the probabilities $q_{i'h'}$.

The resulting probability $P(S_m = (i', h') | S_{m-1} = (i, h))$ of transitioning from state (i, h) to state (i', h') is given in Method S1.

If the target haplotype's ancestry is i and the donor haplotype is from reference panel j, the emitted allele is the donor haplotype's allele with probability $1 - \theta_{ij}$, and is a different allele otherwise. The θ_{ij} are mismatch rates which model recent mutation, genotype error, and gene conversion.

Let $I_m(h) = 1$ if the allele at marker m on haplotype h matches the observed allele on the admixed haplotype, and let $I_m(h) =$

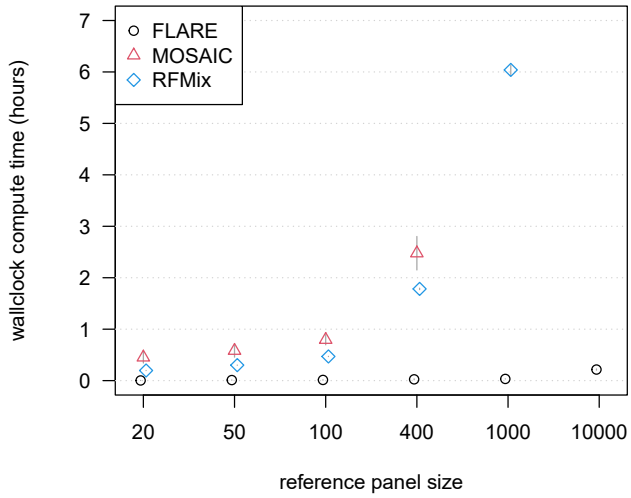

Figure 1. Computation time for simulated sequence data with three-way admixture

Wallclock computation time in hours is shown on the y axis. Reference panel size for each of the three ancestries is shown on the x axis. Each analysis includes 100 admixed individuals, and results are averaged over four replicate simulations. Error bars (± 2 standard errors) are shown as gray lines. Analyses of a simulated chromosome modeled on human chromosome 20 were run with 20 compute threads. MOSAIC could not analyze the data with 1,000 or more individuals per reference panel within the available 384 GB of computer memory. RFMix could not analyze the data with 10,000 individuals per reference panel within 48 h.

0 otherwise. The emission probability $e_m(i,h)$ (probability of the data given the HMM state) for the allele at marker m on the admixed haplotype when the state is (i, h) and the reference haplotype h is from reference panel j is

$$e_m(i,h) = \theta_{ij}^{1-I_m(h)}\left(1-\theta_{ij}\right)^{I_m(h)}.$$

Given the parameter values and genetic data, we calculate the posterior probability of ancestry at each marker using hidden Markov model methods, which are described in Method S2. The assigned ancestry is the ancestry with the highest posterior probability.

Estimating parameter values

A user can optionally specify parameter values. If not specified, values for the reference panel probabilities p_{ij} and the within-ancestry switch rate ρ_i are estimated from the reference panels and other parameters are assigned default values as described in Method S3. If the expectation maximization (EM) option is turned on (the default), the values of $\boldsymbol{\mu}$ and T will be updated based on several iterations of the estimation scheme described in Method S3. In the analyses presented in this paper, we use the default initial parameter values and perform EM updating.

If the user wishes to parallelize their analyses by chromosome, we recommend that the user run one autosomal chromosome first, and then use the output model file to specify the analysis parameters for other autosomes. This will reduce computing time and ensures consistency across chromosomes.

Computational techniques

Many computational techniques have been developed that substantially reduce the computation time and memory requirements for genotype imputation. We incorporate several of these techniques in our local ancestry inference method. These techniques include a compressed representation for reference haplotypes in memory,[31,36] the use of a small custom panel of composite reference haplotypes for each admixed haplotype,[30,32,33] and checkpointing of HMM backward probabilities.[37,38]

The accuracy of local ancestry inference increases significantly with reference panel size (Figure 1), but large reference panels also increase the computational burden. In genotype imputation the use of a small, custom subset of reference haplotypes for each individual can reduce computation time by one or more orders of magnitude with no loss in accuracy.[33] We have developed a fast method for generating a custom chromosome-length reference panel composed of composite reference haplotypes.[33] Each composite reference haplotype is a mosaic of reference haplotype segments that incorporates long identity-by-descent segments between the reference haplotypes and a target haplotype. We create a custom set of composite reference haplotypes for each target haplotype, and we record the source reference panel for each reference haplotype segment so that the appropriate transition and emission probabilities can be used.

We accommodate extremely large reference panels by compressing and storing the phased input data in bref3 format during program execution.[33] This format compresses data for rare variants by storing the indices of haplotypes that carry each rare variant, and it compresses data for other variants by storing the distinct allele sequences in a genomic interval together with an array that maps each haplotype index to its allele sequence.[33] The bref3 format enables an entire chromosome of reference and target haplotypes to be stored in memory and permits rapid lookup of haplotype alleles.

Checkpointing reduces the memory for HMM calculations for M markers from $O(M)$ to $O(\sqrt{M})$ by storing forward probabilities at $\sqrt{M}$ checkpoints and re-calculating backward probabilities from the nearest preceding checkpoint when required.[37,38] Since there can be more than a million markers on a chromosome, checkpointing can produce a 1,000-fold reduction in the memory required for HMM calculations, at the cost of a 2-fold increase in computation time.

Marker filtering

FLARE removes the lowest frequency variants before applying the methodology described below. FLARE applies a minimum minor allele frequency (MAF) threshold of 0.005 across the reference individuals. We found that this filter reduces run time without loss of accuracy, particularly for very large datasets. For sequence data, FLARE also imposes a minimum minor allele count threshold of 50, again calculated using the reference individuals. For analyses with small reference panels, this filter effectively reduces the density of the markers, resulting in better model fit and improved accuracy. The minor allele count threshold is not applied to SNP array data because such data already have reduced marker density relative to sequence data.

Simulated data

We simulated genetic data from human out-of-Africa demographic models for three-way and four-way admixture, using modified versions of the AmericanAdmixture_4B11 and OutOfAfrica_4J17 demographic models implemented in stdpopsim v.0.1.2.[39] We also simulated two-way admixture for populations with various levels of divergence.

The three-way model[18] extends a model of African, European, and Asian demographic history[40] to include admixture occurring 12 generations ago. The new admixed population has 1/6 African, 1/3 European, and 1/2 Asian ancestry, an initial size of 30,000, and a growth rate of 5% per generation. We added population growth in the 10 most recent generations to the unadmixed populations, at rates of 19.3% (African population), 10.8% (European population), and 7.8% (Asian population), so that each population grows to approximately 100,000 individuals in order to permit sampling of large reference panels from these populations. We sampled up to 50,000 individuals from each of the three reference ancestries and up to 10,000 admixed individuals (see below for details).

The four-way model extends the demographic history of African, European, Han Chinese, and Japanese populations inferred by Jouganous et al.[41] As above, we added an admixed population occurring 12 generations ago with 15% African, 15% European, 30% Chinese, and 40% Japanese ancestry, an initial size of 30,000, and a growth rate of 5% per generation. In the four-way admixture analyses, we sampled 400 individuals from each of the four reference ancestries and 100 admixed individuals.

The two-way admixture models have an ancestral population of size 10,000 that split a specified number of generations ago (100, 200, 400, 800, 1,600, or 3,200) into two isolated populations of 10,000 each. These split durations lead to F_{ST} values in the approximate range 0.005–0.15. An admixed population with 30:70 admixture was formed 12 generations ago. We sampled 200 individuals from each of the two reference ancestries and 100 admixed individuals.

We used SLiM[42] (v.3.7.1) for forward simulation of the most recent $10 \times N_e$ generations, followed by simulation of earlier generations with msprime[43] (v.1.1.1) to ensure full coalescence.[44] We used the HapMap II chromosome 20 recombination map[45] to simulate data with characteristics similar to human chromosome 20, and we used this map for analysis as well as for simulating the data. For the very largest dataset (three reference panels of 50,000 individuals each and 10,000 admixed individuals), we simulated only the first 40 cM (approx. 17 Mb) of chromosome 20, due to the size of the simulation. We added mutations at a rate of 1.44×10^{-8} per base pair per generation.[41] During forward simulation, gene conversion was added at a rate of twice the local recombination rate and with a mean tract length of 300 bp.

We constructed multiple datasets with a varying number of sampled individuals and different marker ascertainment schemes. The genetic data for each analysis were generated in three steps: (1) simulation of full demographic history and admixture, (2) site ascertainment, and (3) analysis-specific sampling and site filtering. In the first step, the genotype data were simulated as described above. In the second step, two distinct ascertainment schemes were applied to produce simulated sequence data and simulated array data. For the array data, we removed all sites with mean MAF less than 0.05 in the combined reference populations. For the sequence data, we removed all singletons. In the third step, the datasets were further reduced. After selecting individuals for a specific analysis, variants that were now singletons were removed. If array data had more than 20,000 sites, 20,000 randomly selected sites were retained. Next, we added genotype error to variable sites at a rate of $\varepsilon = 0.0002$, except at sites with MAF $< 2\varepsilon$ where the error rate was set to MAF/2 in order to ensure that the error-added data still contain very low frequency variants. Finally, all individuals (reference and admixed) were phased together with BEAGLE 5.4.

For each demographic history, we conducted four independent simulations and applied array and sequence ascertainment to each. This allowed four fully independent replicate analyses for each scenario, with no overlapping individuals or sites.

We inferred local ancestry with FLARE (v.0.3.0), RFMix (v.2.03-r0), and MOSAIC (v.1.3.9). All programs were supplied with the same phased genotype data, genetic map, and reference panels. Parameters affecting the statistical analyses of the programs were kept at default values, except as noted. FLARE was run with posterior probabilities turned on (probs = true) since these were used to assess accuracy and calibration, and analyses of the array data were performed with array = true. For RFMix, 5 EM iterations were requested (-e). For MOSAIC, the number of grid points per cM (-GpcM) was set to the product of 0.0012 and the number of sites in the analysis. We found that this setting greatly improved the accuracy of MOSAIC for sequence data. If a program didn't report local ancestry at a site, the local ancestry of the closest preceding site was used.

The accuracy of each method was assessed with Pearson's r^2 by comparing the inferred and true local ancestry dose. True local ancestry was defined to be the population of residence of the ancestral chromosome segment 20 generations prior to sampling (8 generations prior to admixture). For each ancestry, we summed the local ancestry posterior probabilities for the two haplotypes to obtain an estimated diploid ancestry dose at each site, and we counted the number of copies of the ancestry in the true local ancestry (0, 1, or 2) to obtain the true diploid ancestry dose. We calculated the squared Pearson correlation of the estimated and true diploid ancestry dose across all individuals and sites. Separate r^2 values were calculated for each ancestry and overall reported values are the unweighted mean r^2 across all ancestries.

Each method reports posterior probabilities, which may be more or less well calibrated. For example, ideally, 90% of sites assigned 90% posterior probability of having ancestry 1 should actually be ancestry 1 and 10% should be another ancestry. Since the simulated data are statistically phased before inferring local ancestry, we cannot check calibration at the haplotype level, but must instead work at the diplotype level. Ideally, the average true ancestry dose for sites with an estimated diploid ancestry dose of 1.8 should be 1.8. To assess the calibration, we divide the range of possible estimated diploid ancestry doses into bins and obtain the average true dose for each bin.

All analyses were run on a 24 core 2.2 GHz server with 384 GB memory and all local ancestry analyses were multi-threaded across 20 threads. We provide a repository containing all code for the generation and analysis of the simulated data presented here (see web resources).

1000 Genomes and Human Genome Diversity Project data

We downloaded high-coverage sequence data for chromosome 1 from the Human Genome Diversity Project (HGDP) and from the 1000 Genomes Project (see web resources).[24,46] We merged the two datasets and excluded variants that were not bi-allelic SNPs with <1% missingness and at least 5 copies of the minor allele in the combined data. After filtering, 2,021,066 SNPs remain on chromosome 1. We phased the data using Beagle 5.2 with the HapMap GRCh38 map (see web resources).[32]

We used the HGDP data for reference panels, assigning panels using the regional labels provided by the HGDP but omitting Oceania due to its smaller size and lack of relevance for the 1000 Genomes

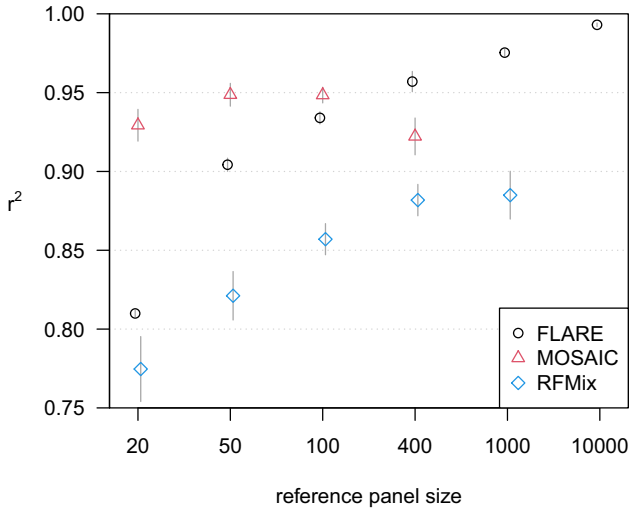

Figure 2. Accuracy when increasing reference panel size for simulated sequence data with three-way admixture

The y axis shows squared correlation between true and inferred local ancestry dose averaged over ancestries and across four replicate simulations (details in subjects and 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, and each analysis includes 100 admixed individuals. MOSAIC could not analyze the data with 1,000 or more individuals per reference panel within the available 384 GB of computer memory. RFMix could not analyze the data with 10,000 individuals per reference panel within the allotted two days.

data. The panels range in size from 61 (America) to 223 (East Asia). We used FLARE with default settings to infer local ancestry in the 26 populations of the 1000 Genomes project, using a separate analysis for each of these populations. Ancestry proportions were obtained by averaging ancestry calls across sites and individuals.

Results

Simulated data

FLARE has much faster computation times than MOSAIC or RFMix (Figure 1). Compute times for FLARE are relatively insensitive to the reference panel size due to its use of composite reference haplotypes (see subjects and methods). FLARE's analyses of 100 admixed individuals in the three-way admixture setting with sequence data take less than 2 min on average for reference panel sizes of up to 1,000 per ancestry and an average of 13 min for 10,000 individuals in each of the three reference panels. FLARE's analysis of 40 cM of data from 10,000 admixed individuals with 50,000 individuals in each of the three reference panels took an average of 35 min.

For the four-way admixture with sequence data, 100 admixed individuals, and 400 individuals in each of the four reference panels, FLARE took an average of 4 min, while RFMix took an average of 273 min. Compute times are expected to scale approximately linearly in the number of admixed individuals for each of the three methods.

Figure 2 shows accuracy results for the three-ancestry simulation with sequence data. With large reference panel sizes, FLARE is the most accurate method ($r^2 = 0.993$ with 10,000 individuals per reference panel). For small reference panel sizes (up to 100 individuals per ancestry), MOSAIC is the most accurate method. For FLARE and RFMix, we see an increase in accuracy with increasing reference panel size.

With simulated array data instead of sequence data, r^2 accuracy generally decreases slightly (Figure S1). With the array data as for the sequence data, FLARE has higher r^2 accuracy than the other two methods for larger reference panel sizes, while MOSAIC has the highest accuracy among the three methods for the smallest reference panel sizes.

Simulation studies typically employ reference panels of equal size for each ancestry, whereas in real analyses some ancestries typically have fewer reference individuals. We thus investigated the accuracy when reference panels have unequal sizes. We found that all three methods performed well, with MOSAIC and FLARE having the highest accuracy performance (Figure S2).

We used the four-ancestry model to assess the ability of the methods to infer local ancestry in situations with less genetic divergence. For the sequence data, we find that FLARE has good resolution to distinguish all four ancestries ($r^2 = 0.888$), whereas RFMix's accuracy is severely reduced ($r^2 = 0.666$), with most of this reduction being driven by lower r^2 for the two Asian ancestries (Figure 3). MOSAIC was excluded from this comparison because it could not analyze the simulated four-ancestry sequence data within the available 384 GB of computer memory. For the array data, FLARE still performs the best ($r^2 = 0.866$), with MOSAIC ($r^2 = 0.820$) and RFMix ($r^2 = 0.779$) being slightly less accurate (Figure S3).

The two-population models include a range of divergence between ancestral populations, with split times ranging from 100 generations ago to 3,200 generations ago. For comparison with human populations, the split time between African and out-of-Africa populations was approximately 4,000 generations ago, while the split time between the Japanese from Tokyo and Han Chinese from Beijing populations was approximately 300 generations ago.[41] All methods have high accuracy for the most diverged ancestral populations (Figure 4). At lower levels of divergence (split times of 200–800 generations ago), FLARE and MOSAIC are most accurate. All methods have difficulty distinguishing the most closely related populations (split time of 100 generations ago).

We found that MOSAIC is well calibrated in terms of its reported posterior ancestry probabilities (Figure S4). In contrast, RFMix's output probabilities are not well calibrated. FLARE's calibration performance is intermediate and improves with increasing reference panel sizes.

1000 Genomes local ancestry analysis

We inferred local ancestry for each of the 26 populations of the 1000 Genomes Project using six regional reference panels from the HGDP. Estimated ancestry proportions

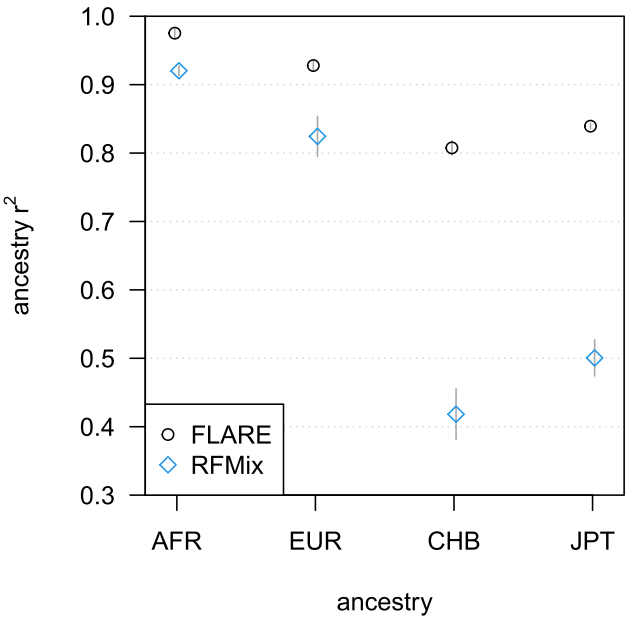

Figure 3. Accuracy by ancestry for simulated sequence data with four-way admixture

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 sequence data have 100 admixed individuals and 400 individuals in each of the four reference panels. Results are averaged over four replicate simulations. MOSAIC could not analyze these data within the available 384 GB of computer memory

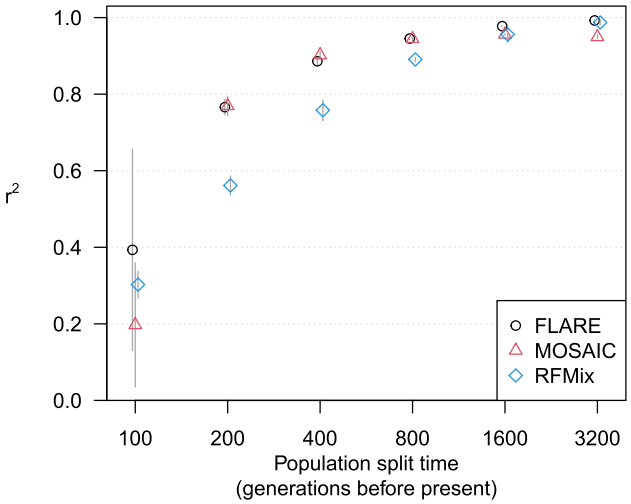

Figure 4. Accuracy for simulated sequence data with two-way admixture

The y axis shows squared correlation between true and inferred local ancestry dose averaged over ancestries and across four replicate simulations (details in subjects and methods). Error bars (±2 standard errors) are shown as gray lines. The x axis shows the split time for the two ancestral populations. Each of the three ancestries is represented by a reference panel of size 200, and each analysis includes 100 admixed individuals.

from this analysis are shown in Table 1. Results generally match expectation. For example, the unadmixed African populations are inferred to have 98%–100% African ancestry. Native American ancestry originally derives from Siberia,[47] which may partially explain the inferred East Asian ancestry in the admixed American populations, although post-colonial migration from Asia may also play a role.[48] Finns (FIN) are inferred to have 2% East Asian ancestry, which is concordant with previous studies that have found evidence of an Asian contribution to the gene pool in Finns.[49,50] Spaniards (IBS) are inferred to have 2% African ancestry, which matches previous reports of gene flow into southern Europe from North Africa.[51]

Our initial analyses of chromosome 1 with parameter estimation took 16.6 h (38 min per population on average). Analyses of other chromosomes that use the parameters estimated from the chromosome 1 data would use much less time. For example, when we repeated the chromosome 1 analysis using the parameters estimated in the first analysis, the second analysis took only 2.4 h (6 min per population on average). The first and second analysis produced identical estimated ancestry probabilities, which is expected because the only randomization that occurs within FLARE is in the parameter estimation. Parameter estimation is needed only for one autosomal chromosome, with analyses of the other autosomal chromosomes using the same parameters.

Discussion

We have presented FLARE, a method for local ancestry inference. We showed that FLARE has superior accuracy, computing efficiency, and scalability compared to RFMix and MOSAIC. FLARE was able to analyze a 40 cM simulated chromosome with 10,000 admixed individuals and three 50,000-member reference panels in 35 min on a computer with 24 CPU cores. FLARE's analysis of simulated chromosome 20 data with 100 admixed individuals and three 10,000-member reference panels took 13 min, while RFMix was unable to complete analysis of these data within two days on the same computer. MOSAIC was even more limited in the data that it could analyze due to memory constraints and was significantly slower than RFMix. Overall, FLARE and MOSAIC had similar accuracy, while RFMix's accuracy was lower. RFMix utilizes called ancestry from the admixed individuals to augment the reference panels, so it is likely that RFMix's accuracy would increase with larger numbers of admixed individuals.

A notable feature of the results of the simulation studies is that FLARE can better distinguish ancestries with lower levels of genetic divergence, such as distinguishing between Japanese and Chinese ancestry. In contrast, RFMix had difficulty distinguishing these ancestries. This suggests the potential to accurately infer local ancestry in

Table 1. Inferred ancestry proportions in 1000 Genomes Project chromosome 1 data for six ancestries using HGDP reference panels

Region	Population	African	East Asian	European	Central/South Asian	American	Middle Eastern
Africa	ACB[a]	0.89*	0.00	0.10	0.01	0.00	0.00
	ASW[a]	0.76*	0.01	0.20*	0.00	0.03	0.00
	ESN	1.00*	0.00	0.00	0.00	0.00	0.00
	GWD	0.99*	0.00	0.01	0.00	0.00	0.00
	LWK	0.98*	0.00	0.00	0.00	0.00	0.02
	MSL	1.00*	0.00	0.00	0.00	0.00	0.00
	YRI	1.00*	0.00	0.00	0.00	0.00	0.00
America	CLM[a]	0.09	0.01	0.63*	0.00	0.26*	0.01
	MXL[a]	0.06	0.03	0.47*	0.01	0.44*	0.00
	PEL[a]	0.04	0.04	0.21*	0.01	0.69*	0.00
	PUR[a]	0.16*	0.01	0.69*	0.00	0.14*	0.01
East Asia	CDX	0.00	1.00*	0.00	0.00	0.00	0.00
	CHB	0.00	1.00*	0.00	0.00	0.00	0.00
	CHS	0.00	1.00*	0.00	0.00	0.00	0.00
	JPT	0.00	1.00*	0.00	0.00	0.00	0.00
	KHV	0.00	1.00*	0.00	0.00	0.00	0.00
Europe	CEU	0.00	0.00	1.00*	0.00	0.00	0.00
	FIN	0.00	0.02	0.98*	0.00	0.00	0.00
	GBR	0.00	0.00	1.00*	0.00	0.00	0.00
	IBS	0.02	0.00	0.98*	0.00	0.00	0.00
	TSI	0.00	0.00	1.00*	0.00	0.00	0.00
South Asia	BEB	0.00	0.00	0.00	1.00*	0.00	0.00
	GIH	0.00	0.00	0.00	1.00*	0.00	0.00
	ITU	0.00	0.00	0.00	1.00*	0.00	0.00
	PJL	0.00	0.00	0.00	1.00*	0.00	0.00
	STU	0.00	0.00	0.00	1.00*	0.00	0.00

Ancestry proportions >10% are indicated with an asterisk (*). Descriptions of the populations can be found in Supplementary Information Table 1 of the 1000 Genomes Project's phase 3 paper.[52]
[a]Recently admixed populations from the Americas.

admixtures that are subtler than the continental-level admixtures that have previously been the focus of attention.

Our model and algorithm borrow heavily from HAPMIX.[27] However, there are several significant differences between FLARE and HAPMIX. The two most obvious differences are that FLARE handles multiple ancestries, while HAPMIX is restricted to two, and that FLARE implements a range of computation techniques that are not included in HAPMIX and that allow FLARE to scale to large-scale whole-genome sequence data. In addition, FLARE has a procedure for estimating the matrix of copying probabilities (the p_{ij}) and the switch rates (the ρ_i) directly from the reference panel data, while HAPMIX provides an option to estimate all parameters iteratively with an EM approach. Also, HAPMIX allows only for bi-allelic SNPs while FLARE can analyze multi-allelic markers.

FLARE is a user-friendly java program with a command-line interface similar to Beagle.[32] When there is a one-to-one matching of ancestries and reference panels, the only input data required by FLARE are phased reference and target VCF files,[53] a genetic map file, and a file specifying the reference panel assignment for each reference individual. FLARE outputs a VCF file containing the input genotypes and phased local ancestry calls. As an option, the posterior local ancestry probabilities can also be included in the output VCF file. FLARE also outputs a model file giving the estimated parameters. The model can be used in future analyses of the same study to reduce computing time and ensure consistency between analyses.

FLARE's user-friendly and robust software implementation, its computational speed and ability to scale to extremely large datasets, and its high accuracy make it a

useful tool for local ancestry inference in the increasingly large and diverse genetic data that are being generated.

Data and code availability

The FLARE software is available from https://github.com/browning-lab/flare. The simulation and analysis pipeline used in this study is available from https://github.com/rwaples/lai-sim.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.12.010.

Acknowledgments

The authors thank Dr. Michael Salter-Townshend for help with MOSAIC. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award numbers HG010869 and HG008359. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of interests

The authors declare no competing interests.

Received: August 3, 2022
Accepted: December 13, 2022
Published: January 6, 2023

Web resources

1000 Genomes Project high-coverage sequence data, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/

Beagle, http://faculty.washington.edu/browning/beagle/beagle.html.

HapMap GRCh38 map, http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/plink.GRCh38.map.zip

Human Genome Diversity Project high-coverage sequence data, ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/

msprime, https://github.com/tskit-dev/msprime.

MOSAIC, https://maths.ucd.ie/~mst/MOSAIC/

Stdpopsim, https://github.com/popsim-consortium/stdpopsim

RFMIX, https://github.com/slowkoni/rfmix.

References

1. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science *343*, 747–751.
2. 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.
3. Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J.K., Muzzio, M., Rodriguez-Flores, J.L., Kenny, E.E., Gignoux, C.R., Maples, B.K., Guiblet, W., et al. (2013). Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet. *9*, e1004023.
4. Homburger, J.R., Moreno-Estrada, A., Gignoux, C.R., Nelson, D., Sanchez, E., Ortiz-Tello, P., Pons-Estel, B.A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C.D., et al. (2015). Genomic insights into the ancestry and demographic history of South America. PLoS Genet. *11*, e1005602.
5. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.
6. Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354–357.
7. Salter-Townshend, M., and Myers, S. (2019). Fine-scale inference of ancestry segments without prior knowledge of admixing groups. Genetics *212*, 869–889.
8. Durand, E.Y., Do, C.B., Wilton, P.R., Mountain, J.L., Auton, A., Poznik, G.D., and Macpherson, J.M. (2021). A scalable pipeline for local ancestry inference using tens of thousands of reference haplotypes. Preprint at bioRxiv.
9. Shriner, D. (2017). Overview of admixture mapping. Curr. Protoc. Hum. Genet. *94*, 1–23.
10. Winkler, C.A., Nelson, G.W., and Smith, M.W. (2010). Admixture mapping comes of age. Annu. Rev. Genomics Hum. Genet. *11*, 65–89.
11. Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T., Conomos, M.P., Schurmann, C., McHugh, C.P., et al. (2016). Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. Am. J. Hum. Genet. *98*, 229–242.
12. Brown, L.A., Sofer, T., Stilp, A.M., Baier, L.J., Kramer, H.J., Masindova, I., Levy, D., Hanson, R.L., Moncrieft, A.E., Redline, S., et al. (2017). Admixture mapping identifies an amerindian ancestry locus associated with albuminuria in Hispanics in the United States. J. Am. Soc. Nephrol. *28*, 2211–2220.
13. Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Uscinski Knob, A.L., et al. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African-Americans. Science *329*, 841–845.
14. 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.
15. 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.
16. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013). Reconstructing the population genetic history of the Caribbean. PLoS Genet. *9*, e1003925.

17. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. Nature *488*, 370–374.
18. Browning, S.R., Browning, B.L., Daviglus, M.L., Durazo-Arvizu, R.A., Schneiderman, N., Kaplan, R.C., and Laurie, C.C. (2018). Ancestry-specific recent effective population size in the Americas. PLoS Genet. *14*, e1007385.
19. Wegmann, D., Kessner, D.E., Veeramah, K.R., Mathias, R.A., Nicolae, D.L., Yanek, L.R., Sun, Y.V., Torgerson, D.G., Rafaels, N., Mosley, T., et al. (2011). Recombination rates in admixed individuals identified by ancestry-based inference. Nat. Genet. *43*, 847–853.
20. Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature *476*, 170–175.
21. Cuadros-Espinoza, S., Laval, G., Quintana-Murci, L., and Patin, E. (2022). The genomic signatures of natural selection in admixed human populations. Am. J. Hum. Genet. *109*, 710–726.
22. Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E.G., and Risch, N.J. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. Am. J. Hum. Genet. *81*, 626–633.
23. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19.
24. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012.
25. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53, 831 diverse genomes from the NHLBI TOPMed program. Nature *590*, 290–299.
26. Li, N., and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics *165*, 2213–2233.
27. 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.
28. Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.
29. Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat. Commun. *10*, 5436–5510.
30. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.
31. Browning, B.L., and Browning, S.R. (2016). Genotype imputation with millions of reference samples. Am. J. Hum. Genet. *98*, 116–126.
32. Browning, B.L., Tian, X., Zhou, Y., and Browning, S.R. (2021). Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. *108*, 1880–1890.
33. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A one-penny imputed genome from next-generation reference panels. Am. J. Hum. Genet. *103*, 338–348.
34. Das, S., Abecasis, G.R., and Browning, B.L. (2018). Genotype imputation from large reference panels. Annu. Rev. Genomics Hum. Genet. *19*, 73–96.
35. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
36. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.
37. Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223.
38. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.
39. Adrion, J.R., Cole, C.B., Dukler, N., Galloway, J.G., Gladstein, A.L., Gower, G., Kyriazis, C.C., Ragsdale, A.P., Tsambos, G., Baumdicker, F., et al. (2020). A community-maintained standard library of population genetic models. Elife *9*, e54967.
40. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., 1000 Genomes Project, and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA. *108*, 11983–11988.
41. Jouganous, J., Long, W., Ragsdale, A.P., and Gravel, S. (2017). Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. Genetics *206*, 1549–1567.
42. Haller, B.C., and Messer, P.W. (2019). SLiM 3: forward genetic simulations beyond the wright–fisher model. Mol. Biol. Evol. *36*, 632–637.
43. Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A.P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E.C., Galloway, J.G., et al. (2022). Efficient ancestry and mutation simulation with Msprime 1.0. Genetics *220*.
44. Haller, B.C., Galloway, J., Kelleher, J., Messer, P.W., and Ralph, P.L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Mol. Ecol. Resour. *19*, 552–566.
45. International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
46. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Preprint at bioRxiv.
47. Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C.,

Ávila-Arcos, M.C., Malaspinas, A.S., et al. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. Science *349*, aab3884.

48. Pinto, J.A., Mas, L.A., and Gomez, H.L. (2017). High epidermal growth factor receptor mutation rates in Peruvian patients with non–small-cell lung cancer: is it a matter of Asian ancestry? J. Glob. Oncol. *3*, 429–430.
49. Kittles, R.A., Perola, M., Peltonen, L., Bergen, A.W., Aragon, R.A., Virkkunen, M., Linnoila, M., Goldman, D., and Long, J.C. (1998). Dual origins of Finns revealed by Y chromosome haplotype variation. Am. J. Hum. Genet. *62*, 1171–1179.
50. Ingman, M., and Gyllensten, U. (2007). A recent genetic link between Sami and the Volga-Ural region of Russia. Eur. J. Hum. Genet. *15*, 115–120.
51. Botigué, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., et al. (2013). Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc. Natl. Acad. Sci. USA. *110*, 11791–11796.
52. 1000 Genomes Project Consortium, Auton, A., 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.
53. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

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

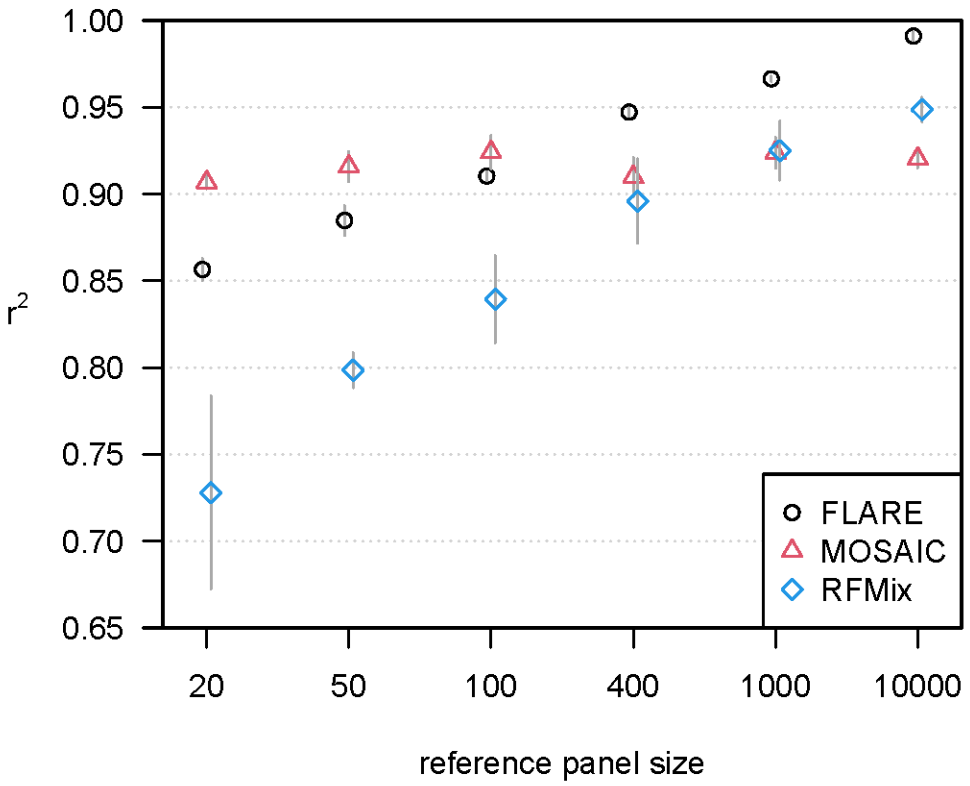

Figure S1: Accuracy when increasing reference panel size for simulated array data with three-way admixture. The y-axis shows squared correlation between true and inferred local ancestry dose averaged over ancestries and across four replicate simulations (details in Methods). Error bars (+/- 2 standard 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.

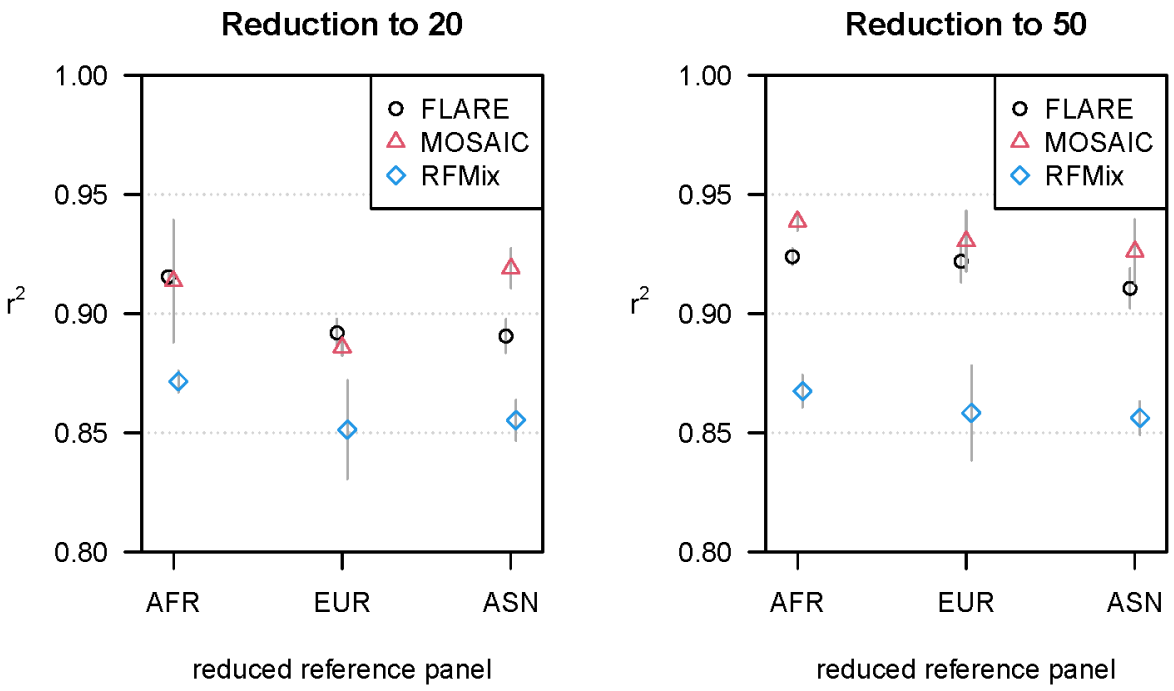

Figure S2: Accuracy when reducing the size of one of three reference panels. The data are simulated sequence data for three-way admixture. Each reference panel has size 400, except for the reference panel that is denoted on the x-axis (AFR is West African, EUR is European, ASN is East Asian) which has 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.

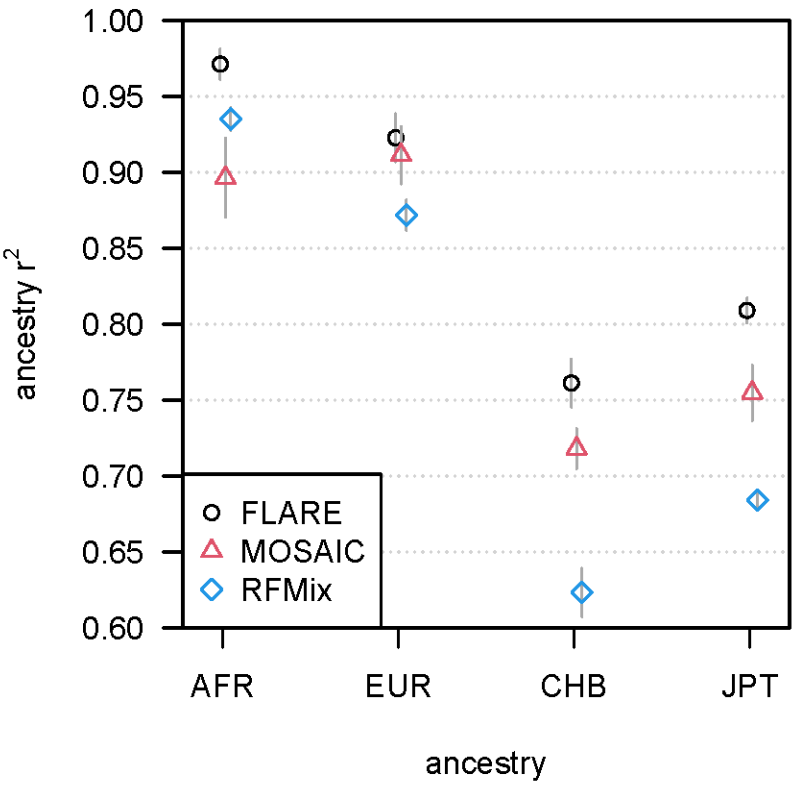

Figure S3: Accuracy by ancestry for the simulated array data with four-way admixture. The y-axis is the squared correlation between the true and inferred ancestry dose for a single ancestry averaged across four replicate simulations. Error bars (+/- 2 standard errors) are shown as gray lines. The ancestry 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.

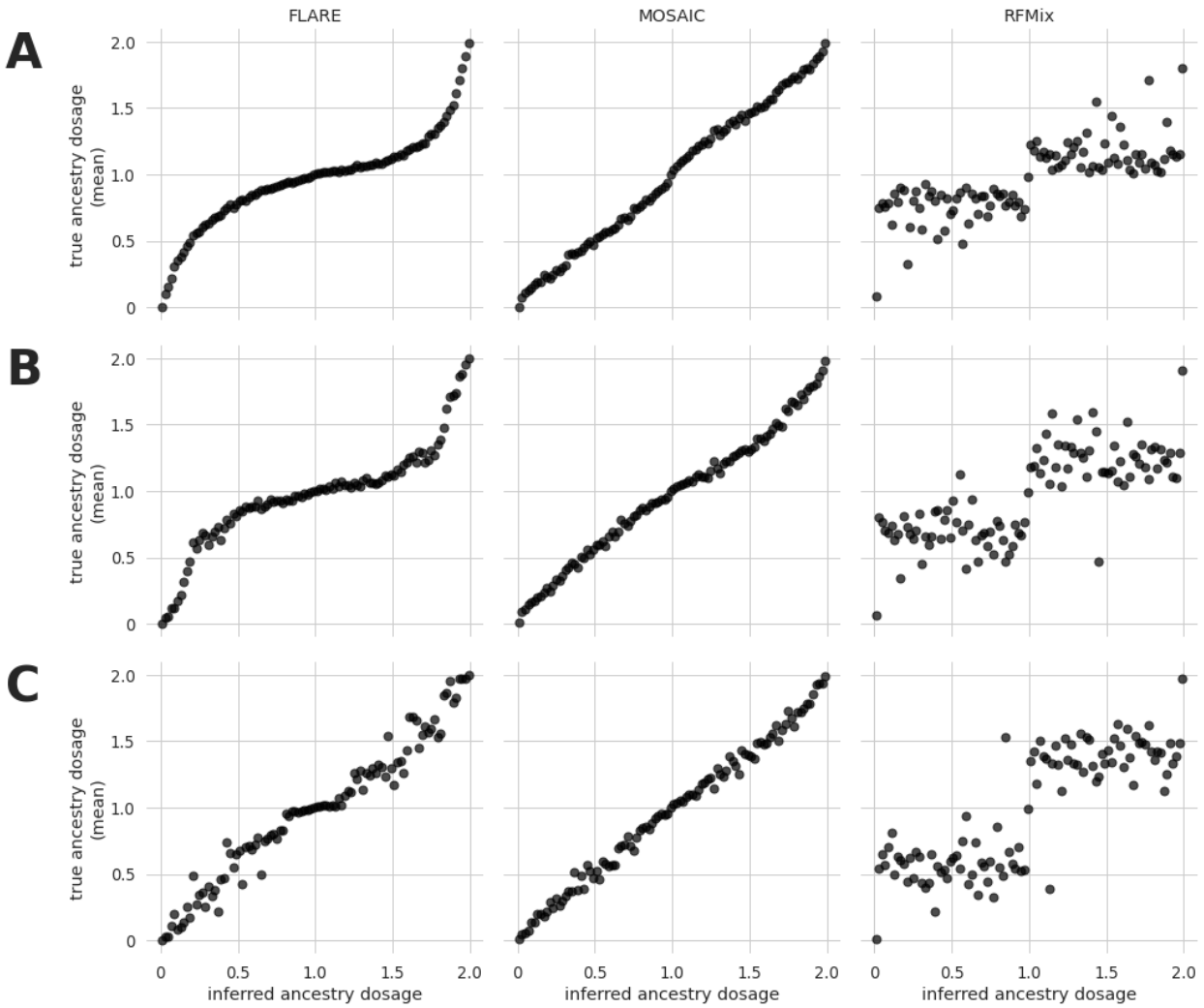

Figure S4: Calibration of estimated diploid ancestry dose on simulated three-way admixture data. Estimated diploid ancestry dose is binned into bins of width 0.02 along the x-axis. The y-axis is the average true diploid ancestry dose for each bin. Results for FLARE, MOSAIC, and RFMix are shown in the 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 $1, \dots, M$. The forward probabilities are

$$\alpha_m(i,h) = P(Y_1, \dots, Y_m, S_m = (i,h)) \tag{S1}$$

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

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

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)

$$\alpha_m(i',h') = e_m(i',h')\sum_{i,h} P\big(S_m=(i',h')\big|S_{m-1}=(i,h)\big)\alpha_{m-1}(i,h)$$
$$= e_m(i',h')\big[\big(1-e^{-d_m T}\big)\mu_{i'}q_{i'h'}s_f + e^{-d_m T}\big(1-e^{-d_m\rho_{i'}}\big)q_{i'h'}f_{i'}$$
$$+ e^{-d_m T}e^{-d_m\rho_{i'}}\alpha_{m-1}(i',h')\big].$$

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)

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

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 μ_i from the value in the preceding iteration, excluding those ancestries for which $\mu_i < 0.001$. A 5% relative change in a μ_i taking value of 0.1 in the previous iteration would be 0.005.

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

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 $\boldsymbol{p_{ij}}$ and switch rates $\boldsymbol{\rho_i}$:

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 θ_{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

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

The updated switch rate ρ_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:

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

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 ρ_i. After we have estimated the $\hat{\tau}_{m,i,k}$ for each marker m and each target haplotype k, we estimate ρ_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 ρ_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 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 γ_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 $\hat{\gamma}_{m,k} \approx T d_m$ 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.