Fast and accurate local ancestry inference with Recomb-Mix

Yuan Wei[1], **Degui Zhi**[2], **and Shaojie Zhang**[1]

[1]Department of Computer Science, University of Central Florida, Orlando, FL, USA

[2]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

Corresponding authors:

shzhang@cs.ucf.edu

degui.zhi@uth.tmc.edu

Running title:

Efficient local ancestry inference

Abstract

The availability of large genotyped cohorts brings new opportunities for revealing high-resolution genetic structure of admixed populations, via local ancestry inference (LAI), the process of identifying the ancestry of each segment of an individual haplotype. Though current methods achieve high accuracy in standard cases, LAI is still challenging when reference populations are more similar (e.g., intra-continental), when the number of reference populations is too numerous, or when the admixture events are deep in time, all of which are increasingly unavoidable in large biobanks. Here, we present a new LAI method, Recomb-Mix. Adopting the commonly used site-based formulation based on the classic Li and Stephens' model, Recomb-Mix integrates the elements of existing methods and introduces a new graph collapsing to simplify counting paths with the same ancestry label readout. Through comprehensive benchmarking on various simulated datasets, we show that Recomb-Mix is more accurate than existing methods in diverse sets of scenarios while being competitive in terms of resource efficiency. We expect that Recomb-Mix will be a useful method for advancing genetics studies of admixed populations.

Introduction

Local ancestry inference (LAI) is a process of assigning the ancestral population labels of each segment on an individual's genome sequence. LAI is not only useful for better study of human demographic history (Martin et al. 2017) but also can enable several downstream tasks, including admixture mapping (Reich et al. 2005), ancestry-aware genome-wide association studies (GWAS) (Pasaniuc et al. 2011), and ancestry-specific polygenic risk scores (Duncan et al. 2019). Recent studies show that local ancestry information improves the resolution of association signals in GWAS (Atkinson et al. 2021), helping to infer the high-resolution of genomic regions containing genes as under selection (Hamid et al. 2023). Local ancestry calls contribute to understanding the impact of genetic variants that cause disease (Hou et al. 2023) and the accuracy of polygenic scores of genetically based predictions (Ding et al. 2023).

Additionally, the inferred result from LAI methods can be used as an intermediate result for other uses such as producing global ancestry inference (GAI) results. LAI results can be easily converted to GAI results by summing up the local ancestry calls per population. RFMix (Maples et al. 2013) was used for the identification of sub-populations highly impacted by the COVID-19 pandemic (Parikh et al. 2022). For each sample, GAI was derived from the called LAI assignments on the chromosome which were summed up as an overall ancestry proportions and later used for association analyses.

The recent availability of biobank-scale genotyped datasets (Bycroft et al. 2018; Kurki et al. 2023) and the rising of enormous databases from direct-to-consumer genetic companies (Durand et al. 2021) create new challenges and opportunities for LAI. Participants in biobanks may be from highly imbalanced source populations. Inferring ancestral components underrepresented in these reference populations may be needed. On the other hand, the admixture to be inferred in the samples may be multi-ways and may be from both recent and distant admixture events. However, opportunities coexist with such challenges. The number of participants in biobanks is much larger and more representative than in previous reference panels. Possible sub-continental ancestral information becomes available from biobanks. With the availability of diverse samples in biobanks and the need for in-depth knowledge of admixed individuals, current LAI methods are unable to fully capitalize on these opportunities.

Existing LAI methods fall into two categories: site-based and window-based. Originally, Hap-

Mix (Price et al. 2009), based on extended Li and Stephens' Hidden Markov model (HMM) (Li and Stephens 2003), is developed to model different transition probabilities for within population and between population jumps. To make it tolerate mismatches, emission probabilities are also introduced and the model is not very efficient and cannot be scaled up (Geza et al. 2018). Later, window-based methods are gaining popularity (e.g., RFMix (Maples et al. 2013), G-Nomix (Hilmarsson et al. 2021), SALAI-Net (Oriol Sabat et al. 2022)). These methods take short stretches of sites as windows and define window-based features. It first makes local ancestry prediction over each window and then uses some post-processing to smooth out the labels across all windows. For each window, typically some machine learning method is used. However, the predefined window boundaries are not necessarily optimized and the noisy initial window labels can be difficult to correct by post-processing. Loter (Dias-Alves et al. 2018) is a recent site-based method. It formulates the LAI problem as a combinatorial best-path problem in a graph, which can be solved efficiently by dynamic programming. However, the problem formulation was simplistic and did not take into account the useful information as encoded in Li and Stephens' model, such as differential transition probabilities for within and between populations, variable recombination rates across sites. Therefore, there is room for improvement over the Loter approach by introducing a Li and Stephens-inspired parametrization of its scoring function.

To tackle the challenges of the availability of biobank-scale datasets and the deficiency of current LAI methods, we developed Recomb-Mix, a graph optimization approach based on Li and Stephens' model. Apart from current trends of window-based approaches, Recomb-Mix is a come-back of site-based methods. We used Li and Stephens-inspired scoring functions to define more relevant problem formation and used simplifications to Li and Stephens' model to make the graph collapsible and thus efficient. The reference panel was converted as a compact one, which contains only population features extracted from the original one. Its size is linear to the number of sites in the reference panel and it does not employ allele frequency.

Our approach was different from HapMix (Price et al. 2009) or Loter (Dias-Alves et al. 2018), which were methods also based on Li and Stephens' model. Table 1 is the summary of their features. Loter and Recomb-Mix were able to handle multi-way admixture inference, as HapMix was only able to tackle two-way admixture. Both HapMix and Recomb-Mix used genetic map to help out the inference while Loter did not take any biological information as input. From the HMM algorithm perspective, HapMix

Feature	HapMix	Loter	Recomb-Mix
Multi-way admixture	no	yes	yes
Use genetic map	yes	no	yes
HMM algorithms	forward-backward and Viterbi	Viterbi-like	forward and Viterbi-like

Table 1: Features of HapMix, Loter, and Recomb-Mix.

used the forward-backward algorithm to update transition and emission probabilities and then used the Viterbi algorithm to estimate the hidden ancestral states (Price et al. 2009; Wu et al. 2021). Loter's approach was to minimize an objective function using dynamic programming, which is a Viterbi-like algorithm (Dias-Alves et al. 2018; Oriol Sabat et al. 2022). Recomb-Mix took advantage of the graph optimization formulation as Loter did, but on a compact graph with population-level information only. All possible paths in a compact graph can be viewed as a set of "combined" paths from the original graph that emit the same population label readout. Thus, it had a flavor of a forward-like algorithm as those "combined" paths were calculated all together.

We propose a fast and efficient LAI method, Recomb-Mix. Our main contributions are the following: our method harnesses penalty insensitivity within populations, and our problem formulation enables the reference panel collapsible. Both improve the precision and speed of the inference. Recomb-Mix's superior accuracy is demonstrated against several leading LAI methods in various datasets from our simulation scenarios, including multi-way admixture and at the intra-continental level. We illustrate that Recomb-Mix is able to achieve high accuracy on individuals from the admixture of imbalanced founder populations, as well as using an imbalanced reference panel for the inference. We also show Recomb-Mix's capability of transforming a reference panel with a large number of individuals into a succinct-size population-oriented reusable one.

Methods

We present a site-based local ancestry inference method Recomb-Mix from a graph optimization perspective. We retain population-level information only from the reference data and build a compact

population graph from it to make inference results accurately and process efficiently. We use an objective function containing penalties of mismatches and across-population path changes to calculate the cost of each path in the graph. The optimal path reveals the estimations of ancestral labels across the genome.

Our approach was inspired by HapMix (Price et al. 2009) and Loter (Dias-Alves et al. 2018). HapMix extended the basic Li and Stephens' HMM (Li and Stephens 2003) to have population labels as augmented HMM state labels and introduce two transition probability parameters: small-scale (between haplotypes from within a reference population) and large-scale (between the reference populations). Loter formulated LAI as a graph optimization problem that finds a best-scoring path over a site-level graph. As stated in the Loter paper, Loter can be viewed as a Li and Stephens' "copying model" with simplified non-probabilistic parameterization (Dias-Alves et al. 2018). However, it applied the same penalty to haplotype template switches in both cases within or across reference populations. Through the unified view of the Li and Stephens' model and the graph optimization formulation (Table 2), Recomb-Mix introduces special parameterization to the Li and Stephens' model to induce graph simplification and more biologically-relevant scoring function. First, by assuming no path change penalty when switching haplotype templates within a reference population, Recomb-Mix enables the collapsing of the reference panel to a compact graph. Generating a compact graph greatly reduces the size of reference populations and retains the ancestry information as most non-ancestry informative markers are collapsed in the compact graph. Different path change penalties were used when switching haplotype templates within a reference population and between the reference populations: the path change penalty within a reference population is set to zero, and the path change penalty between the reference populations is parameterized by recombination rates from a genetic map. The scoring function is a simplified version of HapMix's, but a richer version than that of Loter's: $\sum(d + w \cdot r)$, where d is length of mismatched site between query haplotype and ones in reference panel, w is a tunable weight for relative importance of within and between population haplotype template switches, and r is normalized recombination rate value translated from genetic distance. Table 2 shows the differences between HapMix, Loter, and Recomb-Mix's scoring functions. HapMix incorporated recombination, miscopying, mutation, and genetic distance parameters into their model. With such a number of required biological parameters, lacking accurate population information may lead to biased inference result (Suarez-Pajes

Scoring Function	HapMix	Loter	Recomb-Mix
Within population transition probability / path change penalty	Recombination parameter	1	0
Across population transition probability / path change penalty	Recombination parameter and miscopying parameter	1	r, genetic distance
Emission probability / mismatch penalty	Mutation parameter	Length of mismatched site	d, length of mismatched site
Parameterization of transition probability	$1 - e^p$, where p is the product of genetic distance and recombination parameter	Bootstrap aggregation (bagging)	w, weight

Table 2: Analogy of HapMix, Loter, and Recomb-Mix.

et al. 2021). In contrast to HapMix's complex model, Loter had a simple approach that did not adopt any biological information. Recomb-Mix took Loter's simplicity but added a biological flavor which is genetic distance, as well as utilizing the collapsed graph, to achieve high computability and accuracy at the same time.

Figure 1 is an example showing Recomb-Mix's model, flow chart, and how it makes local ancestry inference. G is an original graph, similar to Loter, built from a reference panel having two populations (red or P_1 is ancestral population 1, blue or P_2 is ancestral population 2), I is a query of an admixed individual haplotype, G' is the graph collapsed from the original graph G, I' is the inferred query of an admixed individual haplotype. The color indicates the inferred population label of the site (i.e., red indicates P_1 and blue indicates P_2). The compact graph G' is constructed from the reference panel $\{P_1, P_2\}$, whose size is independent from $\{P_1, P_2\}$ and I. The top right dotted line box illustrates a zoomed-in version of the construction of nodes of a site in G' from nodes of a site in G. For each site j, population emission nodes $p_{i,j}$ are expanded to genotype emission nodes $g_{i,j}$. They are linked to the nodes $s_{i,j}$ from the original graph G, and then collapsed to nodes $s'_{i,j}$ as ones in the compact graph G'. Then, they are linked to population emission nodes $p_{i,j}$, to make a cross-population connection

for the next site $j+1$. This process has the similarity to a forward-like approach as each node in G' can be viewed as a bundle of nodes in G being consolidated as one. The bottom right dotted line box illustrates values of nodes and edges as a zoomed-in version of a subgraph of G'. The scoring function is the summation of value d of each node in G' (length of mismatched site between query haplotype and ones in reference panel) and value r of each edge in G' (recombination rate value translated from genetic distance). If two nodes share the same population label, their edge's value is zero (i.e., $r = 0$). The optimal path found using dynamic programming or a Viterbi-like approach in G' (i.e., bold edges) has the smallest score, which indicates the inferred ancestral population labels of each site in query I'.

Recomb-Mix is a version of the simplified Li and Stephens' model, and such a lightweight model can outperform a complex model, for instance, HapMix, under some simulated scenarios. Unlike HapMix calculating within population transition probability using recombination parameters, Recomb-Mix takes zero penalties for any within population path changes. This simple setting in our model is based on our observation that individuals within a population who share the same genealogical path have equal weight to present the population. Recomb-Mix is able to locate the optimal path if the reference panel contains several path changes within a population as there is no penalty applied. However, HapMix may identify it as a sub-optimal path as it corresponds to a relatively small transition probability due to the number of path changes it has.

Graph optimization problem

We form the local ancestry inference (LAI) as a graph optimization problem. We assume a haplotype of an admixed individual is modeled as a mosaic of existing parental haplotypes, i.e., Li and Stephens' model. Inputs are the segments from a number of phased individual queries Q, a phased reference panel S, and ancestral population values V_s of the panel S: $Q = \{q_j | j \in \{0, 1, ...n-1\}\}$, $S = \{s_{i,j} | i \in \{0, 1, ..., m-1\}, j \in \{0, 1, ..., n-1\}\}$, and $V_s = \{v_{i,j} | i \in \{0, 1, ..., m-1\}, j \in \{0, 1, ..., n-1\}\}$, where m is total number of haplotypes, n is total number of sites. In the graph, each node represents a site of a haplotype from the reference panel. The solution to the given LAI problem is to find the path having the smallest score from node start to node end in the graph using an objective function. The objective function calculates the penalty of each path from the graph. Output V_Q is a list of the

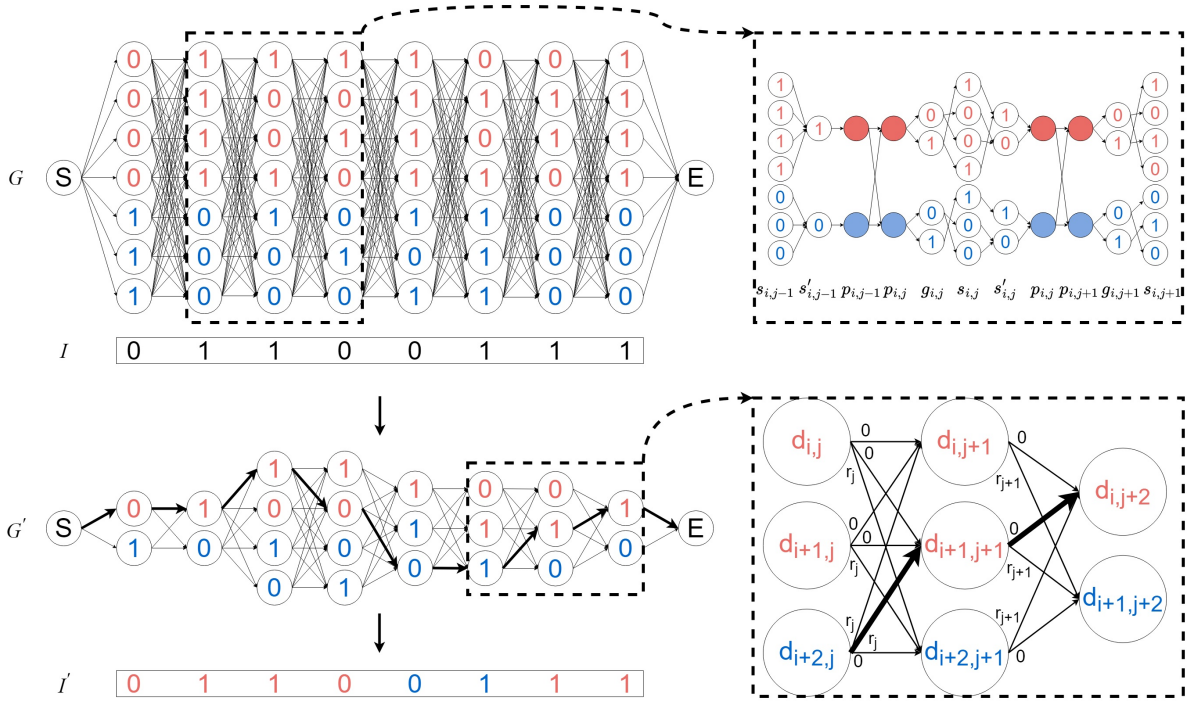

Figure 1: Model and flow chart of Recomb-Mix. G is an original graph built from a reference panel having two populations (shown in red and blue), I is a query of an admixed individual haplotype, G' is the graph collapsed from G, I' is the inferred query of an admixed individual haplotype. The top right dotted line box illustrates the construction on a site from nodes in G to nodes in G' as a zoomed-in version. $p_{i,j}$ is population emission node, $g_{i,j}$ is genotype emission node, $s_{i,j}$ is node from original graph G, and $s'_{i,j}$ is node from compact graph G' at haplotype i and site j. The bottom right dotted line box illustrates values of nodes (d, length of mismatched site between I and one in G') and edges (r, recombination rate value translated from genetic distance between sites in G') as a zoomed-in version of a subgraph of G'. The optimal path (shown in bold), having the smallest score of a summation of all d's and r's values, indicates the inferred ancestral population labels of each site in query I'.

ancestral population values inferred from the optimal path found in the graph: $V_Q = \{v_i | \{i \in \mathcal{P}\}\}$, where $\mathcal{P}$ is a list of the optimal site ids from the panel which forms the most similar haplotype to the query.

One of the advantages of forming LAI as a site-based graph optimization problem is that it works with both phased and unphased reference panels as input. When converting the panel into the graph, the order of the sites from the two haplotypes of an individual is irrelevant, thanks to the property that the out-neighborhood of a node a in a graph is the set of nodes adjacent from a. Thus, our method is flexible to both phased and unphased reference panels.

Compact population graph

We observe that individuals within the same ancestral population cluster are of equal value presenting the population for ancestry inference as they share the same population genealogical path. Figure 2 is an example of schematic population structure showing five known ancestral population clusters (marked as P_1, P_2, P_3, P_4, and P_5), four individuals (marked as I_1, I_2, I_3, and I_4), and population clusters' genealogical paths (shown in blue, red, or black). I_1 and I_2 represent P_1 and P_2 respectively, as they have their own genealogical paths (I_1 is the red path from P_5 to P_1, and I_2 is the red path from P_5 to P_4, then the black path from P_4 to P_2). Both I_3 and I_4 represent P_3 because they share the same genealogical path (the red path from P_5 to P_4, then the red path from P_4 to P_3). When we make the ancestry inference, the path change penalty is only applied to individuals across populations (i.e., when path change happens between individual I_1 and I_2, I_1 and I_3, I_1 and I_4, I_2 and I_3, I_2 and I_4). There is no penalty applied between I_3 and I_4 as they belong to the same population cluster which has a distinct genealogical path. Thus, we set within population path change penalty to zero in our model. The individuals from the reference panel are grouped by population and compiled into data with a much smaller size representing the population (i.e., collapsing the graph). The collapsed graph containing population-level data not only eliminates the noise of penalties when a path is changed between individuals having the same ancestral population label in the graph but also speeds up the entire computation time.

Unlike Loter (Dias-Alves et al. 2018), not taking advantage of the population structure and building

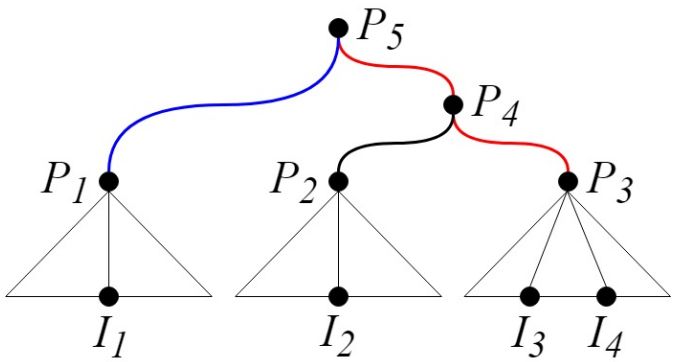

Figure 2: A schematic population structure illustrating five known ancestral population clusters (P_1, P_2, P_3, P_4, and P_5), four individuals (I_1, I_2, I_3, and I_4), and genealogical paths (shown in blue, red, or black). Both individual I_3 and I_4 represent population P_3 as they share the same genealogical path (i.e., the red path from P_5 to P_3).

the reference haplotype graph from the original reference panel, Recomb-Mix builds the graph from a compact version of the original reference panel by keeping the population-level information only. Within each ancestral population, genetic markers having the same allele value are collapsed to form the compact reference panel. Let p be the total number of distinct populations, compact reference panel reduces the space complexity from $O(nm)$ to $O(np)$. Clearly, $p \ll m$, so Recomb-Mix has linear complexity, i.e., independently from the number of individual haplotypes in the reference panel and the query panel. Assuming the haplotypes are bi-allelic, the input panel sizes are reduced to $S' = \{s_{i,j} | i \in \{0, 1, ..., 2p-1\}, j \in \{0, 1, ..., n-1\}\}$, and $V'_s = \{v_{i,j} | i \in \{0, 1, ..., 2p-1\}, j \in \{0, 1, ..., n-1\}\}$. An example of a compact population graph G' collapsed from the original graph G has been shown in Figure 1. A similar approach was used in a phasing method SHAPEIT (Delaneau et al. 2019), where it converts its reference panel into a compact graph in a different way: each node represents its allele value and each edge has the weight of the number of individual haplotypes in its reference panel that traverse it. This approach makes SHAPEIT have its graph $O(nj)$ space complexity (j is the number of conditioning states for each marker), which speeds up its subsequent HMM calculation (Delaneau et al. 2012). FLARE (Browning et al. 2023) used a different technique of haplotype compression to improve its runtime when constructing its HMM state space from its reference panel by leveraging composite reference haplotypes (firstly introduced in an imputation method Beagle (Browning et al. 2018)). For

each query individual haplotype, it finds Identity-By-State (IBS) segments using Positional Burrows-Wheeler Transform (PBWT) algorithm (Durbin 2014) from the reference haplotypes and then makes composite reference haplotypes by stitching those IBS segments. Utilizing the composite reference haplotypes costs $O(nm')$ space complexity, where m' depends on the number and locations of IBS segments of the query individual haplotype against the original reference panel. Usually, m' is relatively small, as it is expected there exist many long IBS segments.

Objective function

In our graph optimization formulation, the ultimate goal is to find an optimal path having a minimal penalty score. Thus, our objective function is to calculate the penalty score of each path in the graph. At a high level, calculating a minimized penalty score as we and Loter do is an approximation of calculating a maximized probability as HapMix does. Generally, the penalty score between the query haplotype and the reference haplotype graph on each path consists of two penalty terms: mismatch penalty and path change penalty. Respectively, they play similar roles to emission probability and transition probability in the HMM approach, and loss function and regularization function in the machine learning approach. For the mismatch penalty, we choose the length of the node to be the penalty score, if such node representing the site is a mismatch from the correlated site in query haplotype. For the path change penalty, we choose the normalized recombination rate across populations to be the penalty score. In the graph, each edge represents the recombination breakpoint between two sites. Thus, the recombination rate is inversely proportional to the probability of path change between two sites. We only penalize it when a path is changed across populations (i.e., a path is switched to another haplotype template having a different population label from the previous one). The cross-population penalty is effective as it considers the population structure through history. In an ancestral recombination graph (ARG), defined populations are clusters, and individuals within each population have a relatively smaller distance from each other than individuals outside their population. thus, in our model, the path change penalty is given between population clusters as those distances are the ones distinguishing populations.

In our model, the optimization problem is minimizing $f(\mathcal{P})$:

$$f(\mathcal{P}) = \sum_{j=0}^{n-1} (|s_{i,j}|_{q_j \neq s_{i,j}} + w r_{j,j+1} {}_{v_{i,j} \neq v_{k,j+1}}), \tag{1}$$

where $\mathcal{P} = \{i\}^n, i \in \{0, 1, ..., 2p-1\}$, which is a n-size list of haplotype template ids from the panel S, indicating the optimal path. q is a site from the query, s is a site from the panel, $|s_{i,j}|$ is the length of site j (i.e., mismatch penalty), $r_{j,j+1}$ is normalized recombination rate between site j and $j+1$ (i.e., path change penalty), $v_{i,j}$ is the ancestral population value of the site $s_{i,j}$, and w is the weight of path change penalty. This is a simplified scoring function like the one HapMix used. To find a path having a minimum score in the compact graph, we use dynamic programming to compute the score with the objective function in Equation 1. It consists of two terms to penalize the paths as mentioned previously. The mismatch penalty term is the penalty of the length of the segment in each node in the graph if such segment is different from the correlated segment in the query. Since each node represents a genetic marker, the penalty value is the length of a marker, i.e., 1. The path change penalty term is the penalty of switching a path to a node representing a different haplotype in the graph. If the outgoing edge of a node points to a node having a different ancestral population from the current one, we penalize the edge with the normalized reciprocal of the recombination rate between the genetic markers represented by the two nodes. For Loter (Dias-Alves et al. 2018), it simply considered the number of switches between haplotypes in the graph. It did not utilize a recombination map, so its path change penalty was a constant value (i.e., 1); since it did not leverage the population structure property, its path change penalty occurred even within the population (i.e. penalizing when the path was shifted to any other haplotype). Recomb-Mix not only considers the quantity of the switches, but the quality of the switches. By utilizing the compact population graph, we treat haplotypes sharing the same population label as one entity and we focus on cross-population penalty only. We also take advantage of the recombination rate as it reflects the probability of a switch occurrence between two genetic markers. The path is encouraged to be switched if the recombination rate is a large value, and vice versa. Thus, we penalize the path using the reciprocal of the recombination rate. Utilizing the recombination rate for the path change penalty helps improve the accuracy of inference results, as we tested (see Supplementary Figure S1). We normalized its value within the range of $[1.0, 2.0]$, so it

is in the same order of magnitude as the mismatch penalty to prevent the domination of any one of them. We adopt the same dynamic programming technique as Loter did to calculate the score but on a compact population graph. The maximum indegree and outdegree of a node in a compact graph is a constant value $2p$ (each genetic marker's allele value is either one or zero, and it belongs to one of the p ancestral populations). Thus, the runtime complexity of computing the scores on the compact graph using dynamic programming is improved from $O(nm)$ to $O(n)$.

Simulated datasets

To evaluate the performance of Recomb-Mix, we simulated admixture datasets using SLiM v4.0 (Haller and Messer 2019). We use Whole Genome Sequencing (WGS) individuals from various populations in the study of 1000 Genomes Project (Auton et al. 2015; Clarke et al. 2016) and Human Genome Diversity Project (HGDP) (Bergström et al. 2020) as the inputs. Each input population is split into disjoint sets of founders and references. The admixture population is simulated as the descendants of admixed founders from different populations. The admixed individuals are sampled from the admixture population, and the referenced individuals are combined ones sampled from each reference population. We simulated a three-way inter-continental dataset of chromosome 18 using YRI, CEU, and CHB individuals (representing African, European, and Asian populations) from the 1000 Genomes Project. The average recombination rate and mutation rate used for the simulation is 1.46455e-08 and 1.29e-08 per base pair per generation, according to stdpopsim library (Adrion et al. 2020). A 0.02% genotyping error, following Browning et al. (2023), is added to both admixed and reference individuals. The ground truth ancestral labels of admixed individuals are extracted from the SLiM output tree sequence (Haller et al. 2019). Additionally, we simulated a seven-way inter-continental dataset of chromosome 18 using AFR, EAS, EUR, NAT, OCE, SAS and WAS individuals (representing African, East Asian, European, American, Oceanian, Central and South Asian, and Middle Eastern (i.e., West Asian) populations) using individuals from HGDP as the input. The HGDP data was phased and imputed using Beagle 5.4 (Browning et al. 2018, 2021). The goal of simulating the seven-way admixture is to explore how well LAI methods are able to distinguish local ancestral segments from the admixture of a large number of ancestral populations.

Besides the conventional inter-continental admixture simulation, we also simulated an intra-continental dataset using TSI, FIN, and GBR individuals (representing Italian, Finnish, and British populations) using the same settings as the inter-continental dataset in order to explore the power of LAI in an intra-continental level. We prepared various sizes of reference panels (i.e., 100, 250, 500, and 1,000) and numbers of generations after admixture (i.e., 15, 50, 100, and 200) for our experiments. To explore the influence of the uneven proportion of individuals per population in founders and references, we simulated two variations of three-way 15-generation intra-continental datasets with uneven founders (i.e. 68 Italian, 32 Finnish, and 100 British individuals) or uneven references (i.e. 170 Italian, 80 Finnish, and 250 British individuals). Both cases are 1/3, 1/6, and 1/2 individuals respectively to the entire population panel.

Results

Recomb-Mix performed well in our simulation scenarios. We ran experiments on Recomb-Mix with the following LAI methods: FLARE (Browning et al. 2023), G-Nomix (Hilmarsson et al. 2021), Loter (Dias-Alves et al. 2018), RFMix (Maples et al. 2013), and SALAI-Net (Oriol Sabat et al. 2022) (Parameters used for each method are in Supplementary Table S1). FLARE was the most recent LAI method. It was a site-based generative method, derived from HapMix (Price et al. 2009), which is one-layer of HMM: hidden local ancestry of a site, and Mosaic (Salter-Townshend and Myers 2019), which is two-layer HMM: hidden local ancestry of a site and hidden haplotype as the source of hidden local ancestry of the site. It showed encouraging speed and accuracy because its performance was optimized for computation resources and its model was designed flexible for optional parameters to deal with various situations. G-Nomix was a window-based discriminative approach using two-layer prediction to perform LAI: Logistic Regression as a base model and XGBoost as a smoother model. It is becoming a current leading LAI method due to its promising speed and accuracy. Loter framed the ancestry prediction problem as a graph optimization problem. It was prioritized for LAI on distant admixture events and good for non-model species as no biological information is required. RFMix, another window-based discriminative approach, was one of the popular LAI methods. It applied a conditional random field model to LAI, particularly using random forest classification. RFMix showed a robust performance on

multi-way admixture datasets. SALAI-Net was also a window-based discriminative approach, which was developed from its predecessor: LAI-Net (Montserrat et al. 2020), a first neural network based LAI. It used a reference matching layer and a smoother layer (a combination of cosine similarity score and neural network) to perform LAI. In addition to the adoption of GPU hardware, SALAI-Net utilized a pre-trained generalized model, making it free from re-training and parameter tuning during the inference; thus, it was considered to be very fast. We selected the above LAI methods for benchmarking because FLARE, G-Nomix, and SALAI-NET were the newly published ones, reporting promising results; RFMix was the most popular one and widely used for ancestry-related applications. Loter had the same problem formulation as Recomb-Mix did. We did not include HapMix because it was not able to tackle multi-way admixture and only produced inference results at the diploid level.

Our experimental runs showed that Recomb-Mix outperformed other LAI methods in both three- and seven-way, inter- and intra-continental admixture datasets. When having an uneven number of individuals per population in the founders or reference panel, all methods' performance was impacted slightly. Under this circumstance, Recomb-Mix was able to achieve the best accuracy.

Local ancestry inference for standard three-way inter-continental admixed populations

For a conventional three-way inter-continental simulated dataset, Recomb-Mix has the best accuracy in reference panel size 100, 250, 500, and 1,000 with 15 generations, as well as in generation 15, 50, 100, and 200 with 500 references. Figure 3 and Figure 4 show the average accuracy of inference results of six LAI methods: FLARE, G-Nomix, Loter, Recomb-Mix, RFMix, and SALAI-Net. Figure 5 illustrates an inferred haplotype sample of an admixed individual from the methods. Overall, as the reference panel size increases, the average accuracy increases for all methods. The large reference panel contains more individual samples than those in small ones, which helps improve the inference result. When using a reference panel having 1,000 individuals, all methods can have at least 92% accuracy, while Recomb-Mix reaches 98.98%. Recomb-Mix performance well for a small reference panel (100 individuals with 97.96% accuracy), while others do not achieve superior accuracy (SALAI-Net and G-Nomix have the second and third high accuracy rates, which are 86.69% and 86.63%).

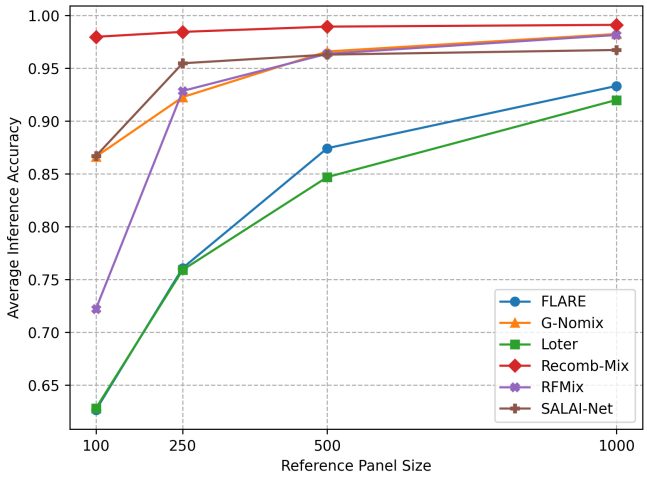

Figure 3: Accuracy with reference panel sizes for three-way 15-generation inter-continental simulated datasets (values are in Supplementary Table S2).

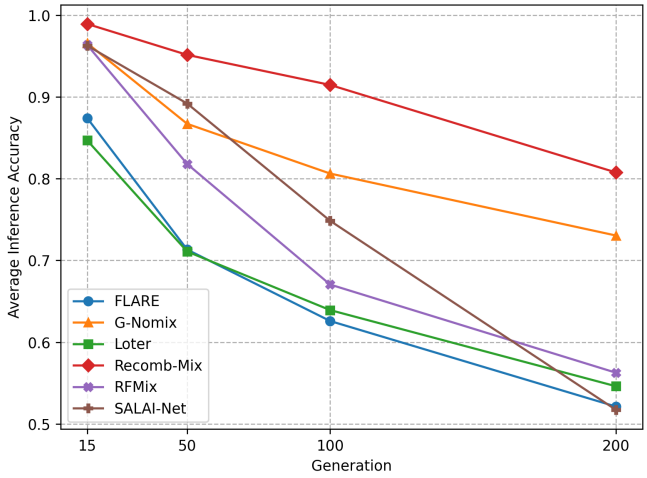

Figure 4: Accuracy with generations for three-way 500-reference inter-continental simulated datasets (values are in Supplementary Table S3).

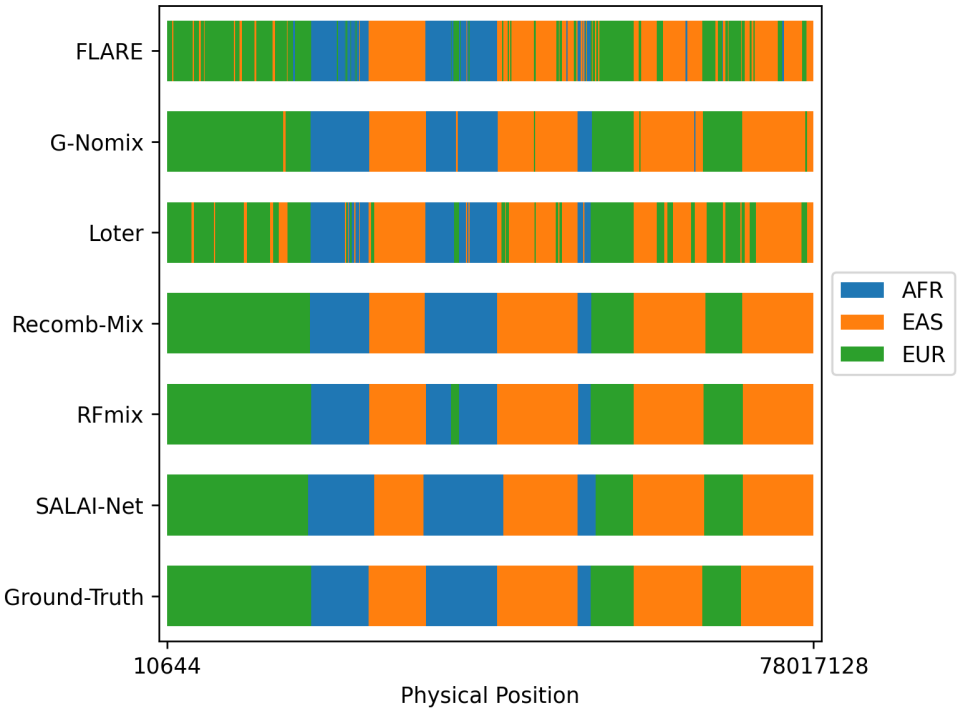

Figure 5: Inferred sample of three-way 15-generation 500-reference inter-continental simulated dataset from LAI methods.

Multi-way admixture

Besides the set of experiments on traditional three-way admixed individuals, we also made a case study on seven-way admixed individuals. We wanted to find out how LAI methods perform on individuals admixed from a large number of founder populations, as in a real case scenario, human individuals are involved in multiple population admixture events. We simulated seven-way inter-continental datasets with various reference panel sizes and generations. For such challenging datasets, the accuracy rates of LAI methods dropped but Recomb-Mix kept performing well. Figure 6 and Figure 7 show the average accuracy of local ancestry inference of six LAI methods: FLARE, G-Nomix, Loter, Recomb-Mix, RFMix, and SALAI-Net with various reference panel sizes and generations, respectively. Figure 8 illustrates an inferred haplotype sample of an admixed individual from the methods. Inferring seven-way admixed individuals is harder than inferring three-way ones, as it involves 2.3 times more population labels. In general, window-based LAI methods perform better than site-based ones, except Recomb-Mix. Using windows helps tolerate errors during the inference since the population label having the highest estimated probability within a window dominates the inference result of the entire window. On the other hand, site-based methods may put more focus on each site's optimal label, which may affect the inference result of one's surrounding region, especially if the number of potential population labels is large. Though Recomb-Mix uses a site-based approach, it achieves high accuracy. Allowing individual variations within the same population helps alleviate overly seeking optimum on local site inference.

For a seven-way 200-generation 500-individual inter-continental WGS dataset, the accuracy of G-Nomix was 1.35% higher than that of Recomb-Mix. G-Nomix was one of the LAI methods good at identifying ancient admixture events, as demonstrated in a recent study (Pearson and Durbin 2023).

Intra-continental admixture

Similar to the result of the inter-continental dataset, Recomb-Mix performs well on the intra-continental dataset. Figure 9 shows the average accuracy of local ancestry inference of six LAI methods: FLARE, G-Nomix, Loter, Recomb-Mix, RFMix, and SALAI-Net on a three-way 15-generation intra-continental simulated dataset. Overall, the average accuracy of each method is worse than the one in the inter-

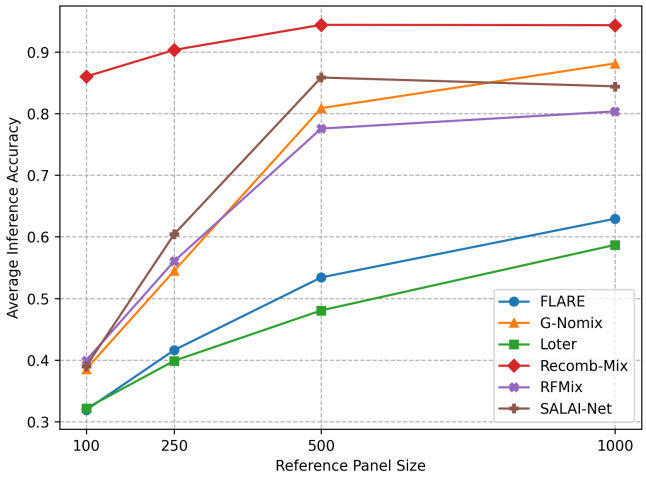

Figure 6: Accuracy with reference panel sizes for seven-way 15-generation inter-continental simulated datasets (values are in Supplementary Table S4).

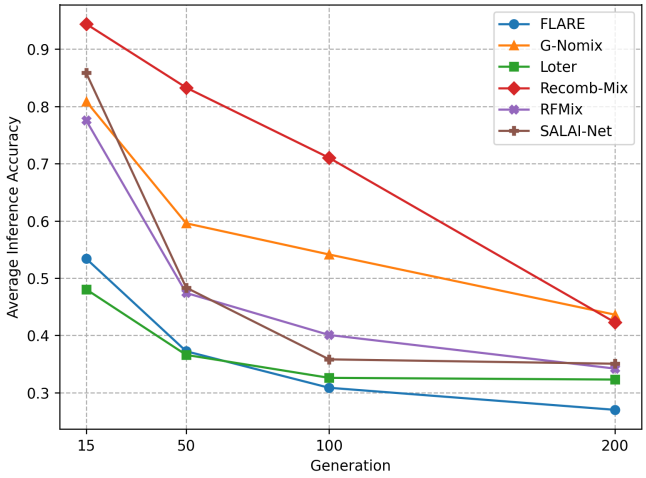

Figure 7: Accuracy with generations for seven-way 500-reference inter-continental simulated datasets (values are in Supplementary Table S5).

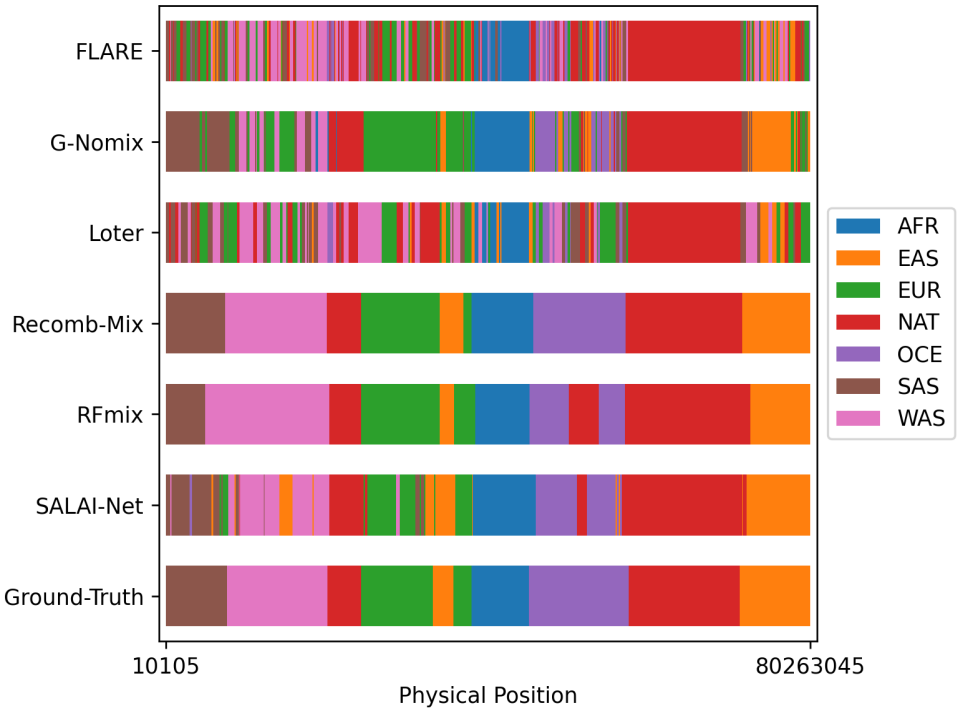

Figure 8: Inferred sample of seven-way 15-generation 500-reference inter-continental simulated dataset from LAI methods.

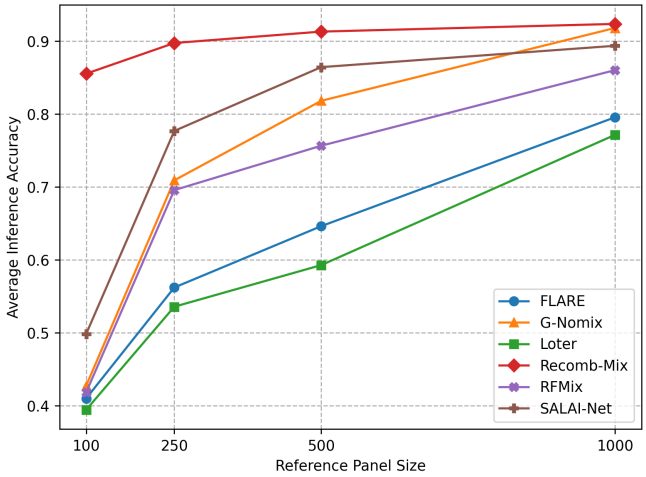

Figure 9: Accuracy with reference panel sizes for three-way 15-generation intra-continental simulated datasets (values are in Supplementary Table S6).

continental dataset. This is expected as the admixture occurring at the intra-continental level generates individuals who resemble each other. Thus, performing LAI on such a dataset is more challenging than that at the inter-continental level. Recomb-Mix has the best accuracy in reference panel sizes 100, 250, 500, and 1,000 with 15 generations. For a 100-individual reference panel the accuracy rate of Recomb-Mix is 85.55% and the second best method, SALAI-Net, only achieves 49.83%. For a 1,000-individual reference panel, the accuracy rate of Recomb-Mix is 92.38%, while the second best method, G-Nomix, achieves 91.82%.

We investigated the impact of the number of generations on LAI. Four three-way 500-reference intra-continental simulated datasets with generations: 15, 50, 100, and 200 were tested, and the result shows the accuracy rate is inversely proportional to the number of generations (see Figure 10). For a 15-generation dataset, Recome-Mix has the highest accuracy at 91.33%. The second and the third high accuracy rates are 86.45% (SALAI-Net) and 81.83% (G-Nomix). For a 200-generation dataset, though the accuracy rate drops for all methods, Recome-Mix keeps the highest accuracy at 72.46%. The second and the third high accuracy rates are 54.38% (SALAI-Net) and 54.07% (G-Nomix).

To verify the robustness of Recomb-Mix handling cases on uneven founder populations and uneven reference panels, We also experimented with LAI using uneven founders for the imbalanced admixture

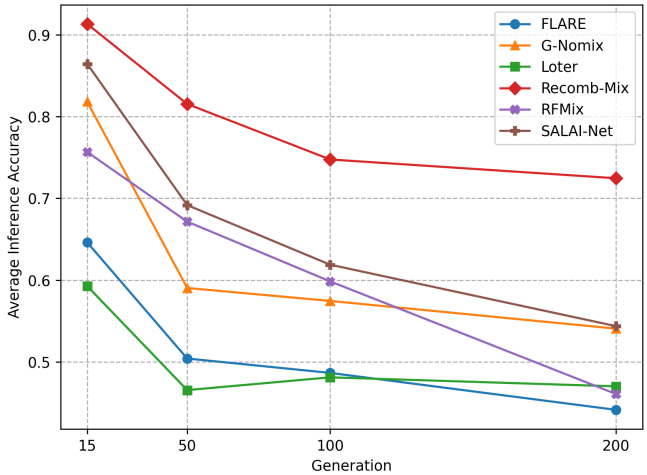

Figure 10: Accuracy with generations for three-way 500-reference intra-continental simulated datasets (values are in Supplementary Table S7).

simulation, as well as using uneven reference panels for the inference. Being uneven means the group consists of one-third of individuals from the first population, one-sixth of individuals from the second population, and half of individuals from the third population. Being even means the numbers of individuals from the populations in the group are divided equally. Three sets of experiments were performed: one three-way admixture dataset simulated using even founders and inferred using even references, one using uneven founders and inferred using even references, and one using even founders but inferred using uneven references.

The result in Table 3 indicates that admixed individuals with uneven founders do not impact the performance of LAI methods much, though the accuracy decreased slightly (on average 3.48%). On the other hand, making LAI calls using an uneven reference panel seems to have a big impact on its outcome as the accuracy decreased some amount (on average 6.65%). Among all LAI methods, RFMix has the smallest impact in both cases (its accuracy decreased by 1.05% and 3.19% respectively), and Recomb-Mix has the highest accuracy in both cases (89.20% and 83.61% respectively). The process of Recomb-Mix generating a compact reference panel helps convert the unbalanced reference populations into balanced ones. Thus, Recomb-Mix keeps a high accuracy of inference results on the unbalanced reference populations by using the collapsed panel.

Method	Even Founders and References	Uneven Founders	Uneven References
Loter	59.28%	51.67%	52.98%
FLARE	64.62%	59.60%	56.85%
RFMix	75.66%	74.61%	72.47%
G-Nomix	81.83%	80.16%	73.88%
SALAI-Net	86.45%	83.07%	79.49%
Recomb-Mix	**91.33%**	**89.20%**	**83.61%**

Table 3: Average accuracy of LAI methods on three-way 15-generation intra-continental simulated datasets with even or uneven number of individuals per population in founder or reference panel.

Recomb-Mix is efficient in memory, space, and runtime

We looked into the runtime and maximum amount of memory LAI methods used for performing LAI on an admixed individual haplotype. Supplementary Table S8 shows the average CPU runtime and maximum amount of physical memory all six LAI methods consumed across different experimental runs. In general, inference on three-way admixed individuals was faster than those on seven-way. This is expected as a seven-way admixture has many more local ancestral segments across the chromosome than ones in a three-way, which costs more time to perform the inference. SALAI-Net was the fastest method and Recomb-Mix was the runner-up but only took 0.31 and 2.04 more seconds than SALAI-Net in three-way and seven-way datasets. From the memory-consuming perspective, Recomb-Mix required the smallest amount of memory, 2.44 and 4.13 GB in three-way and seven-way datasets, respectively.

Recomb-Mix is able to convert the collapsed reference panel graphs into Variant Call Format (VCF) files. Later they can be reused by Recomb-Mix to further save processing time. The size of a compact VCF file is much smaller than the original one since it only contains template individuals with population-level information. For example, the disk space needed to store a 3-way inter-continental 500-reference panel was decreased from 665 to 13.3 MB (and 1.4 MB for a compressed VCF file). Similarly, for a 7-way panel, the disk space was decreased from 787 to 21.8 MB (and 1.9 MB for a compressed VCF file).

Discussion

We presented a new LAI method named Recomb-Mix, based on the formulation of LAI as a graph optimization problem. The between-population penalty and recombination rate penalty, which were heavily weighted in our method, showed promising LAI results under various circumstances. We also introduced a compact form of reference panel used for our method, considering the results of concision and efficiency. Overall, Recomb-Mix is competitive with other state-of-the-art LAI methods in both accuracy and computational performance.

As a site-based LAI method, Recomb-Mix is designed to exploit the site-level information to achieve superior accuracy, especially at the intra-continental level. The number of ancestry informative markers (AIMs) in intra-continental admixed individuals is much less than those in inter-continental admixed individuals as intra-AIMs are a subset of inter-AIMs. Window-based LAI methods may find it difficult to achieve high accuracy during the intra-level ancestry inference. There is a higher probability for each window containing multiple intra-AIMs than that for inter-AIMs. Since the window is the smallest unit representing one ancestral source, windows having intra-AIMs representing different populations may misrepresent the inference result. Decreasing the window size may mitigate the situation, with the potential computational burden. However, its lower bound is one site per window, i.e., site-based. In our method, we capture low-frequency rare variants, which are a subset of conventional AIMs, to distinguish population labels. The step to capture higher frequency AIMs is planned in our future work.

Despite of higher accuracies demonstrated, Recomb-Mix has the following limitations. It does not consider the disparate genetic maps across populations, the allele frequencies, or genotyping errors. Thus, for a well-specified model, other complementary methods may be useful. FLARE takes optional parameters such as minor allele frequency and number of generations since admixture. It may perform well if these biological parameters are correctly estimated from the model. G-Nomix has a few pre-trained models available which may be in handy if one was pre-trained specifically for the given model. SALAI-Net may be a good choice as it employs a pre-trained model that is generalized and applicable to any species and any set of ancestries. Though Recomb-Mix was not designed to handle erroneous panels, the simulated datasets with 0.02% genotyping error rate seem to have no large impact on the

inference results (see Results section). If the error rate is high, a pre-processing step may be needed to correct the noisy data panel.

Software availability

The Recomb-Mix code is available at https://github.com/ucfcbb/Recomb-Mix.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by the National Institutes of Health under grants R01HG010086 and R56HG011509.

References

Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, Kyriazis CC, Ragsdale AP, Tsambos G, Baumdicker F, et al.. 2020. A community-maintained standard library of population genetic models. *eLife* **9**: e54967.

Atkinson EG, Maihofer AX, Kanai M, Martin AR, Karczewski KJ, Santoro ML, Ulirsch JC, Kamatani Y, Okada Y, Finucane HK, et al.. 2021. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nature Genetics* **53**: 195–204.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al.. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Bergström A, McCarthy SA, Hui R, Almarri MA, 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.

464 Browning BL, Tian X, Zhou Y, and Browning SR. 2021. Fast two-stage phasing of large-scale sequence
465 data. *The American Journal of Human Genetics* **108**: 1880–1890.

466 Browning BL, Zhou Y, and Browning SR. 2018. A One-Penny imputed genome from Next-Generation
467 reference panels. *Am J Hum Genet* **103**: 338–348.

468 Browning SR, Waples RK, and Browning BL. 2023. Fast, accurate local ancestry inference with FLARE.
469 *The American Journal of Human Genetics* **110**: 326–335.

470 Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau
471 O, O'Connell J, et al.. 2018. The UK Biobank resource with deep phenotyping and genomic data.
472 *Nature* **562**: 203–209.

473 Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, Tassé AM, and Flicek P. 2016.
474 The international Genome sample resource (IGSR): A worldwide collection of genome variation
475 incorporating the 1000 Genomes Project data. *Nucleic Acids Research* **45**: D854–D859.

476 Delaneau O, Marchini J, and Zagury JF. 2012. A linear complexity phasing method for thousands of
477 genomes. *Nature Methods* **9**: 179–181.

478 Delaneau O, Zagury JF, Robinson MR, Marchini JL, and Dermitzakis ET. 2019. Accurate, scalable
479 and integrative haplotype estimation. *Nature Communications* **10**: 5436.

480 Dias-Alves T, Mairal J, and Blum MGB. 2018. Loter: A Software Package to Infer Local Ancestry for
481 a Wide Range of Species. *Molecular Biology and Evolution* **35**: 2318–2326.

482 Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulier K, Privé F, Vilhjálmsson BJ, Olde Loohuis
483 LM, and Pasaniuc B. 2023. Polygenic scoring accuracy varies across the genetic ancestry continuum.
484 *Nature* **618**: 774–781.

485 Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, and Domingue B. 2019.
486 Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Com-*
487 *munications* **10**: 3328.

488 Durand EY, Do CB, Wilton PR, Mountain JL, Auton A, Poznik GD, and Macpherson JM. 2021. A
489 scalable pipeline for local ancestry inference using tens of thousands of reference haplotypes. *bioRxiv*
490 doi: 10.1101/2021.01.19.427308.

491 Durbin R. 2014. Efficient haplotype matching and storage using the positional Burrows-Wheeler trans-
492 form (PBWT). *Bioinformatics* **30**: 1266–1272.

493 Geza E, Mugo J, Mulder NJ, Wonkam A, Chimusa ER, and Mazandu GK. 2018. A comprehensive
494 survey of models for dissecting local ancestry deconvolution in human genome. *Briefings in Bioin-*
495 *formatics* **20**: 1709–1724.

496 Haller BC, Galloway J, Kelleher J, Messer PW, and Ralph PL. 2019. Tree-sequence recording in SLiM
497 opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources* **19**:
498 552–566.

499 Haller BC and Messer PW. 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher
500 Model. *Molecular Biology and Evolution* **36**: 632–637.

501 Hamid I, Korunes KL, Schrider DR, and Goldberg A. 2023. Localizing Post-Admixture Adaptive
502 Variants with Object Detection on Ancestry-Painted Chromosomes. *Molecular Biology and Evolution*
503 **40**: msad074.

504 Hilmarsson H, Kumar AS, Rastogi R, Bustamante CD, Mas Montserrat D, and Ioannidis AG.
505 2021. High Resolution Ancestry Deconvolution for Next Generation Genomic Data. *bioRxiv* doi:
506 10.1101/2021.09.19.460980.

507 Hou K, Ding Y, Xu Z, Wu Y, Bhattacharya A, Mester R, Belbin GM, Buyske S, Conti DV, Darst BF,
508 et al.. 2023. Causal effects on complex traits are similar for common variants across segments of
509 different continental ancestries within admixed individuals. *Nature Genetics* **55**: 549–558.

510 Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, Reeve MP, Laivuori H,
511 Aavikko M, Kaunisto MA, et al.. 2023. Finngen provides genetic insights from a well-phenotyped
512 isolated population. *Nature* **613**: 508–518.

Li N and Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**: 2213–2233.

Maples BK, Gravel S, Kenny EE, and Bustamante CD. 2013. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics* **93**: 278–288.

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, and Kenny EE. 2017. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**: 635–649.

Montserrat DM, Bustamante C, and Ioannidis A. 2020. Lai-Net: Local-Ancestry Inference with Neural Networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1314–1318.

Oriol Sabat B, Mas Montserrat D, Giro-i-Nieto X, and Ioannidis AG. 2022. SALAI-Net: species-agnostic local ancestry inference network. *Bioinformatics* **38**: ii27–ii33.

Parikh VN, Ioannidis AG, Jimenez-Morales D, Gorzynski JE, De Jong HN, Liu X, Roque J, Cepeda-Espinoza VP, Osoegawa K, Hughes C, et al.. 2022. Deconvoluting complex correlates of COVID-19 severity with a multi-omic pandemic tracking strategy. *Nature Communications* **13**: 5107.

Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Kao WHL, Ruczinski I, Fornage M, Siscovick DS, Zhu X, et al.. 2011. Enhanced Statistical Tests for GWAS in Admixed Populations: Assessment using African Americans from CARe and a Breast Cancer Consortium. *PLOS Genetics* **7**: 1–15.

Pearson A and Durbin R. 2023. Local Ancestry Inference for Complex Population Histories. *bioRxiv* doi: 10.1101/2023.03.06.529121.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, and Myers S. 2009. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLOS Genetics* **5**: 1–18.

Reich D, Patterson N, Jager PLD, McDonald GJ, Waliszewska A, Tandon A, Lincoln RR, DeLoa C, Fruhan SA, Cabre P, et al.. 2005. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nature Genetics* **37**: 1113–1118.

Salter-Townshend M and Myers S. 2019. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics* **212**: 869–889.

Suarez-Pajes E, Díaz-de Usera A, Marcelino-Rodríguez I, Guillen-Guio B, and Flores C. 2021. Genetic Ancestry Inference and Its Application for the Genetic Mapping of Human Diseases. *International Journal of Molecular Sciences* **22**.

Wu J, Liu Y, and Zhao Y. 2021. Systematic Review on Local Ancestor Inference From a Mathematical and Algorithmic Perspective. *Frontiers in Genetics* **12**: 698.