frontiers in Genetics

REVIEW
published: 24 May 2021
doi: 10.3389/fgene.2021.639877

Systematic Review on Local Ancestor Inference From a Mathematical and Algorithmic Perspective

Jie Wu[1,2], *Yangxiu Liu*[1] *and Yiqiang Zhao*[1]*

[1] State Key Laboratory of Agrobiotechnology, China Agricultural University, Beijing, China, [2] Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China

Genotypic data provide deep insights into the population history and medical genetics. The local ancestry inference (LAI) (also termed local ancestry deconvolution) method uses the hidden Markov model (HMM) to solve the mathematical problem of ancestry reconstruction based on genomic data. HMM is combined with other statistical models and machine learning techniques for particular genetic tasks in a series of computer tools. In this article, we surveyed the mathematical structure, application characteristics, historical development, and benchmark analysis of the LAI method in detail, which will help researchers better understand and further develop LAI methods. Firstly, we extensively explore the mathematical structure of each model and its characteristic applications. Next, we use bibliometrics to show detailed model application fields and list articles to elaborate on the historical development. LAI publications had experienced a peak period during 2006–2016 and had kept on moving in the following years. The efficiency, accuracy, and stability of the existing models were evaluated by the benchmark. We find that phased data had higher accuracy in comparison with unphased data. We summarize these models with their distinct advantages and disadvantages. The Loter model uses dynamic programming to obtain a globally optimal solution with its parameter-free advantage. Aligned bases can be used directly in the Seqmix model if the genotype is hard to call. This research may help model developers to realize current challenges, develop more advanced models, and enable scholars to select appropriate models according to given populations and datasets.

Keywords: LAI model, HMM, mathematical structure, bibliometrics, benchmark

OPEN ACCESS

Edited by:
Marcio Dorn,
Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Yuriy L. Orlov,
I.M. Sechenov First Moscow State Medical University, Russia
Manuel Villalobos,
University of Santiago, Chile

***Correspondence:**
Yiqiang Zhao
Yiqiangz@cau.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *10 December 2020*
Accepted: *12 April 2021*
Published: *24 May 2021*

Citation:
Wu J, Liu Y and Zhao Y (2021) Systematic Review on Local Ancestor Inference From a Mathematical and Algorithmic Perspective. Front. Genet. 12:639877. doi: 10.3389/fgene.2021.639877

INTRODUCTION

Rapid advancements in computing technologies, genome sequencing, and single nucleotide polymorphism (SNP) genotyping methods have made it possible to infer the genomic structure at a fine scale (Kidd et al., 2012). It also accelerates the exploration of mixed ancestry or local ancestry inference (LAI) at the individual and population levels (Schumer et al., 2020). In LAI, each chromosome is considered as a mosaic of genomic segments, originated from multiple ancestral groups (Padhukasahasram, 2014). LAI is of great importance in studying population evolution, migration history, or disease risks (Fitak et al., 2018). Up to now, various LAIs have been widely

used; each model comes with its own advantages and disadvantages toward LAI in admixed populations (Geza et al., 2019).

Due to the genetic recombination after interbreeding, the genome consists of mosaic of DNA segments with different genetic ancestries (Dougherty et al., 2017). Genotypes from putative ancestral populations are mostly utilized to infer the local ancestry of admixed individuals (Sankararaman et al., 2008a). Currently, about 70% of LAI models are based on hidden Markov model (HMM), where the hidden states correspond to ancestries and generate the observed haplotypes/genotypes (Baran et al., 2012). LAI models use ancestry informative markers (AIMs) for simplicity or to account for linkage disequilibrium (LD) of variants, i.e., STRUCTURE (Falush et al., 2003), Hapmix (Price et al., 2009), Saber (Tang et al., 2006), and LAMP-LD (Baran et al., 2012). Other models consider rich haplotype information by employing window-based strategies, i.e., RFMix (Maples et al., 2013), PCAdmix (Brisbin et al., 2012), and LAMP (Sankararaman et al., 2008b). **Table 1** presents more details in this regard.

MODELS BASED ON AN ORIGINAL HIDDEN MARKOV MODEL

The challenge of identifying ancestry along each chromosome can be addressed with different approaches. One of the most widely used models is HMM, an extension of a Markov chain, in which the state transformation is generally unobservable (Wu and Zhao, 2019). In HMM, the parameters include initial state distributions, state transition probability matrix, and emission probability matrix. Algorithms were developed to solve three main questions of HMM: evaluation (forward algorithm), decoding (Viterbi algorithm), and training (Baum–Welch algorithm including expectation maximization or maximum likelihood) (Schuster-Böckler and Bateman, 2007).

LAI models based on the original HMM algorithm include Hapmix (Price et al., 2009), Seqmix (Hu et al., 2013), PCAdmix (Brisbin et al., 2012), Supportmix (Omberg et al., 2012), and LAMP-LD (Baran et al., 2012). These models use Baum–Welch to iteratively update the initialized transition probability matrix and the emission probability matrix and use Viterbi for estimating the hidden ancestral states. The designs of the initialized emission and the subsequent calculations mainly differentiate among the models. Supportmix utilizes a support vector machine (SVM) (Haasl et al., 2013) for classifying the chromosome segments of the ancestral group, while PCAdmix calculates Euclidean distances between the ancestral groups and admixed individuals for finding the closest ancestry for each window.

Hapmix Model

The Hapmix model (Price et al., 2009) is based on a combination of the HMM and haplotype. The hidden state for position s is denoted *via* a triplet (i,j,k); here, i denotes the ancestry derived from a different population, while j recorded the population from which the haplotype was copied considering miscopying, and k corresponds to the source of the individual the chromosomal segment was copied from. $p^s(i,j,k;l,m,n)$ is the transition probability from state (i,j,k) to state (l,m,n) between the adjacent sites s and $(s + 1)$. ${e^1}_{ijk}(s)$ denotes the type 1 offspring chromosome probability at site s and t_{jk} represents the parent individual k type in the reference population j. The initialized emission probability matrix is given in Equation (1).

$$e^1_{ijk}(s) = \begin{cases} \theta_i \delta\left(t_{jk} = 0\right) + (1-\theta_i)\,\delta\left(t_{jk} = 1\right) & if\ i = j \\ \theta_3 \delta\left(t_{jk} = 0\right) + (1-\theta_3)\,\delta\left(t_{jk} = 1\right) & if\ i \neq j \end{cases} \quad (1)$$

Here, offspring carrying the identical type to the specific parent is with a probability $(1 - \theta_1)$, while a different type with the probability θ_1, θ_3 denotes the mutation rate in the case that offspring copied from the other population.

Seqmix Model

The Seqmix model (Hu et al., 2013) aligns bases directly rather than relying on genotypic calls. The method implemented in Seqmix consists of three layers: the hidden ancestry state, the hidden genotype, and the observed sequence reads. The genotype is placed in the intermediate layer by connecting the sequence reads and ancestry. In the HMM, the transition matrix denotes the hidden ancestry state q_s as (A_{s1}, A_{s2}), Herein, A_{s1} represents the first chromosome ancestry at site s, while the ancestry of the other chromosomes is represented by A_{s2}. $\gamma_{s,s+1}$ is the rate of recombination per generation between site s and $s + 1$ and T represents the generations since admixture. π_A and π_E correspond to the prior probabilities for populations 1 and 2. The initialized transition probability matrix is given in Equation (2).

$$P_{s,s+1} = \begin{bmatrix} P^{E,E}_{s,s+1} & P^{E,A}_{s,s+1} \\ P^{A,E}_{s,s+1} & P^{A,A}_{s,s+1} \end{bmatrix} = \begin{bmatrix} 1-\left(1-e^{-\gamma_{s,s+1}T}\right)\pi_A & \left(1-e^{-\gamma_{s,s+1}T}\right)\pi_A \\ \left(1-e^{-\gamma_{s,s+1}T}\right)\pi_E & 1-\left(1-e^{-\gamma_{s,s+1}T}\right)\pi_E \end{bmatrix} \quad (2)$$

The initialized emission probability is $P(O_s|\ q_s)$, which is calculated as a sum of the overall possible genotypes, assuming the Hardy–Weinberg equilibrium, and is weighted by ancestry-specific allele frequencies: $P\left(O_s|q_s = (A_{s1}, A_{s2})\right)$. The genotype likelihood $P(O_s|\ q_s)$ is the probability of the observed set of reads given the hidden ancestry state.

PCAdmix Model

The PCAdmix model (Brisbin et al., 2012) is based on a combination of the HMM and principal component analysis (PCA). The principal components (PCs) of the ancestral populations are firstly calculated based on the phased genotypes of the ancestral representatives and the phased genotypes of admixed individuals projected onto the component space. The vector $P(S_{i,w}|anc_{i,w} = j)$ defines the emissions probability, $\text{anc}_{i,w}$ denotes the ancestry of haplotype i at window w from population j and comprises the ancestry scores across the first $K - 1$ PCs, where K is the total count of ancestral populations, the weighted sum $S_{iw} = L_w g_{iw}$ is the ancestry score for haplotype i in window w, g_{iw} represents a column vector of the haplotype's alleles in the window, and L_w represents a matrix in which

TABLE 1 | Detailed deconvolution model description.

Model name	Split format	Algorithm	The number of ancestral populations	Built-in phasing error correction	Reference populations	Admixed populations	LD	References
Hapmix	Haplotype	HMM	2	Yes	Phased	Unphased	Yes	Price et al., 2009
Seqmix	Haplotype	HMM	2	No	Unphased	Unphased	No	Hu et al., 2013
PCAdmix	Window	PCA + HMM	> = 2	No	Phased	Phased	Yes	Brisbin et al., 2012
Supportmix	Window	SVM +HMM	> = 2	No	Phased	Phased	Yes	Omberg et al., 2012
LAMP-LD	Window	HMM	2, 3, 5	No	Phased	Unphased	Yes	Baran et al., 2012
ALLOY	Window	F-HMM	> = 2	Yes	Phased	Phased	Yes	Rodriguez et al., 2013
Saber	Haplotype	M-HMM	> = 2	No	Phased/ Unphased	Phased/ Unphased	Yes	Tang et al., 2006
SWITCH	Haplotype	MCMC	> = 2	No	Phased	Phased	Yes	Sankararaman et al., 2008a
HAPAA	Haplotype	H-HMM	> = 2	Yes	Phased	Phased	Yes	Sundquist et al., 2008
ELAI	Haplotype	Two-layer HMM	> = 2	No	Phased/ Unphased	Phased/ Unphased	Yes	Guan, 2014
RFMix	Window	RF + CRF	> = 2	No	Phased	Phased	No	Maples et al., 2013
Chromopainter	Haplotype	PCA+MCMC	> = 2	No	Phased	Phased	Yes	Lawson et al., 2012
LAMP	Window	ICM	> = 2	No	Unphased	Unphased	No	Sankararaman et al., 2008b
Loter	Haplotype	DP	> = 2	Yes	Phased	Phased	No	Dias-Alves et al., 2018
EILA	Haplotype	FQR + *K*-means	> = 2	No	Unphased	Unphased	No	Yang et al., 2013
LASER 2.0	Haplotype	PCA + PPA	> = 2	No	Phased	Phased	No	Wang et al., 2015
WINPOP	Window	DP	> = 2	Yes	Unphased	Unphased	No	Pasaniuc et al., 2009

the individual columns carry the PC loadings of one SNP in the window; each window is used as the observation value in HMM. The transition probability is $P(\text{anc}_{i,w} = j | \text{anc}_{i,w-1} = k)$. A forward–backward algorithm is applied to find the posterior probability for each window in the admixed haplotype.

Supportmix Model

In the Supportmix model (Omberg et al., 2012), SVM and HMM algorithms are combined, and independent SVM classifiers are firstly applied for each genomic window to retrieve putative ancestry origins. The outputs of the SVMs are then fed to HMM to refine the ancestral assignment for each window. The emission possibilities are p for the hidden state $(1 - p)/(k' - 1)$ and for the other states, where k' is the number of ancestral populations and p is the classification from the SVM at the corresponding window. LD is considered in the HMM where the recombination is modeled as a Poisson process. The transition probability is thus defined as $(1 - e^{-gd})/(k' - 1)$, where d is the genetic distance (in centimorgan) between the windows and g is the generation since admixture.

LAMP-LD Model

The LAMP-LD model (Baran et al., 2012) uses a window-based HMM, which divides the genome into non-overlapping windows of fixed length L with a fixed state space of hidden ancestry of $\binom{K}{2}$. The admixed chromosome is modeled by HMM corresponding to each ancestry pair $S^w = (M_1^w, M_2^w)$. Genotypic block G^w is emitted by each state(M_1^w, M_2^w) with the emission probability: $\sum_{(H_1^w, H_2^w)} P\left(H_1^w | M_1^w\right)\left(H_2^w | M_2^w\right)$. Here, $P(H_1^w | M_1^w)$ is the probability that the haplotype segment H_1^w is emitted under the ancestry M_1 and (H_1^w, H_2^w) is the haplotype pair consistent with the genotypes. The transition probability between the two states in a consecutive window (M_1^w, M_2^w) and $\left(M_1^{w'}, M_2^{w'}\right)$ is set to the average recombination rate per base per generation $\theta = 10^{-8} \times D$ (D denotes the length in base pairs between windows) if the unordered ancestry pairs (M_1^w, M_2^w) and $(M_1^{w'}, M_2^{w'})$ differ by one ancestry, θ^2 if both ancestries differ, or $1 - 2\theta - \theta^2$ if there is no ancestry switch.

MODELS BASED ON A HIDDEN MARKOV MODEL FAMILY

The HMM family, based on an extension of the original algorithm, includes factorial-HMM (F-HMM), hierarchical-HMM (H-HMM), Markov-HMM (M-HMM), conditional random field (CRF), and two-layer HMM. Their transition and emission probabilities have been improved for reinforcing the learning of the original HMM. LAI models based on the HMM family include ALLOY (Rodriguez et al., 2013), Saber (Omberg et al., 2012), HAPAA (Sundquist et al., 2008), ELAI (Guan, 2014), and SWITCH (Sankararaman et al., 2008a). ALLOY applies a F-HMM to get hold of the parallel process, thus giving rise to the paternal and maternal admixed haplotypes. This, in turn, strengthens the correction of the HMM parameters, especially for the emission probabilities. Saber and SWITCH improve and enhance the traditional emission probabilities at a marker by using the joint distribution of alleles at two neighboring markers. SWITCH depends on pairwise SNP allele frequencies between consecutive markers, whereas the Saber model relies on the allele frequencies at the two consecutive markers. Unlike

the M-HMM emission probability models of SWITCH and Saber, HAPAA has an emission probability of a 5 × 5 stochastic matrix and is historically the first model of the series (Sundquist et al., 2008). Most of the transition probabilities still consider the genetic distance and generations in extended HMM. Like Supportmix, RFMix adopts a kind of multi-classification models for investigating chromosome segments of similar ancestry and uses CRF to smooth ancestral window information.

ALLOY Model

The ALLOY model (Rodriguez et al., 2013) uses F-HMM and is an improved form of HMM to capture parallel processes for producing the maternal (m) and paternal (p) admixed haplotypes. This model is denoted by H_l^m, H_l^p, the haplotype cluster membership drawn from $a_l \in A_l$ on the haplotypes at position l. $G_l \in \{0,1,2\}$, which is the observed genotype at the same marker position, represents the count of the minor allele. Across all the positions of the L marker, the presence of vectors of haplotype cluster memberships and genotypes are represented by $H^{\{m,p\}} = (H_1^{\{m,p\}}, H_2^{\{m,p\}}, ..., H_L^{\{m,p\}})$ and $G = (G_1, G_2, \ldots, G_L)$, correspondingly. In the model, the posterior marginal is first computed to infer the emission probability, given the sample of genotypes $P(H_l^m, H_l^p|G)$ by applying the forward–backward algorithm. Local observation is made from the multiplication of the emission probability $P(G_l|H_l^m = a_l, H_l^p = a_l^{'})$ and by incorporating the transition probability of $(H_l\,|H_{l-1})$.

Saber Model

The Saber model (Tang et al., 2006) computes the posterior probability of the hidden states in the M-HMM based on forward and backward algorithms and adds the relationship between the observed genotype along each chromosome. The transition probabilities of the initial state are given in Equation (3).

$$P(Z_1 = i|\pi) = \pi_i,\ (i = 1, \cdots, N), A_{ij}^{\text{struct}}$$
$$(t) = P\left(Z_t = j|Z_{t-1} = i, \tau, \pi\right) \quad (3)$$

where Z_t represents unobserved ancestry, π represents the genome-wide average individual admixture, and τ is the time since admixing.

The distribution of $O_t{}^f$ given $Z_t{}^f$ is described by the emission probability; $O_t{}^f$ represents the observed genotype. The allele frequency in each ancestral population is considered as a natural choice of emission probabilities at a particular marker. In M-HMM, the model further requires the alleles' joint distribution at two neighboring markers. Equation (4) can be defined as the emission probability at marker t.

$$B_t\left(v, u, j, i\right) = P\left(O_t^f = v|O_{t-1}^f = u, Z_t^f = j, Z_{t-1}^f = i\right) \quad (4)$$

SWITCH Model

The SWITCH model (Sankararaman et al., 2008a) uses M-HMM and presents an effective initialization procedure that yields a highly accurate outcome at a notably reduced cost of computation *via* the expectation maximization (EM) algorithm for the estimation of parameters. In each EM iteration, the ancestry information of each haplotype is represented by matrix Z, and matrix W denotes recombination events. The Z and W updates are computed with the help of the Viterbi algorithm having emission probabilities $P_r(X_{i,j}|Z_{i,j}, p_j, q_j)$, which are replaced with an integral of p_j and q_j; the noticed SNP binary matrix has been represented by $X_{i,j}$ at the j-th SNP of the i-th haplotype. The expectation step includes the calculation of the posterior probabilities of p_j and q_j; that is, $P_r(p_j, q_j|X_{i,j}, Z_{i,j}^{(t)})$. The underlined step can be performed *via* Bayes' theorem. The maximization step includes finding a solution to m separate optimization problems in Z_i, W_i, $i\in\{1, m\}$, where the vector of ancestries for the i-th haplotype is represented by Z_i and the complementary vector of recombination events is shown by W_i, as shown in Equation (5).

$$\log\left[\Pr\left(Z_{i,1}|\alpha\right)\right] + I_{1,i}\left(Z_{i,1}\right) + \sum_{j=2}^{n} \{I_{j,i}(Z_{i,j}) + f_{i,j-1,j}\left(Z_{i,j-1}, Z_{i,j}, W_{i,j}\right)\}. \quad (5)$$

where $f_{i,j-1,j}(Z_{i,j-1}, Z_{i,j}, W_{i,j})$. corresponds to the log transition probabilities and $I_{j,i}\left(Z_{i,j}\right)$ represents the expectations of the log emission probabilities. α refers to the fraction of the first population in the ancestral population.

HAPAA Model

In the HAPAA model (Sundquist et al., 2008) based on H-HMM, an integration of the model with multiple HMMs is used. The model assumes the N populations $P = \{P_1, P_2, \ldots, P_N\}$, each P denoted *via* a set of n_p model individuals, $P_p = \{a_{p1}, a_{p2}, \ldots, a_{pn_p}\}$. The probability of emission is given by a 5 × 5 stochastic matrix, $P\left(\bar{a}_i = x|y_i = S_{pkh}\right)$, where the hidden state variable is denoted y_i. S_{pkh} is for the two haplotypes $h \in \{0, 1\}$ of each k individual in the p population. After that, an emitting state starts with an equivalent probability for the individual population, which is provided as $P(y_1 = S_{pkh}) = 1/2Nn_p$. Every S_{pkh} state can exist in three transitions: back to itself and the other presumed haplotype in the very individual $S_{pk(1-h)}$ with a probability of $(1 - w_{pki})e^{-\tau_p R_i}$, and $w_{pki} \cdot e^{-\tau_p R_i}$, respectively, or to the state Out_p exit with probability $1 - e^{-\tau_p R_i}$. Training samples provide the recombination rate τ_p, the probability of a phasing switch error is represented by w_{pki}, R_i represents the genetic distance between the loci, the emission probability is represented by $P\left(\bar{a}_i = x|y_i = S_{pkh}\right)$, and the transition probability is represented by $P(\text{Out}_p \rightarrow In_{p'})$, and using an EM algorithm to update these parameters on the training examples.

ELAI Model

In the ELAI model (Guan, 2014), a two-layer HMM is used: the upper-layer switch probabilities provide the information regarding the switching frequency between various ancestral populations, while the lower-layer switch probabilities are related to the switching frequency between the haplotypes within each ancestral population. For each individual i, let $X_m{}^{(i)}$, $Y_m{}^{(i)}$ be the hidden state of the upper and lower clusters at marker m. Herein, $X_m{}^{(i)}$ obtains values in $1,\cdots S$, S and $Y_m{}^{(i)}$ obtain values in $1,\cdots K$, K. The haplotypic marker $h_m{}^{(i)}$ emission of i at m from

a lower-layer cluster is given in Equation (6).

$$P\left(h_m^{(i)}|X_m^{(i)}, Y_m^{(i)}, \xi\right) = P\left(h_m^{(i)}|Y_m^{(i)}, \xi\right) \quad (6)$$

The complete data likelihood combines with the lower-layer and upper-layer clusters, as shown in Equation (7).

$$P\left(h^{(1)}, \ldots, h^{(N)}, X^{(1)}, Y^{(1)}, \ldots, X^{(N)}, Y^{(N)}|\xi\right) = \prod_{i=1}^{N}\prod_{m=1}^{M} P(h_m^{(i)}|Y_m^{(i)}, \xi)P(X_m^{(i)}, Y_m^{(i)}|\xi) \quad (7)$$

where ξ is defined as the parameter correlating with the HMM.

The first marker and the Markov transitions are expressed as follows because the model takes two scales of LD occurring in admixed individuals into consideration: $P\left(X_1 = \mathrm{s}, Y_1 = \mathrm{k}\right) = P\left(Y_1 = k|X_1 = s\right)P\left(X_1 = s\right)$ and $P\left(X_m = s, Y_m = k|X_{m-1} = s^{'}, Y_{m-1} = k^{'}\right)$.

RFMix Model

In this model (Maples et al., 2013), CRF and the random forest (RF) (Wu and Zhao, 2019) algorithm are combined. In the event of CRF along with its chain structures, all potential functions work on pairs of haplotype label variables, H_i and H_{i+1}, that are adjacent to each other. Firstly, the emission probability is learned and RF is trained with segments (reference haplotypes) in the corresponding window, which is then used for the estimation of the ancestry $A_{i,*}$ posterior probabilities, considering the segment of the admixed haplotype for the window. Secondly, the transition probability is also learned. In adjacent windows, the joint probability of the local ancestries relies primarily on the global proportion of the individual ancestry and the likeliness of recombination between the pair of windows. The joint probability distribution is $P(A_{i,p} = j, A_{i,p+1} = k)$. Thirdly, a linear-chain CRF is independently used to model $P(A_{i,*}|H_{i,*}{:}\Theta)$ for each admixed chromosome. The EM method is used for updating the above parameters. In consideration of a phasing error, $P(A_{i,*},A_{ic,*},H_{i,*},H_{ic,*}|\ O_{i,*},O_{ic,*}{:}\Theta)$ is modeled, wherein i and i_c are the indices representing both copies of the chromosome under evaluation for a specific admixed subject, $O_{i,*}$ represents the phased sequence observed for chromosome i given by phasing algorithms, while $H_{i,*}$ indicates the set of each potential haplotype in the window.

MODELS BASED ON NON-HIDDEN MARKOV MODEL FAMILY

Along with the HMM family models, there are also some other non-HMM family models that are based on the basic algorithm and data mining techniques. For example, Loter is a parameter-free model that uses dynamic programming (DP) to obtain a globally optimal solution. Chromopainter adopts PCA for investigating chromosome segments of similar ancestry and uses Markov chain Monte Carlo (MCMC) (Gilks, 1999) to smooth ancestral segment information.

Chromopainter Model

The Chromopainter model (Lawson et al., 2012) works based on PCA and MCMC (Gilks, 1999). Firstly, PCA uses the co-ancestry matrix $x_{ij.}$ For each element in the matrix, x_{ij} is an estimate of the number of discrete segments of individual i, which is strongly correlated with the individual j corresponding part. The Chromopainter model is built on the assumption that the chunks $P_{q_iq_j}/\hat{n}_{q_j}$ in various individuals are independent; hence, the cross individuals are multiplied, which results in a complete likelihood, as shown in Equation (8).

$$F\left(x|p, q\right) = \prod_{i=1,\, j=1}^{N}\left(\frac{P_{q_iq_j}}{\hat{n}_{qj}}\right)^{x_{ij}/c} \quad (8)$$

where c could be considered as describing an effective number of chunks, N represents the number of individuals, while the individuals are represented by j and i in populations q_j and q_i, accordingly. Probably a single chunk delivered from the j to the i individual is $P_{q_iq_j}/\hat{n}_{q_j}$, and in various individuals, the chunks are independent.

Secondly, a prior value $P_a \sim \mathrm{Dirichlet}(\beta_a = \{\beta_{a1}, \ldots, \beta_{aK}\})$ is selected. β_{ab} values are proportionate to the *a priori* estimated value of each P_{ab}. Eventually, F is updated within the algorithm *via* the updates of standard Metropolis–Hastings MCMC.

LAMP Model

In this model (Sankararaman et al., 2008b), a clustering algorithm called iterated conditional model (ICM) is used to investigate an optimal classification of all individuals regarding probability. The ICM algorithm is different from the traditional EM model. The E step comprises the expected classification θ, given minor allele frequencies f_l, thus resulting in a fractional class membership for each individual i. In the LAMP, it is supposed that a logical answer will be provided by the initial classification, and it determines the maximum *a posteriori* estimate of θ, as indicated here.

For populations A_s and A_t, the underlined model uses G_i, which represents the genotype $(g_{i1}, \ldots, g_{in})$ of the individual i, as shown in Equation (9).

$$\hat{\theta}\,(i) = arg\max_{A_sA^2_{t\in\{1,\ldots,K\}}} P_r\left[\theta\,(i) = A_sA_t|f_1, \ldots, f_k, G_i\right] \quad (9)$$

In the M step, it receives the maximum–likelihood estimation of $f_1, \ldots, f_k$ *via* investigation, as shown in Equation (10).

$$arg\max_{f_1,\ldots,f_k} P_r\left[(G_i)_{i=1}^{m}|f_1, \ldots f_k, \theta\right] = \prod_{i=1}^{m} P_r[G_i|f_1, \ldots, f_k, \theta(i)] \quad (10)$$

Loter Model

The Loter model (Dias-Alves et al., 2018) adopts DP and supposes that ancestral populations contain individuals n, which results in haplotypes $(2n)$ presented *via* $(H_1, \ldots, H_{2n})$. The i-th haplotype value (0 or 1) at the j-th SNP is indicated *via* $H_i{}^j$. The estimation of the haplotype h (admixed individual) is made possible by a vector $(s_1, \ldots, s_p)$ that determines the sequence (haplotype labels). For the j-th SNP in the dataset, $s_j = k$ if haplotype h resulted from the haplotype H_k copy. The optimization problem

comprised reducing the underlined cost function, as shown in Equation (11).

$$C\left(s_1, \ldots, s_p\right) = \sum_{j=1}^{p} |h^j - H^j_{s_j}| + \lambda \sum_{j=1}^{p-1} 1_{s_j \neq s_{j+1}} \quad (11)$$

In consideration of a phasing error, shown in Equation (12)

$$C'(\Theta) = \sum_{j=1}^{p} |a^j - A^j_{s_j}| + \sum_{j=1}^{p} |a'^j - A^j_{s_j}| + \lambda \sum_{j=1}^{p-1} 1_{s_j \neq s_{j+1}} + \lambda \sum_{j=1}^{p-1} 1_{s'_j \neq s'_{j+1}} \quad (12)$$

where $(s_1, \ldots, s_p)$ is in $\{1, \ldots, 2n\}^p$. A regularization parameter, called λ, is involved in an optimization problem. A high λ strongly penalizes the transition between the parental haplotypes of long chunks of the constant local ancestry. $A_1 = (0, \ldots, 0)$ and $A_2 = (1, \ldots, 1)$ represent two possibility ancestry states; haploid local ancestry is represented two by vectors, $a \in \{0, 1\}^P$ and $a' \in \{0, 1\}^P$.

EILA Model

In the EILA model (Yang et al., 2013), fused quantile regression (FQR) and the *k*-means classifier are used and are based on three steps. Firstly, EILA defines a score $e_{j,i}$ (a continuous variable with a range of 0–1) for the admixed genotype $g_{j,i}$ (= 0,1,2) as the probability that $g_{j,i}$ is the descendant of ancestry *A*. This is shown in Equation (13).

$$e_{j,i} = P_r\left[g_{j,i} \in A \mid g^{(A)}_{j,1}, \ldots, g^{(A)}_{j,n_1} \text{ and } g^{(B)}_{j,1}, \ldots, g^{(B)}_{j,n_2}\right] \quad (13)$$

Secondly, $\theta_{j,i}$ is defined as a smooth series and infers the site of breakpoints for ancestral blocks by using FQR and $\theta_{j,i}$ is estimated *via* investigating the value that minimizes $\sum_{j=1}^{m} \left|e_{j,i} - \theta_{j,i}\right| + \lambda \sum_{j=2}^{m} \left|\theta_{j,i} - \theta_{j-1,i}\right|$. Smaller λ will lead to the lowering of penalty effects. The fitted value of $\theta_{j,i}$ is closer to the observed $e_{j,i}$. Thirdly, the breakpoints for all admixed individuals are investigated, and the model infers the local ancestry for all segments between breakpoints *via* *k*-means to obtain a high power of inference.

LASER 2.0 Model

In the LASER 2.0 model (Wang et al., 2015), PCA and projection Procrustes analysis (PPA) are combined. Firstly, PCA is conducted on the genotypes of a set that has been chosen from the *N* reference individuals and results in the construction of a *K*-dimensional ancestry map. For all the evaluated samples, further PCA is carried out on genotypes through overlapping markers between the *N* reference individuals and the evaluated sample and for obtaining a K'-dimensional map corresponding to $N + 1$ individuals (K' greater than or equal to *K*). Furthermore, PPA is performed to determine the transformation optimal set on the PCA map (sample-specific) for the maximization of its resemblance with the reference ancestry map. For the similar *N* reference individuals, the two sets of coordinates are given, i.e., $X_{N \times K'}$ and $Y_{N \times K}$, and the PPA investigates a set of transformations *f* to project *X* from a K'-dimensional space to a *K*-dimensional space and reduces the squared Euclidean distances being added between $f(X)$ and *Y*. Supposing that *X*, as well as *Y*, has been centered toward the origin, the objective of the model is to investigate an isotropic scaling factor, ρ, in such a way that the minimization of $||\rho XA - Y||_F^2$ and the orthonormal projection matrix $A_{K' \times K}$ takes place.

Statistics and Comparison

Here, we performed a bibliometric analysis of the LAI research. "Local ancestry inference" was selected as the search topic from 2000 to 2020 from the NCBI database.[1] Each bibliographic record includes detailed information of published articles, including their titles, abstracts, and keywords. **Figure 1A** shows the number of published articles on the significant increase in LAI from 2012. Since 2000, when Chapman and Thompson (2001) published *Linkage Disequilibrium Mapping: The Role of Population History, Size, and Structure*, 186 articles have been published until 2020. The major topics in LAI research are shown in **Figure 1B**. The visual representation, known as a form tree, was generated using the clustering tool Carrot II (Cost et al., 2002) based on 40 clusters. The leading topics of research are disease association and human history. We analyzed the main contents of the cited articles for each model in **Figure 1C**, which illustrates that research on human history plays a leading role in LAI analysis and model development. Similarly, LAI research is also largely applied in disease risk, wildlife conservation, and domestication. **Figure 1D** shows four original types of research and seven model designs with top citations, which may play a driving role in the research of LAI. During 2006–2016, LAI research had been highly fascinating for various research groups; thus, LAI publications experienced a peak period. This research has gently and extensively infiltrated different fields of science and has kept on moving in the following years (Lao et al., 2006; Sankararaman et al., 2008b; Price et al., 2009; Bryc et al., 2010, 2015; Gravel, 2012; Lawson et al., 2012; Eaton and Ree, 2013; Loh et al., 2013; Maples et al., 2013; Moreno-Estrada et al., 2013; Jeong et al., 2014). To benchmark the computational efficiency and accuracy of the seven most used models (Chromopainter, LAMP, LAMP-LD, Loter, RFMix, Seqmix, and Supportmix), we simulated data using SLiM 3.2 (Messer, 2013) and estimated the average running time (ART), memory footprint size (MFZ), the mean squared error (MSE $= \frac{1}{n}\sum_{i=1}^{n}(\text{observed}_i - \text{predicted}_i)$) for an individual genome, standard deviation (SD), and the coefficient of variation (CV) for each model. In the SLiM one, we initially generated two ancestor populations during 5,000 generations. The use of two initial populations differentiates into five admixed subpopulations with different infiltration rates after 4,000 generations. During the next step, differentiated individuals evolve freely during 5–1,000 generations, and every five generation is an interval. This step is repeated 20 times. Finally, we randomly selected 1,000 ancestral populations and 500 admixed populations to stimulate LAI in seven models. **Table 2** shows further details regarding the simulation parameters and other simulation processes.

As shown in **Table 3**, we adopted seven models in SLiM 1–3 and six models in SLiM 4–5 because Seqmix can only handle two ancestral groups. The most efficient model is LAMP

[1] https://pubmed.ncbi.nlm.nih.gov

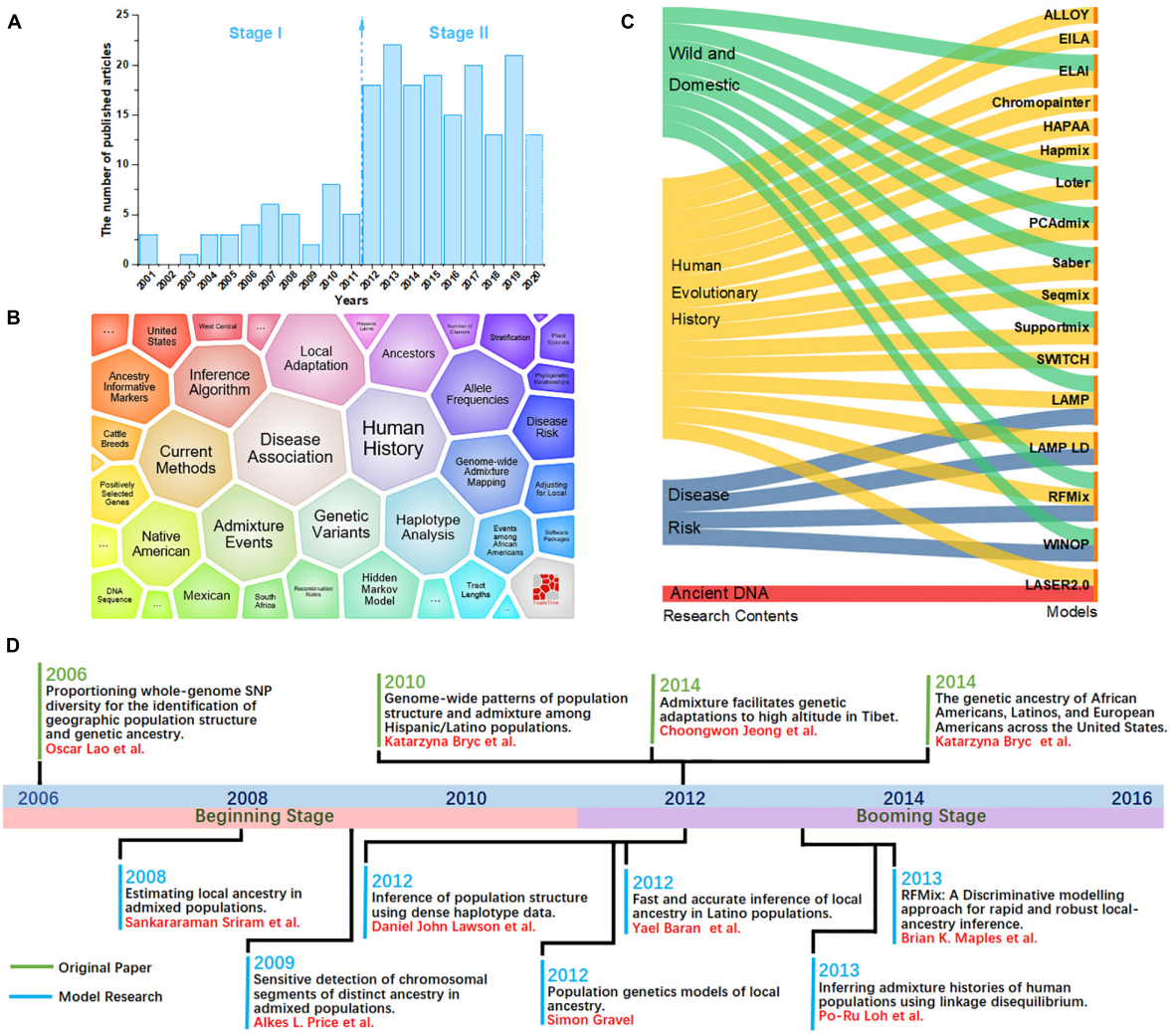

FIGURE 1 | (A) Number of local ancestry inference (LAI) publications from 2000 to 2020. **(B)** A visual survey of the major topics on LAI by the Carrot II system. **(C)** Distribution of models across the top four research fields including wild domestication, history of human evolution, disease risk, and ancient DNA. **(D)** Timeline illustrating the development of LAI.

with respect to the run time (ART = 1.50 s) and memory size (MFZ = 53.74 Mb); however, its accuracy is slightly lower (1 – mean of MSE = 0.67) and the results are not stable (SD = 0.20). The primary reason is the total reliance of this model on biological parameters. Seqmix based on aligned bases turns out to be the most accurate (1 – mean of MSE = 0.86) and stable (SD = 0.08) model, while it is also efficient enough. Loter is the only model with a parameter-free process and general accuracy (1 – mean of MSE = 0.79) and fair stability (SD = 0.10); however, it requires a comparatively longer running time (ART = 2,506.70 s). The RFMix process has general accuracy (1 – mean of MSE = 0.80) and fair stability (SD = 0.10), but it consumes a lot of memory (MFZ = 2,472.29 Mb). A weighing between the pros and cons of the different models is shown in **Table 4**.

As shown in **Figure 2**, the phased data had a higher accuracy in comparison to the unphased data. Besides, there exists a significant difference between the phase and unphased results (1 – mean of MSE) in all the simulated values by each paired comparison in Tukey's HSD (all $P < 0.05$). As shown in **Table 3**, the CV of the phased results is less than that of the unphased results in all simulated values, thus proving the higher stability of phased data.

DISCUSSION

Current Situation and Existing Problems

Various challenges confront the researchers during inferring the local ancestry *via* genome-wide data. Firstly, several models need complex parameters, such as a genetic map and the number of generations since admixture, that are difficult to be supplied, particularly for non-model species. Secondly, some models only use haplotype information and unlinked markers are removed

TABLE 2 | Details for generating slim data.

	Slim1	Slim2	Slim3	Slim4	Slim5
Mutation rate	1e–8	1e–8	1e–8	1e–8	1e–8
Recombination rate	1e–6	1e–6	1e–6	1e–6	1e–6
Chromosome length	1e+6	1e+6	1e+6	1e+6	1e+6
Initial effective population size	2000	2000	2000	2000	2000
The number of generations producing true populations	5000	5000	5000	5000	5000
The number of ancestral populations	2	2	2	3	3
Effective ancestral population size	2000	2000	2000	2000	2000
Divergence time of ancestral population	4000	2000	200	2000	200
The number of admix populations	5	5	5	6	6
The number of selected ancestor individuals	1000	1000	1000	1000	1000
The number of selected admix individuals	500	500	500	500	500
Repetition times	20	20	20	20	20
Infiltration rate		AdmP1 = 0.1\0.9 AdmP2 = 0.2\0.8 AdmP3 = 0.3\0.7 AdmP4 = 0.4\0.6 AdmP5 = 0.5\0.5		AdmP1 = 0.1\0.1\0.8 AdmP2 = 0.2\0.1\0.7 AdmP3 = 0.3\0.1\0.6 AdmP4 = 0.4\0.1\0.5 AdmP5 = 0.5\0.2\0.1 AdmP6 = 0.6\0.2\0.2	
Generation number		5∼1000 (the interval is 5 generations)			

TABLE 3 | Benchmark analysis of most used LAI models.

Model	ART/s	MFZ/Mb	1–Mean of MSE	SD	CV
Chromopainter	2243.56	309.20	0.73	0.18	0.24
LAMP	1.50	53.74	0.67	0.20	0.30
LAMP-LD	97.95	129.40	0.60	0.18	0.30
Loter	2506.70	269.84	0.79	0.10	0.13
RFMix	166.74	2472.29	0.80	0.10	0.13
Seqmix	31.11	1201.27	0.86	0.08	0.09
Supportmix	753.01	130.94	0.80	0.12	0.15

via the trimming step. With this process, many informative SNPs are lost. Thirdly, because some models exclude probable ancestral informative haplotypes, unmodeled LD could cause systematic biases in determining ancestry, which results in false-positive conclusions regarding the deviation in ancestry at specific loci. Lastly, ancestral segments are windows or blocks of varying lengths; however, existing models commonly use a window of fixed size for simplification. The total count of generations since admixture is inversely proportional to the length of ancestral segments. As the number of generations is hardly recognized, it is difficult to investigate the breakpoint or transition point for ancestral haplotypes based on the statistics of the ancestral group or even an individual's genome.

Model-Based Recommendation

We summarize these models with their distinct advantages and disadvantages as follows: (i) We recommend Seqmix if the genotype is hard to call, and aligned bases can be used directly in this model (Hu et al., 2013). (ii) ALLOY utilizes F-HMM and the haplotype structure of the compound state to improve its accuracy. We recommend this model if ancient and complex admixtures need to be analyzed (Rodriguez et al., 2013). (iii) We recommend Saber if high-density SNP panels exist; however, a potential weakness of M-HMM, compared with an HMM, is that when the genetic information on the ancestral populations is not rich, it will weaken the accuracy of the calculations (Tang et al., 2006). (iv) ELAI is appropriate for instances where researchers require detecting further structure of the haplotypes because of the two scales of LD in admixture and a two-layer HMM exists as independent upper-layer latent clusters that enforce structure on the haplotypes and other lower-layer latent clusters depicting ancestral haplotypes (Guan, 2014). (v) We recommend EILA if the researchers are interested in the estimation of recombination events. The model has the advantage of allowing the lack of ancestral populations' high-quality haplotype information; however, a potential weakness of the *k*-means, unsupervised clustering, will weaken the stability of calculations (Yang et al., 2013). (vi) Loter uses DP to obtain a globally optimal solution, and its advantage is its being parameter-free (Dias-Alves et al., 2018).

Integration With Other Methods

LAI incorporates other bioinformatics approaches and is widely used in different research fields, including breeding new varieties, protection of endangered animals and plants, and the prevention and treatment of human genetic diseases. In the study of population structure, the ADMIXTURE (Alexander et al., 2009) and STRUCTURE (Pritchard et al., 2000) models perform population allele frequencies and observe genotype probability by ancestry proportions. Both models can be used to assign global ancestry. They are applied in fine-matched corrected association research and are relatively consistent with the LAI results. Galaverni et al. better estimated the actual admixture proportions of the hybrids according to the combination of global and local ancestry inferences (Galaverni et al., 2017). About up to 50% of blocks of domesticated individuals were identified by

TABLE 4 | Weighing of most used LAI models.

Model	Advantage	Disadvantage
Chromopainter	Moderate memory consumption and certain accuracy	Slower processing speed and unstable analysis results
LAMP	Fast processing speed and low memory consumption	Low accuracy and unstable analysis results
LAMP-LD	Fast processing speed and moderate memory consumption	Low accuracy and unstable analysis results
Loter	Moderate memory consumption, certain accuracy and stable analysis results	Slower processing speed
RFMix	Moderate processing speed, high accuracy and stable analysis results	High memory consumption
Seqmix	Fast processing speed, high accuracy and stable analysis results	High memory consumption
Supportmix	Moderate processing speed and memory consumption, high accuracy and certain stibility	–

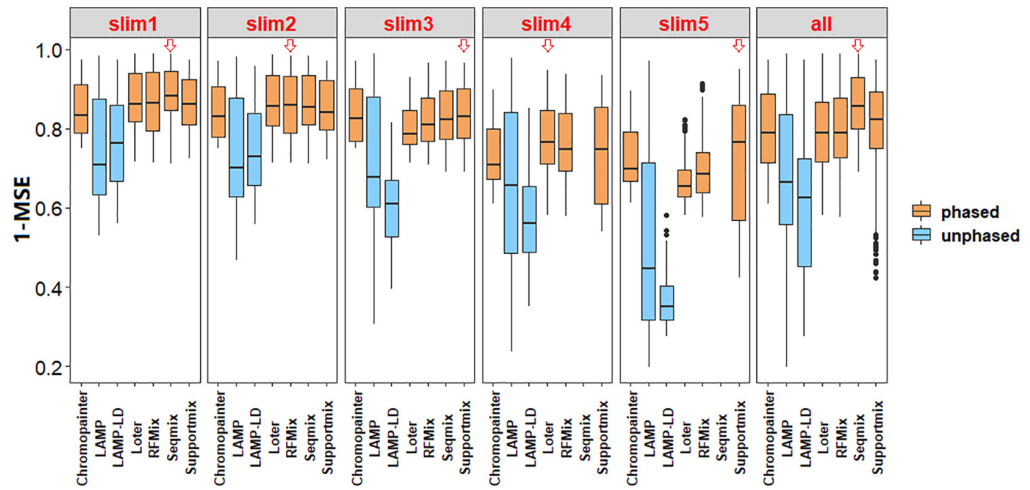

FIGURE 2 | Box plots of the accuracy of local ancestry inference (LAI) using a benchmark. The *red hollow arrows* indicate a higher accuracy by the median comparison in this simulation. The results showed that phased data had higher accuracy in comparison with unphased data.

PCADMIX in the hybrid genome. The results of the analysis were consistent with those estimated in ADMIXTURE at $K = 2$. In the study of domestication, the admixture compositions of select individuals with the minor allele for the peak markers of quantitative trait loci (QTL) were analyzed by LAI. For example, in one study, QTL were located in a chromosome segment substitution line (CSSL) population. This population comes from an interspecific cross between a wild aus-like *Oryza rufipogon* donor accession and cv. *Curinga* (an upland tropical japonica variety from Brazil). It was found that the CSSLs conferred a wild aus-like introgression across the target segment, which was beyond the rest of the CSSLs that carried the tropical japonica genotype (Wang et al., 2017). In the study of ancient DNA, the use of LAI and masking reconstruct population-specific surrogates of the ancestral components to yield entire genome. Yelmen et al. applied this technique to reconstruct population-specific surrogates of South Asian and West Eurasian populations, which complemented low-quantity and low-coverage availability and provided a substantial advantage (Yelmen et al., 2019).

Application and Development

Wild populations significantly contribute to the adaptation of domesticated populations; therefore, their absence or presence is imperative for breeding and genetics-related studies. Many good traits exist in the wild population; however, they were lost during domestication. Some advantageous or disadvantageous alleles were located by constructing a hybrid population and were further assigned the corresponding ancestral source. This can help in understanding the molecular mechanisms behind the traits and in explaining the valuable pool of genetic resources found in wild populations. Domesticated rice (*Oryza sativa*) is adopted as an example. Some traits of wild rice (such as persistent seed dormancy and freely shattering seed) may have high adaptability if introgressed into weedy rice populations. Inversely, some traits of wild rice (prostrate plant architecture and sporadic seed production) are considered inappropriate for survival in domesticated rice. Given the potential combination of the advantageous and disadvantageous traits for weedy rice, it can be expected that introgression evidence of wild rice to weed rice would confer weed rice-adaptive traits to the specific genomic regions. Such as some regions were likely introgressed from wild accessions: *PROG1*, controlling prostrate *versus* erect growth; *qSW5*, controlling seed size; *sh4*, controlling grain shattering; *Bh4*, controlling hull color; *An-1*, controlling awn development; and *Rc*, controlling pericarp pigmentation (Vigueira et al., 2019). In another study, the analysis of wild caprids and whole genomes of domestic goats revealed ancient introgression evidence from a West Caucasian tur-like population to the

ancestor of domestic goats. It was further revealed that the *MUC6* gene was an introgression locus with a strong selection signature and conferred enhanced immune resistance to gastrointestinal pathogens (Zheng et al., 2020). The third case is the wild yeast (*Saccharomyces eubayanus*). The lager-style beers are an interspecies hybrid (*S. eubayanus* × *Saccharomyces cerevisiae)*. It was found that the wild isolates of *S. eubayanus* are not the closest relatives of lager-brewing hybrids. Inversely, the genetic composition of lager yeasts was contributed by *S. eubayanus* strains with continuous variation, thus revealing the complex ancestries of lager yeasts (David et al., 2016). The LAI model can be a powerful tool for protecting wild species by identifying segments of the genomes of hybrids. In the research of Galaverni et al., domestic dogs (*Canis lupus familiaris*) can reproduce with wild wolves (*Canis lupus*), coyotes (*Canis latrans*), and golden jackals (*Canis aureus*). The gene pool of several wild canid populations were threatened by the widespread diffusion of stray dogs in human-dominated areas. Use of the LAI model and genotype–phenotype association procedures identified putative dog-derived causal mutations associated with phenotypic variants, thereby constituting a conservation strategy. Such as the black coat color, this trait is coded by a 3-bp deletion at the β-defensin gene *CDB103* that was possibly introduced into wolves by ancient hybridization with dogs (Galaverni et al., 2017).

The LAI model can be applied to the treatment and prevention of human genetic diseases by assigning ancestry to the chromosomal regions and applying admixture mapping to identify candidate genes. Dengue has become a worldwide health concern due to the increase in virus and vector dispersions. LAI analysis has proven that African ancestry has a protective effect against the dengue haemorrhagic phenotype in admixed Cuban population. This was further authenticated by identifying the corresponding candidate genes (Sierra et al., 2017). A similar study indicates that the Tibetans have a better altitude adaptation, on account of the introgression of associated haplotypes from Denisovans or Denisovan-related populations (Huerta-Sánchez et al., 2014). Besides, a recent example is that about 3,000 coronavirus disease 2019 (COVID-19) patients and control individuals were adopted, and it was found that a gene cluster can cause severe symptoms after SARS-CoV-2 infection. This genetic risk factor was caused by a genomic segment of a size of about 50 kb inherited from Neanderthals (Zeberg and Pääbo, 2020). Furthermore, this genomic segment was carried by about 50% South Asian and about 16% European people. In conclusion, these studies not only enhance our understanding of genetic diversity and natural history but also offer valuable evidence for the source of diversity among human beings, animals, plants, and model organisms.

AUTHOR CONTRIBUTIONS

JW and YZ wrote the paper. YL organized and designed the benchmark. YZ supervised the study and revised the manuscript. All authors have read and commented on the manuscript and approved the final version.

FUNDING

The project was supported by the National Natural Science Foundation of China (U1704233) and the Key-Area Research and Development Program of Guangdong Province (2018B020203001).

ACKNOWLEDGMENTS

We thank the support of the high-performance computing platform of the State Key Laboratory of Agrobiotechnology.

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28, 1359–1367. doi: 10.1093/bioinformatics/bts144

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., et al. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.* 84, 343–364. doi: 10.3378/027.084.0401

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. doi: 10.1016/j.ajhg.2014.11.010

Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., et al. (2010). Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc. Natl. Acad. Sci. U.S.A.* 107, 8954–8961. doi: 10.1073/pnas.0914618107

Chapman, N. H., and Thompson, E. A. (2001). Linkage disequilibrium mapping: the role of population history, size, and structure. *Adv. Genet.* 42, 413–437. doi: 10.1016/s0065-2660(01)42034-7

Cost, R. S., Kallurkar, S., Majithia, H., Nicholas, C., and Shi, Y. (2002). "Integrating distributed information sources with CARROT II, 194-201," in *Cooperative Information Agents VI. CIA 2002. Lecture Notes in Computer Science*, Vol. 2446, eds M. Klusch, S. Ossowski, and O. Shehory (Berlin: Springer), doi: 10.1007/3-540-45741-0_17

David, P., Langdon, Q. K., Moriarty, R. V., Kayla, S., Martin, B., Guillaume, C., et al. (2016). Complex ancestries of lager-brewing hybrids were shaped by standing variation in the wild yeast saccharomyces eubayanus. *PLoS Genet.* 12:e1006155. doi: 10.1371/journal.pgen.1006155

Dias-Alves, T., Mairal, J., and Blum, M. G. B. (2018). Loter: a software package to infer local ancestry for a wide range of species. *Mol. Biol. Evol.* 35, 2318–2326. doi: 10.1093/molbev/msy126

Dougherty, M. L., Nuttle, X., Penn, O., Nelson, B. J., Huddleston, J., Baker, C., et al. (2017). The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Geno. Biol.* 18:49. doi: 10.1186/s13059-017-1163-9

Eaton, D. A. R., and Ree, R. H. (2013). Inferring phylogeny and introgression using RADseq data: an example from flowering plants (pedicularis: orobanchaceae). *Syst. Biol.* 62, 689–706. doi: 10.1093/sysbio/syt032

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587.

Fitak, R. R., Rinkevich, S. E., and Culver, M. (2018). Genome-wide analysis of SNPs is consistent with no domestic dog ancestry in the endangered mexican wolf (*Canis lupus baileyi*). *J. Heredity* 109, 372–383. doi: 10.1093/jhered/esy009

Galaverni, M., Caniglia, R., Pagani, L., Fabbri, E., Boattini, A., and Randi, E. (2017). Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing wolf population. *Mol. Biol. Evol.* 34, 2324–2339. doi: 10.1093/molbev/msx169

Geza, E., Mugo, J., Mulder, N. J., Wonkam, A., Chimusa, E. R., and Mazandu, G. K. (2019). A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. *Brief. Bioinform.* 20, 1709–1724. doi: 10.1093/bib/bby044

Gilks, W. R. (1999). *Markov Chain Monte Carlo*. New York, NY: Springer.

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics* 191, 607–619. doi: 10.1534/genetics.112.139808

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics* 196, 625–642. doi: 10.1534/genetics.113.160697

Haasl, R. J., McCarty, C. A., and Payseur, B. A. (2013). Genetic ancestry inference using support vector machines, and the active emergence of a unique American population. *Eur. J. Hum. Genet.* 21, 554–562. doi: 10.1038/ejhg.2012.258

Hu, Y., Willer, C., Zhan, X., Kang, H. M., and Abecasis, G. R. (2013). Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. *Am. J. Hum. Genet.* 93, 891–899. doi: 10.1016/j.ajhg.2013.10.008

Huerta-Sánchez, E., Jin, X., Asan, B. Z., Peter, B. M., Vinckenbosch, N., et al. (2014). Altitude adaptation in tibetans caused by introgression of denisovan-like DNA. *Nature* 512, 194–197. doi: 10.1038/nature13408

Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D. B., Pritchard, J. K., et al. (2014). Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat. Commun.* 5:3281. doi: 10.1038/ncomms4281

Kidd, J. M., Gravel, S., Byrnes, J., Moreno-Estrada, A., Musharoff, S., Bryc, K., et al. (2012). Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *Am. J. Hum. Genet.* 91, 660–671. doi: 10.1016/j.ajhg.2012.08.025

Lao, O., van Duijn, K., Kersbergen, P., de Knijff, P., and Kayser, M. (2006). Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am. J. Hum. Genet.* 78, 680–690. doi: 10.1086/501531

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet.* 8:e1002453. doi: 10.1371/journal.pgen.1002453

Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., et al. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193, 1233–1254. doi: 10.1534/genetics.112.147330

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. doi: 10.1016/j.ajhg.2013.06.020

Messer, P. W. (2013). SLiM: simulating evolution with selection and linkage. *Genetics* 194, 1037–1039.

Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J. L., Byrnes, J. K., Gignoux, C. R., et al. (2013). Reconstructing the population genetic history of the caribbean. *PLoS Genet.* 9:e1003925. doi: 10.1371/journal.pgen.1003925

Omberg, L., Salit, J., Hackett, N., Fuller, J., Matthew, R., Chouchane, L., et al. (2012). Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC Genet.* 13:10. doi: 10.1186/1471-2156-13-49

Padhukasahasram, B. (2014). Inferring ancestry from population genomic data and its applications. *Front. Genet.* 5:204. doi: 10.3389/fgene.2014.00204

Pasaniuc, B., Sankararaman, S., Kimmel, G., and Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics* 25, i213–i222.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., et al. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* 5:e1000519. doi: 10.1371/journal.pgen.1000519

Pritchard, J. K., Stephens, and Matthew. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Rodriguez, J. M., Bercovici, S., Elmore, M., and Batzoglou, S. (2013). Ancestry inference in complex admixtures via variable-length markov chain linkage models. *J. Comput. Biol.* 20, 199–211. doi: 10.1089/cmb.2012.0088

Sankararaman, S., Kimmel, G., Halperin, E., and Jordan, M. I. (2008a). On the inference of ancestries in admixed populations. *Genome Res.* 18, 668–675. doi: 10.1101/gr.072751.107

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008b). Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.* 82, 290–303. doi: 10.1016/j.ajhg.2007.09.022

Schumer, M., Powell, D. L., and Corbett-Detig, R. (2020). Versatile simulations of admixture and accurate local ancestry inference with mixnmatch and ancestryinfer. *Mol. Ecol. Res.* 20, 1141–1151. doi: 10.1111/1755-0998.13175

Schuster-Böckler, B., and Bateman, A. (2007). An introduction to hidden Markov models. *Int. J. Pattern Recog. Artif. Int.* 15, 9–42.

Sierra, B., Triska, P., Soares, P., Garcia, G., and Guzman, M. G. (2017). OSBPL10, RXRA and lipid metabolism confer African-ancestry protection against dengue haemorrhagic fever in admixed CUBANS. *PLoS Pathogens* 13:e1006220. doi: 10.1371/journal.ppat.1006220

Sundquist, A., Fratkin, E., Do, C. B., and Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Geno. Res.* 18, 676–682. doi: 10.1101/gr.072850.107

Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* 79, 1–12. doi: 10.1086/504302

Vigueira, C. C., Qi, X., Song, B. K., Li, L. F., Caicedo, A. L., Jia, Y., et al. (2019). Call of the wild rice: Oryza rufipogon shapes weedy rice evolution in Southeast Asia. *Evol. Appl.* 12, 93–104. doi: 10.1111/eva.12581

Wang, C., Zhan, X., Liang, L., Abecasis, G. R., and Lin, X. (2015). Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *Am. J. Hum. Genet.* 96, 926–937. doi: 10.1016/j.ajhg.2015.04.018

Wang, D. R., Han, R., Wolfrum, E. J., and Mccouch, S. R. (2017). The buffering capacity of stems: genetic architecture of nonstructural carbohydrates in cultivated Asian rice, Oryza sativa. *New Phytol.* 215, 658–671. doi: 10.1111/nph.14614

Wu, J., and Zhao, Y. (2019). Machine learning technology in the application of genome analysis: a systematic review. *Gene* 705, 149–156. doi: 10.1016/j.gene.2019.04.062

Yang, J. J., Li, J., Buu, A., and Williams, L. K. (2013). Efficient inference of local ancestry. *Bioinformatics* 29, 2750–2756. doi: 10.1093/bioinformatics/btt488

Yelmen, B., Mondal, M., Marnetto, D., Pathak, A. K., Montinaro, F., Gallego Romero, I., et al. (2019). Ancestry-specific analyses reveal differential demographic histories and opposite selective pressures in modern south asian populations. *Mol. Biol. Evol.* 36, 1628–1642. doi: 10.1093/molbev/msz037

Zeberg, H., and Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* 587, 610–612. doi: 10.1038/s41586-020-2818-3

Zheng, Z., Wang, X., Li, M., Li, Y., Yang, Z., Wang, X., et al. (2020). The origin of domestication genes in goats. *Sci. Adv.* 6:eaaz5216. doi: 10.1126/sciadv.aaz5216

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.