A comprehensive survey of models for dissecting local ancestry deconvolution in human genome

Ephifania Geza, Jacqueline Mugo, Nicola J. Mulder, Ambroise Wonkam, Emile R. Chimusa and Gaston K. Mazandu

Corresponding authors: Gaston K. Mazandu, African Institute for Mathematical Sciences, Muizenberg, Cape Town 7945, South Africa. E-mail: kuzamunu@aims.ac.za; Emile R. Chimusa, Division of Human Genetics, Department of Pathology, University of Cape Town, Cape Town 7925, South Africa. E-mail: emile.chimusa@uct.ac.za

Abstract

Over the past decade, studies of admixed populations have increasingly gained interest in both medical and population genetics. These studies have so far shed light on the patterns of genetic variation throughout modern human evolution and have improved our understanding of the demographics and adaptive processes of human populations. To date, there exist about 20 methods or tools to deconvolve local ancestry. These methods have merits and drawbacks in estimating local ancestry in multiway admixed populations. In this article, we survey existing ancestry deconvolution methods, with special emphasis on multiway admixture, and compare these methods based on simulation results reported by different studies, computational approaches used, including mathematical and statistical models, and biological challenges related to each method. This should orient users on the choice of an appropriate method or tool for given population admixture characteristics and update researchers on current advances, challenges and opportunities behind existing ancestry deconvolution methods.

Key words: local ancestry; linkage disequilibrium; admixture; hidden Markov models; ancestry informative markers

Introduction

Recent advances in high-throughput technologies have led to the availability of enormous amounts of publicly available genomic data. These data require appropriate techniques for informed decisions, which are useful in both medical and population genetics. This helps in understanding population history, demography, disease aetiology and the way different individuals respond to drugs. Genetic diversities observed in the human deoxyribonucleic acid (DNA) sequences result from inheritance processes, including mutation and recombination,

and from population migration [1]. Owing to population migration, most modern humans are admixed [2, 3], as a result of the mating between two or more previously isolated populations. This interbreeding yields genetic recombination break points and the formation of variation and mixed DNA segments. Admixed genomes are a mosaic of segments originating from different ancestral populations [4–9]. There is a great interest in understanding the dynamics related to the origin of these variations, the evolutionary process and its consequences in human health. Studying admixed populations may

Ephifania Geza, Masters degree in Mathematical Sciences. She is currently a PhD candidate in Bioinformatics at the Computational Biology Division at University of Cape Town in collaboration with the African Institute for Mathematical Sciences (AIMS).

Jacqueline Mugo, Masters degree in Mathematical Sciences. She is currently a PhD candidate in Bioinformatics at the Computational Biology Division at University of Cape Town.

Nicola J. Mulder, PhD in Medical Microbiology. She is a Professor and Head of the Computational Biology Division at University of Cape Town.

Ambroise Wonkam, MD, DMedSc, PhD degree. He is a Professor at the Division of Human Genetics, -Department of Pathology and senior specialist medical geneticist, HPCSA MP0686980, University of Cape Town.

Emile R. Chimusa, PhD in Bioinformatics. He is currently a Senior Lecturer at the Division of Human Genetics, -Department of Pathology, University of Cape Town.

Gaston K. Mazandu, PhD in Bioinformatics. He is an honorary senior member of the Computational Biology Division at UCT and an associate researcher at AIMS, and holds a senior lecturer position at the Division of Human Genetics, Department of Pathology, University of Cape Town.

Submitted: 11 March 2018; **Received (in revised form):** 16 April 2018

© The Author(s) 2018. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

improve our understanding of the genetic structure of diseases [10–13].

Today, advances in high-throughput sequencing and genotyping technologies, combined with the appropriate computational and statistical methods, have made feasible the inference of the admixture history of human populations. As a result, several local ancestry inference or ancestry deconvolution methods have been implemented. Since 2003, >20 ancestry deconvolution methods have been introduced, some of which have similar statistical models. Here we have categorized them into two main groups based on whether they account for linkage disequilibrium (LD). LD-based models include STRUCTURE [14], SABER [15], HAPMIX [4] and ALLOY [16], and some of non-LD-based models are LAMP [17], WINPOP [18] and LOTER [19]. These local ancestry inference approaches have been used in several applications, mostly in admixed populations from the American continent, including understanding signatures of natural selection, ethnic differences in drug response and the risk of complex diseases. It is worth noting that these methods have merits and drawbacks in inferring local ancestry estimates in multiway admixed populations within a complex or multifaceted admixture model.

Some review studies on ancestry inference exist, but they are limited in describing existing ancestry inference-associated tools, their relationship with one another and applications [20, 21]. Unlike the reviews of Liu et al. [20] and Padhukasahasram [21], in 2013, Gompert and Buerkle [22] provided a mathematical structure of some existing methods, highlighting their similarities and differences. Here, we systematically survey all the ancestry deconvolution methods/tools we are aware of. We dissect approaches for pinpointing local ancestry along the genome of admixed populations and summarize the evolutionary factors they account for. Importantly, this survey explores different statistical model properties, assumptions and results from existing evaluations. This may orient the development of new methods and enable researchers to choose appropriate tools for given population characteristics, depending on where the local ancestry estimates should be applied. Furthermore, this study will enable software developers to identify current challenges and advances in local ancestry deconvolution, providing them with appropriate information for implementing practical and integrative tools, which consider current medical and population genetics demands. It also provides a reference manual on ancestry deconvolution models and on how biologically practical these models are, depending on the use of the inferred local ancestry estimates.

Importance and application of local ancestry deconvolution

Admixture introduces a type of LD, known as ‘admixture LD’ (Box 2). The distribution of admixture at the genetic level and/or along the genome of an admixed individual can exhibit a pattern of single nucleotide polymorphisms (SNPs), which may have either medical or evolutionary implications. Local ancestry deconvolution in human populations has contributed to a wide range of biomedical applications from identifying both local

selection and genetic variants underlying ethnic difference in disease risk and drug/treatment responses to an understanding of population genetic history [23, 24]. These studies have so far shed light on the patterns of genetic variation throughout modern human evolution and have improved our understanding of the demographics and adaptive processes of human populations. As the world has become ‘smaller’, with travel making migration easier, many global populations are likely to become admixed [3]. This is likely to increase the complexity of the population admixture dynamics, resulting in multifaceted admixture events. There is a high chance of the SNPs of admixed populations contributing significantly to the phenotypic variability and differences in drug response therapies, which may not be the case in homogeneous populations (Box 2) [25]. As a result, genetic and clinical studies of admixed populations are complicated by the degree of admixture/traits association [25]. Accordingly, understanding the structure of human genetic admixture is not only of anthropological relevance but also a medical necessity; thus, knowledge of local ancestry estimates is of utmost importance. Local ancestry deconvolution is also of great relevance in gene–environment interactions. For example, local ancestry inference may be a key in pharmacogenetics and personalizing medicine [26, 27], considering potential interactions between genes and a drug (or environment). Moreover, local ancestry estimates are also crucial in gene–gene interactions [28], when trying to understand how multiple genes interact to confer a complex trait or disease. This suggests that local ancestry estimates may also be important in understanding signatures of natural selection and complex diseases through admixture mapping as an alternative approach to the popular genome-wide association studies, particularly in admixed populations.

There are several applications of local ancestry deconvolution in understanding the demographics and admixture structure of various admixed populations. However, as indicated in Table 2, there exist few applications relevant to disease risk and clinical prediction of treatment. These applications are mostly based on admixed populations from the American continent. Table 2 provides more detail on these studies that applied local ancestry estimates, the reference paper, the admixed populations under study, the number of samples used and markers studied and the results obtained. This table suggests that local ancestry inference has been applied in identifying selection signatures, by examining the distribution of genome-wide ancestry in admixed populations [21] or by comparing the local ancestry estimates in cases to local ancestry estimates in control samples or local ancestry in cases at a site compared with the average genome-wide ancestry.

Overview of local ancestry deconvolution

Owing to their popularity in local ancestry inference models and to guide readers, we illustrate the hidden Markov model (HMM)-based local ancestry problem, choosing a general local ancestry concept, which, under a null model of two-way admixture and simplifying assumption, considers an admixed individual, i , with genotypes coded based on the copy of minor allele,

$Y_{it} \in \{0, 1, 2\}$. Unless otherwise stated, throughout this article, we denote:

$k = 1, 2, \dots, K$	the population index for ancestral population, assumed to be previously isolated and mixed g , generations ago
$i = 1, \dots, n_k$	the individual index for ancestral population k , with n_k the number of individuals in ancestral population k , meaning that the total number, n , of ancestral individuals is $n = \sum_{k=1}^K n_k$,
$t = 1, 2, \dots, T$	the site or SNP index
$h \in \{0, 1\}$	the haplotype
π_k	the proportion of ancestry inherited from population k , such that $\sum_{k=1}^K \pi_k = 1$,
π	the vector of ancestry inherited from each ancestral population
Π	the $n \times K$ matrix of proportions of each individual i inherited from population k ,
p_t	allele frequency at a site t and P the matrix of ancestral allele frequencies
X_{it}	the local ancestry at site t , for individual i , copied from k ,
Y_t	the observed admixed genotype at t ,
r	the recombination rate
d_t	the genetic distance between t and $t - 1$, and
$\sigma(k = k')$	the indicator function, with

$$\sigma(k = k') = \begin{cases} 1 & \text{if } k = k' \\ 0 & \text{otherwise} \end{cases} \text{ and } \sigma(k \neq k') = 1 - \sigma(k = k')$$

For window-based methods, a window is indexed by w , and the distance between the centre of the two consecutive windows is d_w .

The fraction of the individual's genome-wide ancestry of ancestral population A (π_A) and the average number of generations since admixture (g) occurred are assumed to be known. Let X_t be the unobserved state 0, 1 or 2, the number of allele copies from one of the ancestral chromosomes of this individual at marker t along the genome. An HMM family-based model is completely defined based on three main parameters, namely, initial, transition and emission probabilities. Thus, we set the left end of chromosome initial probabilities as $P(X_0 = 0) = (1 - \pi_A)^2$, $P(X_0 = 1) = 2\pi_A(1 - \pi_A)$ and $P(X_0 = 2) = \pi_A^2$. Let d be the genetic distance (in Morgans) between markers t and $t + 1$, and the transition probabilities are as follows:

1. Switching probability while in State 0
 $P(X_{t+1} = 0 | X_t = 0) = e^{-2gd} + 2e^{-gd}(1 - e^{-gd})(1 - \pi_A) + (1 - e^{-gd})^2(1 - \pi_A)^2$
 $P(X_{t+1} = 1 | X_t = 0) = 2e^{-gd}(1 - e^{-gd})\pi_A + (1 - e^{-gd})^2 2\pi_A(1 - \pi_A)$
 $P(X_{t+1} = 2 | X_t = 0) = (1 - e^{-gd})^2 \pi_A^2$
2. Switching probabilities while in State 1
 $P(X_{t+1} = 0 | X_t = 1) = 2e^{-gd}(1 - e^{-gd})(1 - \pi_A) + (1 - e^{-gd})^2(1 - \pi_A)^2$
 $P(X_{t+1} = 1 | X_t = 1) = e^{-2gd} + e^{-gd}(1 - e^{-gd}) + (1 - e^{-gd})^2 2\pi_A(1 - \pi_A)$
 $P(X_{t+1} = 2 | X_t = 1) = e^{-gd}(1 - e^{-gd})\pi_A + (1 - e^{-gd})^2 \pi_A^2$
3. Switching probabilities while in State 2
 $P(X_{t+1} = 0 | X_t = 2) = (1 - e^{-gd})^2(1 - \pi_A)^2$
 $P(X_{t+1} = 1 | X_t = 2) = 2e^{-gd}(1 - e^{-gd})(1 - \pi_A) + (1 - e^{-gd})^2 2\pi_A(1 - \pi_A)$
 $P(X_{t+1} = 2 | X_t = 2) = e^{-2gd} + 2e^{-gd}(1 - e^{-gd})\pi_A + (1 - e^{-gd})^2 \pi_A^2$

Let p_A and p_B be the genotype frequencies of marker j in proxy ancestral Populations A and B, respectively, thus for $k \in \{A, B\}$, and $u \in \{0, 2\}$, the likelihoods (the emission probabilities) that the ancestral population k has either 0 (state 0) or both copies of alleles (State 2) are $P(Y_{it} = 0 | X_t = u) = (1 - p_k)^2$, $P(Y_{it} = 1 | X_t = u) = 2p_k(1 - p_k)$ and $P(Y_{it} = 2 | X_t = u) = p_k^2$. It follows that the emission probabilities that both ancestral populations A and B share copy alleles are $P(Y_{it} = 0 | X_t = 1) = (1 - p_A)(1 - p_B)$, $P(Y_{it} = 1 | X_t = 1) = p_A(1 - p_B) + p_B(1 - p_A)$ and $P(Y_{it} = 2 | X_t = 1) = p_A p_B$. The probability of interest is the inference of posterior probabilities $P(X_t | Y_{it})$ and the integration over uncertainty in, π_A , g , p_A and p_B .

In general, HMM family-based models in the context of local ancestry address one or all of the three problems, i.e. (1) determining the probability of observing a particular sequence given HMM parameters, (2) determining the most probable hidden state sequence given the observed sequence and (3) finding the HMM parameters that maximize the probability of observing a given sequence [29]. Methods to address these three respective problems exist and include the forward, forward-backward or Viterbi and expectation maximization (EM) algorithms. All the extensions of HMM family-based models have the three HMM parameters and most of them make use of the forward-backward or Viterbi and EM algorithms for inference. [Box 1](#) describes the forward-backward, Viterbi and EM algorithms. It is worth noting that few local ancestry deconvolution methods make use of other approaches, such as discriminative or generative models. Discriminative models use the conditional distribution to model the decision boundary, including conditional random fields (CRFs) and support vector machines (SVMs). On the other hand, generative approaches, which use joint probability distribution, are the most common in ancestry deconvolution, as they are easy to fit as compared with discriminative approaches, which often lead to solving a convex programming problem ([Supplementary File](#)).

Dissecting current local ancestry deconvolution approaches

This section provides a detailed theoretical overview of existing local ancestry deconvolution and an exploration of their statistical models. Generally, deconvolving genetic ancestry may yield global ancestry and/or local ancestry from a given proxy ancestry ([Box 2](#)). Modelling the complex nature of the linkage patterns and the choice of HMM play a critical role in yielding better performance in inferring local ancestry. As mentioned earlier, existing local ancestry deconvolution approaches can be subdivided into two main groups based on whether the method models admixture/background LD (refer to [Box 2](#) for description of LD types). [Figure 1](#) shows the evolution of different local ancestry deconvolution methods graphically, indicating the time each method was proposed. Below, we discuss these two categories of local ancestry deconvolution approaches in more detail.

LD-based methods

This category consists of all methods that account for either admixture or background LD or both, and model ancestry along the genome of an admixed individual using HMM family-based models [14, 15]. LD is not strictly important in the inference process; however, removing SNPs in LD reduces the estimation accuracy. Moreover, not modelling LD and using all SNPs (linked and unlinked) may cause noise and systematic biases [30]

because of the complex nature of modelling LD patterns in both an admixed population and its proxy ancestral populations. Although it is a challenge, accurately modelling LD in local ancestry deconvolution may improve the accuracy. On the other hand, inaccurately modelled background LD within ancestral populations may yield spurious results [11]. Ancestry deconvolution models assume that ancestry at different sites or windows follow a Markov chain. Recombination is assumed to occur at every generation resulting in Poisson recombination points with a rate that depends on both the recombination rate, r , and number of generations since admixture, g [6, 22, 30, 31]. Individuals are assumed to be independent of each other; hence, for admixed haploid data, the ancestry at a site, $X_t \in \{1, \dots, K\}$ is modelled by:

$$P(X_1 = k | \pi, r) = \pi_k, \quad (1)$$

at the first marker, and the transition probability between consecutive markers is modelled by:

$$P(X_t = k' | X_{t-1} = k, \pi, r) = \sigma(k = k')e^{-d_t r} + (1 - e^{-d_t r})\pi_{k'}, \quad (2)$$

where $\sigma(k = k')$ and d_t is as defined above. On haploid data, the probability of a recombination event is $1 - e^{-d_t r}$, meaning that the probability of no recombination is $e^{-d_t r}$ [13, 31].

Modelling only admixture LD

These include the early methods that use ancestry informative markers (AIMs) to estimate the ancestry of alleles inherited from a particular population at a chromosomal site. Examples include STRUCTURE [14, 21], which pioneered genetic ancestry inference from 2003, ANCESTRYMAP [31] and ADMIXMAP [32]. ANCESTRYMAP and ADMIXMAP were both proposed in 2004, while STRUCTURE was proposed in 2003 (Figure 1). They are all aimed at improving the application of local ancestry estimates in admixture mapping. They are based on the Bayesian inference framework [22] and assume markers are independent. The ancestry estimation process is memoryless by modelling admixture LD as Markov chains. These early methods do not model background LD, and most of them were designed for two-way admixtures, although some deconvolve ancestry in multiway admixtures, e.g. STRUCTURE. Local ancestry is deconvolved by a standard HMM integrated with Markov chain Monte Carlo (MCMC) to account for uncertainty in model parameters [12, 13, 20]. Admixture generations, average genome-wide ancestry proportions and ancestral allele frequencies are assumed to be known [14, 31, 32], which is not always the case in reality. As they are based on HMMs, early method emission probability models depend on the allele frequency of allele a originating from one of the K unknown ancestries at locus t . Assuming ancestry of origin is X_t and the observed minor allele at t is Y_t , then the emission probability is computed as follows:

$$P(Y_t = z | X_t = k) = \frac{1}{2r_{1k}} \sum_{i=1}^{n_k} \sum_{h=0}^1 \sigma(a_{kiht} = z), \quad (3)$$

where σ is the indicator function for a_{kiht} such that $a_{kiht} \in \{A, C, G, T, -\}$, and h as defined previously.

Advances in technology, which produce dense datasets and progress made in designing statistical and computational methods, have brought to light the inadequacy of AIMs and standard HMM methods in modelling ancestry along an admixed genome. Although violating the independence assumption because

of LD [20], high-density SNP data sets are more powerful [11]. In addition to early methods, SUPPORTMIX [33] also models admixture LD only by combining SVMs and HMMs (Supplementary Text S1). The method was proposed in 2012 in an effort to improve on the computational time and also address the challenge of few typed or existing reference panels to improve multiway local ancestry deconvolution overall. Note that SVMs are popular in computational biology [34] because of their flexibility in data types and ability to handle big data. In local ancestry deconvolution, SVMs train admixed individuals on ancestral populations that are bigger in size than the mixed populations. Refer to Table 1 for the biological or statistical parameters of this model and to Supplementary Text S1 for the mathematics of SVMs. Despite using the rich information on haplotypes, this method is faster than other LD-based methods. In the same year, PCADMIX [39], which leverages principal component analysis from proxy ancestral haplotypes in only modelling admixture LD under a similar HMM as above was proposed (Table 1). Unlike LD-based admixture methods, SUPPORTMIX [33] and PCADMIX [39] are much faster. In 2013, as high-quality genotypes are not always available, a method that is based on exome sequence reads and which uses HMMs, namely, SEQMIX [6], was proposed. This method does not model SNPs in LD. As a result, to reduce the noise and systematic biases from the use of all SNPs [30], methods that model both admixture and background LD emerged [11].

Modelling admixture and background LD

Admixture LD HMM-based approaches are often not realistic in modelling biological data, as dense SNPs violate the independence of alleles within the ancestral population (background LD). To model background LD in addition to admixture LD, HMM was extended and rich ancestral haplotype information was also used (Box 3 and 4). Methods modelling background LD include SABER, SWITCH [36], HAPAA [35], HAPMIX [4], MULTIMIX [41], ALLOY [16] and ELAI [44]. These are based on HMM extensions, including the Markov HMM (MHMM), hierarchical HMM (HHMM), factorial HMM (FHMM), two layer HMM and the multivariate normal (MVN) distributions to model background LD. Two existing methods use MHMM: SABER, proposed in 2006, was the first method to account for background LD in the genetic ancestry inference; and SWITCH [36] was later proposed in 2008. Unlike a standard HMM, MHMM emission models depend on the joint distribution of alleles at nearby markers if ancestry is the same at consecutive markers [21, 22]. Otherwise, if ancestry switches, then the model is as in Equation (2). Box 3 discusses the differences in emission models, while Box 4 discusses the switch models of different local ancestry deconvolution models. The expected length of blocks inherited from population k , is inversely related to a parameter τ , which approximately corresponds to the number of generations, as admixture happened with all ancestral populations mixed from a single point back in time, referred to as a hybrid isolation model. SABER extends Equation (2) for $K > 3$ given the different admixing times, g_k [21]. Assuming the inverse of g_k is the expected length of blocks inherited from population k , the transition matrix is given in Box 4.

With a large parameter set, SABER's Markov model only accounts for background LD between consecutive markers [11]; hence, some researchers described it as incomplete modelling of background LD. As a result, in 2008, SWITCH [36], which considers recombination even if it does not result in an ancestry switch, emerged. SWITCH differs from SABER in that it conditions its MHMM on recombination, whereas in SABER, it is

Box 1: Forward-backward algorithm

Given the observed sequence and HMM parameters, λ , we seek to estimate the probability of a hidden state at time or marker t . The forward-backward algorithm is an application of dynamic programming, which reduces the model complexity in the naive approach of the computation of the joint probability of hidden states and observed states, i.e. $P(X_t|Y_{1:T})$ is estimated as follows:

$$\begin{aligned} P(X_t = k' | Y_{1:T}) &\propto P(X_t = k', Y_{1:T}) \\ &= P(Y_{t-1:T} | X_t = k', Y_{1:t}) P(X_t = k' | Y_{1:t}) \\ &\propto P(Y_{t+1:T} | X_t = k', Y_{1:t}) P(X_t = k', Y_{1:t}) \end{aligned} \quad (4)$$

The second factor in the second line of Equation (4) is given by:

$$\begin{aligned} P(X_t = k', Y_{1:t}) &= \sum_k P(X_t = k' | X_{t-1} = k) P(Y_t | X_t = k', X_{t-1} = k, Y_{1:t-1}) P(X_{t-1} = k, Y_{1:t-1}) \\ &= \sum_k P(X_t = k' | X_{t-1} = k) P(Y_t | X_t = k') P(X_{t-1} = k, Y_{t-1}) \\ &= \epsilon_t(k') \sum_k A(k, k') P(X_{t-1} = k, Y_{t-1}) \end{aligned} \quad (5)$$

Letting $P(X_t = k', Y_{1:t}) = \alpha_t(k')$ from Equation (2), we obtain:

$$\alpha_t(k') = \epsilon_t(k') \sum_k A(k, k') \alpha_{t-1}(k) \quad (6)$$

Setting the first factor of the same line of Equation (22) to $\beta_t(k)$, i.e. $\beta_t(k) = P(Y_{t+1:T} | X_t = k)$ we have:

$$\begin{aligned} \beta_t(k) &= \sum_{k'} P(X_{t+1} = k', Y_{t+1}, Y_{t:T} | X_t = k) \\ &= \sum_{k'} P(Y_{t+1} | X_{t+1} = k') P(Y_{t+2:T} | X_{t+1} = k') P(X_{t+1} = k' | X_t = k) \\ &= \sum_{k'} \epsilon_t(k') \beta_{t+1}(k') A(k, k') \end{aligned} \quad (7)$$

Substituting $\alpha_t(k)$, $\beta_t(k)$, and letting $\gamma_t(k') = P(X_t = k' | Y_{1:T})$ in Equation (4), we have $\gamma_t(k') \propto \alpha_t(k') \beta_t(k')$:

At locus t the likelihood is given by $P(Y_t) = \sum_k \alpha_t(k') \beta_t(k')$. However, Equation (6) depends highly on whether the data are phased. The Viterbi algorithm is another dynamic programming technique, similar to back-tracking in the forward backward algorithm [37].

Box 2: Key definitions

Population admixture—It is the mating of two or more previously isolated populations.

n -way admixture—An admixed population resulting from the mixing of n previously distinct population.

Global ancestry—It is the genome-wide fraction of ancestry contributed by each ancestral population.

Candidate ancestry—It is the possible ancestry that admixed to form an admixed population.

Local ancestry (locus-specific ancestry)—It is the number of alleles inherited from a particular candidate ancestry at a site values are 0, 1 or 2.

Proxy ancestry (candidate ancestry)—It is a population used in representing an ancestral population of admixed individuals.

LD—It is the non-random association of alleles between two or more loci. There exist three major type of LDs:

Mixture LD—It is the LD that is caused by the differences in average ancestral proportions contributed by ancestral populations in sampled individuals. Occurs even if alleles are linked.

Admixture LD—It results when ancestry at nearby markers is inherited together in local chromosomes as a result of the admixture process, resulting in large haplotype blocks. It is weak at all distances and decays slowly.

Background LD—It is the LD within ancestral populations, and it depends highly on population history (i.e. generated by genetic drift and population bottlenecks). It is strong at short distances, results in short haplotype block and decay rapidly.

Local ancestry deconvolution—It is the ancestral origin of each of the alleles at a site.

AIMs—Markers that show significant differences in frequency between diverged populations.

HMMs—Hidden Markov models are statistical models that consist of two sets of variables: observed variables that are used to generate the hidden variables assumed to be a Markov or a memoryless process. They are used to model a sequence data sets, for which the hidden variables are a Markov chain.

Figure 1. The evolution of local ancestry deconvolution since 2003–17.

conditioned on the ancestry states. The SWITCH model is as described in Box 3 and 4, and marginalizing this model over recombination yields Equations (1) and (2). The probability of recombination depends on the admixture generations, genetic distance between consecutive SNPs and the recombination rate. Emission models of SABER and SWITCH are the same, except that SABER ignores recombination that does not result in an ancestry change [36]. If ancestry switches between sites, then the emission model of MHMM is as in Box 3. For SABER, the same ancestry at consecutive markers implies the emission model depends on the allele frequencies at the two consecutive markers [11], while in SWITCH, the model depends on pairwise SNP allele frequencies between consecutive markers (Box 3) [36]. Although SWITCH models background LD and estimates recombination rates, the authors recommended the development of richer MHMM or other models different from the pairwise models in SWITCH and SABER [36]. As a result, methods that use both large- and small-scale HMMs, referred to as the HHMM, were proposed. The first model to use the HHMM is a multiway local ancestry deconvolution method, HAPAA [35]. It filters admixed segments by removing those with sizes that do not meet the expected minimum threshold [35] and assumes the phase of admixed individuals is known. It accounts for phase switch errors (Box 4) through a variable, θ , and its transition probability model is illustrated in Box 4, with the indicator function in Equation (2) replaced by a function of phasing error between consecutive markers for each given individual in a given population. Unlike Equation (3), the emission model in HHMM depends on base pairs, capturing mutation and genotyping error (Equations 10 and 11, Box 3). In 2009, a genotype imputation-based method modelling admixture and background LD, GEDI-ADMIX [37], was introduced. It uses HMMs similar to those in Marchini et al. [45], Kimmel and Shamir [46] and Rastas et al. [47], referred to as FHMMs. This method trains the FHMM on ancestral haplotypes for genotype imputation [37]. In addition to estimating the local ancestry, GEDI-ADMIX imputes missing genotypes at typed or untyped SNPs, which is not offered by most of the methods described so far [35]. Furthermore, GEDI-ADMIX HMMs are of fixed state space [37] unlike those leading to a quadratic computational time based on the number of ancestral haplotypes in HHMM models.

HAPMIX [4] was proposed in 2009 and uses an HHMM emission model (Box 3). It assumes the admixed individual's phase information is unknown and uses a built-in phasing algorithm

[4] in phasing the study samples. HAPMIX is computationally expensive, and its processing time is proportional to the square of the number of ancestral haplotypes. This is mainly because of the number of parameters to be estimated under its ancestry switch model (Box 4). Despite outperforming most existing state-of-the-art methods, HAPMIX is for two-way admixture only. HHMM methods face the following issues: (1) the number of ancestral individuals increases with the model complexity [30], and (2) a haplotype is locally more similar to haplotypes from different ancestral populations. As a result, HHMM yields spurious ancestry switches, explaining why HAPMIX allows miscopying from the wrong populations. The problem worsens when some candidate reference populations are small in size, are not good proxies or are closely related [30]. These issues opened opportunities for other HMM-based methods, such as FHMM and two-layer HMM in multiway ancestry deconvolution.

CHROMOPAINTER [38] leverages the candidate ancestral haplotypes similar to the clustering approach in STRUCTURE by extracting more relevant genealogical information from the haplotype data, which is then represented in a co-ancestry matrix [22]. The elements of the co-ancestry matrix are estimated using an HMM [38] suggested by Li and Stephen's [48]. The co-ancestry matrix measures the genetic similarity between individuals while accounting for the genealogical processes and the structure of the genome. The CHROMOPAINTER model can deconvolute ancestry on linked or unlinked SNPs. Although it is a haplotype-based method, CHROMOPAINTER is inaccurate in estimating local ancestry in populations with sub-continental ancestry and does not incorporate familial relationships, information on admixture or genetic drift. Hence, LAMPLD [30], which uses the Li and Stephen model (HHMM) [48] to leverage the structure of LD in ancestral populations (Box 3), as in HAPAA and HAPMIX, was proposed. Unlike the other HHMMs, by integrating the HHMM within a window-based framework (Box 3), LAMPLD compresses the haplotype information in each ancestry reducing the computational time and addressing the ancestry switch problem. It is less prone to misspecification of model parameters, as it estimates parameters from the ancestral haplotypes. The only burden for LAMPLD is that ancestry is assumed constant within each window. Just like the standard HMM in Equation (1) and (2), the ancestry switch model of LAMPLD (Box 5) depends on a constant recombination and length in base pairs between consecutive windows [30]. Refer to

Table 1. Existing ancestry deconvolution tools

Software	Multiway	Account LD	LD model	Biological/statistical parameters	Reference populations	Admixed populations	Reference
STRUCTURE V2*	✓	✓	HMM	Markers and ancestry proportions	Unphased	Unphased	[14]
ANCESTRYMAP*	✗	✓	HMM	Physical map, recombination and ancestry proportions	Unphased	Unphased	[30]
ADMIXMAP*	✓	✓	HMM	Physical map and ancestry proportions	Unphased	Unphased	[31]
SABER	✓	✓	MHMM	Physical map or recombination distance	Phased/unphased	Phased/unphased	[15]
'LAMP'	✓	✗	✗	Admixture generations, LD threshold and physical map	Unphased	Unphased	[17]
HAPAA	✓	✓	HHMM	Admixture generations and genetic divergence	Phased	Phased	[35]
SWITCH	✓	✓	MHMM	Recombination rate	Phased	Phased	[36]
GEDI-ADMIX	✓	✓	Fixed size FHMM	Admixed and ancestral SNPs (physical map)	Phased	Unphased	[37]
WINPOP	✓	✗	✗	Recombination, admixture generations, LD threshold and physical map	Unphased	Unphased	[18]
HAPMIX	✗	✓	HHMM	Genetic map mutation rate and admixed and ancestral SNPs	Phased	Unphased	[4]
CHROMOPAINTER	✓	✓	Co-ancestry matrix	Recombination rate	Phased	Phased	[38]
LAMPLD	✓	✓	HHMM	Number of hidden states, window size and physical map	Phased	Unphased	[30]
SUPPORTMIX*	✓	✓	HMM	Admixture generations and genetic map	Phased	Phased	[33]
PCADMIX*	✓	✓	Windows of blocks of SNPs	Genetic map and window size	Phased	Phased	[39]
mSPECTRUM	✓	✓		SNPs, mutation and recombination rate	Phased	Phased	[40]
MULTIMIX	✓	✓	MVN	Genetic map, legend file and misfitting probabilities	Phased/Unphased	Phased/Unphased	[41]
ALLOY	✓	✓	Non-homogeneous VLMC	Markers, ancestral proportions, admixture generations and genetic map	Phased	Phased	[16]
RFMIX	✓	✗	✗	Genetic map, window size and admixture generations	Phased	Phased	[42]
EILA	✓	✗	✗	Physical map	Unphased (no missing values)	Unphased (no missing values)	[43]
SEQMIX	✓	✗	✗	Genetic map	Unphased	Unphased	[6]
ELAI	✓	✓	Two-layer HMM	Admixture generations, lower and upper cluster	Phased/unphased	Phased/unphased	[44]
LOTER	✓	✗	✗	—	Phased	Phased	[19]

Note: The columns indicate the tool name (Column 1); tools with * account for admixture LD and not background LD and those without * account for both admixture and background LD, the software's ability to model multiway admixtures (Column 2), ability to model LD (Column 3), LD model (Column 4), biological and statistical parameters required (Column 5), population data type required: ancestral (Column 6) and admixed (Column 7) and finally (Column 8), the reference paper from which more details are found, respectively. ✓ indicates the ability of the software to perform a specified task, and ✗ indicates the inapplicability of the task by a particular tool.

Table 2. Applications of local ancestry distribution

Aim of study	Admixed populations	Number of study samples	Markers	Results obtained	Reference paper
Determining if genetic ancestry is associated with complex diseases (type 2 diabetes)	Two admixed populations: Medellin, Columbia and central Mexico	499 cases and 197 controls from Medellin, and 163 cases and 72 controls from central Mexico	67 AIMS	Non-European ancestry proportion is associated with type 2 diabetes and lower socio-economic status in Latino admixed populations from the South and North America	Florez et al. [52]
Detecting natural selection in African Americans	African Americans	5210 individuals (1890 African Americans and 3320 non-African Americans)	491 526	Six positions were identified with excess European or African ancestry chr1: 17409539, chr2: 241750403, chr2: 37451925, chr3: 116930811, chr6: 163653158 and chr16: 61214438, in the following corresponding regions 1p36, 2q37, 2p22, 3q13, 6q26 and 16q2, respectively. The disease genes in the identified regions were known to be highly associated with human disease like prostate cancer and hypertension	Jin et al. [53]
Inferring genome-wide admixture patterns in the Qatar population	Qatar population	156 individuals	71 982 SNPs	The Qatar population was subdivided by SUPPORTMIX into three: Arab-Qataris, a mixture of middle eastern population and Bedouin or Hadar-Arab ancestry; Persian-Qataris, a mixture of Europeans and non-Middle East Asian populations and sub-Saharan Africans; and African-Qataris, African, Persian and fewer Middle Eastern populations	Omberg et al. [33]
Understanding the disease genetics and history of Latino populations by assessing local ancestry performance in Latino populations	GALA data set of mainland United States: Puerto Rico and Mexico, and Multiethnic Cohort (MEC) data set	489 mother-father-child trio families of Puerto and Mexican ethnic groups and 3204 unrelated Latino individuals from MEC	127 935 common SNPs	Using WINPOP, LAMPLD, ALLOY and PCADMIX, it was shown that deviations exist in local ancestry inference. It was also shown that sites/loci with increased deviations in local ancestry Mendelian inconsistencies in local ancestry (MILANC)	Pasaniuc et al. [25]
Finding the signals of gene-gene interaction on human diseases	Simulated admixed population	10 000 admixed individual	1 million SNPs	Local ancestry deconvolution estimates provides promising results in detecting gene-gene interaction	Aschard et al. [28]
Identifying signatures of selection in Latinos	13 admixed Latin American populations	249 individuals	678 genome-wide micro-satellite markers	Three loci were identified as selection candidates. Two regions had excess of European ancestry at 1p36 and 14q32, and one had excess of African ancestry at 6p22	Deng et al. [54]

Box 3

Assuming that paternal and maternal ancestral states are independent of each other, given Y_t^m as the observed maternal haplotype at t and the corresponding ancestral state X_t^m , LD is modelled by the first-order Markov chain, such that

$$P(Y_t^m | X_1^m, X_2^m, \dots, X_t^m, Y_1^m, \dots, Y_{t-1}^m) = \begin{cases} P(Y_t^m | X_t^m, X_{t-1}^m), & \text{for } X_t^m = X_{t-1}^m \\ P(Y_t^m | X_t^m), & \text{Otherwise.} \end{cases}$$

As a result, the emission probability of maternal alleles at two neighbouring markers $t-1$ and t in the SABER model is given by:

$$P(Y_t^m = s | Y_{t-1}^m = r, X_t^m = k', X_{t-1}^m = k) = B_t(s, r, k', k) = \begin{cases} \bar{B}_{k',t}(s, r), & \text{for } k = k' \\ \bar{B}_{k',t}(s), & \text{Otherwise,} \end{cases} \quad (8)$$

where $\bar{B}_{k',t}(s, r)$ is the chance of having alleles at marker t , given there was allele r at $t-1$, given r and s originated from the same population, i.e. $k = k'$. $\bar{B}_{k',t}(s)$ is the allele frequency of alleles at marker t with origins k' , which is evaluated as in Equation (3). In Equation (8), if ancestry switches between consecutive markers, the first-order MHMM emission model reduces to standard HMM emission model, while having same ancestry at consecutive markers makes the chance of observing an allele at t to depend on the alleles at markers t and $t-1$ [20]. Posterior probabilities of the ancestry given observations are estimated by the likelihood approach through the forward-backward algorithm.

In the SWITCH model, the emission probabilities at t depend on those at $t-1$ for instance, given a two-way admixture, the emission probability is defined as:

$$P(Y_{h,t} = 1 | W_{h,t}, X_{h,t}, Y_{h,t-1}, p_t, q_t, p_{t-1}, q_{t-1}) = \begin{cases} P(Y_{h,t} = 1 | X_{h,t}, p_t, q_t), & \text{for } W_{h,t} = 1, \\ P(Y_{h,t} = 1 | X_{h,t}, X_{h,t-1}, p_{t-1}, q_{t-1}, p_t, q_t), & \text{otherwise} \end{cases} \quad (9)$$

where $Y_{h,t} \in (0, 1)$ – matrix is the observed t^{th} marker of haplotype h , and $X_{h,t}$ is the h^{th} hidden haplotype at marker t . p_t and q_t are the allele frequencies of Populations 1 and 2 at marker t , respectively, p_{t-1} and q_{t-1} are pairwise SNP frequencies at t and $t-1$, and $W_{h,t}$ is the recombination indicator. Equations (8) and (9) are the same, except that, in Equation (9), recombination is considered, while Equation (8) ignores all recombinations that do not change ancestry [22]. Unlike the MHMM emission probability models of SWITCH and SABER, the emission probability of HAPAA is a 5×5 stochastic matrix such that

$$P(\bar{a}_t = z | X_t = X_{kih}) = E(a_{kih}, z) \quad (10)$$

where $z \in \{A, C, G, T, -\}$, and $E(z, z' \neq z)$ capture mutation, genotyping error and haplotypes not observed in representative individuals. Unlike Equation (8), HAPMIX emission model is a 3×3 matrix representing the three possible genotypes, i.e. $u = 0, 1$, or 2 where u is the minor allele count of allele ‘1’. Given the underlying hidden state $(ijklmn)$ and the parental individual k in reference population j , p_{jk} denote the probability of observing genotype u as $e_{ijklmn}^u(t)$ that is,

$$e_{ijklmn}^u(t) = \begin{cases} (1 - e_{ijk}^1(t))(1 - e_{lmn}^1(t)), & \text{for } g = 0, \\ e_{lmn}^1(t)(1 - e_{ijk}^1(t)) + e_{ijk}^1(t)(1 - e_{lmn}^1(t)), & \text{for } g = 1, \\ e_{lmn}^1(t)e_{ijk}^1(t) & \text{for } g = 2. \end{cases}$$

where $e_{ijk}^1(t)$ is the probability of having an admixed chromosome being of type ‘1’ allele at marker t denoted by:

$$e_{ijk}^1(t) = \begin{cases} \theta_i \sigma(\rho_{jk} = 0) + (1 - \theta_i) \sigma(\rho_{jk} = 1) & \text{if } i = j \\ \theta_3 \sigma(\rho_{jk} = 0) + (1 - \theta_3) \sigma(\rho_{jk} = 1) & \text{if } i \neq j \end{cases},$$

where θ_i are mutation parameters capturing historical mutation and genotyping error for $i = 1, 2, 3$ and σ indicator function. In the window-based ancestry deconvolution model, LAMPLD, for a pair of ancestry in a window $X^w = (X_1^w, X_2^w)$, the emission probability of observing a genotype Y^w in window w is given by:

$$P(Y^w | X^w = (X_1^w, X_2^w)) = \sum_{(H_1^w, H_2^w)} P(H_1^w | X_1^w) P(H_2^w | X_2^w). \quad (11)$$

where $P(H_i^w | X_i^w)$ is the probability that haplotype H_i is emitted by ancestry X_i in window w , and the haplotype pair (H_1^w, H_2^w) is compatible with observed genotype Y^w .

Box 4

The SABER switch model is given by: $P(X_t = k' | X_{t-1} = k, \pi, r) = e^{-d_t A}$ where A is a $K \times K$ matrix representing rate of transition from ancestral population i to ancestral population j , obtained as follows:

$$A_{ij} = \begin{cases} i \frac{g_i^2}{\sum_{k=1}^K \pi_k g_k} - g_i, & \text{for } i = j, \\ \pi_j \frac{g_i g_j}{\sum_{k=1}^K \pi_k g_k}, & \text{otherwise} \end{cases}$$

The switch model in another MHMM-based model, SWITCH [22], depends on the occurrence of a recombination event. Let $K = 2$, if recombination occurs between markers t and $t - 1$ of haplotype h , define ancestry $X_{h,t} \in \{0, 1\}$ such that ancestry is 0 if it is inherited from the first population with a probability α to be selected and 1 if not from the first population with probability $1 - \alpha$. In the absence of recombination, ancestry will stay the same at consecutive markers, and thus, the transition model is modelled as follows:

$$P(X_{h,t} | X_{h,t-1}, W_{h,t}, \alpha) = \begin{cases} \sigma(X_{h,t}, X_{h,t-1}) & W_{h,t} = 0, \\ (1 - \alpha)^{X_{h,t}} \alpha^{1-X_{h,t}}, & W_{h,t} = 1 \end{cases} \quad (12)$$

where σ is the indicator function, and $W_{h,t}$ is the recombination event between markers t and $t - 1$. Recombination is a Poisson process, and hence, $W_{h,u}$ and $W_{h,v}$ are independent for $u \neq v$, and the probability of at least one recombination is $1 - e^{-(g-1)d_t r_t}$ where g is the number of generations since admixture. d_t and r_t are the genetic distance between markers t and $t - 1$, and the recombination rate in the region, respectively. Marginalizing over $W_{h,t}$, Equation (12) can reduce to Equation (2). As HHMMs are a two-scale HMM, there are two switch models in HHMM, i.e. the switch model at a small scale, which is between haplotypes within a particular reference population, and large scale, which consist of transitions between populations. HAPAA's small-scale transitions depend on the phasing error and the probability of recombinations. From any putative haplotype, transitions can be any one of the following three states: itself, other putative haplotype or to the end state (non-emitting), o_k . If there is no phase error, then, within an individual, there is no transition between putative haplotypes. Given two emitting states, $X_{kih} \in \{X_{kio}, X_{kil}\}$ denoting haplotype $h \in \{0, 1\}$ for an individual i of population k and two non-emitting states for each population k , I_k and o_k activating the emitting states. Given the hidden state X_t at marker t , recombination points follow a Poisson process with rate $d_t r$ where d_t is the genetic distance between marker t and $t - 1$, and r the recombination rate. The transition probability can be expressed as in following equations:

$$P(X_t = X_{kih}) = \frac{1}{2n_k} \quad \text{for the initial state and} \quad P(X_t = a' | X_{t-1} = a) = \theta e^{-d_t r} + (1 - e^{-d_t r})\omega, \quad \text{where}$$

$$\theta = \begin{cases} w_{kit} & \text{if } a', a \in \{X_{kio}, X_{kil}\} \text{ and } a' \neq a \\ 1 - w_{kit} & \text{if } a', a \in \{X_{kio}, X_{kil}\} \text{ and } a' = a \end{cases} \quad \text{for } w_{kit} \text{ the probability of having a phasing error between markers } t \text{ and } t - 1, \text{ for in-}$$

dividual i in population k , and $\omega = 0$ if both states are emitting states and 1 if at least one of the two consecutive states is the non-emitting state O_k . From the end state, O_k , a transition is made into the other non-emitting state, $I_{k'}$ resulting in a $K \times K$ matrix denoted $A(k, k') = P(O_k \rightarrow I_{k'})$ while transitioning from $I_{k'}$ to the emitting state (either same or different population) and $X_{k'ih'}$ is uniform with probability $1/(2n_{k'})$. Despite being a two-way admixture ancestry deconvolution model, HAPMIX has so many parameters, including the miscopying parameters, p_l for $l = 1, 2$ where p_l is defined by the probability of copying ancestry from population 1 given the actual ancestry was copied from population 2. In addition, switches between individuals within a reference population group occur as a Poisson process with rate ρ_l , the recombination rate [5]. As a result, given a haploid admixed chromosome, a hidden state at marker t is represented by (a, b, c) , $a = 1$ or 2 is the index for which ancestry was drawn from, $b = 1$ or 2 index that captures the population the chromosome copies from at t and so b may be different from a [5] and c is the individual from which the chromosomal segment is copied from. The probability of transition from state (a, b, c) to state (d, e, f) between consecutive sites t and $t - 1$ is denoted by

$$p^t(a, b, c; d, e, f) = \begin{cases} a\pi_d \frac{1-p_d}{n_e} & \text{if } d \neq a \text{ and } e = d \\ a\pi_d \frac{p_d}{n_e} & \text{if } d \neq a \text{ and } e \neq d \\ \bar{a}b \frac{1-p_d}{n_e} + a\pi_d \frac{1-p_d}{n_e} & \text{if } d = a \text{ and } e = d \text{ and } (b \neq e \text{ or } c \neq f) \\ \bar{a}b \frac{1-p_d}{n_e} + \bar{a}a \frac{1-p_d}{n_e} & \text{if } d = a \text{ and } e = d \text{ and } (b = e \text{ and } c = f) \\ \bar{a}b \frac{p_d}{n_e} + a\pi_d \frac{p_d}{n_e} & \text{if } d = a \text{ and } e \neq d \text{ and } (b \neq e \text{ and } c \neq f) \\ \bar{a}\bar{b} + \bar{a}b \frac{p_d}{n_e} + a\pi_d \frac{p_d}{n_e} & \text{if } d = a \text{ and } e \neq d \text{ and } (b = e \text{ and } c = f), \end{cases} \quad (13)$$

where $a = 1 - e^{-r_t g}$, $b = 1 - e^{-r_t p_d}$, $\bar{q} = 1 - q$, and g is the generation since admixture, π_k and n_k are as defined before. The HAPMIX switch model has $2(n_1 + n_2)$ states.

Box 5

The window-based method LAMPLD estimates the HMM parameters from the reference populations and has top-level HMM with $\binom{K}{2}$ hidden states, which are local ancestries on each chromosome within each window, $X^w = (X_1^w, X_2^w)$ that generate genotype blocks, Y^w . The probability of transition from state (X_1^w, X_2^w) to state $(X_1^{w'}, X_2^{w'})$ is given by:

$$P(X_t = (X_1^{w'}, X_2^{w'}) | X_{t-1} = (X_1^w, X_2^w)) = \begin{cases} \theta & \text{if } X_1^{w'} \neq X_1^w, \text{ for at least one } i \\ \theta^2 & \text{if } X_i^{w'} \neq X_i^w, \forall i = 1, 2 \\ 1 - 2\theta - 3\theta^2 & \text{if } X_1^{w'} = X_1^w \end{cases} \quad (14)$$

where $i = 1, 2$; $\theta = D * 10^{-8}$, with D the length in base pairs between window $w = [i, i + L]$ and $w' = [i + L, i + 2 * L]$, where L is the number of SNPs in each window. Given the notation as in FHMM and emission model (Supplementary File: Text S3), the probability of occurrence of recombination between two consecutive sites, $P(r_t) = 1 - e^{-\phi((g-1)d_t)}$, and $P(\bar{r}_t)$, then the transition probabilities for ALLOY are given by:

$$\begin{aligned} P(X_t | X_{t-1}) &= P_{anc(X_t)}(X_t | X_{t-1})P(\bar{r}_t) + P(r_t)P_{anc(X_t)}(X_t) \\ &= P_{anc(X_t)}(X_t | X_{t-1})v + P_{anc(X_t)}(X_t)\bar{v} \end{aligned} \quad (15)$$

where $v = e^{-\phi((g-1)d_t)}$, $\bar{v} = 1 - v$, and $P_{anc(X_t)}(X_t)$ is the ancestral haplotype cluster at t and $P_{anc(X_t)}(X_t | X_{t-1})$ model transitions within ancestral population $anc(X_t)$ accounting for LD. d_t and g are as defined above, and $\phi(z)$ is the recombination rate function. The prior probability of the ancestral haplotype cluster depends on the ancestry proportions [33], π , and the intra-population haplotype cluster, $P_{anc(X_t)}(X_t)$, such that

$$P(X_t) = \Pi_{anc(X_t)} P_{anc(X_t)}(X_t) \quad (16)$$

In the ALLOY model, given a set of markers from a single ancestry, markers are processed in chromosomal order. Nodes represent some history of the allele sequences and are split according to the alleles for such nodes. Nodes are then merged based on the Markov criterion, such that at t , two clusters are merged if the allele sequence probabilities at sites $t + 1, t + 2, \dots, t + z$ resemble each other for some parameter z . For each population, a directed acyclic graph is formed with edges e_t^i for cluster i and the number of haplotypes in population sample passing through cluster i as weights denoted by w_t^i at t . Let s_t^i and T_t^i be the source node and target for the edge at t for cluster i , e_t^i , respectively. Hence, the prior and transition probabilities in Equations (15) and (16) are calculated as follows:

$$P_{anc}(X_t = a_t^{anc,i}) = \frac{w_t^i}{\sum w_t^j}, \text{ and } P_{anc}(X_t = a_t^{anc,i} | X_{t-1} = a_{t-1}^{anc,j}) = \begin{cases} \frac{w_t^i}{\sum w_{t-1}^j}, & \text{if } T_{t-1}^j = s_t^i \\ 0, & \text{Otherwise} \end{cases} \quad (17)$$

The specific $P_{anc(a_t)}(X_t)$ and $P_{anc(a_t)}(X_t = a_t | X_{t-1} = a_{t-1})$ are obtained by repeating the above separately for each ancestry. The notation in this paragraph follows the one described in two-layer HMM (Supplementary File: Text S4). For the main HMM, the switching model is given by:

$$P(X_1 = k, Z_1 = s) = P(Z_1 = s | X_1 = k)P(X_1 = k) = \pi_s^{(i)} \beta_{1ks}$$

at the first marker, and the transition from state $t - 1$ to t is modelled as follows:

$$P(X_t = k', Z_t = s' | X_{t-1} = k, Z_{t-1} = s) = j_t \pi_{k'}^{(i)} \beta_{tk's'} + (1 - j_t) r_t \beta_{tk's'} \sigma(k' = k) + (1 - j_t)(1 - r_t) \sigma(k' = k) \sigma(s' = s) \quad (18)$$

where $\pi_{k'}^{(i)}$ is the probability that individual i jumps into the upper cluster k' given the jump occurs, $\beta_{tk's'}$ is the probability that individual i jumps to lower cluster given the jump occurs with the upper cluster being k , β is a $T \times K \times S$ tensor shared by individuals [36] and σ is an indicator function. More details on the assumptions made on Equation (18) are described in [28]. The other HMM layer in the ELAI model referred to as the ancillary HMM is independent of the main HMM given that m ancestral haplotypes have been observed. In ancillary HMM, the hidden state X is replaced by W , which represents the upper cluster, such that $W_t^{(m)} \in \{1, \dots, K\}$. The transition of hidden states is given by:

$$P(W_1^{(m)} = k) = a_k^{(m)}, \text{ and } P(W_t^{(m)} = k | W_{t-1}^{(m)} = k') = \rho_t \alpha_{k'}^{(m)} + (1 - \rho_t) \sigma(k' = k), \quad (19)$$

where there is no association between the jump probabilities ρ in the ancillary HMM and the jump probabilities in the main HMM, and $a_k^{(m)}$, is the probability that the upper cluster is k .

Table 1 for biological/statistical parameters required in running LAMPLD. To account for model correlation between proxy ancestral haplotypes whilst catering for the unavailability of some phased reference panels [30], later in the same year, MULTIMIX [41] was suggested and is also a window-based method. It uses an MVN distribution, whereby model parameters are trained from MCMC, the EM and the classification EM algorithms (Supplementary Text S2). In the same year, a study implemented PCADMIX (Supplementary Text S2), which uses an MVN emission model. Both MULTIMIX (see admixture LD-based methods) and PCADMIX have ancestry switch models as in Equations (1) and (2) above. Despite the close relationship between LD and population structure, existing methods treated these independently. To address this, mSPECTRUM [40] was proposed. It leverages proxy ancestral haplotypes to jointly infer population structure and recombination events using an infinite HMM. Although it is a multiway admixture deconvolution method, it is computationally expensive like HAPMIX [22].

In the race towards accurately modelling the LD structure to improve local ancestry deconvolution, in 2013, ALLOY [16] was developed. It accounts for LD through a non-homogeneous variable length Markov chain (VLMC). The method in ALLOY captures the way admixed maternal and paternal haplotypes can be inherited using an FHMM (Supplementary Text S3) [16]. Technically, its accuracy improved, as it leveraged the haplotype structure in the compound state. The emission and ancestry switch models are described in Box 5. In 2014, a two-layer HMM that models LD at two scales, i.e. one between ancestral populations and the other within ancestral populations between haplotype groups, ELAI [44], was proposed for multiway local ancestry deconvolution. The model is divided into the main HMM and an ancillary HMM (Supplementary Text S4). ELAI was motivated by the need to infer recombination rates using local ancestry deconvolution. It uses either phased or unphased data, limiting phase uncertainty problems. As a result, ELAI can still perform well when the ancestral populations' high-quality haplotype information is unavailable [44]. Given an upper cluster of one, ELAI extends fastPhase, a phasing method. If the upper and lower clusters are equal, then it yields independent markers, while having a lower cluster descending from an upper cluster reduces ELAI to STRUCTURE [44].

Non-LD-based methods

Methods discussed under this section do not model either background or admixture LD. Some of the methods prune or remove SNPs in LD (LAMP [17] and WINPOP [18]), while others do not model LD but include both linked and unlinked SNPs in local ancestry deconvolution. Owing to the complexity of the MHMM, a window-based method, LAMP [17], emerged in 2008. It is fast and robust and can infer local ancestry in the absence of proxy ancestral genotypes. It uses the naive Bayes classifier and the iterative conditional modes, a clustering algorithm, to estimate the most probable ancestry at a particular chromosomal site through the application of the majority vote for each SNP [17]. It does not deal with issues arising from the use of HMMs and their extensions but at the cost of neglecting admixture and background LD information.

WINPOP [18], a dynamic programming algorithm, was proposed to deconvolve ancestry in closely related populations and reduce the computational cost of HMM-based methods. It extended LAMP, by assuming at least one recombination event in each window and varying the window length depending on how similar ancestral populations are. Similar to LAMP,

WINPOP assumes unlinked markers and removes SNPs in LD. For instance, WINPOP outperforms other methods in closely related populations (Table 3) [18], and LAMP outperformed them in distant ancestral populations (Table 3). As sequence data availability increased, in 2013, Maples et al. [42] proposed a discriminative approach, RFMIX [42], to estimate local ancestry. RFMIX uses the information contained in admixed samples, while generative ancestry deconvolution methods use the information in ancestral samples. Using the information contained in admixed samples was advantageous in the case of scarce reference populations, e.g. the Native Americans [42]. RFMIX uses CRFs parametrized with random forests to estimate ancestry. First-order CRFs are shown in Supplementary Text S5. This method seemed to improve and stabilize the local ancestry estimates in various admixture models, in particular multiway admixture scenarios (Table 3). This could be because it models phase switch errors just like HAPMIX.

To address some of the challenges in local ancestry deconvolution that reduce the inference power, including use of three genotype values, identifying break points, which are useful in estimating generations since admixture, and the independence of SNPs assumption, Yang et al. [43] in 2013, proposed a multivariate statistics-based method (Supplementary Text S6), EILA [43]. It uses quantile regression and k-means classifiers to infer ancestry from genotype data. To overcome the limited power from using three genotype values, EILA assigns a numerical score, *ns*, to admixed genotypes. This numerical score is a value between 0 and 1 and measures the similarity of SNPs to certain ancestral populations [43]. Fused quantile regression is then used in identifying break points. After break point identification, k-means classifiers are used to estimate local ancestry using all SNPs (linked and unlinked).

Recently, a software package LOTER [19], which deconvolved local ancestry in multiway admixtures for a wide range of species, was proposed, but the model in LOTER can currently model phase (Table 1) errors in two-way admixture only. Despite the availability of dense genotype and sequence data on a wide range of species [19], deconvolving local ancestry in non-model species is a challenge. Existing methods require biological and statistical parameters to deconvolve ancestry, while LOTER requires none of these parameters. Biological parameters include admixture time, recombination rates, mutation rates, physical map, etc., while statistical parameters include the number of hidden states for the HMM and window size [19]. LOTER is a non-probabilistic approach formulated from an optimization problem (Supplementary Text S7), using Li and Stephen's copying model [48], where the optimal solution is obtained from dynamic programming. In the LOTER model, an optimization problem similar to that of EILA is solved. Table 1 provides a summary of different ancestry deconvolution tools. From a number of previous reviews [20–22, 49], we observed that some non-LD-based methods outperform LD-based methods, which could be because of the estimation of a high number of parameters in some of the LD-based approaches given the complexity of modelling the LD structure. Another reason could be the ability of non-LD-based methods to use the information contained in the haplotypes and also model phase switch errors, for example RFMIX and LOTER.

Challenges, evaluation and opportunities

Despite efforts that have been invested in local ancestry inference or local ancestry deconvolution, several challenges still exist. Previously, we discussed the importance of local ancestry

Table 3. Details of comparing or evaluating local ancestry deconvolution methods

Aim of study	Methods used to deconvolve ancestry	Rank as per findings	Reference paper
Introducing a local ancestry deconvolution method: WINPOP	SABER, HAPAA, LAMP and WINPOP	<ul style="list-style-type: none">• WINPOP• LAMP• HAPAA• SABER	Pasaniuc et al. [18]
Introducing a local ancestry deconvolution method: LAMPLD	WINPOP, GEDI-ADMIX, HAPMIX and LAMPLD	<ul style="list-style-type: none">• LAMPLD• WINPOP• GEDI-ADMIX• HAPMIX	Baran et al. [30]
Introducing a local ancestry deconvolution method: SUPPORTMIX	WINPOP and SUPPORTMIX	<ul style="list-style-type: none">• SUPPORTMIX• WINPOP	Omberg et al. [33]
Introducing a local ancestry deconvolution method: PCADMIX	LAMP, HAPMIX and PCADMIX	<ul style="list-style-type: none">• PCADMIX• LAMP	Brisbin et al. [39]
Introducing a local ancestry deconvolution method: mSPECTRUM	LAMP, HAPMIX, mSPECTRUM	<ul style="list-style-type: none">• HAPMIX• mSPECTRUM• LAMP	Sohn et al. [40]
Introducing a local ancestry deconvolution method: ALLOY	WINPOP and ALLOY	<ul style="list-style-type: none">• HAPMIX• ALLOY	Rodriguez et al. [16]
Introducing a local ancestry deconvolution method: RFMIX	LAMPLD, SUPPORTMIX and RFMIX	<ul style="list-style-type: none">• WINPOP• RFMIX• LAMPLD	Maples et al. [42]
Introducing a local ancestry deconvolution method: EILA	HAPMIX, LAMP and EILA	<ul style="list-style-type: none">• SUPPORTMIX• EILA• LAMP	Yang et al. [43]
Association studies	LAMPLD and WINPOP	<ul style="list-style-type: none">• HAPMIX• LAMPLD	Chimusa et al. [51]
Admixture mapping	LAMP, LAMPLD and MULTIMIX	<ul style="list-style-type: none">• WINPOP• LAMPLD• MULTIMIX	Chen et al. [50]
Introducing a local ancestry deconvolution method: ELAI	HAPMIX, LAMPLD and ELAI	<ul style="list-style-type: none">• LAMP• ELAI	Guan [44]
Introducing a local ancestry deconvolution method: LOTER	LAMPLD, RFMIX and LOTER	<ul style="list-style-type: none">• LAMPLD• HAPMIX• LOTER• RFMIX• LAMPLD	Dias-Alves et al. [19]

Note: Studies considered multiway admixtures. Ranking is for distant ancestral populations. Though all methods underperformed in closely related populations, HAPMIX was better than the other two. Ranking based on three-way ancient admixtures.

and the applications of the estimates from such studies. However, the success of these studies depends heavily on the accuracy of the estimates. Studies in [Tables 2](#) and [3](#) showed that deviations in local ancestry estimates still exist in multiway admixtures. As illustrations, [Chen et al. \[50\]](#) used three state-of-the-art methods (refer to [Table 3](#)) to deconvolve local ancestry. The best two methods ([Table 3](#)) that are all LD-based models had differences in ancestry estimates on almost 20% of the analysed SNPs. This could be because of the way these two approaches account for the biological or statistical parameters within their model, for example, although both methods require window size to approximate ancestry estimates, LAMPLD performed well across all the window sizes used (150, 100, 75 and 50 SNPs), while MULTIMIX performance improves with the decrease in window size [50]. In addition, [Chimusa et al. \[51\]](#) using a different population with two state-of-the-art methods also highlighted that admixture mapping is limited by inaccuracies in local ancestry inference. However, the two methods used do not all fall under the same group category, and in this case, LAMPLD, which uses the haplotypic information in ancestral populations and models LD, outperformed the non-LD-based method that uses allele frequencies of ancestral populations to infer ancestry.

Deviations in local ancestry may be because of several reasons, such as signals of selection, risk of disease, miscalling true ancestry and error in genotyping [24, 50]. As a result, inaccuracies in local ancestry are one of the main challenges in studying genetic ancestry. There are several possible causes of inaccuracies in local ancestry deconvolution related to the fact that (a) most existing methods (1) require biological or statistical parameters to infer ancestry, which are not always accurate if provided, (2) assume information on reference ancestral populations is known and (3) assume markers are unlinked; (b) there are few existing individuals that have been genotyped; (c) existing methods generally underperform in ancient admixtures; (d) genotypes only take three values, which might reduce the power to infer ancestry; (e) existing methods are benchmarked or tested for up to three-way admixture; (f) HMM-based methods have a large parameter set and (g) methods do not model natural selection when estimating ancestry. Some of these challenges have been successfully addressed, while others are still pending. To account for (e), a framework that deconvolves local ancestry using existing tools adjusted to incorporate more than three ancestral populations where possible should be implemented. This framework should also facilitate the analysis of all the autosomal chromosomes. Challenges (a (1)) and (c) have recently been addressed by LOTER, as it performs well in admixed populations that mixed >150 generations ago [19]. The SWITCH model addressed (f) by using LAMP to initialize parameters in HMM. Challenge (g) is still yet to be addressed. SWITCH [36] addressed the non-existence of one of the ancestral reference panels by its capability to infer ancestry, as it estimates the allele frequency of one reference given the genotype data. On the other hand, RFMIX [42] uses the information in admixed samples themselves to address (a(2)) and account for (b), but the user is required to provide the ancestral populations.

Admixture scenarios are diverse, and the local ancestry deconvolution tools are proposed for different applications, under different model assumptions. Therefore, local ancestry inference methods are not easy to compare especially if they also do not fall under the same category. As a result, following [Pasanici et al. \[25\]](#), we compare methods that fall within the same category, that is, LD-based or non-LD-based methods. [Tables 2](#) and [3](#) are used to compare existing methods using existing studies.

To assess if the local ancestry problem is solved in multiway admixtures, [Chen et al.](#) used two LD-based, multiway ancestry deconvolution methods (LAMPLD and MULTIMIX), which both divide the admixed genome into windows before ancestry estimation. It was noted that LAMPLD outperform MULTIMIX. [Maples et al.](#) on introducing RFMIX, they, assessed LAMPLD and SUPPORTMIX and discovered that LAMPLD again performs better. Although LAMPLD outperformed ALLOY and PCADMIX in the Mexican population, in 2014 ([Table 2](#)) [25], it was outperformed by ELAI ([Table 3](#)) [44]. However, as [Guan 2014](#) is the only study that compared ELAI with other local ancestry inference methods, more studies should be done to confirm whether ELAI is the best local ancestry deconvolution method that models LD. Given distant ancestral populations, EILA outperforms LAMP ([Table 3](#)), while because of the use of the haplotype information, LOTER and RFMIX outperform WINPOP, which uses ancestral allele frequencies. LOTER outperformed these in both distant ancestral populations and recent/and ancient admixtures.

Conclusions and perspectives

Over the past decade, studies of admixed populations have increasingly gained interest and today >20 ancestry deconvolution methods have been introduced. It was expected that critical applications out of these ancestry deconvolution approaches would not only shed light on the demographics and admixture structure of various admixed populations but also improve our understanding of signatures of natural selection, ethnic difference in drug response and the risk of complex diseases. Although a few applications of local ancestry deconvolution, mostly in admixed populations from the American continent, have emerged, unusual ancestry deviation and switch error in estimating local ancestry estimates in multiway admixed populations within a complex or multifaceted admixture model are still a challenge. Here, we have dissected approaches for pinpointing ancestry along the genome of admixed populations and summarized evolutionary factors they account for. We discussed different statistical model properties, assumptions and results from existing evaluations. We believe that this survey may orient the development of new methods and enable researchers to choose appropriate tools for given population characteristics depending on where the local ancestry estimates should be applied. Furthermore, this survey will enable software developers to identify current challenges and advances in ancestry deconvolution, providing them with useful information for the implementation of practical tools, which consider current medical and population genetics demands.

Although recent development of the LOTER method [19] provides more insights into previous challenges regarding ancestry deconvolution in ancient admixtures and a complete reliance on biological parameters, such as LAMP [17] and WINPOP [18], the computational accuracy is still a challenge because of the use of optimization problems. On the other hand, methods that can model biological parameters, such as mutations and recombination, and allow some miscopying as in HAPMIX, CHROMOPAINTER, etc., may suffer from ancestry deviation and computational complexity. Thus, taking into account current challenges, we recommend SWITCH or ELAI [44] to researchers interested in estimating recombination events. If the aim is to understand admixture patterns without knowledge of some of the ancestral populations of the study samples, while having access to a pool of reference population panels, it would be

more effective to use SUPPORMIX [33]. In addition, for ancient or recent admixture with large sample sizes of reference panels (large training sets), ALLOY [16] may be more relevant, as the large sample sizes of reference panels improves the LD model used in ALLOY [16]. In cases where the users suspect imperfect reconstruction in the haplotype, LAMPLD [30], RFMIX [42] and LOTER [19] are instrumental, and HAPMIX should be avoided in such scenarios, where possible, when deconvolving local ancestry users should avoid using default parameters, e.g. admixture generations when using RFMIX. This is because the default generations are too recent. In the case of low coverage sequence data sets, such as exome or targeted data for history analysis or genetic association studies, it may be more effective to use SEQMIX [6]. In summary, current implemented local ancestry deconvolution models are not fully accurate in complex admixture modelling and do not fully enable admixture mapping in multifaceted admixed populations. There is still a need to design different admixture scenario simulation framework, which assesses all existing local ancestry deconvolution models. This may enable the development of a novel and/or integrative local ancestry inference framework that captures appropriate biological parameters given a population admixture model, is flexible to handle the presence or absence of proxy ancestral data and allows end users to select the most appropriate models.

Key Points

- Dissecting existing ancestry deconvolution methods and underlying tools.
- Comprehensive survey and consistent classification of existing local ancestry inference approaches.
- Discussion of issues related to application of existing ancestry deconvolution methods.
- Highlighting possible gaps that still need to be filled in the area of local ancestry deconvolution.
- Orienting local ancestry deconvolution tool end users based on the evolution of these tools.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Acknowledgements

The authors thank researchers who have contributed towards advancing the local ancestry inference around the world. Also, the authors thank those who have helped in the preparation of this manuscript.

Funding

Some of the authors are supported by the Organization for Women in Science for the Developing World (OWSD) and Swedish International Development Cooperation Agency (Sida), and the German Academic Exchange Service (DAAD), and others are supported in part by the National Institutes of Health (NIH) Common Fund (grant numbers U24HG006941 and U01HG009716), Wellcome Trust/AESA Ref: H3A/18/001 and 1U54HG009790-01 through H3ABioNet project, HI Genes Africa and IFGeneRA, respectively.

References

1. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. *Nat Genet* 2003;**33**:266–75.
2. Yang JJ, Cheng C, Devidas M, et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nat Genet* 2011;**43**(3):237–41.
3. Koehl A. Estimating ancestry and genetic diversity in admixed populations. The University of Mexico, 2016.
4. Price AL, Tandon A, Patterson N, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 2009;**5**(6):e1000519.
5. Gravel S. Population genetics models of local ancestry. *Genetics* 2012;**191**(2):607–19.
6. Hu Y, Willer C, Zhan X, et al. Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. *Am J Hum Genet* 2013;**93**(5):891–9.
7. Ma Y, Zhao J, Wong JS, et al. Accurate inference of local phased ancestry of modern admixed populations. *Sci Rep* 2015;**4**(1):5800.
8. Durand EY, Do CB, Mountain JL, et al. Ancestry composition: a novel, efficient pipeline for ancestry deconvolution. *BioRxiv* 2014. <https://www.biorxiv.org/content/early/2014/10/18/010512>.
9. Khayatadeh N, Mészáros G, Gredler B, et al. Prediction of global and local simmental and red holstein friesian admixture levels in Swiss fleckvieh cattle. *Poljoprivreda* 2015;**21**(Suppl 1):63–7.
10. Shriner D. Overview of admixture mapping. *Curr Protoc Hum Genet* 2013;**Chapter 1**:Unit 1.23.
11. Seldin MF, Pasaniuc B, Price AL. New approaches to disease mapping in admixed populations. *Nat Rev Genet* 2011;**12**(8):523–8.
12. Reich D, Patterson N, De Jager PL, et al. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat Genet* 2005;**37**(10):1113–18.
13. Zhu X, Cooper RS, Elston RC. Linkage analysis of a complex disease through use of admixed populations. *Am J Hum Genet* 2004;**74**(6):1136–53.
14. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003;**164**(4):1567–87.
15. Tang H, Coram M, Wang P, et al. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* 2006;**79**(1):1–12.
16. Rodriguez JM, Bercovici S, Elmore M, et al. Ancestry inference in complex admixtures via variable-length Markov chain linkage models. *J Comput Biol* 2013;**20**(3):199–211.
17. Sankararaman S, Sridhar S, Kimmel G, et al. Estimating local ancestry in admixed populations. *Am J Hum Genet* 2008;**82**(2):290–303.
18. Pasaniuc B, Sankararaman S, Kimmel G, et al. Inference of locus-specific ancestry in closely related populations. *Bioinformatics* 2009;**25**(12):i213–21.
19. Dias-Alves T, Mairal J, Blum MG. Loter: a software package to infer local ancestry for a wide range of species. *BioRxiv* 2017. <https://www.biorxiv.org/content/early/2018/04/26/213728>.
20. Liu Y, Nyunoya T, Leng S, et al. Softwares and methods for estimating genetic ancestry in human populations. *Hum Genomics* 2013;**7**:1.
21. Padhukasahasram B. Inferring ancestry from population genomic data and its applications. *Front Genet* 2014;**5**:204.
22. Gompert Z, Buerkle CA. Analyses of genetic ancestry enable key insights for molecular ecology. *Mol Ecol* 2013;**22**(21):5278–94.

23. Martin AR, Gignoux CR, Walters RK, et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet* 2017;**100**(4):635–49.
24. Chimusa ER, Daya M, Moller MR, et al. Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *PLoS One* 2013;**8**(9):e73971.
25. Pasaniuc B, Sankararaman S, Torgerson DG, et al. Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics* 2013;**29**(11):1407–15.
26. Duconge J, Ruano G. The emerging role of admixture in the pharmacogenetics of Puerto Rican Hispanics. *J Pharmacogenomics Pharmacoproteomics* 2010;**1**:101. doi: 10.4172/2153-0645.1000101.
27. Goetz LH, Uribe-Bruce L, Quarless D, et al. Admixture and clinical phenotypic variation. *Hum Hered* 2014;**77**(1–4):73–86.
28. Aschard H, Gusev A, Brown R, et al. Leveraging local ancestry to detect gene-gene interactions in genome-wide data. *BMC Genet* 2015;**16**:124.
29. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 1989;**77**(2):257–86.
30. Baran Y, Pasaniuc B, Sankararaman S, et al. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 2012;**28**(10):1359–67.
31. Patterson N, Hattangadi N, Lane B, et al. Methods for high-density admixture mapping of diseases genes. *Am J Hum Genet* 2004;**74**(5):979–1000.
32. Hoggart CJ, Shriver MD, Kittles RA, et al. Design and analysis of admixture mapping studies. *Am J Hum Genet* 2004;**74**(5):965–78.
33. Omberg L, Salit J, Hackett N, et al. Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC Genet* 2012;**13**(1):49.
34. Eilers PH, De Menezes RX. Quantile smoothing of array CGH data. *Bioinformatics* 2005;**21**(7):1146–53.
35. Sundquist A, Fratkin E, Do CB, et al. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res* 2008;**18**(4):676–82.
36. Sankararaman S, Kimmel G, Halperin E, et al. On the inference of ancestries in admixed populations. *Genome Res* 2008;**18**(4):668–75.
37. Paşaniuc B, Kennedy J, Mandoiu I. Imputation-based local ancestry inference in admixed populations. In: Măndoiu I, Narasimhan G, Zhang Y (eds). *Bioinformatics Research and Applications (ISBRA 2009)*, Lecture Notes in Computer Science, Vol. 5542. Springer, Berlin, Heidelberg, 2009, 221–33.
38. Lawson DJ, Hellenthal G, Myers S, et al. Inference of population structure using dense haplotype data. *PLoS Genet* 2012;**8**(1):e1002453.
39. Brisbin A, Bryc K, Byrnes J, et al. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol* 2012;**84**(4):343–64.
40. Sohn KA, Ghahramani Z, Xing EP. Robust estimation of local genetic ancestry in admixed populations using a nonparametric Bayesian approach. *Genetics* 2012;**191**(4):1295–308.
41. Churchhouse C, Marchini J. Multi way admixture deconvolution using phased or unphased ancestral panels. *Genet Epidemiol* 2013;**37**(1):1–12.
42. Maples BK, Gravel S, Kenny EE, et al. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* 2013;**93**(2):278–88.
43. Yang JJ, Li J, Buu A, et al. Efficient inference of local ancestry. *Bioinformatics* 2013;**29**(21):2750–6.
44. Guan Y. Detecting structure of haplotypes and local ancestry. *Genetics* 2014;**196**(3):625–42.
45. Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;**39**(7):906–13.
46. Kimmel G, Shamir R. A block-free hidden Markov model for genotypes and its application to disease association. *J Comput Biol* 2005;**12**(10):1243–60.
47. Rastas P, Koivisto M, Mannila H, et al. Phasing genotypes using a hidden Markov model. In: *Bioinformatics Algorithms: Techniques and Applications*. 2005, 355–73.
48. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 2003;**165**(4):2213–33.
49. Tang H, Choudhry S, Mei R, et al. Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am J Hum Genet* 2007;**81**(3):626–33.
50. Chen M, Yang C, Li C, et al. Admixture mapping analysis in the context of GWAS with gaw18 data. *BMC Proc* 2014;**8**(Suppl 1):S3.
51. Chimusa ER, Zaitlen N, Daya M, et al. Genome-wide association study of ancestry-specific TB risk in the South African coloured population. *Hum Mol Genet* 2014;**23**(3):796–809.
52. Florez JC, Price AL, Campbell D, et al. Strong association of socioeconomic status with genetic ancestry in Latinos: implications for admixture studies of type 2 diabetes. *Diabetologia* 2009;**52**(8):1528–36.
53. Jin W, Shuhua X, Haifeng W, et al. Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res* 2012;**22**(3):519–27.
54. Deng L, Ruiz-Linares A, Xu S, et al. Ancestry variation and footprints of natural selection along the genome in Latin American populations. *Sci Rep* 2016;**6**(1):21766.