Chapter

Orienting Future Trends in Local Ancestry Deconvolution Models to Optimally Decipher Admixed Individual Genome Variations

Gaston K. Mazandu, Ephifania Geza, Milaine Seuneu and Emile R. Chimusa

Abstract

Rapid advances in sequencing and genotyping technologies have significantly contributed to shaping the area of medical and population genetics. Several thousand genomes are completed with millions of variants identified in the human deoxyribonucleic acid (DNA) sequences. These genomic variations highly influence changes in phenotypic manifestations and physiological functions of different individuals or population groups. Of particular importance are variations introduced by admixture event, contributing significantly to a remarkable phenotypic variability with medical and/or evolutionary implications. In this case, knowledge of local ancestry estimates and date of admixture is of utmost importance for a better understanding of genomic variation patterns throughout modern human evolution and adaptive processes. In this chapter, we survey existing local ancestry deconvolution and dating admixture event models to identify possible gaps that still need to be filled and orient future trends in designing more effective models, which account for current challenges and produce more accurate and biological relevant estimates.

Keywords: genomic variations, admixture, local ancestry, dating admixture event, linkage disequilibrium

1. Introduction

Today, advances in high-throughput technologies have generated huge amounts of human genomics data in public domains. These data are useful for medical and population genetics to understand the population history, human evolution and demographics, susceptibility to disease, and response to drug. Over time, humanity has experienced the exchange of genetic materials across populations, mainly due to population migrations [1], which have led to wide human genetic variations as results of interbreeding or mating between different populations previously isolated. These genetic variations observed in the human deoxyribonucleic acid (DNA) sequences are caused by inheritance processes, such as mutation and recombination. Generally, the mating process yields the genetic recombination break points,

introduces some variations, and creates mixed DNA segments. As a consequence, current human populations are admixed [2, 3] with specific genomes displaying a mosaic of segments originating from different ancestral populations [1, 2, 4], wide phenotypic variations, divergent genetic ancestry, and different traits observed among individuals in worldwide population groups. Thus, it is critical to understand the dynamics related to the origin of these variations, the evolution process, and its consequences in human heredity and health.

Studying admixture patterns in human populations consists of characterization of admixture features in human populations, including admixture mapping and date to admixture events. Admixture mapping combines both the identification of genetic variants underlying the ethnic difference in disease risk and inference of ancestry estimates associated with these genetic variants. Estimation of ancestry is commonly known as genetic ancestry inference, which is either global or local ancestry inference. Global ancestry inference estimates the overall proportion contributed by each ancestral population to the admixed genome; while, local ancestry deconvolution (local ancestry inference) estimates the number of copies from a particular population at a given site [5]. Together, admixture mapping and date to admixture events provide a better understanding of the genetic variation features throughout modern human evolution, the demographics, and adaptive processes of human populations. Currently, analyzing admixture patterns has become central to genomics research, contributing to a wide range of biomedical applications. Current advance in technologies is facilitating the movement of people worldwide, thus influencing the complexity of population admixture dynamics and leading to multifaceted admixture events. On the other hand, the determination of local ancestry through genotyping and microarray datasets has empowered the approaches for dating mutation, selection, and admixture events [6, 7].

The significance of the local ancestry inference topic is viewed through the research interests it has raised over the last two decades. Several models exist for local ancestry deconvolution, including ANCESTRYMAP [8], ADMIXMAP [9], SABER [10], LAMP [11], LAMPLD/LAMPHAP [12], SUPPORTMIX [13], EILA [14], LOTER [15], etc. **Figure 1** displays the implementation dynamics of different local ancestry deconvolution models graphically, indicating the time each model was introduced. Local ancestry inference is relevant in personalizing medicines, understanding complex diseases, localizing missing sequences in reference genomes and understanding the population history and demographics. Subsequently, several studies have particularly been focusing on dating past admixture events, relevant to population migrations, heritable genes associated to some diseases, and responses to

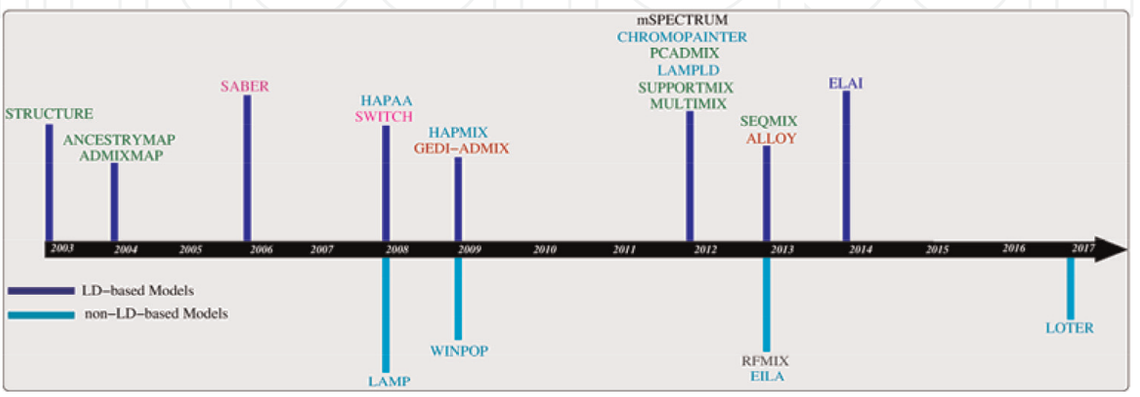

Figure 1.
The evolution of local ancestry deconvolution since 2003 to 2017.

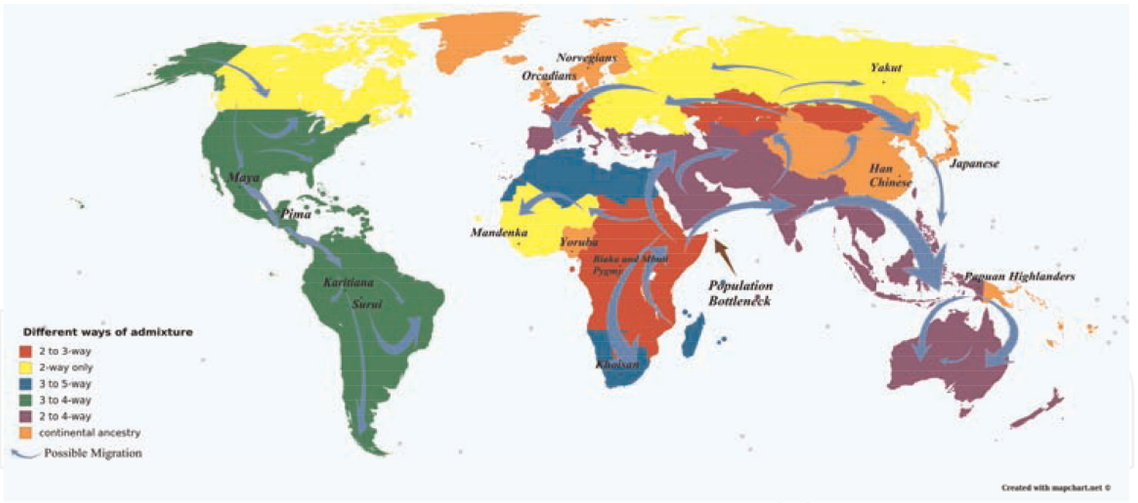

Figure 2.
A partial worldwide admixture painting map. The figure shows several worldwide admixed populations with patterns identified through published paper on population structure from 2008 to 2018. The population migrations within and between continents have resulted in different admixed populations ranging from one- to five-way admixtures.

treatment [16]. The date of admixture in a given population can be predicted by analyzing the ancestral track, break-points, and linkage disequilibrium (LD) [17]. Also, distinction between date of admixture events is made with the use of LD and ancestral tracts in the admixed genomes [17]. Nowadays, there are several models for predicting the age of an admixture event, which are classified into two main groups: LD-based approaches and haplotype-based approaches [17, 18]. These models use information from genomes of several population groups around the world as representative or equivalent ancient populations known to influence the migration and/or admixture processes, yielding observed admixed population patterns worldwide (**Figure 2**).

In this chapter, we survey current models for deconvoluting local ancestry and dating admixture events and explore computational techniques used in these models. We highlight advances made so far in this genomic era and opportunities behind these models and challenges or gaps that still need to be addressed. This informs users and researchers on the current state of research, and orient future trends in designing more effective models, which account for current challenges and produce more accurate and biological relevant estimates. In the subsequent sections, we provide an overview of existing methods used for inferring local ancestry estimates and dating admixture events.

2. Overview of admixture feature inference models

In this section, we survey current models used to elucidate admixture patterns, including local ancestry estimates (deconvolution) and dating admixture events. These models assume that the T genotyped sites are biallelic and the genotype information of the K reference candidate ancestral and admixed populations are considered known. Ancestry at different sites or windows follows a Markov chain. Recombination is assumed to occur at every generation resulting in Poison recombination points with a rate which depends on both the recombination rate, r, and number of generations since admixture, g, and individuals are independent of each other.

2.1 Local ancestry inference models

As pointed out previously, existing local ancestry inference models can be categorized into two main groups based on whether the model makes use of admixture/background linkage disequilibrium (LD) or not.

2.1.1 LD-based models for local ancestry inference

LD-based models account for LD in local ancestry deconvolution, and due to the importance of LD in disease mapping, the first local ancestry methods fall into this category. They assume that ancestry along an admixed individual genome follows a first order Markov chain. This means that the immediate past state captures all the information on past states [19]. As a result, LD-based models assume that, at every site, the observed admixed genotypes are generated by the unobserved ancestry, and hence, Hidden Markov Model (HMM) and its extensions are used to infer the unobserved (hidden) states. Thus, to deconvolute ancestry along the admixed genome, these models have three model parameters, namely the initial, transition and observation, or emission probability models. Due to uncertainty and the number of parameters involved, LD-based methods use Markov Chain Monte Carlo (MCMC), forward-backward, or Viterbi algorithms to determine the hidden ancestry sequence for a given individual. Falush et al. and Patterson et al. modeled ancestry switch between ancestry populations at a given site, $X_t \in \{1, ..., K\}$, by

$$P(X_1 = k|q, r) = q_k, \tag{1}$$

$$P\left(X_t = k|X_{t-1} = k^{'}, q, r\right) = \delta\left(k^{'} = k\right)e^{-d_t r} + \left(1 - e^{-d_t r}\right)q_k \ for \ 1 < t \leq T \tag{2}$$

representing the first marker, and the transition probability between consecutive markers with $\delta\left(k = k^{'}\right)$ is the indicator function and d_t the genetic distance between sites t and $t-1$, above and q_k the proportion of ancestry contributed by candidate ancestral population k such that $q = \left(q_1, ..., q_k\right)$ is a vector of ancestry inherited from each ancestral population. On haploid data, the probability of a recombination event is $1 - e^{-d_t r}$, meaning that the probability of no recombination is $e^{-d_t r}$ [8, 20]. LD-based methods can be subdivided into admixture LD-based and admixture and background LD methods. Note that admixture LD occurs when ancestry at nearby markers is inherited together and background LD is the LD within ancestral populations, and it depends highly on population history (i.e, generated by genetic drift and population bottlenecks).

2.1.1.1 Admixture LD-based models

Admixture LD-based methods are models that account for LD that resulted from the admixture process. They do not model background LD. Admixture LD-based methods include the early methods, for example, STRUCTURE V2 [20], ANCESTRYMAP [8], and ADMIXMAP [9], which are based on the Bayesian framework. Early methods rely on markers that show significant difference in frequency between ancestral populations (AIMs). Admixture LD-based models assume that markers are independent and the global and ancestral allele frequencies are known. They integrate HMM with MCMC, and their switch model and initial and transition models are as in Eqs. (1) and (2), respectively. Since LD-based methods do not model background LD, their observation model depends on only

the allele frequency of the ancestry at that site. For instance, assuming K = 2, Patterson et al. defined the emission probability by

$$P(Y_t = y|X_t = n_a) = \begin{cases} \binom{2}{n_a} p_k^{n_a}(1-p_k)^{2-n_a} & \text{if } n_a = 0 \text{ or } 1 \\ \begin{pmatrix} 2(1-p_1)(1-p_2) \\ p_2(1-p_2)+p_1(1-p_2) \\ p_1 p_2 \end{pmatrix} & \text{if } n_a = 2 \end{cases} \quad (3)$$

where y and $n_a \in \{0,1\}$ are numbers of reference alleles of an admixed individual at t, and that of alleles from population 1, respectively. p_k is the allele frequency of population $k \in \{1,2\}$ at the site t, such that when $n_a = 0, p_k = p_1$ while $p_k = p_2$ when $n_a = 2$. Nowadays, technological, statistical, and computational advances avail enormous amounts of high density SNP data. Although high density SNPs violate the independence assumption due to background LD [21], they contain more information than in AIMs [22]. To loosen the independence assumption and minimize noise and systematic biases from unmodelled LD, more advanced local ancestry inference methods emerged [22]. These methods include SEQMIX [23], PCADMIX [24], and SUPPORTMIX [13].

SUPPORTMIX [11] models only admixture LD by combining support vector machines (SVMs) and HMM. It was proposed in 2012 to improve on the computational time and address the challenge of a few typed or nonexistent reference panels, which overall improve multi-way local ancestry deconvolution. SUPPORTMIX is the first model to allow the learning of ancestral surrogates given a pool of reference panels. As a result, it is capable to train ancestral populations that are bigger in size than those that are mixed. Since SVMs can handle huge datasets, SUPPORTMIX is faster than early methods. It uses the rich haplotype information. Also proposed in 2012, PCADMIX [24] divides the genome into contiguous windows of SNPs as in SUPPORTMIX. It leverages principal component analysis from proxy ancestral haplotypes to model admixture LD under a standard HMM. Similar to SUPPORTMIX, PCADMIX is fast and requires phased data. Nevertheless, SUPPORTMIX and PCADMIX do not model phase switch errors, and as a result, in 2013, SEQMIX [23] was proposed. Unlike all other admixture LD-based methods, SEQMIX is based on exome sequence, reads data, and uses HMM. SEQMIX models only admixture LD and prunes SNPs in background LD. As a result, to reduce noise and systematic biases from using all SNPs [10] whilst not fully modeling LD (background), admixture and background LD methods emerged [22].

2.1.1.2 Admixture and background LD models

Since the biological data often have some dependences that violate the independence assumption in standard HMM, admixture LD-based methods are often not realistic. To relax the independence assumption, the HMM is extended to either Markov HMM, factorial HMM, hierarchical HMM, or two-layer HMM or other multivariate statistical models such as multivariate normal distribution (MVN) and a rich ancestral haplotype data are used unlike early methods. This is the case for SABER [10], SWITCH [25], HAPAA [26], HAPMIX [4], MULTIMIX [27], ALLOY [28], and ELAI [29]. MHMMs were the first HMM extension in local ancestry. They were first implemented in SABER and later in SWITCH. SABER was the first method to model background LD in the genetic ancestry inference. MHMM assumes that the current observed haplotype depends on both the current ancestry

and the immediate past observation. The difference in the MHMM and admixture LD HMM-based is that when ancestry switches between sites t − 1 and t, then the MHMM observation model depends on the joint distribution of allele frequencies at the two sites [6, 30], defined as follows [10]:

$$P\left(Y_t = c|Y_{t-1} = d, X_t = k, X_{t'} = k^{'}\right) = B_t\left(c, d, k^{'}, k\right),$$
$$P\left(Y_t = c|Y_{t-1} = d, X_t = k, X_{t'} = k^{'}\right) = \begin{cases} \tilde{B}_{k',t}(c,d) \ \textit{for} \ \ k^{'} = k \\ \overline{B}_{k',t}(c) \quad \textit{otherwise} \end{cases} \tag{4}$$

where $\tilde{B}_{k,t}(c,d)$ is the probability of having alleles at marker t provided there was allele d at t − 1 and $\overline{B}_{k,t}(c)$ the allele frequency of alleles at marker t have for origins the population k. However, if the ancestry does not switch, then the observation model is like that of models in Section 2.1.1.1. The transition model of the SABER model accounts for the differences in admixture times that are in the real case of continuous gene flow where populations contribute their genetic material to the admixture in different generations [10]. Tang et al. defined the probability of switching from ancestry k at t to k at t as

$$A_{ij} = \begin{cases} q_i \dfrac{g_i^2}{\sum_{k=1}^{K} q_k g_k} - g_i, & \text{for i = j,} \\ q_j \dfrac{g_i g_j}{\sum_{k=1}^{K} q_k g_k}, & \text{otherwise} \end{cases} \tag{5}$$

where g_k is the admixture time when population k started to contribute to the admixture.

However, SABER has a large parameter set, and does not explicitly model background LD as it models background LD using first order Markov chain [22]; other methods such as SWITCH were proposed. SWITCH takes into recombination even if it does not result in an ancestry switch, emerged. In contrast to SABER, SWITCH conditions the MHMM on recombination. Similar to early methods, probability of recombination depends on the admixture generations, genetic distance between consecutive SNPs, and the recombination rate. Thus, if the transition probability model in SWITCH is marginalized over recombination, then it is similar to Eq. (2) for two-way and Eq. (5) for multi-way. Although SWITCH models background LD and estimates recombination rates, the authors recommended richer MHMM or other different models that would outperform the SWITCH and SABER pairwise models [25]. As a result, methods that use both large- and small-scale HMM, referred to as the HHMM, were introduced.

2.1.2 Non-LD-based local ancestry inference models

Non-LD methods neither model background nor admixture LD. They either remove SNPs in LD which is the case for LAMP [11] and WINPOP [31], or use all SNPs (linked and unlinked SNPs) without modeling LD; this is the case for EILA [14], RFMIX [32], and LOTER [15]. Since MHMMs have a large number of parameters and do not model LD explicitly, an algorithmic approach that divides genome into windows of SNPs, LAMP [11], emerged in 2008. LAMP is fast and robust, and can infer local ancestry even without proxy ancestral genotypes. This is the case for

two-way admixtures. It uses the naive Bayes classifier and a clustering algorithm known as the iterative conditional modes. LAMP estimates the most probable ancestry at a site by applying the majority vote for each SNP [11]. Although accuracy is comprised, LAMP does not suffer from challenges of HMM and extension. As a result, LAMP underperforms in closely related populations, and hence it was extended to WINPOP [31], a dynamic programming algorithm. Unlike LAMP, WINPOP assumes at least one recombination event within each window and varies the window length depending on the genetic distance between populations. Hence, WINPOP and LAMP outperform other methods in closely and distantly related populations, respectively. Both LAMP and WINPOP assume unlinked markers and discards SNPs in LD.

As the admixed sequence data availability increases, Maples et al. proposed a discriminative approach to estimate local ancestry, RFMIX [32]. A discriminative approach estimates the posterior probability directly and not via the joint probability distribution. In contrast to generative ancestry inference models, RFMIX uses the information contained in admixed individuals. This is advantageous in cases of genotyped few reference panels. This is the case for Native Americans [32]. RFMIX uses conditional random fields (CRFs) parametrized on random forests. It outperforms in multi-way admixtures maybe due to modeling phase switch errors. In 2013, EILA [14], a multivariate statistic based method, was proposed particularly to increase inference power through addressing three common challenges in local ancestry. Addressed challenges are the independence of SNP assumption, difficulties in identifying break points, and the use of three genotype values. Instead of raw genotypes, EILA uses a numerical value between 0 and 1. The score determines how close SNPs are to the ancestral populations. Breakpoints are a challenge to identify, but EILA identifies them by fused quantile regression facilitating the use of estimates in admixture dating. Finally, k-means classifiers are used to infer ancestry using all genotyped SNPs [14].

Recently, a software package that deconvolves local ancestry in multi-way admixtures for a wide range of species, LOTER [15], was proposed. LOTER can account for phase errors in two-way admixture only. It facilitates the local ancestry inference process and its application in non-model species [15]. Unlike other methods, LOTER needs no biological such as admixture time and recombination rate or statistical parameters such as, number of hidden states and misfit probabilities to deconvolve ancestry [15]. Although it uses the Li and Stephen's copying model [33] as in LAMPLD/LAMPHAP, LOTER is a nonprobabilistic approach formulated from an optimization problem. Its solution is obtained through dynamic programming.

Finally, different existing LD and non-LD-based local ancestry inference models are summarized in **Table 1** extracted from Geza et al. [34].

2.2 Models for dating admixture events in a genome

Several models are now available to determine the date of admixture events in a given admixed genome. Breakpoints of haplotypes are used by some models while others focus on the ancestry blocks. Models based on ancestry blocks for dating admixture are formulated using either an empirical criteria or variants associated with a specific population. In order to determine the average length of the admixture block, these methods then assign ancestry on predefined windows using either wavelet transformation or conditional random fields [35]. On the other hand, there are models requiring rapid decrease in haplotype block sizes to estimate the date of the admixture event [36]. This suggests that, in general,

Software	Multi-way	Account LD	LD model	Biological/statistical parameters	Reference populations	Admixed populations	Year of publication
STRUCTURE V2*	✓	✓	HMM	Markers, and ancestry proportions	Unphased	Unphased	August 2003
ANCESTRYMAP*	✗	✓	HMM	Physical map, recombination and ancestry proportions	Unphased	Unphased	May 2004
ADMIXMAP*	✓	✓	HMM	Physical map and ancestry proportions	Unphased	Unphased	May 2004
SABER	✓	✓	MHMM	Physical map or recombination distance	Phased/unphased	Phased/unphased	July 2006
"LAMP"	✓	✗	✗	Admixture generations, LD threshold, and physical map	Unphased	Unphased	February 2008
HAPAA	✓	✓	HHMM	Admixture generations and genetic divergence	Phased	Phased	February 2008
SWITCH	✓	✓	MHMM	Recombination rate	Phased	Phased	February 2008
GEDI-ADMX	✓	✓	Fixed size FHMM	Admixed and ancestral SNPs (physical map)	Phased	Unphased	May 2009
WINPOP	✓	✗	✗	Recombination, admixture generations, LD threshold, and physical map	Unphased	Unphased	June 2009
HAPMIX	✗	✓	HHMM	Genetic map mutation rate and admixed and ancestral SNPs	Phased	Unphased	June 2009
CHROMOPAINTER	✓	✓	Co-ancestry matrix	Recombination rate	Phased	Phased	January 2012
LAMPLD	✓	✓	HHMM	Number of hidden states, window size and physical map	Phased	Unphased	May 2012
SUPPORTMIX*	✓	✓	HMM	Admixture generations and genetic map	Phased	Phased	June 2012
PCADMIX*	✓	✓	Windows of blocks of SNPs	Genetic map and window size	Phased	Phased	August 2012
mSPECTRUM	✓	✓		SNPs, mutation and recombination rate	Phased	Phased	August 2012
MULTIMIX	✓	✓	MVN	Genetic map, legend file and misfitting probabilities	Phased/unphased	Phased/unphased	November 2012

Software	Multi-way	Account LD	LD model	Biological/statistical parameters	Reference populations	Admixed populations	Year of publication
ALLOY	✓	✓	Non-homogeneous VLMC	Markers, ancestral proportions, admixture generations, and genetic map	Phased	Phased	February 2013
RFMIX	✓	✗	✗	Genetic map, window size, and admixture generations	Phased	Phased	August 2013
EILA	✓	✗	✗	Physical map	Unphased (no missing values)	Unphased (no missing values)	November 2013
SEQMIX	✓	✗	✗	Genetic map	Unphased	Unphased	November 2013
ELAI	✓	✓	Two layer HMM	Admixture generations, lower and upper cluster	Phased/unphased	Phased/unphased	May 2014
LOTER	✓	✗	✗	—	Phased	Phased	November 2017

Table 1.
Existing 20 ancestry deconvolution tools: ✓ indicates the ability of the software to perform a specified task, ✗ indicates the inapplicability of the task by a particular tool. Unless explicitly specified, LD refers to background LD.

models used for dating admixture events can be subdivided in two main classes [17, 18], namely those based on LD and those based on the haplotype distribution, as mentioned earlier.

2.2.1 LD-based models for dating admixture events

An admixture event is mainly characterized by the transfer of genes from the ancestral populations to the admixed ones. This leads to the appearance of linkage disequilibrium with regard to the ancestral populations. However, this LD formed often decreases with time. Also, the rate of decay of LD is a function of recombination and the proportion of the admixture [35]. Inversely, many methods employ this rate to calculate the time since the admixture event occurs.

In 2011, Moorjani et al. introduced a method to determine the weighted correlation for a pair of SNPs [36]. This correlation coefficient is further used to measure the LD with ancestral populations [37]. The time of admixture is then determined by analyzing the correlation with respect to the genetic distance, and also fitting using a least squares method the decay of the correlation [35]. This method got improved in 2011 by Loh et al. [18]. The major improvements are in terms of computation. Loh et al. employed instead a fast Fourier transform and other faster techniques to determine the optimal distance to the fitting curve. This method has another advantage that it reduces considerable biases in the estimation of the time of admixture [18, 36]. Later, Loh et al.'s method was improved by Pickrell et al. [38] by introducing the notion of mixture exponential decay in order to take into account the admixture events in the given admixed population history. It mainly focuses on the decay of the LD.

2.2.1.1 Multiple weighted correlation coefficient

Let us consider three ancestral populations k_1, k_2, and k_3, and Q the admixed population. Let us denote by ω_{1-2}, ω_{1-3}, and ω_{2-3} three weighted linkage disequilibrium scores computed based on all possible pairs of SNPs between the three ancestral populations: $k_1 - k_2$, $k_1 - k_3$, and $k_2 - k_3$, respectively, in the admixed population Q calculated using the method proposed by Loh et al. According to Prickrell et al., the multiple weighted correlation coefficient is [38],

$$C_{k_1-k_2,\,k_1-k_2,\,k_2-k_3} = \sqrt{\frac{\omega_{2-3}^2 + \omega_1^2 - 2\omega_{2-3}\omega_{1-2}\omega_{1-3}}{1-\omega_{2-3}^2}}. \tag{6}$$

The date of admixture between population k_1 and k_3 is

$$D_{k_1,k_2,k_1k_2-k_2k_3} = \begin{cases} w_0 + w_1 e^{-n_1\frac{\delta_n}{100}}, & \text{for one admixture event} - D_{(1)}, \\ w_0 + w_1 e^{-n_1\frac{\delta_n}{100}} + w_2 e^{-n_2\frac{\delta_n}{100}}, & \text{in the case of two admixture events} - D_{(2)}, \end{cases} \tag{7}$$

with n_1 and n_1 the number of generations; δ_n the genetic distance; w_1 and w_2 stand for the value of the multiple weighted LD; and w_0 the affine term. $D_{(1)}$ is the date of admixture of population Q in the case of admixture either between $k_1 - k_2$ or $k_2 - k_3$. On the other hand, if it is assumed that two admixture events took place between $k_1 - k_3$ and either $k_1 - k_2$ or $k_2 - k_3$, the date of the admixed population is given by $D_{(2)}$.

2.2.2 Haplotype distribution-based models for dating admixture events

Among the haplotype-based approaches, there is the likelihood method introduced in 2009 by Price et al. [4]. It basically determines the number of breakpoints using Hidden Markov Model. It is also able to determine the number of alleles at a particular site inherited from a given ancestor in a population. This is done in two steps. First, the method consists in identifying haplotype from the proxy ancestry populations, and secondly, the origin of each haplotype bock is identified by comparing their likelihood for one ancestral population versus the others. Considering an admixed genome, the likelihood of an observed allele is given by

$$H_{uvw}(h) = \begin{cases} \theta_u P(t_{vw} = 0) + (1 - \theta_u) P(t_{vw} = 1), \text{ if } u = v, \\ \theta_3 P(t_{vw} = 0) + (1 - \theta_3) P(t_{vw} = 1), \quad \text{otherwise} \end{cases} \tag{8}$$

with θ_u, $u \in \{1, 2, 3\}$ the mutation parameter is; h represents the haplotype site in the chromosomal offspring; the function t_{vw} is an indicator function. It takes the value 1 if individual w coming from offspring x has the same haplotype with the ancestral population v and 0 otherwise; and P is the probability to inherit a pair of haplotype [4]. The number of generations since admixture is given by

$$G = \frac{C}{4\gamma(1-\gamma)\zeta} \tag{9}$$

where ζ is the total Morgan length, γ the proportion of admixture, and C the observed number of breakpoints [4].

On the other hand, Pugach et al. [17] employed the wavelet transform to design a haplotype block approach. The aim of this method is to derive the time of admixture of a given population using the simple hybrid isolation model. It proceeds in two main steps. First, it obtains a signal of admixture from the admixed data using the principal component technique. The second step consists in deriving the date of admixture using the signal obtained in the first step [17].

Pool and Nielsen also built a haplotype-based approach. It used precautionary ancestral populations to infer the date of admixture from the genome of an admixed population [39]. It assumed that after a number of generation g, the distribution of the ancestral haplotypes follows exponential distribution given by

$$f(\lambda, g) = ge^{-\lambda g} \tag{10}$$

where λ is the length of haplotypes. Also, the mean of this distribution is known and is equal to $\frac{1}{g}$.

Further methods include that of Gravel developed in 2012 for the identification of multiple ancestral populations in a given admixture dataset [40]. Also, Jin et al. [41] came up with a similar method to explain admixture dynamics. The method incorporates several models including gradual admixture (GA), hybrid isolated (HI), and continuous gene flow (CGF) models [41], which can be extended to GA-Isolation (GA-I) and CGF-Isolation (CGF-I) by considering isolation after admixture [42]. Hellenthal et al. [43] on the other hand built up on the work of Lawson et al. [44] on dating admixture. This method particularly considers the genome of an admixed individual to be a set chunk DNA coming from other individuals. The scheme of this method is mainly made of two stages. The first stage consists in dividing the genome into chunks and matching each of them to the proper ancestral individual. This stage is achieved with the help of Hidden Markov

Model. The second stage consists in identifying haplotypes and determining their respective ancestral population [43, 44]. Moreover, the admixture event and its date are derived by fitting the decay of the ancestral haplotype with an exponential distribution curve. Moreover, Ni et al. developed a method based on the observation that the date of admixture events is related to the model used. Their method consists in using the likelihood ratio test to identify the best model for the inference of the date of admixture. Furthermore, they are able to estimate several admixture events with the given optimal model [35].

Finally, different existing models and tools for dating admixture events are summarized in **Table 2** extracted from Chimusa et al. [35].

Tool	Category	Admixture model	Priori proxy ancestral raw data	Multi-way events	Online link
ROLLOFF	LD-based model	HI	Yes	No	https://github.com/DReichLab/AdmixTools/
ALDER		HI	Yes	No	http://cb.csail.mit.edu/cb/alder/
MALDER		HI	Yes	Yes	https://github.com/joepickrell/malder/
CAMer		HI, GA, CGF, GA-I, CGF-I	Yes	Yes	https://github.com/david940408/CAMer
IMAAPs		HI, GA, CGF, GA-I, CGF-I	Yes	Yes	http://www.picb.ac.cn/PGG/resource.php
StepPCO	Haplotype/ancestry block size distribution-based model	HI	Yes	Yes	https://bioinf.eva.mpg.de/download/StepPCO/
Adware		HI, Dual-admixture	Yes	Yes	https://cran.r-project.org/web/packages/adwave/index.html
HAPMIX		HI	Yes	Yes	http://genetics.med.harvard.edu/reichlab/Reich_Lab/Software.html/
MultiWaveIner		HI	Yes	Yes	https://github.com/xyang619/MultiWaveInfer/ or http://www.picb.ac.cn/PGG/resource.php
GLOBBERTROTTER		HI, GA, CGF	No	Yes	https://github.com/maarjalepamets/human-admixture/
Tracts		HI, GA, CGF	No	Yes	https://github.com/sgravel/tracts/
Ancestry_HMM		HI	No	No	https://github.com/russcd/

Table 2.
Existing dating admixture genomic tools.

3. Challenges and perspectives

3.1 Case of local ancestry inference models

Although several models exist to deconvolve local ancestry, most studies that evaluate such models showed that deviations in local ancestry estimates still exist in multi-way admixtures. Deviations in local ancestry also result from genetic drift, miscalling true ancestry, and genotyping errors. However, the signals from these factors affect the whole genome while that of unmodelled natural selection affects particular regions. For example, Chen et al. using four local ancestry inference models to scan for disease-related loci through admixture mapping showed that although all of them are LD based and divide the genome into windows of continuous SNPs, MULTIMIX and LAMPLD estimates differed in almost 20% of the analyzed SNPs. This results from the differences in the biological and statistical parameters they require and the mathematical approaches they use. Another association study by Chimusa et al. [45] also pointed out that admixture mapping is still limited by inaccuracies in multi-way local ancestry deconvolution when they evaluated one LD-based and one non-LD-based local ancestry models, WINPOP and LAMPLD.

Inaccuracies in local ancestry estimates may result from the use of statistical or biological parameters in the estimation process, which are not always accurate when provided. It could also be due to the dependence of models on reference panels which for some populations are few or even not sampled for others. This is the case for the Native Americans. More so for other admixed populations, their history is not well known. When applied to ancient admixtures, existing methods may yield spurious estimates as they were designed for recent admixtures. Existing methods do not account for natural selection; hence, some deviations exist in regions that are under selection [45]. Also, most of them are benchmarked for three-way admixtures.

Since each model was introduced to address a particular challenge that models before it faced, it is clearly expected that no model or tool can achieve the best performance in all admixture scenarios and not trading estimate accuracy with computational speed. Using existing studies by Geza et al. [34], more than 50% of studies that either introduced a model or evaluated methods for association mapping showed that LAMPLD/LAMPHAP outperforms most LD-based methods. And the only LD-based method than outperformed LAMPLD is ELAI; however, this is the only study that assessed ELAI with other models. In cases where LD-based models were compared to non-LD-based models, RFMIX outperformed LAMPLD in three cases highlighted in [34], while another separate study aiming to determine the place of admixture of an admixed population RFMIX also outperformed. This could be because RFMIX can deconvolve ancestry in closely related populations [46]. However, a recent assessment between RFMIX and LOTER resulted in LOTER outperforming in ancient admixtures [15].

Generally, each model is implemented as a tool in local ancestry deconvolution, existing as individual scripts requiring unique inputs and producing unique outputs. This challenges researchers with a limited computational background; thus, there is lack of a unified framework which can require a standard easy to manipulate input files and output results in a way that is easy to process for further application. In conclusion, for informed decisions on models and algorithms, existing models or tools should be assessed within a unified framework. This will allow them to be tested on different admixture scenarios and also incorporating most state-of-the-art LD and non-LD based models.

3.2 Case of the dating admixture models

The evolution of human populations and the history of the mixture of these populations have been deciphered using statistical and computational methods. These methods have been found to perform well when dealing with single point admixture event in two-way admixed populations [35]. However, as any method, they not only have advantages but also pitfalls regarding the estimation of admixture dates in some cases. It is challenging to fit to real admixed populations (for more than 3-way admixture context) in the existing models dating admixture events due to several reasons, including reliance to optimal local ancestry estimates and accurate ancestry breakpoints. This suggests that there is still a need for designing an integrative or a new model to dating admixture events for current multi-way admixed populations to further advance our understanding of human demographics and movement, and facilitate admixture mapping and estimation of the age of a disease locus contributing to disease risk.

In addition, it have been discovered that the mixture exponential decay model over-estimates the date of older admixture events [35] and was suggested to detect at most three admixture events. As mentioned earlier, Ni et al. [47] dealt with the optimization of the method used in dating admixture estimation. They took into account several models but the evaluation of their technique is not effective in the estimation of ancient and multi admixture events [35, 47]. On the other hand, several practical considerations can further limit these approaches including the use of proxy ancestry populations in the estimations which could bias the accuracy of the result. This is the case when dealing for instance with low sample size and inappropriate proxy ancestral populations [35]; the requirement of having accurate LD patterns, ancestry haplotypes distribution, and a big sample size of the admixed population. Thus, there is a need for an adequate model for inferring different dates of admixture events and matching real admixture history using proxy ancestry-based methods [35].

4. Conclusions

Currently, more than 20 models exist and are implemented as software to deconvolve local ancestry and 12 tools for dating admixture events. In this chapter, we discussed in detail and summarized the most commonly used models, the model assumptions, statistical and biological parameters they require, and existing challenges. This discussion highlights the need for designing more effective models, which account for current challenges and produce more accurate and biologically relevant estimates. Furthermore, it provides useful information for the implementation of practical tools, which consider current medical and population genetic demands. More importantly, this may guide users in the choice of appropriate tools for specific applications and can assist software developers in designing more advanced tools for local ancestry deconvolution and dating admixture events.

Acknowledgements

Some of the authors are supported in part by the National Institutes of Health (NIH) Common Fund [grant numbers U24HG006941 (H3ABioNet) and 1U01HG007459–01 (SADaCC)]. One of the authors is fully funded by the Organization for Women in Science for the Developing World (OWSD) and Swedish International Development Cooperation Agency (Sida). The content of this

publication is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Conflict of interest

The authors declare that they have no competing interest.

Author details

Gaston K. Mazandu[1,2,3*], Ephifania Geza[1,3], Milaine Seuneu[1,2] and Emile R. Chimusa[2]

1 African Institute for Mathematical Sciences (AIMS), Cape Town, South Africa

2 Division of Human Genetics, Department of Pathology, Faculty of Health Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town (UCT), Cape Town, South Africa

3 Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, South Africa

*Address all correspondence to: kuzamunu@aims.ac.za

References

[1] Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nature Genetics. 2003;**33**: 266-275

[2] A. Koehl, Estimating Ancestry and Genetic Diversity in Admixed Populations. The University of New Mexico. Thesis 2016

[3] Yang JJ, Cheng C, Devidas M, Cao X, Fan Y, Campana D, et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nature Genetics. 2011;**43**(3):237-241

[4] Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genetics. 2009;**5**(6):e1000519

[5] Thornton TA, Bermejo JL. Local and global ancestry inference and applications to genetic association analysis for admixed populations. Genetic Epidemiology. 2014;**38**(S1): S5-S12

[6] Liu Y, Nyunoya T, Leng S, et al. Softwares and methods for estimating genetic ancestry in human populations. Human Genomics. 2013;**7**(1):1

[7] Bhatia G, Patterson N, Pasaniuc B, et al. Genome-wide comparison of African-ancestry populations from care and other cohorts reveals signals of natural selection. American Journal of Human Genetics. 2011;**89**: 368-381

[8] Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, et al. Methods for high-density admixture mapping of diseases genes. The American Journal of Human Genetics. 2004;**74**(5):979-1000

[9] Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. The American Journal of Human Genetics. 2004;**74**(5):965-978

[10] Tang H, Coram M, Wang P, Zhu X, Risch N. Reconstructing genetic ancestry blocks in admixed individuals. The American Journal of Human Genetics. 2006;**79**(1):1-12

[11] Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. The American Journal of Human Genetics. 2008;**82**(2): 290-303

[12] Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, et al. Fast and accurate inference of local ancestry in latino populations. Bioinformatics. 2012;**28**(10):1359-1367

[13] Omberg L, Salit J, Hackett N, Fuller J, Matthew R, Chouchane L, et al. Inferring genome-wide patterns of admixture in qataris using fifty-five ancestral populations. BMC Genetics. 2012;**13**(1):49

[14] Yang JJ, Li J, Buu A, Williams LK. Efficient enference of local ancestry. Bioinformatics. 2013;**29**(21):2750-2756

[15] Dias-Alves T, Mairal J, Blum MG. Loter: A software package to infer local ancestry for a wide range of species. Molecular Biology and Evolution. 2018; **35**(7):msy126

[16] Cheng R, Lim J, Samocha K, et al. Genome-wide association studies and the problem of relatedness among advanced intercross lines and other highly recombinant populations. Genetics. 2010;**185**:1033-1044

[17] Pugach I, Matveyev R, Wollstein A, et al. Dating the age of admixture via

wavelet transform analysis of genome-wide data. Genome Biology. 2011;**12**:R19

[18] Loh P-R, Lipson M, Patterson N, et al. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013;**193**: 1233-1254

[19] Murphy KP. Machine Learning: A Probabilistic Perspective. Cambridge, Massachusetts, London: MIT press; 2012

[20] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics. 2003;**164**(4):1567-1587

[21] Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012;**99**(6):323-329

[22] Seldin MF, Pasaniuc B, Price AL. New approaches to disease mapping in admixed populations. Nature Reviews Genetics. 2011;**12**(8):523-528

[23] Hu Y, Willer C, Zhan X, Kang HM, Abecasis G. Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. The American Journal of Human Genetics. 2013;**93**(5):891-899

[24] Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, et al. PCAdmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Human Biology. 2012; **84**(4):343

[25] Sankararaman S, Kimmel G, Halperin E, Jordan MI. On the inference of ancestries in admixed populations. Genome Research. 2008;**18**(4):668-675

[26] Sundquist A, Fratkin E, Do CB, Batzoglou S. Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Research. 2008;**18**(4):676-682

[27] Churchhouse C, Marchini J. Multiway admixture deconvolution using phased or unphased ancestral panels. Genetic Epidemiology. 2013; **37**(1):1-12

[28] Rodriguez JM, Bercovici S, Elmore M, Batzoglou S. Ancestry inference in complex admixtures via variable-length Markov chain linkage models. Journal of Computational Biology. 2013;**20**(3): 199-211

[29] Guan Y. Detecting structure of haplotypes and local ancestry. Genetics. 2014;**196**(3):625-642

[30] Padhukasahasram B. Inferring ancestry from population genomic data and its applications. Frontiers in Genetics.**5**:204

[31] Paşaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. Bioinformatics. 2009; **25**(12):i213-i221

[32] Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics. 2013;**93**(2): 278-288

[33] Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;**165**(4):2213-2233

[34] Geza E, Mugo J, Mulder NJ, Wonkam A, Chimusa ER, Mazandu GK. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Briefings in Bioinformatics. 2018. DOI: 10.1093/bib/bby044

[35] Chimusa ER, Defo J, Thami PK, Awany D, Mulisa DD, Allali I, et al. Dating admixture events is unsolved problem in multi-way admixed populations. Briefings in Bioinformatics. 2018:1-58. https://doi.org/10.1093/bib/bby112

[36] Moorjani P, Thangaraj K, Patterson N, et al. Genetic evidence for recent population mixture in India. Human Genetics. 2013;**93**:422-438

[37] Moorjani P, Patterson N, Hirschhorn J, et al. The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genetics. 2011;**7**:e1001373

[38] Pickrell J, Reich D. Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics. 2014;**30**:377-389

[39] Pool J, Nielsen R. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics. 2009;**181**:711-719

[40] Gravel S. Population genetics models of local ancestry. Genetics. 2012; **191**:607-619

[41] Jin W, Li R, Zhou Y, et al. Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping. Human Genetics. 2014;**22**:930

[42] Zhou Y, Qiu H, Xu S. Modeling continuous admixture using admixture-induced linkage disequilibrium. Scientific Reports. 2017;**7**:43054

[43] Hellenthal G, Busby G, Band G, et al. A genetic atlas of human admixture history. Science. 2014;**434**: 747-751

[44] Lawson D, Hellenthal G, Myers S, et al. Inference of population structure using dense haplotype data. PLoS Genetics. 2012;**8**:e1002453

[45] Chimusa ER, Zaitlen N, Daya M, Møller M, van Helden PD, Mulder NJ, et al. Genome-wide association study of ancestry-specific tb risk in the South African coloured population. Human Molecular Genetics. 2014;**23**(3):796-809

[46] Xue J, Lencz T, Darvasi A, Pe'er I, Carmi S. The time and place of European admixture in Ashkenazi Jewish history. PLoS Genetics. 2017; **13**(4):e1006644

[47] Ni X, Yuan K, Yang X, et al. Inference of multiple-wave admixtures by length distribution of ancestral tracks. Heredity (Edinb). 2018;**121**:52-63