

REVIEW

Models, methods and tools for ancestry inference and admixture analysis

Kai Yuan^{1,2,†}, Ying Zhou^{1,2,†}, Xumin Ni^{3,†}, Yuchen Wang^{1,2}, Chang Liu^{1,2} and Shuhua Xu^{1,2,4,5,*}

¹ CAS Key Laboratory of Computational Biology, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, CAS, Shanghai 200031, China

² University of Chinese Academy of Sciences, Beijing 100049, China

³ Department of Mathematics, School of Science, Beijing Jiaotong University, Beijing 100044, China

⁴ School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China

⁵ Collaborative Innovation Center of Genetics and Development, Shanghai 200438, China

* Correspondence: xushua@picb.ac.cn

Received May 5, 2017; Revised July 1, 2017; Accepted July 3, 2017

Background: Genetic admixture refers to the process or consequence of interbreeding between two or more previously isolated populations within a species. Compared to many other evolutionary driving forces such as mutations, genetic drift, and natural selection, genetic admixture is a quick mechanism for shaping population genomic diversity. In particular, admixture results in “recombination” of genetic variants that have been fixed in different populations, which has many evolutionary and medical implications.

Results: However, it is challenging to accurately reconstruct population admixture history and to understand of population admixture dynamics. In this review, we provide an overview of models, methods, and tools for ancestry inference and admixture analysis.

Conclusions: Many methods and tools used for admixture analysis were originally developed to analyze human data, but these methods can also be directly applied and/or slightly modified to study non-human species as well.

Keywords: genetic admixture; ancestry; population structures; demographic history; archaic introgression; incomplete lineage sorting

INTRODUCTION

Population admixture has been a common phenomenon throughout the history of modern humans, and occurs when previously isolated populations come into contact through colonization and migration. Admixed populations have received much attention due to their potential advantages in the discovering disease-associated genes. For instance, a gene mapping strategy known as admixture mapping has been instrumental in identifying disease-associated genetic variants [1–4]. The statistical power of admixture mapping relies on the extended and elevated linkage disequilibrium (LD) in admixed populations, and can be determined by population history and admixture processes [1,5,6]. Therefore, as shown in several theoretical and simulation studies, population

admixture dynamics has a strong effect on the statistical power of admixture mapping [6–9]; however, the in-depth admixture dynamics of highly admixed populations has not been extensively studied or examined. In fact, only a few studies have examined simulated data [10,11] or experimental data with sparse markers [6,9]. Recently, the availability of genome-wide high-density single nucleotide polymorphisms (SNPs) data has facilitated the study of detailed genetic structures in admixed populations [12–17]. Population history in admixed populations can be recovered by utilizing the information in genomes, such as break points of recombination [16], admixture linkage disequilibrium (ALD) [18–20], and the length of ancestral tracks [21–26] because admixed genomes can be explained as the mosaics of segments from different ancestries. To date, there have been many methods and

[†] These authors contributed equally to this work.

corresponding computational tools developed for analyzing population admixture. Most of the studies relied on simplified models that did not take into account the inherent complexity of admixture processes; however, there are some methods that utilize more complex models. In this study, we give an overview of models, methods, and tools for ancestry inference and admixture analysis.

DETECTION OF GENE FLOW

Gene flow detection is the primary task of admixture analysis. An intuitive way to verify the existence of admixture and gene flow is to discover adequate variations in genome that are shared only by target populations and donor populations. D test [27] and f test [18] were developed base on this idea. These two methods are widely used to identify population admixture to detect the gene flow magnitude and direction. These methods could be used to detect admixture and gene flow in large time scales. Additionally, D test and f test can be used to detect not only recent admixture of modern humans [28,29], they can also be used to detect ancient admixture [30,31] and archaic introgression [32–35].

GLOBAL ANCESTRY INFERENCE

After detecting gene flow, we get a rough idea about each ancestral population. Ancestry inference is a way to obtain detailed information about each ancestral population present in admixed populations. There are currently two different paradigms underlying ancestry inference: global ancestry inference and local ancestry inference [36]. Global ancestry inference is more about estimating genome level contribution proportions from each ancestor population, which gives a global view of admixture in target populations. Global ancestry inference can be classified into two main categories: model-based methods and non-parametric approaches. In this study, we used model-based methods.

In recent years, multiple software programs have been developed to study global ancestry inference, STRUCTURE is the most well-known and widely used software program for examining global ancestry inference [37]. STRUCTURE is a model-based clustering method that uses multi-locus genotype data to infer genomic make-up and population structures. STRUCTURE has been developed to several versions and series, based on a Bayesian approach that utilizes a Markov Chain Monte Carlo algorithm to obtain samples from the posterior distribution. STRUCTURE is based on a model in which there are K independent populations, where each K can be unknown, and each of which can be characterized by a set of allele frequencies at each locus. Sample individuals can be assigned into a particular population or be assembled

by several independent populations if they are from admixed populations. This method can be applied to unlinked genetic markers (making use of allele frequency information), which are found in linkage equilibrium. Moreover, STRUCTURE also assumes that there are no particular mutation processes and that the populations are in Hardy–Weinberg equilibrium (HWE) [37].

New STRUCTURE series methods have been developed in recent years. First, Falush *et al.* developed a new prior model for the allele frequencies within each population, and accounted for the correlations between linked loci that arose in admixed populations, which improved the resolution of methods for subtle subpopulation structures and enabled the ability to linked data markers [38]. STRUCTURE 2.0 can be used to detect admixture events further in the past, and is more accurate when linked loci are used. STRUCTURE 2.2 [39] was developed to examine dominant markers such as amplified fragment length polymorphisms (AFLPs), null alleles, and the limitations of genotype calling in polyploids, the presence of which made many conventional analysis methods invalid for many organisms. STRUCTURE 2.3 [40] used new models for both admixed and non-admixed cases, which improved performance in dealing with lower levels of divergence and less data than previous methods. Moreover, by taking advantage of sample group information, the new models in STRUCTURE 2.3 were not biased in detecting false structures.

The newest version of STRUCTURE is fastSTRUCTURE [41]. Raj *et al.* developed efficient algorithms for approximating inference in models underlying the STRUCTURE program using a variational Bayesian framework. The variational algorithms are nearly two orders of magnitude faster than STRUCTURE, which allows it to quickly infer population structures in large datasets. In addition, fastSTRUCTURE uses heuristic scores to identify the number of populations in a dataset, and uses a new hierarchical prior to capture weak structures in populations. The heuristic scores provide a reasonable range number of populations presented in the data, minimizes bias in detecting structures especially when the structure are very weak.

FRAPPE (FRequentist APProach for Estimating individual ancestry proportion) [42], ADMIXTURE [43] and sNMF (sparse nonnegative matrix factorization) [44] software programs use the maximum likelihood (ML) estimation methods to infer population structures. The ML method is computationally faster than the MCMC method commonly used by STRUCTURE series methods [36,42]. FRAPPE is commonly used to estimate individual admixture and allows for uncertainty in ancestral allele frequencies. The full ML method used in FRAPPE demonstrates increased robustness compared to partial

ML approaches, and it is as efficient as Bayesian methods, while requiring only a fraction of the computational time to produce point estimates. As such, FRAPPE and the ML method allow for extensive analysis that cannot be achieved by using Bayesian methods.

ADMIXTURE [43] uses model-based estimations of ancestry in unrelated individuals, and utilizes the likelihood model embedded in STRUCTURE. ADMIXTURE is considerably faster than both STRUCTURE and FRAPPE, and is nearly as fast as EIGENSTRAT, a non-parametric approach. Simulations also show that ADMIXTURE has greater accuracy than FRAPPE, and is as accurate as STRUCTURE's estimates. The computational speed of ADMIXTURE makes it possible to use thousands of individuals with hundreds of thousands of markers in model-based ancestry estimation, and it is suitable for correcting population stratification in association studies.

Snmf [44] was developed in 2014 and uses sparse nonnegative matrix factorization algorithms. Without loss of accuracy, the runtimes of sNMF in computing estimates of ancestry coefficients are ~10–30 times shorter than those of ADMIXTURE.

LOCAL ANCESTRY INFERENCE

Global ancestry inference provides an in-depth analysis of each ancestral population. Local ancestry inference is necessary to investigate admixture on a fine scale. The genomes of admixed individuals can be described as mosaics of different ancestries [45]. Inferring the ancestral origin of chromosomal segments in admixed individuals is useful for studies on human evolutionary history and genetic association studies. Most methods of local ancestry inference are based on hidden Markov models (HMM), where the hidden states correspond to ancestral populations and generate the observed genotypes. Early approaches to local ancestry inference [39,46,47] based on the STRUCTURE framework made use of HMMs and did not explicitly model background LD. These methods assumed that, given the ancestry, the sampled markers were unlinked. An analysis by Tang *et al.* [48] showed the importance of accounting for background LD and proposed a Markov-hidden Markov model (MHMM), which is utilized in SABER software program. Moreover, an expectation maximization (EM) algorithm based on HMM was proposed for inferring local ancestry using continuous gene flow (CGF) model. LD patterns were compared with CGF model and intermixture admixture (IA) model [49].

HAPMIX [50] allows for a more comprehensive account of background LD (i.e., LD within the ancestral population) over longer segments; however, it only considered two ancestral populations at a time. As such,

HAPMIX is useless in estimating the ancestry of Latino and Hispanic populations, such as Mexicans and Puerto Ricans, because they are three-way admixed [51]. HAPAA (HMM-based analysis of polymorphisms in admixed ancestries) [52], ChromoPainter [53], SWITCH-MHMM [54], MULTIMIX [55], and ALLOY [56] are all capable of analyzing admixtures of more than two populations.

Guan [57] presented a two-layer HMM to detect the structures of haplotypes for unrelated individuals. This method models two scales of LD (one within a group of haplotypes and one between groups) and takes advantage of rich haplotype information to infer the local ancestry of admixed individuals. Lanc-CSV [58] is a new method for local ancestry inference that leverages continent-specific variants (CSVs) to attain increased performance over existing approaches in sequenced admixed genomes. As opposed to most previous local ancestry methods that require phased reference panels, this approach only requires allele frequency information for each continental group. This method was developed to deal with genome-wide sequencing data, and it is fast and efficient.

An alternative approach to local ancestry inference is the statistical learning algorithm. LAMP (Local Ancestry in adMixed Populations) [59] computes the ancestry structure of overlapping windows of contiguous SNPs and assigns ancestries based on a clustering algorithm known as iterated conditional modes (ICM). WINPOP [60] is an extension of LAMP that uses a refined model of recombination events and an efficient dynamic programming algorithm to infer locus-specific ancestries. This improvement is most significant when the ancestral populations are closely related. PCAdmix [61] uses a principal components algorithm (PCA) for determining ancestry along each chromosome from a high-density, genome-wide set of phased single-nucleotide polymorphism (SNP) genotypes of admixed individuals. This method first divides the genome into windows of 10 – 50 kb width and then estimates the probability of origin from particular reference panel populations using PCA. The accuracy of this method is heavily influenced by window size. SupportMix [62] uses support vector machines (SVM) to identify the putative ancestral origin of a genomic segment, which can efficiently scale for simultaneous analysis of 50 – 100 putative ancestral populations while being independent of prior demographic information. RFMix [63] is a discriminative modeling approach for rapid and robust local ancestry inference that uses a conditional random field (CRF) parameterized by random forests trained on reference panels. Efficient inference of local ancestry (EILA) [64] is a statistical method that uses fused quantile regression and *k*-means classifiers to infer the local ancestry of admixed individuals. Simulation studies show that EILA has

higher accuracy and lower variations compared to HAPMIX and LAMP when the ancestral distance is large or moderate. Table 1 shows a comparison of all of the methods of local ancestry inference.

MODELLING POPULATION ADMIXTURE

Ancestry inference allows us to determine the genetic composition of admixed populations. Based on ancestry information, numerous methods have been developed to study demographic history, genetic association study, local adaption and evolution. In this section, we explore population admixture dynamics. In broad terms, there are two models for population admixture: (i) the demographic admixture model that attempts to model demographic events that shaped present admixed populations, and (ii) the geographic migration model that attempts to model gene flow among isolated populations. The demographic admixture model assumes that source populations are isolated, and that gene flow only from source populations to admixed populations; however, the geographic migration model does not have such limitations, and any populations (or local groups of individuals) can provide or adopt gene flow. Figure 1 shows a break-down of demographic admixture models and geographic migration models.

Demographic admixture models

Hybrid isolation (HI), also known as intermixture admixture (IA) and the “immediate” admixture model, is the most popular admixture model and is widely employed in admixture inference, including admixture proportion estimation, local ancestry inference, admixture time inference, and mapping disease causing genes in admixed populations. In the HI model, individuals from two or more source populations admix in one generation and form a new randomly mating population. Afterwards, both the newly generated population and the source populations are randomly mating populations [8]. In the HI model, populations are isolated from each other after admixture occurs (Figure 1A). The continuous gene flow (CGF) model was firstly proposed as an alternative model for inferring the demographic history of African Americans [6]. In the CGF model, one of the source populations (donor populations) provides gene flow at a constant rate for every generation (Figure 1C).

The LD pattern of the CGF model is reported different from that of the HI model and simulation studies support that using the HI model on CGF-like admixed populations causes high false-positive rates in admixture mapping [6,49]. The gradual admixture model [65,66] and the general admixture model [10] both describe scenarios in which admixed populations are generated by several

Table 1. A comparison of methods for local ancestry inference.

| Methods | Applicable to more than two populations | Key technique | Model background LD | Phased ancestral data |
|-----------------------------|---|--------------------------------------|---------------------|-----------------------|
| STRUCTURE [37,38] | YES | HMM | NO | NO |
| SABER [48] | YES | MHMM (first-order Markov HMM) | YES | YES |
| HAPAA [52] | YES | HMM | YES | YES |
| LAMP [59] | YES | A window-based method | NO | NO |
| SWITCH and SWITCH-MHMM [54] | YES | MHMM | YES | YES |
| WINPOP [60] | YES | A window-based method | NO | NO |
| HAPMIX [50] | NO | HMM | YES | YES |
| PCAdmix [13] | YES | PCA | NO | NO |
| ChromoPainter [53] | YES | HMM | YES | YES |
| SupportMix [61] | YES | SVM (support vector machine) | NO | YES |
| MULTIMIX [55] | YES | HMM | YES | NO |
| ALLOY [56] | YES | FHMM (factorial hidden Markov model) | YES | YES |
| RFMix [63] | YES | CRF (conditional random field) | NO | YES |
| EILA [64] | YES | <i>k</i> -means | NO | NO |
| EILA [57] | YES | Two-layer hidden Markov model | YES | NO |
| Lanc-CSV [58] | YES | HMM | NO | NO |

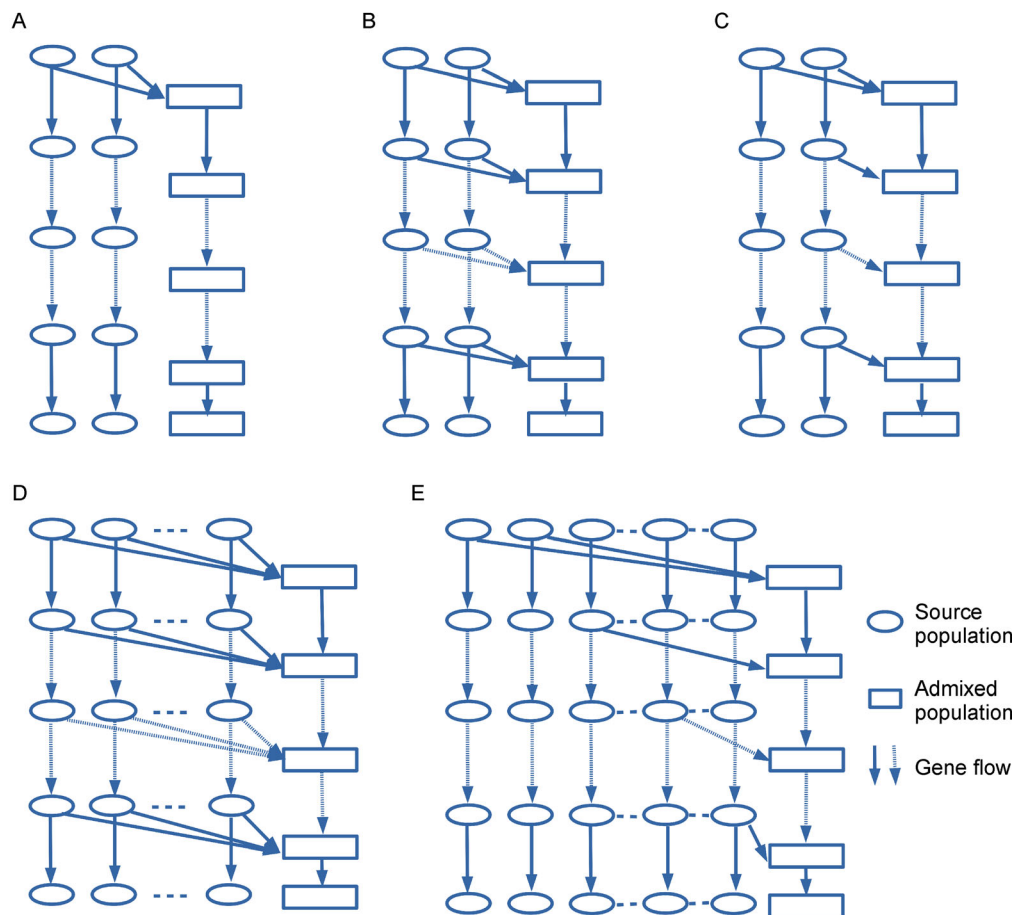


Figure 1. Geographic admixture models. (A) HI model; (B) GA model; (C) CGF model; (D) the general admixture model; (E) Pickrell *et al.*'s multiple-wave model. The ellipse disks represent the source populations, the rectangle represents the target admixed population and the arrows represent the direction of gene flow.

source populations, and gene flow between populations can occur any time after the initial admixture (Figure 1B). In this review, we use GA to refer to a modified gradual admixture model in which there are only two source populations and constant gene flow rates occur in each generation [23] (Figure 1B).

Pickrell proposed a model similar to the general admixture model that considers multiple sources involved in several admixture events [20]. Unlike the general admixture model, only one source population is allowed to provide gene flow in each wave of admixture (Figure 1E). Recently Zhou *et al.* extended the GA and CGF models to create the GA-I and CGF-I models that assume isolation after a period of continuous admixture (Figure 2).

Geographic migration models

In the n -island model, devised by Sewall Wright [67], a metapopulation is split into several islands, or discrete

subpopulations of equal population size N , and the migration rates between any groups of subpopulations are constant (Figure 3A). Given the geographical substructure, the stepping-stone model is the only way to determine migrations between adjacent discrete populations [68] (Figure 3B), which decreases the complexity of the n -island model. Another model describing the migration between discrete populations is the source-sink model [69] (Figure 3C), in which directed gene flow occurs between the source population (the donor population) and the sink population (the population that receives gene flow from the donor population).

The isolation by distance (IBD) model [70] describes migrations within in a spatially continuous population. Unlike the migration models of discrete subpopulations, the IBD model assumes that mating probability decreases as the geographic distance increases. Based on this view, the IBD model acts as a continuous version of the stepping-stone model described above.

Genetic data often exhibits patterns that are broadly

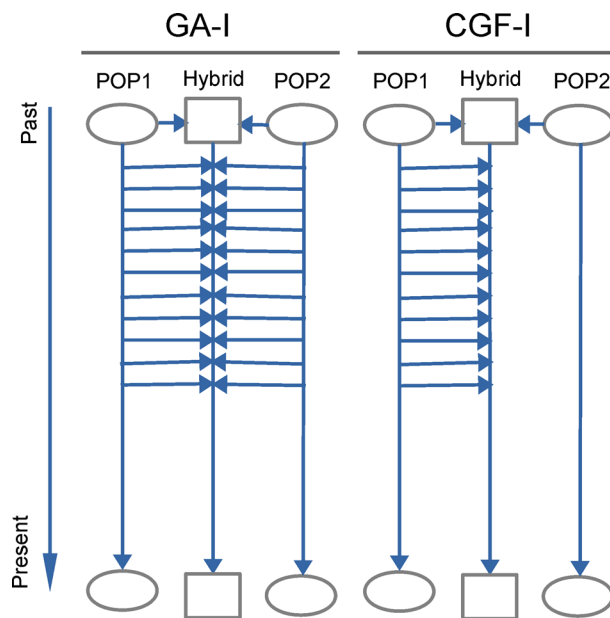


Figure 2. Extended GA and CGF models.

consistent with the IBD model. For instance, genetic similarity tends to decay with geographic distance; however, the decay rate can be heterogeneous. For example, barriers (including extrinsic factors, such as topography and other environmental factors, and intrinsic factors such as mate recognition and reproductive compatibility) to gene flow can accelerate the decay rates between groups located close together in space, leading to some degree of “structure”. PCA [71] is a method commonly used for analyzing population structures. PCA summarizes the main patterns of population structure in explicit visual representations. PCA projections are often interpreted post hoc with geographic information in hand, while PCA itself ignores the geographic information of samples even if they are known. On the other hand, the estimated effective migration surface (EEMS) [72] model, based on the “stepping stone” model, uses both genetic and geographic information from samples to highlight regions deviating from typical isolation by distance patterns, thus identifying genetic barriers to gene flow, if they exist. In order to capture continuous population structures, EEMS uses a

dense regular grid of demes (the number of which can be set by users) spread across a habitat. Gene flow exists only between samples within neighboring demes. Expected genetic dissimilarity between a pair of demes can be calculated by integrating over all possible migration histories in the genetic ancestry of its samples. If the observed genetic similarity decays faster than expected in some parts of the space, it is reflected by a lower value of EEMS in those areas. Therefore, EEMS provides a visual summary of the observed genetic dissimilarities among samples, and how they are related to geographic locations.

DATING ADMIXTURE

In the previous section, we introduced two types of models to describe the various types of population admixture. One of the applications of these models is to understand the admixture history of the target population [18–20,22–25,50,53,73–75]. One of the major benefits of genotyping technology is that high quality genetic data from hundreds of populations are available for studying global patterns of population migrations and population admixture [76,77]. In an admixed genome, haplotypes can be regarded as the chromosomal mosaics of source populations (Figure 4). Consider the ancestry for each genetic marker to be classified, and that ancestral chromosomal tracks (ACTs), sites continuously share the same ancestry and can be measured. Additionally, the number of ancestral switches (ASs), the pairs of sites from different ancestral populations, can be counted. Both ACT and AS convey the basic information of admixture patterns and admixture times. In this section, we discuss the various methods of using this information alongside genetic data to infer admixture history.

Dating admixture based on the ancestral switches

The number of ASs is a function of admixture time regarding constant recombination rates. The longer the admixture history is, the more recombination events accumulate and the higher the number of ASs. Under the two-way HI model, the admixture time can be estimated from the number of observed ASs (n_{AS}), the admixture proportion (m) and the genome wide recombination rate (L recombination per generation) [50,73],

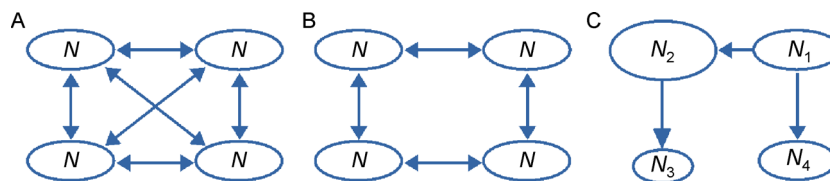


Figure 3. Geographic migration model. (A) The n -island model; (B) The stepping-stone model; (C) The source-sink model. The ellipse disks represent populations, the disk size represents the relative population size, and the arrow represents the direction of the gene flow.

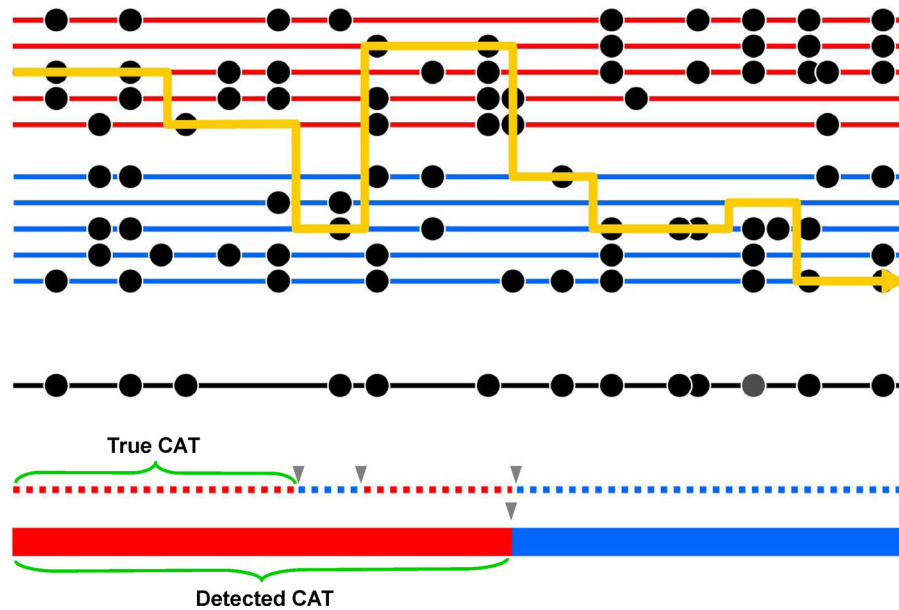


Figure 4. The mosaic of the admixed genome, modified from Refs. [50,73]. The black lower line represents a haplotype segment from an admixed individual, carrying a number of mutations (black pie). The red and blue lines represent the haplotypes from two different source populations are in red and blue lines, respectively. The yellow line shows how the admixed haplotype is constructed by the haplotype segments from source populations. The dashed line shows the true ancestry for each site on the admixed haplotype, and the colored horizontal bar represents the possible inferred ancestry. ASs are also represents by gray arrows. There are three ASs but only one can be inferred.

$$T = \frac{n_{AS}}{2m(1-m)L}.$$

Based on the idea that the admixed chromosomes can be modeled as mosaics of haplotype segments of source populations, a probability model can be built to determine the parameters that best explain the observed admixed haplotypes within the source haplotypes. In this way, the ancestry switch rate can be inferred, as well as the admixture time [50]. GLOBETROTTER extends the ancestry switch rate to the pairwise probability of two sites from the specified ancestries. By fitting the co-ancestry curves against to the distance between the two sites, admixture time can be inferred. ChromePainter can be used for local ancestry inference with multiple source populations, and it can deal with multiple-way and multiple-wave admixture [24,53].

Dating admixture based on continuous ancestry tracts

Another result of recombination accumulation is shortened ACT length. Short ACTs always indicate a long admixture history, whereas long ACTs tend to indicate more recent gene flow from the ancestral populations to which the ACT belongs. Under the HI model, the density function of the ACT from the source population who comprises the m proportion of the admixed genome is

$$f(x; T) = (1-m)Te^{-(1-m)Tx},$$

where T is the admixture time and x is a variable representing the length of the ACT [22,25,78]. Because the length distribution of the ACT is varies for different demographic admixture models, the ACT length can be collected to find the best-fit model for the admixture history [21–23,25,26,78,79].

Pool and Nielsen first used the length of ancestral tracks to infer population history [26]. They introduced a theoretical framework that described the length distribution of ancestral tracts and proposed a likelihood inference method for estimating parameters related to historical change in migration rates. Additionally, Pugach *et al.* introduced a two-part method for inferring admixture history. The first part of the method, called StepPCO, is an extension of PCA and is used to obtain a signal of admixture from an individual genome. The second part of the method relies on the wavelet decomposition of the admixture signal to extract information about the date of the admixture event [21,79]. Jin *et al.* further explored admixture dynamics by comparing the empirical and simulated distributions of ancestral tracks under three typical two-way admixtures models, i.e., the HI model, the GA model, and the CGF model [23]. Jin *et al.* later deduced the theoretical distributions of ancestral tracks under the HI and GA

models [22]. Gravel extended these studies to multiple ancestral populations and discrete migrations, and developed a method called Tracks to infer admixture history [25]. For all these methods, a prior admixture model is required. These methods must be input into an admixture model, and then estimated to determine the parameters of the specific model; however, in data analysis, there is typically very little information on admixture history, and the admixture model is often uncertain in complex admixed populations. Recently, Ni *et al.* introduced model selection into admixture history inference. They first developed a method called *AdmixInfer* [78] to infer admixture history under three typical two-way admixtures models. They then developed a new method called *MultiWaver* [80] to explore multiple-wave admixture histories. Their methods can automatically determine an optimal admixture model based on the length distribution of ancestral tracks, and estimate the corresponding parameters.

Dating admixture based on weighted linkage disequilibrium (LD)

Pairwise LD describes the non-random dependence between markers along chromosomes at the population level. Admixture produces high levels of LD at loci that have different allele frequencies among the involved populations. After admixture occurs, the admixture LD decays at the rate of $1 - d$, where d is the genetic distance (or recombination rate) between two sites [81]. As such, if we know the LD value when the admixture occurred and the LD value in the present admixed population, we can trace the admixture history. Under the HI model, LD in the admixed population after T generations of isolation is

$$D^{(t)} = D^{(0)}(1 - d)^T,$$

where $D^{(0)} = m(1 - m)\delta(x)\delta(y)$, and δ is the allele frequency difference between the two source populations at the site x or site y [82]. This is the basic idea for LD-based admixture time inference algorithms. Moorjani *et al.* firstly applied this idea by aggregating pairwise LD measurements through a weighting scheme [75]. Patterson *et al.* [18] developed the software *rolloff* for this purpose, which was further developed by Loh *et al.* [19] and by Pickrell *et al.* [20]. When studying the genome wide pattern of LD, the sign of LD value must be seriously considered because it can be affected by the coding rules [83]. In *rolloff*, the weight on each site ($w(x)$) is defined as the sign of $\delta(x)$ to correct for the coding effect on LD. In the software program ALDER [19], the weight is directly defined as $\delta(x)$, which signifies the site where source populations are highly differentiated. In this way, the weighted LD statistic is defined as the average of

the weighted LD on the set holding pairs of sites whose genetic distance are similar:

$$a(d) = \frac{\sum_{S(d)} D\delta(x)\delta(y)}{|S(d)|},$$

where $S(d) = \{f(x, y) : d - \varepsilon/2 < |x - y| < d + \varepsilon/2\}$ and ε is the discretization parameter inducing a discretization on d . Then $a(d) = a_0(1 - d)^T$, where a_0 is a constant. Admixture time can be estimated by fitting the decay curve of $a(d)$. The latest version of ALDER can do admixture time inference based on multiple-way and multiple-wave model [20], which is probably the most powerful time inference method available because it can provide the source populations for each wave of admixture. Admixture LD is composed by two parts: (i) LD directly inherited from source populations (SLD) and (ii) LD produced by admixture. Using weighted LD can reduce the effects of SLD. Moreover, most of these algorithms use starting distance to confirm that the SLD's effect is small enough. An alternative method (using software such as iMAAPs) to reduce the SLD's effect is to use the reference populations. Zhou *et al.* illustrated out that the SLD can be estimated by proper reference populations, as well as the weighted SLD [84].

Dating admixture under different models

Dating results are highly dependent on the models being used. Jin *et al.* pointed out that the mean length of ACT in the HI model is approximately half of that in the GA model if both admixture proportions and admixture times in the two models are identical [22]. In recent gene flow, full length ACT is introduced and the average length of ACT increases and the average number of AS decreases. Significantly, the HI model does not consider changes brought on by recent gene flow. As a result, admixture time is under-estimated in the HI model when recent gene flow exists. The AS model can only be used to estimate the time under the HI model. The ACT based methods can model the HI, GA, and CGF models. GLOBETROTTER and *MultiWaver* are based on the model of multiple-way and multiple-wave models, and they can be used for inference under the general admixture model; however, the results are still at risk of bias due to phasing errors and local ancestry inference errors. After several years of development, weighted LD based algorithms are very robust in models that use for admixture time inference. Both *rolloff* [75] and ALDER [19] can date admixture under the HI model, and the latest version of ALDER can do admixture time inference based on multiple-way and multiple-wave models [20], which is probably the most powerful time inference method as it can determine the source populations for each wave of admixture. IMAAPs

can infer the multiple-wave admixture, but it is still under the two-way admixture model [84]. CAMer is specially designed for dating continuous admixture, and extends the continuous admixture model GA and CGF models to GA-I and CGF-I by considering isolation after admixture Figure 2 [85]. Moreover, weighted LD can be calculated based on genotype data [86], so the related methods naturally avoid the risk of phasing errors and local ancestry inference errors.

ANALYSIS OF ARCHAIC INTROGRESSION

Apart from admixture among modern human populations, it is widely believed that introgression from archaic hominids to modern humans occurred during period of coexistence. Many studies have found evidence for “adaptive archaic introgression” [87–92]. Since the advent of whole genome sequencing, evidence has emerged that there was admixture between humans and two archaic hominids, Neanderthals [33,93,94] and Denisovans [95,96]. In recent years, diverse methods have been proposed to identify introgressive sequences in present-day human populations. Since each pair of populations share some DNA segments derived from their common ancestry, an effective method must distinguish true introgression from shared ancestry [97].

D statistics is an applicable method for detecting archaic introgression using genome-wide data [27]. Nevertheless, using D statistics locally to identify whether a specific segment is introgressed from an archaic hominid may not be very accurate. Some other approaches take advantage of sequence divergence to calculate time to the most recent common ancestry (TMRCA), which can be used to infer archaic introgression locally [92]. A DNA sequence that has a more recent TMRCA with archaic hominids than a modern humans is regarded as a product of introgression.

A method based on S^* statistics provides a way to detect extremely distinguished haplotypes that extract LD information without reference populations [98–100]. S^* statistics employs a rule to score each pair of SNPs on a haplotype in order to seek a subset of SNPs with maximum sum of scores segments that indicates the existence of a strong LD. A best fit model under which simulated data has similar values of features with real data is required to eliminate segments of modern humans with strong LD.

Model-based methods under the HMM framework similar to normal local ancestry inference for modern humans have also been employed [33,87,101]. Each SNP can be labeled statistically by one of two hidden states, archaic and modern humans, with the largest likelihood to produce observed data. Prior parameters are applicable in

HMM frameworks due to large deviations between archaic humans and modern humans [33,87].

ArchaicSeeker [32] is a more heuristic method for detecting archaic DNA sequences in present-day human genomes. The basic idea of the method is based on an archaic gene flow model, i.e., for an introgressed segment in modern humans, its haplotype divergence from African sequences would be larger than its divergence from the archaic hominin donor (see Figure 5A). DNA sequences from modern humans are directly compared with archaic genomes. African sequences are used as a reference since archaic hominin introgressions are absent in sub-Saharan Africans [102]. This method is based on haplotype data and all of the data involved in the analysis are assumed to be phased (haplotypes are known).

In real data analysis, E -allele information [87,88,103] (markers can be only observed in non-African populations) is used to determine the boundaries of the archaic segments and false positive results are removed by filtering segments without E -alleles or with only one E -allele. Figure 5B displays how one candidate archaic segment (the long red segment without any detected E -alleles) is removed and the other (left, with three E -alleles covered) is kept within boundaries determined by the location of the two leftmost and rightmost E -alleles. As a general rule, the locations of the first and the last E -alleles in those segments with more than two E -alleles are treated as the left and right boundaries, respectively.

Incomplete lineage sorting

In the analysis of archaic introgression (or more generally, ancient admixture), some confounding effects need to be considered and controlled, such as incomplete lineage sorting (ILS). ILS, also known as deep coalescence or ancestral polymorphism, is the result of the retention of a genetic polymorphism along several speciation or divergence events. The posterior sorting of polymorphic lineages can make gene and population trees incongruent [104]. Having been well studied at the species level [104–106], ILS has not been thoroughly investigated within human populations, especially in admixed populations [50,107]. In the formula of the probability of ILS between human and chimpanzees provided by previous studies [108], the shorter time between three species (populations) and two species (populations) and the larger the sample size of the two species’ (populations’) common ancestor, the more likely ILS occurred.

Because global ancestry inference often considers genome-wide information, the results are unlikely to be influenced by ILS. With for local ancestry inference with reference populations, ILS affects the results of ancestry inference and admixture analysis (Figure 6). Directly

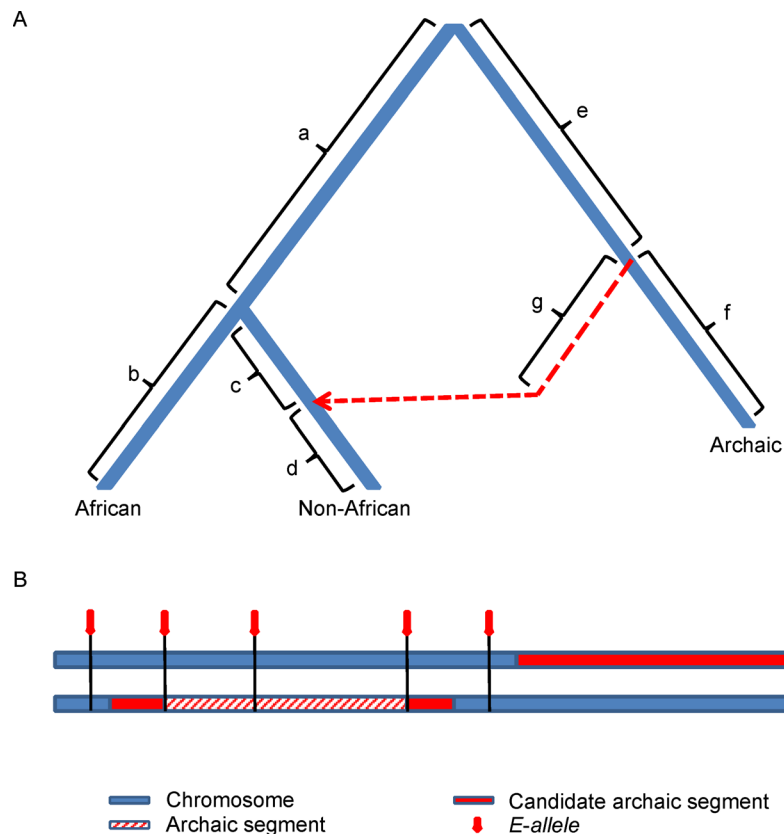


Figure 5. A schematic of analysis design of ArchaicSeeker. (A) Illustration of a phylogeny with a gene flow model from archaic hominins to modern humans. The figure illustrates the phylogenetic relationship among archaic groups, Africans and Non-Africans with archaic ancestry contribution. The characters a–g each denote the branch length (or genetic distance) as indicated. (B) Illustration of boundary refinement for the candidate archaic-like segment. First, the candidate archaic-like segments without *E-alleles* or with only one *E-allele* were filtered out (upper panel). Then, the first and the last *E-alleles* in the remaining candidate archaic-like segments with more than two *E-alleles* were regarded as the boundaries (lower panel).

comparing the admixed sequences with reference can cause inference errors.

Significantly, most of the local ancestry inference methods do not take ILS into consideration. The reason could be, for continental admixed populations, that ILS did not frequently occur since the effective population size of the common ancestors of modern humans is smaller compared to the passage of time. Although ILS can be considered a nuisance that should be treated carefully, it can facilitate the inferences. To solve this problem, a direct method is to model ILS into the algorithm. For example, HAPMIX [50] considers the ancestries as two parts, “real” ancestry and “copied” ancestry. The former stands for the sequence similarity, while the latter means from which ancestral population the haplotype comes from. A miscopy rate parameter is used to account for the differences between “real” and

“copied” ancestry. In HAPMIX, an EM algorithm is used to estimate those parameters for African Americans [109].

In archaic introgression analysis, the ILS test is necessary for regions with high introgression proportions [34,107]. There are several methods that can be used for testing ILS. One method is to examine the segment length. The ILS segments have a deep coalescence and are expected to be very short due to recombination. The introgression segments are expected to be significantly longer than segments resulting from ILS. With the ILS length distribution, the *P*-value of a candidate introgression segment can be obtained [107]. In addition, the divergence time between introgression segments and archaic hominids is expected to be smaller than that between modern humans and archaic hominids. The other method to exclude the ILS effect is using a “null model”, a model that shares all the demographic parameters by all of

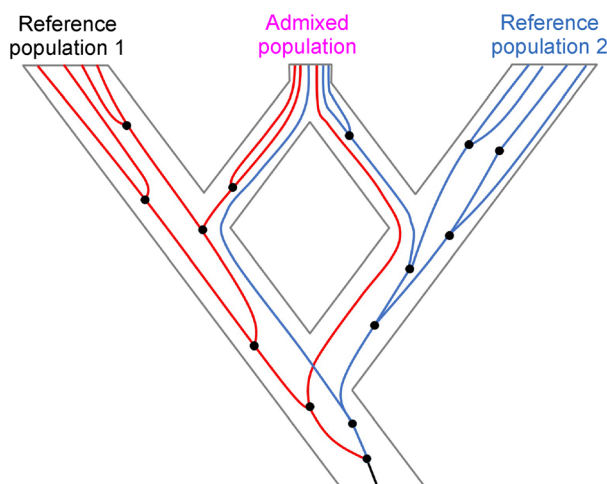


Figure 6. Incomplete lineage sorting in an admixed population. The figure shows a population tree (in grey) and a gene tree (in blue and red) tracking the evolutionary history of two ancestral populations and their corresponding admixed population. Due to ILS, specific haplotypes of reference population 1 (red lines) could flow into and admixed with population 2 and vice versa.

the real populations, except introgression, to control for ILS and gene flow [34]. When dating admixture with continuous ancestry tracts, ILS can lead to an over-estimation of admixture time, and determination of admixture models can also be affected.

PERSPECTIVE

In this review, we provided an overview of models, methods, and tools for ancestry inference and admixture analysis. Recent advances in genotyping and sequencing technologies have facilitated genome-wide investigations of human genetic variations and provided new insights into population structures and admixture history. Admixed populations are attracting more and more attention from both evolutionary and medical studies as well as from the other fields. Without thoroughly understanding the genetic structures and history of admixed populations well, our knowledge about human genetics will remain incomplete. Further efforts are needed to reveal local adaptation signatures, and to apply admixture mapping in many admixed populations. In the future, more powerful methods are expected to be developed and applied to admixed populations with longer histories and more complex admixture scenarios.

ACKNOWLEDGEMENTS

S.X. acknowledges financial support from the National Natural Science Foundation of China (NSFC) grant (Nos. 91331204 and 31711530221), the

Strategic Priority Research Program (No. XDB13040100) and Key Research Program of Frontier Sciences (No. QYZDJ-SSW-SYS009) of the Chinese Academy of Sciences (CAS), the National Science Fund for Distinguished Young Scholars (No. 31525014), and the Program of Shanghai Academic Research Leader (No. 16XD1404700); S.X. is Max-Planck Independent Research Group Leader and member of CAS Youth Innovation Promotion Association. S.X. also gratefully acknowledges the support of the National Program for Top-notch Young Innovative Talents of The “Wanren Jihua” Project. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Kai Yuan, Ying Zhou, Xumin Ni, Yuchen Wang, Chang Liu and Shuhua Xu declare that they have no conflict of interest.

This article is a review article and does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Chakraborty, R. and Weiss, K. M. (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. USA*, 85, 9119–9123
2. McKeigue, P. M. (1997) Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *Am. J. Hum. Genet.*, 60, 188–196
3. McKeigue, P. M. (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am. J. Hum. Genet.*, 63, 241–251
4. Montana, G. and Pritchard, J. K. (2004) Statistical tests for admixture mapping with case-control and cases-only data. *Am. J. Hum. Genet.*, 75, 771–789
5. Stephens, J. C., Briscoe, D. and O’Brien, S. J. (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am. J. Hum. Genet.*, 55, 809–824
6. Pfaff, C. L., Parra, E. J., Bonilla, C., Hiester, K., McKeigue, P. M., Kamboh, M. I., Hutchinson, R. G., Ferrell, R. E., Boerwinkle, E. and Shriver, M. D. (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.*, 68, 198–207
7. Long, J. C. (1991) The genetic structure of admixed populations. *Genetics*, 127, 417–428
8. Ewens, W. J. and Spielman, R. S. (1995) The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.*, 57, 455–464
9. Parra, E. J., Kittles, R. A., Argyropoulos, G., Pfaff, C. L., Hiester, K., Bonilla, C., Sylvester, N., Parrish-Gause, D., Garvey, W. T., Jin, L., *et al.* (2001) Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *Am. J. Phys. Anthropol.*, 114, 18–29
10. Verdu, P. and Rosenberg, N. A. (2011) A general mechanistic model for admixture histories of hybrid populations. *Genetics*,

- 189, 1413–1426
11. Guo, W. and Fung, W. K. (2006) The admixture linkage disequilibrium and genetic linkage inference on the gradual admixture population. *Yi Chuan Xue Bao* (in Chinese), 33, 12–18
12. Zakharia, F., Basu, A., Absher, D., Assimes, T. L., Go, A. S., Hlatky, M. A., Iribarren, C., Knowles, J. W., Li, J., Narasimhan, B., *et al.* (2009) Characterizing the admixed African ancestry of African Americans. *Genome Biol.*, 10, R141
13. Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S. A., *et al.* (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci. USA*, 107, 786–791
14. Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J., Fernandez-Lopez, J. C., Uribe-Figueroa, L., Contreras, A., Balam-Ortiz, E., del Bosque-Plata, L., Velazquez-Fernandez, D., Lara, C., *et al.* (2009) Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proc. Natl. Acad. Sci. USA*, 106, 8611–8616
15. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C. D. and Ostrer, H. (2010) Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc. Natl. Acad. Sci. USA*, 107, 8954–8961
16. Xu, S., Huang, W., Qian, J. and Jin, L. (2008) Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *Am. J. Hum. Genet.*, 82, 883–894
17. Xu, S. and Jin, L. (2008) A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *Am. J. Hum. Genet.*, 83, 322–336
18. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. and Reich, D. (2012) Ancient admixture in human history. *Genetics*, 192, 1065–1093
19. Loh, P. R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D. and Berger, B. (2013) Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193, 1233–1254
20. Pickrell, J. K., Patterson, N., Loh, P. R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B. and Reich, D. (2014) Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. USA*, 111, 2632–2637
21. Pugach, I., Matveyev, R., Wollstein, A., Kayser, M. and Stoneking, M. (2011) Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol.*, 12, R19
22. Jin, W., Li, R., Zhou, Y. and Xu, S. (2014) Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping. *Eur. J. Hum. Genet.*, 22, 930–937
23. Jin, W., Wang, S., Wang, H., Jin, L. and Xu, S. (2012) Exploring population admixture dynamics via empirical and simulated genome-wide distribution of ancestral chromosomal segments. *Am. J. Hum. Genet.*, 91, 849–862
24. Hellenthal, G., Busby, G. B., Band, G., Wilson, J. F., Capelli, C., Falush, D. and Myers, S. (2014) A genetic atlas of human admixture history. *Science*, 343, 747–751
25. Gravel, S. (2012) Population genetics models of local ancestry. *Genetics*, 191, 607–619
26. Pool, J. E. and Nielsen, R. (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181, 711–719
27. Durand, E. Y., Patterson, N., Reich, D. and Slatkin, M. (2011) Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.*, 28, 2239–2252
28. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. and Singh, L. (2009) Reconstructing Indian population history. *Nature*, 461, 489–494
29. Deng, L., Hoh, B. P., Lu, D., Fu, R., Phipps, M. E., Li, S., Nur-Shafawati, A. R., Hatin, W. I., Ismail, E., Mokhtar, S. S., *et al.* (2014) The population genomic landscape of human genetic structure, admixture history and local adaptation in Peninsular Malaysia. *Hum. Genet.*, 133, 1169–1185
30. Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T. W. Jr, Orlando, L., Metspalu, E., *et al.* (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, 505, 87–91
31. Jones, E. R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin, R. L., Gallego Llorente, M., Cassidy, L. M., Gamba, C., *et al.* (2015) Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.*, 6, 8912
32. Lu, D., Lou, H., Yuan, K., Wang, X., Wang, Y., Zhang, C., Lu, Y., Yang, X., Deng, L., Zhou, Y., *et al.* (2016) Ancestral origins and genetic history of Tibetan highlanders. *Am. J. Hum. Genet.*, 99, 580–594
33. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505, 43–49
34. Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N. and Reich, D. (2014) The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507, 354–357
35. Qin, P. and Stoneking, M. (2015) Denisovan ancestry in East Eurasian and native American populations. *Mol. Biol. Evol.*, 32, 2665–2674
36. Padhukasahasram, B. (2014) Inferring ancestry from population genomic data and its applications. *Front. Genet.*, 5, 204
37. Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959
38. Falush, D., Stephens, M. and Pritchard, J. K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164, 1567–1587
39. Falush, D., Stephens, M. and Pritchard, J. K. (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes*, 7, 574–578
40. Hubisz, M. J., Falush, D., Stephens, M. and Pritchard, J. K.

- (2009) Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.*, 9, 1322–1332
41. Raj, A., Stephens, M. and Pritchard, J. K. (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197, 573–589
 42. Tang, H., Peng, J., Wang, P. and Risch, N. J. (2005) Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.*, 28, 289–301
 43. Alexander, D. H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, 19, 1655–1664
 44. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. and François, O. (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196, 973–983
 45. Li, N. and Stephens, M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165, 2213–2233
 46. Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G. and McKeigue, P. M. (2004) Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.*, 74, 965–978
 47. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., *et al.* (2004) Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.*, 74, 979–1000
 48. Tang, H., Coram, M., Wang, P., Zhu, X. and Risch, N. (2006) Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.*, 79, 1–12
 49. Zhu, X., Cooper, R. S. and Elston, R. C. (2004) Linkage analysis of a complex disease through use of admixed populations. *Am. J. Hum. Genet.*, 74, 1136–1153
 50. Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D. and Myers, S. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.*, 5, e1000519
 51. Benirschke, K. (2002). The evolution and genetics of Latin American populations. *J Hered.*, 93, 387
 52. Sundquist, A., Fratkin, E., Do, C. B. and Batzoglou, S. (2008) Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res.*, 18, 676–682
 53. Lawson, D. J., Hellenthal, G., Myers, S. and Falush, D. (2012) Inference of population structure using dense haplotype data. *PLoS Genet.*, 8, e1002453
 54. Sankararaman, S., Kimmel, G., Halperin, E. and Jordan, M. I. (2008) On the inference of ancestries in admixed populations. *Genome Res.*, 18, 668–675
 55. Churchhouse, C. and Marchini, J. (2013) Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet. Epidemiol.*, 37, 1–12
 56. Rodriguez, J. M., Bercovici, S., Elmore, M. and Batzoglou, S. (2013) Ancestry inference in complex admixtures via variable-length Markov chain linkage models. *J. Comput. Biol.*, 20, 199–211
 57. Guan, Y. (2014) Detecting structure of haplotypes and local ancestry. *Genetics*, 196, 625–642
 58. Brown, R. and Pasaniuc, B. (2014) Enhanced methods for local ancestry assignment in sequenced admixed individuals. *PLoS Comput. Biol.*, 10, e1003555
 59. Sankararaman, S., Sridhar, S., Kimmel, G. and Halperin, E. (2008) Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.*, 82, 290–303
 60. Pasaniuc, B., Sankararaman, S., Kimmel, G. and Halperin, E. (2009) Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25, i213–i221
 61. Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G. and Bustamante, C. D. (2012) PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.*, 84, 343–364
 62. Omberg, L., Salit, J., Hackett, N., Fuller, J., Matthew, R., Chouchane, L., Rodriguez-Flores, J. L., Bustamante, C., Crystal, R. G. and Mezey, J. G. (2012) Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC Genet.*, 13, 49
 63. Maples, B. K., Gravel, S., Kenny, E. E. and Bustamante, C. D. (2013) RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.*, 93, 278–288
 64. Yang, J. J., Li, J., Buu, A. and Williams, L. K. (2013) Efficient inference of local ancestry. *Bioinformatics*, 29, 2750–2756
 65. Chakraborty, R. (1986) Gene admixture in human populations: models and predictions. *Am. J. Phys. Anthropol.*, 29, 1–43
 66. Guo, W., Fung, W. K., Shi, N. and Guo, J. (2005) On the formula for admixture linkage disequilibrium. *Hum. Hered.*, 60, 177–180
 67. Wright, S. (1990) Evolution in Mendelian populations. *Bull. Math. Biol.*, 52, 241–295
 68. Kimura, M. and Weiss, G. H. (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49, 561–576
 69. Beerli, P. and Felsenstein, J. (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. USA*, 98, 4563–4568
 70. Wright, S. (1943) Isolation by distance. *Genetics*, 28, 114–138
 71. Patterson, N., Price, A. L. and Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.*, 2, e190
 72. Petkova, D., Novembre, J., and Stephens, M. (2014) Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.*, 48, 94–100
 73. Johnson, N. A., Coram, M. A., Shriver, M. D., Romieu, I., Barsh, G. S., London, S. J. and Tang, H. (2011) Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genet.*, 7, e1002410
 74. Ni, X., Yang, X., Guo, W., Yuan, K., Zhou, Y., Ma, Z. and Xu, S. (2016) Corrigendum: length distribution of ancestral tracks under a general admixture model and its applications in population history inference. *Sci. Rep.*, 6, 26367
 75. Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A. L. and Reich, D.

- (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.*, 7, e1001373
76. Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. and Feldman, M. W. (2002) Genetic structure of human populations. *Science*, 298, 2381–2385
77. Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, 467, 52–58
78. Ni, X., Yang, X., Guo, W., Yuan, K., Zhou, Y., Ma, Z. and Xu, S. (2016) Length distribution of ancestral tracks under a general admixture model and its applications in population history inference. *Sci. Rep.*, 6, 20048
79. Pugach, I., Matveev, R., Spitsyn, V., Makarov, S., Novgorodov, I., Osakovsky, V., Stoneking, M. and Pakendorf, B. (2016) The complex admixture history and recent southern origins of Siberian populations. *Mol. Biol. Evol.*, 33, 1777–1795
80. Ni, X., Yang, X., Yuan, K., Feng, Q., Guo, W., Ma, Z. and Xu, S. (2016) Inference of multiple-wave admixtures by length distribution of ancestral tracks. *bioRxiv* 096560
81. Hill, W. G. and Robertson, A. (2007) The effect of linkage on limits to artificial selection. *Genet. Res.*, 89, 311–336
82. Chakraborty, R. and Weiss, K. M. (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. USA*, 85, 9119–9123
83. Winkler, C. A., Nelson, G. W. and Smith, M. W. (2010) Admixture mapping comes of age. *Annu. Rev. Genomics Hum. Genet.*, 11, 65–89
84. Zhou, Y., Yuan, K., Yu, Y., Ni, X., Xie, P., Xing, E. P. and Xu, S. (2017) Inference of multiple-wave population admixture by modeling decay of linkage disequilibrium with polynomial functions. *Heredity (Edinb)*, 118, 503–510
85. Zhou, Y., Qiu, H. and Xu, S. (2017) Modeling continuous admixture using admixture-induced linkage disequilibrium. *Sci. Rep.*, 7, 43054
86. Rogers, A. R. and Huff, C. (2009) Linkage disequilibrium between loci with unknown phase. *Genetics*, 182, 839–844
87. Ding, Q., Hu, Y., Xu, S., Wang, J. and Jin, L. (2014) Neanderthal introgression at chromosome 3p21.31 was under positive natural selection in East Asians. *Mol. Biol. Evol.*, 31, 683–695
88. Ding, Q., Hu, Y., Xu, S., Wang, C. C., Li, H., Zhang, R., Yan, S., Wang, J. and Jin, L. (2014) Neanderthal origin of the haplotypes carrying the functional variant Val92Met in the MC1R in modern humans. *Mol. Biol. Evol.*, 31, 1994–2003
89. Huerta-Sánchez, E., Jin, X., Asan, B., Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., *et al.* (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512, 194–197
90. Abi-Rached, L., Jobin, M. J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F. A., *et al.* (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, 334, 89–94
91. Mendez, F. L., Watkins, J. C. and Hammer, M. F. (2012) Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. *Mol. Biol. Evol.*, 29, 1513–1520
92. Mendez, F. L., Watkins, J. C. and Hammer, M. F. (2012) A haplotype at STAT2 introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am. J. Hum. Genet.*, 91, 265–274
93. Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, 328, 710–722
94. Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468, 1053–1060
95. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., *et al.* (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338, 222–226
96. Castellano, S., Parra, G., Sánchez-Quinto, F. A., Racimo, F., Kuhlweilm, M., Kircher, M., Sawyer, S., Fu, Q., Heinze, A., Nickel, B., *et al.* (2014) Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad. Sci. USA*, 111, 6666–6671
97. Racimo, F., Sankararaman, S., Nielsen, R. and Huerta-Sánchez, E. (2015) Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.*, 16, 359–371
98. Plagnol, V. and Wall, J. D. (2006) Possible ancestral structure in human populations. *PLoS Genet.*, 2, e105
99. Wall, J. D., Lohmueller, K. E. and Plagnol, V. (2009) Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.*, 26, 1823–1827
100. Vernot, B. and Akey, J. M. (2014) Resurrecting surviving Neandertal lineages from modern human genomes. *Science*, 343, 1017–1021
101. Seguin-Orlando, A., Korneliussen, T. S., Sikora, M., Malaspinas, A.-S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., *et al.* (2014) Genomic structure in Europeans dating back at least 36,200 years. *Science*, 346, 1113–1118
102. Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, 328, 710–722
103. Hu, Y., Ding, Q., He, Y., Xu, S. and Jin, L. (2015) Reintroduction of a homocysteine level-associated allele into east asians by Neanderthal introgression. *Mol. Biol. Evol.*, 32, 3108–3113
104. Mallo, D. and Posada, D. (2016) Multilocus inference of species trees and DNA barcoding. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 371, 20150335
105. Xu, B. and Yang, Z. (2016) Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, 204, 1353–1368

-
106. Mailund, T., Munch, K. and Schierup, M. H. (2014) Lineage sorting in apes. *Annu. Rev. Genet.*, 48, 519–535
107. Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., *et al.* (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512, 194–197
108. Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H. and Mailund, T. (2011) Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.*, 21, 349–356
109. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., Chen, G. K., Wang, K., Buxbaum, S. G., Akylbekova, E. L., *et al.* (2011) The landscape of recombination in African Americans. *Nature*, 476, 170–175