IJE TRANSACTIONS B: Applications Vol. 37 No. 02, (February 2024) 412-424

International Journal of Engineering

Journal Homepage: www.ije.ir

Genomic Ancestry Inference of Admixed Population by Identifying Approximate Boundaries of Ancestry Change

F. Alizadeh[a], H. Jazayeriy*[a], O. Jazayeri[b], F. Vafaee[c,d]

[a] *Faculty of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran*
[b] *Department of Molecular and Cell Biology, Faculty of Science University of Mazandaran, Babolsar, Iran*
[c] *School of Biotechnology and Biomolecular Sciences, University of New South Wales (UNSW), Sydney, Australia*
[d] *UNSW Data Science Hub, University of New South Wales (UNSW), Sydney, Australia*

PAPER INFO

Paper history:
Received 30 August 2023
Received in revised form 25 September 2023
Accepted 07 October 2023

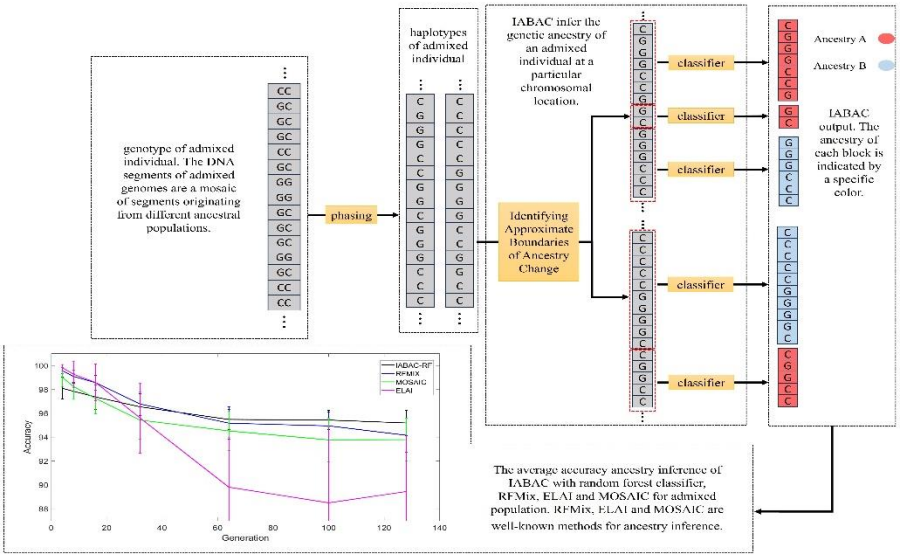
Keywords:
Admixed Haplotype
Admixed Population
Ancestry Inference
Classification
Haplotype Block
Local Ancestry

ABSTRACT

Admixture is a common phenomenon in human populations, resulting from the mating of individuals from two or more previously isolated populations. This can lead to the formation of mosaic DNA segments, with each segment originating from a different ancestral population. Local ancestry inference methods are used to identify the ancestry of each segment, which can provide insights into the history of admixture in a population. Many local ancestry inference (LAI) methods require the determination of various parameters that may be difficult to obtain, which can hamper using LAI methods. In this paper, we present a novel method for identifying approximate boundaries of ancestry change (IABAC) in admixed haplotypes and then determining the ancestry between boundaries. Unlike many LAI methods, our method does not rely on many statistical or biological parameters, therefore more robust to variations in admixture patterns. We evaluate our method on human data, and show that it is more accurate than existing methods for ancestry detection. Our results suggest that IABAC is a promising new method for identifying ancestry boundaries in admixed haplotypes. This method could be used to study the history of admixture in human populations, and to identify genetic variants that are associated with different ancestral populations.

doi: *10.5829/ije.2024.37.02b.16*

Graphical Abstract



*Corresponding Author Email: *jhamid@nit.ac.ir* (H. Jazayeriy)

Please cite this article as: Alizadeh F, Jazayeriy H, Jazayeri O, Vafaee F. Genomic Ancestry Inference of Admixed Population by Identifying Approximate Boundaries of Ancestry Change. International Journal of Engineering, Transactions B: Applications. 2024;37(02):412-24.

1. INTRODUCTION

Genetic diversity in the DNA sequences of humans is the result of inheritance processes, including mutation and recombination (1). When two or more previously isolated populations mate, the resulting offspring are admixed, meaning that their genomes contain DNA segments from both populations (2, 3). This admixture creates new genetic recombination breakpoints, which can lead to the formation of diverse genomes with mixed DNA segments. The DNA segments of admixed genomes are a mosaic of segments originating from different ancestral populations (4-10).

As travel around the world becomes easier, admixed populations and their complexity are increasing. This is because more and more people are having contact with people from other populations, which can lead to interbreeding and the formation of new admixed populations (4). Figure 1 shows that how the genomes of new populations are created from their ancestors. The chromosomes of more recent generations are a mosaic of ancestral chromosomes.

The relationship between genetic variation and disease risk can vary between ancestral populations. This is because different ancestral populations have different allele frequencies at specific genetic variants (11). Ancestry inference is the process of determining the ancestral populations that contributed to an individual's genome. This is important for a variety of applications, including pharmacogenomics and the study of human demography (2, 6, 12-14).

The availability of genotype and haplotype data has made it possible to statistically infer the admixture history of human populations (4). Several computational methods and tools have been developed for this purpose. One approach, known as local ancestry inference (LAI), identifies the ancestry of each segment of an individual's genome. LAI methods infer the genetic ancestry of an individual at a particular chromosomal location. This information can be used to study the history of human migrations and to identify genetic variants that are associated with specific ancestries. LAI has been used in a variety of models and tools, including SupportMix (15), RFMix (16), and LAMP (17).

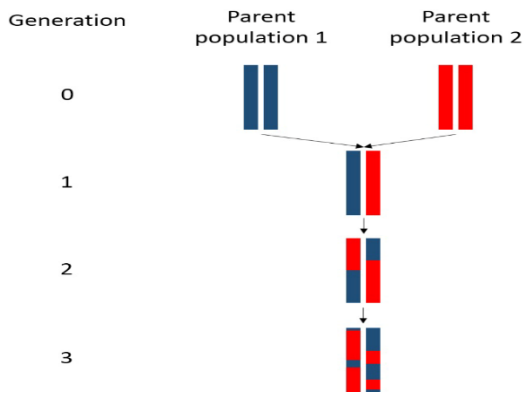


Figure 1. The process of combining the genomes of ancestral populations after generations

Local ancestry inference (LAI) methods subdivide chromosomes into smaller segments, or blocks, to infer the ancestry of each block. The choice of block size is an important factor in the accuracy of LAI. If the block size is too large, it may contain segments from multiple ancestries, which can lead to inaccurate ancestry inference. On the other hand, if the block size is too small, it may not contain enough information to accurately identify the ancestry. The ideal block size should be large enough to contain enough information to identify the ancestry, but small enough to ensure that each block contains only one ancestry (18). This can be a challenging task, as the ancestry of each block can vary depending on the individual's genetic makeup. A number of studies have investigated the optimal block size for LAI. However, the optimal block size may vary depending on the dataset and the LAI method used (17, 19).

Hidden Markov models (HMMs) are a popular approach for local ancestry inference (LAI). HMMs can model the correlation between the ancestries of blocks, which is due to linkage disequilibrium (LD) (15). LD is a phenomenon where genetic variants that are close together on a chromosome are more likely to be inherited together. Some HMM-based LAI methods are SupportMix, PCAdmix (20), MOSAIC (21) and ELAI (22). SupportMix first divides the genome into blocks with a fixed length. Then, it uses a support vector machine (SVM) to determine the ancestry of each block. PCAdmix also divides the genome into blocks with a fixed length. However, it uses a principal components algorithm to determine the ancestry of each block. RFMix determines local ancestry by using a conditional random field (CRF). A CRF is a statistical model that can model the dependencies between multiple variables. RFMix divides the genome into blocks with a fixed length. Then, it uses a conditional random field, parameterized by random forest trained on reference panels, to infer local ancestry within each block. MOSAIC uses nested HMMs to model the correlation between the ancestries of blocks. This allows MOSAIC to infer the ancestry of each segment more accurately than methods that do not use nested HMMs. ELAI employs a two-layer hidden Markov model to obtain local ancestry of each admixed individual.

XGMIX, LAI-NET, LAMP, WINPOP and EILA are also local ancestry inference (LAI) methods that use different approaches to infer the ancestry of each segment of an individual's genome. XGMIX (23) divides the genome into blocks with a fixed length and then infers

local ancestry within each window by using gradient boosting trees. The initial estimates are then smoothed using a sliding gradient boosting tree. LAI-NET (24) also divides the genome into blocks with a fixed length, but it uses a neural network model to infer local ancestry within each window. The initial estimates are then smoothed afterwards. Similarly to XGMIX and LAI-NET methods, SALAI-Net (25) follows a two-stage approach: a reference matching layer and then a smoother layer. The reference matching layer infers initial estimates of ancestry for each block, and then the smoother layer improves the initial estimates of ancestry by using neighbouring block information and smoothing ancestry. Wang et al. (26) first divides the genome into blocks with a fixed length and then clusters ancestry within each window by use of localized haplotype clustering (27). The initial estimates are then smoothed using the HMM. LAMP and WINPOP (19) first find the optimal window length on the basis of recombination events and then use a clustering algorithm to find ancestral populations. In these methods, the number of recombination events over time is considered as a Poisson distribution. EILA (28) uses fused quantile regression to identify breakpoints of the ancestral haplotypes. Then, to infer the ancestry of each segment between breakpoints, it utilizes the k-means classifier. machine learning is being increasingly used to analyze genetics data and detect diseases (29-35). As mentioned, some ancestry inference methods also use machine learning to classify haplotype and genotype data.

It is important to note that the accuracy of LAI can be improved by selecting blocks that do not contain any breakpoints. This is because breakpoints can lead to inaccurate ancestry inference (17). LAI methods require the specification of statistical or biological parameters, such as the recombination rate, genetic maps, and average number of generations since admixture. These parameters can affect the accuracy of the LAI results. Among LAI methods, RFMix has been shown to have high accuracy in inferring admixed individuals with two and three ancestral populations.

In this study, we introduce a method called IABAC (Inferring Ancestry using Boundaries of Ancestry Change) that first infers the approximate boundaries of the ancestry change based on the distance between ancestral populations. Then, IABAC identifies the ancestry between boundaries, which are called haplotype blocks. LAI methods require various parameters to be determined, which can make LAI practical uses difficult. It is usually difficult to access these parameters. For example, RFMix requires a genetic map, a window size and the average number of generations since admixture, MOSAIC requires a recombination rates files and SNP files, ELAI requires a SNP position file and the number of upper and lower layers of clusters. Unlike many LAI methods, IABAC does not required to many statistical or biological parameters. the only input parameter of IABAC is the length of the IB-block, which the optimal value of it for ancestry inference is investigated in the next sections.

SupportMix, PCAdmix, RFMix, XGMIX, SALAI-Net and LAI-NET unlike IABAC divide the genome into blocks with a fixed length and determine the ancestry of each block. LAMP and WINPOP first find the optimal block length on the basis of recombination events and then find the ancestral population between two recombination events. LAMP and WINPOP need parameters such as recombination rate to find recombination events. EILA, like IABAC, identifies boundaries of ancestry change. The main difference between EILA and IABAC is that EILA uses fused quantile regression to identify boundaries of ancestry change and IABAC uses distance between ancestral populations to identify boundaries of ancestry change. EILA is used for genotype data and IABAC is used for haplotype data.

We used four classification methods to identify the ancestry of each haplotype block: decision tree (DT), support vector machine (SVM), random forest (RF), and logistic regression (LO). We named IABAC with four different classifiers (IABACs) as IABAC-SVM, IABAC-DT, IABAC-RF, and IABAC-LO, respectively.

We compared the accuracy of the ancestry detection of admixed individuals by IABAC with the fixed window method. In the fixed window method, haplotypes were divided into blocks with a fixed length.

Finally, we compared the performance of IABACs with three well-known benchmark methods: RFMix, ELAI, and MOSAIC.

2. MATERIALS AND METHODS

In this section the research method for identifying the ancestry of admixed individuals is presented.

2. 1. Identifying Boundaries of Ancestry Change

To identify the boundaries of ancestry changes, we consider the allele frequencies of single-nucleotide polymorphisms (SNPs). This method calculates the distance between the alleles of an admixed individual by taking the mean of the alleles of the ancestral populations in a number of predefined SNPs. This distance is denoted by D. The mean of the alleles of each ancestral population in an SNP is obtained from Equation 1.

$$\mu = \frac{\sum_{i=1}^{N} h_i}{N} \quad (1)$$

In this equation, μ represents the mean of the alleles of the ancestral population per SNP. It is a value between 0 and 1. h represents the haplotype allele of each individual in each SNP. It is either 0 or 1. N represents the number of individuals in the population.

If the mean of the alleles of the ancestral population A is denoted by μ_a and the allele of the admixed individual is denoted by h_{ad}, the distance between the admixed individual allele and the mean of the alleles of the ancestral population A in a SNP can be obtained from Equation 2.

$$d_a = |(h_{ad} - \mu_a)| \quad (2)$$

In this equation, d_a represents the distance between the admixed individual allele and the mean of the alleles of the ancestral population A.

When examining the existence or non-existence of the boundaries of ancestry change between two desired SNPs, the distance of alleles in one SNP alone does not have enough information. Therefore, several SNPs need to be considered together. The information of the neighbors of the two SNPs can be used to examine the points of ancestry change between the two SNPs. For example, if we want to examine the ancestry change between two adjacent SNPs i and j, where $j > i$, the number of L_w SNPs from the left neighbor i is considered as one block, and the number of L_w SNPs from the right neighbor j is considered as another block.

In this paper, we call these blocks IB-blocks. The selection of IB-blocks is shown in Figure 2. L_w is the number of neighbors of each SNP that can be selected for different sizes. Figure 2a presents IB-blocks with SNPs i and j (SNPs in these IB-blocks are shown in the red square). In Figure 2b, position i to $i-4$ is considered as one IB-blocks and position j to $j+4$ is considered as another IB-blocks. The value of L_w in Figure 2 is 4. The top row indicates the alleles, and the bottom row indicates the location of the SNPs.

After determining the IB-blocks, the distance between the alleles of the admixed individual in each IB-block and the mean of the alleles of the ancestral population equivalent to that IB-block in each ancestral population is calculated. For the left neighbors (left IB-block) of location i, the distance is calculated using Equation 3. For the right neighbors (right IB-block) of location j, the distance is calculated using Equation 4.

$$D_{aL} = \sum_{k=-L_w}^{0} d_{a(i+k)} \quad (3)$$

$$D_{aR} = \sum_{k=0}^{L_w} d_{a(j+k)} \quad (4)$$

The length of IB-block is shown as L_w and D_{aL} represents the distance between the left IB-block of the admixed individual and the ancestral population A, and D_{aR} represents the distance between the right IB-block of the admixed individual and the ancestral population A. The following algorithm is used to check the boundaries of ancestry changes between location i and j of an admixed individual with ancestry population A and B.

Step 1: Determine the IB-block and calculate distance D between the admixed individual and the ancestral populations.

Step 2: Select the smallest distance between the ancestral populations and admixed individual for each IB-block. In Equation (5), D_{aR} represents the distance between the admixed individual and mean of the alleles of the ancestral population A in the right IB-block of location j, D_{bR} is the distance between the admixed individual and mean of the alleles of the ancestral population B in the right IB-block of the location j, D_R is the minimum distance between the admixed individual and mean of the alleles of ancestral populations A and B in the right IB-block of location j.

$$D_R = min\,(D_{aR}, D_{bR})$$
$$D_L = min\,(D_{aL}, D_{bL}) \quad (5)$$

where D_{aL} represents distance between admixed individual and mean of the alleles of ancestral population A in the left IB-block of location i, D_{bL} is distance between admixed individual and mean of the alleles of ancestral population B in the left IB-block of location i and D_L is the minimum distance between the admixed individual and mean of the alleles of the ancestral populations A and B in the left IB-block of location i.

Step 3: Select the smallest distance between the ancestral populations and admixed individual for the total IB-blocks, the blue IB-block specified in Figure 3c, the IB-blocks introduced for both location i and j are considered as one IB-block.

$$D_T = min\,((D_{aL} + D_{bL}), (D_{aR} + D_{bR})) \quad (6)$$

In this equation, D_T represents the minimum distance between the admixed individual and mean of the alleles of the ancestral populations A and B in sum of two IB-blocks of the right and left of location i and j.

Step 4: Compare sum of the minimum distance between the left and right IB-blocks specified in step 2 and the IB-block specified in step 3. If the minimum distance between the total IB-blocks and the value of D_T is not the same, the location between location i and j is selected as the boundary.

$$boundary = \begin{cases} false & D_T = D_R + D_L \\ true & D_T \neq D_R + D_L \end{cases} \quad (7)$$

The steps of determination of the existence or non-existence of the boundaries of ancestry change between locations i and j are shown in Figure 3. In Figure 3a, IB-blocks are determined for locations i and j. The alleles of

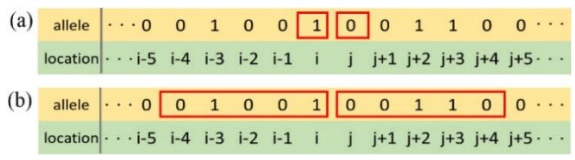


Figure 2. Investigating the existence or non-existence of the boundaries of ancestry change between locations i and j

the admixed individual are displayed with allele AD, the mean of the alleles of the ancestral population A are presented with allele mean A, and the mean of the alleles of the ancestral population B are displayed with allele mean B. In Figure 3b, the distance between the admixed individual and the ancestral population A for each IB-block is shown by the red rectangle, and the distance between the admixed individual and the ancestral population B for each IB-block is shown by the blue rectangle. The distance value for this hypothetical example is shown on the rectangle of each IB-block. An ancestral population with a smaller distance (rectangle) is selected for each IB-block.

Determination of total IB-blocks are shown in Figure 3c. This IB-block is the sum of the left neighbors for i and the right neighbors for j. In Figure 3d The distance between the admixed individual and the ancestral population A for the total IB-block is shown by the red rectangle, and the distance between the admixed individual and the ancestral population B for the total IB-block is shown by the blue rectangle. An ancestral population with a smaller distance is selected. Figure 3e presents Comparison of the sum of the minimum distance between the left and right IB-blocks specified in Figure 3b and the total IB-block specified in Figure 3d. In this example, there is no boundary of ancestry change between locations i and j, because the sum of the minimum distance between the left and right IB-blocks is equal with the total IB-block.

In the same way, the existence or non-existence of the boundaries of ancestry change between all SNPs are investigated. These IB-blocks are placed as sliding windows between all the SNPs and their distance is calculated. For example, to examine the boundary of ancestry changes between locations j and $j+1$, as shown in Figure 4b, location j to $i-3$ is considered as one IB-block (left IB-block) and position $j+1$ to $j+5$ are considered as another IB-block (right IB-block). The value of L_w in this Figure is 4.

The IB-blocks shown in Figure 4a are the IB-blocks defined to determine the boundary of ancestry changes between location i and j, the IB-blocks indicated in Figure 4b are the IB-blocks defined to determine the boundary of ancestry changes between location j and $j+1$. The left neighbors (left IB-block) of location j and the right neighbors (right IB-block) of location $j+1$ in an

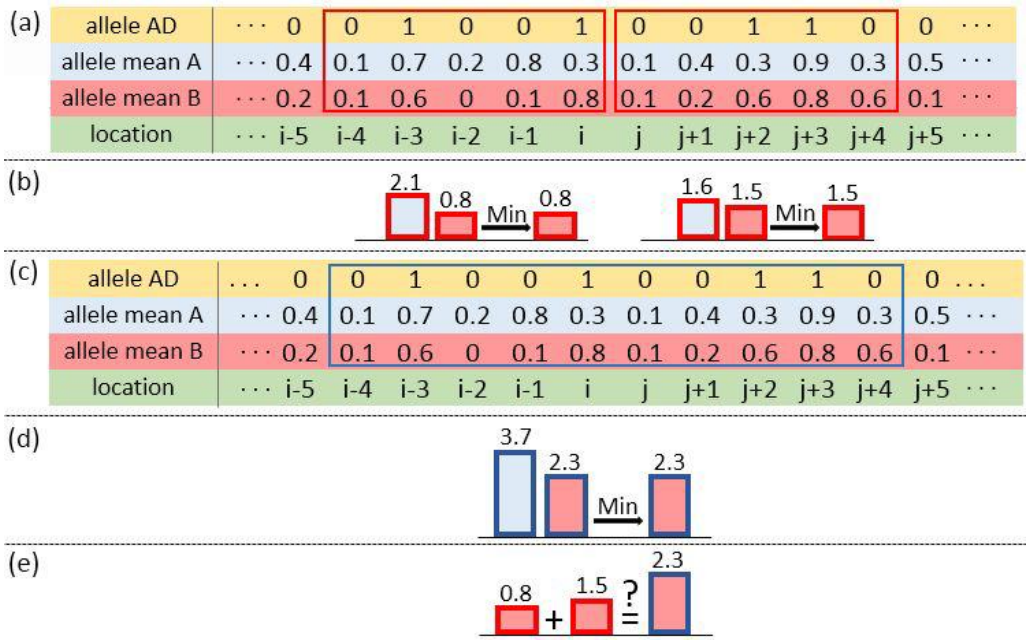


Figure 3. Determination of the existence or non-existence of the boundaries of ancestry change between locations i and j

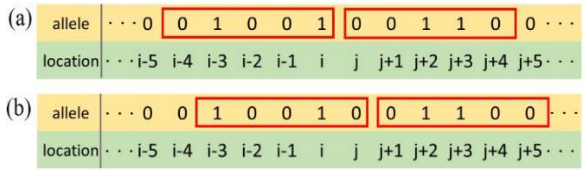


Figure 4. Investigating the existence or non-existence of the boundaries of ancestry change between locations j and $j+1$

IB-block with L_w length for the admixed individual and the ancestry A are obtained from Equations 8 and 9, respectively.

$$D_{aL} = \sum_{k=-L_w}^{0} \mu_{a(j+k)} \quad (8)$$

$$D_{aR} = \sum_{k=0}^{L_w} \mu_{a(j+1+k)} \quad (9)$$

In these equations, D_{aL} represents the distance between the left IB-block of the admixed individual and the ancestral population A, and D_{aR} represents the distance between the right IB-block of the admixed individual and the ancestral population A. Similarly, IABAC can also determine the existence or non-existence of the boundaries of ancestry change between SNPs with more than two ancestral populations.

2. 2. Classification Once the haplotype IB-blocks with appropriate length have been determined using the IABAC method, the ancestry of each of these haplotype IB-blocks must be classified. We used four well-known classification methods to do this: support vector machine (SVM), decision tree (DL), random forest (FR), and logistic regression (LR). We named the four methods as IABAC-SVM, IABAC-DL, IABAC-RF, and IABAC-LR, respectively.

2. 3. Data and Simulation We used the genotypes of chromosome 10 from the HapMap project (36), which is a database of genetic variation in humans. The genotypes were phased with SHAPEIT (37), a software that estimates haplotypes from genotype data. Admixed individuals were simulated from these haplotypes using a simple hybrid isolation (HI) model. In this model, all individuals in the first generation can mate with each other, but after that, only admixed individuals from the previous generation can mate (38).

We used data from eight populations in the HapMap project, the names of the populations and their IDs are shown in Table 1. We selected 160 unrelated samples from each population, which resulted in 160 haplotypes for each sample (80 genotypes became 160 haplotypes after phasing). We created admixed individuals by randomly mating samples from the ancestral populations. The probability of recombination in chromosome 10 for each generation was set to 1.8, based on the HapMap data (39). We simulated admixed individuals from their ancestors for 2, 4, 8, 16, 32, 64, 100, and 128 generations.

TABLE 1. populations and their IDs.

Population	ID
Northern and Western European Ancestry	CEU
Toscani in Italia	TSI
Gujarati Indians in Houston, Texas	GIH
Yoruba in Ibadan, Nigeria	YRI
Luhya in Webuye, Kenya	LWK
Maasai in Kinyawa, Kenya	MKK
Chinese in Metropolitan Denver, Colorado	CHD
Han Chinese in Beijing, China	CHB

TABLE 2. admixed populations and their ancestral populations.

Number of ancestral populations	Ancestral Population	Admixed population
Two populations	CEU, CHB	CEU-CHB
	CEU, TSI	CEU-TSI
	CEU, YRI	CEU-YRI
	CHD, TSI	CHD-TSI
	LWK, YRI	LWK-YRI
	LWK, MKK	LWK-MKK
	CHD, GIH	CHD-GIH
	GIH, TSI	GIH-TSI
	LWK, TSI	LWK-TSI
Three populations	CHD, TSI, LWK	CHD-TSI-LWK
	CHD, GIH, TSI	CHD-GIH-TSI
	MKK, GIH, LWK	MKK-GIH-LWK

From the 160 single-population individuals, 140 were selected for training and 20 were used to generate admixed individuals for testing. We simulated 20 admixed individuals by test samples from each pair of ancestral populations. We also simulated 30 admixed individuals by test samples from each triplet of ancestral populations, the names of the admixed populations and their ancestral populations are shown in Table 2.

The simulations were performed with Python and MATLAB software, and each haplotype contained 73,832 SNPs.

3. RESULTS

This section presents the results of the proposed IABAC method. Two important parameters that affect the quality of the IABAC method are the length of the IB-blocks and the choice of classification algorithm. To investigate the effect of IB-block length on the results, we performed ancestry inference for admixed individuals from the CEU-TSI, CEU-YRI, and LWK-MKK populations with different IB-block lengths.

The average accuracy of IABACs for these admixed individuals with 10 samples for each population is shown in Figure 5. Accuracy is measured by the percent of SNPs whose ancestries have been correctly identified. Admixed individuals with 32 generations from the admixed time were considered.

As shown in Figure 5, L_w with value of 100 – 300 SNPs is a good IB-block for all classifiers, and as the number of L_w gets higher or lower, the accuracy of the method decreases. The purpose of IABAC is to provide

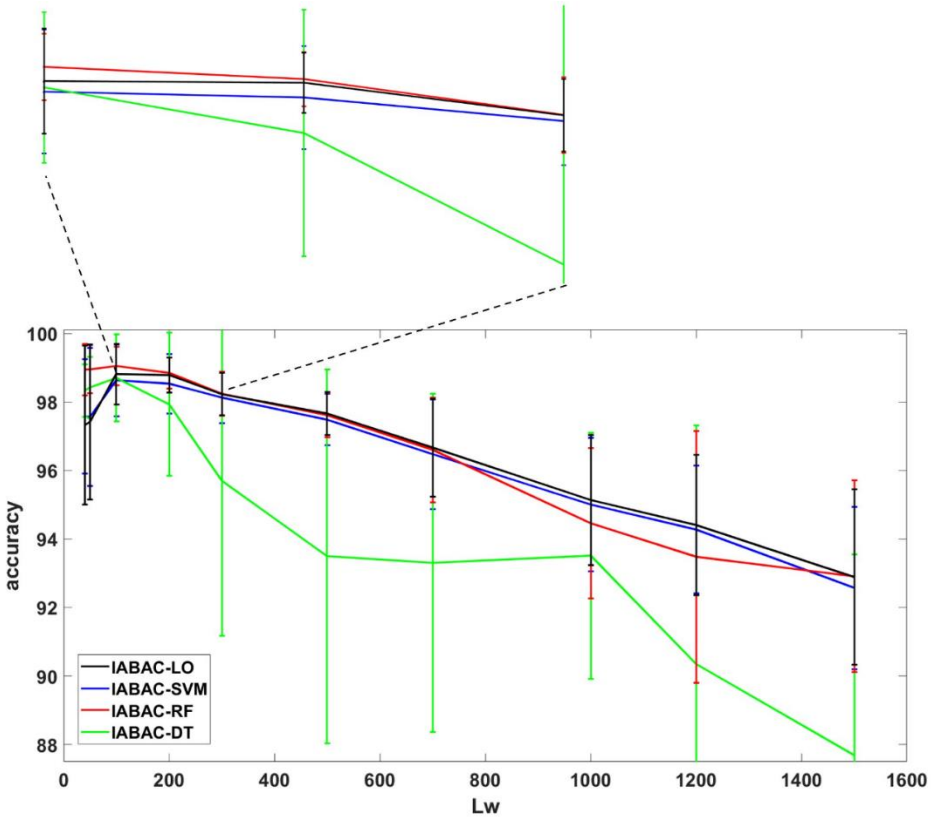


Figure 5. Average accuracy of ancestry inference of admixed individuals with CEU-TSI, CEU-YRI, and LWK-MKK ancestral population in different L_w

an appropriate way to divide chromosomes into smaller haplotype IB-blocks.

To evaluate IABAC and its effect on the accuracy of ancestry inference, the results of IABAC and fixed window with length of 500 SNPs for 20 sample from each admixed individual of CEU-CHB, CHD-TSI and LWK-YRI were compared with each other. The reason for choosing fixed window with length of 500 SNPs is that other ancestry inference methods such as XGMIX (23) and LAI-NET (24) use fixed window with length of 500 SNPs. L_w with a value of 150 SNPs were selected.

Admixed individuals with 2, 4, 8, 16, 32, 64, 100 and 128 generations from the admixed time were considered. The average accuracies of the ancestry detection of admixed individuals using IABAC-SVM and fixed window with SVM classifier (FW-SVM), IABAC-DT and fixed window with DT classifier (FW-DT), IABAC-RF and fixed window with RF classifier (FW-RF), and IABAC-LO and fixed window with LO classifier (FW-LO) are shown in Figure 6.

The results revealed that with increasing admixture times, IABAC-SVM is more accurate than FW-SVM, IABAC-DT is more accurate than FW-DT, IABAC-RF accuracy is better than FW-RF and IABAC-LO accuracy is more valid than FW-LO for ancestry inference. Overall, as the time admixture became longer, the performance of the AICRF algorithm is better than the fixed window in all four classifiers of SVM, DT, RF and LO.

To evaluate IABAC relative to other ancestry inference methods, RFMix, ELAI and MOSAIC are compared with IABACs. Figure 7 presents the average accuracy ancestry inference of IABACs, RFMix, ELAI and MOSAIC for 20 sample from each admixed population of CHD-GIH, GIH-TSI and LWK-TSI (admixed individuals with two ancestral populations). Admixed individuals with 4, 8, 16, 32, 64, 100 and 128 generations from the admixed time are considered. L_w with a value of 150 SNPs is selected.

The results indicated that in low generations (G < 32), ELAI performs better than other methods, and with increasing admixture times (G > 32), IABAC-RF accuracy is more precise than other methods.

The average accuracy ancestry inference of IABACs, RFMix, ELAI and MOSAIC for 30 sample from each admixed populations of CHD-TSI-LWK, CHD-GIH-TSI and MKK-GIH-LWK (admixed individuals with three ancestral populations) is shown in Figure 8. As shown in Figure 8, in low generations (G < 16), ELAI is more accurate than other methods, and with increasing

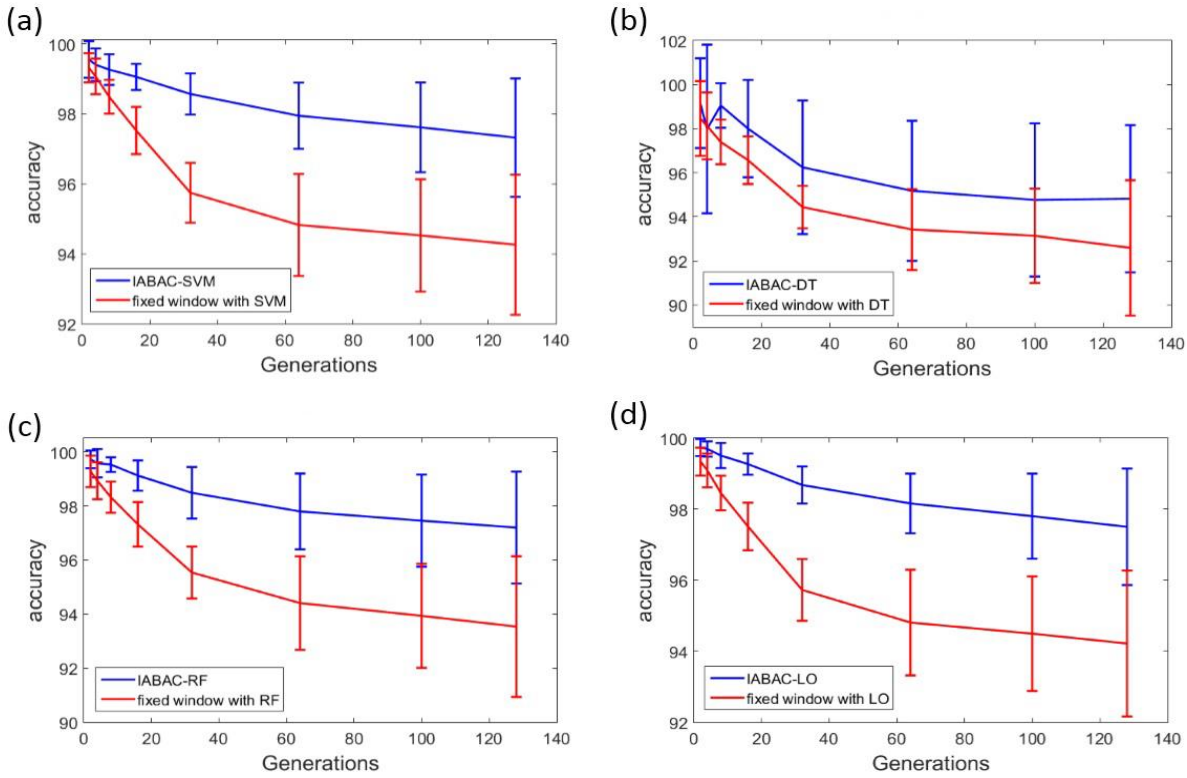


Figure 6. Average accuracy of ancestry inference of admixed individuals using IABAC and fixed window. (a) The average accuracy of ancestry inference using IABAC-SVM and FW-SVM. (b) The average accuracy of ancestry inference using IABAC-DT and FW-DT. (c) The average accuracy of ancestry inference using IABAC-RF and FW-RF. (d) the average accuracy of ancestry inference using IABAC-LO and FW-LO

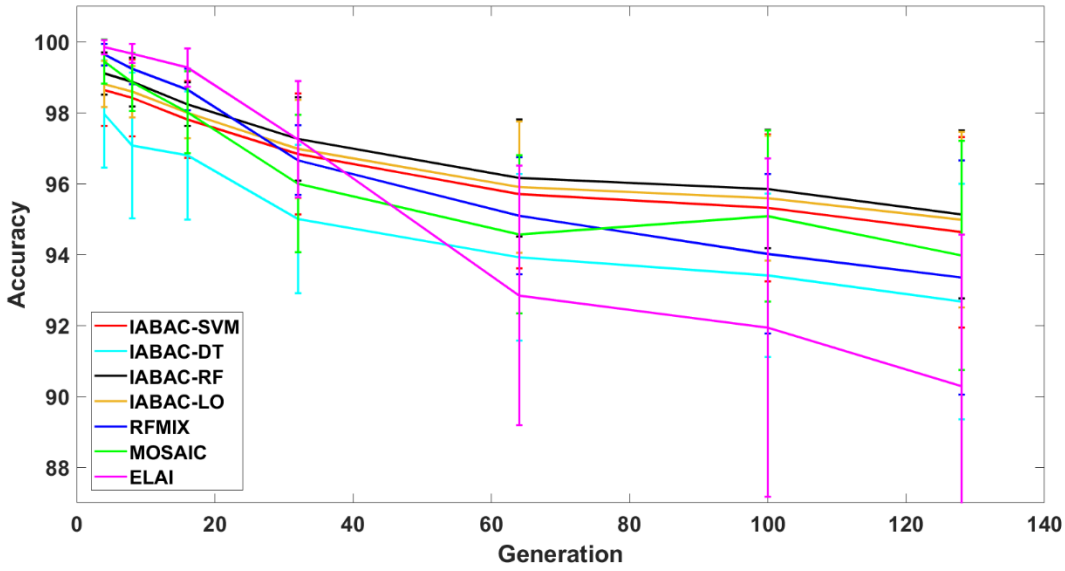


Figure 7. the average accuracy ancestry inference of IABACs, RFMix, ELAI and MOSAIC for 20 sample from each admixed population of CHD-GIH, GIH-TSI and LWK-TSI

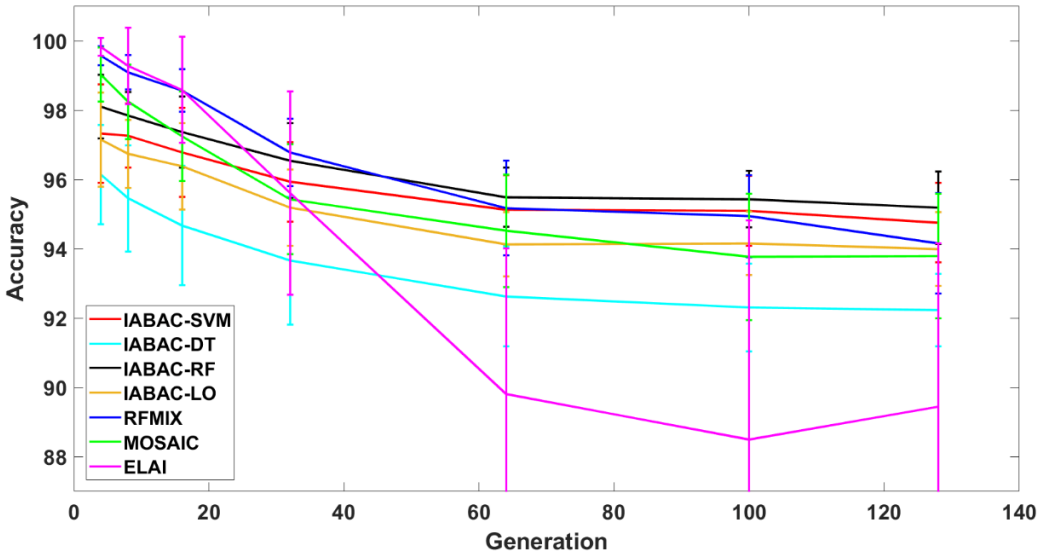


Figure 8. The average accuracy ancestry inference of IABACs, RFMix, ELAI and MOSAIC for 30 sample from each admixed population of CHD-TSI-LWK, CHD-GIH-TSI and MKK-GIH-LWK

admixture times (G > 64), IABAC-RF accuracy is better than other methods.

The results show that with increasing admixture time (G > 64), IABAC-RF can more accurately identify the ancestors of admixed individuals with two and three ancestral population.

An example of an admixed individual of CEU-CHB with its original ancestral population and estimates of IABAC-RF is shown in Figure 9. Admixed individuals with 32 generations from the admixed time are considered. L_w with a value of 150 SNPs are selected. Figure 9a is shown true ancestry of admixed individual, the red blocks represent the ancestral population of CEU and the blue blocks represent the ancestral population of CHB. Figure 9b is presented ancestry estimates of admixed individual, and Figure 9c is indicated difference between true ancestry and estimated ancestry that shown by the red block. The Y-axis represents the probability that one allele is derived from a specific ancestry and the possibility of error in any SNP; the X-axis indicates the physical locations of SNPs. The results of Figure 9 show that most errors occur at ancestry change boundaries. Additionally, some narrow ancestral mosaics have not been correctly detected.

One of the important and influential factors in the accuracy of ancestry inference using IABAC is the chromosome length of the ancestral population constituting the admixed individual chromosome (ancestral mosaics). To examine the effect of the length of the ancestral mosaics, we simulated new admixture populations of CEU-CHB and LWK-YRI. In these admixture people, the ancestral mosaics are equal in length and are repeated alternately.

The results of average accuracy of ancestry inference using IABAC-RF for 20 samples from each admixed population of CEU-CHB and LWK-YRI with different lengths of mosaics are shown in Figure 10. L_w with a value of 150 SNPs were selected.

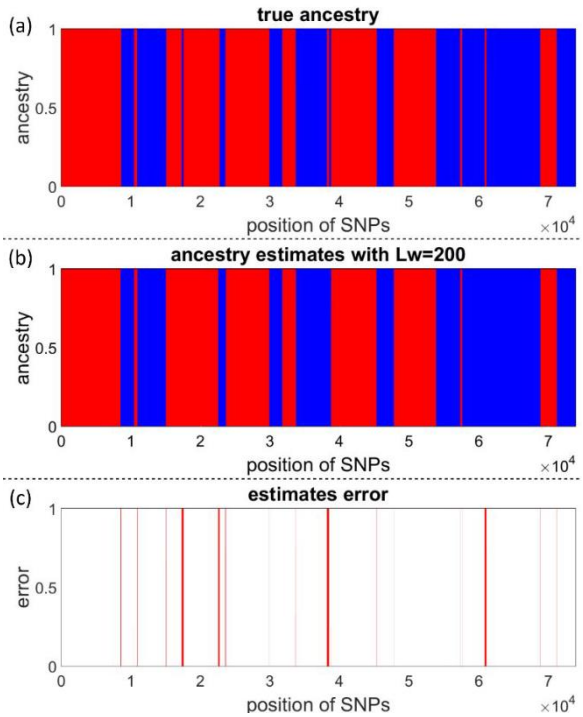


Figure 9. An example of an admixed individual of CEU-CHB with its original ancestral population and estimates of IABAC-RF

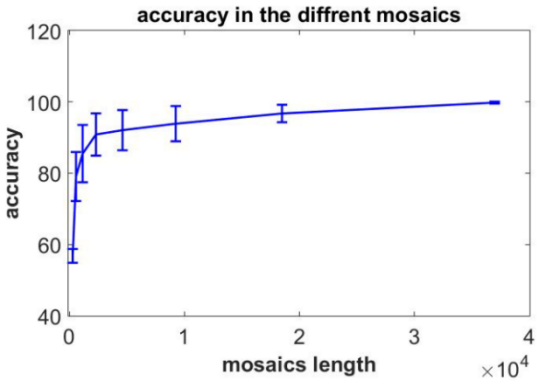


Figure 10. Average accuracy of ancestry inference using IABAC-RF for 20 samples from each admixed population of CEU-CHB and LWK-YRI with different lengths of mosaics

The results show that as the length of ancestral mosaics decreases, the accuracy of ancestry inference decreases. As the admixing time increases, the length of the ancestral mosaics constituting the admixed individual chromosome decreases, so with increasing admixing time, the accuracy of the IABAC-RF decreases.

4. DISCUSSION

Inferring the ancestry of admixed individuals is used in fields of historical demographic, anthropological, and pharmacogenomics. We present a method that uses the distance between the haplotype of ancestral populations to identify boundaries of ancestry change and then determining the ancestry between boundaries. The purpose of IABAC is to provide an appropriate way to divide chromosomes into smaller haplotype blocks. The major strength of IABAC is that it identifies approximate boundaries of ancestry change and ancestry inference without the need to determine many statistical or biological parameters. Many LAI methods require the determination of various parameters that may be difficult to obtain, which can hamper using LAI methods. For example, RFMix requires a genetic map, a window size and the average number of generations since admixture, while the only input parameter of IABAC is the length of the IB-block (L_w).

The results of this study show that the length of IB-blocks in Equations 3 and 4 are influential in the IABAC method. As the length of IB-blocks increases, the accuracy of the IABAC also increases. However, with excessive growth of the length of IB-blocks, the accuracy of the IABAC decreases. The reason for the decrease of IABAC accuracy with increasing IB-blocks length is that in places where the length of ancestral haplotype mosaics is small, large IB-blocks ignore them. In fact, in determining the points of ancestry change, the IB-blocks defined with L_w length play a role in smoothing ancestry change of the SNPs. The longer the L_w length is, the greater smoothing would be, and with the smaller L_w length, less smoothing will be observed. The ancestry inference of IABAC, which is based on ancestral change points, is preciser than the fixed window method where the length of the haplotype blocks is constant.

In determining the points of ancestry change, the correlation between the SNPs is considered. Due to the fact that the alleles within the ancestral population are not independent from each other in dense SNPs (background LD) (4), and also in examining the possibility of ancestry change between two adjacent SNPs using the distance between the alleles of an admixed individual and the mean of the alleles of ancestral populations, the distance of alleles in an SNP alone cannot have much information, so several SNPs must be considered together. As the results have shown, with the increase of the number of SNPs in each block, the accuracy of IABAC increases. In this case, the adjacent SNPs play a role in determining SNPs ancestral information and influence it. IB-blocks are sliding and move between all SNPs, so adjacent SNPs will be in the same IB-block at least once and will play a role in determining ancestral information of each other. As mentioned, the genomes of admixed individuals are a mosaic of different ancestral populations, so close SNPs are more likely to be from the same ancestors than far SNPs. For this reason, in determining the IB-block, in addition to the desired SNP, its nearest neighbors are considered.

In the stated algorithm to investigate the existence of ancestry change between two SNPs, in addition to the left and right IB-blocks of the desired SNPs (red IB-blocks in Figure 3a), the total IB-blocks (blue IB-blocks in Figure 3c) are also considered. The distance between different ancestral populations and admixed individual in the left, right and total IB-blocks is calculated. In fact, IABAC investigates the association between the ancestry of two IB-blocks and the sameness of their ancestral population. In other words, the proposed way for considering the association between the ancestry of IB-blocks is defined based on the statistic D explained for LD, which examines the probability of two alleles occurring at two chromosomal sites and calculates their correlation (40, 41). The value of D between alleles A and B at two chromosomal sites is presented by Equation 10.

$$D_{AB} = P_{AB} - P_A P_B \quad (10)$$

In this equation, D_{AB} represents the value of LD between alleles A and B, P_A denotes the frequencies of alleles A in the population, P_B denotes the frequencies of alleles B in the population, and P_{AB} is the frequencies of alleles A and B together (AB) in the population. Bigger D means more dependency between alleles and smaller D means more independency between alleles. In Equation 10, the frequencies of occurrence of each allele and the sum of alleles are considered, while in the IABAC, the distance of each IB-block and the sum of IB-blocks is considered.

Among the four classification methods mentioned, random forest method more accurately classifies haplotype blocks. With increasing admixture times, the ancestry inference accuracy decreases in all methods. However, the accuracy of IABAC-RF is better than other methods of ancestry inference.

5. CONCLUSION

IABAC infers ancestry by identifying the approximate boundaries of the ancestry change. To identify the boundaries of ancestry changes, IABAC uses distance between ancestral populations by considering the information of the neighbors in SNPs. After identifying boundaries of ancestry change, the ancestry between boundaries is determined. The important features of the IABAC compared to the former methods is that IABAC does not require many statistical or biological parameters. The only input parameter is the length of the IB-block. The results were shown that IABAC with IB-block length of 100-300 SNPs had the highest accuracy of ancestry inference and with increasing admixture times, IABAC with random forest classifier was more accurate that other methods for the ancestry inference. When IABAC and fixed window use the same classifiers, IABAC are more precise than fixed window for ancestry inference. In the present study, we studied haplotype data. We used SHAPEIT to convert genotype to haplotype. Future work will include adding a phasing step to IABAC. With the addition of the phasing step, if we have genotype data, phasing step converts genotype into haplotypes and then IABAC infers the ancestry of haplotypes. Sometimes we need to infer the ancestry of genotype data (similar to EILA). The algorithm used in IABAC, with minor changes, can infer the ancestry of genotype data, which can be done as future work.

6. REFERENCES

1. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nature genetics. 2003;33(Suppl 3):266-75. https://doi.org/10.1038/ng1113
2. 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):237-41. https://doi.org/10.1038/ng.763
3. Koehl AJ. Estimating ancestry and genetic diversity in admixed populations: The University of New Mexico; 2016.
4. 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. 2019;20(5):1709-24. https://doi.org/10.1093/bib/bby044
5. 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. https://doi.org/10.1371/journal.pgen.1000519
6. Gravel S. Population genetics models of local ancestry. Genetics. 2012;191(2):607-19.
7. Hu Y, Willer C, Zhan X, Kang HM, Abecasis GR. Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. The American Journal of Human Genetics. 2013;93(5):891-9. https://doi.org/10.1016/j.ajhg.2013.10.008
8. Ma Y, Zhao J, Wong J-S, Ma L, Li W, Fu G, et al. Accurate inference of local phased ancestry of modern admixed populations. Scientific reports. 2014;4(1):5800. https://doi.org/10.1038/srep05800
9. Durand EY, Do CB, Mountain JL, Macpherson JM. Ancestry composition: a novel, efficient pipeline for ancestry deconvolution. biorxiv. 2014:010512. https://doi.org/10.1101/010512
10. Khayatzadeh N, Mészáros G, Gredler B, Schnyder U, Curik I, Sölkner J. Prediction of global and local Simmental and Red Holstein Friesian admixture levels in Swiss Fleckvieh cattle. Poljoprivreda. 2015;21(1 SUPPLEMENT):63-7. https://doi.org/10.18047/poljo.21.1.sup.14
11. Alizadeh F, Jazayeriy H, Jazayeri O, Vafaee F, editors. SMIA: a simple way for inference of admixed population ancestors. 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE); 2020: IEEE. https://doi.org/10.1109/ICCKE50421.2020.9303686
12. Pool JE, Nielsen R. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics. 2009;181(2):711-9. https://doi.org/10.1534/genetics.108.098095
13. Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Kao WL, et al. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS genetics. 2011;7(4):e1001371. https://doi.org/10.1371/journal.pgen.1001371
14. Wang X, Zhu X, Qin H, Cooper RS, Ewens WJ, Li C, et al. Adjustment for local ancestry in genetic association analysis of admixed populations. Bioinformatics. 2011;27(5):670-7. https://doi.org/10.1093/bioinformatics/btq709
15. 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;13:1-10. https://doi.org/10.1186/1471-2156-13-49
16. 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):278-88. http://dx.doi.org/10.1016/j.ajhg.2013.06.020
17. 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. https://doi.org/10.1016/j.ajhg.2007.09.022
18. Alizadeh F, Jazayeriy H, Jazayeri O, Vafaee F. AICRF: Ancestry Inference of Admixed Population with Deep Conditional Random Field. Journal of Genetics. accepted for puplication, 2023. 10.1007/s12041-023-01445-7
19. Paşaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. Bioinformatics. 2009;25(12):i213-i21. https://doi.org/10.1093/bioinformatics/btp197
20. 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. https://doi.org/10.3378%2F027.084.0401

21. Salter-Townshend M, Myers S. Fine-scale inference of ancestry segments without prior knowledge of admixing groups. Genetics. 2019;212(3):869-89. https://doi.org/10.1534/genetics.119.302139
22. Guan Y. Detecting structure of haplotypes and local ancestry. Genetics. 2014;196(3):625-42. https://doi.org/10.1534/genetics.113.160697
23. Kumar A, Montserrat DM, Bustamante C, Ioannidis A. Xgmix: Local-ancestry inference with stacked xgboost. BioRxiv. 2020:2020.04. 21.053876. https://doi.org/10.1101/2020.04.21.053876
24. Montserrat DM, Bustamante C, Ioannidis A, editors. Lai-net: Local-ancestry inference with neural networks. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020: IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053662
25. Oriol Sabat B, Mas Montserrat D, Giro-i-Nieto X, Ioannidis AG. SALAI-Net: species-agnostic local ancestry inference network. Bioinformatics. 2022;38(Supplement_2):ii27-ii33. https://doi.org/10.1093/bioinformatics/btac464
26. Wang Y, Song S, Schraiber JG, Sedghifar A, Byrnes JK, Turissini DA, et al. Ancestry inference using reference labeled clusters of haplotypes. BMC bioinformatics. 2021;22(1):1-14. https://doi.org/10.1186/s12859-021-04350-x
27. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics. 2007;81(5):1084-97. https://doi.org/10.1086/521987
28. Yang JJ, Li J, Buu A, Williams LK. Efficient inference of local ancestry. Bioinformatics. 2013;29(21):2750-6. https://doi.org/10.1093/bioinformatics/btt488
29. Gaurav K, Kumar A, Singh P, Kumari A, Kasar M, Suryawanshi T. Human Disease Prediction using Machine Learning Techniques and Real-life Parameters. International Journal of Engineering. 2023;36(6):1092-8. https://doi.org/10.5829/ije.2023.36.06c.07
30. Hamidi H, Qaribpour F. An efficient predictive model for probability of genetic diseases transmission using a combined model. International Journal of Engineering. 2017;30(8):1152-9. 10.5829/ije.2017.30.08b.06
31. Kumar S, Sahoo G. A random forest classifier based on genetic algorithm for cardiovascular diseases diagnosis. International Journal of Engineering, Transactions B: Applications. 2017;30(11):1723-9. 10.5829/ije.2017.30.11b.13
32. Zamani F, Mohammadjani A. A Multiple Kernel Learning based Model with Clustered Features for Cancer Stage Detection using Gene Datasets. International Journal of Engineering, Transactions B: Applications. 2023. https://doi.org/10.5829/ije.2023.36.11b.08
33. Shedthi B S, Shetty V, Chadaga R, Bhat R, Bangera P, Kini K P. Implementation of Chatbot that Predicts an Illness Dynamically using Machine Learning Techniques. International Journal of Engineering. 2023. IJE Article in press
34. Anbananthen KSM, Busst MBMA, Kannan R, Kannan S. A Comparative Performance Analysis of Hybrid and Classical Machine Learning Method in Predicting Diabetes. Emerging Science Journal. 2022;7(1):102-15. https://doi.org/10.28991/ESJ-2023-07-01-08
35. Muthaiyah S, Singh VA, Zaw TOK, Anbananthen KS, Park B, Kim MJ. A Binary Survivability Prediction Classification Model towards Understanding of Osteosarcoma Prognosis. Emerging Science Journal. 2023;7(4):1294-314. https://doi.org/10.28991/ESJ-2023-07-04-018
36. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851-61. https://doi.org/10.1038%2Fnature06258
37. Delaneau O, Coulonges C, Zagury J-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC bioinformatics. 2008;9(1):1-14. https://doi.org/10.1186%2F1471-2105-9-540
38. Geza E, Mulder NJ, Chimusa ER, Mazandu GK. FRANC: a unified framework for multi-way local ancestry deconvolution with high density SNP data. Briefings in bioinformatics. 2020;21(5):1837-45. https://doi.org/10.1093/bib/bbz117
39. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005;310(5746):321-4. https://doi.org/10.1126/science.1117196
40. Slatkin M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics. 2008;9(6):477-85. https://doi.org/10.1038/nrg2361
41. Smith RD. The nonlinear structure of linkage disequilibrium. Theoretical Population Biology. 2020;134:160-70. https://doi.org/10.1016/j.tpb.2020.02.005

Persian Abstract

چکیده

ترکیب ژنوم یک پدیده رایج در جمعیت های انسانی است که نتیجه آمیزش افراد از دو یا چند جمعیت مستقل است. این می تواند منجر به تشکیل DNA جدیدی شود که هر بخش آن از جمعیت اجدادی متفاوت منشا می گیرد. روش‌های استنباط تبار محلی برای شناسایی تبار هر بخش از ژنوم استفاده می‌شوند که می‌تواند بینشی در مورد تاریخچه جمعیت‌ها ارائه دهد. بسیاری از روش‌های استنباط تبار محلی (LAI) نیاز به تعیین پارامترهای مختلفی دارند که ممکن است بدست آوردن آنها دشوار بوده و استفاده از روش‌های LAI را با مشکل مواجه کند.در این مقاله، ما یک روش جدید برای شناسایی مرزهای تقریبی تغییر تبار (IABAC) در هاپلوتیپ‌های مخلوط و سپس تعیین تبار بین مرزها ارائه می‌دهیم. بر خلاف بسیاری از روش های LAI، روش ما به پارامترهای آماری یا بیولوژیکی زیادی متکی نیست، بنابراین در برابر تغییرات الگوهای اختلاط قوی‌تر است .ما روش خود را بر روی داده های انسان ارزیابی می کنیم و نشان می دهیم که از روش های موجود تشخیص تبار دقیق تر است. نتایج ما نشان می‌دهد که IABAC یک روش جدید امیدوارکننده برای شناسایی مرزهای تغییر تبار در هاپلوتیپ‌های مخلوط است. این روش می تواند برای مطالعه تاریخچه اختلاط در جمعیت های انسانی و شناسایی واریانت های ژنتیکی که با جمعیت های اجدادی مختلف مرتبط هستند مورد استفاده قرار گیرد.