

Allele frequency patterns of 78,122,254 SNPs illuminate diverse evolutionary trajectories of human populations: Supplementary Text S1

Yu-Li Tung¹, Khong-Loon Tiong¹, Chen-Hsiang Yeang^{1*}

¹Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Taipei, Taiwan.

*corresponding author, email: chyeang@stat.sinica.edu.tw

In this Supplementary Text we provide detailed descriptions of seven algorithms used in this study.

Inference of continental-level SNP allele frequency patterns from 1000G data

The centerpiece of this study is an algorithm to infer the allele frequency pattern at continental level. In this subsection we describe the algorithm in detail. To facilitate communication we first introduce the following notations. Suppose the data contains subjects from P populations belonging to C continental groups, and denote $P2C: \{1, \dots, P\} \rightarrow \{1, \dots, C\}$ the mapping from the population label to the continental label. In the 1000G data, there are $P = 20$ populations (excluding the mixed populations from Americas) and $C = 4$ continental groups (1: AFR or Africa, 2: EUR or Europe, 3: SAS or South Asia, 4: EAS or East Asia). For each locus and each population, we count the number of subjects possessing each allele (0, 1 or 2) and normalize the counts by the number of subjects belonging to the population. The normalized allele frequencies of 0 (homozygote major allele) and 2 (homozygote minor allele) fully specify the data since the normalized allele frequency of 1 (heterozygote allele) is one minus the remaining two components. Therefore, we represent the population allele frequency data on a particular locus as a $P \times 2$ matrix X , where $X_{ij} \in [0,1]$ denotes the normalized frequency of population i on allele j . The continental-level allele frequency pattern comprises two parts. The first part specifies the super-continental groups that possibly combine multiple continental groups (e.g., AFR and EUR may form one group) and the memberships of individual populations to these groups. To accommodate the presence of outliers we allow a population not assigned to any group. Two more restrictions are imposed: (1) populations of each continent are assigned to at most one super-continental group, (2) at most two populations of each continent have empty memberships. Consequently, the first part is a membership function $\mathcal{M}: \{1, \dots, P\} \rightarrow \{0,1, \dots, N \leq C\}$ with the constraints that for $S_c \equiv \{i: P2C(i) = c\}$, $|\{i \in S_c: \mathcal{M}(i) = 0\}| \leq 2$, and $\forall i, j \in S_c, (\mathcal{M}(i) = \mathcal{M}(j)) \vee (\mathcal{M}(i) = 0) \vee (\mathcal{M}(j) = 0)$. The second part specifies the geometric relations between the super-continental groups on the two dimensional plane of X . We represent these relations by an undirected graph $G = (V, E)$, where $V = \{1, \dots, N\}$ denotes the super-continental groups and E specifies the relations that two super-continental groups have no other groups between them. These two parts are combined to form an encode vector. An encode vector ε has $2 + P + \frac{C(C+1)}{2}$ components. The first two components specify the numbers of super-continental groups N

or $(|V|)$ and edges $|E|$. The next P components specify the membership function $\mathcal{M}$. The following $\frac{C(C+1)}{2}$ components specifies the edges in E .

We now describe the algorithm.

Algorithm 1: Inferring the continental level allele frequency patterns

Inputs: Population allele frequencies X , population-to-continent mapping $P2C$.

Outputs: The encode vector ε .

Procedures:

1. Check whether all the 2D vectors in X are encompassed in a circle with a small radius. If so, then set $N = 1, |E| = 0, \mathcal{M}(i) = 1 \forall i$, and return.
2. Find the outliers of each continental group.
 - 2.1 Extract the vectors X_c for a continental group c .
 - 2.2 Retrieve the unique pairwise distances between vectors in X_c and sort them in an increasing order.
 - 2.3 Vary the distance threshold d_θ from small to large values. For each d_θ value, construct an undirected graph by connecting the vectors whose distances $\leq d_\theta$, and obtain connected components of the graph.
 - 2.4 Mark the birth and death of each cluster, and identify the stable clusters with the longest lifespans.
 - 2.5 If the distinct stable clusters are distant on the 2D plane, then mark the minority cluster(s) as the outliers.
3. Enumerate all possible arrangements of the membership function.
 - 3.1 Vary the number of super-continental groups $N = 1, \dots, C$.
 - 3.2 For each N enumerate all possible mappings from C continental groups to N super-continental groups.
 - 3.3 For each continent-super continent mapping enumerate all possible assignments of outliers such that the number of outliers in each continent ≤ 2 .
4. Apply a series of filters to rule out the membership functions incompatible with X .
 - 4.1 For a membership function $\mathcal{M}$, if there exists a population i and a super continental group j , such that $\mathcal{M}(i) \neq j$ and X_i is in the interior of $X_{S_j}, S_j \equiv \{k: \mathcal{M}(k) = j\}$, then discard $\mathcal{M}$. In other words, if the allele frequencies of a population are in the interior of the allele frequencies of populations assigned to different super continental populations, then discard the membership function.
 - 4.2 For a membership function $\mathcal{M}$, if there exist two continental groups which do not have interface points but are assigned to the same super continental groups, then discard $\mathcal{M}$.
 - 4.3 If an outlier population according to a membership function $\mathcal{M}$ is neither an outlier of the data (obtained at 2.5) nor at the interior of any other super continental group members, then discard $\mathcal{M}$.
 - 4.4 If an outlier population according to $\mathcal{M}$ is not an interface between pairs of continental groups assigned to different super continental groups, then discard $\mathcal{M}$.

5. Among the membership functions that pass filters 4.1-4.4, calculate their Rand indices: the number of population pairs which belong to the same continent and are assigned to the same super continental group plus the number of population pairs which belong to the different continents and are assigned to different super continental groups. Choose the membership function with the highest Rand index.
6. Consider the populations which are assigned to non-outliers according to the selected membership function. Enumerate all triplets of these populations.
7. For each triplet (i, j, k) check whether population k is between populations i and j on the 2D plane. Betweenness is determined by (1) checking whether the angle spanned by two vectors $\vec{ki}$ and $\vec{kj} \geq \frac{\pi}{2}$, (2) if (1) is passed, then projecting k onto the line $\vec{ij}$ and checking whether the projection k' lies between i and j . If both conditions hold, then k is between i and j .
8. Construct a graph G_P with nodes as populations. Two nodes (i, j) are adjacent if their corresponding populations have at most one population between them.
9. Collapse G_P to the graph G with nodes as super continental groups. Two super continental groups s_1 and s_2 are adjacent if any pair of their member populations is adjacent in G_P .
10. Report $\mathcal{M}$ and G in the encode vector ε .

Verification of the inferred encode vectors on HGDP data

We verify the SNP allele frequency patterns inferred from the 1000G data on another dataset of Human Genome Diversity Project (HGDP). Similar to 1000G, HGDP collects the genotype data of diverse populations across all continents. Moreover, nearly all subjects of HGDP are from the indigenous populations, hence they are a more suitable reference for the population genotypes. Yet the main drawback of HGDP is the small sample size of each population. Unlike 1000G where each population typically contains around 100 subjects, in HGDP each population includes only around 8-20 subjects. The small sample size makes the allele frequencies in most populations rather unstable. To mitigate this problem, rather than giving each population a pair of (homozygote) major and minor allele frequencies, we collapse all subjects from the same continent in one group and count the normalized allele frequencies of each continent in the HGDP data. We then verify the 1000G-inferred SNP allele frequency patterns by comparing the graphs derived from the encode vectors and the graphs derived from the betweenness relations of the continental level HGDP data. The comparison algorithm is described in detail below.

Algorithm 2: Verifying the SNP allele frequency patterns derived from 1000G data on HGDP data.

Inputs: 1000G SNP data of a locus, HGDP SNP data of the same locus.

Outputs: The encode vector ε inferred from 1000G data. Verification outcome on HGDP data.

Procedures:

1. Convert the 1000G SNP data into the population allele frequencies X_1 where the allele frequencies aggregated at population levels (e.g., GBR, YRI) are reported.
2. Apply Algorithm 1 to infer the encode vector ε from X_1 .
3. Convert the HGDP SNP data into the continental allele frequencies X_2 where the allele frequencies aggregated at continental levels (e.g., EUR, AFR) are reported. Restrict HGDP data to the four continents appeared in 1000G data (AFR, EUR, SAS, EAS).
4. Retrieve from ε the conversion function $C2S$ from continents to super continents and the graph G_1 of super continental groups.
5. From X_2 evaluate the betweenness relations for all triplets of continents. Construct the graph G_2 of continents according to the betweenness relations. Incur the same procedures as steps 7-8 in Algorithm 1.
6. Collapse G_2 to the graph G'_2 with nodes as super continental groups. Follow the same procedure as step 9 of Algorithm 1.
7. The verification outcome is 1 if $G_1 = G'_2$, and is 0 otherwise.

Inference of continental-level SNP allele frequency patterns jointly from 1000G and HGDP data

1000G data covers the subjects from only four continents (AFR, EUR, SAS, EAS). Instead HGDP data covers the subjects from the same four continents and three more continents (Middle East or MID, America or AMR, Oceania or OCE), and the South and Central Asia continent includes populations from the Indian subcontinent and Central Asia such as Afghanistan and Iran. By joining 1000G and HGDP data we can infer the SNP allele frequency patterns from the seven continents. However, since the sample sizes of HGDP populations are rather small, and the 1000G and HGDP data are not necessarily compatible, it is inadequate to simply concatenate the two datasets and apply Algorithm 1 to the joint data. Instead we apply Algorithm 2 to identify the loci with compatible encode vectors of the two datasets and then infer the SNP allele frequency patterns from the full HGDP data at continental level. Below we describe the algorithm in detail.

Algorithm 3: Inferring the SNP allele frequency patterns derived from 1000G and HGDP data.

Inputs: 1000G SNP data of a locus, HGDP SNP data of the same locus.

Outputs: The encode vector ε inferred from both 1000G and HGDP data.

Procedures:

1. Apply Algorithm 2 to verify whether the encode vector ε_1 inferred from 1000G data is compatible with HGDP data on the four continents.
2. If ε_1 is incompatible with HGDP data, then return a null encode vector.
3. Retrieve from ε the conversion function $C2S$ from continents to super continents and the graph G_1 of super continental groups.

4. Convert the HGDP SNP data into the continental allele frequencies X_2 where the allele frequencies aggregated at continental levels are reported. Apply the conversion to the complete data covering seven continents.
5. Check whether each of the three additional continents in HGDP data is encompassed by the super continents obtained at step 1. If so, then absorb the continent into the corresponding super continent. Otherwise assign the continent to a new super continent group.
6. Evaluate the betweenness relations of all continental triplets in X_2 . Construct the graph G_2 according to the betweenness relations. Collapse G_2 into G according to the continent-to-super continent mapping. Repeat steps 6-9 of Algorithm 1.
7. Report the continent-to-super continent mapping and G .

Inference of population-level SNP allele frequency linear patterns within a continent

Beyond the continental-level allele frequency patterns we are also interested in the population-level patterns within each continent. We consider only the linear patterns of populations in 1000G data. Since each continent includes only five populations (for instance, within East Asia there are KHV, CDX, CHS, CHB and JPT) and each population comprises around 100 subjects, the inference outcomes are more reliable and interpretable. The algorithm resembles Algorithm 3 by constructing a linear graph according to the betweenness relations. Below we describe the algorithm in detail.

Algorithm 4: Inferring the SNP allele frequency patterns derived from 1000G and HGDP data.

Inputs: 1000G SNP data of a locus within a continent.

Outputs: The encode vector ε inferred from 1000G data.

Procedures:

1. Convert the 1000G SNP data into the population allele frequencies X_1 where the allele frequencies aggregated at population levels are reported.
2. From X_1 evaluate the betweenness relations for all triplets of populations. Construct the graph G_1 of populations according to the betweenness relations. Incur the same procedures as steps 7-8 in Algorithm 1.
3. Discard G_1 if it is not a linear graph.
4. Construct ε according to G_1 .

Inference of the evolutionary histories from the SNP allele frequency patterns

A SNP allele frequency linear pattern can be converted into (one or multiple) evolutionary patterns of the four continental populations. In this subsection we describe the algorithm to construct this mapping. To streamline inference we use the following compact representation for the allele frequency linear pattern (rather than the aforementioned

encode vector). A chain of K continental populations ($K = 4$ in 1000G data) c is a K -component vector where each component indicates the order of the corresponding continent in the linear pattern. For instance, if the continental graph of a linear pattern is $1 - (2,3) - 4$, then the chain vector is $(1,2,2,3)$ (or equivalently $(3,2,2,1)$ by reversing the order of nodes in the chain). We follow the convention that 1: AFR, 2: EUR, 3: SAS, 4: EAS.

Like most studies in evolutionary biology, an evolutionary pattern is represented by a tree. The structure of a tree can be commonly represented by a *regular expression* comprising the notations of indices, parentheses and commas. Siblings with a common parent are separated by commas, while left and right parentheses indicate the start and end of descendant nodes of a specific branch. For instance, the regular expression $(1,((2,3),4))$ depicts a binary tree where nodes 2 and 3 are adjacent to a common parent 5, nodes 4 and 5 are adjacent to a common parent 6, and nodes 1 and 6 are adjacent to the common parent and root 7. Furthermore, one tree structure may yield multiple edge length configurations. Hence a tree is the combination of a tree structure and edge length configurations.

The mapping between SNP allele frequency linear patterns (the chain vectors) and evolutionary patterns (the regular expressions of tree structures and their edge length configurations) is not a straightforward and one-to-one relation. Below we describe the assumptions and the algorithm to build this mapping.

Algorithm 5: Constructing the mapping from an allele frequency linear pattern to the compatible evolutionary patterns.

Inputs: A set of chain vectors C of K continental populations.

Outputs: A set of compatible trees comprising regular expressions $\mathcal{E}$ and their edge length configurations $\mathcal{F}$.

Procedures:

1. Enumerate three binary tree structures of the four continental populations: $(1,((2,3),4))$, $(1,(2,(3,4)))$, $(1,(3,(2,4)))$. We consider these tree structures because only they are compatible with the history that in the vast majority of loci Africans (1) split early from the Eurasians (2-4).
2. Enumerate the edge length configurations of each tree structure. Each edge has only three possible lengths 0, 1 and 2 since the three values suffice to discern the four continental populations in the linear patterns.
3. For each chain $c \in C$, tree $T \equiv (e, f)$ of structure e and edge length configuration f , check its compatibility with c by comparing the pairwise distance orders between continental populations.
 - 3.1 In c , the distance between continents i and j is 0 if they are assigned to the same super continental group ($c(i) = c(j)$). Otherwise the distance is the number of steps between their super continental groups ($|c(i) - c(j)|$).
 - 3.2 In T , the distance between continents i and j is their shortest path lengths.

- 3.3 T is compatible with c if their orders of continental pairwise distances are identical (ties are distinguished from strict orders).
4. Impose parsimonious constraints to reduce the number of compatible trees.
 - 4.1 If the two branches of an internal node have zero lengths, then no descendant nodes have nonzero branch lengths. This constraint attempts to rule out the counterintuitive outcomes that continents are separated by more branches but have shorter distances because most of the early branches have zero lengths.
 - 4.2 Keep the compatible edge length configurations which have the minimum number of unique edge length values. This constraint incorporates parsimony by introducing the minimum number of allele changes to fit the linear pattern.
 - 4.3 Find the compatible edge length configurations within the minimum number of 1s and 2s. Among the selected configurations, if some have 1s and 0s only, then exclude the configurations with 2s. This constraint also incorporates parsimony by minimizing the number of allele changes to fit the linear pattern.
5. Impose global constraints across multiple chains on the compatible trees.
 - 5.1 Group the chains (linear patterns) such that pairs of chains can be reduced to each other by collapsing and expanding members of the super-continental groups. For instance, the three chains (1,3,2,1), (1,4,3,2), (3,1,2,4) are in the same group (Table 3 in the main text) because the first chain can be reduced from the second and third chains by collapsing continents 1 and 4 in the same super-continental group. The chains in the same group can be equivalent by varying the parameters of the pattern inference algorithm (Algorithm 1).
 - 5.2 Among the compatible trees that pass the parsimonious criteria at step 4, keep the ones that are compatible with all the chains in some chain groups. This constraint requires that the compatible chains of a selected tree should be robust against the parameter values of the inference algorithm.
6. Report the selected trees and their compatible chains.

Local ancestry inference from the SNP allele frequency patterns

The SNP allele frequency patterns can distinguish populations at continental level and within continents and are widely distributed along all chromosomes. Therefore, we propose an algorithm to incorporate these allele frequency patterns to perform a classical analysis in population genomics: inferring the ancestry tracts of mixed descendants. Similar to other local ancestry inference algorithms, we subdivide each chromosome into small windows and build a classifier for each window to assign the ancestry label according to the training data. Yet our algorithm differs from previous methods in several aspects. First, we adopt a two-tier approach to enhance granularity of prediction. In the first tier we build a continental-level predictor based on the extended SNP allele frequency patterns of seven continents inferred from the 1000G and HGDP data. In the second tier we build a population-level predictor based on the linear SNP allele frequency patterns within each continent inferred from 1000G data alone. Second, we combine the predictive power of all loci possessing all the SNP allele frequency patterns to distinguish the populations. The predictive power of one allele frequency pattern is often limited. For instance, the allele frequency pattern # 2 separates Africans from Eurasians hence cannot distinguish the populations within Eurasia.

However, by combining all the SNP allele frequency patterns we can distinguish populations from different “angles” and better predict the population labels.

Before explaining the local ancestry inference algorithm, we first describe an algorithm to predict the population label according to the genotype data of selected loci. Each locus is annotated with a SNP allele frequency pattern (according to algorithms 3 or 4). For each allele frequency pattern, we calculate the weighted sum of the genotype values of the member loci, and use the aggregate value to quantify the probability of each predicted label. The joint probabilities of the predicted labels are obtained by multiplying the probabilities over all allele frequency patterns.

Algorithm 6: Predicting the population labels from the SNP allele frequency patterns and genotype data.

Inputs: The training genotype data G^1 of M loci and N_1 subjects from K populations. An $U \times K$ matrix L representing the linear orders of U unique SNP allele frequency patterns. Each row in L indicates the order of a linear pattern. For instance, $(1,2,2,3)$ represents the chain $1 - (2,3) - 4$. A $1 \times M$ vector v_1 indicating the linear pattern indices of all loci. A $1 \times N_1$ vector v_2 indicating the population labels of each subject. The test genotype data G^2 of M loci and N_2 subjects.

Outputs: A $N_2 \times K$ matrix P_2 of predicted population label probabilities of each test subject, and a $1 \times N_2$ vector F_2 of predicted population labels.

Procedures:

1. For each locus i with linear pattern index l_i , calculate the weight w_i with the following steps.
 - 1.1 Find the corresponding row L_{l_i} of L , and denote K_i the number of classes in L_{l_i} . Without loss of generality, suppose classes $1, \dots, K_i$ follow a sequential order.
 - 1.2 For each class $j = 1, \dots, K_i$, find the member populations $C_j \equiv \{k: L_{l_i}(k) = j\}$. Collect the subjects whose population labels belong to C_j : $S_j \equiv \{k: v_2(k) \in C_j\}$. Calculate the mean genotype value m_j over subjects of S_j .
 - 1.3 Check whether $m_1, \dots, m_{K_i}$ follow a sequential order $m_1 \geq m_2 \geq \dots \geq m_{K_i}$ or $m_1 \leq m_2 \leq \dots \leq m_{K_i}$. If so, then assign the weight of locus i to $w_i = \min_{j=1, \dots, K_i-1} (m_j - m_{j-1})$. The weight quantifies the minimum margin between the mean genotype values of adjacent classes. Otherwise $w_i = 0$.
2. For each unique linear pattern L_{l_i} , construct a probabilistic predictor of population labels with the following steps.
 - 2.1 Denote K_i the number of classes in L_{l_i} . For each class $j = 1, \dots, K_i$, find the member populations C_j and the corresponding subjects S_j . For each training subject $l = 1, \dots, N_1$, calculate the weighted sum of the genotype values (projection values) $\pi_{jl} \equiv \sum_{k: v_1(k)=i} w_k G_{kl}^1$.
 - 2.2 Collect the projection values $\Pi_1, \dots, \Pi_{K_i}$ of subjects belonging to each class. Denote the range of projection values $x \in [x_{min}, x_{max}]$. For each class $j =$

- 1, $\dots$, K_i , apply kernel density estimation to infer the probability density function $p_j(x)$ from Π_j .
- 2.3 For each possible projection value $x \in [x_{min}, x_{max}]$, renormalize the pdf of each class to calculate the probability of assigning each class: $\tilde{p}_i(y = j|x) = \frac{p_j(x)}{\sum_{j'=1}^{K_i} p_{j'}(x)}$.
- 2.4 Adjust $\tilde{p}_i(y|x)$ to assign the prediction probabilities of K populations. If a class j includes multiple member populations C_j , then assign the probability of a population $k \in C_j$ to $\hat{p}_i(y = k|x) = \tilde{p}_i(y = j|x) / |C_j|$, indicating an equal probability for each population in class j .
3. Combine the probabilistic predictors of all unique linear patterns in L to predict the population labels of each test subject G_t^2 .
 - 3.1 For each L_i of L , calculate the probability vector of population labels $\hat{p}_i(y_t = k|G_t^2)$ for $k = 1, \dots, K$.
 - 3.2 Evaluate the joint log-likelihood by adding the population label log probabilities over all unique linear patterns. Reweight each log probability by the number of loci belonging to each linear pattern: $\ln \hat{L}_t(y = k|G_t^2) = \sum_{L_i \in L} \#(v_1 = i) \ln \hat{p}_i(y_t = k|G_t^2)$ for $k = 1, \dots, K$.
 - 3.3 Renormalize $\hat{P}_t(y = k|G_t^2)$ according to $\hat{L}_t(y = k|G_t^2)$ and concatenate the rows of $\hat{P}_t(y = k|G_t^2)$ over the test subjects to form P_2 .
 - 3.4 Return $\hat{k} = \arg \max_{k=1, \dots, K} \hat{P}_t(y = k|G_t^2)$ and concatenate the predicted labels over the test subjects to form F_2 .

The probabilistic predictor of the population labels forms the basis for our algorithm of inferring the ancestry tracts of subjects from a mixed lineage. In brief, we subdivide a chromosome into smaller windows, apply the probabilistic predictor to the genotype data in each window, and integrate the prediction outcomes of consecutive windows. The window size is determined by the training error rate of predictions.

Algorithm 7. Inferring the ancestry tracts of subjects from the SNP allele frequency patterns and genotype data.

Inputs: The 1000G and HGDP training data of genotypes G^1 of M loci and N_1 subjects from known populations. The continental-level encode vectors $\mathcal{E}_1$ of all loci in the training data. The population-level encode vectors $\mathcal{E}_2$ of all loci in the training data. The matrices of unique linear patterns L_1 and L_2 at continental and population levels respectively. $1 \times M$ vectors v_{11} and v_{12} indicating the linear pattern indices of all loci at continental and population levels. $1 \times N_1$ vectors v_{21} and v_{22} indicating the continental and population labels of each subject. The test genotype data G^2 of M loci and N_2 subjects.

Outputs: For each subject l of G^2 and each chromosome γ , partition chromosome γ into sets of triplets $(s_{l\gamma 1}, t_{l\gamma 1}, c_{l\gamma 1}), \dots, (s_{l\gamma R}, t_{l\gamma R}, c_{l\gamma R})$, where each triplet $(s_{l\gamma i}, t_{l\gamma i}, c_{l\gamma i})$ denotes the start and end positions of a tract and the predicted label at continental or population level.

Procedures:

1. Infer the tracts of test subjects at continental level.
 - 1.1 For chromosome γ , vary the window size W and subdivide the chromosome into windows of the specified size. For each window apply Algorithm 6 to the training data G^1 restricted to the specified window and predict the target class labels. Calculate the training error rate of the target class predictions. So a given window size yields a number of training error rates for the windows. Choose the smallest window size that minimizes the maximum of the window training error rates.
 - 1.2 Denote $\hat{W}$ the selected window size according to step 1.1. Subdivide chromosome γ into windows of size $\hat{W}$. If the size of the last window is not close to $\hat{W}$, then append it to the precedent window. Apply Algorithm 6 on the restricted training data to build the classifier for each window.
 - 1.3 Apply the classifier of each window to the test data of each subject and assign a window to the predicted label with the maximum log likelihood score.
 - 1.4 Combine the consecutive windows with identical predicted class labels. The merged windows and their predicted class labels form the inferred tracts at continental level.
2. Conditioned on the inferred tracts at continental level, infer the tracts of test subjects at population level. Suppose we want to infer the population labels of tracts belonging to a specific continental class (e.g., Africa).
 - 2.1 For each chromosome γ and test subject, collect the tracts belonging to the specific continent.
 - 2.2 Apply step 1.1 to the training data of the collected loci to determine the window size of predictors at population level. The training data is restricted to the subjects from the specific continent, and the class labels are the population labels within the specific continent.
 - 2.3 Apply step 1.2 to construct the population label classifiers of each window on the restricted training data.
 - 2.4 Incur step 1.3 to apply the classifier of each window to the test data of each subject and assign a predicted population label to each window.
 - 2.5 Incur step 1.4 to combine the consecutive windows with identical predicted class labels.