

1 **Length Distribution of Ancestral Tracks under a General Admixture Model and**
2 **Its Applications in Population History Inference**

3 Xumin Ni,^{*,1} Xiong Yang,^{†,1} Wei Guo,[‡] Kai Yuan,[†] Ying Zhou,[†] Zhiming Ma,^{‡,2}
4 Shuhua Xu^{†,§,**,2}

5 *Department of Mathematics, School of Science, Beijing Jiaotong University, Beijing
6 100044, China;

7 † Chinese Academy of Sciences (CAS) Key Laboratory of Computational Biology, Max
8 Planck Independent Research Group on Population Genomics, CAS-MPG Partner
9 Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences,
10 CAS, Shanghai 200031, China;

11 ‡ Institute of Applied Mathematics, Academy of Mathematics and Systems Science,
12 Chinese Academy of Sciences, Beijing 100190, China;

13 § School of Life Science and Technology, ShanghaiTec University, Shanghai 200031,
14 China;

15 ** Collaborative Innovation Center of Genetics and Development, Shanghai 200438,
16 China.

17

¹ These authors contributed equally to this work.

² Correspondence and requests for materials should be addressed to S.X. (xushua@picb.ac.cn) or Z.M. (mazm@amt.ac.cn).

1 *Running title:* Inference of population admixture history

2 **Keywords:** Genetic admixture; Ancestral tracks; Population history; SNP

3

4 Shuhua Xu: 320 Yueyang Road, Chinese Academy of Sciences (CAS) Key Laboratory of

5 Computational Biology, Max Planck Independent Research Group on Population

6 Genomics, CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai

7 Institutes for Biological Sciences, CAS, Shanghai 200031, China. Tel:

8 +86-21-549-20479. E-mail: xushua@gmail.com.

9 Zhiming Ma: Zhongguancun East Road, Chinese Academy of Sciences, Academy of

10 Mathematics and Systems Science(South Building), Beijing 100190, China. E-mail:

11 mazm@amt.ac.cn.

12

1 **ABSTRACT**

2 As a chromosome is sliced into pieces by recombination after entering an admixed
3 population, ancestral tracks of chromosomes are shortened with the pasting of
4 generations. The length distribution of ancestral tracks reflects information of
5 recombination and thus can be used to infer the histories of admixed populations.
6 Previous studies have shown that inference based on ancestral tracks is powerful in
7 recovering the histories of admixed populations. However, population histories are
8 always complex, and previous studies only deduced the length distribution of
9 ancestral tracks under very simple admixture models. The deduction of length
10 distribution of ancestral tracks under a more general model will greatly elevate the
11 power in inferring population histories. Here we first deduced the length distribution
12 of ancestral tracks under a general model in an admixed population, and proposed
13 general principles in parameter estimation and model selection with the length
14 distribution. Next, we focused on studying the length distribution of ancestral tracks
15 and its applications under three typical admixture models, which were all special
16 cases of our general model. Extensive simulations showed that the length distribution
17 of ancestral tracks was well predicted by our theoretical models. We further developed
18 a new method based on the length distribution of ancestral tracks and good
19 performance was observed when it was applied in inferring population histories under
20 the three typical models. Notably, our method was insensitive to demographic history,
21 sample size and threshold to discard short tracks. Finally, we applied our method in
22 African Americans and Mexicans from the HapMap dataset, and several South Asian

1 populations from the Human Genome Diversity Project dataset. The results showed
2 that the histories of African Americans and Mexicans matched the historical records
3 well, and the population admixture history of South Asians was very complex and
4 could be traced back to around 100 generations ago.

5

1 INTRODUCTION

2 Population admixture is a common phenomenon in human populations when
3 previously isolated populations start to contact and interact with each other,
4 accompanied by population migration, rising and falling of empires, trading of goods
5 and services, and so on (HELLENTHAL *et al.* 2014). The history of population
6 admixture does not fade with time, but leaves a great deal of information in the
7 genomes of individuals from admixed populations. Population history in admixed
8 populations thus can be recovered by utilizing the information in the genome, such as
9 break points of recombination (XU *et al.* 2008), admixture linkage disequilibrium
10 (ALD) (PATTERSON *et al.* 2012; LOH *et al.* 2013; PICKRELL *et al.* 2014) and the length
11 of ancestral tracks (ancestral tracks) (POOL and NIELSEN 2009; PUGACH *et al.* 2011;
12 GRAVEL 2012; JIN *et al.* 2012; HELLENTHAL *et al.* 2014; JIN *et al.* 2014).

13 The information of ancestral tracks was first used by Pool and Nielsen (which they
14 called migration tracts) to infer the history of populations, and they also deduced the
15 distribution of ancestral tracks under a hybrid isolation (HI, or a one pulse admixture)
16 model with a small migration rate (POOL and NIELSEN 2009). A subsequent study of
17 Pugach *et al.* performed wavelet transform on the ancestral tracks in admixed
18 populations to obtain the dominant frequency of ancestral tracks and compared it to
19 those obtained from extensive simulations, to estimate the admixture time, still under
20 the simple HI model (PUGACH *et al.* 2011). Then, the study of Jin *et al.* explored
21 admixture dynamics by comparing the empirical and simulated distribution of
22 ancestral tracks under three typical two-way admixture models; i.e. HI model, gradual

1 admixture (GA) model, and continuous gene flow (CGF) model (JIN *et al.* 2012). Jin
2 *et al.* later deduced the theoretical distributions of ancestral tracks under HI and GA
3 models (JIN *et al.* 2014). Gravel extended these studies to multiple ancestral
4 populations and discrete migrations. However, he failed to explicitly deduce the
5 theoretical distribution of ancestral tracks under a general situation (GRAVEL 2012).

6 Here we proposed a general model that can cover all the scenarios of an admixed
7 population with an arbitrary number of ancestral populations and (or) arbitrary
8 number of admixture events. In this study, we first described the general admixture
9 model and deduced a general formula for the theoretical distribution of ancestral
10 tracks. With this distribution, we can use maximum likelihood estimation (MLE) to
11 estimate model parameters, and the Akaike information criterion (AIC) (AKAIKE 1998)
12 or the likelihood ratio test (LRT) (WILKS 1938) to select an optimal model from
13 candidates for the given data. We next demonstrated that the three aforementioned
14 admixture models, namely HI, GA and CGF models in previous studies (JIN *et al.*
15 2012), are all special cases under our general model. Then, under these three models,
16 we developed a method called *AdmixInfer* to estimate the admixture proportion and
17 admixture time, and simultaneously selected the optimal model according to the
18 principles of AIC. We further conducted extensive simulation studies to demonstrate
19 the accuracy of the theoretical distribution of ancestral tracks under the general model,
20 and the effectiveness of our method to estimate the parameters and select an optimal
21 model. Finally we applied our method to African Americans and Mexicans from the
22 HapMap phase III dataset (INTERNATIONAL HAPMAP *et al.* 2010), and several South

1 Asian populations from the Human Genome Diversity Project (HGDP) (Li *et al.* 2008)
2 dataset.

3 **GENERAL MODEL**

4 **Length Distribution of Ancestral Tracks**

5 In our general model, population admixture is accomplished by receiving gene flow(s)
6 from ancestral populations either continuously or discontinuously. We model this
7 process generation by generation, in which, if the admixed population does not
8 receive further gene flow(s) in a particular generation, we set the strength of gene
9 flow(s) to 0. Specifically, given an admixed population started T generations ago,
10 with K ancestral populations, let $m_i(t)$ denotes the ancestry proportion from the
11 i th ancestral population that entered at the t generation (Figure 1), then the general
12 model is only determined by a $K \times T$ matrix $M = \{m_i(t)\}_{1 \leq i \leq K, 1 \leq t \leq T}$, which
13 satisfies three properties: (a) $m_i(t) \geq 0$; (b) $\sum_{i=1}^K m_i(1) = 1$ and (c) $\sum_{i=1}^K m_i(t) \leq$
14 $1, \text{ if } 2 \leq t \leq T$.

15 Let $I(t)$ be the ancestry proportion of the admixed population at generation t
16 inherited from the previous generation, and then we get $I(t)$ as

$$I(t) = 1 - \sum_{j=1}^K m_j(t), 1 \leq t \leq T. \quad (1)$$

17 Denote $H_i(t)$ as the total ancestry proportion of the i th ancestral population in the
18 admixed population at t generation, then

$$H_i(t) = \begin{cases} m_i(1), & t = 1 \\ H_i(t-1)I(t) + m_i(t), & 2 \leq t \leq T. \end{cases}$$

19 Recursively, we can get

$$H_i(t) = \sum_{k=1}^{t-1} m_i(k) \left(\prod_{l=k+1}^t I(l) \right) + m_i(t). \quad (2)$$

1 Define $s_i(t)$ as the survival proportion of the ancestral tracks at generation T from
2 the i th ancestral population that entered at generation t , then

$$s_i(t) = m_i(t) \left(\prod_{l=t+1}^T I(l) \right). \quad (3)$$

3 Generally, we assume that the ancestral populations are homogeneous, and
4 recombination among segments from the same ancestry does change the length of the
5 tracks, but it is not “observable” to us, thus the length of tracks seems unchanged, and
6 only these recombination events among different ancestries produce “observable”
7 changes. (Figure S1) Here we explicitly take these “unobservable” changes into
8 consideration and adjust the recombination rate accordingly as following. Define the
9 recombination among tracks from different ancestral populations as effective
10 recombination, let $u_i(t)$ be the effective recombination rate for tracks from the i th
11 ancestral population that entered at generation t , then

$$u_i(t) = \sum_{k=t}^T (1 - H_i(k)). \quad (4)$$

12 If the end is ignored, a chromosome from the i th ancestral population that entered at
13 generation t is expected to be split into $u_i(t)$ pieces per unit length (unit in
14 Morgan). Then the contribution of ancestral tracks from i th ancestral population that
15 entered at generation t to the admixed population is proportional to $u_i(t)s_i(t)$. Let
16 X_i be the length of ancestral tracks of the i th ancestral population at generation T ,
17 and $f_i(x; M)$ be the distribution of X_i , then

$$f_i(x; M) = \sum_{t=1}^T \frac{u_i(t)s_i(t)}{\sum_{t=1}^T u_i(t)s_i(t)} u_i(t)e^{-u_i(t)x}. \quad (5)$$

1 Due to the limited accuracy in local ancestry inference, only those relatively long
 2 tracks are reliable (POOL and NIELSEN 2009; GRAVEL 2012). Therefore, we are also
 3 interested in the conditional length distribution of ancestral tracks longer than a
 4 specific threshold C ,

$$f_i(x; M|x > C) = \frac{f_i(x; M)}{\int_C^\infty f_i(x; M)dx} = \sum_{t=1}^T \frac{u_i(t)s_i(t)}{\sum_{t=1}^T u_i(t)s_i(t)e^{-u_i(t)C}} u_i(t)e^{-u_i(t)x}. \quad (6)$$

5 With the length distribution of ancestral tracks (Formula (5)), we can easily deduce
 6 the expectation and variance of X_i as

$$E(X_i) = \int_0^\infty x f_i(x; M)dx = \frac{\sum_{t=1}^T s_i(t)}{\sum_{t=1}^T u_i(t)s_i(t)} \quad (7)$$

7 and

$$Var(X_i) = E(X_i^2) - E(X_i)^2 = \frac{\sum_{t=1}^T \left(\frac{2s_i(t)}{u_i(t)}\right)}{\sum_{t=1}^T u_i(t)s_i(t)} - \left(\frac{\sum_{t=1}^T s_i(t)}{\sum_{t=1}^T u_i(t)s_i(t)}\right)^2. \quad (8)$$

8 Parameter Estimation and Model Selection

9 As the parameters of admixture events are fully determined by the matrix M , once M
 10 was accurately estimated, we could fully recover the history of the admixed
 11 population. MLE can be used to obtain the estimation of M . By utilizing the ancestral
 12 tracks inferred from the data, a likelihood function can be computed with the length
 13 distributions of ancestral tracks. The log-likelihood function of ancestral tracks from
 14 i th ancestral population $L_i(M)$ has the form

$$L_i(M) = \sum_{j=1}^{n_i} \log f_i(x_{ij}; M),$$

1 where $\{x_{ij}\}_{1 \leq j \leq n_i}$ are the observed length of tracks from the i th ancestral
2 population in the admixed population. Then the log-likelihood function of ancestral
3 tracks of the admixed population $L(M)$ is

$$L(M) = \sum_{i=1}^K L_i(M).$$

4 Then the estimator of M is

$$\hat{M} = \underset{M}{\operatorname{arg\,max}} L(M).$$

5 where M satisfies the properties in the above subsection.

6 However, with the increase in the number of parameters, it is complex and
7 time-consuming to find the optimal solution and too many parameters could lead to
8 over-fitting. In a real situation, we can propose several candidate models with prior
9 knowledge in which the number of parameters is dramatically reduced, thus the
10 problem is simplified to estimating parameters for each candidate model and selecting
11 the most suitable one. If we have obtained the parameters of the candidate models, we
12 can compare the models in a pair-wise fashion by using either AIC or LRT. Here, the
13 two models are regarded to be ‘nested’ if one of the models constitutes a special case
14 of the other (LEWIS F *et al.* 2011). When the two competing models are nested, we use
15 LRT to select the model; otherwise we use AIC.

16 **THREE TYPICAL MODELS**

17 **The Distribution of Ancestral Tracks under a HI, GA, and CGF Model**

18 In this subsection, we demonstrate that, with the length distribution of the general
19 model, we can easily deduce the length distributions of ancestral tracks under three
20 typical models aforementioned in previous studies (JIN *et al.* 2012). By restricting the

1 number of ancestral populations to be two in the general model, if only one pulse of
2 gene flow is allowed, the model turns into a HI model; if extra equal gene flows from
3 both ancestral populations are allowed, the model turns into a GA model; if extra
4 equal gene flows from only one of the ancestral populations are allowed, the model
5 turns into a CGF model (Figure S2). Thus these three models are all special cases of
6 our general model. There are only two parameters in each of the three models: the
7 admixture proportion m and the admixture time T . Easily we can obtain the
8 distribution of the ancestral tracks of these three models from Formula (5). The
9 detailed calculation is in Supporting Information.

10 For simplicity, we define the ancestral population with the minor ancestry
11 contribution as population 1, and the corresponding proportion is m . For the HI
12 model (Figure S2 [A]), the ancestry proportions from population 1 and population 2 at
13 generation t are

$$m_1(t) = \begin{cases} m, & t = 1 \\ 0, & 2 \leq t \leq T \end{cases} \text{ and } m_2(t) = \begin{cases} 1 - m, & t = 1 \\ 0, & 2 \leq t \leq T \end{cases}$$

14 Then the length distribution of ancestral tracks from population 1 is

$$f_1(x; m, T) = (1 - m)T e^{-(1-m)Tx}. \quad (9)$$

15 We can also get the expectation and variance of the length of the ancestral tracks from
16 Formula (7) and Formula (8),

$$E(X_1) = \frac{1}{(1 - m)T}; \quad (10)$$

$$Var(X_1) = \frac{1}{(1 - m)^2 T^2}. \quad (11)$$

17 Substituting m with $1 - m$ in Formulas (9), (10), and (11), we can obtain the length
18 distribution, expectation, and variance of ancestral tracks from population 2,

1 respectively. These two distributions are identical to the ones in previous studies
 2 ([GRAVEL 2012](#); [JIN *et al.* 2014](#)).

3 For the GA model ([Figure S2 \[B\]](#)), the ancestry proportions from population 1 and
 4 population 2 at generation t are

$$m_1(t) = \begin{cases} m, & t = 1 \\ m/T, & 2 \leq t \leq T \end{cases} \text{ and } m_2(t) = \begin{cases} 1 - m, & t = 1 \\ (1 - m)/T, & 2 \leq t \leq T \end{cases}$$

5 Then the length distribution of ancestral tracks from population 1 is

$$f_1(x; m, T) = (1 - m) \left[\frac{T^2 e^{-T(1-m)x} + \frac{1}{T} \sum_{t=2}^T (T - t + 1)^2 W_t e^{-(T-t+1)(1-m)x}}{T + \frac{1}{T} \sum_{t=2}^T (T - t + 1) W_t} \right], \quad (12)$$

6 where $W_t = \left(1 - \frac{1}{T}\right)^{1-t}$. The expectation and variance of the ancestral tracks are

$$E(X_1) = \frac{T + \sum_{t=2}^T W_t}{(1 - m)[T^2 + \sum_{t=2}^T (T - t + 1) W_t]}, \quad (13)$$

$$Var(X_1) = \frac{1}{(1 - m)^2} \left[\frac{2 \left(1 + \sum_{t=2}^T \frac{W_t}{T - t + 1}\right)}{T^2 + \sum_{t=2}^T (T - t + 1) W_t} - \left(\frac{T + \sum_{t=2}^T W_t}{T^2 + \sum_{t=2}^T (T - t + 1) W_t} \right)^2 \right]. \quad (14)$$

7 Substituting m with $1 - m$ in Formulas (12), (13), and (14), we can get the length
 8 distribution, expectation and variance of ancestral tracks from population 2,
 9 respectively.

10 For the CGF model ([Figure S2 \[C\]](#)), the ancestral population that contributes only
 11 one pulse of gene flow is treated as a gene flow recipient and the one that contributes
 12 continuously as gene flow donor. Here, we divide the CGF model into two
 13 sub-models. If population 1 is a gene flow recipient, we denote it as a CGFR model;
 14 otherwise we denote it as a CGFD model.

15 In the case of a CGFR model, the ancestry proportions from population 1 and
 16 population 2 at t generation are

$$m_1(t) = \begin{cases} 1 - \alpha, & t = 1 \\ 0, & 2 \leq t \leq T \end{cases} \text{ and } m_2(t) = \alpha, 1 \leq t \leq T,$$

1 where $\alpha = 1 - m^{1/T}$. Then the length distributions of ancestral tracks from the two

2 ancestral populations are

$$f_1(x; m, T) = \left(T - \frac{(1-m)m^{1/T}}{1-m^{1/T}} \right) e^{-\left(T - \frac{(1-m)m^{1/T}}{1-m^{1/T}} \right) x}, \quad (15)$$

$$f_2(x; m, T) = \frac{\sum_{t=1}^T m^{-t/T} (m^{t/T} - m^{(T+1)/T})^2 e^{-\left(\frac{m^{t/T} - m^{(T+1)/T}}{1-m^{1/T}} \right) x}}{\sum_{t=1}^T (1 - m^{(T+1-t)/T}) (1 - m^{1/T})}. \quad (16)$$

3 The expectations and variances of the ancestral tracks are

$$E(X_1) = \frac{(1 - m^{1/T})}{T(1 - m^{1/T}) - (1 - m)m^{1/T}},$$

$$E(X_2) = \frac{(1 - m)}{m} \frac{(1 - m^{1/T})}{T(1 - m^{1/T}) - (1 - m)m^{1/T}};$$

$$\text{Var}(X_1) = \left(\frac{(1 - m^{1/T})}{T(1 - m^{1/T}) - (1 - m)m^{1/T}} \right)^2,$$

$$\text{Var}(X_2) = \frac{2(1 - m^{1/T})^3 \sum_{t=1}^T \frac{m^{-t/T}}{m^{t/T} - m^{(T+1)/T}}}{T(1 - m^{1/T}) - (1 - m)m^{1/T}} - \left(\frac{(1 - m)(1 - m^{1/T})}{mT(1 - m^{1/T}) - m(1 - m)m^{1/T}} \right)^2.$$

4 In the case of a CGFD model, we just replace m for $1 - m$ in Formula (15) and

5 (16), and obtain the length distribution of ancestral tracks from population 2 and

6 population 1, respectively.

7 **Parameter Estimation and Model Selection under HI, GA, and CGF Models**

8 As discussed above, there are only two parameters m and T in the HI, GA and CGF

9 models. As for m , with the inferred ancestral tracks in an admixed population, we

10 divide the total length of tracks from population 1 by the total length of tracks, and

11 obtain an estimator \hat{m} of the admixture proportion. Interestingly, from the expectation

12 of the ancestral tracks from two ancestral populations in the HI, GA and CGF models,

1 we find that the expectation ratio between population 1 and population 2 relies only
2 on m ,

$$\frac{E(X_1)}{E(X_2)} = \frac{m}{1-m}, \quad (17)$$

3 thus we provide an alternative way to obtain the estimator \hat{m} ,

$$\hat{m} = \frac{E(X_1)}{E(X_2) + E(X_1)}. \quad (18)$$

4 Generally, if there are only two ancestral populations, Formula (17) always holds
5 whatever the admixture model is. The proof is in [Supporting Information](#).

6 As for admixture time T , the estimation relies on the model assumed. Different
7 models give different estimations of T so that we first need to assume a model. Here
8 we regard the HI, GA, CGFR and CGFD models as the candidate models, and use
9 MLE to estimate the admixture time T and AIC to select the optimal model as
10 following: First, by utilizing the inferred ancestral tracks, we calculate the admixture
11 proportion and determine population 1; secondly, with the estimator \hat{m} , maximizing
12 the likelihood under model assumption HI, GA, CGFR, and CGFD, we obtain the
13 maximum likelihoods $L_{max}(\text{HI})$, $L_{max}(\text{GA})$, $L_{max}(\text{CGFR})$ and $L_{max}(\text{CGFD})$ and
14 corresponding optimal times \hat{T} . Because each pair of these models is not nested, thus
15 here we use AIC to select the optimal model. The value AIC can be calculated by the
16 formula

$$AIC = 2k - 2l n(L_{max}),$$

17 where k is the number of parameters and L_{max} is the maximized value of the
18 likelihood function. The number of parameter of these models are the same, thus at
19 the end of the comparison, we find that the problem is equivalent to finding the model

1 with the highest likelihood. Thus, the model with the highest likelihood is chosen as
2 the optimal model, and the corresponding parameters as the final results. These
3 routines are implemented in our *AdmixInfer*. We also apply the bootstrapping
4 technique in *AdmixInfer* and give a bootstrap estimation and confidence interval (CI)
5 of the admixture time.

6 **MATERIALS AND METHODS**

7 **General Settings of Simulation Studies**

8 Simulation studies were performed to evaluate the correctness of the length
9 distribution of ancestral tracks under the general model, and the performance of
10 *AdmixInfer* under three typical models. The following settings remained the same
11 under all situations simulated if with no further modification: The population size of
12 the admixed population was simply set to 5,000 and remained constant in our
13 simulations and the length of simulated chromosome was 3.0 Morgan, which
14 approximated the length of chromosome 1 of the human genome. We simulated one
15 chromosome each time and a pair of chromosomes represented an “individual.” At the
16 end of simulation, 400 “individuals” (genome length of an “individual” approximated
17 1/10 of the length of a human individual) were sampled and the ancestral tracks were
18 directly recorded.

19 **Evaluate the Theoretical Distribution under the General Model**

20 To test the accuracy of the length distribution of ancestral tracks under the general
21 model, we simulated several general representative cases with three or four ancestral
22 populations and one or more waves of admixture. Simulation 1A was to simulate one

1 pulse admixture with three ancestral populations: admixture started 50 generations
2 ago with proportions 10%, 30% and 60%, respectively. Simulation 1B was to simulate
3 discrete multiple-waves admixture with three ancestral populations: the admixture
4 started 50 generations ago with initial proportions 10%, 40% and 50%, respectively,
5 and population 1 contributed extra 10% ancestries each 10 generations later.
6 Simulation 1C was to simulate continuous multiple-waves admixture with three
7 ancestral populations: the admixture started 50 generations ago with initial
8 proportions 10%, 30% and 60%, respectively; population 1 contributed an extra 0.2%
9 ancestries for each generation afterwards. And simulation 1D was to simulate
10 multiple-waves admixture with four ancestral populations that arrived at different
11 times: populations 1 and 2 firstly admixed 50 generations ago with proportions 40%
12 and 60%, population 3 entered 40 generations ago with proportion 20%, and
13 population 4 entered 30 generations ago with proportion 10%. The simulations were
14 repeated 5 times, and the ancestral tracks in the admixed population were recorded
15 and the length distribution was compared to the theoretical distribution. Detailed
16 parameters for the simulations were provided in [Table S1](#).

17 **Evaluate the Performance of *AdmixInfer***

18 Then we focused on evaluating the performance of *AdmixInfer* under three typical
19 models; i.e. HI, GA and CGF. The proportions of admixture varied from 10% to 50%
20 in steps of 10% for the symmetric admixture models (HI and GA) and varied from 10%
21 to 90% in steps of 10% for the asymmetric admixture model (CGF). We set the
22 admixture time as 5, 10, 20, 30, ..., 200 generations. The ancestral tracks in admixed

1 populations were also recorded as previous simulations. Each simulation here was
2 repeated 10 times and, in total, 4,200 simulations were carried out under HI, GA and
3 CGF models. *AdmixInfer* was applied to the simulated data with the default settings;
4 the results were recorded and summarized.

5 In real situations, we could only accurately infer the ancestral tracks longer than a
6 specific threshold due to methods' limitations in local ancestry inferences. To make
7 our method more feasible to real cases, we also evaluated the robustness of our
8 method under different thresholds ranging from 0 centi-Morgan (cM) to 2 cM in step
9 of 0.1 cM, with the dataset simulated in previous evaluations.

10 We also evaluated the performance of *AdmixInfer* with different sample sizes. We
11 simulated populations starting with the admixture of 50 and 100 generations ago
12 under HI, GA, and CGF models, with admixture proportions 30%:70%. At the end of
13 the simulation, 10, 20, 50, 100, 200, and 500 "individuals" were sampled,
14 corresponding to 1, 2, 5, 10, 20, and 50 human samples. *AdmixInfer* was applied to
15 the simulated dataset without discarding short tracks.

16 Finally, we tested the performance of our method with data simulated by real
17 populations and inferred ancestral tracks. Simulations were carried out with real
18 populations YRI and CEU as parental populations under different models (30% YRI
19 ancestry and 70% CEU ancestry) with admixture time 10, 20, 50 and 100 generations.
20 Here we simulated with the data of chromosome 1 and sampled 25 "individuals" at
21 the end of the simulation, and each simulation was repeated 10 times. Then the local
22 ancestry of the simulated populations was inferred by HAPMIX (PRICE *et al.* 2009).

1 With the derived ancestral tracks, *AdmixInfer* was used to select the optimal model
2 and estimate generations accordingly with the tracks longer than 1 cM.

3 **Apply to Real Datasets**

4 We applied our method to some real datasets. First, the histories of African Americans
5 and Mexicans are relatively clear, thus they can be used to test the performance of our
6 method. The data of African Americans, Mexicans and reference populations CEU
7 and YRI were obtained from HapMap project phase III ([INTERNATIONAL HAPMAP *et*](#)
8 [al. 2010](#)), and the reference populations that represented American Indian ancestry
9 were obtained from HGDP dataset. Then we also applied our method to several
10 HGDP populations from South Asia ([LI *et al.* 2008](#)), which showed evidence of
11 population admixture from previous studies ([PATTERSON *et al.* 2012](#); [HELLENTHAL *et*](#)
12 [al. 2014](#)). Haplotype phasing was performed by SHAPEIT 2 ([DELANEAU *et al.* 2012](#)).
13 Local ancestry was inferred by HAPMIX ([PRICE *et al.* 2009](#)). According to the prior
14 knowledge, the generations settings in HAPMIX inference were 10, 20 and 50 for
15 African Americans, Mexicans and South Asian populations, respectively
16 ([HELLENTHAL *et al.* 2014](#)). *AdmixInfer* was used to select the optimal model and
17 admixture time accordingly with the tracks longer than 1 cM. We also performed
18 bootstrapping 100 times to obtain confidence of model selection, and calculated the
19 95% confidence intervals of the generations inferred.

20 **RESULTS**

21 **Theoretical and Simulated Distributions of Ancestral Tracks Match Well**

1 With the length distributions of ancestral tracks under the general model, we could
2 easily sketch the curves of theoretical length distributions of ancestral tracks under a
3 given model (Figure 2, solid line). By putting the theoretical and simulated length
4 distributions of ancestral tracks together, we clearly observed that the theoretical and
5 simulated distributions of ancestral tracks matched well, for all the situations
6 simulated and all the replicates (Figure 2). It showed that the theoretical length
7 distribution of ancestral tracks for the general model, which we deduced, was
8 reasonable and accurate.

9 ***AdmixInfer* Performs well in Parameters Estimation and Model Selection**

10 With the extensively simulated data, we could systematically evaluate the
11 performance of our method in parameters estimation and model selection. For the
12 simulated admixed populations under the HI model or CGFD model, our method
13 could always distinguish the right model in all our simulations; for the CGFR model,
14 our method could distinguish the right model with accuracy of 97.0%; and for GA
15 model, our method could distinguish the right model with accuracy of 93.0% (Table
16 1). Moreover, the specificity of our method was over 97% for all the situations
17 simulated. The sensitivity of the GA and CGFR models and the specificity of the HI
18 and CGFD models under different admixture proportions and different admixture
19 times were shown in Figure S3. We found that the simulations in which our method
20 could not distinguish the right ones were mostly observed in these simulations with
21 very recent admixture times and small admixture proportions (Figure S3 and Table
22 S2-S6). We also found that almost all CGFR models were only wrongly distinguished

1 as HI models, and GA models as CGFD models (Table S6). This is also reasonable,
2 because the CGFR model is close to the HI model, so that CGFR model is more likely
3 to be distinguished as a HI model. The same reason applies for that the GA model
4 being wrongly distinguished as the CGFD model.

5 Note that there were only two parameters m and T for the three typical models.
6 Our method also performed well in estimating parameters m and T for the three
7 typical models. Regarding the admixture proportion m , estimations were very close
8 to the pre-settings in simulations, and the small deviation was due to random drift in
9 simulation with finite population sizes and only a subset of individuals being sampled
10 at the end of the simulation. Time estimations for wrongly distinguished models were
11 meaningless, and thus should be discarded. Results showed that our method can
12 estimate admixture times with high accuracy (Figure 3, Figure S4-S7, and Table
13 S2-S5). For HI and CGFD models, results showed high consistency with the time
14 simulated, while a slight underestimation occurred for GA and CGFR models. For all
15 these models that were simulated, the time estimated for a small proportion was less
16 accurate than that for a larger proportion. We defined the relative errors between true
17 admixture times and inferred admixture times as

$$E(T) = \frac{|T - \hat{T}|}{T},$$

18 where T is the true admixture time and \hat{T} is the estimation of the admixture time.
19 Under the situation of a certain admixture proportion and a certain model, we defined
20 the average relative error \bar{E} on different values of admixture time as

$$\bar{E} = \frac{1}{n} \sum_{i=1}^n E(T_i).$$

1 We found that when the admixture proportion was 0.1, the relative errors of CGFR
2 and CGFD were 6.43% and 5.89%, respectively. For the other cases, the relative
3 errors were all less than 4% (Table 2). In conclusion, no matter the model selection or
4 parameters estimation, our method performed well.

5 **Robustness for Different Thresholds**

6 To test the robustness of our method for different thresholds, we tested our method
7 under different thresholds varying from 0 cM to 2 cM in steps of 0.1 cM. The results
8 showed that our method was robust to thresholds, except the GA model with a larger
9 time (Figure 4). When a larger threshold is taken, less information is kept for ancient
10 admixture events. Although keeping all the information to estimate admixture times is
11 better, we must balance the trade-off between information and accuracy, because the
12 accuracy of local ancestry inference is not so good for short ancestral tracks due to
13 method limitations. Take HI for an example, the probability p of ancestral tracks
14 larger than a specific threshold C is

$$p = 1 - \int_0^C (1-m)T e^{-(1-m)Tx} dx = e^{-(1-m)TC}.$$

15 Therefore, with an increase in threshold C , less information is kept. Here, we
16 provided a reference table of the probability p under different admixture times and
17 proportions (Table S7). For example, when $T = 200$, $m = 0.1$, if we set the threshold
18 C as 2cM, the probability that the tracks exceed C is only 2.7%.

19 **Small Sample Sizes also Give Good Estimations**

1 To test the performance of *AdmixInfer* with different sample sizes, we evaluated the
2 models with 10, 20, 50, 100, 200, and 500 “individuals” (corresponding to 1, 2, 5, 10,
3 20, and 50 human samples). Results showed that *AdmixInfer* was insensitive to
4 sample sizes. Even with only one human sample, it could distinguish the right model
5 and estimate the admixture time with high accuracy (Figure 5). However, considering
6 the accuracy of local ancestry, short tracks were usually discarded. The information
7 kept for extremely small sample sizes might not be sufficient to give a clear picture of
8 the history of a population. Therefore, relatively larger sample sizes were
9 recommended.

10 **Error Analysis**

11 When we use our method to infer the history of a real admixed population, there are
12 two kinds of errors that may influence the accuracy of inference. The first kind of
13 error is caused by the assumptions in deducing the length distribution of ancestral
14 tracks. In the derivation, for simplicity, we ignored the end of the chromosome and
15 the drift. For this kind of error, we have used the simulation data to demonstrate that
16 the accuracy of the inference was neglectable (Figure 3 and Table 1). The second kind
17 of error is caused by the local ancestry inference. In our study, local ancestral tracks
18 are inferred by HAPMIX software. Here we used the simulation data with ancestry
19 populations YRI and CEU to analyze the influence of this kind of error. For all the
20 cases we simulated, only the HI model with an admixture time 100 generations was
21 wrongly distinguished as a GA model (Table S8). We also found that for the case of
22 large admixture times, the error of local ancestry inference will cause underestimation

1 of admixture time. When the admixture time is large, the ancestral tracks will be short.
2 However, the method of inferring ancestral tracks cannot effectively determine short
3 tracks. Thus, it will influence the accuracy of our method in inferring admixture times
4 and model selections.

5 **Real Data Analysis**

6 Our method for parameters estimation and model selection under three typical models
7 implemented in *AdmixInfer* was applied to a real dataset. First, African American was
8 inferred as a GA model and the admixture time was 12 generations ago ([Table 3](#)).
9 When 29 years per generation was assumed according to previous investigation
10 ([FENNER 2005](#)), it was about 350 years ago, which was consistent with the recorded
11 history that most African ancestors arrived in America (via slave trading) after the
12 seventeenth century. The slave trade continued until the nineteenth century and after
13 that, many African people settled down in America. Gene flows from Africa and
14 Europe would have continuously contributed to the African American gene pool and
15 thus the GA model matched the recorded history well. Similarly, the model for
16 Mexicans was inferred as a GA model and the admixture time was 18 generations ago
17 (522 years before present), which was also consistent with the time of the exploration
18 of the new world and the arrival of Europeans. The GA model indicates continuous
19 contact and admixture of Europeans and American Indians.

20 Finally, we studied the admixture histories of several HGDP populations from
21 South Asia. Previous studies have shown that the populations from South Asia have
22 admixed ancestries mainly from Europe and East Asia ([LI et al. 2008](#); [HELLENTHAL et](#)

1 *al.* 2014). Regarding the admixture proportions, our estimations based on *AdmixInfer*
2 (Table 3) were consistent with previous estimations from the three-population test
3 (PATTERSON *et al.* 2012). Regarding the admixture model and time, the populations
4 Balochi, Brahui, Burusho, Kalash, Pathan and Sindhi from Pakistan in South Asia
5 were all inferred to as CGF (Sardinian as donor), which indicated extra gene flows
6 from European ancestry after initial admixture. The initial admixture times estimated
7 ranged from 107 to 96 generations ago. When 29 years per generation was assumed
8 according to previous investigations (FENNER 2005), these estimations
9 (1103BC-784BC) coincided with the migration of Indo-Aryan speaking people into
10 the Indian subcontinent. Extra gene flows from European ancestry might be
11 contributed to by the rising of empires during the following centuries. In the case of
12 the populations of Hazara and Uygur, they not only showed very similar admixture
13 proportions, but also showed the same admixture model and very close admixture
14 times; i.e., 67 and 70 generations ago, respectively. The Hazara population mainly
15 settled in Afghanistan and Pakistan, while the Uygur population mainly settled in
16 West China, and both populations are connected along the Silk Road. It was feasible
17 to receive continuous gene flows from both European and East Asian ancestries.
18 These similarities also indicated a possible close relationship or shared histories
19 between these two populations.

20 In summary, our method showed good performance in inferring the admixture
21 history of African Americans and Mexicans. The admixture scenarios in South Asian
22 populations were more complex than expected. However, with our method, the

1 analysis could shed light on the mysterious histories of these populations.

2 **DISCUSSION**

3 In this work, we proposed a general model to describe the admixture history with
4 multiple ancestral source populations and multiple-wave admixtures. We showed the
5 length distribution of ancestral tracks and some of its useful properties under this
6 general model. We thus provided a theoretical framework to study population
7 admixture history. With the general framework, we focused on studying three special
8 cases of the admixture models (HI, GA, and CGF) and developed a method to
9 estimate the admixture proportion, admixture time and determine the optimal model
10 simultaneously. Our simulations showed that the theoretical distribution of ancestral
11 tracks was consistent with our theoretical prediction, and our method was precise and
12 efficient in inferring population history under three typical models.

13 In the efforts of model selection, we found that the simulations in which our
14 method was not able to determine the correct model, were mostly those cases with
15 recent admixture times and minor admixture proportions. The possible reason for
16 incorrect determination was that we ignored the chromosome ends in deducing the
17 theoretical length distribution. When the admixture proportion and times were small,
18 the chromosomes without “observable” recombination were over-represented in the
19 ancestral tracks. (Figure S8). Our further simulations showed that when the
20 chromosome length increased, the accuracy of our method was enhanced.
21 Furthermore, we note that the length distributions of ancestral tracks have no
22 relationship with the population size. Thus, the change of population size does not

1 affect the time estimation. Simulations under different demographic models also
2 supported it (Figure S9).

3 The efficiency of our method could also be influenced by the validity of the local
4 ancestry inference. To improve the reliability of the inference, we suggest using the
5 ancestral tracks longer than a certain threshold C . However, when the threshold
6 became large, some ancient admixture information disappeared rapidly. In principle, if
7 short ancestral tracks could be precisely detected, our method is promising in
8 recovering even more ancient admixture history, such as the admixture between
9 modern human and Neanderthals (PRÜFER *et al.* 2014; SANKARARAMAN *et al.* 2014).

10 Though we proposed a general framework and relevant principles to infer the
11 population history under the general model, finding optimal estimation for parameters
12 is a challenging work with high dimensionality. Currently, our method implemented
13 in *AdmixInfer* is focusing on the three typical models. For the real admixed
14 populations, the admixture history is always complex, such as discrete multiple-waves
15 admixture. Under such circumstances, the length distribution of ancestral tracks under
16 the general model is still broadly useful and applicable. Therefore, based on this
17 framework, to infer more complicated admixture history is a problem to be solved in
18 the future.

19 **ACKNOWLEDGEMENTS**

20 This work was supported by the Strategic Priority Research Program of the Chinese
21 Academy of Sciences (CAS) (XDB13040100), by National Natural Science

1 Foundation of China (NSFC) grants (91331204 and 31171218), 973 Project
2 (2011CB808000), the Fundamental Research Funds for the Central Universities
3 (2011JBZ019), the National Excellent Doctoral Dissertation Foundation of PR China
4 (FANEDD 201312), National Center for Mathematics and Interdisciplinary Sciences
5 of CAS, and the Key Laboratory of Random Complex Structures and Data Science,
6 CAS (2008DP173182). S.X. also gratefully acknowledges the support of the National
7 Program for Top-notch Young Innovative Talents of The "*Wanren Jihua*" Project.

8

1 **Figure Legends**

2 **Figure 1.** The general admixture model. Here we illustrated an admixed population
3 with K ancestral populations, which started to admix T generations ago. The gene
4 flows from each ancestral population could be zero at a specific generation. POP k
5 represents the reference population k .

6 **Figure 2.** Theoretical and simulated distributions of ancestral tracks under some
7 representative admixture scenarios. A: three reference populations admixed once at 50
8 generations ago; B: three reference populations admixed at 50 generations ago and
9 population 1 contributed an extra 10% ancestry each 10 generations later; C: three
10 reference populations admixed at 50 generations ago and population 1 contributed an
11 extra 0.2% ancestry every generation later; D: two reference populations admixed 50
12 generations ago, the third reference population contributed 20% ancestry 40
13 generations ago, and the fourth reference population contributed 10% ancestry 30
14 generations ago.

15 **Figure 3.** Mean generations estimated from simulation. Each dot denotes the mean
16 generation of ten simulation replicates. A: mean generations estimated under the HI
17 model; B: mean generations estimated under the GA model; C: mean generations
18 estimated under the CGF model (population 1 as gene flow recipient); and D: mean
19 generations estimated under the CGF model (population 1 as gene flow donor).
20 Different colors represent different simulated proportions of population one.

1 **Figure 4.** Generation estimated with different thresholds from simulation. Models
2 simulated were HI(A), GA(B), CGFR(C) and CGFD(D). The simulated admixture
3 proportion was 30%. Different colors represent different simulated generations.

4 **Figure 5.** Admixture time in generations estimated with different sample size. Models
5 simulated are HI (A), GA (B), CGFR (C), and CGFD (D). Different colors represent
6 different simulated generations. The simulated admixture proportions was 30%.

7

1 **References**

- 2 AKAIKE, H., 1998 Information theory and an extension of the maximum likelihood
3 principle, pp. 199-213 in *Selected Papers of Hirotugu Akaike*. Springer.
- 4 DELANEAU, O., J. MARCHINI and J. F. ZAGURY, 2012 A linear complexity phasing method
5 for thousands of genomes. *Nat Methods* **9**: 179-181.
- 6 FENNER, J. N., 2005 Cross-cultural estimation of the human generation interval for use
7 in genetics-based population divergence studies. *Am J Phys Anthropol* **128**:
8 415-423.
- 9 GRAVEL, S., 2012 Population genetics models of local ancestry. *Genetics* **191**: 607-619.
- 10 HELLENTHAL, G., G. B. BUSBY, G. BAND, J. F. WILSON, C. CAPELLI *et al.*, 2014 A genetic atlas
11 of human admixture history. *Science* **343**: 747-751.
- 12 INTERNATIONAL HAPMAP, C., D. M. ALTSHULER, R. A. GIBBS, L. PELTONEN, D. M. ALTSHULER *et*
13 *al.*, 2010 Integrating common and rare genetic variation in diverse human
14 populations. *Nature* **467**: 52-58.
- 15 JIN, W., R. LI, Y. ZHOU and S. XU, 2014 Distribution of ancestral chromosomal segments
16 in admixed genomes and its implications for inferring population history and
17 admixture mapping. *Eur J Hum Genet* **22**: 930-937.
- 18 JIN, W., S. WANG, H. WANG, L. JIN and S. XU, 2012 Exploring population admixture
19 dynamics via empirical and simulated genome-wide distribution of ancestral
20 chromosomal segments. *Am J Hum Genet* **91**: 849-862.
- 21 LEWIS F, BUTLER A and G. L., 2011 A unified approach to model selection using the
22 likelihood ratio test. *Methods in Ecology and Evolution* **2**: 155-162.

- 1 LI, J. Z., D. M. ABSHER, H. TANG, A. M. SOUTHWICK, A. M. CASTO *et al.*, 2008 Worldwide
2 human relationships inferred from genome-wide patterns of variation.
3 *Science* **319**: 1100-1104.
- 4 LOH, P. R., M. LIPSON, N. PATTERSON, P. MOORJANI, J. K. PICKRELL *et al.*, 2013 Inferring
5 admixture histories of human populations using linkage disequilibrium.
6 *Genetics* **193**: 1233-1254.
- 7 PATTERSON, N., P. MOORJANI, Y. LUO, S. MALLICK, N. ROHLAND *et al.*, 2012 Ancient
8 admixture in human history. *Genetics* **192**: 1065-1093.
- 9 PICKRELL, J. K., N. PATTERSON, P. R. LOH, M. LIPSON, B. BERGER *et al.*, 2014 Ancient west
10 Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci U S A* **111**:
11 2632-2637.
- 12 POOL, J. E., and R. NIELSEN, 2009 Inference of historical changes in migration rate from
13 the lengths of migrant tracts. *Genetics* **181**: 711-719.
- 14 PRÜFER, K., F. RACIMO, N. PATTERSON, F. JAY, S. SANKARARAMAN *et al.*, 2014 The complete
15 genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**:
16 43-49.
- 17 PRICE, A. L., A. TANDON, N. PATTERSON, K. C. BARNES, N. RAFAELS *et al.*, 2009 Sensitive
18 detection of chromosomal segments of distinct ancestry in admixed
19 populations. *PLoS Genet* **5**: e1000519.
- 20 PUGACH, I., R. MATVEYEV, A. WOLLSTEIN, M. KAYSER and M. STONEKING, 2011 Dating the age
21 of admixture via wavelet transform analysis of genome-wide data. *Genome*
22 *Biol* **12**: R19.

- 1 SANKARARAMAN, S., S. MALLICK, M. DANNEMANN, K. PRUFER, J. KELSO *et al.*, 2014 The
- 2 genomic landscape of Neanderthal ancestry in present-day humans. *Nature*
- 3 **507**: 354-357.
- 4 WILKS, S. S., 1938 The large-sample distribution of the likelihood ratio for testing
- 5 composite hypotheses. *The Annals of Mathematical Statistics* **9**: 60-62.
- 6 XU, S., W. HUANG, J. QIAN and L. JIN, 2008 Analysis of genomic admixture in Uyghur and
- 7 its implication in mapping strategy. *Am J Hum Genet* **82**: 883-894.
- 8
- 9
- 10

1 **Table 1. The accuracy of our methods in model selection under three typical models.**

Model	TP	FN	TN	FP	Sensitivity (%)	Specificity (%)
HI	1050	0	3120	30	100.0	99.0
GA	977	73	3150	0	93.0	100.0
CGFR	1019	31	3150	0	97.0	100.0
CGFD	1050	0	3076	74	100.0	97.7

2 HI: Hybrid isolation model; GA: Gradual admixture model; CGFR: Continuous gene flow model
3 (population 1 as recipient); CGFD: Continuous gene flow model (population 1 as donor). TP: True
4 positive; FP: False positive; TN: True negative; FN: False negative. Sensitivity=TP / (TP+FN);
5 Specificity=TN / (TN+FP).
6
7

8 **Table 2. The average relative error \bar{E} for different values of admixture proportions.**

m	HI	GA	CGFR	CGFD
0.1	0.02304	0.02245	0.06433	0.05893
0.2	0.01327	0.03171	0.03917	0.02396
0.3	0.01149	0.02939	0.02439	0.01629
0.4	0.01260	0.02969	0.01704	0.01386
0.5	0.01158	0.02547	0.01474	0.01475

9 m : Admixture proportion; \bar{E} : Average relative error on different values of admixture times. Here we
10 discard the cases of 5 and 10 generations admixture time because wrongly distinguished models
11 mainly focused on these two admixture times.

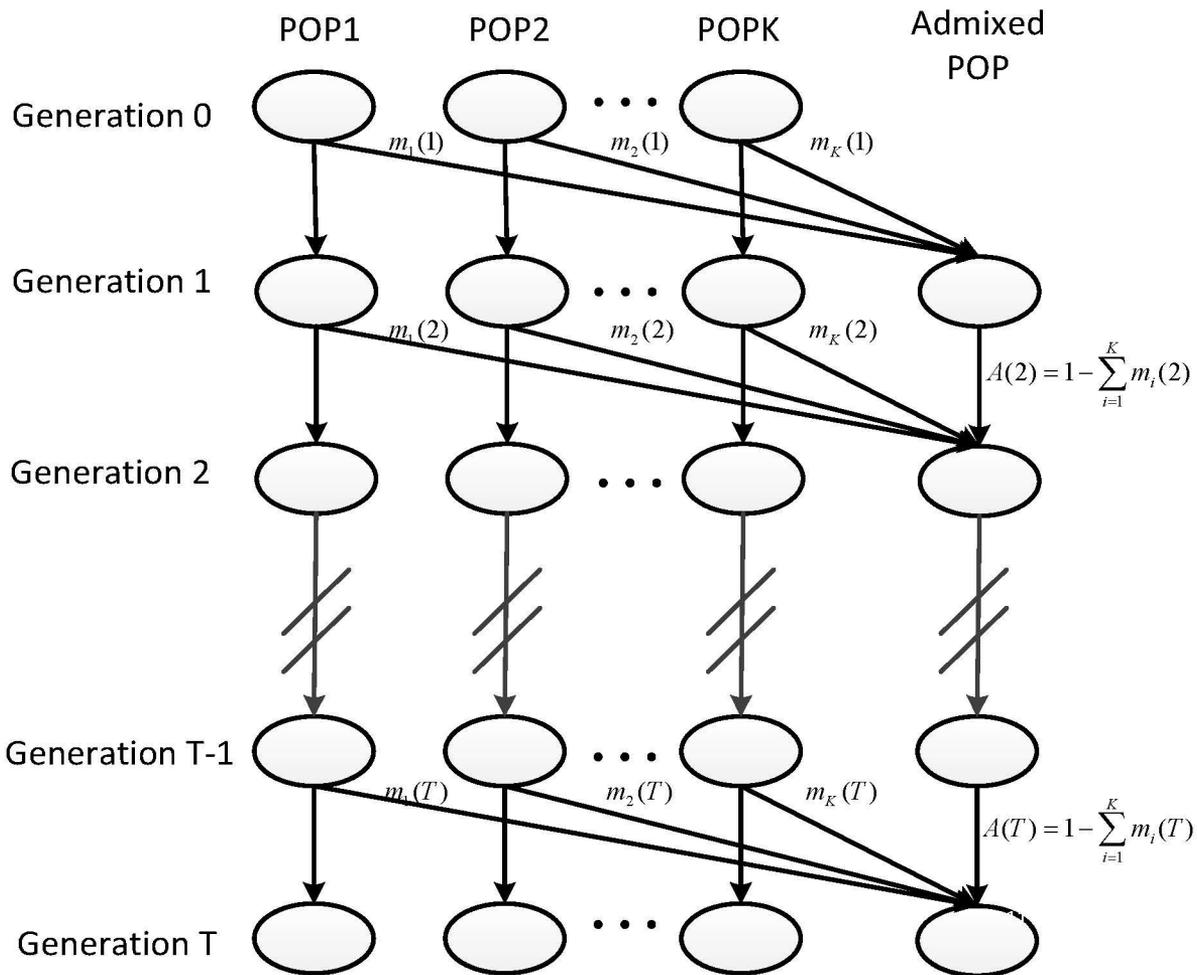
1 **Table 3. Admixture time and model inferred for real datasets.**

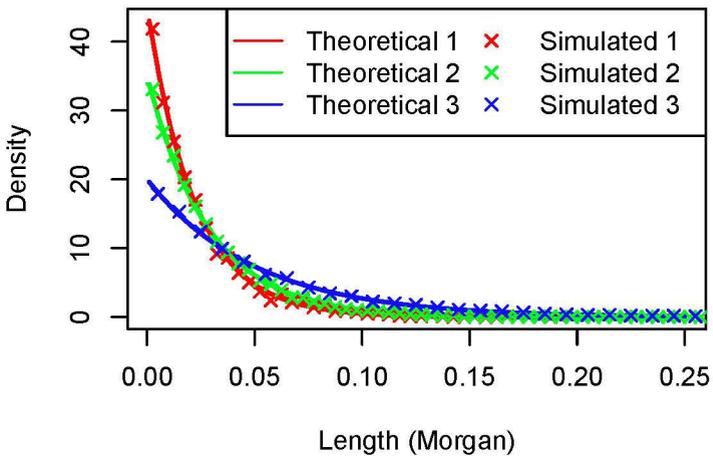
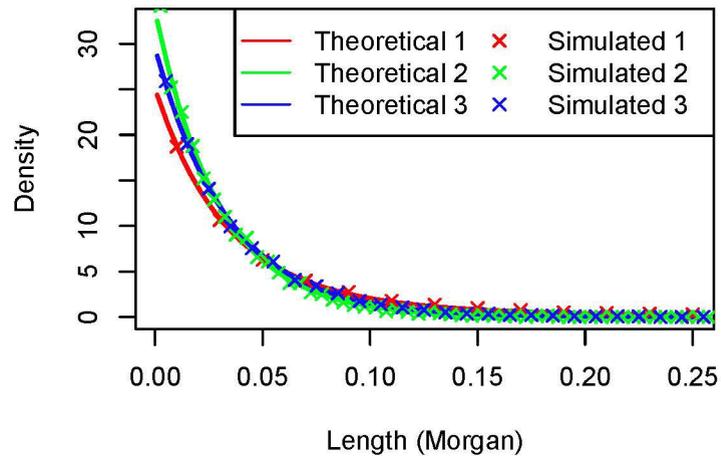
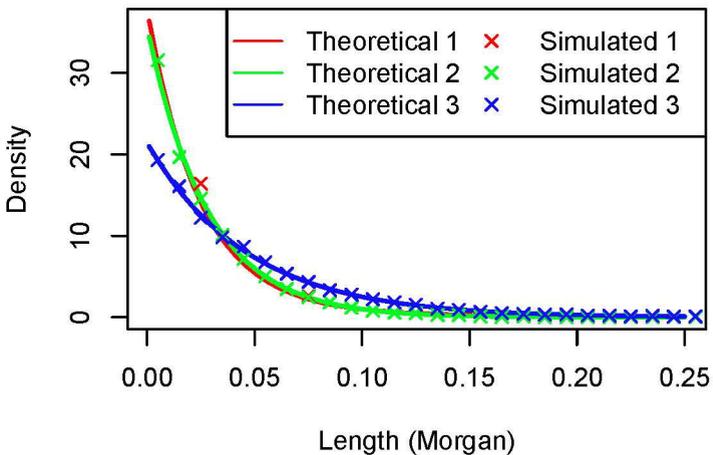
POP1	POP2	Admixed pop	m	Model	T	95%CI
CEU	YRI	ASW	0.246748	GA (99%)	12	[12,12]
CEU	AMI	MEX	0.477949	GA (100%)	18.02	[17.99, 18.05]
Japanese	Sardinian	Uygur	0.569544	GA (100%)	70.43	[70.2, 70.66]
Japanese	Sardinian	Hazara	0.558653	GA (100%)	67.16	[67.01, 67.31]
Japanese	Sardinian	Balochi	0.197582	CGF (100%) (Sardinian as donor)	104.7	[104.5, 104.8]
Japanese	Sardinian	Brahui	0.189526	CGF (100%) (Sardinian as donor)	105.2	[105, 105.3]
Japanese	Sardinian	Burusho	0.339483	CGF (100%) (Sardinian as donor)	96.52	[96.34, 96.7]
Japanese	Sardinian	Kalash	0.173549	CGF (100%) (Sardinian as donor)	100.5	[99.27, 101.6]
Japanese	Sardinian	Pathan	0.257682	CGF (100%) (Sardinian as donor)	98.49	[98.28, 98.7]
Japanese	Sardinian	Sindhi	0.271986	CGF (100%) (Sardinian as donor)	107.5	[107.4, 107.7]

2 POP1: Reference population one; POP2: Reference population two; Admixed pop: Admixed population; m : Admixture proportion of reference population one; Model:
 3 Inferred admixture model, percentage in the parenthesis is the support rate in 100 times bootstrapping; T : estimated admixture time in generation; 95%CI: 95%
 4 confidence interval of the estimated admixture time; AMI: Combined dataset of populations Maya and Pima which represent American Indian ancestry.

5

6



A**B****C****D**