

Inference of Multiple-wave Admixtures by Length Distribution of Ancestral Tracks

Xumin Ni¹⁺, Xiong Yang²⁺, Kai Yuan^{2,3+}, Qidi Feng^{2,3}, Wei Guo⁴, Zhiming Ma^{1,4*},
Shuhua Xu^{2,3,5,6*}

¹Department of Mathematics, School of Science, Beijing Jiaotong University, Beijing 100044, China;

²Chinese Academy of Sciences (CAS) Key Laboratory of Computational Biology, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences, CAS, Shanghai 200031, China;

³University of Chinese Academy of Sciences, Beijing 100049, China;

⁴Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;

⁵School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China;

⁶Collaborative Innovation Center of Genetics and Development, Shanghai 200438, China.

*Corresponding author

E-mail: xushua@picb.ac.cn (S.X.) and mazm@amt.ac.cn (Z.M.)

⁺These authors contributed equally to this work.

1 ABSTRACT

2 The ancestral tracks in admixed genomes are of valuable information for population
3 history inference. A few methods have been developed to infer admixture history
4 based on ancestral tracks. Nonetheless, these methods suffered the same flaw that
5 only population admixture history under some specific models can be inferred. In
6 addition, the inference of history might be biased or even unreliable if the specific
7 model is deviated from the real situation. To address this problem, we firstly proposed
8 a general discrete admixture model to describe the admixture history with multiple
9 ancestral populations and multiple-wave admixtures. We next deduced the length
10 distribution of ancestral tracks under the general discrete admixture model. We further
11 developed a new method, *MultiWaver*, to explore the multiple-wave admixture
12 histories. Our method could automatically determine an optimal admixture model
13 based on the length distribution of ancestral tracks, and estimate the corresponding
14 parameters under this optimal model. Specifically, we used a likelihood ratio test
15 (LRT) to determine the number of admixture waves, and implemented an
16 expectation–maximization (EM) algorithm to estimate parameters. We used
17 simulation studies to validate the reliability and effectiveness of our method. Finally,
18 good performance was observed when our method was applied to real datasets of
19 African Americans, Mexicans, Uyghurs, and Hazaras.

20

1 INTRODUCTION

2 Admixture among previously isolated populations has been a common phenomenon
3 throughout the evolution of modern humans¹⁻³. The history of population admixture
4 has a strong influence on the landscape of genetic variation in individuals from
5 admixed populations. Therefore, the population history of admixed populations can be
6 reconstructed by utilizing genetic variation information⁴⁻¹⁶.

7 A few methods have been developed to infer admixture history based on ancestral
8 tracks information¹⁰⁻¹⁶. Pool and Nielsen firstly used the length of ancestral tracks to
9 infer population history¹⁰. They introduced a theoretical framework describing the
10 length distribution of ancestral tracts and proposed a likelihood inference method to
11 estimate parameters related to historical change in migration rates. Additionally,
12 Pugach *et al.* performed wavelet transforms on the ancestral tracks in an admixed
13 population to obtain the dominant frequency of ancestral tracks to estimate the
14 admixture time^{11,16}. Jin *et al.* further explored admixture dynamics by comparing the
15 empirical and simulated distribution of ancestral tracks under 3 typical two-way
16 admixtures models, i.e., the hybrid isolation (HI) model, gradual admixture (GA)
17 model, and continuous gene flow (CGF) model¹². They later deduced the theoretical
18 distributions of ancestral tracks under HI and GA models¹⁴. Gravel extended these
19 studies to multiple ancestral populations and discrete migrations, and provided a
20 numerical estimation of tract length distribution¹³.

21 However, there was a significant shortcoming for all these methods. Before
22 estimating the parameters of admixture history, a prior admixture model was required.

1 The method by Pool and Nielsen considered a model that a target population received
 2 migrants from a source population¹⁰. Pugach *et al.*'s method was under an HI model,
 3 and Jin *et al.*'s methods were under HI, GA, and CGF models^{11,12,14}. While Gravel
 4 considered models of multiple ancestral populations and discrete migrations, a prior
 5 admixture model was also required when dealing with the problem of admixture
 6 history inference¹³. However, in data analysis, we always have little information of
 7 admixture history, and the admixture model is often uncertain for some complex
 8 admixed populations^{4,17-19}. Therefore, when the prior model deviates from the real
 9 history, these methods might be unreliable.

10 In our previous work¹⁵, we proposed some general principles in parameter
 11 estimation and model selection with the length distribution of ancestral tracks under a
 12 general model. However, with the increase of the number of parameters, it is complex
 13 and time-consuming to find the optimal solution, and too many parameters can lead to
 14 over-fitting. Thus, we only developed a method to infer admixture history under 3
 15 typical two-way admixtures models.

16 In this work, we introduced a new method to select the optimal admixture model
 17 and estimate the corresponding parameters under a general model. Firstly, we
 18 proposed a general discrete admixture model with an arbitrary number of ancestral
 19 populations and arbitrary number of admixture events. This was similar to the general
 20 model in our previous work¹⁵. Then, we deduced the theoretical distribution of
 21 ancestral tracks with some reasonable approximations under the general discrete
 22 admixture model. We selected an optimal admixture model based on the length

1 distribution of ancestral tracks. Specifically, we used a likelihood ratio test (LRT)^{20,21}
2 to determine the number of admixture waves, and employed an exhaustion method to
3 determine the order of admixtures. We then applied an expectation–maximization
4 (EM) algorithm²² to estimate parameters under the optimal model. In our method, no
5 prior knowledge about the admixture history was required, and the admixture model
6 and its corresponding parameters could both be inferred by ancestral tracks. Finally,
7 we conducted simulation studies to demonstrate the effectiveness of our method, and
8 then applied our method to African Americans and Mexicans from the HapMap
9 project phase III dataset²³, and Uyghurs and Hazaras from the Human Genome
10 Diversity Project (HGDP) dataset¹.

11 **METHODS AND MATERIALS**

12 **General Discrete Admixture Model**

13 In our previous study¹⁵, we modeled admixture history generation by generation and
14 proposed a general admixture model. The model was determined by a $K \times T$
15 admixture proportion matrix $M = \{m_i(t)\}_{1 \leq i \leq K, 1 \leq t \leq T}$, where K is the number of
16 ancestral populations, T is the time the admixed population arose, and $m_i(t)$ is the
17 ancestry contribution of i th ancestral population at time t . If the admixed
18 population did not receive any gene flows of i th ancestral population at time t , we
19 set $m_i(t)$ as 0. This general model covers all scenarios of an admixed population
20 with an arbitrary number of ancestral populations and arbitrary number of admixture
21 events. However, the parameters for this general model are redundant, and will lead to
22 over-fitting for most cases. For example, if we consider an HI model of 2 ancestral

1 populations and the admixed population arose T generations ago, the number of
 2 parameters is $2T$. However, $2(T - 1)$ parameters should be equal to 0, since ancestral
 3 populations contribute nothing after the first admixture. Thus, in fact, only 2
 4 parameters must be estimated. Thus, to reduce redundancy and maintain the
 5 universality of our model, we proposed a general discrete admixture model that only
 6 records the information of actual admixture events (see Fig. 1).

7 We considered an admixed population with K ancestral populations and n -wave
 8 discrete admixtures. Here, the time of the admixture in generations increase over time,
 9 with T being the present time. For the first wave admixture ($i = 1$), there are 2
 10 ancestral populations. We denote one ancestral population as population k_0 and the
 11 other as population k_1 . When $i \geq 2$, we denote k_i as the ancestral population of
 12 i th admixture. Then, we denote a vector $O = (k_0, k_1, \dots, k_n)$ as the admixture order
 13 of ancestral populations. Let α_i be the admixture proportion of the i th admixture
 14 and t_i be the admixture time of the i th admixture. We note that $0 \leq \alpha_i \leq$
 15 1 for $1 \leq i \leq n$, and $t_1 \leq t_2 \leq \dots \leq t_n \leq t_{n+1} = T$. For convenience in our later
 16 description, we denote the admixture event from population k_0 as the 0th
 17 admixture, which means the ancestral population k_0 is regarded as an admixed
 18 population before the first wave admixture. Thus, we set the corresponding admixture
 19 proportion $\alpha_0 = 1$ and admixture time $t_0 = t_1$. With this definition, each wave (i th
 20 wave) of admixture can be determined by 3 parameters, k_i , α_i , and t_i .

21 Now, denote $I_k = \{i: k_i = k\}$, then, $I_k(j)$ represents the wave ordinal of the j th
 22 admixture from ancestral population k . Let n_k denote the number of admixture

1 waves from ancestral population k , and thus we have $n_k = |I_k|$ and $\sum_{k=1}^K n_k = n +$
2 1, where n is total number of admixture waves. The general discrete admixture
3 model is determined by the admixture order $O = (k_0, k_1, \dots, k_n)$, the admixture
4 proportion $\{\alpha_i\}_{0 \leq i \leq n}$, and the admixture time $\{t_i\}_{0 \leq i \leq n+1}$. If we set

$$5 \quad m_{k_i}(t_i) = \begin{cases} 1 - \alpha_1, & i = 0 \\ \alpha_i, & i \geq 1 \\ 0, & otherwise \end{cases},$$

6 we can get the admixture proportion matrix of the previous general model¹⁵. This
7 shows that our new model is similar to the previous model. Furthermore, this new
8 model can also cover all scenarios of an admixed population with an arbitrary number
9 of ancestral populations and arbitrary number of admixture events.

10 Length Distribution of Ancestral Tracks

11 Next, we deduced the length distribution of ancestral tracks from ancestral population
12 k . The deduction process is similar to that of our previous work¹⁵. The wave ordinals
13 of admixture from ancestral population k are $I_k(1), I_k(2), \dots, I_k(n_k)$, respectively.
14 Denote $H_k(t)$ as the total ancestry proportion of the k th ancestral population in the
15 admixed population at t generation, and then we have

$$H_k(t) = \sum_{j=1}^w \alpha_{I_k(j)} \prod_{i=I_k(j)+1}^h (1 - \alpha_i), \quad (1)$$

16 where $w = \max\{j: t_{I_k(j)} \leq t, 1 \leq j \leq n_k\}$ and $h = \max\{i: t_i \leq t, 1 \leq i \leq n\}$.
17 Define s_i as the survival proportion of the ancestral tracks from the i th admixture
18 at generation T . Then,

$$s_i = \alpha_i \prod_{k=i+1}^n (1 - \alpha_k). \quad (2)$$

19 For simplicity, we assumed that chromosome length was infinite and there was no

1 genetic drift. Additionally, we defined the recombination among tracks from different
 2 ancestral populations as effective recombination because we only observed these
 3 recombination events among different ancestries. The length of ancestral tracks was
 4 changed by these recombination events. For the tracks from ancestral population k ,
 5 the effective recombination rate is $1 - H_k(t)$ at t generation. Let $u_k(j)$ be the
 6 total effective recombination rate for ancestral tracks from the j th admixture of
 7 ancestral population k . Then, we have

$$u_k(j) = \sum_{h=I_k(j)}^n (1 - H_k(t_h)) (t_{h+1} - t_h). \quad (3)$$

8 The length distribution of ancestral tracks from the j th admixture of ancestral
 9 population k is an exponential distribution with a rate of $u_k(j)$ ^{10,13,15}. A
 10 chromosome from the j th admixture of ancestral population k is expected to be
 11 split into $u_k(j)$ pieces per unit length (unit in Morgan). Thus, for the admixed
 12 population at T generation, the number of ancestral tracks from the j th admixture
 13 of ancestral population k is proportional to $s_{I_k(j)} u_k(j)$. Let X_k be the length of
 14 ancestral tracks from ancestral population k at generation T , and $f_k(x)$ is the
 15 probability density of X_k . Then,

$$\begin{aligned} f_k(x) &= \sum_{j=1}^{n_k} P \left(\begin{array}{c} \text{ancestral tracks are from the } j\text{th admixture} \\ \text{of ancestral population } k \end{array} \right) u_k(j) \exp(-u_k(j)x) \\ &= \sum_{j=1}^{n_k} \frac{s_{I_k(j)} u_k(j)}{\sum_{j=1}^{n_k} s_{I_k(j)} u_k(j)} u_k(j) \exp(-u_k(j)x). \end{aligned} \quad (4)$$

16 The length distribution of ancestral tracks was a mixed exponential distribution, and
 17 consisted with the results from our previous study ¹⁵.

18 Model Selection and Parameter Estimation

1 If the admixture model is determined, the length distribution of ancestral tracks can be
 2 written as [Formula \(4\)](#), which is a mixed exponential distribution. The
 3 EM-algorithm²² can be used to estimate the parameters in this distribution. However,
 4 the admixture model is often unclear in real situations, which means the number of
 5 admixture waves (n_k) and the order of admixtures (O) are unknown. Thus, we must
 6 first determine n_k and O . Here, we used LRT²⁰ to select the optimal n_k . After that,
 7 we used the exhaustion method to validate the accuracy of O . For any order of
 8 admixtures, we estimated the admixture proportion $\{\alpha_i\}_{0 \leq i \leq n}$ and admixture time
 9 $\{t_i\}_{0 \leq i \leq n+1}$ using the EM-algorithm. However, these parameter estimations must
 10 satisfy the following constraint conditions:

11 (a) $0 \leq \alpha_i \leq 1$, for $1 \leq i \leq n$;

12 (b) $t_1 \leq t_2 \leq \dots \leq t_n \leq t_{n+1}$.

13 If the estimations don't satisfy these conditions, the order is incorrect. After traversing
 14 all admixture orders, we could determine the correct ones.

15 The detailed procedures are as follows:

16 Step 1: Estimate the total admixture proportion m_k of ancestral population k .

17 With the inferred ancestral tracks, divide the total length of tracks from population k
 18 by the total length of tracks in the admixed population.

19 Step 2: Determine the number of admixture waves (n_k) for each ancestral
 20 population and estimate the parameters of the mixed exponential distribution. For
 21 ancestral population k , use LRT to select the optimal number of admixture waves and
 22 then estimate the parameters $\{(\omega_{k1}, \lambda_{k1}), (\omega_{k2}, \lambda_{k2}), \dots, (\omega_{kn_k}, \lambda_{kn_k})\}$ of the mixed

- 1 exponential distribution using the EM-algorithm, where $\omega_{kj} = \frac{s_{I_k(j)}u_k(j)}{\sum_{j=1}^{n_k}s_{I_k(j)}u_k(j)}$,
- 2 $\lambda_{kj}=u_k(j)$. Details of the EM-algorithm and LRT procedures are in Supplementary
- 3 Information ([Supplementary Text S1](#)).

4 Step 3: Select an admixture order O without replacement from set

$$5 \quad \Omega = \left(O : \text{a permutation of sequence } \left(\underbrace{1, \dots, 1}_{n_1}, \dots, \underbrace{k, \dots, k}_{n_k}, \dots, \underbrace{K, \dots, K}_{n_K} \right), \text{ where } O(1) \neq O(2) \right).$$

6 Get I_k for each k base on the selected O .

- 7 Step 4: Determine $\{s_i\}_{0 \leq i \leq n}$ and $\{u_k(j), 1 \leq j \leq n_k, 1 \leq k \leq K\}$ from the
- 8 following equations:

$$\begin{cases} u_k(j) = \lambda_{kj}, \\ \frac{s_{I_k(j)}u_k(j)}{\sum_{j=1}^{n_k}s_{I_k(j)}u_k(j)} = \omega_{kj}, \\ \sum_{j=1}^{n_k}s_{I_k(j)} = m_k, \end{cases}$$

9 where $1 \leq j \leq n_k, 1 \leq k \leq K$.

10 Step 5: Determine $\{\alpha_i\}_{0 \leq i \leq n}$ from the following equations:

$$s_i = \alpha_i \prod_{k=i+1}^n (1 - \alpha_k),$$

11 where $0 \leq i \leq n$.

12 Step 6: Determine $\{t_i\}_{0 \leq i \leq n+1}$ from the following equations:

$$\begin{aligned} H_k(t) &= \sum_{j=1}^w \alpha_{I_k(j)} \prod_{i=I_k(j)+1}^h (1 - \alpha_i), \\ u_k(j) &= \sum_{i=I_k(j)}^n (1 - H_k(t_i)) (t_{i+1} - t_i), \end{aligned}$$

13 where $1 \leq j \leq n_k, 1 \leq k \leq K$.

1 Step 7: Judge whether $\{\alpha_i\}_{0 \leq i \leq n}$ and $\{t_i\}_{0 \leq i \leq n+1}$ satisfy the following conditions:

2 (a) $0 \leq \alpha_i \leq 1$, for $1 \leq i \leq n$;

3 (b) $t_1 \leq t_2 \leq \dots \leq t_n \leq t_{n+1}$.

4 If these conditions are satisfied, record the corresponding admixture order O ,
 5 admixture proportion $\{\alpha_i\}_{0 \leq i \leq n}$, and admixture time $\{t_i\}_{0 \leq i \leq n+1}$. Then return to Step
 6 3 until all possible admixture orders are checked.

7 Through these above procedures, we obtained all reasonable admixture orders O ,
 8 the corresponding estimators of admixture proportion $\{\alpha_i\}_{0 \leq i \leq n}$, and admixture time
 9 $\{t_i\}_{0 \leq i \leq n+1}$. Based on the estimations of these parameters, we could recover the
 10 history of the admixed population.

11 However, due to a lack of accuracy in local ancestry inference, only these relatively
 12 long tracks are reliable^{10,13}. Therefore, we are interested in the conditional length
 13 distribution of ancestral tracks longer than a specific threshold C . As we know, the
 14 length distribution of ancestral tracks from each ancestral population is a mixed
 15 exponential distribution. When we consider only tracks larger than C , the length
 16 distribution from ancestral population k becomes

$$f_k(x|x \geq C) = \sum_{j=1}^{n_k} \frac{\omega_{kj}}{\sum_{j=1}^{n_k} \omega_{kj} \exp(-u_k(j)C)} u_k(j) \exp(-u_k(j)x),$$

17 where $\omega_{kj} = \frac{s_{I_k(j)} u_k(j)}{\sum_{j=1}^{n_k} s_{I_k(j)} u_k(j)}$. However, since this condition distribution is not a mixed
 18 exponential distribution, we cannot use the EM-algorithm to estimate the parameters.
 19 Fortunately, when we consider the random variable $Y_k = X_k - C$, we find that the
 20 distribution of Y_k is a mixed exponential distribution, which can be written as
 21 follows:

$$f_k(y) = \sum_{j=1}^{n_k} \frac{\omega_{kj} \exp(-u_k(j)C)}{\sum_{j=1}^{n_k} \omega_{kj} \exp(-u_k(j)C)} u_k(j) \exp(-u_k(j)y).$$

1 To take the threshold C into consideration, we must change the procedures of the
2 aforementioned Step 2. We can easily obtain samples of Y_k from samples of X_k .
3 Then, we can use the EM-algorithm and LRT to obtain the distribution parameters of
4 Y_k . Furthermore, by the relationship between $f_k(x)$ and $f_k(y)$, we can obtain the
5 parameters of the mixed exponential distribution of X_k . Then, the following Steps are
6 the same as those aforementioned in Steps 3-7. These procedures were all
7 implemented in our *MultiWaver*.

8 In the software of *MultiWaver*, 2 estimations of admixture time for the first wave
9 were output. One was an estimation of t_0 , while the other was an estimation of t_1 . In
10 theory, t_1 is equal to t_0 , but in real data analysis, the estimations may be not equal
11 because of random errors and tracks inference errors. Thus, we presented 2
12 estimations of admixture time for the first admixture wave in our results.

13 **SIMULATION**

14 **Performance Evaluation of *MultiWaver***

15 We conducted simulations to evaluate the performance of *MultiWaver*. The simulation
16 data were generated by forward-time simulator *AdmixSim*²⁴. General settings of our
17 simulation were the same as those in our previous study¹⁵.

18 Here, we divided multiple-wave admixture models into 2 different types of models.
19 We denoted the model as a simple model if each ancestral population could contribute
20 only once to the admixed population. The others were denoted as a complex model. In
21 the complex model, at least 1 ancestral population donates more than once admixture.

1 It is important to note that when we infer the admixture history under the complex
2 model, it is very challenging to distinguish the different admixture waves from the
3 same ancestral population.

4 We focused on evaluating the performance of *MultiWaver* under these 2 types of
5 models. For the simple model, we considered a scenario of 3 ancestral populations
6 (Fig. S1 Scenario (I)), and a scenario of 5 ancestral populations (Fig. S1 Scenario (II)).
7 For the complex model, we considered a scenario of 2 ancestral populations with
8 2-wave admixtures (Fig. S1 Scenario (III)). We evaluated the performance of
9 *MultiWaver* with different admixture times and admixture proportions. For simplicity,
10 we supposed the admixture proportions ($\alpha_i, 1 \leq i \leq n$) were equal. We set 3 different
11 values of admixture proportion, 0.1, 0.3, and 0.5, for each scenario. For Scenario (I),
12 the admixture time was set as 2 different cases: (a) $t_2 = 20, T = 40$, and (b)
13 $t_2 = 40, T = 60$. For Scenario (II), the admixture time was also set as 2 different
14 cases: (a) $t_2 = 20, t_3 = 40, t_4 = 60, T = 80$, and (b) $t_2 = 40, t_3 = 80, t_4 =$
15 $120, T = 140$. However, for Scenario (III), the admixture time was set as 4 different
16 cases: (a) $t_2 = 20, T = 40$, (b) $t_2 = 40, T = 60$, (c) $t_2 = 60, T = 80$, and (d)
17 $t_2 = 80, T = 100$. Each case was repeated 10 times for a total of 240 simulations
18 across these 3 scenarios. *MultiWaver* was applied to the simulated data with the
19 default settings; the results were recorded and summarized.

20 In real situations, due to the limitations of local ancestry inference, only the
21 ancestral tracks longer than a special threshold can be accurately inferred. Thus, to
22 make our method more available to real situations, we chose the thresholds ranging

1 from 0 cM to 2 cM in steps of 0.25 cM, and then evaluated the robustness of our
2 method under different thresholds.

3 **Application to Real Datasets**

4 Firstly, we applied our method to some real datasets of African Americans and
5 Mexicans. These 2 populations are typical admixed populations and their histories are
6 relatively clear. Therefore, they could be used to test the performance of our method
7 for real data. We obtained the datasets of African Americans (ASW), Mexicans
8 (MEX), and reference populations African (YRI) and European (CEU) from the
9 HapMap Project Phase III dataset²³. Meanwhile, Maya and Pima populations
10 represented American Indian ancestry, which were obtained from the HGDP dataset¹.
11 According to prior knowledge, African Americans and Mexicans have more than 2
12 ancestries^{14,25}. However, the proportion of Native American ancestry of African
13 Americans is less than 5%²⁶, and thus, we only considered 2 dominant ancestries
14 (African and European ancestry) of African Americans. For Mexicans, we considered
15 3 ancestries: African, European, and American Indian ancestry²⁷.

16 Then, our method was used to reconstruct the population history of Uyghurs and
17 Hazaras. The histories of these 2 populations are more complex. Uyghurs and Hazaras
18 populations were obtained from the HGDP dataset. Previous studies have shown that
19 Uyghurs and Hazaras had admixed ancestries mainly from Europe and East Asia^{1,5}.
20 Here, we used Han and French as the proxies of Asian ancestry and European ancestry,
21 respectively⁸. These reference populations were also obtained from the HGDP dataset.

1 To enhance the reliability of our analysis, HAPMIX⁶ was selected as the local
2 ancestry inference method since it shows good performance in admixture break points
3 inference²⁸. However, HAPMIX can only be used to detect ancestral tracks for
4 two-way admixtures, and thus it might not be proper for the Mexican population.
5 PCAdmix²⁹ has shown great power in inferring the local ancestry of Mexican
6 populations²⁷, and thus we used PCAdmix in this study. The generations pre-set in
7 HAPMIX inference were 10 for African Americans and 80 for Uyghurs and Hazaras.
8 The window size set in PCAdmix was default. Since phasing data was required for
9 both HAPMIX and PCAdmix, SHAPEIT 2³⁰ was used to infer the haplotype phase.
10 Finally, *MultiWaver* was used to determine the optimal model and estimate the
11 admixture time accordingly with tracks longer than 1 cM.

12 RESULTS

13 *MultiWaver* Performed Well under Simple and Complex Models

14 We compared the admixture histories inferred by *MultiWaver* with the histories set in
15 simulations, and then evaluated the performance of our method in model selection and
16 parameters estimation. For Scenario (I) and (II), results showed that estimations of
17 admixture time were high consistency with the time simulated if we pre-set the
18 admixture model as the simple model (-s option in *MultiWaver*) (see Fig. 2). Our
19 method also performed well when we did not pre-set this option (-s) (see Fig. S2).
20 Only a few models in our simulations were wrongly selected. When the model was
21 correctly selected, the admixture time estimated was consistent with the simulated
22 time.

1 For the complex model, we found that our method could select the right model with
2 high accuracy (see Fig. 3). Model selection was incorrect for only 3 simulations. In
3 these 3 cases, the numbers of admixture waves were wrongly estimated, which led to
4 inaccurate estimation of admixture time. Thus, selecting a correct model is of crucial
5 importance for admixture history inference. When the admixture model could be
6 correctly selected, only a slight overestimation occurred for the admixture time.

7 We also evaluated the performance of *MultiWaver* with different admixture
8 proportions. We found that the overestimation of admixture time in the complex
9 model was related to the admixture proportions (see Fig. S3). When the proportions of
10 each admixture wave became smaller, estimation error decreased. However, for the
11 simple model, our method performed well for all situations.

12 In conclusion, regardless of which type of admixture model, our method performed
13 well for model selection. Furthermore, the admixture time was estimated well for the
14 simple model, and with only slight overestimation for the complex model.

15 **Robustness for Different Thresholds of Track Length.**

16 We tested the robustness of our method for different thresholds of track length.
17 Results showed that our method was robust to thresholds for both the simple model
18 and complex model (see Fig. 4). Due to the limitations of our method, the local
19 ancestry inference was not so accurate for short ancestral tracks. Thus, in real data
20 analysis, we had to discard tracks smaller than a threshold. However, short ancestral
21 tracks contain ancient admixture information, and if the threshold was too large, lots
22 of information would be lost. Therefore, we had to balance the trade-off between

1 information and accuracy. In our real data analysis, we set the thresholds as 1cM.

2 **Real Data Analysis**

3 We applied our method to infer the admixture histories of some real datasets. For
 4 African Americans, HAPMIX was used to infer the ancestries with Africans (YRI)
 5 and European (CEU) as the 2 ancestral populations. The admixture model was
 6 inferred as 2 ancestral populations with a 2-wave admixtures model (see Fig. 5(a)).
 7 The African population (YRI) contributed 2 wave admixtures, and the admixture time
 8 was 11 generations ago and 7 generations ago, respectively. The time of the first
 9 admixture was about the 17th century, which was consistent with the time that most
 10 African ancestors arrived in America via slave trading. This inferred time was close to
 11 previous findings^{12,13,15,25,26}. After the slave trading, many African people settled down
 12 in America. The second admixture wave might have been caused by these people or
 13 by recent migrations from Africa to America. The admixture model inferred by our
 14 method pointed out that the admixture history of African Americans was not 1 pulse
 15 admixture, which was also reported in previous studies^{12,25,26}.

16 For Mexicans, we used PCAdmix to infer the local ancestries, and a 2-wave
 17 admixtures model was inferred (see Fig. 5(b)). Each ancestral population contributed
 18 once to the admixed population. The time of the first admixture wave was about 18
 19 generations ago, which was close to previous findings^{25,27,31-33}. The time of the second
 20 admixture was 12 generations ago. This time period (12~18 generation ago) was
 21 consistent with the time of the exploration of the new world. For our analysis of
 22 African Americans and Mexicans, the admixture histories inferred by our method

1 were consistent with recorded histories, thus showing the power of our method in real
2 data analysis.

3 Finally, we applied our method to reconstruct the admixture histories of Uyghurs
4 and Hazaras (see Fig. 5(c) and (d)). Results showed that these 2 populations shared a
5 similar admixture model, except the admixture time of Hazaras was more ancient. The
6 earliest admixture event of Uyghurs occurred about 144 generations ago, with
7 subsequent admixture waves from both ancestries 20-50 generations ago. While the
8 earliest admixture event of Hazaras occurred around 173 generations ago, with
9 following gene flows occurred 20~70 generations ago. Compared with the results
10 inferred by the admixture history inference method ALDER⁸, our method found an
11 additional ancient admixture event in Uyghurs and Hazaras. To explain the
12 discrepancies in theory, ALDER considers only the decay curve of weighted linkage
13 disequilibrium (LD) between pairs of sites whose genetic distance were larger than
14 0.5 cM⁸, and thus the information of ancient signals within shorter loci pairs would be
15 lost. Meanwhile, our method saved some of these ancient signals by deducing the
16 conditional length distribution of ancestral tracks even if we discarded ancestral tracks
17 shorter than 1 cM.

18 Conclusively, Uyghurs and Hazaras had a similar admixture history. The ancient
19 admixture might have been caused by the migrations of Indo-Aryan speaking people
20 into the Indian subcontinent (1500 BC). Uyghurs mainly settled in West China, and
21 Hazaras mainly settled in Afghanistan and Pakistan. The residences of these 2
22 populations were all near the Silk Road, and thus we thought the recent multiple

1 admixtures might have been caused by the trades or migrations along the Silk Road.
2 In fact, the real history of Uyghurs and Hazaras might be more complex than inferred.
3 However, our method could detect some effective admixtures and provide some
4 useful information to understand the origin and development of these complex
5 populations.

6 **DISCUSSION**

7 Complex admixture history inference has long been a challenging problem in
8 population genetics. In this work, we proposed a general discrete admixture model to
9 describe admixture history with multiple ancestral populations and multiple-wave
10 admixtures. We deduced that the length distribution of ancestral tracks was a mixed
11 exponential distribution. Based on this distribution, we developed a new method,
12 *MultiWaver*, to infer the multiple-wave admixture histories. We used LRT to select the
13 number of admixture waves, and implemented an exhaustion method to determine the
14 order of admixtures. When the admixture model was determined, we applied the
15 EM-algorithm to estimate parameters. Simulations and real data analysis showed that
16 *MultiWaver* was precise and efficient in inferring admixture history.

17 Comparing with previous methods, our method showed superiority in 2 aspects.
18 Firstly, our method could be used to infer multiple-wave admixture history, while
19 previous methods could only infer admixture history under some simple models.
20 Secondly, no prior admixture model was required in our method, while previous
21 methods needed to assume a special admixture model when trying to infer admixture
22 history. Therefore, the inferred history might be biased or even unreliable if the

1 provided model deviates from real history. However, our method avoided this
2 problem by selecting an optimal admixture model based on ancestral tracks.

3 Our method introduced an elegant solution to the complex admixture history
4 inference. However, some problems still exist. When inferring admixture history
5 under the complex model, overestimation occurred for the admixture time. In our
6 method, we assumed chromosome length was infinite and there was no genetic drift,
7 and then we found that the length distribution was a mixed exponential distribution.
8 However, Liang and Nielsen pointed out that the length distribution did not follow an
9 exponential distribution when the admixture time was too small or too large³⁴. In the
10 complex model, the non-exponential property would be accumulated, which might be
11 the reason behind the overestimation we observed with our method. We also found
12 that the overestimation was related to the admixture proportion of each admixture
13 wave. We performed simple linear regression analysis on the errors for admixture
14 time estimations and admixture proportion (Fig. S4).

15 In our method, it is possible that more than 1 optimal admixture model satisfied the
16 constraint conditions and should be recorded. However, for all simulations we
17 conducted, this phenomenon did not appear. This was reasonable because the
18 admixture history had a one-to-one correspondence with the length distribution of
19 ancestral tracks. If 1 situation had more than 1 optimal model, it implied the ancestral
20 tracks were not accurately inferred.

21 The efficiency of our method was also influenced by the validity of the local
22 ancestry inference. We tested the performance of our method with the inferred

1 ancestral tracks (see Supplementary Text 2). We found that *MultiWaver* tended to
 2 overestimate the number of waves, and thus led to overestimating the admixture time
 3 with the ancestral tracks inferred by HAPMIX (Fig. S5 (a)). For multiple-way
 4 admixtures, the inaccuracy of ancestral tracks inferred by PCAdmix led to
 5 underestimating the time of the first admixture wave (Fig. S5 (b)). It was very
 6 difficult to obtain relatively accurate ancestral tracks with a small length for all local
 7 ancestry inference methods. To improve the effectiveness of the inference, we suggest
 8 using the ancestral tracks longer than a certain threshold C in our method. However,
 9 when the threshold became large, ancient admixture information would be lost rapidly.
 10 With the development of sequencing technology and computational methods, short
 11 ancestral tracks could be precisely detected in the near future. Then, our method
 12 would be promising in recovering even more ancient admixture history, such as the
 13 admixture between modern humans and ancient humans^{35,36}.

14

15 ACKNOWLEDGEMENTS

16 This work was supported by National Natural Science Foundation of China (NSFC)
 17 grants (91331204 and 11426237), the Strategic Priority Research Program
 18 (XDB13040100) and Key Research Program of Frontier Sciences
 19 (QYZDJ-SSW-SYS009) of the Chinese Academy of Sciences (CAS), the National
 20 Science Fund for Distinguished Young Scholars (31525014), the Program of Shanghai
 21 Academic Research Leader (16XD1404700); 973 Project (2011CB808000), the
 22 Fundamental Research Funds for the Central Universities (2011JBZ019, 2014RC008
 23 and 2015IBM099), the National Excellent Doctoral Dissertation Foundation of PR

1 China (201213), National Center for Mathematics and Interdisciplinary Sciences of
 2 CAS, and the Key Laboratory of Random Complex Structures and Data Science, CAS
 3 (2008DP173182). S.X. also gratefully acknowledges the support of the National
 4 Program for Top-notch Young Innovative Talents of The "*Wanren Jihua*" Project. We
 5 thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of
 6 this manuscript.

7

8

1 **Figure Legends**

2 **Figure 1. The general discrete admixture model.** Here, we illustrated an admixed
3 population with K ancestral populations and n -wave discrete admixtures, which
4 started to admix T generations ago. POP_{ki} is the ancestral population of the i th
5 admixture, α_i is the admixture proportion of the i th admixture, and t_i is the
6 admixture time of the i th admixture.

7 **Figure 2. Admixture time estimated under simple model.** Admixture time
8 estimated under Scenario (I) for (a) $\alpha_1 = \alpha_2 = 0.3$ and $t_2 = 20, T = 40$; (b)
9 $\alpha_1 = \alpha_2 = 0.3$ and $t_2 = 40, T = 60$. Admixture time estimated under Scenario (II)
10 for (c) $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.3$ and $t_2 = 20, t_3 = 40, t_4 = 60, T = 80$; (d)
11 $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.3$ and $t_2 = 40, t_3 = 80, t_4 = 120, T = 140$. X-coordinate
12 is the admixture time in generations ago, with 0 being the present time. Each case was
13 repeated for 10 times, and $Y = i$ means the i th simulation. The points in the line
14 ($Y = i$) represent the admixture time estimated from the i th simulation, and the
15 color of the points indicates the ancestral population. The dashed lines represent the
16 simulated admixture time.

17 **Figure 3. Admixture time estimated under complex model.** Admixture time
18 estimated under Scenario (III) for (a) $\alpha_1 = \alpha_2 = 0.3$ and $t_2 = 20, T = 40$; (b)
19 $\alpha_1 = \alpha_2 = 0.3$ and $t_2 = 40, T = 60$; (c) $\alpha_1 = \alpha_2 = 0.3$ and $t_2 = 60, T = 80$; and
20 (d) $\alpha_1 = \alpha_2 = 0.3$ and $t_2 = 80, T = 100$. X-coordinate is the admixture time in
21 generations ago, with 0 being the present time. Each case was repeated for 10 times,
22 and $Y = i$ means the i th simulation. The points in the line ($Y = i$) represent the

1 admixture time estimated from the i th simulation, and the color of the points
2 indicates the ancestral population. The dashed lines represent the simulated admixture
3 time.

4 **Figure 4. Admixture time estimated with different thresholds.** (a) Admixture time
5 estimated under Scenario (I), where $\alpha_1 = \alpha_2 = 0.3$ and $t_2 = 40, T = 60$; (b)
6 Admixture time estimated under Scenario (III), where $\alpha_1 = \alpha_2 = 0.3$ and
7 $t_2 = 40, T = 60$. X-coordinate is the admixture time in generations ago, with 0 being
8 the present time. Y-coordinate represents the thresholds, and the color of the points
9 indicates the ancestral population. The dashed lines represent the simulated admixture
10 time.

11 **Figure 5. Inferred admixture history of real datasets.** Inferred admixture history of
12 (a) African Americans, (b) Mexicans, (c) Uyghurs, And (d) Hazaras. The time of the
13 first admixture wave was the average of estimations for time t_0 and t_1 . AMI:
14 combined dataset of populations Maya and Pima which represent American Indian
15 ancestry; Han: Han population, represent Asian ancestry; Fre: French population,
16 represent European ancestry.

17

18

REFERENCE

- 1 Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104, doi:10.1126/science.1153717 (2008).
- 2 Wall, J. D., Lohmueller, K. E. & Plagnol, V. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular biology and evolution* **26**, 1823-1827, doi:10.1093/molbev/msp096 (2009).
- 3 Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489-U450, doi:10.1038/nature08365 (2009).
- 4 Xu, S., Huang, W., Qian, J. & Jin, L. Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *American journal of human genetics* **82**, 883-894, doi:10.1016/j.ajhg.2008.01.017 (2008).
- 5 Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747-751, doi:10.1126/science.1243518 (2014).
- 6 Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics* **5**, e1000519, doi:10.1371/journal.pgen.1000519 (2009).
- 7 Moorjani, P. *et al.* The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS genetics* **7**, e1001373, doi:10.1371/journal.pgen.1001373 (2011).
- 8 Loh, P. R. *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233-1254, doi:10.1534/genetics.112.147330 (2013).
- 9 Pickrell, J. K. *et al.* Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 2632-2637, doi:10.1073/pnas.1313787111 (2014).
- 10 Pool, J. E. & Nielsen, R. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711-719, doi:10.1534/genetics.108.098095 (2009).
- 11 Pugach, I., Matveyev, R., Wollstein, A., Kayser, M. & Stoneking, M. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome biology* **12**, R19, doi:10.1186/gb-2011-12-2-r19 (2011).
- 12 Jin, W., Wang, S., Wang, H., Jin, L. & Xu, S. Exploring population admixture dynamics via empirical and simulated genome-wide distribution of ancestral chromosomal segments. *American journal of human genetics* **91**, 849-862, doi:10.1016/j.ajhg.2012.09.008 (2012).
- 13 Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607-619, doi:10.1534/genetics.112.139808 (2012).
- 14 Jin, W., Li, R., Zhou, Y. & Xu, S. Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping. *European journal of human genetics : EJHG* **22**, 930-937, doi:10.1038/ejhg.2013.265 (2014).
- 15 Ni, X. *et al.* Length Distribution of Ancestral Tracks under a General Admixture Model and Its Applications in Population History Inference. *Scientific reports* **6**, 20048, doi:10.1038/srep20048 (2016).
- 16 Pugach, I. *et al.* The Complex Admixture History and Recent Southern Origins of Siberian Populations. *Molecular biology and evolution* **33**, 1777-1795, doi:10.1093/molbev/msw055 (2016).
- 17 Xu, S. & Jin, L. A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *American journal of human genetics* **83**, 322-336,

1 doi:10.1016/j.ajhg.2008.08.001 (2008).

2 18 Lipson, M. *et al.* Reconstructing Austronesian population history in Island Southeast Asia.

3 *Nature communications* **5**, 4689, doi:10.1038/ncomms5689 (2014).

4 19 Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry of

5 African Americans, Latinos, and European Americans across the United States. *American*

6 *journal of human genetics* **96**, 37-53, doi:10.1016/j.ajhg.2014.11.010 (2015).

7 20 Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite

8 hypotheses. *The annals of mathematical statistics* **9**, 60-62 (1938).

9 21 Likelihood-ratio test, <https://en.wikipedia.org/wiki/Likelihood-ratio_test> (

10 22 Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the

11 EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38 (1977).

12 23 International HapMap, C. *et al.* Integrating common and rare genetic variation in diverse

13 human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).

14 24 Yang, X. *et al.* AdmixSim: a forward-time simulator for various and complex scenarios of

15 population admixture. *bioRxiv*, 037135 (2016).

16 25 Kidd, J. M. *et al.* Population genetic inference from personal genome data: impact of ancestry

17 and admixture on human genomic variation. *American journal of human genetics* **91**,

18 660-671, doi:10.1016/j.ajhg.2012.08.025 (2012).

19 26 Baharian, S. *et al.* The Great Migration and African-American Genomic Diversity. *PLoS*

20 *genetics* **12**, e1006059, doi:10.1371/journal.pgen.1006059 (2016).

21 27 Moreno-Estrada, A. *et al.* Reconstructing the population genetic history of the Caribbean.

22 *PLoS genetics* **9**, e1003925, doi:10.1371/journal.pgen.1003925 (2013).

23 28 Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170-175,

24 doi:10.1038/nature10336 (2011).

25 29 Brisbin, A. *et al.* PCAdmix: principal components-based assignment of ancestry along each

26 chromosome in individuals with admixed ancestry from two or more populations. *Human*

27 *biology* **84**, 343-364 (2012).

28 30 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of

29 genomes. *Nature methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2012).

30 31 Tian, C. *et al.* A genomewide single-nucleotide-polymorphism panel for Mexican American

31 admixture mapping. *American journal of human genetics* **80**, 1014-1023 (2007).

32 32 Wang, S. *et al.* Geographic patterns of genome admixture in Latin American Mestizos. *PLoS*

33 *genetics* **4**, e1000037, doi:10.1371/journal.pgen.1000037 (2008).

34 33 Price, A. L. *et al.* A genomewide admixture map for Latino populations. *American journal of*

35 *human genetics* **80**, 1024-1036, doi:10.1086/518313 (2007).

36 34 Liang, M. & Nielsen, R. The lengths of admixture tracts. *Genetics* **197**, 953-967,

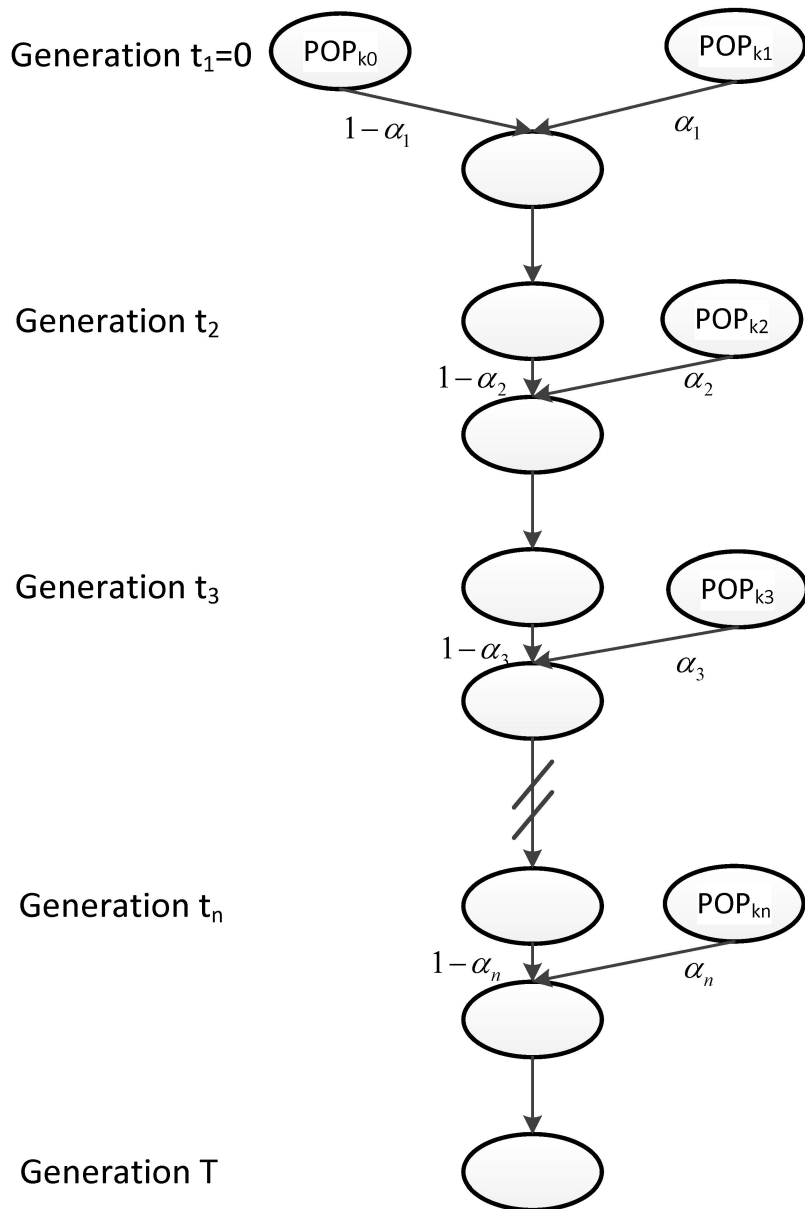
37 doi:10.1534/genetics.114.162362 (2014).

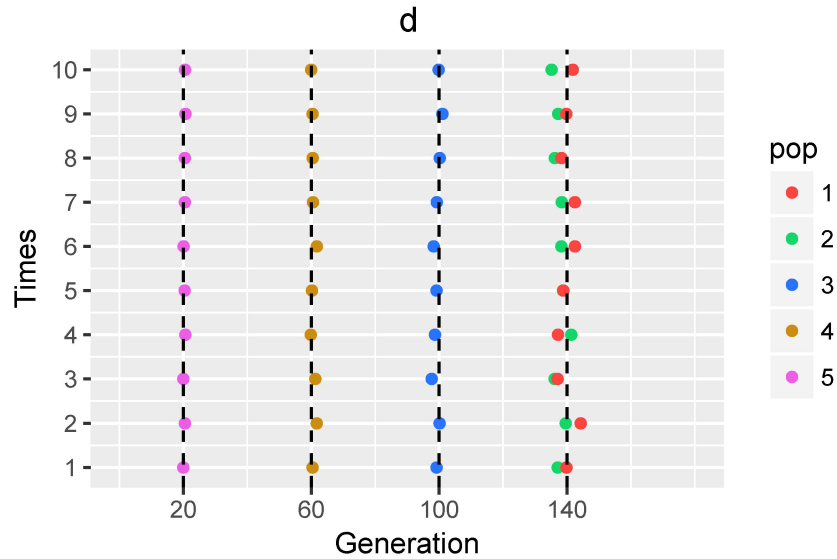
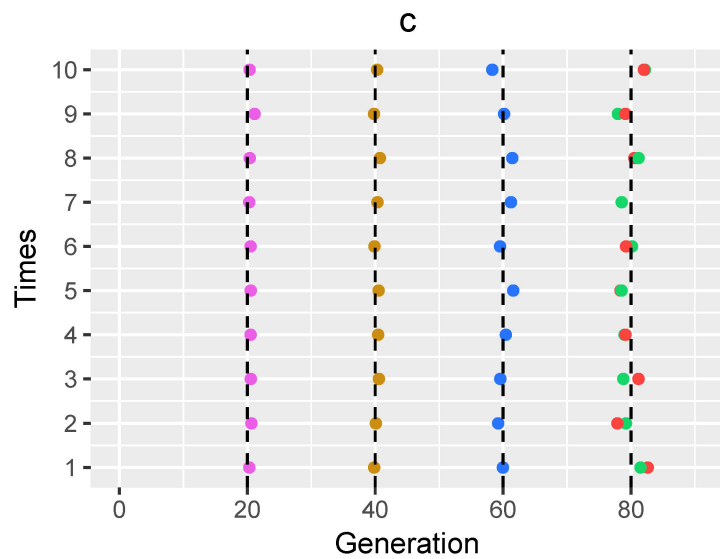
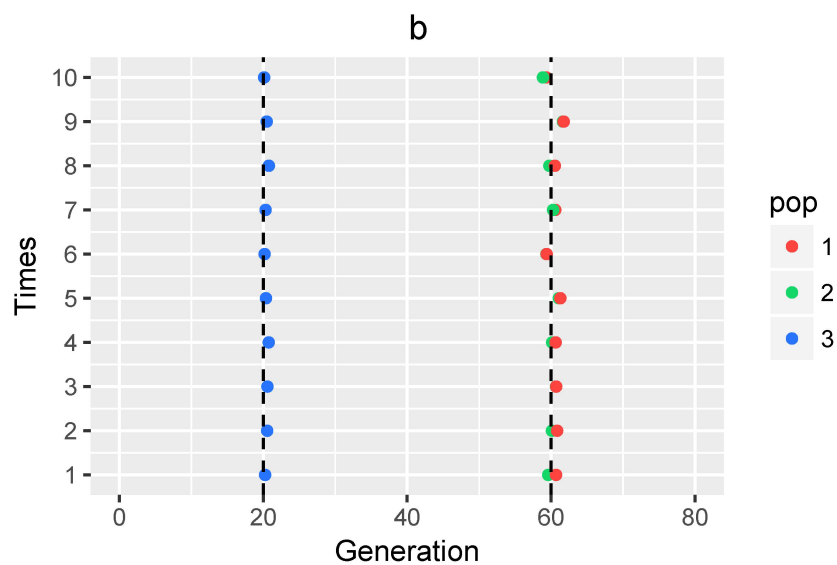
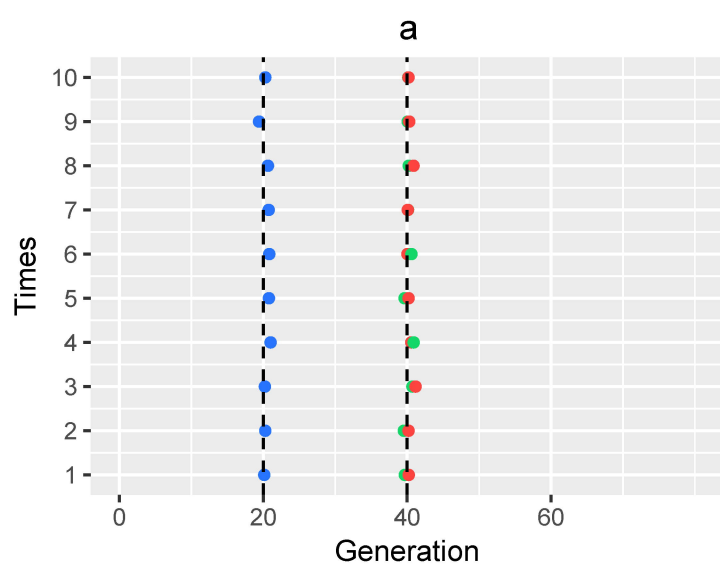
38 35 Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day

39 humans. *Nature* **507**, 354-357, doi:10.1038/nature12961 (2014).

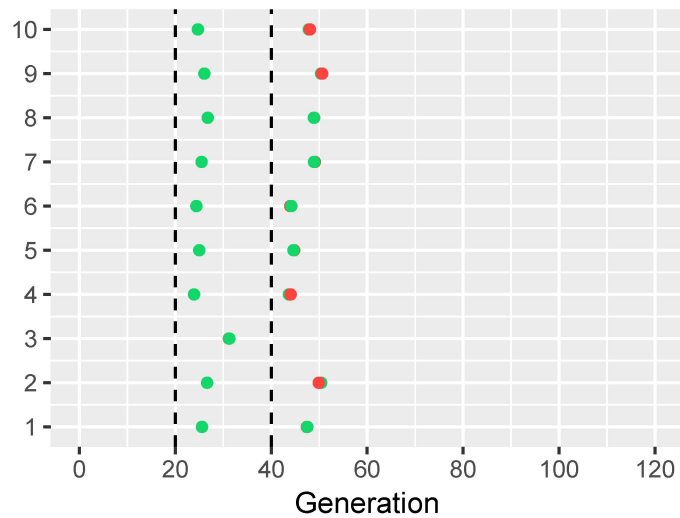
40 36 Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains.

41 *Nature* **505**, 43-49 (2014).

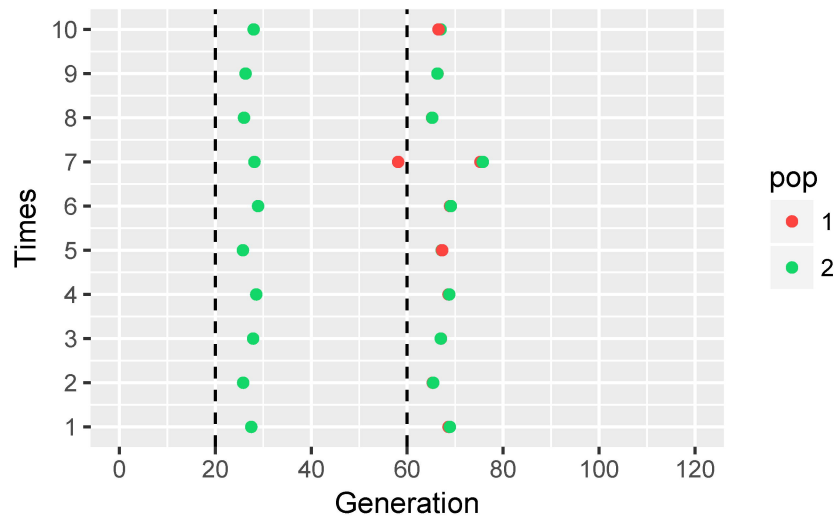




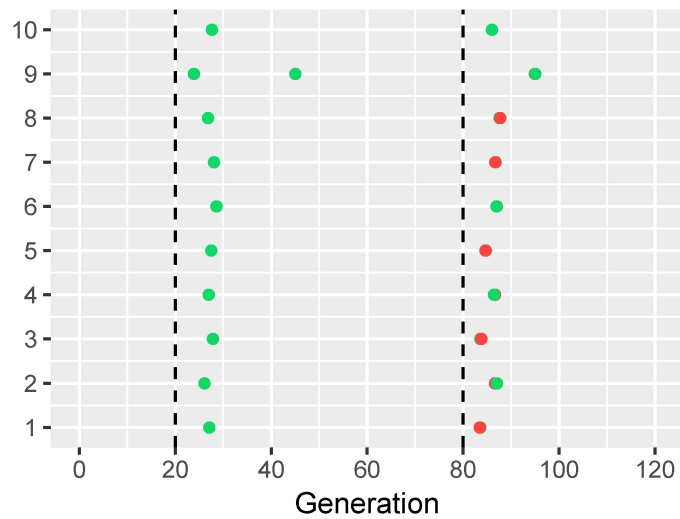
a



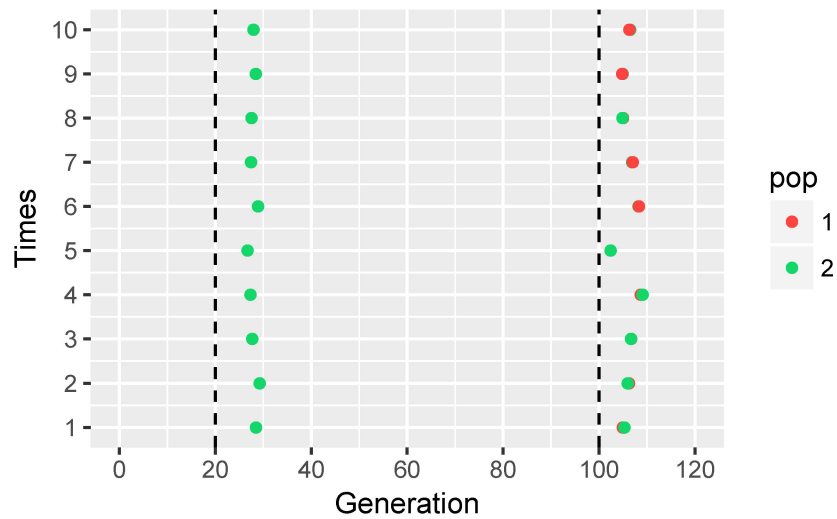
b



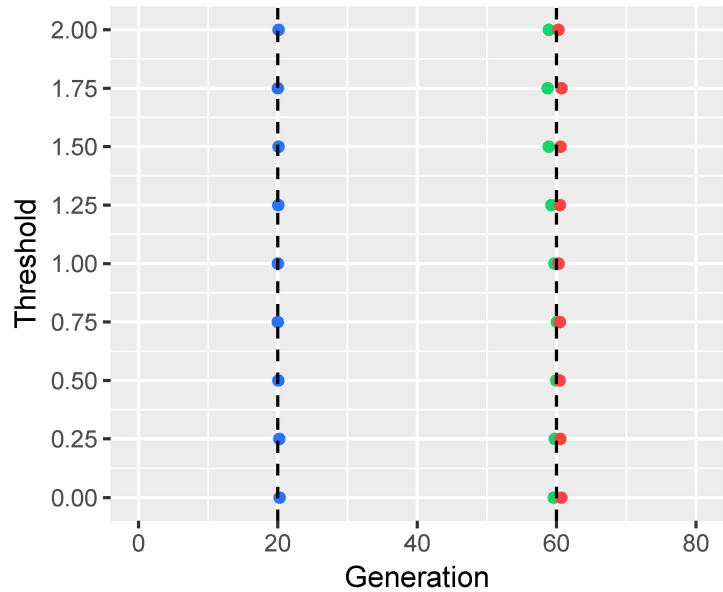
c



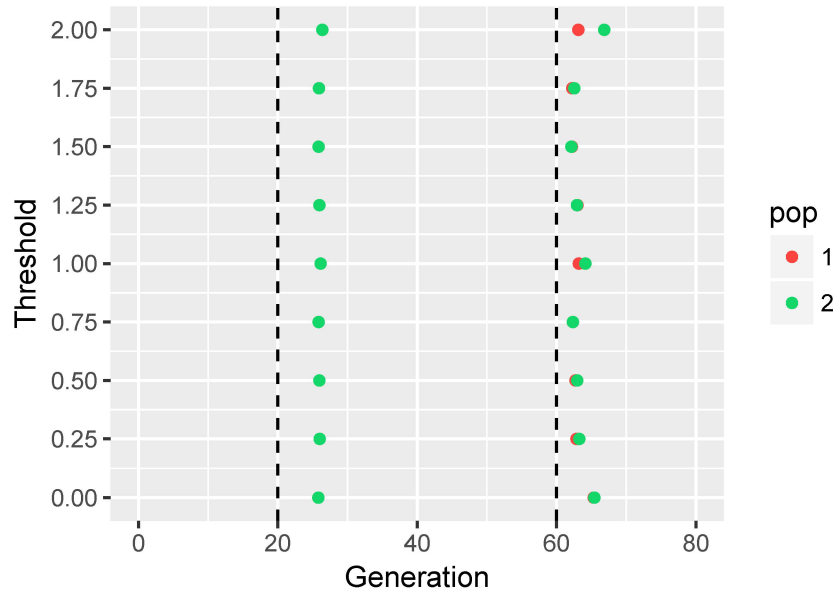
d



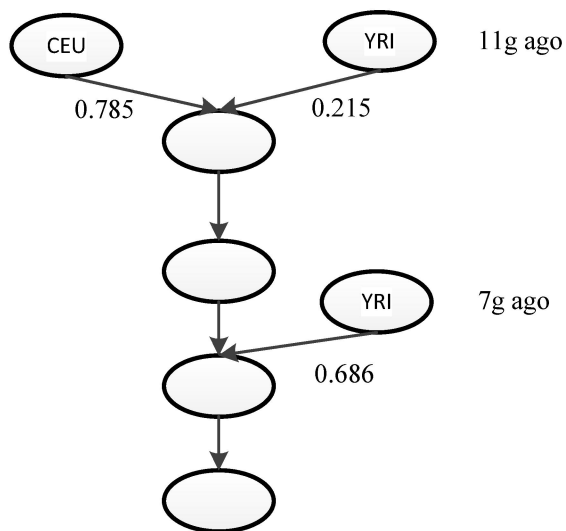
a



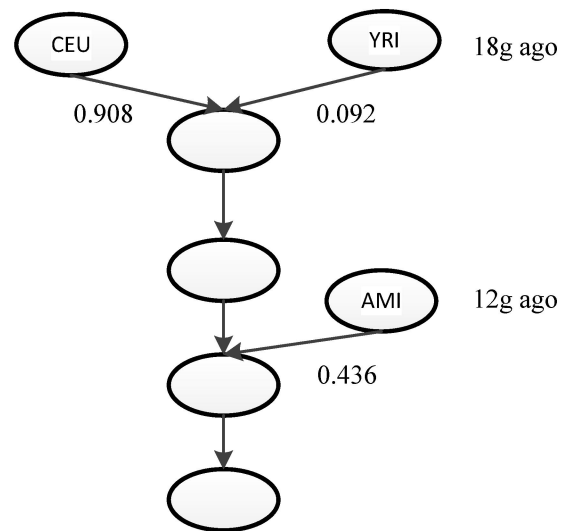
b



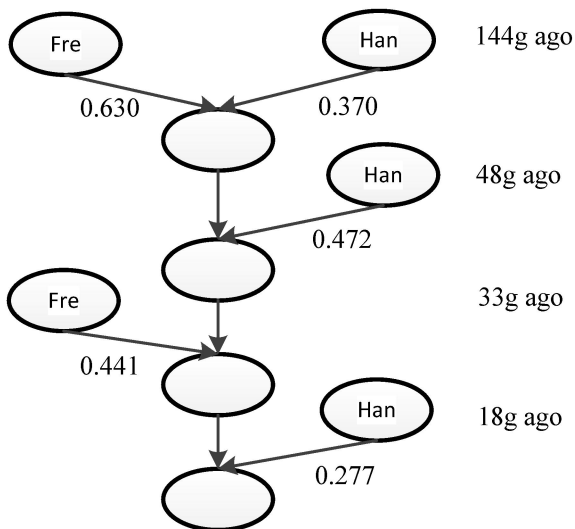
a



b



c



d

