

Inference of multiple-wave population admixture by modeling decay of linkage disequilibrium with polynomial functions

Running title: Inference of multiple-wave population admixture

Ying Zhou^{*,**,\$}, Kai Yuan^{*,**,\$}, Yaoliang Yu[†], Xumin Ni[‡], Pengtao Xie[†], Eric P. Xing[†], Shuhua Xu^{*,**,††,‡‡,\$\$}

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

^{**}University of Chinese Academy of Sciences, Beijing 100049, China;

[†]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, 15213, USA;

[‡]Department of Mathematics, School of Science, Beijing Jiaotong University, Beijing 100044, China;

^{††}School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China;

^{‡‡}Collaborative Innovation Center of Genetics and Development, Shanghai 200438, China.

^{\$}These authors contributed equally to this work.

^{\$\$} Correspondence and requests for materials should be addressed to xushua@picb.ac.cn (S.X.)

The authors declare no competing financial interests.

Shuhua Xu: 320 Yueyang Road, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China. Tel: +86-21-5492-0479. E-mail: xushua@picb.ac.cn

Abstract

To infer the histories of population admixture, one important challenge with methods based on the admixture linkage disequilibrium (ALD) is to get rid of the effect of source LD (SLD) which is directly inherited from source populations. In previous methods, only the decay curve of weighted LD between pairs of sites whose genetic distance were larger than a certain starting distance was fitted by single or multiple exponential functions, for the inference of recent single- or multiple-wave of admixture. However, the effect of SLD has not been well defined and no tool has been developed to estimate the effect of SLD on weighted LD decay. In this study, we defined the SLD in the formularized weighted LD statistic under the two-way admixture model, and proposed polynomial spectrum (p-spectrum) to study the weighted SLD and weighted LD. We also found reference populations could be used to reduce the SLD in weighted LD statistic. We further developed a method, iMAAPs, to infer Multiple-wave Admixture by fitting ALD using Polynomial spectrum. We evaluated the performance of iMAAPs under various admixture models in simulated data and applied iMAAPs into analysis of genome-wide single nucleotide polymorphism data from the Human Genome Diversity Project (HGDP) and the HapMap Project. We showed that iMAAPs is a considerable improvement over other current methods and further facilitates the inference of the histories of complex population admixtures.

Keywords: Population admixture; Admixture Linkage Disequilibrium (ALD); Source Linkage Disequilibrium (SLD); Admixture dating; Polynomial spectrum.

Introduction

The “Out of Africa” human migrations result in population differentiation in different continents, while subsequent migrations that have occurred over the past millennia led to gene flow among previously separated human sub-populations. As a consequence, admixed populations come into being when previously mutually isolated populations met and intermarry. Population admixture has received a great deal of attention recently. Many studies based on genome-wide data have shown that gene flow is common among inter-continental and intra-continental populations and that admixture of populations often leads to extended linkage disequilibrium (LD), which can greatly facilitate the mapping of human disease genes (McKeigue 2005; Reich and Patterson 2005; Smith and O’Brien 2005).

The high levels of LD are produced by admixture at loci that have different allele frequencies among the involved populations (Nei and Li 1973). Because of recombination, this particular type of admixture LD, or ALD, decays as a function of time since admixture. Consequently, it is possible to infer population admixture by modeling the dynamic changes of ALD. Moorjani et al. proposed such an approach by

aggregating pairwise LD measurements through a weighting scheme (Moorjani *et al.* 2011) and its software ROLLOFF was fully explained by Patterson N. *et al.* (Patterson *et al.* 2012), which was further developed by Loh *et al.* (Loh *et al.* 2013) and by Pickrell *et al.* (Pickrell *et al.* 2014). This ALD-based approach is particularly useful for admixture dating.

Under the hybrid isolation (HI) model, the expected value of LD decreases at the rate of $1 - d$ (Chakraborty and Weiss 1988; Pfaff *et al.* 2001), where d is the genetic distance (in Morgan) between two sites. And after g generation, the LD decays to $(1 - d)^g$ of its original value, assuming that the admixed population is engaged in random mating and has infinite effective population size (Hill and Robertson 1966). Recently, Pickrell *et al.* considered the situation of multiple waves of admixture from different source populations and showed that LD was comprised of multiple exponential terms, each of which refers to a single admixture event (Pickrell *et al.* 2014). Zhou *et al.* confirmed the polynomial expression (taking e^{-ld} as approximation of $(1 - d)^l$) for each wave of admixture and added the effect of source LD (SLD) from source populations into the LD's expression under the general admixture model (Zhou *et al.* 2016). Based on this LD framework, dating admixture becomes a problem of fitting the polynomial terms in the ALD decay.

When dating admixture in empirical populations, two major factors affect the accuracy of estimation: background LD (or SLD in the context of this work) and representative reference populations. Pickrell *et al.* presented a method based on weighted LD to deal with multiple-wave admixture. In their method, they used starting distance strategy (abandon loci whose genetic distance are shorter than a certain distance) to reduce the bias caused by SLD and they scanned global populations to find out the best pair of reference populations for each wave of admixture (Pickrell *et al.* 2014). The key assumption of their method is that the only effect by different pairs of reference populations was resulted from the relative value of exponential/polynomial coefficients of weighted LD decay. However, they neither validated this assumption nor considered the possible effect from SLD. Here, we introduced the polynomial spectrum (p-spectrum), the fitting results with polynomial functions, to reveal polynomial property of the weighted LD decay. With simulated admixed population, we confirmed the weighted LD decay curves with different pairs of source populations had similar p-spectrum and we also found that starting distance strategy could only partly reduce the SLD (Figure 3).

An alternative way to reduce SLD is to use derived source populations to estimate SLD (Zhou *et al.* 2016). Based on this idea, a new approach is developed here to infer multiple-wave admixture, which has been implemented in iMAAPs. After evaluating this method under various admixture models, we applied it to the well-known admixed populations in HGDP (Rosenberg *et al.* 2002) and HapMap (The

International HapMap Consortium 2010) data and demonstrate that the current study has greatly facilitated the understanding of the admixture history of human populations.

Materials and Methods

Data sets

Data for simulation and empirical analysis were obtained from two public resources: Human Genome Diversity Panel (HGDP) (Rosenberg *et al.* 2002) and the International HapMap Project phase III (The International HapMap Consortium 2010). Data filtering was performed within each population with Plink (Wigginton *et al.* 2005): Samples with missing rate > 5% per individual, SNPs with missing rate > 50% and SNPs failing in Hardy-Weinberg Equilibrium test ($p\text{-value} < 1 \times 10^{-6}$) were permanently removed from subsequent analyses.

The abbreviation of populations used in this study are: YRI: The Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya; MKK: Maasai in Kinyawa, Kenya; ASW: African Ancestry in SW USA; CEU: U.S. Utah residents with ancestry from northern and western Europe; TSI: Tuscans in Italy; MXL: Mexican Ancestry in LA, CA, USA; CHB: Han Chinese in Beijing, China; CHD: Chinese in metropolitan Denver, CO, USA. Haplotypes used as source populations in simulations are from 113 unrelated individuals of CEU and 113 unrelated individuals of YRI.

Simulations

In order to evaluate our method in dating admixture, we employed forward-time simulations to generate haplotypes under variant admixture scenarios: HI model, two-wave model (including the cases of one donor population and two donor populations for the second wave admixture), and the model of isolation after a period of continuous admixture. Our simulations were under the framework of copying model that new haplotypes are assembled from the segments of the source populations' haplotypes generation by generation (Li and Stephens 2003; Price *et al.* 2009), which has been used in previous work (Price *et al.* 2009). In our simulation, no mutation was considered in generating new haplotypes.

Under the HI model, the admixture events were set as having occurred 20, 50, 100, and 200 generations ago. For the two-wave (TW) model, the simulated admixed population experienced two pulses of admixture, which were at 100 and 20 generations ago, respectively, and was isolated in the remaining time. For the recent admixture in the TW model, a scenario in which only one of the source populations donated genetic materials (TW-1 model) and the other scenario where both source populations provided gene flow (TW-2 model) were simulated.

We also simulated admixed haplotypes in the scenarios of continuous migration, in which only gene flow from source populations to the admixed population is allowed and after that the admixed population is isolated to the present. In our simulation, we used modified gradual admixture (GA) (Jin *et al.* 2012) and continuous gene flow (CGF) (Pfaff *et al.* 2001) models to shape the gene flow in the migration window, which separately resulted in GA-I model and CGF-I model. Under these two models, we set the window size of migration as 80 generations and the isolation duration as 20 generations for the long last migration, and we set the window size of migration as 30 generations and the isolation duration as 70 generations for the short last migration.

Source populations also evolved in isolation so that both the reference populations and admixed population were of the same age. The sample sizes for both source populations and admixed populations were set as 5,000. More details of parameters for simulation were given in Table S1–3.

Weighted LD statistic and its estimator under the two-way admixture model

Under the two-way admixture model (Figure 1), two source populations provide genetic materials to the newly formed admixed population. Following the notations of Zhou *et al.* (Zhou *et al.* 2016), the LD in the admixed population of $(n + 1)$ -st generation is composed of SLD and admixture produced LD:

$$D_0 = \sum_{i=1}^2 m_i D_i + \delta_{12}(x) \delta_{12}(y) \sum_{l=1}^n c^{(l)} (1 - d)^l, \quad Eq 1$$

where m_i is the genetic proportion derived from the source population i , serving as the weight for linear combination of D_i (LD in source populations i) to form the SLD; $\delta_{12}(x)$ is the allele frequency difference between population 1 and population 2 at site x and d is the genetic distance between site x and site y ; $c^{(l)}$ is a natural admixture indicator whose positive value means that admixture happened at l generations ago:

$$c^{(l)} = \left(m_1^{(n+1-l)} m_2^{(n+1-l)} + m_0^{(n+1-l)} m_1^{(n+1-l)} \left(w_2^{(n-l)} \right)^2 + m_0^{(n+1-l)} m_2^{(n+1-l)} \left(w_1^{(n-l)} \right)^2 \right) \times \left(\prod_{j=n+2-l}^n m_0^{(j)} \right),$$

where $w_i^{(l)}$ is the total genetic contribution from source population i in the admixed population $B^{(l)}$.

Using allele frequency difference $\delta_{12}(x) \delta_{12}(y)$ as weight, the weighted LD statistic is defined as the average of LD with the weight over a set holding pairs of SNPs whose pairwise genetic distance is d (Loh *et al.* 2013):

$$a_i(d) = \frac{\sum_{S(d)} D_i \delta_{12}(x) \delta_{12}(y)}{|S(d)|},$$

1 where

$$S(d) = \left\{ (x, y) : d - \frac{\varepsilon}{2} < |x - y| < d + \frac{\varepsilon}{2} \right\},$$

2 and ε is a discretization parameter inducing a discretization on d . By taking average with weight on both
3 sides of **Error! Reference source not found.** over the set $S(d)$, we have

$$a_0(d) = \sum_{i=1}^2 m_i a_i(d) + F(d) \sum_{l=1}^n c^{(l)} (1 - d)^l, \quad \text{Eq 2}$$

4 where

$$F(d) = \frac{\sum_{S(d)} (\delta_{12}(x) \delta_{12}(y))^2}{|S(d)|}.$$

5 And the estimators for the weighted LD statistic for admixed population and source populations are given
6 by Loh P. R. et. al (Loh *et al.* 2013):

$$\widehat{a_i(d)} = \frac{\sum_{S(d)} \widehat{cov(X, Y)} \widehat{\delta_{12}(x)} \widehat{\delta_{12}(y)}}{|S(d)|}, \quad i = 0, 1, 2;$$

7 $\widehat{\delta_{12}(x)}$ is the observed allele frequency difference and $\widehat{cov(X, Y)}$ is the estimator of D_i between site x and
8 site y . For the source populations, $i = 1$ or 2 , and $\widehat{a_i(d)}$ defined above is a biased estimator when the
9 same group of samples are used for calculating both LD and the weight for the LD. Fortunately, two ways
10 can be used for eliminating the bias: 1), divide the target population into two groups, one group is used
11 for calculating the allele frequency difference, while the other group is used for calculating the LD
12 (Moorjani *et al.* 2011); 2), employ the unbiased statistics (Loh *et al.* 2013). In this study, we used the
13 second method to correct the bias in the SLD estimation. Besides, $F(d)$ can be independently estimated
14 by:

$$\widehat{F(d)} = \frac{\sum_{S(d)} (\widehat{\delta_{12}(x)} \widehat{\delta_{12}(y)})^2}{|S(d)|}.$$

15 Here, we separated $F(d)$ from the coefficients of polynomial functions to avoid the possible effect to the
16 process of polynomial fitting.

17 **Factorizing of weighted LD statistic with polynomial functions**

Based on formula of weighted LD statistic in the admixed population (Eq 2), admixture events are recorded in the polynomial function $\sum_{l=1}^n c^{(l)} (1-d)^l$, where positive value of $c^{(l)}$ indicates the admixture at l th generation before present. So the direct way to date the admixture is to find out the positive value of $c^{(l)}$. However, two possible risks may affect the results when fitting $a_0(d)$ with polynomial functions: one is $a_i(d)$, $i = 1, 2$, which represents the SLD; the other is $F(d)$, which is a function decaying as d increases (Figure S3). Inspired by the Weierstrass approximation theorem, which says continuous function can be approximated by polynomial function, we used polynomial function to approximate $a_i(d)$, $i = 0, 1, 2$, and $F(d)$ to explore the possible interaction between them. Actually, by fitting the decay curve with polynomial function $\sum_{l \in S_g} b^{(l)} (1-d)^l$, we got the spectrum that the $b^{(l)}$ values on set S_g , which we defined as *polynomial spectrum* (p-spectrum) (Figure 3). In the polynomial fitting, $b^{(l)}$ needs to be non-negative and S_g is a finite set holding the candidate time points for the possible admixture signals. This numeric method to generate the p-spectrum is illustrated in **Appendix**. Replacing $a_i(d)$, $i = 1, 2$, and $F(d)$ with polynomial functions

$$a_i(d) = \sum_{l \in S_g} b_{a_i}^{(l)} (1-d)^l, i = 1, 2$$

$$F(d) = \sum_{l \in S_g} b_F^{(l)} (1-d)^l,$$

$a_0(d)$ turns to be

$$a_0(d) = \sum_{l \in S_g} (m_1 b_{a_1}^{(l)} + m_2 b_{a_2}^{(l)}) (1-d)^l + F(d) \sum_{l=1}^n c^{(l)} (1-d)^l.$$

This expression of $a_0(d)$ tells us that SLD, linear combination of $a_i(d)$, $i = 1, 2$, would bring in false positive admixture signals while $F(d)$ has the potential to destroy the admixture time inference when we try to fit $a_0(d)$ directly with polynomial functions. So it is essential to evaluate the effect from $a_i(d)$, $i = 1, 2$, and $F(d)$. Fortunately, $a_i(d)$, $i = 1, 2$, and $F(d)$ can be estimated with the source populations so that the effect can be evaluated.

To evaluate the effect from $a_i(d)$, $i = 1, 2$, and $F(d)$, we simulated a 100-generation-old admixed population under HI model and the simulated admixed population was initiated with the haplotype of YRI and CEU of the proportion 50%:50%. Derived source populations of YRI and CEU were also generated, separately. Based on the simulated genotype data in both source populations and admixed population, both $\widehat{a_i(d)}$ and $\widehat{F(d)}$ can be calculated through a Fast Fourier Transform algorithm, which is able to increase the computational efficiency (Loh *et al.* 2013). After that, the p-spectrum was constructed on a

time set ranging from 0 to 2000 generations accordingly. In the spectrum of $\widehat{a_0(d)}$, we found three bunches of signals: two sharp bunches appeared around 100 and 1,250 generations and one flat bulb lay around 180 generations (Figure 2). Both in the spectrum of $\widehat{a_1(d)}$ and the spectrum of $\widehat{a_2(d)}$, we found signals around the value 1,250 generations and signals close to 250 generations (Figures S1, S2). And in the p-spectrum of $\widehat{F(d)}$, we found only a strong peak on the time 0 and two weak signal peaks over 250 generations ago, which explained the sharp decay in its decay curve (Figures S3), suggesting that we need to consider this effect to precisely resolve admixture. In the time spectrum of $\widehat{a_0(d)}$, signals around 100 could be explained easily by the designed admixture and both signals around 1250 and signals around 180 are probably introduced by the SLD. To test this explanation, we directly constructed $z(d)$ as

$$z(d) = \frac{a_0(d) - \sum_{i=1}^2 m_i a_i(d)}{F(d)} = \sum_{l=1}^n c^{(l)} (1-d)^l,$$

which can be estimated with the simulated admixed population and derived source populations by

$$\widehat{z(d)} = \frac{\widehat{a_0(d)} - \sum_{i=1}^2 m_i \widehat{a_i(d)}}{\widehat{F(d)}}. \quad \text{Eq 3}$$

In the p-spectrum of $\widehat{z(d)}$, the relative strength of noise-like signals outside the bunch of signals around 100 generation become much weaker than that in the p-spectrum of $\widehat{a_0(d)}$ (Figure S4). This result confirms our explanation on the p-spectrum of $\widehat{a_0(d)}$ and indicates source populations can be used to reduce effect of SLD.

Time inference for multiple-wave admixture

Because $c^{(l)}$ is a natural indicator for admixture event, the natural extension for the p-spectrum is to infer the admixture time. In this section, we are going to present the time inference method based on p-spectrum. In empirical populations, both the $\widehat{a_i(d)}$, $i = 0, 1, 2$ and $\widehat{F(d)}$ can be calculated based on the genotype data of the admixed population and reference populations. Meanwhile, the population admixture proportions are estimated by

$$\widehat{m_2} = \frac{\sum_x \widehat{\delta_{01}(x)} \widehat{\delta_{21}(x)}}{\sum_x (\widehat{\delta_{21}(x)})^2}, \quad \widehat{m_1} = \frac{\sum_x \widehat{\delta_{02}(x)} \widehat{\delta_{12}(x)}}{\sum_x (\widehat{\delta_{12}(x)})^2},$$

then $\widehat{z(d)}$ can be calculated so as its p-spectrum $\{c^{(l)}\}_{l \in S_g}$.

Next, we are going to date admixture and evaluate the existence for each wave of admixture with a Jackknife-based method. Suppose we have 22 autosomes for the target admixed population, and each chromosome is excluded one at each time to calculate decay curve of $\widehat{z(d)}$ (Loh *et al.* 2013). This means

1 when chromosome i is excluded, the remaining 21 chromosomes are used to calculate $\hat{z}_i$ and the p-
2 spectrum $\{c_i^{(l)}\}_{l \in S_g}$. Then the p-values are attained on each l by testing whether the values in the set
3 $\{c_i^{(l)}\}_{i=1, \dots, 22}$ is bigger than zero, and we used $c_m^{(l)}$, the median of $\{c_i^{(l)}\}_{i=1, \dots, 22}$, as the summary p-
4 spectrum for the target population. We could also use the mean value to construct $c_m^{(l)}$, but it would lead
5 to more false admixture signals. Based on the summary p-spectrum, l with positive values of $c_m^{(l)}$ are
6 gathered as the candidate admixture time points and then they are clustered into groups as different waves
7 of admixture, say $S_g^{(k)}$ for k th wave of admixture. Once these time points are grouped, the mean and
8 variance of the time for that wave of admixture can be calculated by

$$\begin{aligned} \text{mean}(T^{(k)}) &= \frac{\sum_{l \in S_g^{(k)}} l \times c_m^{(l)}}{\sum_{l \in S_g^{(k)}} c_m^{(l)}}, \\ \text{var}(T^{(k)}) &= \frac{\sum_{l \in S_g^{(k)}} (l - \text{mean}(T^{(k)}))^2 \times c_m^{(l)}}{\sum_{l \in S_g^{(k)}} c_m^{(l)}}. \end{aligned}$$

9 Meanwhile, we used the minimum p-value on each time point in that group to measure the
10 significance for each wave of admixture. In this way, we could date multiple-wave admixture and
11 measure the significance of each wave of admixture. This method has been implemented in the Software
12 named iMAAPs, which is available at <http://www.picb.ac.cn/PGG/resource.php>

13 Results

14 Dating multiple-wave admixture with weighted LD statistic

15 There are two main difficulties for dating admixture in empirical analysis: reference populations selection
16 and SLD reduction. To deal with these two difficulties, Pickrell et. al claimed that different pairs of
17 reference populations often have different relative values but always have the same sign of the coefficient
18 of $(1 - d)^l$ so that they can traverse all pairs of reference populations to test the presence for the possible
19 admixture and estimate the time for each wave of admixture. And they used the LD whose pairwise
20 genetic distance is longer than 0.5 centiMorgans (cM), which was supposed to reduce the effect of SLD.
21 Meanwhile, they also claimed that their algorithm was not very powerful in detecting multiple admixture
22 (Loh et al. 2013; Pickrell et al. 2014). Under our framework of weighted LD (Eq 2), we confirmed that
23 $c^{(l)}$ is an admixture determined constant that it is independent to the selection of reference populations;

we pointed out both the $a_i(d)$, $i = 1, 2$ and $F(d)$ have the potential to affect the p-spectrum of $a_0(d)$, which directly affect the estimation of the coefficient of $(1 - d)^l$; we also noticed that the effect of SLD reduction with starting distance was also not evaluated in Pickrell et. al's work, which may be the reason that their method is not so powerful in detecting multiple admixture. To verify our conjecture, we constructed the summary p-spectrum for weighted LD decay curves on a simulated admixed population with different pairs of reference populations.

A 100-generation-old admixed population was generated under HI model, with YRI and CEU as source populations of admixture proportion 0.5:0.5. Total 55 pairs of HapMap populations (YRI, LWK, MKK, ASW, CEU, TSI, MXL, CHB, CHD, GIH, JPT) were used as references to calculate weighted LD $a_0(d)$ for further p-spectrum construction. In the summary, p-spectrum with full weighted LD decay (Figure 3A) for nearly all pairs of reference populations arose three main peaks around 100, 180 and 1250 generations. In the p-spectrum with weighted LD decay began at 0.5 cM (Figure 3B), the peak around 180 and 1250 generations disappeared but a new peak around 120 generations arose, which was probably the remaining SLD and it may bias the time estimation of admixture. The remaining SLD should be the reason why ALDER did not work well for multiple admixture. Meanwhile, we also observed that weighted LD decay with the pairs of reference populations close to the true source populations would have similar p-spectrum to what we want to reveal (Figure 3), which indicated us to use populations not the exact but similar to the source populations as references to construct the p-spectrum. This observation also supported that using proper reference populations could increase the accuracy of ALDER in resolving weighted LD decay.

Evaluation of iMAAPs

In our p-spectrum based method iMAAPs, we used reference populations to estimate SLD and $F(d)$, and separated their effect from the weighted LD of admixed population. Thus, we could directly estimate the admixture determined parameters $c^{(l)}$ and the time and the waves of admixture. A workable method in empirical admixture analysis should be robust to the proxy source populations. We have observed the robustness of p-spectrum to different pairs of reference populations, so we will evaluate the performance of iMAAPs to different reference pairs. Here, with the simulated 100-generation-old African-European admixed population, generated by YRI and CEU, we showed that iMAAPs is very robust with African-European pairs (YRI, CEU; LWK, CEU; MKK, CEU; LWK, TSI; YRI, TSI; MKK, TSI) as reference populations to infer the admixture time (Table S4).

We also tested our method under various admixture model, iMAAPs is able to reconstruct the history of the admixture population well. For the one-pulse and two-pulse admixture models, iMAAPs gave the

time close to the true admixture; for the continuous migration models, it was able to place most of the signals in a particular migration time window (Figure 4, Figure S5-13).

Empirical analysis

The current method was first applied to a few well-known admixed populations from available public databases: HGDP (Rosenberg *et al.* 2002) and HapMap Project phase III (The International HapMap Consortium 2010). Our method is currently designed under the framework of two-way admixture and source populations or the populations similar to which are required in empirical analysis. Besides, two principles should be considered for interpretation:

(1) Existence of estimations for longer than 500 generations indicates the SLD has not been well removed so that some of the admixture signals, especially for ancient ones, are probably generated by the SLD instead of the admixture.

(2) Existence of estimations close to generation 1 is always considered as the result of the population substructure but not the admixture.

Based on these principles, we first analyzed three well-known admixed populations: African American (57 ASW individuals from HapMap), Mexican (86 MXL individuals from HapMap) and Uygur (10 Uygur individuals from HGDP). We also used ALDER to analyze these admixed populations (Table S5). For each admixed population, we conducted three rounds of estimations. In the first round, we used all the populations in the full data set as the references to infer the admixture; in the second round, we used population pairs with the highest amplitude for each wave of admixture in the first round as the reference populations to re-run ALDER; in the last round, we selected populations according to the admixture pattern based on the population inference in the first round. That is to say, if CEU and YRI are inferred as the best pair of populations to explain the admixture, then we selected all populations that could represent European ancestries and African ancestries as reference populations in the third run of ALDER, which we believed would increase the estimation accuracy.

In our analysis with iMAAPs, reference populations were selected based on the results of ALDER's inference on each wave of admixture. CEU (n = 113) and YRI (n = 147) were chosen as the ancestral populations of ASW. YRI (n = 147), TSI (n = 102) and American Indian (7 Colombians, 14 Karitiana, 21 Maya, 14 Pimas and 8 Suruis from HGDP) were used as the ancestral populations of MXL. Basque (n = 24), Sardinian (n = 28), Japanese (n = 28), Han (n = 34) and French (n = 28) were used as the ancestral populations of Uygur.

The estimation of admixture time for ASW is 5.4 ± 0.4 generations ago and the SLD has been well reduced with YRI and CEU as reference populations (Table 1, Figure S15). In the meantime, ALDER gives us two different results: 12.0 ± 4.4 generations with all populations in HapMap as references; 6.3 ± 3.3 and 77.0 ± 65.9 generations with selected reference populations from HapMap (Table S4). In this estimation of ALDER, generation 6.3 is very close to our result, which can be interpreted as the admixture time of the population ASW. And the result on generation 77.0 reflects the failure of SLD reduction with starting distance of 0.5 cM.

MXL seems to have experienced the main admixture of 7.0 ± 0.2 generations ago with TSI and American Indian as reference populations; 8.2 ± 0.4 generations ago with YRI and American Indian as reference populations (Table 1). More admixture time points would be detected using the mean to construct summary p-spectrum (Table S5), which needs to be confirmed by further studies.

The Uyghur population has been reported to have much longer admixture history than ASW and MXL (Xu and Jin 2008; Xu *et al.* 2008; Qin *et al.* 2015). The admixture is resolved at 33.3 ± 0.5 generations ago with Han and French as reference populations. This admixture event has also been detected with Basque, Han, Sardinian and Japanese as reference populations, suggesting that the major admixture formatting ALD in the present population happened around 825 years ago.

Loh *et al.* speculated that there could have been multiple waves of admixture in the history of MKK (Loh *et al.* 2013). Here, both our method and ALDER detected at least two waves of admixture (Table 1, Table S4, and Figure S15). We used reference pair of YRI and CEU and pair of YRI and TSI to resolve the admixture of the MKK. With YRI and CEU as references, admixtures around 16.2 and 68.4 generations ago were detected; with YRI and TSI as references, admixtures around 17.9 and 70.3 generations ago were detected. However, both detections have estimations longer than 500 generations ago, indicating that we need to take care of SLD when interpreting the time of admixture.

Discussion

Available methods based ALD for admixture dating have shown their robustness in dealing with genotype data (Loh *et al.* 2013) and complicated admixture history (Pickrell *et al.* 2014). However, the effect of SLD in these methods has not been well defined and efficiently reduced, which may bias the estimation. In this study, we defined the SLD in the weighted LD statistic of the target admixed population under two-way admixture model, and introduced the p-spectrum to study the weighted LD decay for both source populations and admixed population. We found that SLD pretends to have higher degree of p-spectrum than the LD introduced by recent admixture and using starting distance can partly reduce the effect of SLD (Figure 3). We also found that SLD can be well compensated by the LD of

source populations (Figure S5-S13), which motivated us to develop a new method iMAAPs to infer multiple wave admixture. In this method, we used reference populations to estimate the SLD and then reduced its effect and gave the accurate estimation.

We applied both iMAAPs and ALDER to date several well-known admixed populations. When running ALDER, we conducted estimation in three rounds with different sets of reference populations. Based on the ALDER's assumption of the effect of different pairs of reference populations on weighted LD, time estimations in these three rounds should be much close to each other. However, for the population of ASW, the result of first round is different from the results of other rounds, indicating the potential risk for using ALDER with global references. In the estimations of the second and third round, we found signals around 6 and 70 generations before present, and the signal around 70 generations should be the false admixture signal caused by the remaining SLD, because iMAAPs only detected the signal close to the 6 and the SLD was well reduced. And we found signals were close to 0 in all these three rounds of estimation, which, as we suggested, should be interpreted as population substructure instead of admixture time. We also ran iMAAPs in populations of MXL, Uygur, and MKK. We found MXL seems to experience admixture at 7 generations ago. For the Uygur, the major admixture happened at about 33 generations ago. This date is a little earlier than the date with ALDER with three rounds, in which the most significant admixture is around 40 generations ago. The difference is probably caused by the remaining SLD with starting distance strategy. Both ALDER and iMAAPs confirmed the estimation that the population MKK experienced multiple waves of admixture, which had been predicted in the work of Loh et al. (Loh *et al.* 2013)

One of the fundamental idea in this algorithm is to take advantage of proper representative reference populations to reduce the effect of SLD. Since using improper reference populations may bias the final estimation, it is crucial to carefully select reference populations in empirical analysis. Another issue that should be noticed is that this algorithm might give multiple pulses of signals even though the true population history was long-last continuous admixture (Figure S5-8). Nevertheless, this work built the framework of weighted LD under two-way admixture and provided an alternative way to reduce the effect of SLD in estimation of admixture time.

Author contributions:

Conceived and designed the study: SX. Led this study: YZ. Developed methods and computer tools: YZ KY. Analyzed the data: YZ and KY. Interpreted the data and wrote the paper: SX YZ KY YY XN PX EX.

Competing interests: The authors have declared that no competing interests exist.

Acknowledgements:

We are grateful to Dr. Li Jin, Dr. Yungang He, Dr. XiaoMing Liu and Dr. Minxian Wang for their valuable suggestions. This work is supported by the Strategic Priority Research Program (XDB13040100) and Key Research Program of Frontier Sciences (QYZDJ-SSW-SYS009) of the Chinese Academy of Sciences (CAS), the National Natural Science Foundation of China (NSFC) grant (91331204), the National Science Fund for Distinguished Young Scholars (31525014), the Program of Shanghai Academic Research Leader (16XD1404700), and the National Key Research and Development Program (2016YFC0906403). S.X. is Max-Planck Independent Research Group Leader and member of CAS Youth Innovation Promotion Association. S.X. also gratefully acknowledges the support of the National Program for Top-notch Young Innovative Talents of The "*Wanren Jihua*" Project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reference

- Chakraborty R., Weiss K. M., 1988 Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. U. S. A.* **85**: 9119–9123.
- Hill W. G., Robertson a, 1966 The effect of linkage on limits to artificial selection. *Genet. Res.* **8**: 269–294.
- Jin W., Wang S., Wang H., Jin L., Xu S., 2012 Exploring population admixture dynamics via empirical and simulated genome-wide distribution of ancestral chromosomal segments. *Am. J. Hum. Genet.* **91**: 849–862.
- Li N., Stephens M., 2003 Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165**: 2213–2233.
- Loh P. R., Lipson M., Patterson N., Moorjani P., Pickrell J. K., Reich D., Berger B., 2013 Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**: 1233–1254.
- McKeigue P. M., 2005 Prospects for admixture mapping of complex traits. *Am. J. Hum. Genet.* **76**: 1–7.
- Moorjani P., Patterson N., Hirschhorn J. N., Keinan A., Hao L., Atzmon G., Burns E., Ostrer H., Price A. L., Reich D., 2011 The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.* **7**: e1001373.
- Nei M., Li W. H., 1973 Linkage disequilibrium in subdivided populations. **75**: 213–219.
- Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y., Genschoreck T., Webster T., Reich D., 2012 Ancient admixture in human history. *Genetics* **192**: 1065–1093.
- Pfaff C. L., Parra E. J., Bonilla C., Hiester K., McKeigue P. M., Kamboh M. I., Hutchinson R. G., Ferrell R. E., Boerwinkle E., Shriver M. D., 2001 Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.* **68**: 198–207.
- Pickrell J. K., Patterson N., Loh P.-R., Lipson M., Berger B., Stoneking M., Pakendorf B., Reich

- 1 D., 2014 Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad.*
2 *Sci. U. S. A.* **111**: 2632–7.
- 3 Price A. L., Tandon A., Patterson N., Barnes K. C., Rafaels N., Ruczinski I., Beaty T. H.,
4 Mathias R., Reich D., Myers S., 2009 Sensitive detection of chromosomal segments of
5 distinct ancestry in admixed populations. *PLoS Genet.* **5**.
- 6 Qin P., Zhou Y., Lou H., Lu D., Yang X., Wang Y., Jin L., Chung Y.-J., Xu S., 2015
7 Quantitating and Dating Recent Gene Flow between European and East Asian Populations.
8 *Sci. Rep.* **5**: 9500.
- 9 Reich D., Patterson N., 2005 Will admixture mapping work to find disease genes? *Philos. Trans.*
10 *R. Soc. Lond. B. Biol. Sci.* **360**: 1605–1607.
- 11 Rosenberg N. a, Pritchard J. K., Weber J. L., Cann H. M., Kidd K. K., Zhivotovsky L. a,
12 Feldman M. W., 2002 Genetic structure of human populations. *Science* **298**: 2381–2385.
- 13 Smith M. W., O’Brien S. J., 2005 Mapping by admixture linkage disequilibrium: advances,
14 limitations and guidelines. *Nat. Rev.* **6**: 623–632.
- 15 The International HapMap Consortium, 2010 Integrating common and rare genetic variation in
16 diverse human populations. *Nature* **467**: 52–58.
- 17 Wigginton J. E., Cutler D. J., Abecasis R., 2005 A Note on Exact Tests of Hardy-Weinberg
18 Equilibrium.
- 19 Xu S., Jin L., 2008 A Genome-wide Analysis of Admixture in Uyghurs and a High-Density
20 Admixture Map for Disease-Gene Discovery. *Am. J. Hum. Genet.* **83**: 322–336.
- 21 Xu S., Huang W., Qian J., Jin L., 2008 Analysis of Genomic Admixture in Uyghur and Its
22 Implication in Mapping Strategy. *Am. J. Hum. Genet.* **82**: 883–894.
- 23 Zhou Y., Liu X., Xu S., 2016 Dissecting admixture linkage disequilibrium under a general model
24 of population admixture. *Under Rev.*

1 Appendix

2 Construct the p-spectrum for decay curve

3 The decay curve of $z(d)$ can be fitted by a numerical routine known as the proximal gradient
4 (Beck A. and Teboulle 2009) to minimize the objective function to fit out the parameter c :

$$\min_{c^{(j)} \geq 0 \text{ and } \sum c^{(j)} < C} \|Z - Ac\|_2,$$

5 where $Z = (z(d_1), z(d_2), \dots, z(d_{s-1}), z(d_s))^T$ is a vector of $z(d)$ with different d values.
6 $c = (c^{(0)}, c^{(1)}, \dots, c^{(n-1)}, c^{(n)})^T$ is the coefficient of the polynomial functions and all entries in
7 vector c are non-negative. The (i, j) th entry of the matrix $A_{s \times (n+1)}$ is $A_{ij} = (1 - d_i)^j$, where
8 $j = 0, 1, \dots, n$. In empirical analysis, the value of j ranges from 0 to thousands, say 0 to 2000 in
9 our analysis, which would lead the matrix A_{ij} to be too large to compute efficiently. In order to
10 increase the computation efficiency, we use the set S_g , a subset containing time candidates
11 sampled from the range of 0 to 2000, as the j 's value set. With this method, it was possible to
12 find the vector c so as the fitting curve Ac . Next, we construct the p-spectrum from denoising the
13 vector c .

14 For the vector c , each entry represents the magnitude of the signal and large magnitude
15 indicates that the signal contains information rather than noise. Based on this idea, only the top
16 $c^{(l)}$ values that together composed 99.9% of Z were retained for the p-spectrum construction.
17 This means that we need to find out the smallest set $\Omega \subseteq S_g$ subject to the condition

$$\frac{\sum_{j=1}^s \sum_{l \in \Omega} c^{(l)} e^{-ld_j}}{\sum_{j=1}^s z(d_j)} > 0.99,$$

18 and then let $c^{(l)}$, $l \in S_g \setminus \Omega$, to be zero. In this way, we constructed the p-spectrum for the decay
19 curve Z .

20 Clustering the candidate time points

21 Suppose we have the candidate time points generated from summary p-spectrum as increasing
22 series $\{t_i\}_{i=1,2,\dots,I}$, then we say t_i and t_{i+1} belong to the same cluster if only one of these two
23 conditions stands:

$$t_{i+1} - t_i = 1;$$

1 or

$$\frac{t_{i+1} - t_i}{t_{i+1} + t_i} < 0.1.$$

2 Reference

3 Beck A., Teboulle M., 2009 A Fast Iterative Shrinkage-Thresholding Algorithm for Linear
4 Inverse Problems. {SIAM} J. Imaging Sci. **2**: 183–202.

5

6

Table 1: Time of admixture (generations) determined using iMAAPs with selected reference populations.

Admixed population	Reference populations	Time 1	Time 2	Time 3	Time 4
ASW	CEU;YRI	0.0±0.0	5.4±0.4		
		3.92×10^{-40}	4.15×10^{-17}		
MXL	American Indian;TSI	0.1±0.3	7.0±0.2		
		8.69×10^{-27}	1.70×10^{-13}		
MXL	American Indian;YRI	0.0±0.0	8.2±0.4		
		1.82×10^{-31}	8.15×10^{-19}		
MKK	CEU;YRI	1.0±0.0	16.2±0.38	68.4±1.5	653.0±15.0
		1.33×10^{-28}	1.10×10^{-4}	5.47×10^{-5}	3.43×10^{-20}
MKK	TSI;YRI	1.1±0.3	17.9±0.23	70.3±1.4	632.1±15.1
		5.26×10^{-28}	7.11×10^{-5}	4.29×10^{-5}	3.36×10^{-13}
Uygur	Basque;Han	0.0±0.0	2.4±0.5	30.0±0.0	
		4.37×10^{-4}	4.22×10^{-7}	2.48×10^{-6}	
Uygur	French;Han	0.0±0.0	4.4±0.5	33.3±0.5	
		5.81×10^{-12}	4.54×10^{-5}	6.23×10^{-5}	
Uygur	Sardinian; Japanese	2.8±0.1	34.0±0		
		9.71×10^{-12}	7.65×10^{-4}		

In each time cell, the mean value ± standard derivation is listed with P-value for the significance of existence of that wave of admixture.

Titles and legends to figures

Figure 1: Two-way admixture model with n waves of admixture.

Figure 2: p -spectrum for $\widehat{a_0(d)}$ in a simulated admixed population. The observed weighted LD decay (gray points at top right) are fitted by hundreds of polynomial functions (gray curves in the bottom panel), and a few of them whose coefficients are positive (highlighted in heat color) and the amplitude for each positive coefficient are plotted along the value of l (generation before present) at top left.

Figure 3: Summary p -spectrum for $\widehat{a_0(d)}$ with all pairs of populations from HapMap. Spectrum with true source populations (CEU, YRI) is in red lines; spectrum with selected pairs of reference populations (CEU,LWK; CEU, MKK; TSI, LWK; TSI, YRI; TSI, MKK;) are in black lines; spectrums with other pairs of reference populations are in gray lines. **A:** Summary p -spectrum for full LD decay; **B** Summary spectrum for LD decay of starting distance 0.5 cM.

Figure 4: The performance of iMAAPs under various admixture models. The black vertical dash lines represent the true simulated admixture time and gray areas represent the time window for continuous admixture. The summary p -spectrum of $\widehat{Z(d)}$ for each simulated admixed population is plotted in heat color and the estimated admixture times are plotted in blue points, the mean values, and lines, the ranges of 3 times of standard derivation.

Figure S1: p -spectrum for $\widehat{a_1(d)}$ in one of the reference populations (CEU). The observed weighted LD decay (gray points at top right) are fitted by hundreds of polynomial functions (gray curves in the bottom panel), and a few of them whose coefficients are positive (highlighted in heat color) and the amplitude for each positive coefficient are plotted along the value of l (generation before present) at top left.

Figure S2: p -spectrum for $\widehat{a_2(d)}$ in one of the reference populations (YRI). The observed weighted LD decay (gray points at top right) are fitted by hundreds of polynomial functions (gray curves in the bottom panel), and a few of them whose coefficients are positive (highlighted in heat color) and the amplitude for each positive coefficient are plotted along the value of l (generation before present) at top left.

Figure S3: p -spectrum for $\widehat{F(d)}$. The observed $\widehat{F(d)}$ values (gray points at top right) are fitted by hundreds of polynomial functions (gray curves in the bottom panel), and a few of them whose coefficients are positive (highlighted in heat color) and the amplitude for each positive coefficient are plotted along the value of l (generation before present) at top left.

Figure S4: p -spectrum for $\widehat{Z(d)}$. The observed $\widehat{Z(d)}$ values (gray points at top right) are fitted by hundreds of polynomial functions (gray curves in the bottom panel), and a few of them whose coefficients

are positive (highlighted in heat color) and the amplitude for each positive coefficient are plotted along the value of l (generation before present) at top left.

Figure S5: The performance of iMAAPs on 10 independent simulated admixed populations under CGF-I model (case 1). The gray area represents the time window for continuous admixture. The summary p-spectrum of $\widehat{Z(d)}$ for each simulated admixed population is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.

Figure S6: The performance of iMAAPs on 10 independent simulated admixed populations under CGF-I model (case 2). The gray area represents the time window for continuous admixture. The summary p-spectrum of $\widehat{Z(d)}$ for each simulated admixed population is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.

Figure S7: The performance of iMAAPs on 10 independent simulated admixed populations under GA-I model (case 1). The gray area represents the time window for continuous admixture. The summary p-spectrum of $\widehat{Z(d)}$ for each simulated admixed population is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.

Figure S8: The performance of iMAAPs on 10 independent simulated admixed populations under GA-I model (case 2). The gray area represents the time window for continuous admixture. The summary p-spectrum of $\widehat{Z(d)}$ for each simulated admixed population is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.

Figure S9: The performance of iMAAPs on 10 independent simulated admixed populations under HI model. The black vertical dash lines represent the true simulated admixture time. The summary p-spectrum of $\widehat{Z(d)}$ for each simulated admixed population is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.

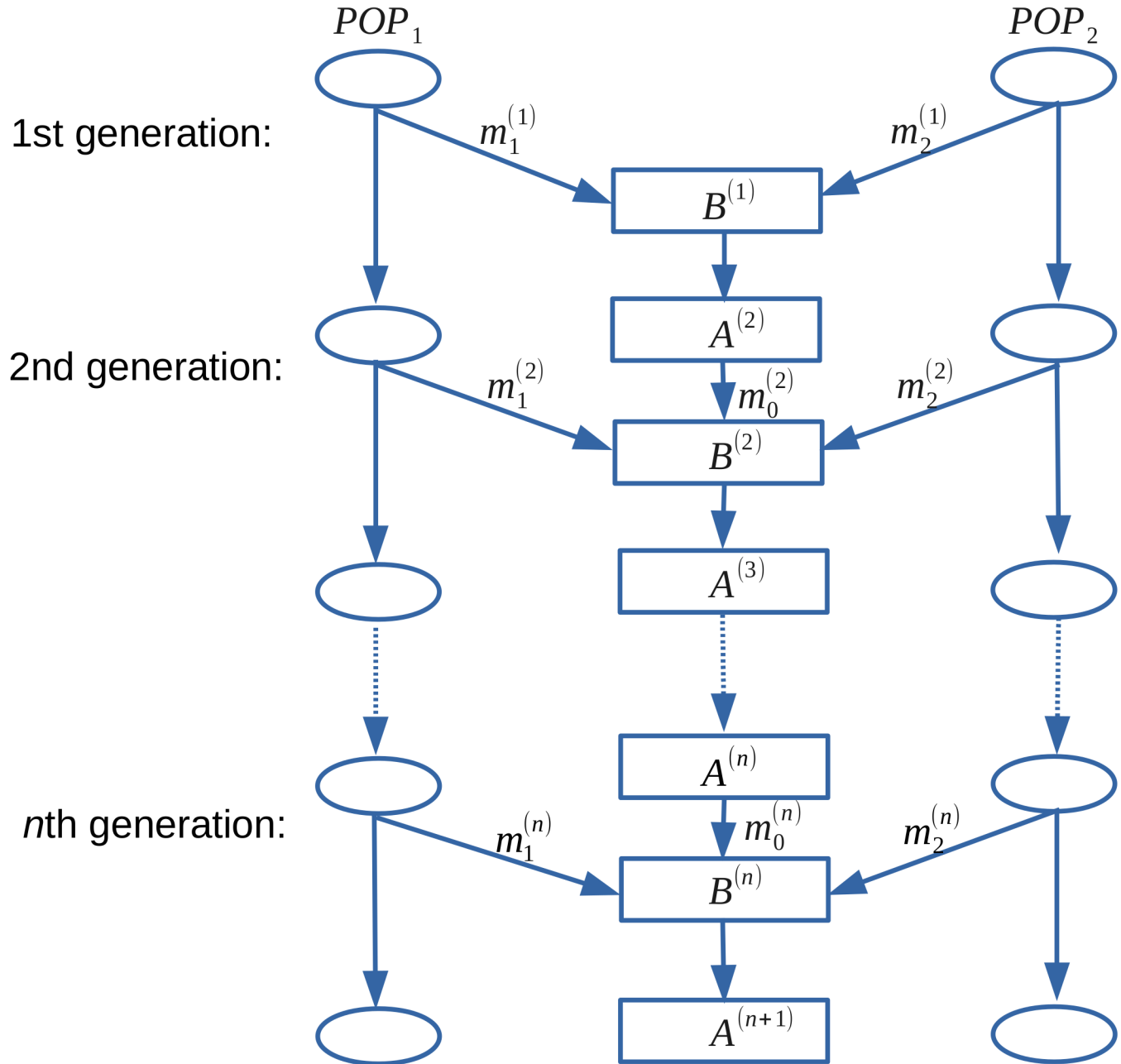
Figure S10: The performance of iMAAPs on 10 independent simulated admixed populations under TW-1 model (case 1). The black vertical dash lines represent the true simulated admixture time. The summary p-spectrum of $\widehat{Z(d)}$ for each simulated admixed population is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.

Figure S11: The performance of iMAAPs on 10 independent simulated admixed populations under TW-1 model (case 2). The black vertical dash lines represent the true simulated admixture time. The summary p-spectrum of $\widehat{Z(d)}$ for each simulated admixed population is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.

Figure S12: The performance of iMAAPs on 10 independent simulated admixed populations under TW-2 model (case 1). The black vertical dash lines represent the true simulated admixture time. The summary p-spectrum of $\widehat{Z}(d)$ for each simulated admixed population is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.

Figure S13: The performance of iMAAPs on 10 independent simulated admixed populations under TW-2 model (case 2). The black vertical dash lines represent the true simulated admixture time. The summary p-spectrum of $\widehat{Z}(d)$ for each simulated admixed population is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.

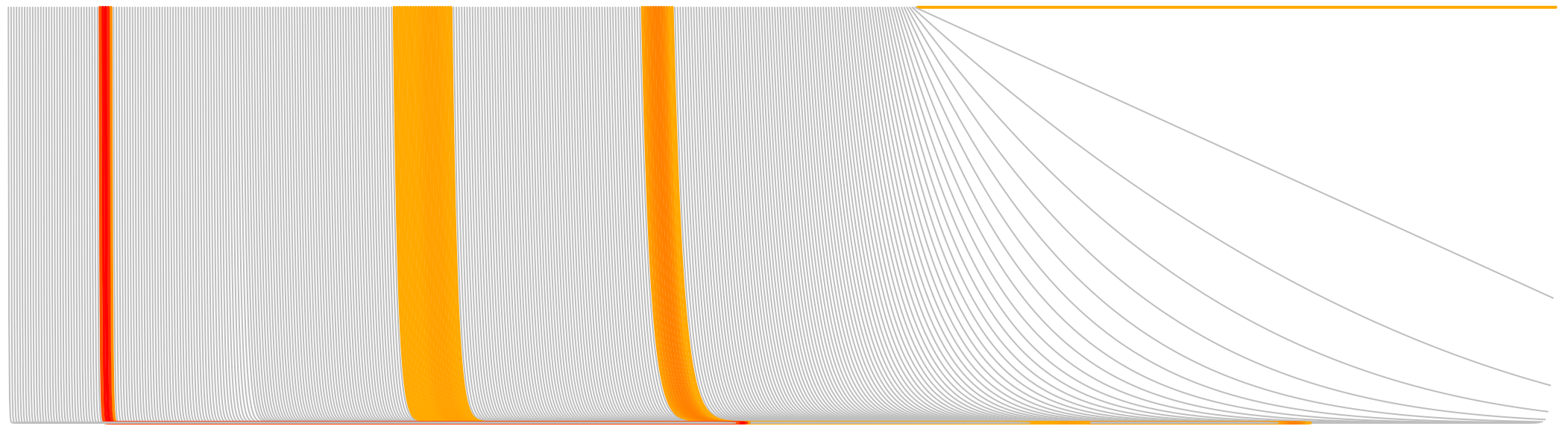
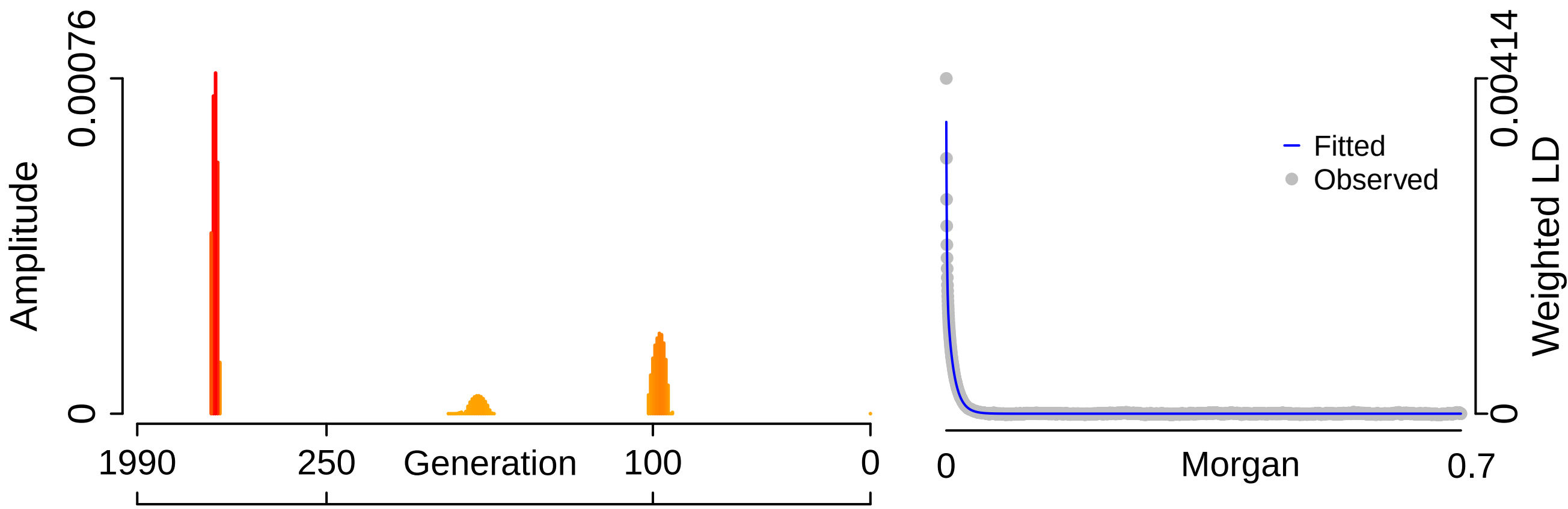
Figure S14: The performance of iMAAPs on empirical admixed populations. The summary p-spectrum of $\widehat{Z}(d)$ for each simulated admixed population, calculated with the median, is plotted in heat color. The blue points represent the mean values, and the blue lines represent the ranges of 3 times of standard derivation.



○ Source population

□ Admixed population

→ Gene flow



Weak Strong
Amplitude Strength

