

1 **Inference of multiple-wave population admixture by**
2 **modeling decay of linkage disequilibrium with**
3 **multiple exponential functions**

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

6
7 ^{*}Chinese Academy of Sciences (CAS) Key Laboratory of Computational Biology,
8 Max Planck Independent Research Group on Population Genomics, CAS-MPG
9 Partner Institute for Computational Biology, Shanghai Institutes for Biological
10 Sciences, Chinese Academy of Sciences, Shanghai, 200031, China;

11 [†]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA,
12 15213, USA;

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

15 ^{††}School of Life Science and Technology, ShanghaiTec University, Shanghai
16 200031, China;

17 ^{‡‡}Collaborative Innovation Center of Genetics and Development, Shanghai
18 200438, China.

19 [§]These authors contributed equally to this work.

20 ^{§§} Correspondence and requests for materials should be addressed to
21 [:xushua@picb.ac.cn](mailto:xushua@picb.ac.cn) (S.X.)

22
23

1 *Running title:* Inference of multiple-wave population admixture

2

3 **Keywords:** Population admixture; Linkage Disequilibrium (LD); Admixture

4 model; Multiple exponential functions; SNP

5

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

7 Computational Biology, Max Planck Independent Research Group on Population

8 Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai

9 Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai,

10 200031, China. Tel: +86-21-5492-0479. E-mail: xushua@gmail.com

11

1

Abstract

2 Admixture-introduced linkage disequilibrium (LD) has recently been
3 introduced into the inference of the histories of complex admixtures. However,
4 the influence of ancestral source populations on the LD pattern in admixed
5 populations is not properly taken into consideration by currently available
6 methods, which affects the estimation of several gene flow parameters from
7 empirical data. We first illustrated the dynamic changes of LD in admixed
8 populations and mathematically formulated the LD under a generalized
9 admixture model with finite population size. We next developed a new
10 method, MALDmef, by fitting LD with multiple exponential functions for
11 inferring and dating multiple-wave admixtures. MALDmef takes into account
12 the effects of source populations which substantially affect modeling LD in
13 admixed population, which renders it capable of efficiently detecting and
14 dating multiple-wave admixture events. The performance of MALDmef was
15 evaluated by simulation and it was shown to be more accurate than MALDER,
16 a state-of-the-art method that was recently developed for similar purposes,
17 under various admixture models. We further applied MALDmef to analyzing
18 genome-wide data from the Human Genome Diversity Project (HGDP) and
19 the HapMap Project. Interestingly, we were able to identify more than one
20 admixture events in several populations, which have yet to be reported. For
21 example, two major admixture events were identified in the Xinjiang Uyghur,
22 occurring around 27–30 generations ago and 182–195 generations ago,
23 respectively. In an African population (MKK), three recent major admixtures
24 occurring 13–16, 50–67, and 107–139 generations ago were detected. Our
25 method is a considerable improvement over other current methods and further
26 facilitates the inference of the histories of complex population admixtures.

27

1

Introduction

2 The “Out of Africa” human migrations resulted in population differentiation
3 in different continents, while subsequent migrations that have occurred over
4 the past millennia have resulted in gene flow among previously separated
5 human sub-populations. As a result, admixed populations come into being
6 when previously mutually isolated populations meet and intermarry.
7 Population admixture has received a great deal of attention recently. Many
8 studies based on genome-wide data have shown that gene flow has been
9 common among inter-continental and intra-continental populations and that
10 admixture of populations often leads to extended linkage disequilibrium (LD),
11 which can greatly facilitate the mapping of human disease genes(McKeigue
12 2005; Reich and Patterson 2005; Smith and O’Brien 2005; Seldin 2007).

13 The high levels of LD in recently admixed populations are due to
14 associations between pairs of loci co-inherited on an intact chromosomal
15 chunk from one of the ancestral source populations(Chakraborty and Weiss
16 1988). This particular type of admixture-introduced LD, or ALD, decays as a
17 function of time since admixture because of recombination. Consequently, it is
18 possible to infer population admixture by modeling the dynamic changes of
19 ALD. Patterson N. et al. recently proposed such an approach by aggregating
20 pairwise LD measurements through a weighting scheme (Patterson *et al.*
21 2012).This was further developed by Loh et al.(Loh *et al.* 2013) This ALD
22 based approach is particularly useful for admixture dating.

23 In the hybrid isolation (HI) model, the expected value of LD decreases at
24 the rate of $1-d$ (Chakraborty and Weiss 1988; Pfaff *et al.* 2001), where d is the
25 genetic distance (in Morgans) between two sites and g (generation) is the time

1 elapsed since the initial admixture event. For convenience, $(1-d)^g$ is
2 considered approximately equal to e^{-d^g} in computation, assuming that the
3 admixed population is engaged in random mating and has infinite effective
4 population size(Hill and Robertson 1966).

5 Pickrell et al. considered the situation of multiple waves of admixture from
6 different source populations and showed that LD was comprised of multiple
7 exponential terms, each of which refers to a single admixture event(Pickrell *et*
8 *al.* 2014). However, the confounding effect was not fully taken into account,
9 especially the effects of source populations on the LD of an admixed
10 population. In this study, an accurate mathematical expression of LD is given
11 under a general admixture model(Verdu and Rosenberg 2011) (Figure 1). Our
12 analysis showed that the effect of the LD of source populations constitutes the
13 major part of the LD in an admixed population (Table 1), which should not be
14 ignored in the case of modeling or estimating ALD from empirical data.

15 Based on our mathematical description of the weighted LD, Eq 14, we
16 proposed a new approach to separate the confounding effects of LD of source
17 populations under a two-way admixture model. This approach was
18 implemented in MALDmef, which was newly developed for this study. Unlike
19 previous methods of estimating the time of admixture based on weighted LD,
20 such as ALDER or MALDER, our method considers the influence of LD from
21 source populations and there is no need to set the starting distance d_0 . The
22 current method also involves a fast numerical method to fit the LD to
23 hundreds of exponential functions to cover the admixture events across a
24 range of hundreds of generations. Applying this method to the well-known
25 admixed populations in HGDP(Rosenberg *et al.* 2002) and HapMap(Altshuler

1 *et al.* 2010) data demonstrated that the current study have greatly facilitated
2 understanding of the admixture history of human populations.

3 **Materials and Methods**

4 **Data sets**

5 Data for simulation and empirical analysis were obtained from two public
6 resources: Human Genome Diversity Panel (HGDP)(Rosenberg *et al.* 2002)
7 and the International HapMap Project phase III (Altshuler *et al.* 2010). Data
8 filtering was performed within each population: Samples with missing rate >
9 5% per individual, SNPs with missing rate > 50% and SNPs failing in Hardy-
10 Weinberg Equilibrium test (p -value < $1.0E-6$) were permanently removed from
11 subsequent analysis. Source populations for simulations used the haplotypes
12 from 113 Utah residents with Northern and Western European ancestries from
13 the CEPH collection (CEU) and 113 Africans in Yoruba (YRI).

14 **LD for general admixture model**

15 A mathematical description of LD decay based on a general admixture model
16 is given here (Figure 1). The relevant notation is summarized in Table S1. To
17 produce the current model and render deduction more general and realistic,
18 the population was assumed to be finite in size, and the effect of natural
19 selection, if exists, is negligible. The details of math derivation can be found in
20 Supplementary Material. However, the math expression of LD under the
21 general assumption (Eq S10) is too complicated to be used in empirical
22 analysis. It was here assumed that all populations were of infinite population
23 size since the initial admixture. This assumption leads to a brief version of LD

1 and it would have only limited effect on admixture time inference in empirical
2 data analysis.

3 When the population size is infinite, the allele frequency on each site
4 should always be considered as constant in the source populations: so for the
5 i^{th} source population at the l^{th} wave admixture on the site x (Eq 14) and the
6 allele frequency difference, Eq 14, between two source populations i and j :

$$f_i^{(l)}(x) = f_i^{(1)}(x), \forall i > 0; \quad \text{Eq 1}$$

$$\delta_{ij}^{(l)}(x) = \delta_{ij}^{(1)}(x) = f_i^{(1)}(x) - f_j^{(1)}(x), i > 0, j > 0, l > 1. \quad \text{Eq 2}$$

7 The total genetic contribution from source population i in the admixed
8 population with n^{th} waves of admixture is as follows:

$$w_i^{(n)} = \sum_{l=1}^n m_i^{(l)} \left(\prod_{j=l+1}^n m_0^{(j)} \right). \quad \text{Eq 3}$$

9 In this way, the expectation of LD for the target admixed population is as
10 follows:

$$D_0^{(n+1)}(d) = \sum_{i=1}^K w_i^{(n)} D_i^{(n+1)}(d) + \sum_{l=1}^n a^{(l)}(d) \left(\prod_{j=l+1}^n m_0^{(j)} \right) \exp(-\sum_{j=l}^n g^{(j)} d). \quad \text{Eq 4}$$

11 If we set $g^{(l)} = 1$ and

$$b^{(l)}(d) = a^{(n+1-l)}(d) \left(\prod_{j=n-l+2}^n m_0^{(j)} \right), \quad \text{Eq 5}$$

12 then we have the expected LD value for the $(n+1)^{\text{th}}$ generation's admixed
13 population is as follows:

$$D_0^{(n+1)}(d) = \sum_{i=1}^K w_i^{(n)} D_i^{(n+1)}(d) + \sum_{l=1}^n b^{(l)}(d) \exp(-ld). \quad \text{Eq 6}$$

14 This equation tells us that the LD in admixed population is composed of
15 two parts: one is from its source populations, and the other is formed by the
16 admixture events.

1 LD for specific admixture models

2 The mathematical description of LD for a general admixture model is
3 provided above. Here, the LD of specific previously reported admixture
4 models are described(Chakraborty and Weiss 1988; Ewens and Spielman
5 1995; Pfaff *et al.* 2001; Guo and Fung 2006; Jin *et al.* 2012). It is shown that
6 these models can be regarded as special cases of our general model when
7 parameters are specified (Tables S2-6).

8 The two-way hybrid-isolation (HI) model is the most popular admixture
9 model and most admixture analyses are based on this model. Under this
10 model, only two source populations are assumed to be involved in any one
11 admixture event, and all the populations are isolated without further gene flow.
12 In the current mathematical framework, description of LD for two-way HI
13 model is as follows:

$$D_0^{(2)}(d) = m_1 D_1^{(2)}(d) + m_2 D_2^{(2)}(d) + m_1 m_2 E(\delta_{12}(x) \delta_{12}(y) | |x - y| = d) \exp(-g^{(1)} d). \quad \text{Eq 7}$$

14 Two typical continuous admixture models have also been frequently
15 discussed. These include the gradual admixture (GA) model(Ewens and
16 Spielman 1995; Guo and Fung 2006) and continuous gene flow (CGF)
17 model(Pfaff *et al.* 2001). Here, the two-way admixture of these models is
18 discussed, but LD for the multiple-way admixture models are also
19 incorporated into the current general admixture model. In these admixture
20 models, we set the number of generations between two admixtures to be 1,
21 $g^{(j)} = 1$ and the LD for the $(n+1)^{\text{th}}$ generation of the admixed population
22 (admixture began at the 1st generation) is given below.

23 Under the GA model,

$$\begin{aligned}
 D_0^{(n+1)}(d) &= m_1 D_1^{(n+1)}(d) + m_2 D_2^{(n+1)}(d) \\
 &+ m_1 m_2 E(\delta_{12}(x)\delta_{12}(y) || x - y| = d) \left(\frac{n-1}{n}\right)^{n-1} \exp(-nd) \\
 &+ m_1 m_2 E(\delta_{12}(x)\delta_{12}(y) || x - y| = d) \\
 &\quad \times \sum_{l=2}^n \frac{(n-1)^{n-l}}{n^{n-l+1}} \exp(-(n-l+1)d);
 \end{aligned}
 \tag{Eq 8}$$

1 Under the CGF model,

$$\begin{aligned}
 D_0^{(n+1)}(d) &= m_1 D_1^{(n+1)}(d) + m_2 D_2^{(n+1)}(d) \\
 &+ \alpha m_2 E(\delta_{12}(x)\delta_{12}(y) || x - y| = d) \\
 &\quad \times \sum_{l=1}^n (1 - \alpha)^{l-1} \exp(-(n-l+1)d), \alpha = 1 - m_2 \frac{1}{n}.
 \end{aligned}
 \tag{Eq 9}$$

2 The GA model describes a scenario in which both source populations
 3 continuously contributed genetic materials to the admixed population after the
 4 admixture event, and the admixture proportion remained constant throughout
 5 all of the generations. However, the CGF model only allows one of the two
 6 source populations to continuously contribute genetic materials to the
 7 admixed population.

8 The two-wave admixture model describes that the admixed population is
 9 formed by two waves of admixture, one ancient and one recent. The ancient
 10 admixture event produces the admixed population and the recent admixture
 11 causes new migration in the admixed population. The LD (Eq S1) in the
 12 admixed population is as follows:

$$\begin{aligned}
 D_0^{(3)}(d) &= m_1 D_1^{(3)}(d) + m_2 D_2^{(3)}(d) \\
 &+ m_1^{(1)} m_2^{(1)} m_0^{(2)} E(\delta_{12}(x)\delta_{12}(y) || x - y| = d) \\
 &\quad \times \exp(-(g^{(1)} + g^{(2)})d) \\
 &+ (m_1^{(2)} m_2^{(2)} + m_0^{(2)} m_1^{(2)} (m_2)^2 + m_0^{(2)} m_2^{(2)} (m_1)^2)
 \end{aligned}
 \tag{Eq 10}$$

$$\times E(\delta_{12}(x)\delta_{12}(y)||x - y| = d) \exp(-g^{(2)}d).$$

1 In a more complicated scenario, the genetic materials of the admixed
2 population are inherited from more than two populations. One general
3 scenario could be that two of the source populations (populations 1 and 2)
4 meet first and then the third source population (population 3) joins the
5 admixed population during the second wave. This phenomenon is here called
6 the three-way-two-wave model. The LD in the admixed population under this
7 model is as follows:

$$\begin{aligned} D_0^{(3)}(d) &= m_1 D_1^{(3)}(d) + m_2 D_2^{(3)}(d) + m_3 D_3^{(3)}(d) \\ &+ m_1^{(1)} m_2^{(1)} m_0^{(2)} E(\delta_{12}(x)\delta_{12}(y)||x - y| = d) \exp(-(g^{(1)} + g^{(2)})d) \quad \text{Eq 11} \\ &+ m_3^{(2)} m_0^{(2)} \sum_{1 \leq i, j \leq 2} m_i m_j E(\delta_{i3}(x)\delta_{j3}(y)||x - y| = d) \exp(-g^{(2)}d). \end{aligned}$$

8 In order to provide an intuitive understanding of the dynamic decay of LD
9 among these models, they were plotted with specific parameters (Figure S1).

10 **Modeling LD with consideration of influence of ancestral source** 11 **populations**

12 In order to demonstrate how the LD in admixed population is influenced by
13 its ancestral source populations, we take the two-way HI model was here
14 used as an example to illustrate the effect of the LD of source populations on
15 the LD in the admixed population. Because, in a two-way HI model, no
16 recombination occurs in the population immediately after admixture, the LD
17 (Eq S2) in the first generation admixed population can be expressed as
18 follows.

$$D_{B^{(1)}}(x, y) = m_1^{(1)} D_1^{(1)}(x, y) + m_2^{(1)} D_2^{(1)}(x, y) + m_1^{(1)} m_2^{(1)} \delta_{12}(x)\delta_{12}(y). \quad \text{Eq 12}$$

1 Here, the absolute ratio, r_{sa} , to measure the effect of the LD in the source
2 populations, is defined as follows:

$$r_{sa}(x, y) = \left| \frac{m_1^{(1)}D_1^{(1)}(x, y) + m_2^{(1)}D_2^{(1)}(x, y)}{D_{B(1)}(x, y)} \right|. \quad \text{Eq 13}$$

3 Here, the admixture between Europeans and Africans was considered.
4 The scenario started with 100 individuals of CEU and 100 individuals of YRI
5 as the source populations of Europeans and Africans. They were simply put
6 together and treated as a population immediately after admixture. Pairwise LD
7 was calculated using the sampled SNPs on chromosome 1. The SNPs were
8 sampled in two rounds. In the first round, one SNP was sampled for every
9 twenty SNPs. In the second round, SNPs on the right side of every SNP
10 sampled in the first round (with greater physical position numbers) as follows:
11 All the SNPs whose genetics distance from the chosen SNP were below 0.05
12 Morgans were sampled; one in every three SNPs in the region whose
13 genetics distance from the chosen SNP was between 0.05 and 0.1 Morgans
14 were samples; and one in every five SNPs in the region whose genetic
15 distance from the chosen SNP was between 0.1 and 0.5 Morgans were
16 sampled. LD was calculated only between the SNPs sampled in the first
17 round and the related SNPs sampled in the second round. The r_{sa} was
18 calculated with raw LD of YRI, CEU and the manufactured admixed
19 population and these were classified into bins according to the genetic
20 distance between each pair of SNPs. Results showed the median value of r_{sa}
21 to be around 0.45 for all bin intervals (Table 1). The LD of source populations
22 made up about 45% of the LD in source populations. However, recent work
23 with the LD provide insight into population admixture history, even though the
24 effect of LD from the source populations was ignored(Loh *et al.* 2013).

1 Actually, using the average value of weighted LD can reduce the effect from
2 source populations. To verify this conclusion, we first calculated r_{sa} on the
3 average value of LD (for every 100 pairs of SNPs) in each bin (classified
4 according to the genetic distance between each pair of SNPs) and then
5 calculated r_{sa} on the average value of weighted LD in each bin. Results
6 showed LD from the source populations to make up about 33% of the average
7 LD and only about 2.8% of the average weighted LD in the admixed
8 population, and this proportion did not deduce to be negligible as the genetic
9 distance increasing (Table 1). In summary, this analysis indicates that
10 weighted LD is more proper to be used to infer admixture because it contains
11 less effect of confounding LD. However, using a starting distance with
12 weighted LD might not be the best way to reduce the confounding effect of
13 source populations.

14 **Weighted LD in a two-way admixed population**

15 The effective population size of the admixed population was assumed to
16 be large enough to be regarded as infinity, the LD can be described by Eq 6.
17 The effect of admixture events could be formalized with the sum of bunches of
18 exponential functions. The coefficient of exponential functions, Eq 6, if positive,
19 indicates the admixture event t generations ago. This can be used to estimate
20 the time of admixture. Moreover, it gives us an opportunity to infer the history
21 of admixture without any prior assumption of particular admixture models
22 (such as HI, GA, or CGF). In this way, it is important to study the exponential
23 property of the LD decay in admixed populations. If the ancestral source
24 populations are known for the two-way admixed population, the LD is as
25 follows:

$$D(d) = \sum_{i=1}^2 m_i D_i(d) + \sum_{l=1}^n b^{(l)}(d) \exp(-ld), \quad \text{Eq 14}$$

1 Here, $D(d)$ is the LD of the admixed population, m_i is the genetic
 2 contribution from source population i to the admixed population and $D_i(d)$ is
 3 the present LD of the ancestral source population i .

4 To determine the time of admixture actually using the exponential term
 5 $\exp(-ld)$, two important things must be taken into consideration. First, the
 6 coefficient of the exponential function, $b^{(l)}(d)$, must be relatively constant and
 7 bigger than 0; second, the effect of the LD of source populations can be
 8 separated. This confirmed that using weight, and the difference in allele
 9 frequencies between two source populations, could leave the coefficients of
 10 the exponential functions constantly bigger than zero (Appendix) and it may
 11 be possible to deduce the effect from source populations (Table 1). The
 12 weighted LD is examines in further detail below. The formula of weighted LD
 13 is as follows:

$$\hat{D}(d) = \sum_{i=1}^2 m_i \hat{D}_i(d) + \sum_{l=1}^n \hat{b}^{(l)}(d) \exp(-ld) \quad \text{Eq 15}$$

14 Here,

$$\begin{aligned} \hat{D}(d) &= E \left((D(x, y) \delta_{12}(x) \delta_{12}(y)) \middle| |x - y| = d \right), \\ \hat{D}_i(d) &= E \left((D_i(x, y) \delta_{12}(x) \delta_{12}(y)) \middle| |x - y| = d \right), \\ \hat{b}^{(l)}(d) &= c^{(l)} \times E \left((\delta_{12}(x) \delta_{12}(y))^2 \middle| |x - y| = d \right), \end{aligned} \quad \text{Eq 16}$$

15 and

$$\begin{aligned} c^{(l)} &= m_1^{(n+1-l)} m_2^{(n+1-l)} \\ &\quad + m_0^{(n+1-l)} m_2^{(n+1-l)} w_1^{(n+1-l)} w_1^{(n+1-l)} \end{aligned} \quad \text{Eq 17}$$

$$+m_0^{(n+1-l)}m_1^{(n+1-l)}w_2^{(n+1-l)}w_2^{(n+1-l)} .$$

1 Here, $c^{(l)}$ is determined by the admixture history and it is a natural
2 indicator for admixture events.

3 **Factorizing of weighted LD with exponential functions**

4 In order to show the exponential properties of the weighted LD, the
5 admixed population was simulated using forward time simulation with
6 haplotype data of YRI and CEU. A 100-generation old admixed population
7 with 50%:50% proportion was constructed. Ancestral source populations
8 based on the haplotype data of YRI and CEU were also constructed in the
9 simulation, separately. Next, $\widehat{D}(d)$, $\widehat{D}_i(d)$, and $E\left(\left(\delta_{12}(x)\delta_{12}(y)\right)^2\middle| |x-y|=d\right)$
10 were calculated based on the genotype data. The LD decay was fit with
11 hundreds of exponential functions. In this way, the coefficient spectrum on
12 exponential functions was determined and used to describe the decay of
13 weighted LD. The fitting method was able to model the decay of weighted LD
14 well and to provide the amplitude in every exponential function. The results of
15 fitting $\widehat{D}(d)$ primarily showed three bunches of exponential functions
16 composing the LD decay, with l values around 100, 180, and 1,250
17 generations (Figure 2A). As mentioned above, the l value corresponds directly
18 to the time of admixture, so the fact that the signal was around 100 could be
19 explained easily by the admixture in that the designed admixture is 100
20 generations ago. $\widehat{D}_i(d)$ and $E\left(\left(\delta_{12}(x)\delta_{12}(y)\right)^2\middle| |x-y|=d\right)$ were also fit, and
21 both $\widehat{D}_1(d)$ and $\widehat{D}_2(d)$ showed signals on l around the value 1,250 (Figures
22 S2–3). No significant signal peak was observed with

1 $E\left((\delta_{12}(x)\delta_{12}(y))^2 \mid |x-y|=d\right)$, even though there was a sharp decay over a
 2 very short distance (Figure S4). Based on these fitting results, it is here
 3 speculated that the signals in $\widehat{D}(d)$ around 1,250 and 180 might have resulted
 4 from confounding LD of the ancestral populations. To test this hypothesis,
 5 $Z(d)$, the optimized ALD, was fit so that LD of source populations was
 6 separated from the admixed population as the follows:

$$Z(d) = \frac{\widehat{D}(d) - \sum_{i=1}^2 m_i \widehat{D}_i(d)}{E\left((\delta_{12}(x)\delta_{12}(y))^2 \mid |x-y|=d\right)} = \sum_{l=1}^n c^{(l)} \exp(-ld). \quad \text{Eq 18}$$

7 The fitting results of $Z(d)$ showed only signal peak around the 100 of l svalue
 8 (Figure 2B). This supported the present hypothesis.

9 In summary, the weighted LD of source populations may affect the
 10 exponential properties of weighted LD of admixed population for relatively
 11 large l values, which may be the cause of signals of ancient admixture when
 12 employing a LD-based method for time estimation. However, we can use the
 13 derived LD of source populations to reduce the fake admixture signals.

14 **Calculation of weighted LD and fitting LD decay**

15 The basic algorithm used to calculate weighted LD was the same as that
 16 used with ALDER and MALDER. It was coded in C++. The weighted LD for
 17 admixed population was calculated as follows:

$$D(d) = \frac{\sum_{S(d)} c \widehat{cov}(x,y) \delta_{12}(x) \delta_{12}(y)}{|S(d)|}. \quad \text{Eq 19}$$

18 Here, $\widehat{cov}(x,y)$ is a non-bias estimator of the covariance on the genotype
 19 between two SNPs at x and y . $S(d)$ is the set holding pairs of SNPs with inter-
 20 site distance of d Morgan. $\delta_{12}(\ast)$ is the allele frequency difference, defined in
 21 Eq S3. The weighted LD for the source populations are calculated as follows:

$$D_i(d) = \frac{\sum_{S(d)} c \delta v(x,y) \delta_{0i}(x) \delta_{0i}(y)}{|S(d)|}, i = 1,2; \quad \text{Eq 20}$$

1 Here, '0' represents admixed population.

$$E \left((\delta_{12}(x) \delta_{12}(y))^2 \middle| |x - y| = d \right) = \frac{\sum_{S(d)} (\delta_{12}(x) \delta_{12}(y))^2}{|S(d)|}. \quad \text{Eq 21}$$

2 In this way, the algorithm and calculations relied only on genotype data,
 3 which prevented the introduction of phasing errors. The fast Fourier transform
 4 algorithm was also used to increase the computational efficiency. The
 5 population admixture proportions were estimated using the following formula.

$$m_2 = \frac{\sum_x \delta_{01}(x) \delta_{21}(x)}{\sum_x (\delta_{21}(x))^2}, m_1 = \frac{\sum_x \delta_{02}(x) \delta_{12}(x)}{\sum_x (\delta_{12}(x))^2}. \quad \text{Eq 22}$$

6 Once it is possible to calculate $Z(d)$, we can fit it using a numerical routine
 7 known as the proximal gradient (Beck A. and Teboulle *et al.* 2009). The object
 8 function to minimize is as follows:

$$\min_{c_j \geq 0 \text{ and } \sum c_j < c} \|Z - Ac\|_2,$$

9 where $Z = (z(d_1), z(d_2), \dots, z(d_{s-1}), z(d_s))^T$ is a vector of $Z(d)$ with
 10 different d values. These can be obtained from genetic data.
 11 $c = (c_0, c_1, \dots, c_{n-1}, c_n)^T$ is the coefficient of the exponential functions. The
 12 (i,j) th entry of the matrix $A_{s \times (n+1)}$ is $A_{ij} = \exp(-d_i G_j)$, where $\{G_j\}_{j=1, \dots, n+1}$ is the
 13 chosen subset of data in the generations from 0 to G . With this method, it was
 14 possible to find the positive values in the vector c .

15 **Determination of the significance of admixture signals and denoising**

16 The significance of admixture signal c was measured using a Jackknife-
 17 based approach. For the target population, each chromosome was excluded
 18 one at a time and the value of $Z(d)$ was calculated using the remaining
 19 chromosomes. After fitting $Z(d)$ with the sum of exponential functions,

1 denoising was performed on the coefficients of exponential functions. Only the
2 top signals that composed 99.9% of $Z(d)$ were retained. For the coefficient of
3 each exponential function, 22 observed values were tested to determine
4 whether they were larger than 0. In this way, P-values were used to measure
5 the significance of the admixture signal. In the spectrum plot of coefficients,
6 the mean of the 22 coefficients with P-values smaller than 0.05 were plotted
7 for each exponential function. This was specified using the candidate
8 admixture time points. (Figure 2 and Figure 5)

9 **Simulations**

10 In order to evaluate the performance of our method for estimating the time
11 of admixture, forward-time simulation was used to generate haplotypes of
12 admixed populations under different admixture models and different
13 scenarios: HI model, two-wave model (including the situation of one source
14 population and the situation of two source populations which contribute
15 genetic materials to the admixed population in the second wave), and
16 continuous gene flow model were used in a combination of GA and isolation
17 models and of CGF and isolation models. In our simulation, the newly
18 generated haplotypes are assembled with the segments of haplotypes in
19 source populations(Li and Stephens 2003; Price *et al.* 2009), which in this
20 particular case come from the haplotypes of CEU, and YRI.

21 In the HI model, the admixture event was set as having occurred 100
22 generations earlier. In the two-wave model, the first admixture event was set
23 as having occurred 100 generations earlier, after which the admixed
24 population was isolated for the next 80 generations, until the second wave of
25 admixture, after which the admixed population was isolated another 20

1 generations. In the second admixture event, a scenario in which only one of
2 the source populations donated genetic materials (TW-CGF model) and
3 another scenario where both source populations provided gene flows (TW-GA
4 model) were simulated.

5 Scenarios with continuous migration were also simulated. In one scenario,
6 the migration window was set to 80 generations: Continuous gene flow was
7 simulated from one of the source populations was simulated for 80
8 generations and then isolated for the next 20 generations (CGF-I model), and
9 continuous gene flow was simulated from both of the two-source populations
10 for 80 generations and then the admixed population was isolated for 20
11 generations (GA-I model); in the other scenario, the migration window lasted
12 30 generations: 30 generations' continuous gene flow and 70 generations'
13 isolation were used in both the CGF-I model and GA-I model. The evolution of
14 the source populations was also simulated under random mating with sample
15 size of 5,000 for 100 generations. The parameter details are given in Table
16 S7–8.

17 **Results**

18 **Robustness of the new method when used on proxy source populations**

19 A way of estimating the time of admixture was developed by separating
20 the LD from source populations directly from that of the admixed population,
21 which requested us to know the true ancestral source populations. However,
22 identifying the true source populations is also a complicated and difficult
23 problem. In most cases, only limited populations are available. In the
24 appendix, it is shown that using the proxy populations similar to the true

1 source populations can show the exponential properties of weighted LD's
2 decay well. It is here claimed that using proxy populations can also reduce the
3 confounding effect attributable to source populations' LD. Here, YRI and CEU
4 served as source populations from Africa and Europe, respectively, and 100
5 admixed individuals were simulated using 100 generations' admixture. This
6 method was able to show the admixture time with the true source populations
7 and pairs of source populations representing different parts of Africa and
8 Europe, i.e. CEU-LWK, CEU-MKK, TSI-LWK, TSI-MKK, and TSI-YRI, very
9 well (Figure S5–9).

10 **Robustness of the new method in various admixture models**

11 This method also involved using different admixture models. In order to
12 render the results of the evaluation reliable, 10 independent admixed
13 populations with haplotypes obtained from 113 unrelated CEU individuals and
14 113 unrelated YRI individuals were simulated with data from all of 22
15 autosomes. When the weighted LD was calculated, 100 individuals were
16 sampled from current source populations (isolated for 100 generations with
17 random mating) and admixed population separately. Under various admixture
18 models, MALDmef was able to reconstruct the history of the admixture
19 population well. For the one-pulse and two-pulse admixture models,
20 MALDmef gave the time close to the true time of admixture for the continuous
21 migration models, MALDmef was able to place most of the signals in a
22 particular migration time interval (Figure 3).

23 MALDER was also run on the same simulation data. MALDER is the only
24 software that can deal with multiple-wave admixture. CHB and CHD were
25 selected as the extra source populations and the starting distance was set to

1 0.005 Morgans and the bin size was set to 0.0002 Morgans. Under the HI
2 model, our method revealed significant signals around the generation (100)
3 we set for simulation (Figure S10), while MALDER gave us 8 significant
4 signals of 10 independent simulations around 100 generations. Here, 5 of the
5 8 signals were accompanied by significant signals above 250 generations,
6 which could be caused by the LD from the source populations (Figure S11).
7 For the multiple-wave admixture, we defined an estimation deviation (ED) to
8 measure difference (distance) of the location between the true admixture
9 signals and the detected admixture signals, which is defined as follows:

$$ED_s = \min (|T_s - T_{true1}|, |T_s - T_{true2}|, \dots, |T_s - T_{truen}|). \quad \text{Eq 23}$$

10 Where $\{T_{truej}, j = 1, 2, 3, \dots, n\}$ is the set of true admixture time and $\{T_s\}$ is
11 the estimated times of admixture. In our simulations, the following was true:

$$12 \quad ED_s = \min(|T_s - 20|, |T_s - 100|).$$

13 The results under HI model can also be estimated as follows

$$14 \quad ED_s = |T_s - 100|.$$

15 Under repeating simulations for the other various admixture models, the
16 *ED* values of MALDmef were significantly smaller than the *ED* values based
17 on MALDER, indicating that the current method is more precise and stable
18 than MALDER (Figure 4). The details of estimation with MALDER and
19 MALDmef on repeating simulations are shown in Figure S10–27. The current
20 method was also used on the empirical admixed populations.

21 **Estimating admixture time using empirical data**

22 The current method was first applied to a few well-known admixed
23 populations as in analysis reported by Alder(Loh *et al.* 2013) using an
24 available public database(Rosenberg *et al.* 2002; Altshuler *et al.* 2010).

1 MALDmef can currently only deal with the two-way admixture when derived
2 source populations or with the populations similar to the true derived source
3 populations are available. However, the real admixture history could be much
4 more complicated than assumed and many factors may affect the results of
5 estimation. In order to interpret the time spectrum, three principles should be
6 followed:

7 (1) A signal with larger amplitude is more reliable than one with smaller
8 amplitude.

9 (2) A signal that remains on the time spectrum for longer than 250
10 generations indicates that the chosen source populations are probably not
11 similar enough to the real ancestral source populations or that the general
12 two-way admixture model does not fit the data well.

13 (3) A signal in the continuous time interval containing generation 1 may
14 reflect the substructure of the admixed population but not the admixture.

15 Based on these principles, MALDmef was first applied to the well-known
16 admixed populations: African American (57 ASW individuals from HapMap),
17 Mexican (86 MEX individuals from HapMap) and Uygur (10 Uygur individuals
18 from HGDP). MALDER was also used to analyze these admixed populations.

19 In our analysis with MALDmef, CEU ($n = 113$) and YRI ($n = 113$) were
20 chosen as the ancestral populations of ASW. CEU (64 individuals) and
21 American Indian (7 Colombians, 14 Karitiana, 21 Maya, 14 Pimas and 8
22 Suruis) were chosen as the ancestral populations of MEX. Han ($n = 34$) and
23 French ($n = 28$) were chosen as the ancestral populations of Uygur. The time
24 of admixture of ASW was found to be about 4 to 5 generations ago (100–125
25 years before present, assuming 25 years per generation) (Figure 5A). MEX

1 seems to experience two wave of admixture: ranging from 6 to 8 generations
2 (150–200 years) ago and from 22 to 24 generations (550–600 years) ago,
3 respectively (Figure 5B). Neither inferences on ASW nor those on MEX
4 showed any significant signals on the time earlier than 250 generations ago,
5 suggesting that the source populations chosen for ASW and MEX are similar
6 enough to the true derived ancestral populations. The Uygur population has
7 been reported to have much longer admixture history than ASW and MEX(Xu
8 and Jin 2008; Xu *et al.* 2008; Jin *et al.* 2012; Qin *et al.* 2015). It showed three
9 time intervals of admixture: 2 to 3, 27 to 30, and 182 to 195
10 (182,189,190,194,195) (generations before present) (Figure 5C). The most
11 significant signals lay in the interval of 27–30, suggesting that the major
12 admixture creating the current population happened around 575 to 750 years
13 ago, which is consistent with previous result on the recent admixture.
14 However, the signals on the time interval from 182 to 195 generations ago
15 may indicate the ancient admixture events. Because of the sample size
16 limitation of LD-based method, an accurate inference on the ancient wave of
17 admixture may be difficult to find for Uygur. Private screening of 92 fully
18 sequenced individuals of Uygur indicated two highly confident admixture
19 intervals: from 18 to 20 generations ago and from 83 to 95 generations ago,
20 supporting the conclusion that an ancient admixture event helped create the
21 modern Uygur.

22 Loh *et al.* speculated that there could have been multiple waves of
23 admixture in the history of MKK (Loh *et al.* 2013). In the current analysis, 113
24 YRI and 113 CEU individuals were used to represent the source populations
25 of the MKK ($n = 156$). Here 5 candidate intervals of admixture were identified,

1 having occurred: 13–16, 50–67, 107–139, 310–410, and 1190–1210
2 (generations before the present). Only three of them, 13–16, 50–67, and 107–
3 139, seemed to be admixture signals, indicating that the three admixture
4 events happened from 325–400, 1,250–1,675 and 2,675–2,475 years ago.
5 The signals from between 1190 and 1210 generations ago may also indicate
6 that the ancestral populations are not good enough to infer the time of
7 admixture (Figure S28)

8 Paralyzed analysis with MALDER was also conducted on these admixed
9 populations (Table 2). For each admixed population, first all the populations in
10 the full data were set as the reference populations to infer the admixture, and
11 then the pair of populations with the highest amplitude for each wave of
12 admixture were collected and used as the reference populations to re-run
13 MALDER. MALDER showed good consistency in two rounds of inference.
14 From the results with global populations as references, MALDER was able to
15 determine the pair of populations that fit each wave of admixture best. Then
16 we MALDmef was run with each pair of populations as the reference to the
17 corresponding admixed population (Table3). Results showed that MALDmef
18 was also very robust to the reference populations and it showed considerable
19 consistency in each wave of admixture. However, with different reference
20 populations suggested by MALDER, it detected signals from even earlier than
21 250 generations back. This result indicated that those populations suggested
22 by MALDER might not be the proper source populations to study the
23 admixture under a two-way model.

24 The MALDmef's results were comparable to MALDER's on the same
25 admixed populations. In recently admixed populations, such as ASW and

1 MEX, MALDmef and MALDER had similar results but MALDmef had shorter
2 time interval ranges. In the admixed population with related ancient admixture,
3 such as Uygur, MALDmef was more powerful and stable than MALDER in
4 detecting the ancient admixture. MALDmef was able to predict the possible
5 ancient admixture events that produced the modern Uygur population. With
6 MALDmef, this can be supported by large sample sizes and dense markers.
7 This is not the case with MALDER.

8 **Discussion**

9 A general admixture model was here used to demonstrate the dynamics of
10 LD in admixed populations. A method was developed based on this model
11 and applied to estimating the time of admixture of populations that has
12 undergone two-way admixture, and for the first time, the admixture time
13 spectrum was used to reveal numbers and times of admixture events having
14 occurred in a particular population. This method and results should shed new
15 light into the drawing of inferences in the population genetics of admixed
16 populations.

17 In this study, results confirmed that the extent of LD was composed of
18 multiple exponential curves as a function of genetic distance in a certain
19 admixed population formed by multiple waves of admixture from multiple
20 source populations. Moreover, the confounding effects of source populations
21 were demonstrated using mathematical description of the LD.

22 In previous studies, LD from source populations was usually assumed to
23 be negligible in admixture time inference (Patterson *et al.* 2012; Loh *et al.*
24 2013; Pickrell *et al.* 2014). However, it was here shown to make up a major

1 part of the LD in admixed population, but the composition in the average
2 values (in bins, determined by genetic distance) of weighted LD was rather
3 small. The results of the current study showed the average proportion to be
4 about 33% in the values of pairwise LD and the proportion to be about 2.8% in
5 the average values of weighted LD. However, both these values decreased
6 slightly as genetic distance increased. This indicated that using weighted LD
7 could really help reduce the confounding effect of source populations, but
8 choosing a starting distance might not be the best way to reduce the
9 confounding effect of source populations.

10 Based on these considerations, the new method we developed in this
11 study was able to infer the time for multiple-wave admixture. In this method,
12 the reference populations were first used to reduce the confounding effect of
13 true. This method was very robust when the true ancestral source populations
14 were not available, and it worked well by using reference populations similar
15 to the source populations. The admixture-induced weighted LD extent curve
16 was fit with hundreds of exponential functions, which provided a signal
17 spectrum on time points ranging from 0–2,000 generations, which were within
18 the range of most possible admixture events after the migration of modern
19 humans “Out of Africa”. The jackknife method was used to produce a *P*-value
20 on each time point and then determined the time intervals for the possible
21 admixtures. Using various simulations, this method was demonstrated to be
22 more accurate in estimating admixture time especially in the scenario of two-
23 way admixture than other available methods.

24 This method, MALDmef, was used to analyze simulated data generated
25 with a continuous admixture model. Results could also be interpreted as

1 multiple-wave admixture instead of the continuous admixture model, here
2 simulated as the true model. This constitutes potential bias that could affect
3 the interpretation of the results. Considering this issue, we propose a
4 particular concept, namely **effective admixture**, i.e., continuous admixture
5 that can be treated as a few single-pulse hybrid events. This treatment should
6 at least make sense in the admixture modeling. The rationale is that the
7 inference of admixture should be still valid even with a model assuming a
8 scenario of hybrid-isolation, in a sense of critical parameters could still be
9 estimated effectively, but the admixture process could not be assessed
10 precisely. The method developed in this study, i.e., MALDmef, can be used to
11 infer the effective admixture events.

12 In the current analysis, it was observed that the weighted LD (one
13 reference) of source populations (no admixture events) could also fit
14 exponential functions closely with the major signals indicating very ancient
15 events, such as, 1,250 generations ago, but the mechanism has yet to be well
16 explained.

17 Even though the current method infers the time of multiple-wave admixture
18 under the two-way model, it still has many limitations regarding the
19 complicated demographic history of admixed populations. For example,
20 during time estimation with weighted LD, the effects of nature selection and
21 inbreeding were not considered here; neither was the situation of multiple-way
22 admixture. This was because the means by which the weight terms affect the
23 LD in multiple-way admixture have yet to be worked out. However, the
24 improved method developed in this study may facilitate the inference of
25 admixture history, particularly in determining multiple-wave population

- 1 admixture. Future work should focus on developing methods capable of
- 2 inferring multiple-way admixture.
- 3

1

Appendix

2 This appendix describes the properties of the coefficients of the
3 exponential terms with or without difference in allele frequency as weights.
4 The source populations are assumed to be infinite in size and to remain so
5 after the first admixture, so the unweighted coefficient is as follows:

$$b^{(l)}(d) = c^{(l)} E(\delta_{12}(x)\delta_{12}(y) | |x - y| = d) \quad \text{Eq A1}$$

6 and the weighted coefficient is as follows:

$$\hat{b}^{(l)}(d) = c^{(l)} E\left(\left(\delta_{12}(x)\delta_{12}(y)\right)^2 \middle| |x - y| = d\right) \quad \text{Eq A2}$$

7 $c^{(l)}$ is defined using Eq 17. Strictly speaking, both $b^{(l)}(d)$ and $\hat{b}^{(l)}(d)$ are
8 the functions of d , and they consist of two parts: The part in the first
9 parenthesis is determined by the admixture; the other is the conditional
10 expectation, determined by the allele frequency difference between the two
11 source populations. The part determined by the admixture is independent of
12 the genetic distance between two markers. It can be regarded as constant
13 during the extraction of time information from the exponential terms. The part
14 that matters is the conditional expectation. If the genetic distance d is large
15 enough that $\delta_{12}(x)$ is independent of $\delta_{12}(y)$, the value of
16 $E(\delta_{12}(x)\delta_{12}(y) | |x - y| = d)$ will be zero, but the value of
17 $E\left(\left(\delta_{12}(x)\delta_{12}(y)\right)^2 \middle| |x - y| = d\right)$ should be a certain constant that measures
18 the genetic distance between source populations 1 and 2. In this way, $\hat{b}^{(l)}(d)$
19 can help infer the exponential terms related to the admixture more effectively
20 than $b^{(l)}(d)$

1 Another problem addressed here is the property of $\hat{b}^{(l)}(d)$, when only get
2 the populations close to the true source populations were available to
3 calculate the weighted LD. Suppose population 3 is similar to population 1
4 and population 4 to population 2, they follow a genealogical relationship
5 described in Figure S29. The weight used here is as follows:

$$\delta_{34}(x)\delta_{34}(y). \quad \text{Eq A3}$$

6 Then

$$\hat{b}^{(l)}(d) = c^{(l)} E \left((\delta_{12}(x)\delta_{12}(y)\delta_{34}(x)\delta_{34}(y)) \middle| |x - y| = d \right) \quad \text{Eq A4}$$

7 Under the same assumption, the genetic distance d is considered large
8 enough that $\delta_{12}(x)$ and $\delta_{34}(x)$ are independent of $\delta_{12}(y)$ and $\delta_{34}(y)$.

$$\hat{b}^{(l)}(d) = c^{(l)} E \left((\delta_{ef}(x)\delta_{ef}(y))^2 \middle| |x - y| = d \right), \quad \text{Eq A5}$$

9 Here, e represents the shared ancestral population of 1 and 3 and f is the
10 shared ancestral population of 2 and 4. In situation in which the true source
11 populations are not available or cannot be identified, populations similar to the
12 source populations can be used to calculate weighted LD.

13 Software

14 C++ Source codes of MALDmef can be found at:

15 <http://www.picb.ac.cn/PGG/resource.php>

16 Acknowledgements

17 These studies were supported by the Strategic Priority Research Program of
18 the Chinese Academy of Sciences (CAS) (XDB13040100), by the National

1 Natural Science Foundation of China (NSFC) grants (91331204 and
2 31171218). S.X. is Max-Planck Independent Research Group Leader and
3 member of CAS Youth Innovation Promotion Association. S.X. also gratefully
4 acknowledges the support of the National Program for Top-notch Young
5 Innovative Talents of The "*Wanren Jihua*" Project. The funders had no role in
6 study design, data collection and analysis, decision to publish, or preparation
7 of the manuscript.

8

9

10

11

12

1

Reference

- 2 Altshuler D. M., Gibbs R. a, Peltonen L., Altshuler D. M., Gibbs R. a, Peltonen
3 L., Dermitzakis E., Schaffner S. F., Yu F., Peltonen L., Dermitzakis E.,
4 Bonnen P. E., Altshuler D. M., Gibbs R. a, Bakker P. I. W. de, Deloukas
5 P., Gabriel S. B., Gwilliam R., Hunt S., Inouye M., Jia X., Palotie A.,
6 Parkin M., Whittaker P., Yu F., Chang K., Hawes A., Lewis L. R., Ren Y.,
7 Wheeler D., Gibbs R. a, Muzny D. M., Barnes C., Darvishi K., Hurles M.,
8 Korn J. M., Kristiansson K., Lee C., McCarrol S. a, Nemesh J.,
9 Dermitzakis E., Keinan A., Montgomery S. B., Pollack S., Price A. L.,
10 Soranzo N., Bonnen P. E., Gibbs R. a, Gonzaga-Jauregui C., Keinan A.,
11 Price A. L., Yu F., Anttila V., Brodeur W., Daly M. J., Leslie S., McVean
12 G., Moutsianas L., Nguyen H., Schaffner S. F., Zhang Q., Ghorri M. J. R.,
13 McGinnis R., McLaren W., Pollack S., Price A. L., Schaffner S. F.,
14 Takeuchi F., Grossman S. R., Shlyakhter I., Hostetter E. B., Sabeti P. C.,
15 Adebamowo C. a, Foster M. W., Gordon D. R., Licinio J., Manca M. C.,
16 Marshall P. a, Matsuda I., Ngare D., Wang V. O., Reddy D., Rotimi C. N.,
17 Royal C. D., Sharp R. R., Zeng C., Brooks L. D., McEwen J. E., 2010
18 Integrating common and rare genetic variation in diverse human
19 populations. *Nature* **467**: 52–58.
- 20 Beck A. amd Tebouille M., Beck A., Tebouille M., 2009 A Fast Iterative
21 Shrinkage-Thresholding Algorithm for Linear Inverse Problems. {SIAM} J.
22 Imaging Sci. **2**: 183–202.
- 23 Chakraborty R., Weiss K. M., 1988 Admixture as a tool for finding linked
24 genes and detecting that difference from allelic association between loci.
25 *Proc. Natl. Acad. Sci. U. S. A.* **85**: 9119–9123.
- 26 Ewens W. J., Spielman R. S., 1995 The transmission/disequilibrium test:
27 history, subdivision, and admixture. *Am. J. Hum. Genet.* **57**: 455–464.
- 28 Guo W., Fung W.-K., 2006 The admixture linkage disequilibrium and genetic
29 linkage inference on the gradual admixture population. *Yi Chuan Xue Bao*
30 **33**: 12–8.
- 31 Hill W. G., Robertson a, 1966 The effect of linkage on limits to artificial
32 selection. *Genet. Res.* **8**: 269–294.
- 33 Jin W., Wang S., Wang H., Jin L., Xu S., 2012 Exploring population admixture
34 dynamics via empirical and simulated genome-wide distribution of
35 ancestral chromosomal segments. *Am. J. Hum. Genet.* **91**: 849–862.
- 36 Li N., Stephens M., 2003 Using Single-Nucleotide Polymorphism Data.
37 *Genetics* **2233**: 2213–2233.
- 38 Loh P. R., Lipson M., Patterson N., Moorjani P., Pickrell J. K., Reich D.,
39 Berger B., 2013 Inferring admixture histories of human populations using
40 linkage disequilibrium. *Genetics* **193**: 1233–1254.

- 1 McKeigue P. M., 2005 Prospects for admixture mapping of complex traits.
2 Am. J. Hum. Genet. **76**: 1–7.
- 3 Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y.,
4 Genschoreck T., Webster T., Reich D., 2012 Ancient admixture in human
5 history. *Genetics* **192**: 1065–1093.
- 6 Pfaff C. L., Parra E. J., Bonilla C., Hiester K., McKeigue P. M., Kamboh M. I.,
7 Hutchinson R. G., Ferrell R. E., Boerwinkle E., Shriver M. D., 2001
8 Population structure in admixed populations: effect of admixture
9 dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.*
10 **68**: 198–207.
- 11 Pickrell J. K., Patterson N., Loh P.-R., Lipson M., Berger B., Stoneking M.,
12 Pakendorf B., Reich D., 2014 Ancient west Eurasian ancestry in southern
13 and eastern Africa. *Proc. Natl. Acad. Sci. U. S. A.* **111**: 2632–7.
- 14 Price A. L., Tandon A., Patterson N., Barnes K. C., Rafaels N., Ruczinski I.,
15 Beaty T. H., Mathias R., Reich D., Myers S., 2009 Sensitive detection of
16 chromosomal segments of distinct ancestry in admixed populations.
17 *PLoS Genet.* **5**.
- 18 Qin P., Zhou Y., Lou H., Lu D., Yang X., Wang Y., Jin L., Chung Y.-J., Xu S.,
19 2015 Quantitating and Dating Recent Gene Flow between European and
20 East Asian Populations. *Sci. Rep.* **5**: 9500.
- 21 Reich D., Patterson N., 2005 Will admixture mapping work to find disease
22 genes? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **360**: 1605–1607.
- 23 Rosenberg N. a, Pritchard J. K., Weber J. L., Cann H. M., Kidd K. K.,
24 Zhivotovsky L. a, Feldman M. W., 2002 Genetic structure of human
25 populations. *Science* **298**: 2381–2385.
- 26 Seldin M. F., 2007 Admixture mapping as a tool in gene discovery. *Curr. Opin.*
27 *Genet. Dev.* **17**: 177–181.
- 28 Smith M. W., O'Brien S. J., 2005 Mapping by admixture linkage
29 disequilibrium: advances, limitations and guidelines. *Nat. Rev. Genet.* **6**:
30 623–632.
- 31 Verdu P., Rosenberg N. a., 2011 A general mechanistic model for admixture
32 histories of hybrid populations. *Genetics* **189**: 1413–1426.
- 33 Xu S., Huang W., Qian J., Jin L., 2008 Analysis of Genomic Admixture in
34 Uyghur and Its Implication in Mapping Strategy. *Am. J. Hum. Genet.* **82**:
35 883–894.
- 36 Xu S., Jin L., 2008 A Genome-wide Analysis of Admixture in Uyghurs and a
37 High-Density Admixture Map for Disease-Gene Discovery. *Am. J. Hum.*
38 *Genet.* **83**: 322–336.

1

2

1 **Figure legends**

2

3 Figure 1: A general admixture model with K source populations and n
4 waves of admixture. See Table S1 for notation.

5

6 Figure 2: Full exponential spectrum for fitting extending weighted LD in a
7 simulated admixed population. A) Exponential spectrum without separating
8 the confounding LD from the source populations. B) Exponential spectrum
9 with separating the confounding LD from the source populations.

10

11 Figure 3: Evaluation of the performance of MALDmef under various
12 admixture models. The blue vertical dash line represents the true simulated
13 admixture time. Sim-1 to Sim-4: one-pulse admixture of 200, 100, 50 and 20
14 generations ago. Sim-5 to Sim-8: two-pulse admixture at 20 and 100
15 generations ago. Sim-9 to Sim-12: continuous admixture from 20 to 100
16 generations ago (Sim-9 and Sim-10) and from 70 to 100 generations ago
17 (Sim-11 and Sim-12).

18

19 Figure 4: Performance of MALDmef and MALDER in various simulations.
20 One-side t-test was performed to calculate the significance of the differences
21 between ED values of MALDmef and MALDER. The detailed parameters for
22 each simulation are given in Table S7.

23

24 Figure 5: Full exponential spectrum for fitting extending weighted LD in
25 admixed populations. A) ASW; B) MEX; C) Uyghur.

26

27

1 **Table 1: Absolute ratio of LD from source population to the LD of the**
 2 **admixed population**

Bin interval (Morgan)	Raw LD		Average [#] Raw LD		Average Weighted LD	
	Prop ^{&}	Median [§]	Prop	Median	Prop	Median
[0,0.05)	39.4%	0.48	39.4%	0.36	39.4%	0.032
[0.05, 0.1)	12.4%	0.45	12.4%	0.33	12.4%	0.028
[0.1, 0.15)	7.1%	0.46	7.1%	0.33	7.1%	0.028
[0.15, 0.2)	6.7%	0.45	6.7%	0.33	6.7%	0.028
[0.2, 0.25)	6.3%	0.45	6.3%	0.33	6.3%	0.029
[0.25, 0.3)	6.1%	0.45	6.1%	0.33	6.1%	0.029
[0.3, 0.35)	5.8%	0.45	5.8%	0.33	5.8%	0.028
[0.35, 0.4)	5.6%	0.45	5.6%	0.33	5.6%	0.028
[0.4, 0.45)	5.4%	0.45	5.4%	0.33	5.4%	0.028
[0.45, 0.5)	5.3%	0.45	5.3%	0.33	5.3%	0.027

3 For each value of pairwise LD, there is a corresponding genetic distance between the
 4 pair of SNPs. For convenience, the LD value and the corresponding genetic distance
 5 are called as the LD point. According to the value of the genetic distance, the LD
 6 point can be placed in bins.

7 [&]: Proportion of the number of LD points in the bin to all the number of LD points.

8 [§]: Median value of the ratio measured using Eq 17 with the LD points in each bin.

9 [#]: Average values are calculated from raw or weighted LD. First, the LD points were
 10 sorted according to the genetic distance. Then, with every 100 sorted LD points, the
 11 averages of the LD value and genetic distance were taken as the LD value and
 12 genetic distance as the new LD points. At last, these new LD points were placed in
 13 the bins and the proportion and median were calculated.

14
 15

1 **Table 2 Time of admixture (generations) determined using MALDmf**
 2 **with selected reference populations.**

Admixed population	Reference populations	Time 1	Time 2	Time 3	Time 4	Time 5
ASW	CEU;YRI	4-5 2.02×10^{-13}				
MEX	CEU;American Indian	6-8 0.00506	22-24 0.00561			
MEX	American Indian; YRI	8-9 3.38×10^{-5}	32-38 0.0481	810-830 0.00129		
MKK	CEU;YRI	13-16 0.0205	50-67 0.0456	107-139 0.0499	310-410 0.0284	
MKK	TSI;YRI	17-19 0.00271	64-79 0.0337	115-160 0.0494	185-188 0.0481	590-670 0.0147
Uyghur	French;Han	2-3 1.82×10^{-6}	27-30 0.0448	182-195 0.0484		
Uyghur	Basque;Han	2-4 0.0385	29-31 0.00377	157-164 0.0427	690-750 0.0441	
Uyghur	Sardinian; Japanese	2-3 2.79×10^{-6}	32-35 0.0149	209-211 0.0421	710-790 0.0253	

3 In each time cell, the biggest P-value is provided under the detected time
 4 interval.

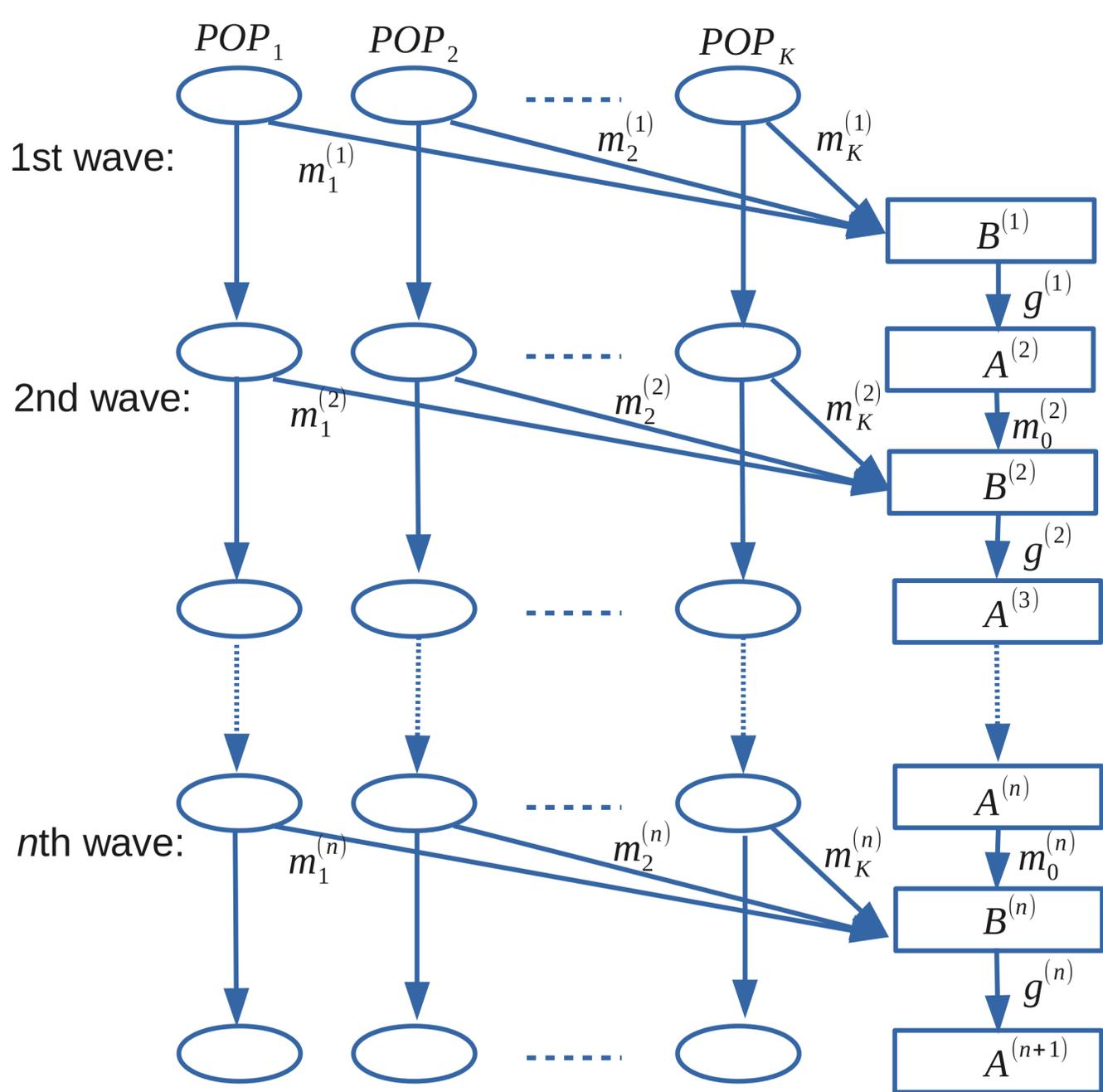
5
 6

1 **Table 3: Time of admixture (generations) determined using MALDER**
 2 **with global scanning and selected populations**

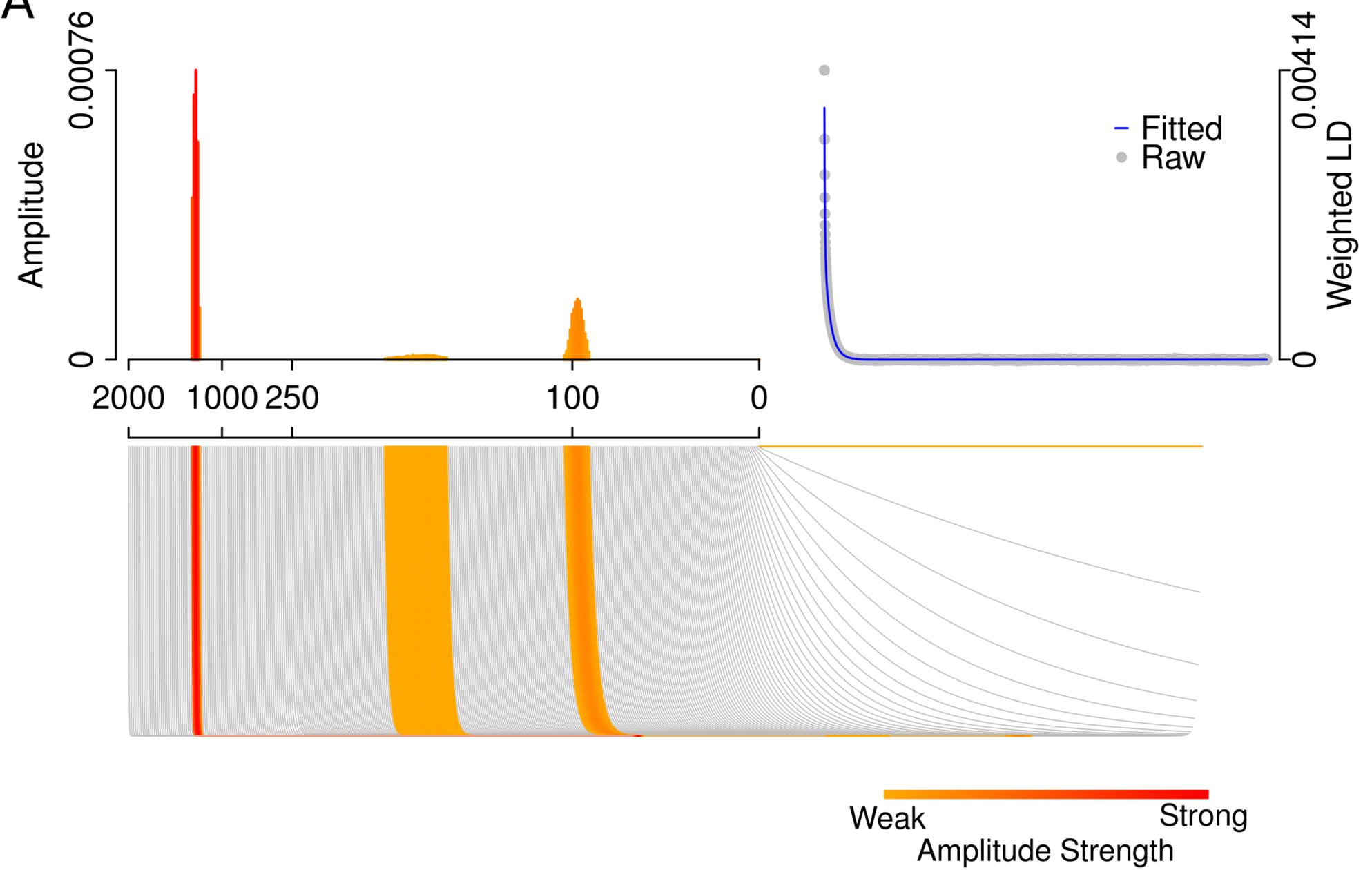
Admixed pop	Reference pop	Time 1	Time 2	Time 3	Time 4
ASW	HapMap	2.7 +/- 0.5 (Z=5.27) CEU;YRI	12.0 +/- 4.4 (Z=2.76) CEU;YRI		
ASW	CEU;TSI;YRI	1.3 +/- 2.8 (Z=0.47) CEU;YRI	6.1 +/- 3.2 (Z=1.92) CEU;YRI	70.3 +/- 60.7 (Z=1.16) TSI;YRI	
MEX	HapMap+ American Indian (HGDP)	5.2 +/- 1.9 (Z=2.78) Indian;TSI	19.8 +/- 9.7 (Z=2.05) Indian;YRI		
MEX	American Indian; TSI;YRI	4.2 +/- 1.6 (Z=2.67) Indian;TSI	15.1 +/- 4.0 (Z=3.82) Indian;YRI		
MKK	HapMap	3.3 +/- 0.7 (Z=4.51) TSI;YRI	23.7 +/- 6.4 (Z=3.69) CEU;YRI	107.8 +/- 32.7 (Z=3.29) TSI;YRI	495.2 +/- 255.1 (Z=1.94) TSI;YRI
MKK	YRI;CEU;TSI	3.1 +/- 0.8 (Z=3.87) TSI;YRI	20.3 +/- 6.4 (Z=3.19) YRI;CEU	80.3 +/- 28.3 (Z=2.84) TSI;YRI	292.7 +/- 121.5 (Z=2.41) TSI;YRI
Uyghur	HGDP	4.3 +/- 2.4 (Z=1.79) Basque; Han	40.9 +/- 7.4 (Z=5.52) Japanese; Sardinian		
Uyghur	Basque;Han; Japanese; Sardinian	4.6 +/- 2.6 (Z=2.12) Basque;Han	41.4 +/- 7.0 (Z=5.92) Japanese; Sardinian		

3 For each wave of admixture, we listed the pair of populations with the biggest
 4 amplitude value under the Z score.

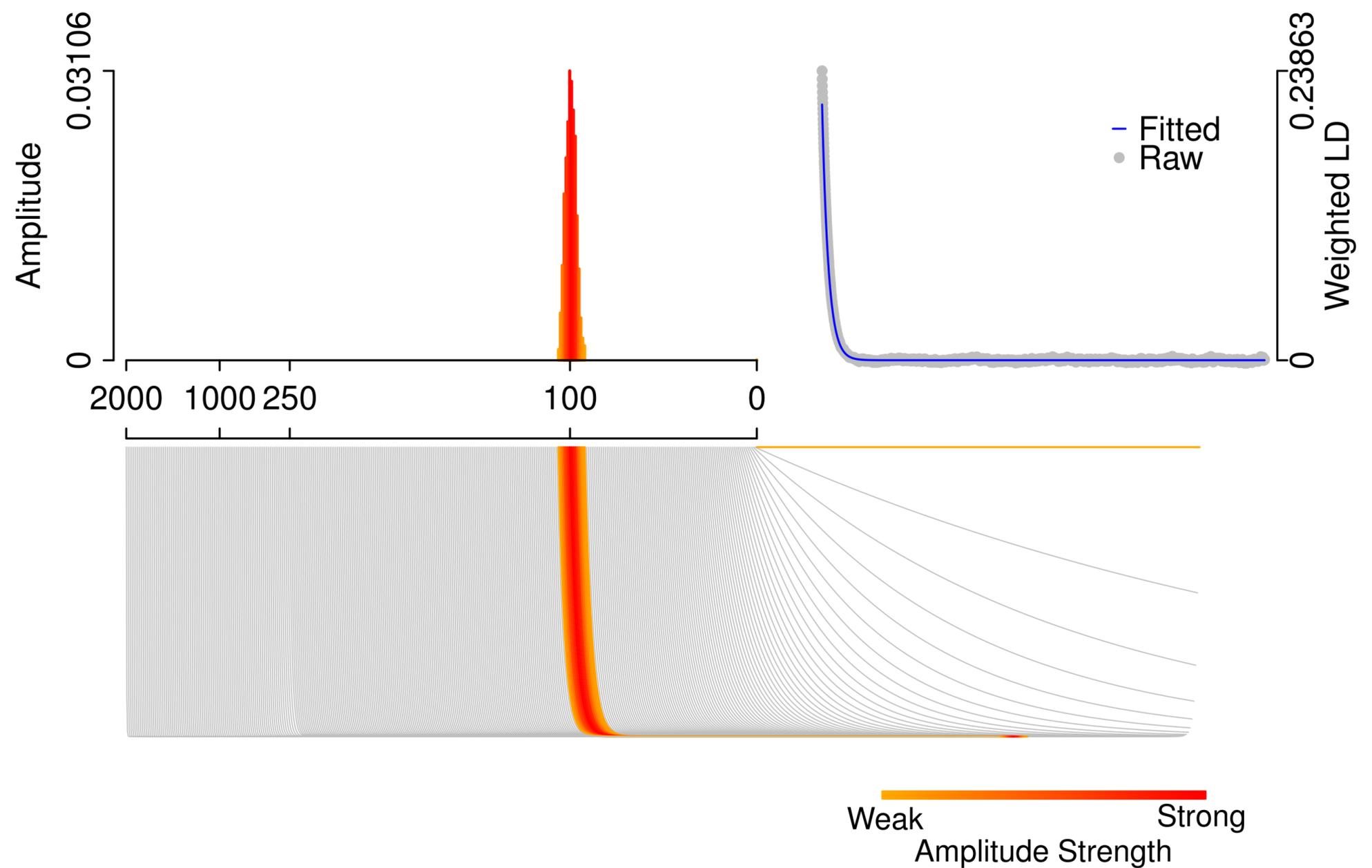
5
 6



Ancestral population
 Admixed population
 Gene flow

A

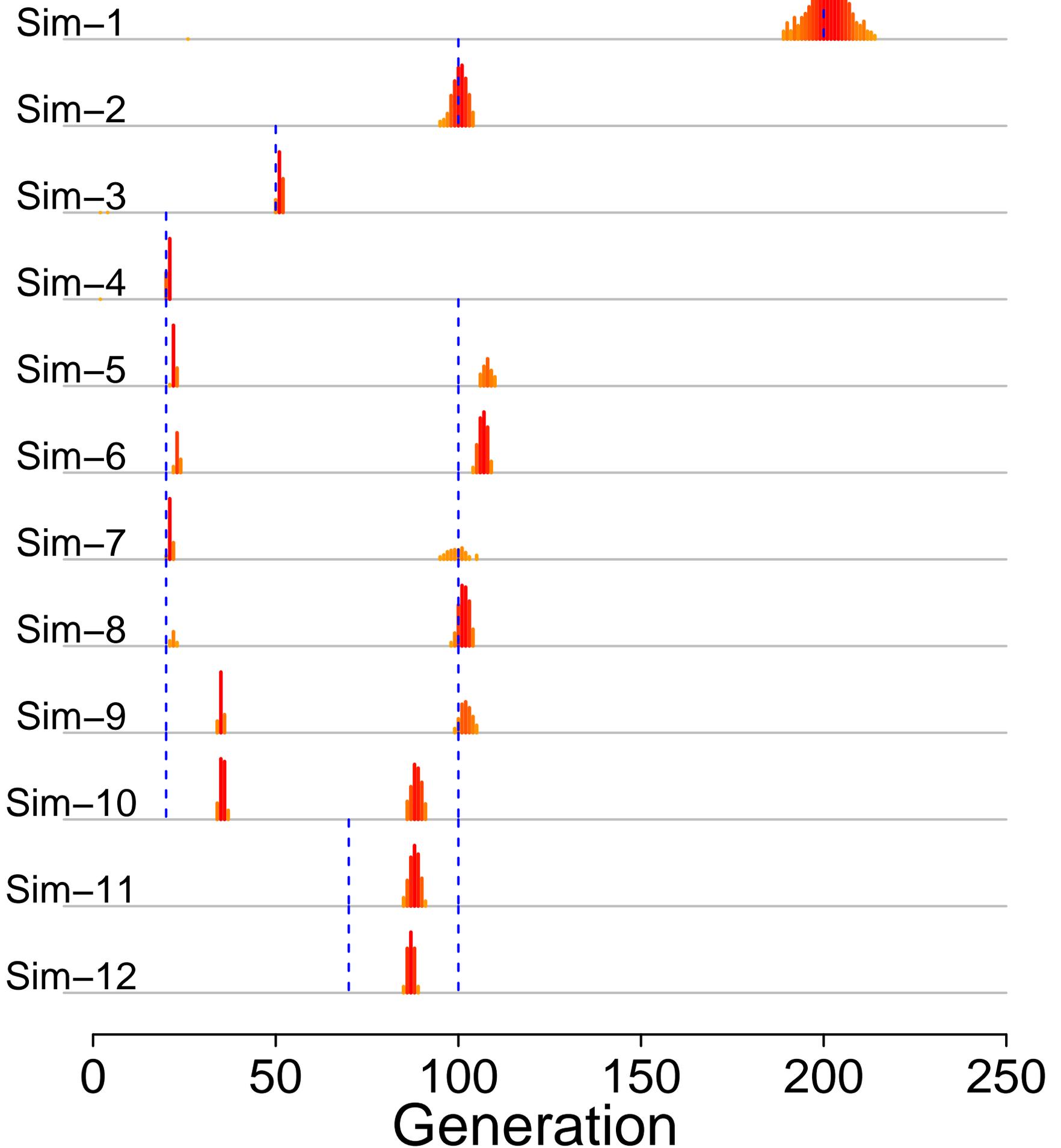
bioRxiv preprint doi: <https://doi.org/10.1101/026757>; this version posted September 14, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

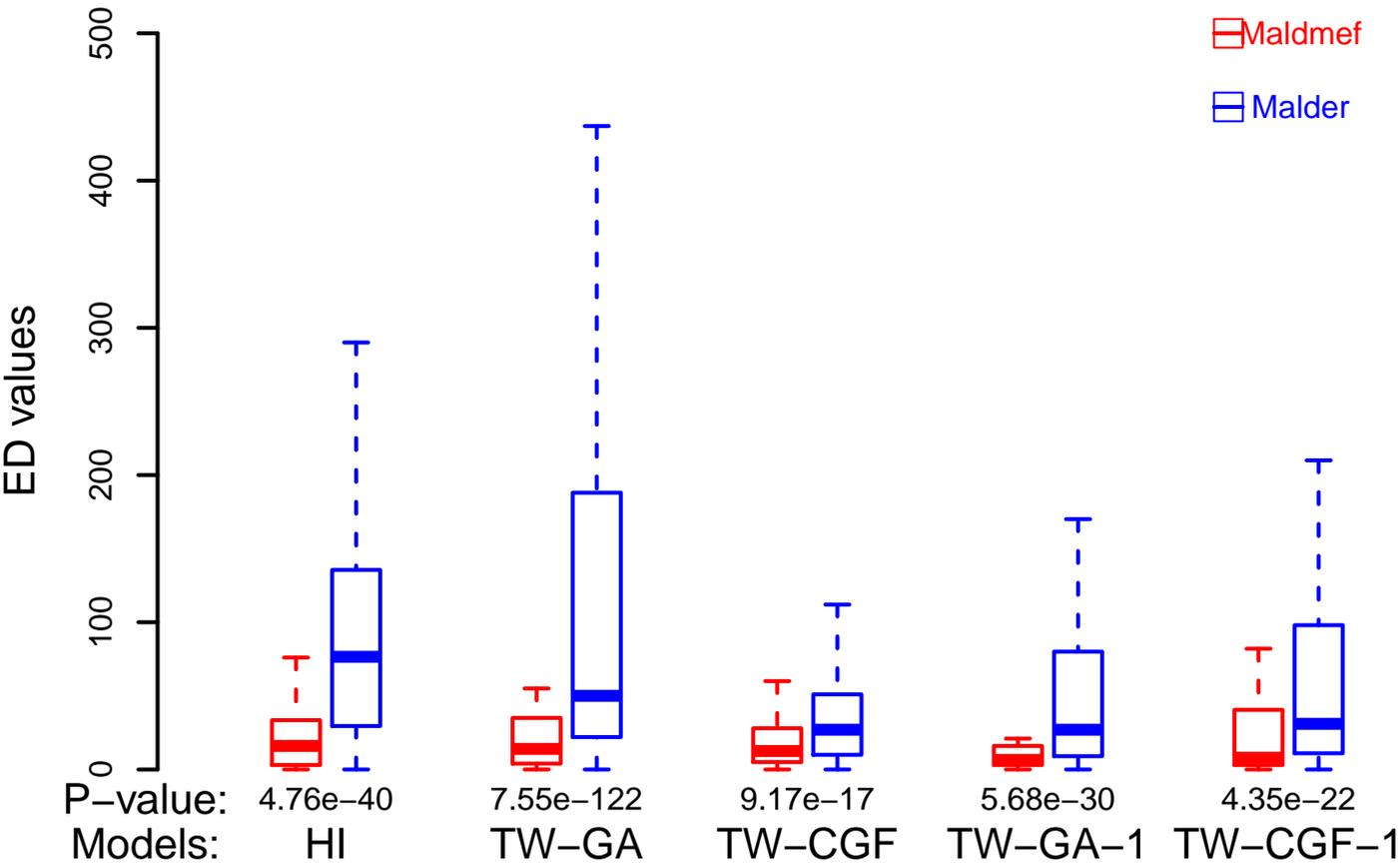
B

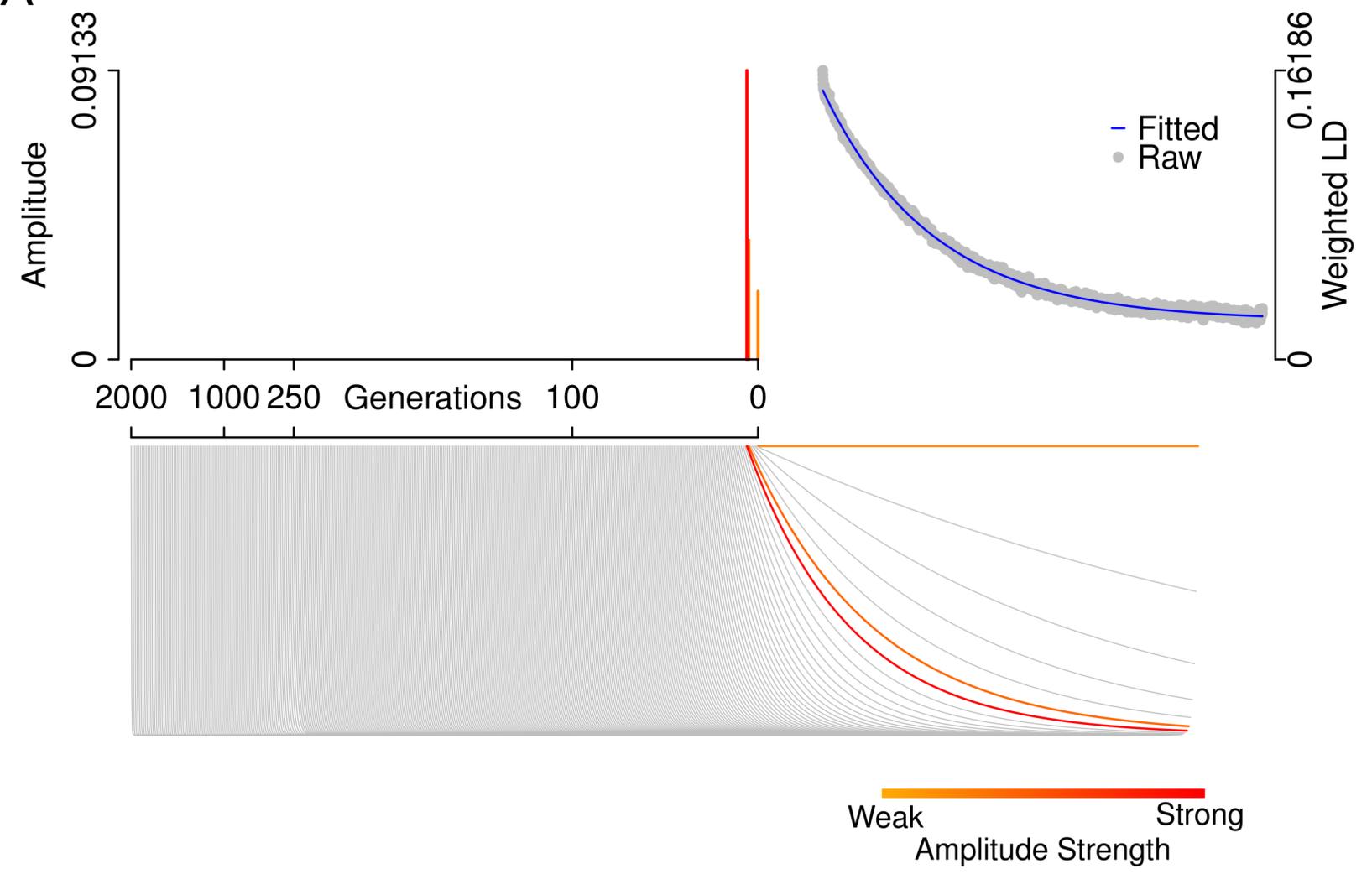
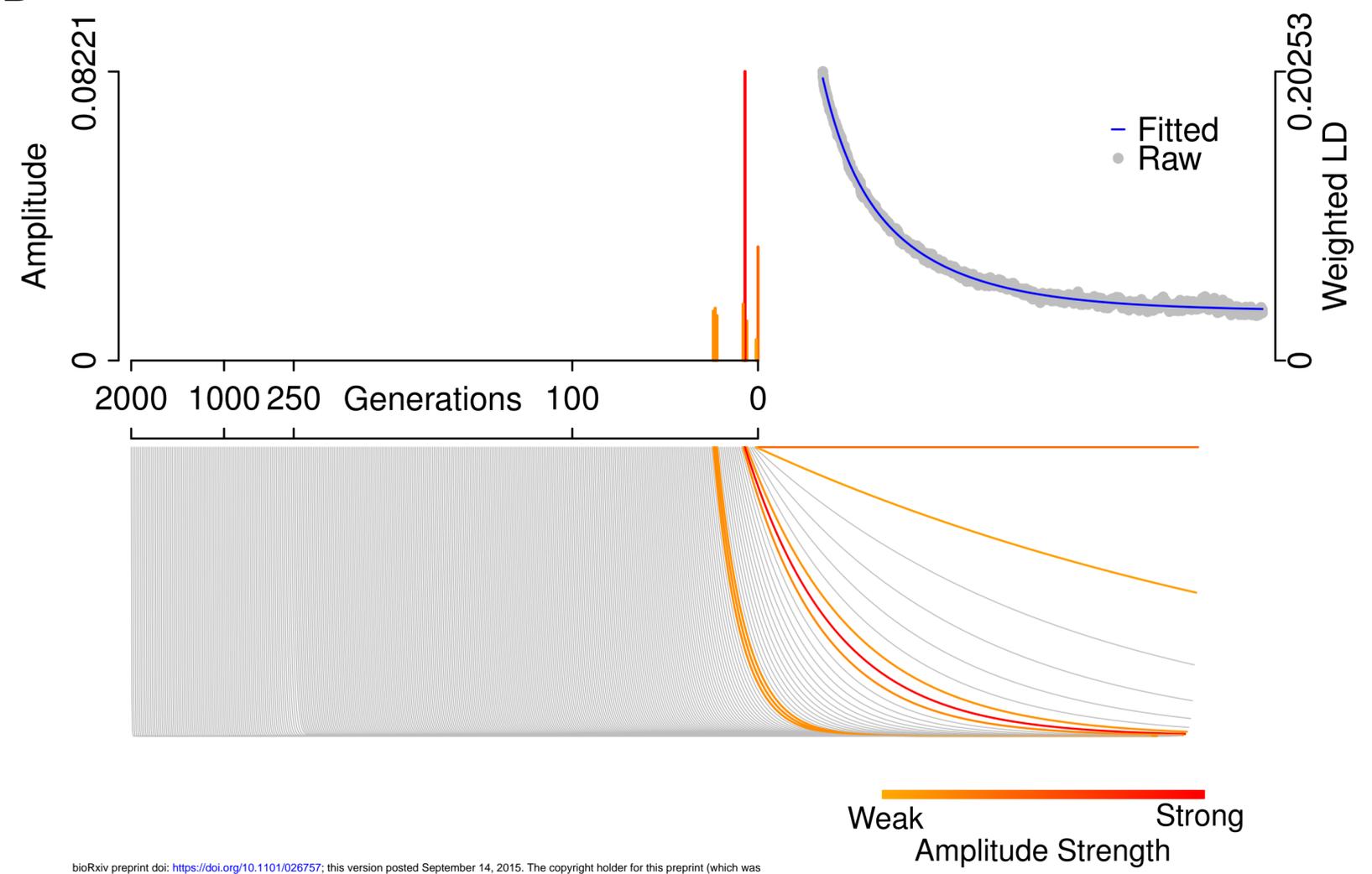
Amplitude Strength

Weak Strong

bioRxiv preprint doi: <https://doi.org/10.1101/026757>; this version posted September 14, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.





A**B**

bioRxiv preprint doi: <https://doi.org/10.1101/026757>; this version posted September 14, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

C