

1 Modeling Continuous Admixture

2 **Keywords:** Admixture-induced linkage disequilibrium; Continuous admixture;
3 Admixture model; Admixture inference; SNP

4

5 *Ying Zhou^{†, §}, Hongxiang Qiu^{†, ‡, §}, Shuhua Xu^{†, ††, ‡‡, *}*

6

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

11 [‡] Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong
12 Kong, China;

13 ^{††} School of Life Science and Technology, ShanghaiTech University, Shanghai
14 200031, China;

15 ^{‡‡} Collaborative Innovation Center of Genetics and Development, Shanghai
16 200438, China.

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

18 * Correspondence and requests for materials should be addressed to
19 xushua@picb.ac.cn (S.X.)

20

21

Abstract

Human migration and human isolation serve as the driving forces of modern human civilization. Recent migrations of long isolated populations have resulted in genetically admixed populations. The history of population admixture is generally complex; however, understanding the admixture process is critical to both evolutionary and medical studies. Here, we utilized admixture induced linkage disequilibrium (LD) to infer occurrence of continuous admixture events, which is common for most existing admixed populations. Unlike previous studies, we expanded the typical continuous admixture model to a more general admixture scenario with isolation after a certain duration of continuous gene flow. Based on the extended models, we developed a method based on weighted LD to infer the admixture history considering continuous and complex demographic process of gene flow between populations. We evaluated the performance of the method by computer simulation and applied our method to real data analysis of a few well-known admixed populations.

Introduction

Human migrations involve gene flow among previously isolated populations, resulting in the generations of admixed populations. In both evolutionary and medical studies of admixed populations, it is essential to understand admixture history and accurately estimate the time since population admixture because genetic architecture at both population and individual levels are determined by admixture history, especially the admixture time. However, the estimation of admixture time is largely dependent on the precision of the applied admixture models. Several methods have

1 been developed to estimate admixture time based on the Hybrid Isolation (HI) model
2 (Xu and Jin 2008; Price *et al.* 2009; Loh *et al.* 2013; Qin *et al.* 2015) or intermixture
3 admixture model (IA) (Zhu *et al.* 2004), which assumes that the admixed population
4 is formed by one wave of admixture at a certain time. However, the one-wave
5 assumption often leads to under-estimation when the progress of the true admixture
6 cannot be well modeled by the HI model. Jin *et al.* showed earlier that under the
7 assumption of HI, the estimated time is half of the true time when the true model is a
8 gradual admixture (GA) model (Jin *et al.* 2013).

9 Admixture models can be theoretically distinguished by comparing the length
10 distribution of continuous ancestral tracts (CAT) (Gravel 2012; Jin *et al.* 2012; Ni *et*
11 *al.* 2015), which refer to continuous haplotype tracts that were deviated from the same
12 ancestral population. CAT inherently represents admixture history as it accumulates
13 recombination events. Short CAT always indicates long admixture histories of the
14 same admixture proportion, whereas long CAT may indicate a recent gene flow from
15 the ancestral populations to which the CAT belongs. Based on the information it
16 provides, CAT can be used to distinguish different admixture models and estimate
17 corresponding admixture time. However, accurately estimating the length of CAT is
18 often very difficult.

19 Weighted linkage disequilibrium (LD) is an alternative tool that can be used to
20 infer admixture (Loh *et al.* 2013; Pickrell *et al.* 2014). Previous studies have indicated
21 that this tool is more efficient than CAT because it requires neither ancestry
22 information inference nor haplotype phasing, which often provides false
23 recombination information, thus decreasing the power of estimation. Weighted LD
24 has already been used in inferring multiple-wave admixtures (Pickrell *et al.* 2014;
25 Zhou *et al.* 2015) However, these methods tend to summarize the admixture into

1 different independent waves, even if the true admixture is continuous. In our previous
2 work (Zhou *et al.* 2015), we mathematically described weighted LD under different
3 continuous models, allowing us to determine admixture history using these models.

4 In the present study, we first developed a weighted LD-based method to infer
5 admixture with HI, GA, and continuous gene flow (CGF) models (Pfaff *et al.* 2001),
6 (Fig 1). Both GA and CGF models assume that gene flow is a continuous process.
7 Next, we extended the GA and CGF models to the GA-I and CGF-I models,
8 respectively (Fig 1), which model a scenario with a continuous gene flow duration
9 followed by a period of isolation to present. We applied our method to a number of
10 well-known admixed populations and provided information that would help better
11 understanding the admixture history of these populations.

12 **Material and Methods**

13 **Datasets**

14 Data for simulation and empirical analysis were obtained from three public
15 resources: Human Genome Diversity Panel (HGDP) (Li *et al.* 2008), the
16 International HapMap Project phase III (The International HapMap Consortium
17 2007) and the 1000 Genomes Project (1KG) (The 1000 Genomes Project
18 Consortium 2012). Source populations for simulations are the haplotypes from
19 113 Utah residents with Northern and Western European ancestries from the
20 CEPH collection (CEU) and the 113 Africans from Yoruba (YRI).

21 **Inferring Admixture Histories by using the HI, GA, and CGF Models**

22 The expectation of weighted LD under a two-way admixture model has been
23 described in detail in another work (Zhou *et al.* 2015). Following the previous

1 notation, the expectation of weighted LD statistic between two sites separated by
2 a distance d (in Morgan) is as follows:

$$E[a_0(d)] = \sum_{i=1}^2 w_i^{(n)} E[a_i(d)] + F(d) \sum_{l=1}^n c^{(l)} \exp(-ld), \quad \text{EQ 1}$$

3 where $F(d) = \frac{\sum_{S(d)} (\delta_{12}(x) \delta_{12}(y))^2}{|S(d)|}$; $a_i(d)$, $i = 0, 1, 2$ are the weighted LD statistic of
4 the admixed population ($i = 0$) and the source population i , ($i = 1, 2$),
5 respectively; m_i is the admixture proportion from the source population i ; and
6 $\delta_{12}(x)$ is the allele frequency difference between populations 1 and 2 at site
7 x ; $S(d)$ is the set holding pairs of SNPs of distance d ; $c^{(l)}$ is admixture indicator
8 for the admixture event of l generations ago, and n is supposed to be the number
9 of generations ago when the source populations first met. To eliminate the
10 confounding effect due to background LD from the source populations, we used
11 the quantity, $z(d)$, defined as follows, to represent the admixture induced LD
12 (ALD) (Zhou *et al.* 2015).

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

13 We presented it in a more compact form using the inner product of two vectors
14 as follows:

$$z(d) = Ex(d)^T C;$$

16 where

$$C = (c^{(1)}, \dots, c^{(n-1)}, c^{(n)})^T;$$

18 and

$$Ex(d) = (\exp(-d), \dots, \exp(-(n-1)d), \exp(-nd))^T.$$

1 For different admixture models where admixture began n generations
2 ago, $z(d)$ varies in terms of the vector of coefficients of exponential functions
3 (Zhou *et al.* 2015):

$$\text{HI} \quad C_{\text{HI}} = (0, \dots, 0, m_1 m_2)^T$$

$$\text{GA} \quad C_{\text{GA}} = m_1 m_2 \left(\frac{(n-1)^0}{n}, \frac{(n-1)^1}{n^2}, \dots, \frac{(n-1)^{n-2}}{n^{n-1}}, \frac{(n-1)^{n-1}}{n^{n-1}} \right)^T$$

$$\text{CGF1} \quad C_{\text{CGF1}} = (1 - m_1^{1/n}) m_1 (m_1^{(n-1)/n}, m_1^{(n-2)/n}, \dots, 1)^T$$

$$\text{CGF2} \quad C_{\text{CGF2}} = (1 - m_2^{1/n}) m_2 (m_2^{(n-1)/n}, m_2^{(n-2)/n}, \dots, 1)^T$$

4 where the vector C_{model} has length n using the HI, GA, CGF1, or CGF2 model; and
5 n represents when the admixture occurred (HI) or began (GA and CGF) in terms
6 of generations. For different models, the coefficient vectors have different
7 patterns (Fig 2), which can be used to infer the best-fit model for a certain
8 admixed population.

9 In the CGF model, CGF1 represents the admixture where source
10 population 1 is the recipient of the gene flow from population 2, whereas CGF2
11 indicates source population 2 as gene flow recipient from population 1. Inference
12 of the admixture time assuming the true admixture history is one of these
13 different models that can be regarded as minimizing the objective function as
14 follows:

$$\text{ssE}(\theta_0, \theta_1, C_{\text{model}}) = ||\theta_0 \cdot \mathbf{1} + \theta_1 A C_{\text{model}} - Z||_2^2. \quad \text{EQ 2}$$

15 The optimization problem is therefore expressed as follows:

$$\min_{\theta_0, \theta_1 \text{ and } C_{\text{model}}} \text{ssE}(\theta_0, \theta_1, C_{\text{model}}), \quad \text{EQ 3}$$

16 where $Z = (z(d_1), z(d_2), \dots, z(d_I))^T$ is the observed ALD calculated from the
17 single nucleotide polymorphism (SNP) data of both the parental populations and

1 the admixed population; θ_0 is a real number used to correct the population
2 substructure; θ_1 is a scalar that improves estimation robustness; $\mathbf{1} \in R^I$ is a
3 vector with each entry being 1; A is an $I \times J$ matrix with the i th row vector
4 defined as $Ex(d_i)^T$, i.e., $A = (Ex(d_1), Ex(d_2), \dots, Ex(d_I))^T$, and $J \geq n$ is a pre-
5 specified upper bound of n . Our definitions are consistent since we can let all
6 entries be 0 after the n -th entry in C_{model} .

7 Next, we tried to estimate the parameters θ_0, θ_1 , and C_{model} , where C_{model}
8 has the information of the admixture model and the related admixture time n (in
9 generations). In our analysis, the value of n is assumed to be a positive integer;
10 therefore, our method is to go through all possible n values (with a reasonable
11 upper limit J) to estimate n with the minimum value of the objective function.
12 Given n , we used linear regression to estimate (θ_0, θ_1) such that the objective
13 function was minimized. Using this approach, the value of n in relation to the
14 minimal objective function value for each model was determined, which
15 represents the time of admixture occurrence under each model.

16 **Admixture Inference under HI, GA-I, and CGF-I Models**

17 GA and CGF models assume that the admixture is strictly continuous from
18 the beginning of admixture to present. This assumption seems too strong to be
19 valid in empirical studies. Here, we extended the GA model and CGF model to GA-
20 I model and CGF-I model, respectively, by considering continuous admixture
21 followed by isolation. In this case, the admixture event lasts from G_{start}
22 generations ago to G_{end} generations ago. Similar to the previous case, the
23 coefficients of exponential functions can be represented as the vector of length

- 1 G_{start} for each model, whose first $G_{\text{end}} - 1$ entries are filled with zeros. Suppose
- 2 the admixture lasted for n generations, then

$$\text{GA-I} \quad C_{\text{GA-I}} = m_1 m_2 \left(0, \dots, 0, \frac{(n-1)^0}{n}, \frac{(n-1)^1}{n^2}, \dots, \frac{(n-1)^{n-2}}{n^{n-1}}, \frac{(n-1)^{n-1}}{n^{n-1}} \right)^T$$

$$\text{CGF1-I} \quad C_{\text{CGF1-I}} = (1 - m_1^{1/n}) m_1 \left(0, \dots, 0, m_1^{(n-1)/n}, m_1^{(n-2)/n}, \dots, 1 \right)^T$$

$$\text{CGF2-I} \quad C_{\text{CGF2-I}} = (1 - m_2^{1/n}) m_2 \left(0, \dots, 0, m_2^{(n-1)/n}, m_2^{(n-2)/n}, \dots, 1 \right)^T$$

- 3 In this case, we can also try to find the parameters to minimize the
- 4 objective function (EQ 2) under new models. By examining all possible pairs of
- 5 $(G_{\text{end}}, G_{\text{start}})$, it is possible to determine the global minimum of the objective
- 6 function, although this might not be computationally efficient. Here, we used a
- 7 faster algorithm (**Algorithm 1**) to determine the starting and ending time points
- 8 of admixture.

- 9 Let E and S be the ending and starting time points (in generations, prior
- 10 to the present) of the admixture, which we want to search for to minimize the
- 11 objective function. The search starts from $(E^0, S^0) = (1, J)$, where J is the upper
- 12 bound for the beginning of the admixture event, which can be set to be a large
- 13 integer to seek for a relatively ancient admixture event. In our analysis of recent
- 14 admixed populations, we set $J = 500$. For $k = 1, 2, \dots$, (E^k, S^k) is updated from
- 15 (E^{k-1}, S^{k-1}) by two alternative proposals. For convenience, we define

$$f(E^k, S^k) := \min_{\theta_0, \theta_1} \text{ssE}(\theta_0, \theta_1, E^k, S^k), \quad \text{EQ 4}$$

- 16 where θ_0, θ_1 can be determined by linear regression.

- 17 We choose the proposal that results in a smaller value for f . The search
- 18 stops when the value of f with (E^{k-1}, S^{k-1}) is no larger than that of either
- 19 proposal or $E^k = S^k$. In this way, we can readily estimate the time interval of the
- 20 admixture event $(G_{\text{end}}, G_{\text{start}})$ quickly.

Algorithm 1:

for k *in* $1, 2, \dots$

$$(E_1^k, S_1^k) := (E^{k-1} + 1, S^{k-1})$$

$$(E_2^k, S_2^k) := (E^{k-1}, S^{k-1} - 1)$$

$$(E^k, S^k) := \underset{(E, S) \in \{(E_1^k, S_1^k), (E_2^k, S_2^k), (E^{k-1}, S^{k-1})\}}{\operatorname{argmin}} f(E, S)$$

$$\text{if } (E^k, S^k) = (E^{k-1}, S^{k-1}) \text{ or } E^k = S^k$$

$$(G_{\text{end}}, G_{\text{start}}) := (E^k, S^k)$$

stop

1 Result evaluation

2 To check our assumption of the true history and evaluate the inference, an
3 intuitive way is to compare empirical weighted LD with the fitted LD. Here, we
4 use two quantities: msE and Quasi F, defined by the following:

5 1) Let $e = \theta_0 \cdot 1 + \theta_1 AC_{\text{model}} - Z$. We look at $\text{msE} = \frac{\sum_1^I e_i^2}{I - 1}$ with e_i

6 being the i th entry of e . This reflects goodness of fit and strength of
7 background noise. A smaller msE indicates less background noise and
8 better fit.

9 2) Let $e' = \hat{Z} - Z$, where \hat{Z} is the fitted weighted LD obtained from
10 MALDmef, which theoretically can be regarded as the de-noised weighted
11 LD. e' is a vector of length I , with the i th entry denoted by e'_i . We look at

12 the quasi-F statistic $F = \frac{\sum_1^I e_i^2}{\sum_1^I (e'_i)^2}$. A small F indicates that the current fit

13 does not significantly deviate from the previous fit.

1 A reliable result should have both small msE and small F values. Particularly, F is
 2 involved in model comparison: when F is too large, one would suspect that the
 3 true admixture history is far from any one of these models. Both F and msE are
 4 involved in revealing data quality. If F is small but msE is large, one would
 5 suspect that the quality of data is not good enough to draw convincing
 6 conclusions. Further explanation of these statistics is in Results and Discussion
 7 sessions.

8 Identification of the best-fit model

9 For the convenience of illustration, we define the **core model** as the
 10 model used to infer admixture time. When inferring admixture of a target
 11 population, HI, GA, CGF1, CGF2, GA-I, CGF1-I and CGF2-I are used as the core
 12 models for conducting inference. Because GA-I, CGF1-I and CGF2-I describe more
 13 general admixture models than GA, CGF1, and CGF2, we classified model
 14 selection into two cases: one case is to identify the best-fit model(s) among the
 15 HI, GA, CGF1, and CGF2 models, whereas the more general case is to determine
 16 the best-fit model(s) among HI, GA-I, CGF1-I and CGF2-I models. In both cases,
 17 the same strategy is adopted, which depends on the pairwise paired difference of
 18 pseudo $\log(\text{msE})$ values associated with each core model, which will be defined
 19 later. For an admixed population, there are $N + 1$ observed weighted LD curves
 20 obtained as follows: N (typically 22) autosomal chromosomes are considered in
 21 an individual genome, and one weighted LD curve is calculated from all these N
 22 chromosomes while the other N weighted LD curves are obtained by jackknife
 23 resampling, leaving out one chromosome for each LD curve (Loh *et al.* 2013;
 24 Pickrell *et al.* 2014; Zhou *et al.* 2015). Next, we fit each observed weighted LD

1 curve for each core model by estimating θ_0, θ_1 and the time interval, which in
2 turn allowed us to obtain the msE value associated with the optimal parameters
3 for each weighted LD curve. Taken together, a total of $N + 1$ msE values
4 associated with $N + 1$ LD curves were evaluated in each core model. For model
5 M , the $\log(\text{msE})$ obtained from all N chromosomes is denoted by ϵ_0^M and that
6 from the LD curve with the q -th chromosome left out by ϵ_q^M . Following Tukey
7 (Tukey 1958), we defined the q -th pseudo $\log(\text{msE})$ for model M to be
8 $\hat{\epsilon}_q^M = N\epsilon_0^M - (N - 1)\epsilon_q^M$ and treated these pseudo values approximately as
9 independent. Next, we defined the best-fit core model(s) to be the model(s) with
10 significantly small $\hat{\epsilon}_q^M$. A pairwise Wilcoxon signed-rank test was conducted for
11 the pseudo $\log(\text{msE})$ of the four models. More precisely, Wilcoxon signed-rank
12 test is applied to all pairs of models with the $\hat{\epsilon}_q^M$ being paired by index q , and
13 then the p-values are adjusted to control familywise error rate (Table 1). We
14 used the Holm-Bonferroni method to adjust p-values (Holm 1979). When
15 $\hat{\epsilon}_q^{\text{HI}}$ were not significantly larger than those of the best model, i.e., the model
16 associated with the smallest sample median of pseudo $\log(\text{msE})$ values, HI was
17 selected. Otherwise the models whose $\hat{\epsilon}_q^M$ were not significantly larger than
18 those of the best model were selected (the best model was selected as well). The
19 significance level was set to be 0.05. Here, we paired the pseudo values
20 according to index q and used Wilcoxon signed-rank test on the paired
21 differences because according to our experience, $\hat{\epsilon}_q^M$ are strongly correlated with
22 q and hence q is a major covariate that must be controlled in the test to gain
23 higher power. This is also why even though theoretically there are examples
24 where the best model according to our definition can be significantly worse than

1 another model in our process, we believe such extreme cases are unlikely in
2 practice and still use this method. In addition, $\log(\text{msE})$ rather than msE were
3 used because after logarithm transformation small values of msE can also have
4 large effect to the comparison. That is, we could better detect the difference
5 between small msE, thus gaining greater power in the test. This claim is also
6 justified by our experience.

7 **Software**

8 Our algorithm has been implemented in an R package (R Core Team 2014),
9 named CAMer (Continuous Admixture Modeler). The package is available on the
10 website of population genetic group: <http://www.picb.ac.cn/PGG/resource.php>
11 or on Github: <https://www.github.com/david940408/CAMer>.

12 **Results**

13 **Simulation studies**

14 Admixed populations were simulated in a forward-time way under
15 different admixture models with the software **AdmixSim** (Yang 2015).
16 Simulation was initiated with the haplotypes from source populations (YRI and
17 CEU) and haplotypes for the admixed population were generated by resampling
18 haplotypes with recombination from source populations and the admixed
19 population of last generation. During the simulation, population size was kept as
20 5000 and migration rates was controlled by the admixture model with the final
21 admixture proportion in the admixed population to be 0.3. We employed a mono
22 recombination map in our simulation, which means recombination rate between

1 two markers is positively proportional to their physical distance. For each model,
2 simulation was performed using 10 replicates; each replicate contained 10
3 chromosomes with a total length of 3 Morgans. To evaluate the performance of
4 our algorithm, we simulated admixed populations under the following
5 conditions:

- 6 1) HI of 50 and 100 generations, designated as HI (50) and HI (100),
- 7 2) GA of 50 and 100 generations, designated as GA (1-50) and GA (1-
8 100), respectively,
- 9 3) CGF of 50 and 100 generation, population 1 as the recipient,
10 designated as CGF1 (1-50) and CGF1 (1-100) respectively,
- 11 4) CGF-I of a 70-generation admixture followed by 30-generation
12 isolation, and a 30-generation admixture followed by a 70-generation
13 isolation, with population 1 as the recipient, designated as CGF1-I (30-
14 100) and CGF1-I (70-100) respectively, and,
- 15 5) GA-I of a 70-generation admixture followed by a 30-generation
16 isolation and a 30-generation admixture followed by a 70-generation
17 isolation, designated as GA-I (30-100) and GA-I (70-100), respectively.

18
19 With simulated admixed populations, we first used the HI, GA and CGF
20 models as core models to conduct inference (Fig S1). When the simulated model
21 was a HI, GA, or CGF model, our method was able to accurately estimate the
22 simulated admixture time, as well as to determine the correct model, with an
23 accuracy of 73.33%. When the simulated model was a CGF-I or GA-I model, the
24 estimated time based on the core model HI was within the time interval of the
25 admixture, whereas all best-fit models were HI (Table 2). This result has

1 indicated the limitation of using the GA and CGF models in inferring admixture
2 history.

3 Using the same simulated admixed populations, we then employed GA-I,
4 CGF-I and HI as core models for performing inference (Figs 3 and S2-S11). With
5 HI, GA, or CGF considered as the true model, our estimation of the optimal model
6 remained highly accurate. On the other hand, when the true model was GA-I or
7 CGF-I, the failure rate decreased by 25%, compared to the estimation in the
8 previous setting. Furthermore, the estimated time intervals were wider than
9 those of the true ones, although the findings were still more accurate than those
10 using GA and CGF as core models (Table 2).

11 Empirical analysis

12 We applied CAMer to the selected admixed populations from HapMap,
13 HGDP, and 1KG. For each target population, we first used MALDmef to calculate
14 the weighted LD and fit the weighted LD with hundreds of exponential functions
15 (Zhou *et al.* 2015). Next, with the weighted LD of target populations, we
16 determined the admixture model and estimated admixture time with CAMer.
17 Quasi F and msE are designed for evaluating the inference with CAMer. The value
18 of msE usually indicates data quality: small msE may indicate a high signal-to-
19 noise ratio (SNR) and vice versa. The quasi F value measures the goodness of fit
20 of the model we employed to fit the admixture event. A small F value indicates
21 that the model we used was of satisfactory performance in modeling an
22 admixture event. In our analysis, we used 10^{-5} as the threshold for msE and 1.5
23 as the threshold for F . Therefore, when the msE value $\leq 10^{-5}$ and the F value
24 ≤ 1.5 , we could not “reject the null hypothesis” that the related model was the

1 true model, i.e., the model well fit the admixture event. On the other hand, an
2 msE value $\geq 10^{-5}$ indicates low- quality data that is incapable of identifying the
3 best-fit model, whereas an F value ≥ 1.5 prompts us to “reject the null
4 hypothesis” and concludes that the model did not well fit the admixture. In the
5 case of the same population from different databases, the data with smaller msE
6 values were given more credits. For example, we obtained samples of ASW from
7 the HapMap and the 1KG. With the ASW data (CEU and YRI as source
8 populations) from HapMap, the best-fit model was HI of 6 generations, and both
9 msE and F values indicated that the inference was acceptable (Fig S12).
10 Similarly, using the ASW data (CEU and YRI as source populations) from 1KG, the
11 best-fit model was HI of 6 generations (Fig S13). However, a quasi F value of 2.54
12 indicated that HI model did not satisfactorily fit the admixture event. Because the
13 msE value of the data set from 1KG was smaller, the conclusion using ASW was
14 as follows: based on the best data we had, the time intervals estimated under the
15 HI, GA-I, CGF1-I, and CGF2-I model were 6 generations, 1–9 generations, 1–13
16 generations, and 1–9 generations, respectively. Furthermore, none of these
17 models satisfactorily modeled the admixture, whereas the HI model showed
18 better performance. We also applied CAMer to other admixed populations (Table
19 3, Figs S14–17). MEX (source populations: CEU (64 individuals) and American
20 Indian (7 Colombians, 14 Karitiana, 21 Maya, 14 Pimas and 8 Suruis)) was
21 satisfactorily modeled by the CGF1 model or GA-I model, with the estimated
22 admixture time interval being 1–17 or 2–16 generations, respectively. We also
23 analyzed Eurasian populations, which showed that the Uyghurs (source
24 populations: Han ($n = 34$) and French ($n = 28$)) most likely fit a continuous
25 model, with a gene flow lasting for more than 60 generations to the present or

1 near present. We cannot determine which model fits best. However, the values of
2 msE were all larger than 10^{-5} , indicating that the results were not so reliable.
3 The Hazara population (source populations: Han (n = 34) and French (n = 28))
4 experienced a GA-I-like admixture event that lasted for about 58 generations,
5 which started 63 generations ago and ended approximately 5 generations ago.

6 **Discussion**

7 Modeling the demographic history of an admixed population and estimating time
8 points of this particular event are essential components of evolutionary and
9 medical research studies (Zhu *et al.* 2004; Zhu and Cooper 2007; Gravel 2012; Jin
10 *et al.* 2012, 2013; Ni *et al.* 2015; Zhou *et al.* 2015). Previous methods have
11 employed the length distribution of ancestral tracts (Gravel 2012; Jin *et al.* 2012,
12 2013), which highly depends on the result of local ancestral inference and
13 haplotype phasing. Another limitation of earlier methods is that only HI, GA, and
14 CGF models were utilized to fit the admixture as well as in identifying the best-fit
15 model. In the present study, our simulations showed that when the true model
16 was not HI, GA, or CGF, the generated inferences were relatively difficult to
17 interpret.

18 Our method, CAMer, can be utilized in inferring admixture histories by
19 using weighted LD, which can be calculated using genotype data with MALDmef
20 (Zhou *et al.* 2015). Furthermore, we extended the GA and CGF models to the GA-I
21 and CGF-I models in order to infer the time interval for a period of continuous
22 admixture events followed by isolation. Although HI model is a degenerate case
23 for both GA-I and CGF-I models, where the admixture window becomes 1
24 generation, we kept it in our method because it is the most popular model

1 employed in previous admixture studies. Considering the difficulty in the fitting
2 problem with exponential functions, it is in our expectation that CAMer was not
3 consistently very accurate in determining the admixture model based on the
4 weighted LD decay. However, its natural advantage of independence of both
5 haplotype phasing and local ancestry inference makes it privilege to other CAT
6 based method. And our simulations indicated that its time interval estimations
7 were reliable when its assumption that the true admixture history could be well
8 approximated by one of the core models is valid.

9 Two quantities, namely msE and quasi F , were used to check the
10 assumption of our method stated above and evaluate the credibility of the
11 models' inference. These two quantities should both be taken into consideration
12 to identify whether the models well describe the admixture history. Both the
13 data quality and the goodness of fitting of models can affect the value of msE,
14 although the F value mainly measures the goodness of modeling. Informally, for
15 the convenience of interpretation, msE is considered to reveal the data quality
16 and F value is considered to check model assumption on admixture history. In
17 our analysis, we suggested thresholds for msE and F to determine whether the
18 null hypothesis should be rejected or not, which may be too strict in empirical
19 analysis. Actually, msE and F values together measure whether the observed
20 weighted LD can be well fit by the best-fit model(s). For example, the fitting
21 process showed poor performance in the MKK population, which was
22 accompanied by exaggerated msE and F values, showing significant
23 inconsistencies between the observed and fitted weight LD curves, which
24 indicates that the true admixture history cannot be well explained by any of the
25 core models (Fig S17). Therefore, in empirical analysis, one can informally think

1 that the msE value reflects the quality of the data, whereas F value describes the
2 performance of the model, although both of them measure the goodness of
3 fitting.

4 In our previous study (Zhou *et al.* 2015), we fit the weighted LD with
5 hundreds of exponential functions. However, this approach did not fully reveal
6 the occurrence of continuous admixture. To address this issue, the present study
7 developed CAMer to model admixture as a continuous process. CAMer also
8 employed extensions of the classic continuous models, GA-I and CGF-I, which
9 may bring the bias to have a wider admixture window when the real admixture
10 exists in a short time. But it is still proved to be able to give more credible
11 estimations in modeling population admixture.

12 Taken together, CAMer is a powerful method to model a continuous
13 population admixture, which in turn would help us elucidate the complex
14 demographic history of population admixture.

15

16 **Author contributions**

17 Conceived and designed the study: **SX**. Developed methods and computer tools: **YZ**
18 **HQ**. Analyzed the data: **YZ** and **HQ**. Interpreted the data and wrote the paper: **SX YZ**
19 **HQ**.

20 **Funding:** These studies were supported by the Strategic Priority Research
21 Program of the Chinese Academy of Sciences (CAS) (XDB13040100), by the
22 National Science Fund for Distinguished Young Scholars (31525014), by the
23 National Natural Science Foundation of China (NSFC) grants (91331204,

1 31171218, 31501011), and by Science and Technology Commission of Shanghai
 2 Municipality (14YF1406800). S.X. is Max-Planck Independent Research Group
 3 Leader and member of CAS Youth Innovation Promotion Association. S.X. also
 4 gratefully acknowledges the support of the National Program for Top-notch
 5 Young Innovative Talents of The "*Wanren Jihua*" Project. The funders had no role
 6 in study design, data collection and analysis, decision to publish, or preparation
 7 of the manuscript.

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

10 **Acknowledgements:** None.

11

12

13

1 Fig Legends

2 Fig 1: Classic admixture models (HI, GA and CGF) and the models we extended
3 (GA-I and CGF-I). For each model, the simulated admixed population (Hybrid) is
4 in the middle of two source populations (POP1 and POP2). Each horizontal
5 arrow represents the direction of gene flow from the source populations to the
6 admixed population. Once the genetic components flow into the admixed
7 population, the admixed population randomly hybridizes with other existing
8 components. The existence of horizontal arrows indicates gene flow from the
9 corresponding source population.

10 Fig 2: Coefficient vector of exponential functions for each model. For each
11 admixture model, the starting time of the population admixture is 50 generations
12 ago.

13 Fig 3: Evaluation of CAMer under various simulated admixture models. Here, the
14 core models are HI, GA-I, CGF1-I, and CGF2-I. The simulated models (True
15 Model) are listed on the left, with the admixture time interval depicted in the
16 parentheses. The gray area on the middle vertical panel is the simulated time
17 interval, whereas colored lines indicate the estimated time intervals under
18 different core models. HI: pink; CGF1-I: green; CGF2-I: purple; GA-I: blue. The
19 intensity of lines means the number each point is covered by the time intervals
20 estimated from all jackknives. Lighter colors represent fewer covers while
21 darker colors mean more.

22

Table 1: Adjusted p-values of pairwise Wilcoxon signed-rank test among core models: HI, GA-I, CGF1-I, CGF2-I.

True Model	Best Model(s)	Adjusted p-Values of Pairwise Wilcoxon Signed Rank Test					
		HI: GA-I	HI: CGF1-I	HI: CGF2-I	GA-I: CGF1-I	GA-I: CGF2-I	CGF2-I: CGF1-I
HI (100)	HI	0.97	0.14	0.97	0.012	0.15	0.97
HI (50)	HI	0.98	0.98	0.52	0.70	0.16	0.16
CGF1 (1-100)	CGF1-I	0.012	0.012	0.012	0.012	0.012	0.012
CGF1 (1-50)	CGF1-I, CGF2-I	0.012	0.012	0.012	0.041	0.064	0.055
GA (1-100)	GA-I	0.012	0.012	0.012	0.012	0.012	0.012
GA (1-50)	GA-I	0.012	0.012	0.012	0.012	0.020	0.012
CGF1-I (30-100)	HI	0.19	0.55	0.15	0.012	0.55	0.012
CGF1-I (70-100)	HI	0.97	0.97	0.97	0.020	0.012	0.20
GA-I (30-100)	CGF1-I, GA-I	0.012	0.012	0.012	0.30	0.012	0.020
GA-I (70-100)	HI	0.52	0.52	0.52	0.029	0.012	1

In each column, the adjusted p-values of the Wilcoxon signed-rank test comparing the two models are presented for all simulation cases. Simulated true model is followed by the parenthesis of time interval for the corresponding gene flow, where the first term in the parenthesis is the ending time of the admixture and the second term is the beginning time of the admixture. They are in the measurements of generation before present. For HI model, only one time point is included in the parenthesis.

1 **Table 2: Accuracy of model detection**

True models	Core models	Counts			Rates		
		Correct	Undeter-mined	Wrong	Correct	Undeter-mined	Wrong
HI;GA;CGF	HI;GA;CGF	44	15	1	73.3%	25.0%	1.7%
GA-I;CGF-I	HI;GA;CGF	0	0	40	0.0%	0.0%	100.0%
HI;GA;CGF	HI;GA-I; CGF-I	37	22	1	61.7%	36.7%	1.6%
GA-I;CGF-I	HI;GA-I; CGF-I	1	9	30	2.5%	22.5%	75.0%

2 Here, as our method can hardly distinguish CGF1 from CGF2 model, we regard
3 CGF1, CGF2 as the CGF model; CGF1-I and CGF2-I as the CGF-I model, which is
4 different from GA-I and HI models.
5

Table 3: Results of CAMer on empirical populations

Population	Core model	End time	Start time	msE	Quasi.F
ASW-HapMap (57)	HI*	6	6	2.72×10^{-6}	1.19
	CGF1-I	2	10	3.02×10^{-6}	1.41
	CGF2-I	1	9	2.98×10^{-6}	1.40
	GA-I	3	8	2.97×10^{-6}	1.19
ASW-1KG (56)	HI*	6	6	2.19×10^{-6}	<u>2.54</u>
	CGF1-I	1	11	1.88×10^{-6}	<u>2.19</u>
	CGF2-I	1	9	1.84×10^{-6}	<u>2.12</u>
	GA-I	2	9	1.86×10^{-6}	<u>2.13</u>
MEX (86)	HI	9	9	6.73×10^{-6}	<u>2.19</u>
	CGF1-I*	1	17	3.57×10^{-6}	1.13
	CGF2-I	1	18	3.57×10^{-6}	1.15
	GA-I*	2	16	3.60×10^{-6}	1.15
MKK (143)	HI*	6	6	2.36×10^{-5}	<u>11.68</u>
	CGF1-I	1	16	2.04×10^{-5}	<u>10.24</u>
	CGF2-I	1	11	2.15×10^{-5}	<u>10.82</u>
	GA-I	1	17	1.97×10^{-5}	<u>9.83</u>
UIG (10)	HI	26	26	4.73×10^{-5}	1.29
	CGF1-I*	1	66	4.01×10^{-5}	1.08
	CGF2-I*	1	64	4.01×10^{-5}	1.08
	GA-I*	3	63	4.03×10^{-5}	1.09
Hazara (24)	HI	27	27	1.26×10^{-5}	<u>1.95</u>
	CGF1-I	3	69	8.78×10^{-6}	1.35
	CGF2-I	3	65	8.87×10^{-6}	1.36
	GA-I*	5	63	8.53×10^{-6}	1.30

Number of individuals listed in the parentheses. Values underlined do not pass our threshold. The time interval is summarized from 22 jackknives, which is shared by more than half of all estimated intervals for continuous models or the nearest integer to the mean of estimated time point for HI model. The best-fit model is marked by an asterisk “*”. For the HI model, the beginning time is the same as the ending time.

1 **Reference**

- 2 Gravel S., 2012 Population genetics models of local ancestry. *Genetics* **191**: 607–
3 619.
- 4 Holm S., 1979 A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J.*
5 *Stat.* **6**: 65–70.
- 6 Jin W., Wang S., Wang H., Jin L., Xu S., 2012 Exploring population admixture
7 dynamics via empirical and simulated genome-wide distribution of
8 ancestral chromosomal segments. *Am. J. Hum. Genet.* **91**: 849–862.
- 9 Jin W., Li R., Zhou Y., Xu S., 2013 Distribution of ancestral chromosomal segments
10 in admixed genomes and its implications for inferring population history
11 and admixture mapping. *Eur. J. Hum. Genet.* **22**: 930–937.
- 12 Li J. Z., Absher D. M., Tang H., Southwick A. M., Casto A. M., Ramachandran S.,
13 Cann H. M., Barsh G. S., Feldman M., Cavalli-Sforza L. L., Myers R. M., 2008
14 Worldwide human relationships inferred from genome-wide patterns of
15 variation. *Science* **319**: 1100–1104.
- 16 Loh P. R., Lipson M., Patterson N., Moorjani P., Pickrell J. K., Reich D., Berger B.,
17 2013 Inferring admixture histories of human populations using linkage
18 disequilibrium. *Genetics* **193**: 1233–1254.
- 19 Ni X., Yang X., Guo W., Yuan K., Zhou Y., Ma Z., Xu S., 2015 Length distribution of
20 ancestral tracts under a general admixture model and its applications in
21 admixture history inference. *Under Rev.*
- 22 Pfaff C. L., Parra E. J., Bonilla C., Hiester K., McKeigue P. M., Kamboh M. I.,
23 Hutchinson R. G., Ferrell R. E., Boerwinkle E., Shriver M. D., 2001 Population
24 structure in admixed populations: effect of admixture dynamics on the
25 pattern of linkage disequilibrium. *Am. J. Hum. Genet.* **68**: 198–207.
- 26 Pickrell J. K., Patterson N., Loh P.-R., Lipson M., Berger B., Stoneking M.,
27 Pakendorf B., Reich D., 2014 Ancient west Eurasian ancestry in southern
28 and eastern Africa. *Proc. Natl. Acad. Sci. U. S. A.* **111**: 2632–7.
- 29 Price A. L., Tandon A., Patterson N., Barnes K. C., Rafaels N., Ruczinski I., Beaty T.
30 H., Mathias R., Reich D., Myers S., 2009 Sensitive detection of chromosomal
31 segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**.
- 32 Qin P., Zhou Y., Lou H., Lu D., Yang X., Wang Y., Jin L., Chung Y.-J., Xu S., 2015
33 Quantitating and Dating Recent Gene Flow between European and East
34 Asian Populations. *Sci. Rep.* **5**: 9500.
- 35 R Core Team, 2014 R: A Language and Environment for Statistical Computing. **0**.

- 1 The 1000 Genomes Project Consortium, 2012 An integrated map of genetic
2 variation from 1,092 human genomes. *Nature* **135**: 0–9.
- 3 The International HapMap Consortium, 2007 A second generation human
4 haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.
- 5 Tukey J. W., 1958 Bias and Confidence in Not-Quite Large Samples. *Ann. Math.*
6 *Stat.* **29**: 614.
- 7 Xu S., Jin L., 2008 A Genome-wide Analysis of Admixture in Uyghurs and a High-
8 Density Admixture Map for Disease-Gene Discovery. *Am. J. Hum. Genet.* **83**:
9 322–336.
- 10 Yang X., 2015 AdmixSim-v1.0.2, <http://www.picb.ac.cn/PGG/resource.php>.
- 11 Zhou Y., Yuan K., Yu Y., Ni X., Xie P., Xing E. P., Xu S., 2015 Inference of multiple-
12 wave population admixture by modeling decay of linkage disequilibrium
13 with multiple exponential functions. Under Rev.
- 14 Zhu X., Cooper R. S., Elston R. C., 2004 Linkage analysis of a complex disease
15 through use of admixed populations. *Am. J. Hum. Genet.* **74**: 1136–1153.
- 16 Zhu X., Cooper R. S., 2007 Admixture mapping provides evidence of association
17 of the VNN1 gene with Hypertension. *PLoS One* **2**.
- 18

HI

GA

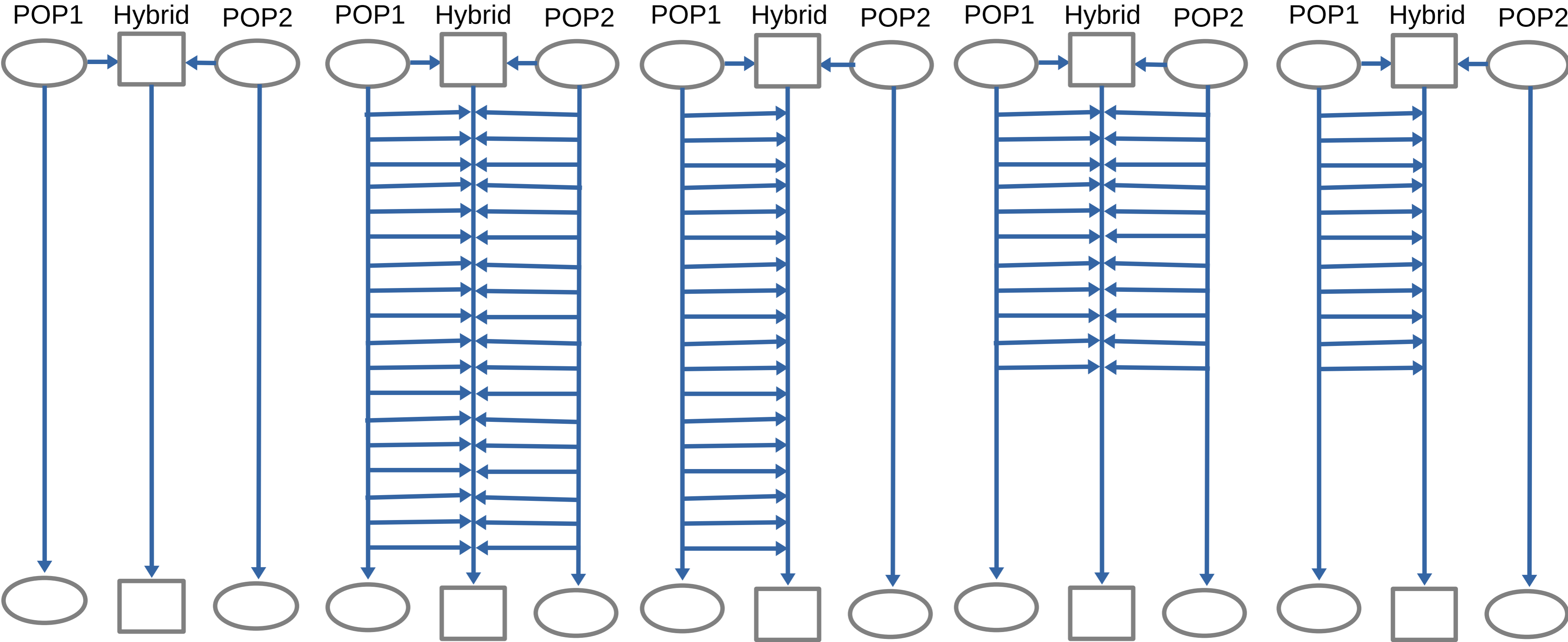
CGF

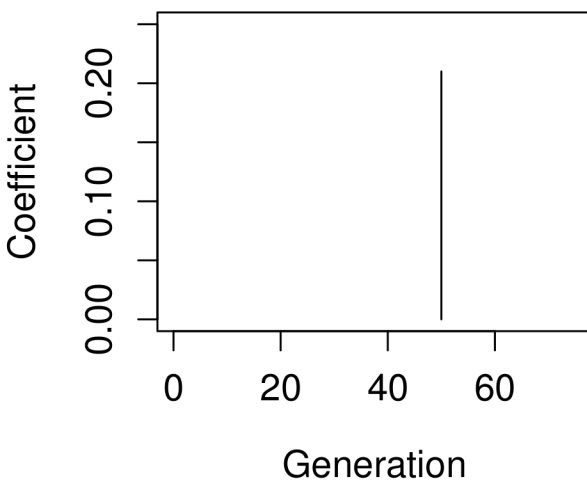
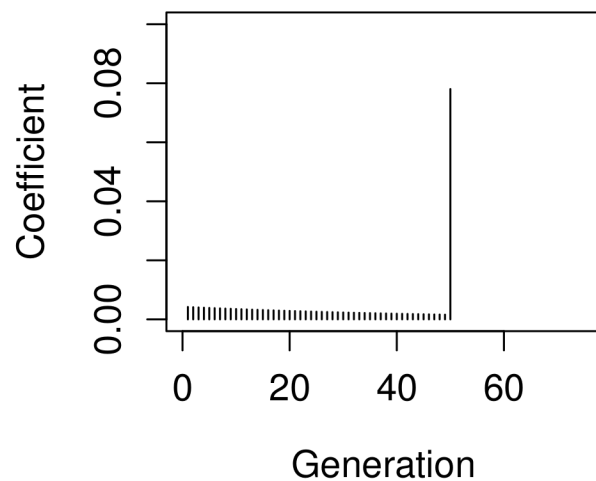
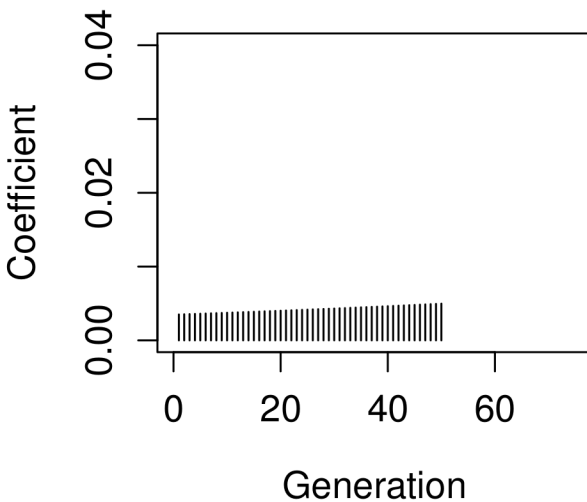
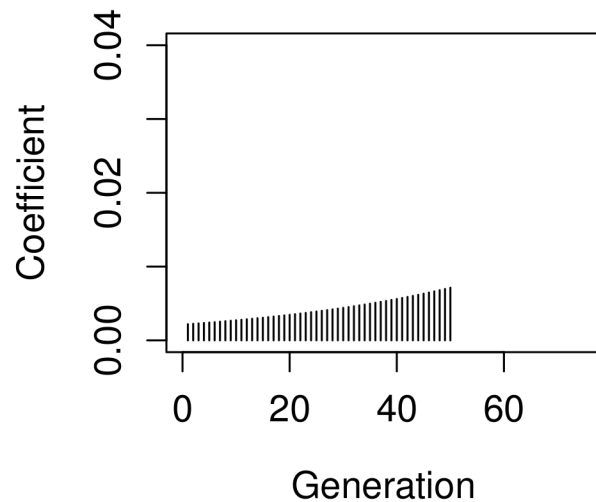
GA-I

CGF-I

Past

Present



A**HI****B****GA****C****CGF1****D****CGF2**

True Model**Time Intervals****Best Model(s)**