

# 1 Modeling Continuous Admixture

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

4

5 *Ying Zhou<sup>†, §</sup>, Hongxiang Qiu<sup>†, ‡, §</sup>, Shuhua Xu<sup>†, ††, ‡‡, ††, \*</sup>*

6

7 <sup>†</sup> 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 <sup>‡</sup> Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong  
12 Kong, China;

13 <sup>††</sup> School of Life Science and Technology, ShanghaiTec University, Shanghai  
14 200031, China;

15 <sup>‡‡</sup> Collaborative Innovation Center of Genetics and Development, Shanghai  
16 200438, China.

17 <sup>§</sup> These authors contributed equally to this work.

18 \* Correspondence and requests for materials should be addressed to

19 [xushua@picb.ac.cn](mailto:xushua@picb.ac.cn) (S.X.)

20

21

1

## Abstract

2 Human migration and human isolation serve as the driving forces of modern  
3 human civilization. Recent migrations of long isolated populations has resulted  
4 in genetically admixed populations. The history of population admixture is  
5 generally complex; however, understanding the admixture process is critical to  
6 both evolutionary and medical studies. Here, we utilized admixture induced  
7 linkage disequilibrium (LD) to infer occurrence of continuous admixture events,  
8 which is common for most existing admixed populations. Unlike previous  
9 studies, we expanded the typical continuous admixture model to a more general  
10 admixture scenario with isolation after a certain duration of continuous gene  
11 flow, and we demonstrated that such treatment significantly improved the  
12 accuracy of inference under complex admixture scenarios. Based on the  
13 extended models, we developed a method based on weighted LD to infer the  
14 admixture history considering continuous and complex demographic process of  
15 gene flow between populations. We evaluated the performance of the method by  
16 computer simulation and applied our method to real data analysis of a few well-  
17 known admixed populations.

18

## Introduction

19 Human migrations involve gene flow among previously isolated populations,  
20 resulting in the generation of admixed populations. In both evolutionary and medical  
21 studies of admixed populations, it is essential to understand admixture history and  
22 accurately estimate the time since population admixture because genetic architecture  
23 at both population and individual levels are determined by admixture history,

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

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

21 Weighted linkage disequilibrium (LD) is an alternative tool that can be used to  
22 infer admixture (Loh *et al.* 2013; Pickrell *et al.* 2014). Previous studies have indicated  
23 that this tool is more efficient than CAT because it requires neither ancestry  
24 information inference nor haplotype phasing, which often provides false  
25 recombination information, thus decreasing the power of estimation. Weighted LD

1 has already been used in inferring multiple-wave admixtures (Zhou *et al.* 2015).  
2 However, these methods tend to summarize the admixture into different independent  
3 waves, even if the true admixture is continuous. In our previous work (Zhou *et al.*  
4 2015), we mathematically described weighted LD under different continuous models,  
5 allowing us to determine admixture history using these models.

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

## 14 **Materials and Methods**

### 15 **Data Sets**

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

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

2           The expectation of weighted LD under two-way admixture model has  
 3 been described earlier (Zhou *et al.* 2015). Following the previous notation, the  
 4 expectation between two sites separated by a distance  $d$  (in Morgan) is as  
 5 follows:

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

6 where  $b^{(l)}(d) = c^{(l)} E\left(\left(\delta_{12}(x)\delta_{12}(y)\right)^2 \mid |x - y| = d\right)$ ;  $D(d)$  and  $D_i(d)$  are the  
 7 expected weighted LD of the admixed population and the source population  $i$ ,  
 8 respectively;  $m_i$  is the admixture proportion from the source population  $i$ ; and  
 9  $\delta_{12}(x)$  is the allele frequency difference between populations 1 and 2 at site  $x$ . To  
 10 eliminate the confounding effect due to background LD from the source  
 11 populations, we used the quantity,  $Z(d)$ , defined as the follows, to represent the  
 12 admixture induced LD (ALD).

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

13 We present it in a more compact form using the inner product of two vectors as  
 14 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           Therefore, for different admixture models where admixture began  
2  $n$  generations ago,  $Z(d)$  varies in terms of the vector of coefficients of  
3 exponential functions:

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

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

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

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

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

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

14           The optimization problem is therefore expressed as follows:

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

15 where  $Z = (Z(d_1), Z(d_2), \dots, Z(d_l))^T$  is the observed ALD calculated from the  
16 single nucleotide polymorphism (SNP) data of both the parental populations and  
17 the admixed population;  $\theta_0$  is a real number used to correct the population

1 substructure;  $\theta_1$  is a scalar that improves estimation robustness;  $\mathbf{1} \in R^I$  is a  
2 vector with each entry being 1;  $A$  is an  $I \times J$  matrix with the  $i$ th row vector  
3 defined as  $Ex(d_i)^T$ , i.e.,  $A = (Ex(d_1), Ex(d_2), \dots, Ex(d_I))^T$ .

4 Next, we tried to estimate the parameters  $\theta_0$ ,  $\theta_1$ , and  $C_{\text{model}}$ , where  $C_{\text{model}}$   
5 has the information of the admixture model and the related admixture time  $n$  (in  
6 generations). In our analysis, the value of  $n$  is assumed to be a positive integer;  
7 therefore, our method is to go through all possible  $n$  values (with a reasonable  
8 upper limit) to estimate  $n$  with the minimum value of the objective function.  
9 Given  $n$ , we used linear regression to estimate  $(\theta_0, \theta_1)$  such that the objective  
10 function was minimized. Using this approach, the value of  $n$  in relation to the  
11 minimal objective function value for each model was determined, which  
12 represents the time of admixture occurrence under each model.

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

14 GA and CGF models assume that the admixture is strictly continuous from  
15 the beginning of admixture to present. This assumption seems too strong to be  
16 valid in empirical studies. Here, we extended the GA model and CGF model to GA-  
17 I model and CGF-I model, respectively, by considering continuous admixture  
18 followed by isolation. In this case, the admixture event lasts from  $G_{\text{start}}$   
19 generations ago to  $G_{\text{end}}$  generations ago. Similar to the previous case, the  
20 coefficients of exponential functions can be represented as the vector of length  
21  $G_{\text{start}}$  for each model, whose first  $G_{\text{end}} - 1$  entries are filled with zeros. Suppose  
22 the admixture lasted for  $n$  generations, then

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

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

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

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

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

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

14 where  $\theta_0, \theta_1$  can be determined by linear regression.

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

---

**Algorithm 1:**

---

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

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


---



---


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

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

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

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

**stop**

---

## 1 Result evaluation

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

5 1) Let  $e = \theta_0 \cdot \mathbf{1} + \theta_1 AC_{\text{model}} - Z$ . 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.

14 A reliable result should have both small msE and small  $F$  values.

## 1 Identification of the best-fit model

2 For convenience of illustration, we define the **core model** as the model  
3 used to infer admixture time. When inferring admixture of a target population,  
4 HI, GA, CGF1, CGF2, GA-I, CGF1-I and CGF2-I are used as the core models for  
5 conducting inference. Because GA-I, CGF1-I and CGF2-I describe more general  
6 admixture models than GA, CGF1, and CGF2, we classified model selection into  
7 two cases: one case is to identify the best-fit model(s) among the HI, GA, CGF1,  
8 and CGF2 models, whereas the more general case is to determine the best-fit  
9 model(s) among HI, GA-I, CGF1-I and CGF2-I models. In both cases, the same  
10 strategy is adopted, which depends on the pairwise paired difference of  $\log(\text{msE})$   
11 values associated with each core model. For an admixed population, there are 22  
12 observed weighted LD curves (considering 22 autosomal chromosomes in an  
13 individual genome; using jackknife by leaving out one chromosome to calculate  
14 each LD curve) (Zhou *et al.* 2015). Next, we fitted the observed weighted LD  
15 curve for each core model by estimating  $\theta_0$ ,  $\theta_1$  and the time interval, which in  
16 turn allowed us to obtain the msE value associated with the optimal parameters  
17 for each weighted LD curve. Taken together, a total of 22 msE values associated  
18 with 22 LD curves were evaluated in each core model. Based on msE values  
19 calculated with different core models, the best-fit core model(s) are those with  
20 significantly small msE values. A pairwise paired t-test was conducted for the  
21  $\log(\text{msE})$  of the four models (Table 1). When the  $\log(\text{msE})$  of HI were not  
22 significantly larger than the those of the best model, i.e., the model associated  
23 with the smallest mean of the  $\log(\text{msE})$  values, HI was selected. Otherwise the  
24 models whose  $\log(\text{msE})$  were not significantly larger than those of the best  
25 model were selected (the best model is selected as well). To control the family-

1 wise error rate, the Holm-Bonfferoni method (Holm 1979) was used to adjust p-  
2 values, and the significance level was set to 0.05 in the present study. Here,  
3 log(msE) rather than msE were used because from our experience, log(msE) was  
4 more similar to the normal distribution.

## 5 **Results**

### 6 **Simulation studies**

7 Admixed populations were simulated in a forward-time way under  
8 different admixture models. For each model, simulation was performed using 10  
9 replicates; each replicate contained 10 chromosomes with a total length of 3  
10 Morgans. To evaluate the performance of our algorithm, we simulated admixed  
11 populations under the following conditions:

- 12 1) HI of 50 and 100 generations, designated as HI (50) and HI (100),
- 13 2) GA of 50 and 100 generations, designated as GA (1-50) and GA (1-  
14 100), respectively,
- 15 3) CGF of 50 and 100 generation, population 1 as the recipient,  
16 designated as CGF1 (1-50) and CGF1 (1-100) respectively,
- 17 4) CGF-I of a 70-generation admixture followed by 30-generation  
18 isolation, and a 30-generation admixture followed by a 70-generation  
19 isolation, with population 1 as the recipient, designated as CGF1-I (30-  
20 100) and CGF1-I (70-100) respectively, and,
- 21 5) GA-I of a 70-generation admixture followed by 30-generation isolation  
22 and a 30-generation admixture followed by a 70-generation isolation,  
23 designated as GA-I (30-100) and GA-I (70-100), respectively.

1

2           With simulated admixed populations, we first used the HI, GA and CGF  
3 models as core models to conduct inference (Fig S1). When the simulated model  
4 was a HI, GA, or CGF model, our method was able to accurately estimate the  
5 simulated admixture time, as well as determine the correct model, with an  
6 accuracy of 86.67%. When the simulated model was a CGF-I or GA-I model, the  
7 estimated time based on the core model HI was within the time interval of the  
8 admixture, whereas all best-fit models were HI (Table 2). This result has  
9 indicated the limitation of using the GA and CGF models in inferring admixture  
10 history.

11           Using the same simulated admixed populations, we then employed GA-I,  
12 CGF-I and HI as core models for performing inference (Figs 3 and S2-S11). With  
13 HI, GA, or CGF considered as the true model, our estimation of the optimal model  
14 remained highly accurate. On the other hand, when the true model was GA-I or  
15 CGF-I, the failure rate decreased from 100% to 30%, compared to the estimation  
16 in the previous setting. Furthermore, the estimated time intervals were wider  
17 than those of the true ones, although the findings were still more accurate than  
18 those using GA and CGF as core models (Table 2).

## 19 **Empirical analysis**

20           We applied CAMer to the selected admixed populations from HapMap,  
21 HGDP, and 1KG. For each target population, we first used MALDmef to calculate  
22 the weighted LD and fitted the weighted LD with hundreds of exponential  
23 functions (Zhou *et al.* 2015). Next, with the weighted LD of target populations,  
24 we determined the admixture model and estimated admixture time with CAMer.

1 Quasi  $F$  and  $msE$  are designed for evaluating the inference with CAMer. The value  
2 of  $msE$  usually indicates data quality: small  $msE$  may indicate a high signal-to-  
3 noise ratio (SNR) and vice versa. The quasi  $F$  value measures the goodness of fit  
4 of the model we employed to fit the admixture event. A small  $F$  value indicates  
5 that the model we used was of satisfactory performance in modeling an  
6 admixture event. In our analysis, we used  $10^{-5}$  as the threshold for  $msE$  and 1.5  
7 as the threshold for  $F$ . Therefore, when the  $msE$  value  $\leq 10^{-5}$  and the  $F$  value  
8  $\leq 1.5$ , we could not “reject the null hypothesis” that the related model was the  
9 true model, i.e., the model well fitted the admixture event. On the other hand, an  
10  $msE$  value  $\geq 10^{-5}$  indicates low- quality data that is incapable of identifying the  
11 best-fit model, whereas an  $F$  value  $\geq 1.5$  prompts us to “reject the null  
12 hypothesis” and conclude that the model did not well fit the admixture. In the  
13 case of the same population from different databases, the data with smaller  $msE$   
14 values were given more credits. For example, we obtained samples of ASW from  
15 the HapMap and the 1KG. With the ASW data from HapMap, the best-fit model  
16 was HI of 6 generations, and both  $msE$  and  $F$  values indicated that the inference  
17 was acceptable (Fig S12). However, using the ASW data from 1KG, the best-fit  
18 models were CGF2-I of 1 to 9 generations and GA-I of 1 to 9 generations (Fig  
19 S13). A quasi  $F$  value of 2.12 indicated that neither the CGF2-I nor GA-I model  
20 satisfactorily fitted the admixture event. Because the  $msE$  value of the data set  
21 from 1KG was smaller, the conclusion using ASW was as follows: based on the  
22 best data we had, the time intervals estimated under the HI, GA-I, CGF1-I, and  
23 CGF2-I model were 6 generations, 1–9 generations, 1–13 generations, and 1–9  
24 generations, respectively. Furthermore, none of these models satisfactorily  
25 modeled the admixture, whereas the HI model and continuous models (CGF2-I

1 and GA-I) showed better performance. We also applied CAMer to other admixed  
2 populations (Table 3, Figs S14–17). MEX was satisfactorily modeled by the CGF  
3 model, with the estimated admixture time interval being 1–17 or 1–18  
4 generations, although the identification of the recipient population could not be  
5 determined. We also analyzed Eurasian populations, which showed that the  
6 Uygurs most likely fit a CGF model with Europeans as donors, with a gene flow  
7 lasting for 66 generations to the present. However, the value of  $msE$  was larger  
8 than  $10^{-5}$ , indicating that the results were not so reliable. The Hazara population  
9 experienced a GA-I-like admixture event that lasted for about 58 generations,  
10 which started 63 generations ago and ended approximately 5 generations ago.

## 11 **Discussion**

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

23 Our method, CAMer, can be utilized in inferring admixture histories by  
24 using weighted LD, which can be calculated using genotype data with MALDmef

1 (Zhou *et al.* 2015). Furthermore, we extended the GA and CGF models to the GA-I  
2 and CGF-I models in order to infer the time interval for a period of continuous  
3 admixture events followed by isolation. Even though CAMer was not consistently  
4 very accurate in determining the admixture model, its time interval estimations  
5 were reliable.

6 Two quantities, namely msE and quasi  $F$ , were used to evaluate the  
7 models inference. These two quantities should both be taken into consideration  
8 to identify the best-fit model(s) or the models well fit the admixture. Both the  
9 data quality and the goodness of fitting of models can affect the value of msE,  
10 although the  $F$  value mainly measures the goodness of modeling. Therefore, for  
11 convenience of interpretation, msE is considered to reveal the data quality and  $F$   
12 value is considered to evaluate the performance of the model. In our analysis, we  
13 suggested thresholds for msE and  $F$  to determine whether the null hypothesis  
14 should be reject or not, which may be too strict in empirical analysis. Actually,  
15 both msE and  $F$  values measure whether the observed weighted LD can be well  
16 fitted by the best-fit model(s). For example, the fitting process showed poor  
17 performance in the MKK population, which was accompanied by exaggerated  
18 msE and  $F$  values, indicating significant inconsistencies between the observed  
19 and fitted weight LD curves (Fig S17). Therefore, in empirical analysis, the msE  
20 value reflects the quality of the data, whereas  $F$  value describes the performance  
21 of the model.

22 In our previous study (Zhou *et al.* 2015), we fitted the weighted LD with  
23 hundreds of exponential functions. However, this approach did not fully reveal  
24 the occurrence of continuous admixture. To address this issue, the present study  
25 developed CAMer to model admixture as a continuous process. We have

1 observed that CAMer performed better than previous admixture model  
2 classification methods because it does not require conducting inferences using  
3 ancestry tracts, and thus can deal with missing genotypes in the data. CAMer also  
4 employed extensions of the classic continuous models, GA-I and CGF-I, which  
5 proved to be more flexible in modeling population admixture.

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

9

## 10 **Software**

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

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 Natural Science Foundation of China (NSFC) grants (91331204 and



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

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

9 **Acknowledgements:** None.

10

11

12

## 1 **Fig Legends**

2

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

11

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

15

16 Fig 3: Evaluation of CAMer under various simulated admixture models. Here, the  
17 core models are HI, GA-I, CGF1-I, and CGF2-I. The simulated models (True  
18 Model) are listed on the left, with the admixture time interval depicted in the  
19 parentheses. The gray area on the middle vertical panel is the simulated time  
20 interval, whereas colored lines indicate the estimated time intervals under  
21 different core models. HI: pink; CGF1-I: green; CGF2-I: purple; GA-I: blue.

22

23

1 **Table 1:** Adjusted p-values of pairwise paired t-test among core models: HI,  
 2 GA-I, CGF1-I, CGF2-I.

True Model	Best Model(s)	Adjusted p-Values of Pairwise Paired t-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	0.92	1	1	0.57	0.15	1
HI (50)	HI	1	1	1	0.87	0.55	0.87
CGF1 (1-100)	CGF1-I	$2.4 \times 10^{-13}$	$3.4 \times 10^{-14}$	$1.3 \times 10^{-13}$	$3.1 \times 10^{-9}$	$1.1 \times 10^{-8}$	$3.1 \times 10^{-9}$
CGF1 (1-50)	CGF1-I	$1.7 \times 10^{-12}$	$4.3 \times 10^{-13}$	$1.0 \times 10^{-12}$	$2.6 \times 10^{-9}$	$1.6 \times 10^{-6}$	$2.4 \times 10^{-9}$
GA (1-100)	GA-I	$1.2 \times 10^{-17}$	$6.9 \times 10^{-17}$	$3.0 \times 10^{-17}$	$3.2 \times 10^{-13}$	$1.7 \times 10^{-10}$	$6.9 \times 10^{-17}$
GA (1-50)	GA-I	$6.9 \times 10^{-15}$	$1.2 \times 10^{-15}$	$4.3 \times 10^{-14}$	$6.7 \times 10^{-10}$	$1.6 \times 10^{-4}$	$2.5 \times 10^{-9}$
CGF1-I (30-100)	CGF1-I, CGF2-I, GA-I	$1.1 \times 10^{-6}$	$8.4 \times 10^{-7}$	$8.4 \times 10^{-7}$	0.48	0.70	0.63
CGF1-I (70-100)	GA-I	$8.7 \times 10^{-4}$	$8.6 \times 10^{-3}$	0.017	0.027	$4.9 \times 10^{-4}$	0.027
GA-I (30-100)	CGF1-I	$1.6 \times 10^{-10}$	$2.1 \times 10^{-10}$	$1.6 \times 10^{-10}$	$8.9 \times 10^{-4}$	$5.3 \times 10^{-3}$	$1.7 \times 10^{-3}$
GA-I (70-100)	CGF1-I, CGF2-I, GA-I	0.036	0.076	0.18	0.52	0.076	0.13

3 Simulated true model is followed by the parenthesis of time interval for the  
 4 corresponding gene flow, where the first term in the parenthesis is the ending  
 5 time of the admixture and the second term is the beginning time of the  
 6 admixture. They are in the measurements of generation before present. For HI  
 7 model, only one time point included in the parenthesis.

8  
 9

10

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

True models	Core models	Counts			Rates		
		Correct	Undetermined	Wrong	Correct	Undetermined	Wrong
HI;GA;C GF	HI;GA;C GF	52	3	5	86.7%	5.0%	8.3%
GA-I; CGF-I	HI;GA;C GF	0	0	40	0.0%	0.0%	100.0%
HI;GA;C GF	HI;GA- I; CGF-I	47	11	2	78.3%	18.3%	3.4%
GA-I; CGF-I	HI;GA- I; CGF-I	6	22	12	15.0%	55.0%	30.0%

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

6

7

1 **Table 3: Results of CAMer on empirical populations**

Population	Best-fit model	End time	Start time	msE	Quasi.F
ASW-HapMap (57)	HI	NA	6	$2.72 \times 10^{-6}$	1.19
ASW-1KG (56)	CGF2-I	1	9	$1.83 \times 10^{-6}$	<u>2.12</u>
	GA-I	1	9	$1.86 \times 10^{-6}$	<u>2.12</u>
MEX (86)	CGF1-I	1	17	$3.56 \times 10^{-6}$	1.13
	CGF2-I	1	18	$3.57 \times 10^{-6}$	1.14
MKK (143)	GA-I	1	17	<u><math>1.97 \times 10^{-5}</math></u>	9.82
UIG (10)	CGF1-I	1	66	<u><math>4.01 \times 10^{-5}</math></u>	1.08
Hazara (24)	GA-I	5	63	$8.53 \times 10^{-6}$	1.30

2 Number of individuals listed in the parentheses. Values colored in red do not  
3 pass our threshold. The time interval is summarized from 22 jackknives, which is  
4 shared by more than half of all estimated intervals for continuous models or the  
5 nearest integer to the mean of estimated time point for HI model.  
6

7

8

9

1

2 Reference

3 Frazer K. a, Ballinger D. G., Cox D. R., Hinds D. a, Stuve L. L., Gibbs R. a, Belmont J.  
4 W., Boudreau A., Hardenbol P., Leal S. M., Pasternak S., Wheeler D. a, Willis T.  
5 D., Yu F., Yang H., Zeng C., Gao Y., Hu H., Hu W., Li C., Lin W., Liu S., Pan H.,  
6 Tang X., Wang J., Wang W., Yu J., Zhang B., Zhang Q., Zhao H., Zhao H., Zhou J.,  
7 Gabriel S. B., Barry R., Blumenstiel B., Camargo A., Defelice M., Faggart M.,  
8 Goyette M., Gupta S., Moore J., Nguyen H., Onofrio R. C., Parkin M., Roy J.,  
9 Stahl E., Winchester E., Ziaugra L., Altshuler D., Shen Y., Yao Z., Huang W.,  
10 Chu X., He Y., Jin L., Liu Y., Shen Y., Sun W., Wang H., Wang Y., Wang Y., Xiong  
11 X., Xu L., Waye M. M. Y., Tsui S. K. W., Xue H., Wong J. T.-F., Galver L. M., Fan J.-  
12 B., Gunderson K., Murray S. S., Oliphant A. R., Chee M. S., Montpetit A.,  
13 Chagnon F., Ferretti V., Leboeuf M., Olivier J.-F., Phillips M. S., Roumy S.,  
14 Sallée C., Verner A., Hudson T. J., Kwok P.-Y., Cai D., Koboldt D. C., Miller R. D.,  
15 Pawlikowska L., Taillon-Miller P., Xiao M., Tsui L.-C., Mak W., Song Y. Q., Tam  
16 P. K. H., Nakamura Y., Kawaguchi T., Kitamoto T., Morizono T., Nagashima A.,  
17 Ohnishi Y., Sekine A., Tanaka T., Tsunoda T., Deloukas P., Bird C. P., Delgado  
18 M., Dermitzakis E. T., Gwilliam R., Hunt S., Morrison J., Powell D., Stranger B.  
19 E., Whittaker P., Bentley D. R., Daly M. J., Bakker P. I. W. de, Barrett J.,  
20 Chretien Y. R., Maller J., McCarroll S., Patterson N., Pe'er I., Price A., Purcell S.,  
21 Richter D. J., Sabeti P., Saxena R., Schaffner S. F., Sham P. C., Varilly P.,  
22 Altshuler D., Stein L. D., Krishnan L., Smith A. V., Tello-Ruiz M. K., Thorisson  
23 G. a, Chakravarti A., Chen P. E., Cutler D. J., Kashuk C. S., Lin S., Abecasis G. R.,  
24 Guan W., Li Y., Munro H. M., Qin Z. S., Thomas D. J., McVean G., Auton A.,  
25 Bottolo L., Cardin N., Eyheramendy S., Freeman C., Marchini J., Myers S.,  
26 Spencer C., Stephens M., Donnelly P., Cardon L. R., Clarke G., Evans D. M.,  
27 Morris A. P., Weir B. S., Tsunoda T., Mullikin J. C., Sherry S. T., Feolo M., Skol  
28 A., Zhang H., Zeng C., Zhao H., Matsuda I., Fukushima Y., Macer D. R., Suda E.,  
29 Rotimi C. N., Adebamowo C. a, Ajayi I., Aniagwu T., Marshall P. a,  
30 Nkwodimmah C., Royal C. D. M., Leppert M. F., Dixon M., Peiffer A., Qiu R.,  
31 Kent A., Kato K., Niikawa N., Adewole I. F., Knoppers B. M., Foster M. W.,  
32 Clayton E. W., Watkin J., Gibbs R. a, Belmont J. W., Muzny D., Nazareth L.,  
33 Sodergren E., Weinstock G. M., Wheeler D. a, Yakub I., Gabriel S. B., Onofrio R.  
34 C., Richter D. J., Ziaugra L., Birren B. W., Daly M. J., Altshuler D., Wilson R. K.,  
35 Fulton L. L., Rogers J., Burton J., Carter N. P., Clee C. M., Griffiths M., Jones M.  
36 C., McLay K., Plumb R. W., Ross M. T., Sims S. K., Willey D. L., Chen Z., Han H.,  
37 Kang L., Godbout M., Wallenburg J. C., L'Archevêque P., Bellemare G., Saeki K.,  
38 Wang H., An D., Fu H., Li Q., Wang Z., Wang R., Holden A. L., Brooks L. D.,  
39 McEwen J. E., Guyer M. S., Wang V. O., Peterson J. L., Shi M., Spiegel J., Sung L.  
40 M., Zacharia L. F., Collins F. S., Kennedy K., Jamieson R., Stewart J., 2007 A  
41 second generation human haplotype map of over 3.1 million SNPs. *Nature*  
42 **449**: 851–861.

43 Gravel S., 2012 Population genetics models of local ancestry. *Genetics* **191**: 607–  
44 619.

- 1 Holm S., 1979 A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J.*  
2 *Stat.* **6**: 65–70.
- 3 Jin W., Wang S., Wang H., Jin L., Xu S., 2012 Exploring population admixture  
4 dynamics via empirical and simulated genome-wide distribution of  
5 ancestral chromosomal segments. *Am. J. Hum. Genet.* **91**: 849–862.
- 6 Jin W., Li R., Zhou Y., Xu S., 2013 Distribution of ancestral chromosomal segments  
7 in admixed genomes and its implications for inferring population history  
8 and admixture mapping. *Eur. J. Hum. Genet.* **22**: 930–937.
- 9 Li J. Z., Absher D. M., Tang H., Southwick A. M., Casto A. M., Ramachandran S.,  
10 Cann H. M., Barsh G. S., Feldman M., Cavalli-Sforza L. L., Myers R. M., 2008  
11 Worldwide human relationships inferred from genome-wide patterns of  
12 variation. *Science* **319**: 1100–1104.
- 13 Loh P. R., Lipson M., Patterson N., Moorjani P., Pickrell J. K., Reich D., Berger B.,  
14 2013 Inferring admixture histories of human populations using linkage  
15 disequilibrium. *Genetics* **193**: 1233–1254.
- 16 Ni X., Yang X., Guo W., Yuan K., Zhou Y., Ma Z., Xu S., 2015 Length distribution of  
17 ancestral tracts under a general admixture model and its applications in  
18 admixture history inference. Under Rev.
- 19 Pfaff C. L., Parra E. J., Bonilla C., Hiester K., McKeigue P. M., Kamboh M. I.,  
20 Hutchinson R. G., Ferrell R. E., Boerwinkle E., Shriver M. D., 2001 Population  
21 structure in admixed populations: effect of admixture dynamics on the  
22 pattern of linkage disequilibrium. *Am. J. Hum. Genet.* **68**: 198–207.
- 23 Pickrell J. K., Patterson N., Loh P.-R., Lipson M., Berger B., Stoneking M.,  
24 Pakendorf B., Reich D., 2014 Ancient west Eurasian ancestry in southern  
25 and eastern Africa. *Proc. Natl. Acad. Sci. U. S. A.* **111**: 2632–7.
- 26 Price A. L., Tandon A., Patterson N., Barnes K. C., Rafaels N., Ruczinski I., Beaty T.  
27 H., Mathias R., Reich D., Myers S., 2009 Sensitive detection of chromosomal  
28 segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**.
- 29 Qin P., Zhou Y., Lou H., Lu D., Yang X., Wang Y., Jin L., Chung Y.-J., Xu S., 2015  
30 Quantitating and Dating Recent Gene Flow between European and East  
31 Asian Populations. *Sci. Rep.* **5**: 9500.
- 32 R Core Team, 2014 R: A Language and Environment for Statistical Computing. **0**.
- 33 The 1000 Genomes Project Consortium, 2012 An integrated map of genetic  
34 variation from 1,092 human genomes. *Nature* **135**: 0–9.
- 35 Xu S., Jin L., 2008 A Genome-wide Analysis of Admixture in Uyghurs and a High-  
36 Density Admixture Map for Disease-Gene Discovery. *Am. J. Hum. Genet.* **83**:  
37 322–336.

- 1 Zhou Y., Yuan K., Yu Y., Ni X., Xie P., Xing E. P., Xu S., 2015 Inference of multiple-  
2 wave population admixture by modeling decay of linkage disequilibrium  
3 with multiple exponential functions. Under Rev.
- 4 Zhu X., Cooper R. S., Elston R. C., 2004 Linkage analysis of a complex disease  
5 through use of admixed populations. *Am. J. Hum. Genet.* **74**: 1136–1153.
- 6 Zhu X., Cooper R. S., 2007 Admixture mapping provides evidence of association  
7 of the VNN1 gene with Hypertension. *PLoS One* **2**.
- 8



# HI

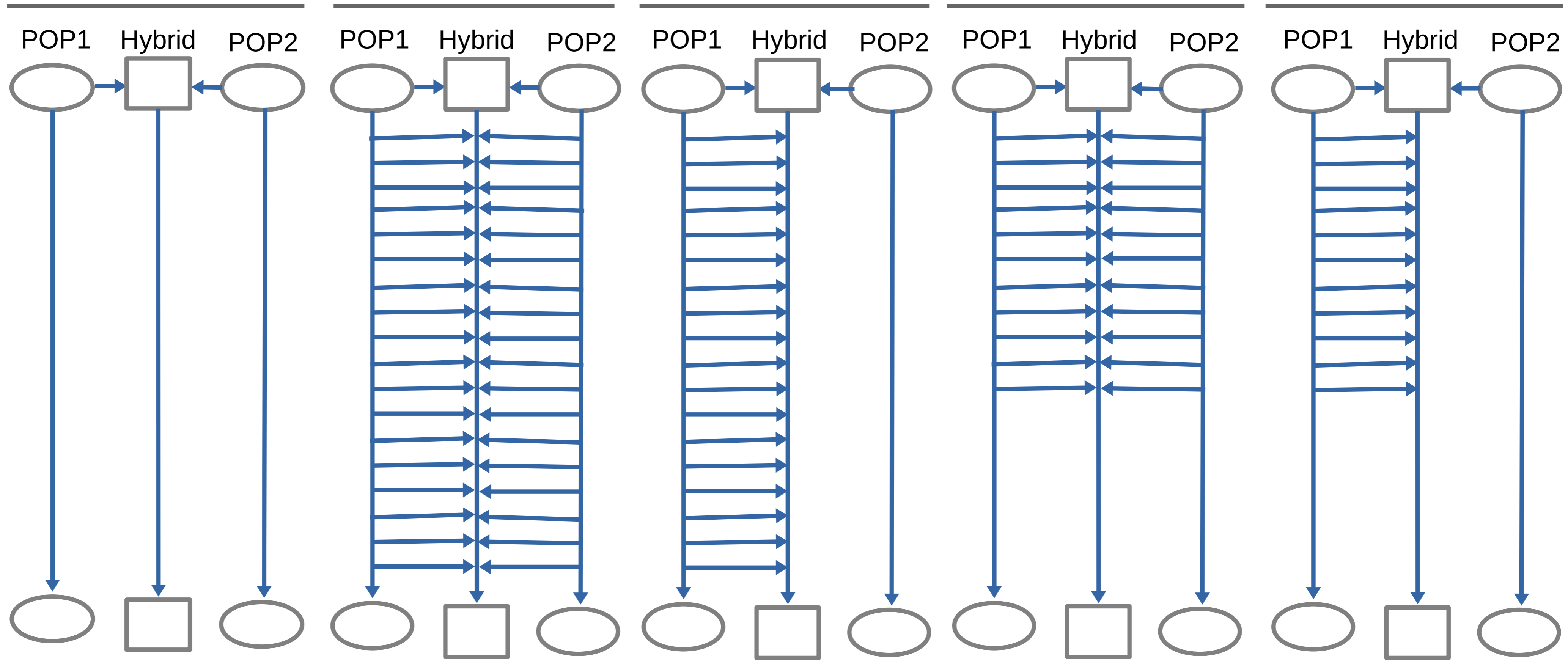
# GA

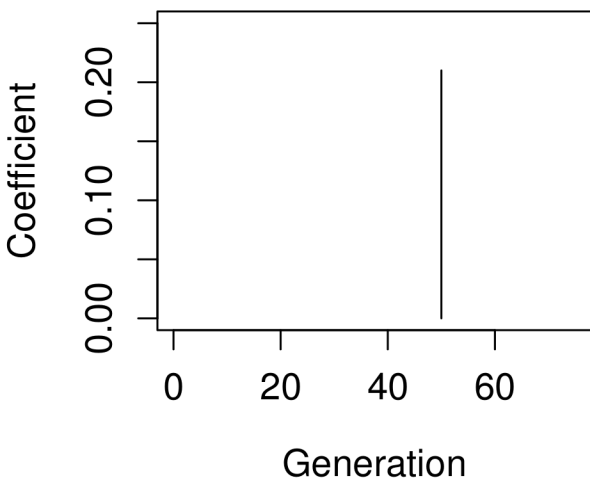
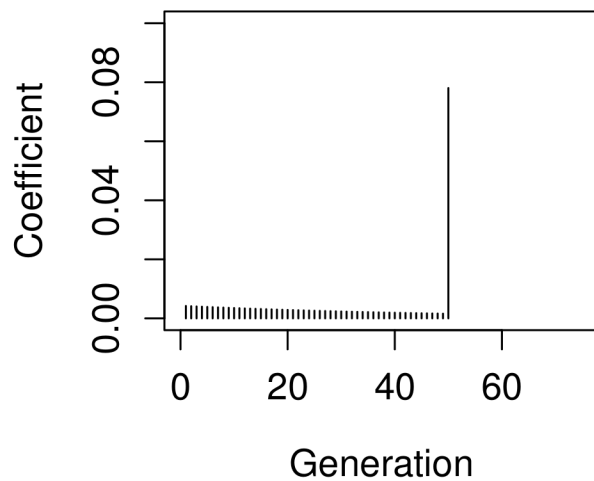
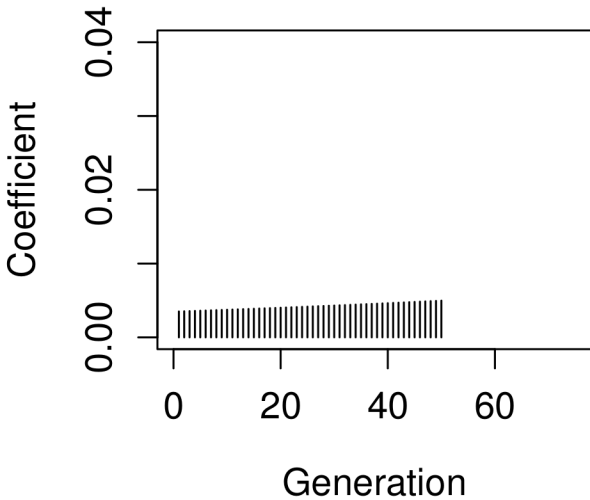
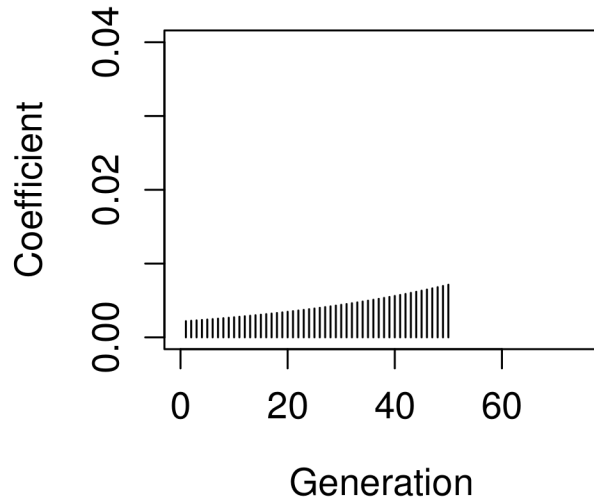
# CGF

# GA-I

# CGF-I

Past  
Present



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

**True Model**

**Time Intervals**

**Best Model(s)**

HI (100)

HI

HI (50)

HI

CGF1 (1-100)

CGF1-I

CGF1 (1-50)

CGF1-I

GA (1-100)

GA-I

GA (1-50)

GA-I

CGF1-I (30-100)

CGF1-I,CGF2-I,GA-I

CGF1-I (70-100)

GA-I

GA-I (30-100)

CGF1-I

GA-I (70-100)

CGF1-I,CGF2-I,GA-I

0 50 100 150 200

Generation

