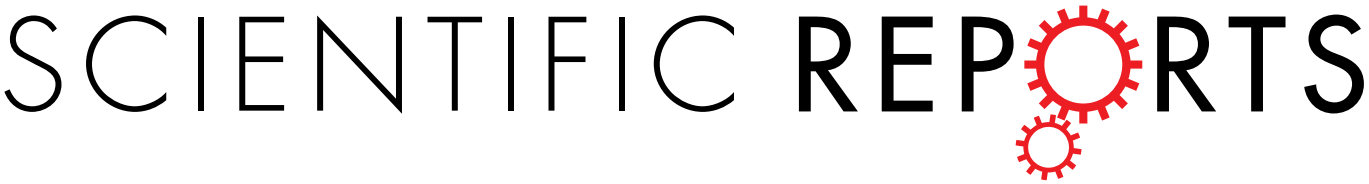

OPEN

Modeling Continuous Admixture Using Admixture-Induced Linkage Disequilibrium

Received: 31 October 2016
Accepted: 18 January 2017
Published: 23 February 2017

Ying Zhou[1,2,*], Hongxiang Qiu[1,3,*] & Shuhua Xu[1,2,4,5]

Recent migrations and inter-ethnic mating of long isolated populations have resulted in genetically admixed populations. To understand the complex population admixture process, which is critical to both evolutionary and medical studies, here we used admixture induced linkage disequilibrium (LD) to infer 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 scenario with isolation after a certain duration of continuous gene flow. Based on the new models, we developed a method, CAMer, to infer the admixture history considering continuous and complex demographic process of gene flow between populations. We evaluated the performance of CAMer by computer simulation and further applied our method to real data analysis of a few well-known admixed populations.

Human migrations involve gene flow among previously isolated populations, resulting in 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 depends largely on the precision of the applied admixture models. Several methods have been developed to estimate admixture time based on the hybrid isolation (HI) model[1–4] or intermixture admixture model (IA)[5], which assume that the admixed population is formed by one wave of admixture at a certain time. However, the one-wave assumption often leads to under-estimation when the progress of the true admixture cannot be well modeled by the HI model. Jin *et al.* showed earlier that under the assumption of HI, the estimated time is half of the true time when the true model is a modified gradual admixture (GA) model[6].

Admixture models can be theoretically distinguished by comparing the length distribution of continuous ancestral tracts (CAT)[7–9], which refers to continuous haplotype tracts that were deviated from the same ancestral population. CAT inherently represents admixture history as it accumulates recombination events. Short CAT always indicates long admixture history of the same admixture proportion, whereas long CAT may indicate a recent gene flow from the ancestral population to which the CAT belongs. Based on the information it provides, CAT can be used to distinguish different admixture models and estimate corresponding admixture time. However, accurately estimating the length of CAT is often very difficult.

Weighted linkage disequilibrium (LD) is an alternative type of information that can be used to infer admixture[1,10]. Previous studies have indicated that it is more efficient than CAT because it requires neither ancestry inference nor haplotype phasing, which often introduces false recombination thus decreasing the power of estimation. Weighted LD has already been used in inferring multiple-wave admixtures[10,11]. However, these methods tend to summarize the admixture into different independent events, even if the true admixture is continuous. In our previous work[11,12], we mathematically described LD under different continuous models, allowing us to determine admixture history using these models.

[1]Chinese Academy of Sciences (CAS) Key Laboratory of Computational Biology, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China. [2]University of Chinese Academy of Sciences, Beijing 100049, China. [3]Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong, China. [4]School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China. [5]Collaborative Innovation Center of Genetics and Development, Shanghai 200438, China. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to S.X. (email: xushua@picb.ac.cn)

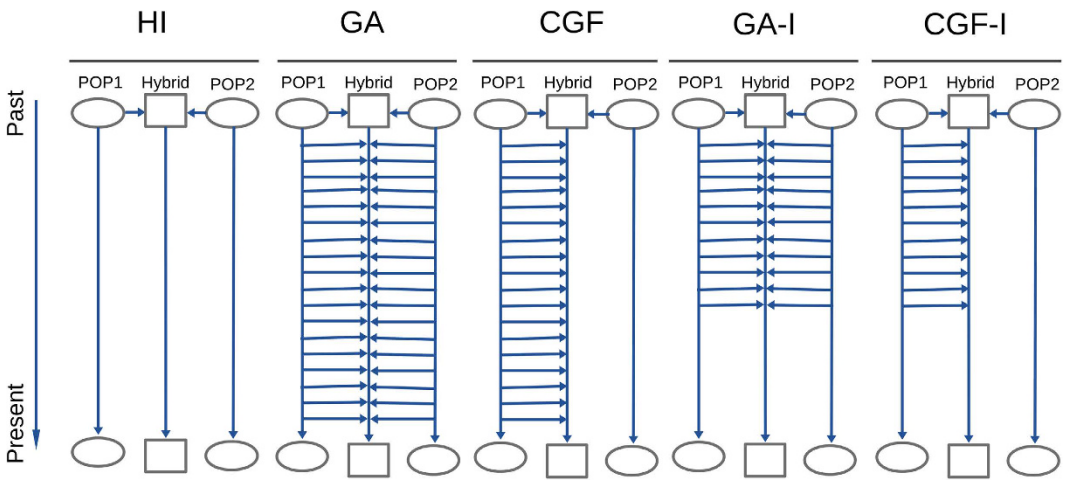

Figure 1. Classic admixture models (HI, GA and CGF) and the models we extended (GA-I and CGF-I). For each model, the simulated admixed population (Hybrid) is in the middle of two source populations (POP1 and POP2). Each horizontal arrow represents the direction of gene flow from the source populations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

In the present study, we first developed a weighted LD-based method to infer admixture with HI, GA, and continuous gene flow (CGF)[13] models (see Fig. 1). Both GA and CGF models assume that gene flow is a continuous process. Next, we extended the GA and CGF models to GA-I and CGF-I models, respectively (see Fig. 1), which models a scenario with a continuous gene flow duration followed by a period of isolation to present. We applied our method to a number of well-known admixed populations and provided information that would help better understanding the admixture history of these populations.

Material and Methods

Datasets. Data for simulation and empirical analysis were obtained from three public resources: Human Genome Diversity Panel (HGDP)[14], the International HapMap Project phase III[15] and the 1000 Genomes Project (1KG)[16]. Source populations for simulations are the haplotypes from 113 Utah residents with Northern and Western European ancestries from the CEPH collection (CEU) and 113 Africans from Yoruba (YRI).

Inferring Admixture Histories by Using HI, GA, and CGF Models. The expectation of weighted LD under a two-way admixture model has been described in detail in another work[11]. Following the previous notation, the expectation of weighted LD statistic between two sites separated by a distance d (in Morgan) is as follows:

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

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

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

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

$$z(d) = \mathrm{Poly}(d)^T C;$$

where

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

and

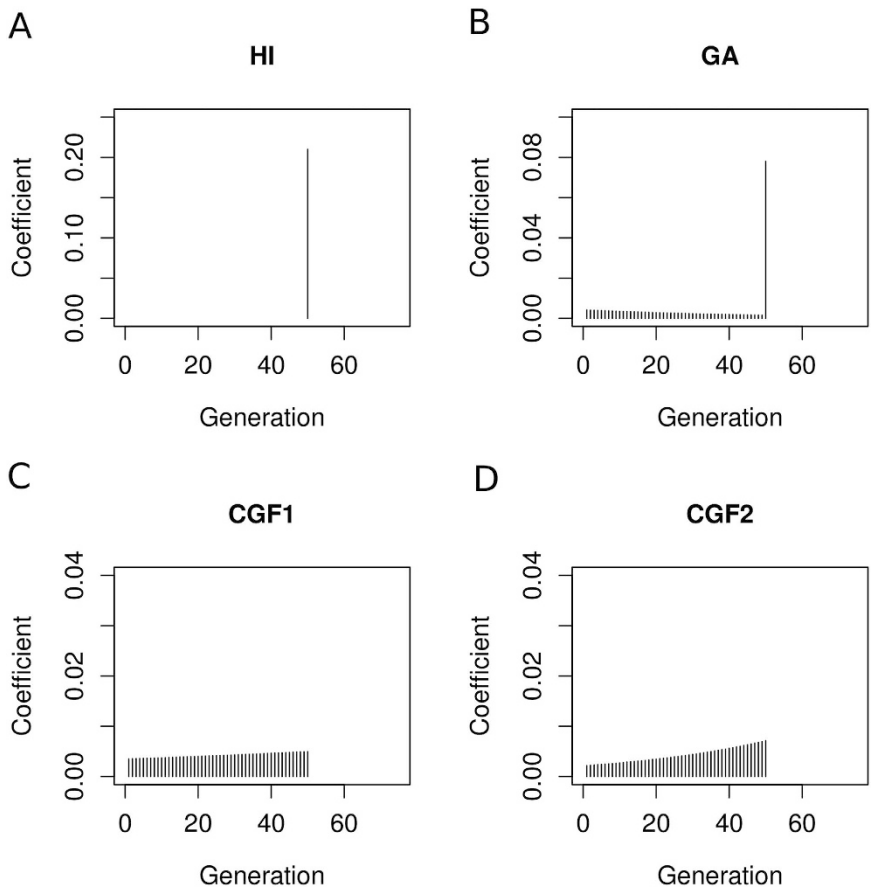

Figure 2. Coefficient vector of polynomial functions for each model. For each admixture model, the starting time of the population admixture is 50 generations ago. The admixture proportion in the final admixed population is set as 0.3:0.7 for population 1 and population 2.

$$\mathrm{Poly}(d) = ((1-d),\ \ldots, (1-d)^{n-1}, (1-d)^n)^T.$$

For different admixture models where admixture began $n$generations ago, $z(d)$ varies in terms of the vector of coefficients of polynomial functions[12]:

$$\begin{aligned}
&\mathrm{HI} \quad C_{\mathrm{HI}} = (0, \ldots, 0, m_1 m_2)^T \\
&\mathrm{GA} \quad C_{\mathrm{GA}} = m_1 m_2 \left(\frac{(n-1)^0}{n}, \frac{(n-1)^1}{n^2}, \ldots, \frac{(n-1)^{n-2}}{n^{n-1}}, \frac{(n-1)^{n-1}}{n^{n-1}} \right)^T \\
&\mathrm{CGF1} \quad C_{\mathrm{CGF1}} = (1 - {m_1}^{1/n}) m_1 ({m_1}^{(n-1)/n}, {m_1}^{(n-2)/n}, \ldots, 1)^T \\
&\mathrm{CGF2} \quad C_{\mathrm{CGF2}} = (1 - {m_2}^{1/n}) m_2 ({m_2}^{(n-1)/n}, {m_2}^{(n-2)/n}, \ldots, 1)^T
\end{aligned}$$

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

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

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

The optimization problem is therefore expressed as follows:

$$\min_{\theta_0, \theta_1, C_{\mathrm{model}}} \mathrm{ssE}(\theta_0, \theta_1, C_{\mathrm{model}}), \tag{3}$$

where $Z = (z(d_1), z(d_2), \ldots, z(d_I))^T$ is the observed ALD calculated from the single nucleotide polymorphism (SNP) data of both the parental populations and the admixed population, both Z and admixture proportion m_i can be calculated by the algorithm iMAAPs[12]; θ_0 is a real number used to correct the population substructure; θ_1 is a scalar that improves estimation robustness; $\mathbf{1} \in R^I$ is a vector with each entry being 1; A is an $I \times J$ matrix with the ith row vector defined as $\text{Poly}(d_i)^T$, i.e., $A = (\text{Poly}(d_1), \text{Poly}(d_2), \ldots, \text{Poly}(d_I))^T$, and $J \geq n$ is a pre-specified upper bound of n. Our definitions are consistent since we can let all entries after the n-th entry be 0 in C_{model}.

Next, we tried to estimate the parameters θ_0, θ_1, and C_{model}, where C_{model} has the information of the admixture model and the related admixture time n (in generations). In our analysis, the value of n is assumed to be a positive integer; therefore, our method is to go through all possible n values (with a reasonable upper limit J) to estimate n with the minimum value of the objective function. Given n, we used the ordinary least squares method to estimate (θ_0, θ_1) such that the objective function was minimized. Using this approach, the value of n in relation to the minimal objective function value for each model was determined, which represents the time of admixture occurrence under each model. The method to conclude which models are the best is described in Identification of the best-fit model session.

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

$$\text{GA-IC}_{\text{GA-I}} = m_1 m_2 \left(0, \ldots, 0, \frac{(n-1)^0}{n}, \frac{(n-1)^1}{n^2}, \ldots, \frac{(n-1)^{n-2}}{n^{n-1}}, \frac{(n-1)^{n-1}}{n^{n-1}}\right)^T$$

$$\text{CGF1-IC}_{\text{CGF1-I}} = (1 - {m_1}^{1/n}) m_1 (0, \ldots, 0, {m_1}^{(n-1)/n}, {m_1}^{(n-2)/n}, \ldots, 1)^T$$

$$\text{CGF2-IC}_{\text{CGF2-I}} = (1 - {m_2}^{1/n}) m_2 (0, \ldots, 0, {m_2}^{(n-1)/n}, {m_2}^{(n-2)/n}, \ldots, 1)^T$$

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

Let E and S be the ending and starting time points (in generations, prior to the present) of the admixture, which we wanted to search for to minimize the objective function. The search starts from $(E^0, S^0) = (1, J)$, where J is the upper bound for the beginning of the admixture event, which can be set to be a large integer to seek for a relatively ancient admixture event. In our analysis of recent admixed populations, we set $J = 500$. For $k = 1, 2, \ldots$, (E^k, S^k) is updated from (E^{k-1}, S^{k-1}) by two alternative proposals denoted by (E_1^k, S_1^k) and (E_2^k, S_2^k). For convenience, we defined

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

where θ_0, θ_1 can be determined by ordinary least squares.

We chose the proposal that resulted in a smaller value for f. The search stopped when the value of f with (E^{k-1}, S^{k-1}) was no larger than that of either proposal or $E^k = S^k$. In this way, we could readily estimate the time interval of the admixture event $(G_{\text{end}}, G_{\text{start}})$ quickly.

Algorithm 1:
for k in **1, 2,** …
$(E_1^k, S_1^k) := (E^{k-1} + \mathbf{1}, S^{k-1})$
$(E_2^k, S_2^k) := (E^{k-1}, S^{k-1} - \mathbf{1})$
$(E^k, S^k) := \underset{(E,S) \in \left\{\left(E_1^k, S_1^k\right), \left(E_2^k, S_2^k\right), \left(E^{k-1}, S^{k-1}\right)\right\}}{\textbf{argmin}} f(E, S)$
if $(E^k, S^k) = (E^{k-1}, S^{k-1})$ ***or*** $E^k = S^k$
$(G_{\text{end}}, G_{\text{start}}) := (E^k, S^k)$
stop

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

(1) Let $e = \theta_0 \cdot 1 + \theta_1 A C_{\text{model}} - Z$. We looked at $\text{msE} = \frac{\sum_1^I e_i^2}{I-1}$ with e_i being the ith entry of e. This reflects goodness of fit and strength of background noise. A smaller msE indicates less background noise and better fit.

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	1	0.20	1	0.068	1	0.059
HI (50)	HI	0.054	0.83	0.83	0.049	0.023	0.83
CGF1 (1–100)	CGF1-I, CGF2-I	0.012	0.012	0.012	0.055	0.018	0.28
CGF1 (1–50)	CGF1-I, CGF2-I, GA-I	0.012	0.012	0.012	0.074	0.018	1
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	0.012	0.084
CGF1-I (30–100)	CGF1-I	0.049	0.049	0.049	0.035	0.43	0.049
CGF1-I (70–100)	HI	0.70	1	1	0.70	0.012	0.19
GA-I (30–100)	HI	0.049	0.15	0.15	0.15	0.012	0.15
GA-I (70–100)	HI	1	1	1	0.22	0.12	1

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

(2) Let $e' = \hat{Z} - Z$, where $\hat{Z}$ is the fitted weighted LD obtained from iMAAPs, which theoretically can be regarded as the de-noised weighted LD. $e^{'}$ is a vector of length I, with the ith entry denoted by e_i'. We looked at the quasi-F statistic $F = \frac{\sum_1^I e_i^2}{\sum_1^I (e_i')^2}$. A small F indices that the current fit does not significantly deviate from the previous fit.

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

Identification of the best-fit model. For the convenience of illustration, we defined the core model as the model used to infer admixture time. When inferring admixture of a target population, HI, GA, CGF1, CGF2, GA-I, CGF1-I and CGF2-I are used as the core models for conducting inference. Because GA-I, CGF1-I and CGF2-I describe more general admixture models than GA, CGF1, and CGF2, we classified model selection into two cases: one case is to identify the best-fit model(s) among the HI, GA, CGF1, and CGF2 models, whereas the more general case is to determine the best-fit model(s) among HI, GA-I, CGF1-I and CGF2-I models. In both cases, the same strategy is adopted, which depends on the pairwise paired difference of pseudo log(msE) values associated with each core model, which will be defined later. For an admixed population, there are $N+1$ observed weighted LD curves obtained as follows: N (typically 22) autosomal chromosomes are considered in an individual genome, and one weighted LD curve is calculated from all these N chromosomes while the other N weighted LD curves are obtained by jackknife resampling, leaving out one chromosome for each LD curve[1,10,11]. Next, we fit each observed weighted LD curve for each core model by estimating θ_0, θ_1 and the time interval, which in turn allowed us to obtain the msE value associated with the optimal parameters for each weighted LD curve. Taken together, a total of $N+1$ msE values associated with $N+1$ LD curves were evaluated in each core model. For model M, the log(msE) obtained from all N chromosomes was denoted by ϵ_0^M and that from the LD curve with the q-th chromosome was left out by $\epsilon_q^M (q = 1, \ldots, N)$. Following Tukey[17], we defined the q-th pseudo log(msE) for model M to be $\hat{\epsilon}_q^M = N\epsilon_0^M - (N-1)\epsilon_q^M$ and treated these pseudo values approximately as independent. Next, we defined the best-fit core model(s) to be the model(s) with significantly small $\hat{\epsilon}_q^M$. A pairwise Wilcoxon signed-rank test was conducted for the pseudo log(msE) of the four models. More precisely, Wilcoxon signed-rank test was applied to all pairs of models with the $\hat{\epsilon}_q^M$ being paired by index q, and then the p-values were adjusted to control family-wise error rate (see Table 1). We used the Holm-Bonferroni method to adjust p-values[18]. When $\hat{\epsilon}_q^{\mathrm{HI}}$ was not significantly larger than those of the best model, i.e., the model associated with the smallest sample median of pseudo log(msE) values, HI was selected because HI is a simpler model compared with the others. Otherwise, the models whose $\hat{\epsilon}_q^M$ was not significantly larger than those of the best model were selected (the best model was selected as well). The significance level was set to be 0.05. Here, we paired the pseudo values according to index q and used Wilcoxon signed-rank test on the paired differences. $\hat{\epsilon}_q^M$ is strongly correlated with q and hence q is a major covariate that must be controlled in the test to gain higher power. This is also the reason that even though theoretically there are examples where the best model, according to our definition, can be significantly worse than other models in our process, we still use this method considering that such extreme cases are unlikely in practice. In addition, log(msE) rather than msE was used because after logarithm transformation, the small values of msE could also have huge effect to the comparison. That is to say, we could better detect the

True models	Core models	Counts			Rates		
		Correct	Undetermined	Wrong	Correct	Undetermined	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	30	29	1	50.0%	48.3%	1.7%
GA-I;CGF-I	HI;GA-I;CGF-I	3	11	26	7.5%	27.5%	65%

Table 2. Accuracy of model determination. Here, as our method can hardly distinguish CGF1 from CGF2 model, we regard CGF1, CGF2 as the CGF model; CGF1-I and CGF2-I as the CGF-I model, which are different from GA-I and HI models. Here, "correct" denotes the best-fit model is the true model; "Undetermined" means the true model can not be determined from the best-fit models; "Wrong" denotes the true model is not given.

difference between small msE, thus gaining greater power in the test. This claim is also justified by our experience. In Table 1, we listed the adjusted p-values to determine the best-fit model(s) under various scenarios. In the simulation of HI (100), HI model was inferred as the best-fit model because $\hat{\epsilon}_q^{\mathrm{HI}}$ is not significantly larger than $\hat{\epsilon}_q^{\mathrm{GA\text{-}I}}$, $\hat{\epsilon}_q^{\mathrm{CGF1\text{-}I}}$, and $\hat{\epsilon}_q^{\mathrm{CGF2\text{-}I}}$. In the cases of CGF1 (1–50), GA-I, CGF1-I, and CGF2-I were inferred as the set of best-fit models because we cannot distinguish the best fit model from GA-I, CGF1-I, and CGF2-I models. This case was marked as "Undetermined" in Table 2.

Software

Our algorithm has been implemented in an R package[19], named CAMer (Continuous Admixture Modeler). The package is available on the website of population genetic group: http://www.picb.ac.cn/PGG/resource.php or on Github: https://www.github.com/david940408/CAMer.

Results

Simulation studies. Admixed populations were simulated in a forward-time way under different admixture models with the software **AdmixSim**[20], which is under the framework of copying model that new haplotypes are assembled from the segments of the source populations' haplotypes generation by generation[4,21], and the same simulation strategy has been used in the previous work[4]. Simulation was initiated with the haplotypes from source populations (YRI and CEU) and the haplotypes for the admixed population were generated by resampling haplotypes with recombination from source populations and the admixed population of last generation. During the simulation, population size was kept as 5000 and migration rate was controlled by the admixture model with the final admixture proportion in the admixed population to be 0.3. We employed a uniform recombination map in our simulation, which means recombination rate between two markers is positively proportional to their physical distance. For each model, simulation was performed using 10 replicates; each replicate contained 10 chromosomes with a total length of 3 Morgans. To evaluate the performance of our algorithm, we simulated admixed populations under the following conditions:

(1) HI of 50 and 100 generations, designated as HI (50) and HI (100),
(2) GA of 50 and 100 generations, designated as GA (1–50) and GA (1–100), respectively,
(3) CGF of 50 and 100 generation, population 1 as the recipient, designated as CGF1 (1–50) and CGF1 (1–100) respectively,
(4) CGF-I of a 70-generation admixture followed by 30-generation isolation, and a 30-generation admixture followed by a 70-generation isolation, with population 1 as the recipient, designated as CGF1-I (30–100) and CGF1-I (70–100) respectively, and,
(5) GA-I of a 70-generation admixture followed by a 30-generation isolation and a 30-generation admixture followed by a 70-generation isolation, designated as GA-I (30–100) and GA-I (70–100), respectively.

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

With the same set of simulated admixed populations, we also used **AdmixInfer**[9] to determine the admixture model and estimate admixture time, which is based on the length distribution of CAT. To avoid any errors introduced by haplotype phasing and local ancestry inference, we analyzed the ancestral segments generated from **AdmixSim**. We found that **AdmixInfer** attained pretty accuracy in determining the admixture model and estimating admixture time when the simulation is under HI, GA, or CGF model. However, it could only give HI model as the best-fit model when the simulated admixture is under GA-I or CGF-I model. (see Supplementary Table S2) These results indicated the limitation of using the GA and CGF models in inferring admixture history, no matter the information from LD or CAT is used for inference.

We next employed GA-I, CGF-I and HI as core models for performing inference (see Fig. 3 and Supplementary Figs S2–11). With HI, GA, or CGF being considered as the true model, our estimation of the optimal model remained accurate. On the other hand, when the true model was GA-I or CGF-I, the failure rate decreased by 35%, compared to the estimation in the previous setting, but it was still at a very high level.

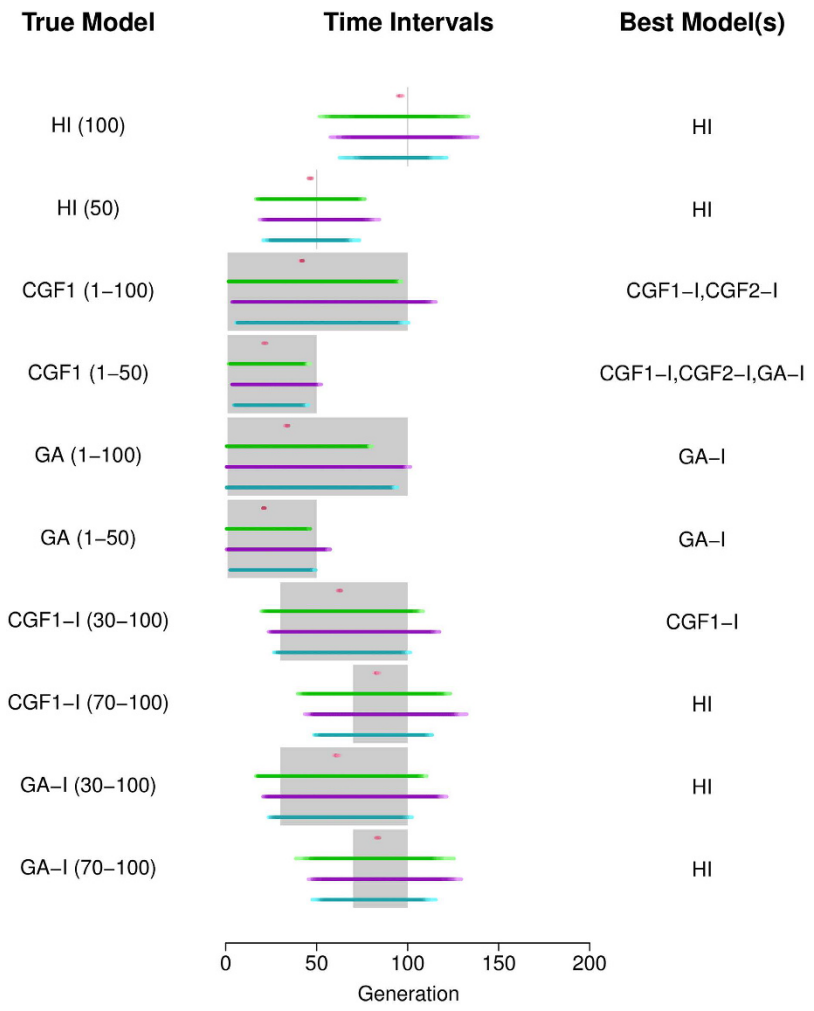

Figure 3. Evaluation of CAMer under various simulated admixture models. Here, the core models are HI, GA-I, CGF1-I, and CGF2-I. The simulated models (True Model) are listed on the left, with the admixture time interval depicted in the parentheses. The gray area on the middle vertical panel is the simulated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

Furthermore, the estimated time intervals were wider than those of the true ones, although the results were still more accurate than those using GA and CGF as core models (see Table 2).

By introducing the GA-I and CGF-I models as core models, CAMer can resolve the admixture into continuous time interval. Considering that CAMer is not so powerful in determining the best-fit admixture model (see Tables 1 and 2), in empirical studies, we presented the results from CAMer with estimations on all core models and the model(s) fitting best the data.

Empirical analysis. We applied CAMer to the selected admixed populations from HapMap, HGDP, and 1KG. For each target population, we first used iMAAPs to calculate the weighted LD and fit the weighted LD decay curve with a numeric method[11]. Next, with the weighted LD of target populations, we determined the admixture model and estimated admixture time with CAMer. Quasi F and msE are designed for evaluating the inference with CAMer. The value of msE usually indicates data quality: small msE may indicate a high signal-to-noise ratio (SNR) and vice versa. The quasi F value measures the goodness of fit of the model we employed to fit the admixture event. A small F value indicates that the model we used was of satisfactory performance in modeling an admixture event. In our analysis, we used 10^{-5} as the threshold for msE and 1.5 for F. Therefore, when the msE value $\leq 10^{-5}$ and the F value ≤ 1.5, we could not "reject the null hypothesis" that the related model was the true model, i.e., the model well fit the admixture event. On the other hand, an msE value $\geq 10^{-5}$ indicates low-quality data that is incapable of identifying the best-fit model, whereas a F value ≥ 1.5 prompts us to "reject the null hypothesis" and concludes that the model does not well fit the admixture. In the case of the same population from different databases, the data with smaller msE values were given more credits. For example, we obtained samples of ASW from the HapMap and the 1KG. With the

Population	Core model	End time	Start time	msE	Quasi.F
ASW-HapMap (57)	HI	5	5	3.44×10^{-6}	1.60
	CGF1-I	1	10	2.87×10^{-6}	1.34
	CGF2-I	1	8	2.47×10^{-6}	1.15
	GA-I*	2	8	2.51×10^{-6}	1.17
ASW-1KG (56)	HI	5	5	4.12×10^{-6}	4.93
	CGF1-I*	1	11	1.96×10^{-6}	2.34
	CGF2-I*	1	9	2.17×10^{-6}	2.60
	GA-I*	2	9	2.04×10^{-6}	2.44
MEX (86)	HI	8	8	1.05×10^{-5}	3.52
	CGF1-I	1	17	3.74×10^{-6}	1.25
	CGF2-I	1	17	3.60×10^{-6}	1.20
	GA-I*	2	15	3.50×10^{-6}	1.17
MKK (143)	HI*	5	5	2.57×10^{-5}	12.66
	CGF1-I	1	19	2.04×10^{-5}	10.24
	CGF2-I	1	12	2.15×10^{-5}	10.82
	GA-I	1	23	1.99×10^{-5}	9.78
Uyghur (10)	HI	26	26	4.65×10^{-5}	1.31
	CGF1-I*	1	65	3.85×10^{-5}	1.08
	CGF2-I*	1	63	3.85×10^{-5}	1.08
	GA-I*	2	64	3.88×10^{-5}	1.09
Hazara (24)	HI	26	26	1.28×10^{-5}	2.05
	CGF1-I	2	70	8.52×10^{-6}	1.35
	CGF2-I	2	65	8.61×10^{-6}	1.37
	GA-I*	4	64	8.19×10^{-6}	1.30

Table 3. Results of CAMer on empirical populations. Number in parentheses denotes the sample size for each population. 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 HI model, the beginning time is the same as the ending time.

ASW data (CEU and YRI as source populations) from HapMap, the best-fit model was GA-I of 2–8 generations, and both msE and F values indicated that the inference was acceptable (see Supplementary Fig. S12). Similarly, using the ASW data (CEU and YRI as source populations) from 1KG, the best-fit model failed to be determined among GA-I, CGF1-I, and CGF2-I (see Supplementary Fig. S13). However, all the quasi F values bigger than 1.5 indicated that these models did not satisfactorily fit the admixture event. Because the msE value of the data set from 1KG was smaller, the conclusion using ASW was as follows: based on the best data we had, the time intervals estimated under the HI, GA-I, CGF1-I, and CGF2-I model were 5 generations, 2–9 generations, 1–11 generations, and 1–9 generations, respectively. Furthermore, none of these models satisfactorily modeled the admixture, whereas the HI model showed better performance. We also applied CAMer to other admixed populations (see Table 3, Supplementary Figs S14–17). MEX (source populations: CEU [n = 64] and American Indian including 7 Colombians, 14 Karitiana, 21 Maya, 14 Pimas and 8 Suruis) was satisfactorily modeled by the GA-I model, with the estimated admixture time interval being 2–15 generations, respectively. We also analyzed Eurasian populations, which showed that the Uyghurs (source populations: Han [n = 34] and French [n = 28]) most likely fit a continuous model, with a gene flow lasting for more than 60 generations to the present or near present. We cannot determine which model fits best. However, the values of msE were all larger than 10^{-5}, indicating that the results were not so reliable. The Hazara population (source populations: Han [n = 34] and French [n = 28]) experienced a GA-I-like admixture event that lasted for about 60 generations, which started 64 generations ago and ended approximately 4 generations ago. It seemed that CAMer failed to reconstruct the admixture history of population MKK (Maasai in Kinyawa, Kenya), giving extreme msE and quasi F values.

Discussion

Modeling the demographic history of an admixed population and estimating time points of admixture event are essential components of evolutionary and medical research studies[5–9,11,22]. Previous methods have employed the length distribution of ancestral tracts[6–8], which highly depends on the accuracy of local ancestral inference and haplotype phasing. Another limitation is that only HI, GA, and CGF models were utilized to fit the admixture as well as in identifying the best-fit model. In the present study, our simulations showed that when the true model was not HI, GA, or CGF, the generated inferences were relatively difficult to interpret.

Our method, CAMer, can be utilized in inferring admixture histories based on weighted LD, which can be calculated using genotype data with iMAAPs[11]. Furthermore, we extended the GA and CGF models to the GA-I and CGF-I models in order to infer the time interval for a period of continuous admixture events followed by

isolation. Although HI model is a degenerate case for both GA-I and CGF-I models, where the admixture window becomes 1 generation, we kept it in our method because it is the most popular model employed in previous admixture studies. Considering the difficulty in fitting problem with polynomial functions, it is in our expectation that CAMer was not consistently accurate in determining the admixture model based on the weighted LD decay. However, its natural advantage of independence of both haplotype phasing and local ancestry inference makes it privilege to other CAT based methods. Our simulations indicated that its time interval estimations were reliable when its assumption that the true admixture history could be well approximated by one of the core models is valid.

Two quantities, namely msE and quasi F, were used to check the assumption of our method stated above and evaluate the credibility of the models' inference. These two quantities should both be taken into consideration to determine whether the models well described the admixture history. Both the data quality and the goodness of fitting of models can affect the value of msE, although the *F* value mainly measures the goodness of modeling. Informally, for the convenience of interpretation, msE can be an indicator of data quality, while *F* value can be used to check model assumption on admixture history. In our analysis, we suggested thresholds for msE and *F* to determine whether the null hypothesis should be rejected or not, which may be too strict in empirical analysis. Actually, msE and *F* values together measure whether the observed weighted LD can be well fit by the best-fit model(s). For example, the fitting process showed poor performance in the MKK population, which was accompanied by exaggerated msE and *F* values, showing significant inconsistencies between the observed and fitted weight LD curves, which indicates that the true admixture history cannot be well explained by any of the core models (see Supplementary Fig. S17). Therefore, in empirical analysis, it can be informally considered that the msE value reflects the quality of the data, whereas *F* value describes the performance of the model, although both of them measure the goodness of fitting.

In our previous study[11], we fit the weighted LD with high degree polynomial functions. However, this approach did not fully reveal the occurrence of continuous admixture. To address this issue, the present study developed CAMer to model admixture as a continuous process. CAMer also employed extensions of the classic continuous models, GA-I and CGF-I, which may bring the bias to have a wider admixture window when the real admixture exists in a short time. As we discussed earlier, another limitation for CAMer is its poor performance to determine the correct admixture model. Therefore, in empirical data analysis, we suggest all core models, rather than the best-fit model(s), should be examined. Taken together, despite there is space to further improve in the future, CAMer is a powerful method to model a continuous population admixture, which in turn would help us elucidate the complex demographic history of population admixture.

References

1. Loh, P. R. *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193,** 1233–1254 (2013).
2. Qin, P. *et al.* Quantitating and Dating Recent Gene Flow between European and East Asian Populations. *Sci. Rep*. **5,** 9500 (2015).
3. Xu, S. & Jin, L. A Genome-wide Analysis of Admixture in Uyghurs and a High-Density Admixture Map for Disease-Gene Discovery. *American Journal of Human Genetics* **83,** 322–336 (2008).
4. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5** (2009).
5. Zhu, X., Cooper, R. S. & Elston, R. C. Linkage analysis of a complex disease through use of admixed populations. *Am. J. Hum. Genet.* **74,** 1136–1153 (2004).
6. Jin, W., Li, R., Zhou, Y. & Xu, S. Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping. *Eur. J. Hum. Genet.* **22,** 930–937 (2013).
7. Jin, W., Wang, S., Wang, H., Jin, L. & Xu, S. Exploring population admixture dynamics via empirical and simulated genome-wide distribution of ancestral chromosomal segments. *Am. J. Hum. Genet.* **91,** 849–862 (2012).
8. Gravel, S. Population genetics models of local ancestry. *Genetics* **191,** 607–619 (2012).
9. Ni, X. *et al.* Length Distribution of Ancestral Tracks under a General Admixture Model and Its Applications in Population History Inference. *Sci Rep* **6,** 20048 (2016).
10. Pickrell, J. K. *et al.* Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. USA* **111,** 2632–7 (2014).
11. Zhou, Y. *et al.* Inference of multiple-wave population admixture by modeling decay of linkage disequilibrium with polynomial functions. *Under Rev*.
12. Zhou, Y., Liu, X. & Xu, S. Dissecting admixture linkage disequilibrium under a general model of population admixture. *Under Rev.* (2016).
13. Pfaff, C. L. *et al.* Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.* **68,** 198–207 (2001).
14. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).
15. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
16. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135,** 0–9 (2012).
17. Tukey, J. W. Bias and Confidence in Not-Quite Large Samples. *The Annals of Mathematical Statistics* **29,** 614 (1958).
18. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat*. **6,** 65–70 (1979).
19. R Core Team. R: A Language and Environment for Statistical Computing. **0** (2014).
20. Yang, X. AdmixSim-v1.0.2, http://www.picb.ac.cn/PGG/resource.php. (2015). at http://www.picb.ac.cn/PGG/resource.php.
21. Li, N. & Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165,** 2213–2233 (2003).
22. Zhu, X. & Cooper, R. S. Admixture mapping provides evidence of association of the VNN1 gene with Hypertension. *PLoS One* **2** (2007).

Acknowledgements

We thank Dr. Li Jin, Dr. Yungang He, and Dr. Raymond Chan for their comments. This work is supported by the Chinese Academy of Sciences (CAS) (XDB13040100; QYZDJ-SSW-SYS009), the National Natural Science Foundation of China (NSFC) grant (91331204), the National Science Fund for Distinguished Young Scholars (31525014), the Program of Shanghai Academic Research Leader (16XD1404700), and the National Key Research and Development Program (2016YFC0906403). S.X. is Max-Planck Independent Research Group Leader and member of CAS Youth Innovation Promotion Association. S.X. also gratefully acknowledges the support of the National Program for Top-

notch Young Innovative Talents of The "*Wanren Jihua*" Project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived and designed the study: S.X. Developed methods and computer tools: Y.Z., H.Q. Analyzed the data: Y.Z. and H.Q. Interpreted the data and wrote the paper: S.X., Y.Z., H.Q.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zhou, Y. *et al.* Modeling Continuous Admixture Using Admixture-Induced Linkage Disequilibrium. *Sci. Rep.* **7**, 43054; doi: 10.1038/srep43054 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.