Exploring Population Admixture Dynamics via Empirical and Simulated Genome-wide Distribution of Ancestral Chromosomal Segments

Wenfei Jin,[1] Sijia Wang,[2] Haifeng Wang,[3] Li Jin,[4] and Shuhua Xu[1,*]

The processes of genetic admixture determine the haplotype structure and linkage disequilibrium patterns of the admixed population, which is important for medical and evolutionary studies. However, most previous studies do not consider the inherent complexity of admixture processes. Here we proposed two approaches to explore population admixture dynamics, and we demonstrated, by analyzing genome-wide empirical and simulated data, that the approach based on the distribution of chromosomal segments of distinct ancestry (CSDAs) was more powerful than that based on the distribution of individual ancestry proportions. Analysis of 1,890 African Americans showed that a continuous gene flow model, in which the African American population continuously received gene flow from European populations over about 14 generations, best explained the admixture dynamics of African Americans among several putative models. Interestingly, we observed that some African Americans had much more European ancestry than the simulated samples, indicating substructures of local ancestries in African Americans that could have been caused by individuals from some particular lineages having repeatedly admixed with people of European ancestry. In contrast, the admixture dynamics of Mexicans could be explained by a gradual admixture model in which the Mexican population continuously received gene flow from both European and Amerindian populations over about 24 generations. Our results also indicated that recent gene flows from Sub-Saharan Africans have contributed to the gene pool of Middle Eastern populations such as Mozabite, Bedouin, and Palestinian. In summary, this study not only provides approaches to explore population admixture dynamics, but also advances our understanding on population history of African Americans, Mexicans, and Middle Eastern populations.

Introduction

Admixed populations come into being when previously mutually isolated populations meet and sexually reproduce. This has been a common phenomenon throughout the history of modern humans as previously isolated populations come into contact through colonization and migration.[1–3] Human diasporas over the past millennium have resulted in even more frequent population admixtures. Many recently admixed populations, such as African Americans and Mestizos (individuals with genetic ancestry from Europeans and Amerindians), have received much attention because of their potential advantages in the discovery of disease-associated genes. Specifically, a gene-mapping strategy for identifying disease-associated genetic variants named admixture mapping has been developed.[4–7] The statistical power of admixture mapping relies on the extended and elevated linkage disequilibrium (LD) in the admixed population that was determined by population history and admixture processes.[4,8,9] Therefore, as shown in several theoretical and simulation studies, population admixture dynamics has a strong effect on the statistical power of admixture mapping.[9–12]

In fact, accurate understanding of population admixture dynamics is important not only to admixture mapping but also to other applications, such as elucidating population history[13] and detecting natural selection signatures in admixed populations.[10,14] However, the fine admixture dynamics of the well-known admixed populations have not been well established, although some studies have examined the simulated data[1,15] or experimental data with sparse markers.[9,12] Recently, the availability of genome-wide high-density single-nucleotide polymorphism (SNP) data has facilitated the study of detailed genetic structures of admixed populations.[16–21] However, most of these studies relied on simplified models that do not take into account the inherent complexity of the admixture processes. Moreover, the haplotype and chromosomal segment patterns shaped by recombination within each individual have been deliberately ignored in most studies because of many inherent challenges.[22]

For individuals from admixed populations that have existed for a long time, their chromosomes resemble a mosaic of chromosomal segments of distinct ancestry (CSDAs). The CSDAs in the admixed population would have been reshaped and rearranged by recombination in each generation, which should provide valuable information about the population history. In other words, the CSDAs will be spliced into smaller pieces as the number of generations since admixture increases, while the chromosomes from recently admixed individuals contain many more long CSDAs. Information regarding the average CSDA length

[1]Max Planck Independent Research Group on Population Genomics, Chinese Academy of Sciences and Max Planck Society (CAS-MPG) Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; [2]Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA; [3]Chinese National Human Genome Center at Shanghai, Shanghai 201203, China; [4]Ministry of Education (MOE) Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China
*Correspondence: xushua@picb.ac.cn
http://dx.doi.org/10.1016/j.ajhg.2012.09.008.

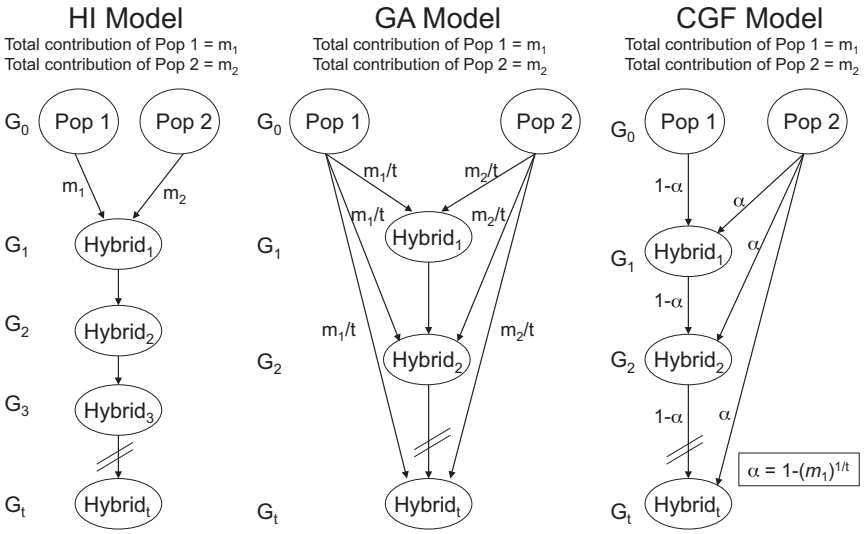

Figure 1. Admixture Models Used to Simulate the Population Admixture Process

Hybrid isolation (HI) model and continuous gene flow (CGF) model were adapted from Long,[10] and graduate admixture (GA) model was adapted from Ewens and Spielman.[11] In each model, the final genetic contributions from pop1 and pop2 are m_1 and m_2, respectively. The admixed population experienced G_i generation, which range from 1 to t generations.

has been used to infer the number of generations since admixture in various studies.[20,23–26] However, the distribution of CSDA length may contain more valuable information concerning population admixture history and admixture dynamics, which has not yet been explored.

In this study, we first performed forward-time simulations to investigate the effect of admixture dynamics on the distributions of CSDA length based on three distinctive admixture models. Our analysis indicated that the distribution of CSDA length could provide much useful information regarding the fine population admixture dynamics. This approach was found to be robust and insensitive to most of the problematic issues relating to the demographic histories of the parental populations and the admixed population. Then we applied this approach to several admixed populations with different histories to explore their admixture dynamics. As a result, we revealed the admixture dynamics of African Americans and Mexicans by comparing the empirical distribution of CSDA length with the simulated data. Our analysis also showed that there had been a few recent gene flows from Sub-Saharan Africa contributing to the gene pools of the admixed populations in Middle East such as Mozabite, Bedouin, and Palestinian.

Materials and Methods

Data Sets and Population Samples

In this study, the admixture dynamics of African Americans, Mexicans, and four admixed populations in the Greater Middle East (Mozabite, Bedouin, Druze, and Palestinian) were investigated. All together, the genotypes of 3,398 individuals were obtained from International Haplotype Map Project (HapMap),[27,28] Human Genome Diversity Project (HGDP),[29] NIGMS (National Institute of General Medical Sciences) Human Variation Panels (HVP), Mexican Genetic Diversity Project (MGDP),[18] and Illumina iControlDB. The combined data set includes the genotypic data of 580 samples from 5 HapMap populations, which comprised individuals of African ancestry from the southwest USA (ASW, n = 87), individuals of Mexican ancestry in Los Angeles, CA (USA) (MEX, n = 77), Yoruba in Ibadan, Nigeria (YRI, n = 167), Utah residents with northern and western European ancestry from the CEPH collection (CEU, n = 165), and Han Chinese from Beijing (CHB, n = 84). In addition, there are 2,161 African Americans genotyped by Illumina 550K Beadarray from iControlDB, 100 Mexican Americans genotyped by Affymetrix SNP 6.0 from HVP, 300 Mexican Mestizos genotyped on Affymetrix 100K GeneChip from MGDP, 30 Zapotecas representing pure Amerindians from MGDP, 64 Amerindians (AMI) genotyped by Illumina 650K Beadarray from HGDP, and four admixed populations from HGDP (30 Mozabite, 45 Bedouin, 46 Palestinian, and 42 Druze). In order to keep as many markers as possible for the analyses, each admixed population and its putative parental populations were merged to perform population genetic analysis separately. For each data set, we filtered out the related individuals, performed a quality control analysis, and removed individuals with >10% missing genotypes and SNPs with >10% missing data.

Admixture Models and Simulations

In reality, population admixture processes are too complex to study directly. Here, we attempted to explore population admixture dynamics by examining the distributions of CSDA length in three typical admixture models that can summarize most of the possible scenarios (Figure 1): hybrid isolation (HI) model,[4] gradual admixture (GA) model,[11] and continuous gene flow (CGF) model.[9,10] The genetic structure and LD pattern of the admixed population under these models have been investigated systematically in several previous studies.[9–11,15] In all three admixture models, m_1 and m_2 denote the genetic contributions of the two parental populations (pop1 and pop2) to the admixed population, respectively, and t denotes the number of generations. In the HI model, admixture occurs only in the first generation and is followed by recombination and genetic drift, with no further genetic contribution from either of the parental populations. In the GA model, admixture occurs at a fixed rate in each generation, with continuous genetic contributions from both of the parental populations. The rates of continuous gene flow from pop1 and pop2 are m_1/t and m_2/t, respectively, with the rest of the genetic contribution being from the admixed population of the previous generation, which ensures the same genetic contribution of a parental population at each generation. The CGF model can be regarded as an extension of the GA model, in which the recipient/admixed population receives a constant but reduced rate of gene flow (α) from the other parental

population (genetic donor) in each generation. In order to make the CGF model compatible with HI model and GA model, we let the cumulative genetic contribution from both parental populations be equal to that under the HI model and GA model. The gene flow that the admixed population receives from the genetic donor in each generation is calculated by $\alpha = 1 - (m_1)^{1/t}$.

A forward-time simulation program was developed based on the three aforementioned admixture models considering only the autosomal data. We used phased genotype data of YRI and CEU from HapMap as the initial statuses of the parental populations.[28] We first sampled the haploid chromosomes of YRI and CEU according to their estimated contributions. Then we combined each pair of haploid chromosomes from the two parental populations to construct a diploid admixed individual. We modeled the processes of genetic drift and recombination under a Wright-Fisher neutral model.[30] In this simulation, recombination was introduced according to the genetic map adapted from HapMap (release #22; with 3,540 cM in total on the 22 autosomes),[28] and mutation was ignored given the short population history in the simulation. The effective population size (N_e) of each population was set at 10,000. We labeled the genotypes from different parental populations in order to accurately know the genetic origin of each locus. In the CGF model, for a given specific admixture proportion, the parental population can serve either as genetic donor or as genetic recipient. In this way, the genetic donor in CGF model was referred to as CGFD, and the genetic recipient was referred to as CGFR. The number of generations for each model was set at 10, 20, 50, and 100, respectively. Extensive simulations were performed to explore the influence of admixture dynamics on the distribution of CSDAs by modifying various parameters such as proportions of ancestry contribution and N_e.

Population Genetic Analysis and Inference of CSDAs

We conducted population genetic analysis on each set of filtered data. In order to mitigate the effects of strong LD blocks, SNPs were removed until $r^2 < 0.5$, which was calculated in a sliding window of 50 SNPs and shifted by 5 SNPs each time. Based on the thinned markers, we conducted principal component analysis (PCA) at the individual level to reveal the population structure by EIGENSOFT.[31] The genetic contribution of the parental population to the admixed population was inferred with FRAPPE[32] and STRUCTURE.[26,33] FRAPPE, which implements an expectation-maximization (EM) algorithm, was run on all the available SNPs with 10,000 iterations. STRUCTURE, which implements a model-based clustering method to infer population structure, was run with 100,000 burn-ins and 100,000 iterations by setting admixture model.

We chose HAPMIX to infer CSDAs in the admixed populations in this study because previous studies with simulated data had reported that HAPMIX outperformed the other available methods and software.[24,34] The haplotypes of parental populations, the inputs for HAPMIX, were either downloaded directly from HapMap website or inferred with fastPHASE[35] when the phased data were unavailable. We found that the short CSDAs inferred by different methods were not consistent with each other, which might result from some uncertainty in statistical inference. We removed those very short CSDAs in order to improve the reliability of the distribution because no method and software can accurately infer very short segments.[13] We also considered the results of LAMP because it also performs well.[34,36]

Measurement of the Differences between Two Distributions

In this study, statistical computing and graphics generating were mainly performed with R version 2.13.[37] The 10-based logarithms of CSDA length were calculated because they were assumed to follow the normal distributions. The log-normal distribution was displayed with the probability density function, which describes the relative likelihood that a given random variable occurs at a given point. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one. To assess the differences of CSDA distributions among different models and empirical data, Kolmogorov-Smirnov (K-S) tests[38] were performed. K-S test is a nonparametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution or to compare two samples. To quantify the differences between these distributions, earth mover's distance (EMD)[39,40] was calculated. EMD is a method for evaluating the dissimilarity between two probability distributions. Intuitively, given two distributions, one can be seen as a mass of earth properly spread in space and the other as a collection of holes in the same space. The EMD calculates the least amount of work needed to fill the holes with the earth. In this way, the value of EMD corresponds to the amount of earth multiplied by the distance by which it moved. Therefore, the lower the EMD between two distributions, the higher the similarities between them.

Analysis of 1,890 African American Samples

It is evident that populations from Europe and West Africa dominantly contributed to the African American gene pool, whereas Amerindian and East Asian contributed to only a small fraction of the gene pool. In order to simplify the analysis to a two-way admixture, we filtered out samples with obvious Amerindian or East Asian ancestry. After a series of quality control procedures, 1,890 unrelated African American individuals and 354 samples from their putative parental populations sharing 491,557 autosomal SNPs were kept. Detailed information on the data set and data processing has been shown in Jin et al.[34] Among the 1,890 African Americans, 52 individuals were from HapMap ASW and the other 1,838 individuals were from iControlDB. Four populations from different continents (112 unrelated CEU, 112 unrelated YRI, 84 unrelated CHB, and 44 unrelated AMI) represented putative parental populations of African Americans.

After filtering out the high-linkage SNPs by using PLINK,[41] we reduced the original SNPs to 341,672 SNPs. PCA was performed on all the samples at the individual level with these thinned markers. FRAPPE was performed on the 491,557 autosomal SNPs successfully genotyped in all the samples of African American, CEU, and YRI. For the inference of CSDA, haploid data of 88 YRI and 88 CEU from HapMap were taken to represent their African and European parental populations, respectively. The genetic contribution of Europeans to African Americans was set at 21.65% at the population level based on FRAPPE results.[34] The number of generations since admixture (λ) with the highest overall likelihood was taken as its estimation. By running HAPMIX in diploid model, we obtained the haplotypes and CSDA of each African American individual.

The aforementioned forward-time simulation program was used to simulate admixture dynamics of African Americans. We inferred the population admixture dynamics considering the distribution of CSDAs for both ancestries simultaneously instead of considering only those of a single ancestry. For the CGF model, the

case in which the European population served as CGFD and the African population as CGFR was called CGF1 model, and the case in which the European population served as CGFR and the African as CGFD was called CGF2 model. The effective population sizes (N_e) of each population were set according to the HapMap.[28] Specifically, the N_e of African, European, and African American populations were set at 17,094, 11,418, and 17,094, respectively. The contribution of European ancestry to African Americans was set at 21.65% according to the observation of the 1,890 African Americans. Based on the recorded history of African Americans, the time of admixture (in generations) was set from 10 to 17, stepped by one generation for each model. For both empirical and simulated data, CSDAs <0.5 cM for African ancestry and CSDAs <0.8 cM for European ancestry were filtered out. We also estimated the influence of assortative mating[42] in the African Americans and removed the regions with significant assortative mating. The EMDs between the empirical distribution of CSDAs and that of each simulated data set were calculated. Only the simulation showing the lowest EMD with empirical distributions for both ancestries was regarded as fitting the corresponding model.

Analysis of 413 Mexican Samples

Mexican Mestizos and Mexican Americans are recently admixed populations mainly composed of Amerindians and Europeans with similar history, both of which were referred to as Mexicans for the convenience of presentation in this study. First, 300 unrelated self-identified Mexican Mestizo individuals and 239 unrelated individuals from their putative parental populations in MGDP were downloaded from the INMEGEN website.[18] These individuals were genotyped with Affymetrix 100K SNP array and the population structure has been systemically analyzed in Silva-Zolezzi et al.[18] Mexican American samples were obtained from two data sets. The first data set contained 100 Mexicans from the Coriell HD100MEX panel genotyped on the Affymetrix SNP 6.0 GeneChip. The second data set contained 58 unrelated Mexicans from HapMap3 genotyped on both Affymetrix SNP 6.0 GeneChip and Illumina 1M Beadarray. The two data sets of Mexican Americans were merged because both of them were collected in Los Angeles and genotyped on Affymetrix SNP 6.0 GeneChip. Overall, 767,454 autosomal SNPs on the 158 Mexican Americans were kept after quality control.

Overall, 458 unrelated Mexicans including 300 Mexican Mestizos and 158 Mexican Americans were collected. After filtering out SNPs with >10% missing genotypes, we had 36,000 autosomal SNPs shared by the remaining 652 samples (413 Mexicans and 239 samples from parental populations). PCA was performed at the individual level with all 36,000 autosomal SNPs shared by the 652 samples. We also performed STRUCTURE analysis on the 652 samples with SNPs with intermarker distance >2 Mb. Because both historical records and genetic evidence indicate that African populations have contributed only mildly to the Mexican gene pool, 45 samples with pronounced African ancestral component (5%, STRUCTURE analysis) were filtered out. Thus we simplified the following analysis to a two-way admixture between European and Amerindian populations.

Taking advantage of the genome-wide high-density SNP data, we analyzed Mexican Americans and Mexican Mestizos separately to infer the CSDAs. For Mexican Mestizos, we ran HAPMIX with haploid data of 30 Zapotecas (Amerindian) and 30 CEU as representatives of their Amerindian and European ancestry, respectively. For Mexican Americans, haploid data of 44 AMI from HGDP and 44 CEU from HapMap were taken to represent Amerindian and European ancestry, respectively. After filtering out the SNPs with >10% missing genotypes and removing the monomorphic SNPs in the data set combined by CEU and AMI, we used 183,042 autosomal SNPs to infer the CSDAs of Mexican Americans. The CSDAs of Mexicans were obtained by simply pooling together those of Mexican Mestizos and Mexican Americans.

The simulations performed on Mexicans were the same as those performed on African Americans, but some population parameters were changed to fit the known history of Mexicans. For the CGF model, the case in which the European population served as CGFD and the Amerindian as CGFR was named CGF1 model, and the case in which European population served as CGFR and Amerindian as CGFD was named CGF2 model. The N_e of the European, Mexican, and Amerindian populations were set at 11,418, 15,000, and 11,418, respectively. Based on the known history of the Mexican population, the admixture time in generations was set from 15 to 25, stepped by one generation for each model. The genetic contribution of European ancestry to Mexicans was set at 49.2%, according to the empirical analysis of the 413 Mexicans. For both empirical and simulated data, CSDAs <1.2 cM for either ancestry were filtered out. We also estimated the influence of assortative mating in the Mexicans and removed the regions with significant assortative mating. Finally, we also compared the distributions of CSDAs between Mexican Mestizos and Mexican Americans.

Analysis of Populations in Middle East

Four Middle Eastern populations (Mozabite, Bedouin, Palestinian, and Druze) have been reported to have both European and Sub-Saharan African ancestries.[24,43] All samples from the four populations have been genotyped with Illumina 650K Beadarray. CEU, YRI, and CHB from HapMap were taken to represent their putative parental populations and were merged together. Overall, 112 CEU, 112 YRI, 84 CHB, 27 Mozabite, 46 Palestinian, 45 Bedouin, and 42 Druze samples were kept after quality control, and PCA analyses were performed on these samples. We ran HAPMIX on each of the four populations by using haploid genomes of 88 YRI and 88 CEU as their reference parental populations. Simulations of each of the four populations were performed following the procedure similar to that used for African Americans, except that the parameters for simulations were adjusted to fit each corresponding population. For the CGF model, the case in which the European population served as CGFD and the African as CGFR was called CGF1 model, and the case in which the European population served as CGFR and the African as CGFD was called CGF2 model. Specifically, the time of admixture for Mozabite, Bedouin, Palestinian, and Druze populations were set at 100, 90, 75, and 60 generations, respectively, according to a previous study.[24] Comparisons of the empirical distributions of CSDA length in each population with those of corresponding simulations were used to determine whether recent gene flow contributed to the gene pools of those admixed populations.

Results

Distribution of CSDA Length under Different Admixture Models

The distributions of CSDA length were examined under three admixture models (HI model, GA model, and CGF model) (Figure 1).[9–11,15] For the CGF model, the parental population acting as genetic donor was referred to as

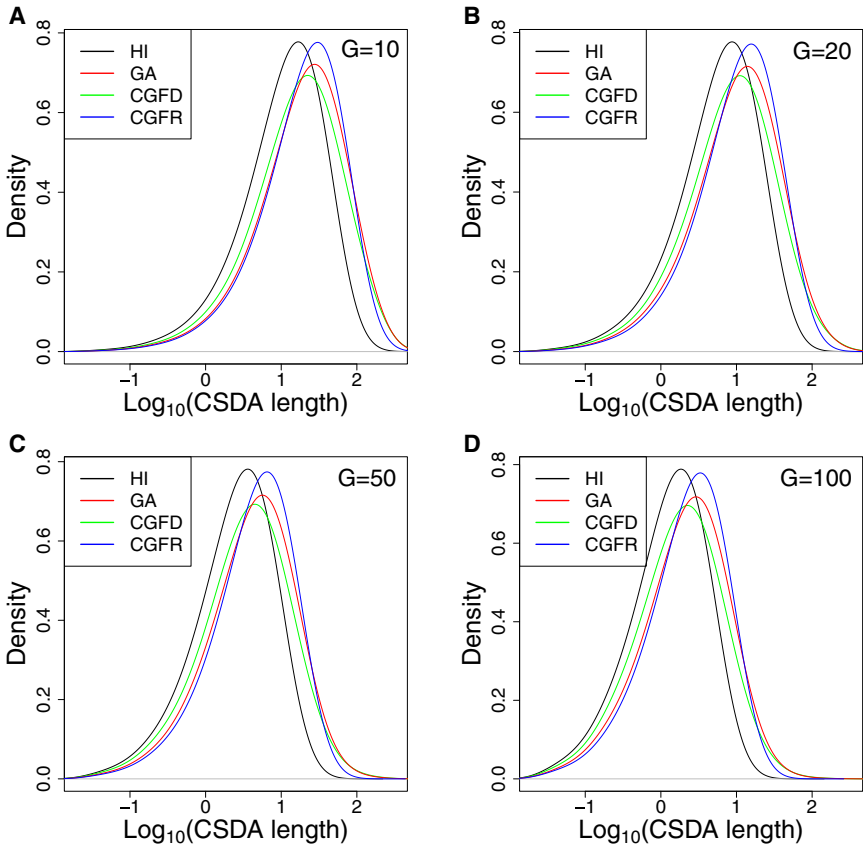

Figure 2. Distributions of Chromosomal Segments of Distinct Ancestry Length when Genetic Contribution of the Parental Population to the Admixed Population Is 50%

G, number of generations since admixture. Number of generations since admixture was set to 10 (A), 20 (B), 50 (C), or 100 (D).

CGFD, and that acting as genetic recipient was referred to as CGFR. We set N_e for each population at 10,000 and simulated a scenario in which the genetic contribution of the parental population to the admixed population was 50%. The number of generations since admixture for each model was set at 10, 20, 50, and 100. The basic information regarding the distribution of CSDAs indicated that different models and generations led to different distributions of CSDAs (Table S1 available online). We observed that the distribution of CSDAs under the HI model was considerably different from those under the other models (Table S1) because of the higher relative number of short CSDAs obtained under the HI model. Analysis showed that distributions of CSDA length between different models differed significantly when the numbers of generations since admixture were the same ($p < 2.2 \times 10^{-16}$, two-sample K-S test). Distributions of CSDA length also differed significantly when the generations since admixture were different under the same model ($p < 2.2 \times 10^{-16}$, two-sample K-S test).

The 10-based logarithms of CSDA length were calculated (Table S2) under the assumption that the distribution of CSDAs approximates a log-normal distribution. After this transformation, the distributions of CSDA length between different models (Figure 2) were still significantly different ($p < 2.2 \times 10^{-16}$, two-sample K-S test). In each simulated scenario, the distributions of CSDA length under the GA model and those under the CGFR model were the most similar among the four models, whereas the HI model and the CGFR were the most dissimilar (Figure 2). This observation was confirmed by EMDs between these distributions (Table S3). Within each model, the EMDs between two distributions were found increased as the number of generations since admixture increased (Table S3). The distribution of CSDAs was still significantly different even when the short ones were filtered out (Figure S1), which indicated that admixture dynamics could be distinguished even if only long CSDAs were available.

In order to analyze the influence of ancestral contribution on the distributions of CSDA length, we allowed the hypothetical contribution of the parental population to range from 10% to 90%. These simulated results suggested that the distribution of CSDA length under the GA model was very similar to that of CGFD when genetic contribution of the parental population was very low (Figure S2), and it became very similar to that of CGFR when genetic contribution of parental population was very high (Figure S3). All simulations showed that long CSDAs were retained if gene flow from the parental populations continuously contributed to the admixed population (GA and CGFD model), and the proportion of short CSDAs under the HI model was much higher than those observed under other models. In brief, these analyses indicated that the distribution of CSDA length could provide valuable information about population admixture dynamics and history.

Distribution of Individual Ancestry Proportions under Different Admixture Models

Individual ancestry proportion in admixed populations can be directly estimated by various methods and software.[26,32,33,44] We investigated whether the distribution of individual ancestral proportions could be used to evaluate the admixture dynamics of admixed populations. We first investigated a scenario in which a parental population contributed to 50% of the genetic components of the admixed population. Basic information on the distributions of individual ancestry proportions was shown in Table S4. When the number of generations since

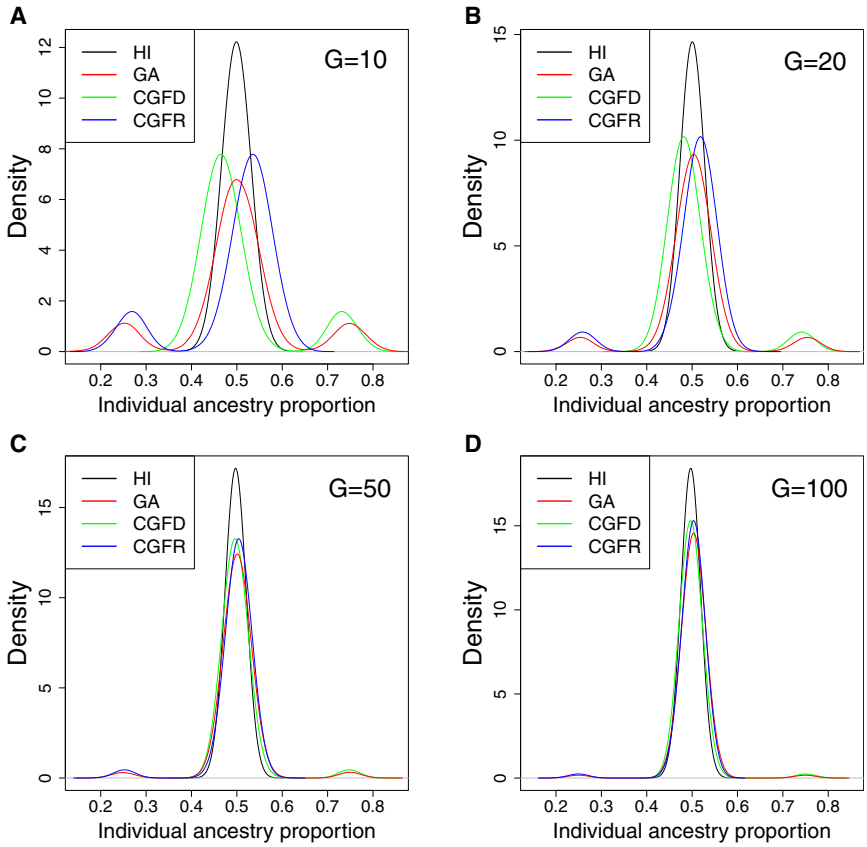

Figure 3. Distributions of Individual Ancestral Proportion when Genetic Contribution of the Parental Population to the Admixed Population Is 50%

G, number of generations since admixture. Number of generations since admixture was set to 10 (A), 20 (B), 50 (C), or 100 (D).

admixture was set at 10, distributions of individual ancestry proportions were completely distinguished from each other (Figure 3A), and these differentiations were statistically significant ($p < 2.2 \times 10^{-16}$, K-S test). However, these differentiations decreased continuously as the number of generations since admixture increased (Figure 3). Eventually, it was almost impossible to distinguish the distributions from each other when the time of admixture was set at 100 generations (Figure 3D). These observations were confirmed by the quantitative measures in that EMDs between any two distributions decreased when the number of generations since admixture increased (Table S5).

Extended simulations on a series of different genetic contribution from parental populations to admixed population were performed. All these results showed that the distribution of individual ancestry proportions could reveal the admixture dynamics of the recently admixed populations but not those populations with a long history (Figures S4–S7). It also indicated that the admixture dynamics of recently admixed populations, such as African Americans and Mexicans, could be inferred by comparing the simulated and empirical distributions of the individual ancestry proportions. Although this approach sounds appealing theoretically, it requires a very large sample size. Another limitation of this approach is that the sampling processes can significantly affect the distribution of individual ancestry proportion, which is unpredictable in most cases.

Robustness of the Distributions of CSDA Length in Inferring Admixture Dynamics

We investigated the feature of CSDA distribution by changing the population parameters and performing extended simulations before applying this approach on empirical data. First, our analysis showed that neither the N_e of the admixed populations nor those of their parental populations affected the distributions of CSDA length (Figure S8A). Further analysis showed that neither population expansion nor bottlenecks affected the distribution of CSDA length (Figure S8B). Therefore, these results essentially indicated that the distributions of CSDA length were not influenced by these common demographic events. However, our analysis showed that the distributions of CSDA length could be affected by nonrandom mating in admixed populations, which flattened the peak of the distribution (Figure S9). Second, as long as distance between the contiguous markers was significantly less than the sizes of the short CSDAs, the distribution of CSDA length remained unaffected by marker density. Third, chromosome length had an obvious influence on the distribution of CSDAs during the first few generations, but its influence became weak as the number of generations increased. However, the effect of chromosome length was ignorable because all these simulations used exactly the same chromosome length and loci. Finally, although the statistical error has a mild effect on very recently admixed population, we found that the statistical error in locus ancestry inference flattened the distribution compared with the expected when the history is very long (Figure S10). We attempted to reduce the influence of possible systematic statistical error by inferring the CSDAs of simulated data alongside the empirical data.

The distribution of CSDA length takes advantage of the information created by recombination and is independent of the allele frequencies of parental populations. Because each individual contains many CSDAs, the distribution of CSDA length is much more useful than the distribution of individual ancestry proportion because of the fact that it requires much fewer samples to create a reliable distribution. This approach is also very robust because it is not significantly affected by most demographic events.

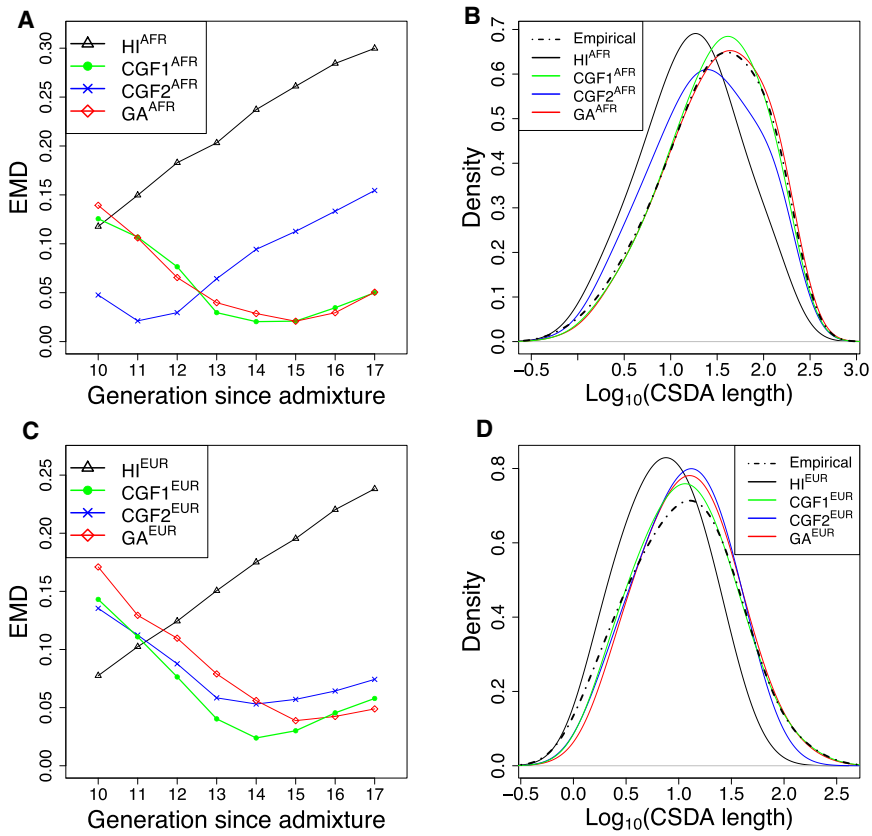

Figure 4. Admixture Dynamics of European and African Ancestry in African Americans

For the CGF model, the case in which Europeans continually served as genetic donor was considered as CGF1 model, whereas Africans as genetic donor was considered as CGF2 model. To find the model that fits the empirical distribution best, earth mover's distance (EMD) between empirical data and that of each model was calculated. The model showing the lowest EMD with the empirical data was considered best fit.

(A) Distribution of EMDs for African ancestral component between empirical data and each model.

(B) Empirical distribution of CSDA length for African ancestral component and simulated distributions when the number of generations was set to 14.

(C) Distribution of EMDs for European ancestral component between empirical data and each model.

(D) Empirical distribution of CSDA length for European ancestral component and the simulated distributions when the number of generations was set to 14.

Admixture Dynamics of African Americans

It is important to illuminate the fine admixture dynamics of African Americans because of the fact that they have wide applications and have already been widely used in admixture mapping. Overall, 1,890 African American samples, which contained negligible ancestral components other than European and African, were investigated (Figures S11A and S11B). The genetic contribution of European ancestry to African Americans was estimated to be 21.65% by FRAPPE. Based on overall likelihood given by HAPMIX, the time of admixture (λ) was estimated to be seven generations, which could be considered as an average value in all samples based on the HI model. The estimated λ was similar to those reported in other studies with different data sets.[23,24,45,46] However, it has been almost 300 years (15 generations assuming 20 years per generation) since the 18th century slave trade that brought most of the African ancestors of current African Americans to the New World. In this way, the time of admixture estimated by genetic data seemed to contradict historical records.

To resolve the apparent contradiction between the estimated date and the historical date, we simulated an African American population by setting the genetic contribution of European ancestry to the African American gene pool at 21.65% and setting the time of admixture at 10–17 generations, stepped by one generation. The empirical distributions of CSDA length for both African and European ancestries were calculated based on HAPMIX output. We compared the empirical distributions of CSDAs with those of simulated data under the four models: HI model, GA model, CGF1 model (European population serving as CGFD and African as CGFR), and CGF2 model (European population serving as CGFR and African as CGFD). We found that the CGF1 model with 14 generations for both ancestries fit the empirical data best (Figures 4A and 4C). The lowest EMDs for both African and European ancestral components (EMD = 0.0204 and 0.0239, respectively) were observed when 14 generations since admixture were set under CGF1 model (Figure 4), in which gene flow from European populations continuously contributed to the African American gene pool.

For both European and African ancestries, the EMDs between empirical distributions and those under HI model increased as the number of generations increased (Figures 4A and 4C). Even the distributions of ten-generation HI model were deficient in long CSDAs compared with empirical distributions for both European and African ancestral components, and neither of their distribution peaks overlapped with those of empirical distributions. In this way, the HI model can be excluded because the history of the African American population goes back more than 200 years (10 generations). For the CGF2 model in which gene flow from an African population continuously contributed to the African American gene pool, the lowest EMD for African ancestral component between the empirical distribution and the simulated distributions was obtained when time of admixture was set at 11 generations. However, this was not consistent with that of European ancestry (14 generations) (Figures 4A and 4C). The

lowest EMD for European ancestral component between the empirical and the simulated distributions in the CGF2 model was also higher than that under either the CGF1 or the GA model (Figure 4C). In this way, the CGF2 model did not hold when both African and European ancestries were considered. The distribution of EMD between the GA model and the empirical data was similar to that of CGF1 model when African ancestral component was investigated (Figure 4A), but the CGF1 model had a much lower minimum EMD with the empirical distribution than that under the GA model considering the European ancestral component.

Although the actual population admixture of African Americans might be more complex than what our simulation suggested, the CGF1 model setting at 14 generations was found to be reasonably representative, capturing the main pattern of the population admixture dynamics. Direct comparison of the empirical CSDA distribution with the simulated distributions at 14 generations also supported the CGF1 model (Figures 4B and 4D), although the empirical distribution was slightly flatter than the simulated distribution, which possibly resulted from nonrandom mating or higher error rates during the inference of CSDAs from the empirical data than from the simulated data. Considering that the migration of Africans to the United States has been rare during the past 200 years and admixture has occurred gradually between African American and European American populations, this model also fits the recorded history well. In addition, gene flow from Europe would have continuously contributed to the African American gene pool because children with one European parent and one African American parent were generally regarded as African Americans. Because the gene flow from the European population is expected to continuously contribute to the African American gene pool, it is very likely that the proportion of European ancestral component in African Americans will continuously increase in the future.

The distributions of individual ancestry proportions for African Americans fit none of the four models perfectly (Figure S11C). This may have been due to the small sample size, sampling error, or substructure within the African American population. By carefully examining the distribution of individual ancestry proportions, we found that a small fraction of African Americans had a much higher proportion of European ancestry (with very little African ancestry) than that of any simulated individuals (Figure S11C), indicating substructures of African American population in terms of ancestry proportion. This might have resulted when African American individuals from particular lineages (integrated into the European American community) were apt to intermarry with people of European ancestry or of dominant European ancestry. Generation by generation, the European ancestral component was continually enriched in these specific African American lineages, and therefore a few African American individuals with a much higher proportion of European ancestry than expected under the assumption of random admixture could be observed in our data. Second, we found that individuals with extremely high proportion of either African ancestry or European ancestry tended to have more estimated generations since admixture (Figure S11D), which might suggest that the individuals with only a little European or African ancestral component tended to inherit them from much earlier admixture events. In contrast, the individuals who received roughly even genetic contributions from both parental populations tended to have fewer estimated generations (Figure S11D), indicating that these individuals were more likely to be descendants of recent interethnic marriages.

Admixture Dynamics of Mexicans

Mexicans (including Mexican Americans and Mexican Mestizos) are the second most well studied population in admixture mapping. Based on a simplified model without considering the complex population admixture process, the time of admixture for Mexican Americans or Mexican Mestizos has been estimated to be 15 generations or fewer in previous studies.[47–49] However, the real admixture history of Mexicans could be much longer, considering that Europeans first colonized the New World more than 500 years ago (>25 generations assuming 20 years per generation). After removing the individuals with an obvious African ancestral component, 413 Mexican individuals with negligible recent ancestry other than Amerindian and European were used to investigate the admixture dynamics of Mexican populations (Figures S12A and S12B). In the PCA plot, four continental populations (YRI, CHB, CEU, and AMI) were located at the corner of the ladder-shaped plot, whereas Mexican individuals all sat between CEU and Amerindian clusters (Figure S12A). The genetic contribution of European ancestry to the 413 Mexican samples was estimated to be 49.2% according to PCA and STRUCTURE.

We simulated a Mexican population by setting the genetic contribution of Europeans to the Mexican gene pool at 49.2% and 15–25 generations since admixture, stepped by one generation. The empirical distributions of CSDA length for both Amerindian and European ancestries were obtained by merging CSDAs from Mexican Americans and Mexican Mestizos that had been inferred by HAPMIX, respectively. The EMDs between the HI model and the empirical distribution for both Amerindian and European ancestral components increased as the number of generations increased (Figures 5A and 5C), and the distributions under HI model at 15 generations still lacked long CSDAs. For both the CGF1 model (European population serving as CGFD and Amerindian as CGFR) and the CGF2 model (European population serving as CGFR and Amerindian as CGFD), the lowest EMDs for Amerindian and European ancestral components were inconsistent. Specifically, both of the lowest EMDs generated by these two models were still

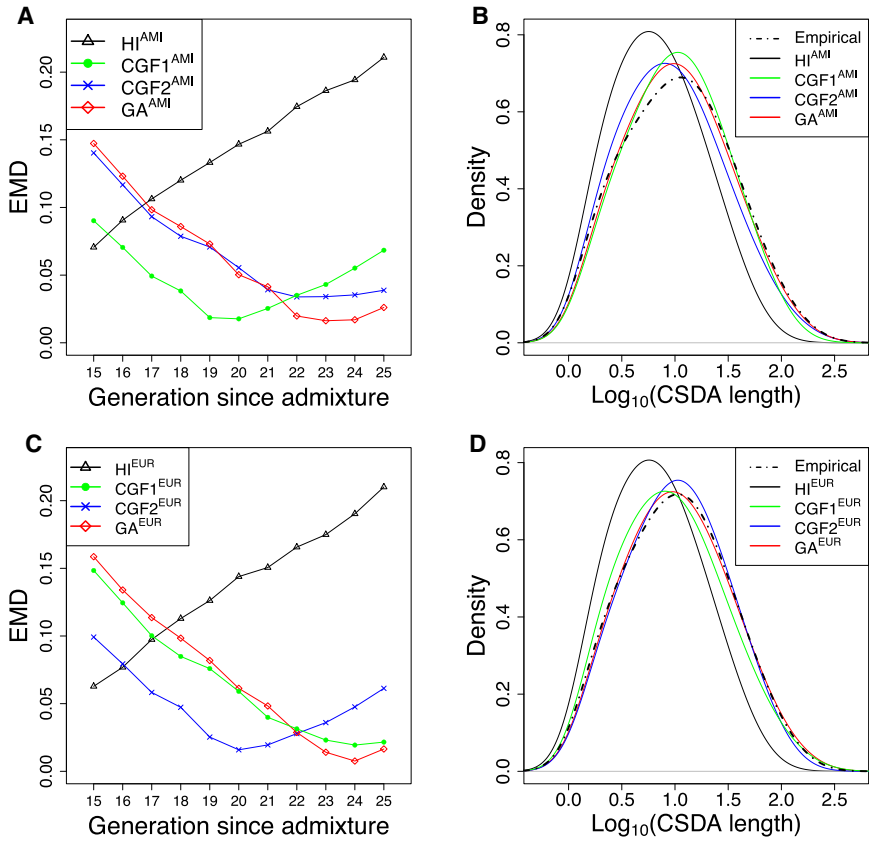

Figure 5. Admixture Dynamics of European and Amerindian Ancestry in Mexicans

The model showing the lowest EMD with the empirical data was considered as best fit. The GA model, in which both European and Amerindian populations continuously contributed to the Mexican gene pool over about 24 generations, fit the empirical data best.

(A) Distribution of EMDs for Amerindian ancestral component between empirical data and each model.

(B) Empirical distribution of CSDA length for Amerindian ancestral component and the simulated distributions when the number of generations was set to 24.

(C) Distribution of EMDs for European ancestral component between empirical data and each model.

(D) Empirical distribution of CSDA length for European ancestral component and the simulated distributions when the number of generations was set to 24.

higher than that generated by the GA model, indicating that the GA model fit the empirical data best among the four models. The EMDs between the empirical distributions and the distributions under the GA model for both Amerindian and European ancestral components reached the lowest value (EMD = 0.0163 and 0.0076, respectively) at 23 and 24 generations, respectively (Figures 5A and 5C). In short, the GA model at 24 generations fit the empirical data best among all these simulated scenarios, as indicated by the distribution of EMDs. Direct observation also showed that empirical distribution of CSDAs essentially fit the GA model at 24 generations (Figures 5B and 5D). The results were essentially consistent with that of an alternative analysis in which the Mexicans with >1% African ancestry were excluded.

Considering that both pure Amerindian and pure European migrants have coexisted in Mexico, the GA model is intuitively much more reasonable than the others. Considering the Mexican Americans and Mexican Mestizos separately, we found the genetic contribution of European ancestry to Mexican Americans to be 53.9%, which was significantly higher than that of the 268 Mexican Mestizos (46.7%, $p = 0.0018$, t test) (Figure S12C). Further analysis showed that the distribution of CSDAs of the Amerindian ancestral component in Mexican Americans was essentially identical to that of Mexican Mestizos. However, the CSDAs of European ancestry in Mexican Americans were much longer than those present in Mexican Mestizos (Figure S12D), which suggested recent gene flow from European to Mexican American populations. In other words, the fact that European populations have contributed more to Mexican Americans than to Mexican Mestizos was likely to have resulted from recent intermarriage between European Americans and Mexican immigrants in the U.S.

Analysis of Admixed Populations in Middle East

We also explored the admixture dynamics of four admixed populations from HGDP (Mozabite, Bedouin, Palestinian, and Druze) by using the same procedure as used for African American and Mexican populations.[24,43] In the PCA plot, three putative parental populations (YRI, CEU, and CHB) were located on the peaks of the triangular-like plot and the four admixed populations were dispersed between YRI and CEU on PC1 (Figure S13). We are interested in whether recent gene flow from their parental populations (African and European) contributed to the gene pools of these admixed populations. We compared the empirical CSDA distributions of each population with those under the four models: the HI model, GA model, CGF1 model (European population serving as CGFD and African as CGFR), and CGF2 model (European population serving as CGFR and African as CGFD). If the empirical distributions contain more long CSDAs than the simulated distributions, it could be taken as indicative of recent gene flow from the parental populations.

Mozabites residing in North Africa have previously been reported to inherit a mixture of European-related and Sub-Saharan-African-related ancestries.[24,43] It has also been reported that recent gene flow from Sub-Saharan African population has contributed to the

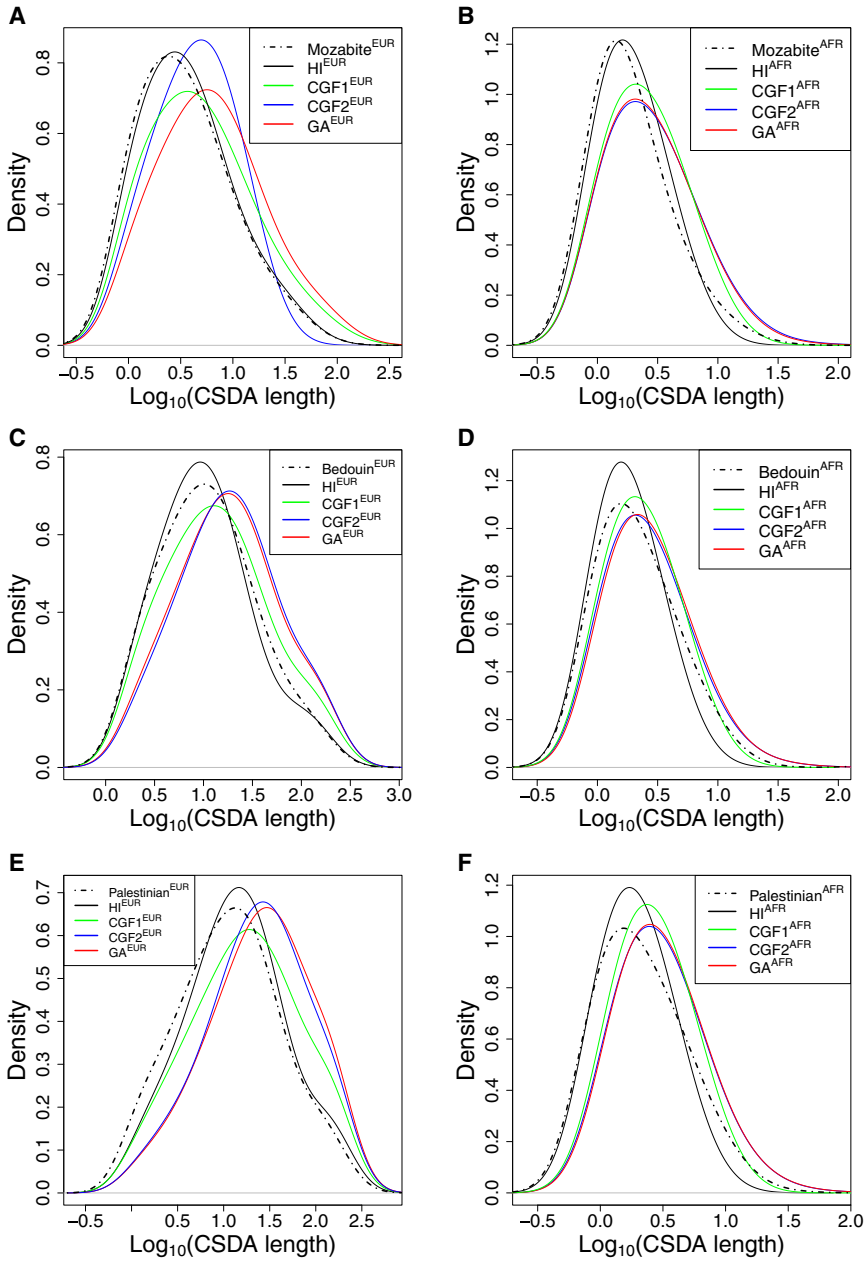

Figure 6. Recent Gene Flows from Sub-Saharan Africa Contributed to the Gene Pools of Mozabite, Bedouin, and Palestinian

The empirical distributions of CSDA length for European ancestral component in Mozabite, Bedouin, and Palestinian were found to fit the HI model best. Although the HI model is essentially fit for the empirical distributions of CSDAs for Sub-Saharan African ancestral component, there have been recent gene flows from Sub-Saharan Africa to each of the admixed populations, as there are more long CSDAs in the empirical distributions of CSDAs for Sub-Saharan African ancestral component than in the HI model.

The empirical distribution and the simulated distributions of CSDA length for (A) European ancestral component in Mozabite, for (B) Sub-Saharan African ancestral component in Mozabite, for (C) European ancestral component in Bedouin, for (D) Sub-Saharan African ancestral component in Bedouin, for (E) European ancestral component in Palestinian, and for (F) Sub-Saharan African ancestral component in Palestinian.

Mozabite gene pool based on analysis of two individuals with the highest proportion of African ancestral component.[24] In this study, the CSDAs of the Mozabite population were obtained with HAPMIX by setting $\lambda = 100$ based on the previous report.[24] Comparing the empirical distribution of CSDAs with that simulated, we found that the Mozabite admixture process essentially fit the HI model with 100 generations since admixture. There was an almost complete absence of recent gene flow from European populations to the Mozabite gene pool (Figure 6A). For the Sub-Saharan African ancestral component, there were more long CSDAs at the tail of empirical distribution than those in the HI model, which confirmed that recent gene flow from African populations had contributed to the Mozabite gene pool (Figure 6B). In summary, we suggest that the Mozabite population could have been formed mainly through one admixture event about 100 generations ago, with a few Sub-Saharan Africans intermarried with Mozabites recently.

Analyses of European ancestral component in Bedouin and Palestinian populations also showed that the empirical distributions essentially fit the HI model for both populations (Figures 6C and 6E). Although the empirical CSDA distribution of Sub-Saharan African ancestral component also fit the HI model best, both distributions showed a long tail at the right compared with those under the HI model, indicating that recent gene flow from Sub-Saharan Africans also contributed to the two admixed populations (Figures 6D and 6F). In short, the three admixed populations were likely to be formed by an earlier admixture, followed by a few subsequent recent gene flows from Sub-Saharan African populations. For Druze, their European component of ancestry fit the HI model very well. However, their African ancestral component contained much shorter CSDAs than those of simulated (Figure S14), which might indicate that previous studies had underestimated the admixture time of Druze. In addition, populations receiving recent gene flow from their parental populations showed higher variation of individual ancestral proportions than those who did not (Figure S13).

Discussion

Interethnic marriage is influenced by various social, cultural, economic, and geographical factors, such as population migration, recolonization, ethnic conflict, ethnic discrimination, and caste systems, which can lead to very complex admixture processes. Therefore, we did not expect that the actual admixture processes could be fully explained by any single simplified model. However, in practice, to facilitate the evolutionary and medical studies that rely on the knowledge of admixture dynamics, we suggest that the primary admixture pattern should be revealed. In this study, we proposed two distinct approaches for the inference of population admixture dynamics. Theoretically, distribution of individual ancestry proportion is particularly powerful in revealing the admixture dynamics of recently admixed population. However, this approach requires a very large sample size and is strongly influenced by sampling error. In contrast, we proposed and demonstrated that genome-wide distribution of CSDAs with moderate sample size could reveal the population admixture dynamics. The distribution of CSDA length has been shown to be powerful in distinguishing different admixture models in various scenarios, including relatively ancient admixture. This approach is also insensitive to general demographic events and to fluctuations and uncertainty of effective population size. The distribution of CSDA length, which takes advantage of recombination information, can serve as a good framework to infer population admixture dynamics.

In this study, by comparing the empirical distribution of CSDA length with those of the simplified models, we determined the primary admixture pattern of admixed populations and provided new insights into the admixture dynamics of several typical admixed populations with different admixture histories. First, we showed that two-way admixture dynamics of African Americans best fit the 14-generation CGF model, in which European ancestry continuously contributed to the African American gene pool, among all the four possible scenarios. Second, we showed that Mexican data fitted the 24-generation GA model best, and recent gene flows from European population might have contributed to the Mexican American gene pool. Finally, recent gene flows from Sub-Saharan Africa were found to have contributed to the gene pool of relatively ancient admixed populations such as Mozabite, Bedouin, and Palestinian populations. Some of these gene flows have not yet been reported. These results suggested that admixture might have been more common in human history than previously determined. Our limited knowledge on interethnic marriage may be due to the fact that many populations have not yet been well studied. These results may also indicate that population admixtures that experienced continuous gene flow from one or multiple parental populations could be more common in human history than the most commonly used scenarios, which were simply described and explained with the HI model.

Although the aforementioned analyses were based on two-way admixtures, our approach could easily be extended to multiple-way admixture. For example, we could explore the three-way admixture of Mexicans and African Americans (African, European, and Amerindian ancestral components) by a similar but slightly modified approach. To estimate the genetic contribution of Amerindians to African Americans, we first combined the African and European parental populations and treated them as a single parental population. Then we analyzed the admixture dynamic between Amerindian ancestry and this combined ancestry (Figure S15). In this way, we found that Amerindian ancestry admixed with the combined ancestry were likely to fit the HI model with about 15 generations. However, there were also a few recent gene flows from Amerindians to African Americans. In fact, our observation was, to some extent, supported by historical records.[50,51] The African and Amerindian ancestral populations both were enslaved in the European colonies during the 17th century and Amerindians might contribute most of the gene flow at that period. However, the gene flows from both Amerindian and African populations to African Americans significantly decreased at the end of the Amerindian slave trade around 1730[50] and the abolishment of the transatlantic slave trade in the beginning of the 19th century, respectively.

For Mexicans, we first combined the European and Amerindian parental populations and treated them as one single parental population. In this way, we could analyze the admixture process between the African parental population and this combined parental population. We found that the admixture dynamics of Mexicans could be explained by 16-generation continuous gene flow (CGF) model, in which African populations contributed all their genetic components to Mexicans at about 16 generations ago and the Mexicans continuously received gene flow from both European population and Amerindian populations (Figure S16). The CGF model was also very reasonable compared with the other models considering that Atlantic slave trade mainly occurred before the end of the 18th century and the continuous inflow of European immigrants. Analysis of Mexicans based on 3-wave admixture model via LAMP was essentially consistent with the results of HAPMIX.

The admixed populations in the New World such as African Americans are widely used in the identification of disease-associated genetic variants through admixture mapping. The effects of admixture dynamics on the pattern of LD have been analyzed in many studies.[9,15,52,53] However, most previous studies simulated the African American population simply with the HI model and assumed the admixture time of only 6–8 generations, which were the average values indicated by genetic data.[23,46] The real statistical power in admixture mapping may have been significantly affected in those studies because the admixture dynamic of African Americans, as shown in this study, are more likely to fit the 14-generation

CGF model in which European ancestry continuously contributed to the African American gene pool. We suggest that future studies should simulate African Americans with the CGF model for accurately evaluating the statistical power of admixture mapping. We also explored the admixture dynamic of Mexicans and obtained useful parameters for the designation of admixture mapping with Mexicans. Until now, the relative ancient admixed populations have not been used for admixture mapping. People generally assume that the extended LD has significantly decayed given the long history of these populations, thus providing limited power in admixture mapping. Here, we demonstrated that three ancient admixed populations have received recent gene flow from their putative parental populations. These results suggested that populations such as the Mozabite, Bedouin, and Palestinian populations might still be suitable for admixture mapping given that the recent gene flow from their putative parental populations could in theory have created new LD.

The efficiency of the CSDA distribution in revealing the population admixture dynamics depends on the accuracy of the inferred CSDAs. In this study, we used existing methods to infer population ancestry and locus-specific ancestry for obtaining the CSDAs in admixed populations. We mainly used HAPMIX for CSDA inference because it outperforms other methods and software in most cases.[24,34] Although HAPMIX is highly accurate and sensitive for inferring CSDA in recent two-way admixed populations, there are many short/tiny CSDAs that may come from a third population which are unavoidable in reality or are due to the limited resolution and accuracy for the inference of breakpoint boundaries. We removed the short CSDAs because the long CSDAs alone were sufficient to reveal the population admixture dynamics. Our approach should be especially helpful in revealing the main admixture patterns in recently admixed populations and in distinguishing the ancient and recent gene flows from their parental populations as we have demonstrated in the empirical analysis of samples from African Americans, Mexican Mestizo, Mexican American, and HGDP populations. This approach could be easily applied to other admixed populations such as the Uyghurs.[20,21,54] We believe that our approach can be continually improved because the accuracy of CSDA inference would be improved with new methods that are under development, allowing even more elaborate evaluations of population admixture dynamics in the future.

Supplemental Data

Supplemental Data include 16 figures and 5 tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

These studies were supported by the National Science Foundation of China (NSFC) grants 31171218 (S.X.), 30971577 (S.X.), and 30890034 (L.J.), by the Shanghai Rising-Star Program 11QA1407600 (S.X.), and by the Science Foundation of the Chinese Academy of Sciences (CAS) (KSCX2-EW-Q-1-11; KSCX2-EW-R-01-05; KSCX2-EW-J-15-05) (S.X.). This research was supported in part by the Ministry of Science and Technology (MoST) International Cooperation Base of China. S.X. is a 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 and the support of K.C.Wong Education Foundation, Hong Kong.

Received: April 5, 2012
Revised: June 1, 2012
Accepted: September 11, 2012
Published online: October 25, 2012

Web Resources

The URLs for data presented herein are as follows:

Coriell Cell Repositories, http://ccr.coriell.org
EIGENSOFT, http://genepath.med.harvard.edu/~reich/Software.htm
fastPHASE, http://stephenslab.uchicago.edu/software.html
FRAPPE and SABER, http://med.stanford.edu/tanglab/software/
HAPMIX, http://www.stats.ox.ac.uk/~myers/software.html
HGDP-CEPH project, ftp://www.cephb.fr/hgdp_supp1/
illumina iControl Database (iControlDB), http://www.illumina.com/science/icontroldb.ilmn
International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
Mexican Genetic Diversity Project (MGDP), ftp://ftp.inmegen.gob.mx/
NIGMS Human Genetic Cell Repository at Coriell, http://ccr.coriell.org/Sections/Collections/NIGMS/GenotypeCopyData.aspx?PgId=564&coll=GM
PLINK v1.06, http://pngu.mgh.harvard.edu/~purcell/plink/
R, http://www.r-project.org/
STRUCTURE, http://pritch.bsd.uchicago.edu/software.html

References

1. Verdu, P., and Rosenberg, N.A. (2011). A general mechanistic model for admixture histories of hybrid populations. Genetics *189*, 1413–1426.
2. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.
3. Abdulla, M.A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S.K., Calacal, G.C., Chaurasia, A., Chen, C.H., Chen, J., Chen, Y.T., et al.; HUGO Pan-Asian SNP Consortium; Indian Genome Variation Consortium. (2009). Mapping human genetic diversity in Asia. Science *326*, 1541–1545.
4. Chakraborty, R., and Weiss, K.M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. USA *85*, 9119–9123.
5. McKeigue, P.M. (1997). Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am. J. Hum. Genet. *60*, 188–196.
6. McKeigue, P.M. (1998). Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am. J. Hum. Genet. *63*, 241–251.

7. Montana, G., and Pritchard, J.K. (2004). Statistical tests for admixture mapping with case-control and cases-only data. Am. J. Hum. Genet. *75*, 771–789.
8. Stephens, J.C., Briscoe, D., and O'Brien, S.J. (1994). Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am. J. Hum. Genet. *55*, 809–824.
9. Pfaff, C.L., Parra, E.J., Bonilla, C., Hiester, K., McKeigue, P.M., Kamboh, M.I., Hutchinson, R.G., Ferrell, R.E., Boerwinkle, E., and Shriver, M.D. (2001). Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am. J. Hum. Genet. *68*, 198–207.
10. Long, J.C. (1991). The genetic structure of admixed populations. Genetics *127*, 417–428.
11. Ewens, W.J., and Spielman, R.S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. Am. J. Hum. Genet. *57*, 455–464.
12. Parra, E.J., Kittles, R.A., Argyropoulos, G., Pfaff, C.L., Hiester, K., Bonilla, C., Sylvester, N., Parrish-Gause, D., Garvey, W.T., Jin, L., et al. (2001). Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am. J. Phys. Anthropol. *114*, 18–29.
13. Pool, J.E., and Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics *181*, 711–719.
14. Adams, J., and Ward, R.H. (1973). Admixture studies and the detection of selection. Science *180*, 1137–1143.
15. Guo, W., and Fung, W.K. (2006). The admixture linkage disequilibrium and genetic linkage inference on the gradual admixture population. Yi Chuan Xue Bao *33*, 12–18.
16. Zakharia, F., Basu, A., Absher, D., Assimes, T.L., Go, A.S., Hlatky, M.A., Iribarren, C., Knowles, J.W., Li, J., Narasimhan, B., et al. (2009). Characterizing the admixed African ancestry of African Americans. Genome Biol. *10*, R141.
17. Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S.A., and Bustamante, C.D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA *107*, 786–791.
18. Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J., Fernandez-Lopez, J.C., Uribe-Figueroa, L., Contreras, A., Balam-Ortiz, E., del Bosque-Plata, L., Velazquez-Fernandez, D., Lara, C., et al. (2009). Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. Proc. Natl. Acad. Sci. USA *106*, 8611–8616.
19. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., and Ostrer, H. (2010). Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc. Natl. Acad. Sci. USA *107* (*Suppl 2*), 8954–8961.
20. Xu, S., Huang, W., Qian, J., and Jin, L. (2008). Analysis of genomic admixture in Uyghur and its implication in mapping strategy. Am. J. Hum. Genet. *82*, 883–894.
21. Xu, S., and Jin, L. (2008). A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. Am. J. Hum. Genet. *83*, 322–336.
22. Xu, S., and Jin, L. (2011). Chromosome-wide haplotype sharing: a measure integrating recombination information to reconstruct the phylogeny of human populations. Ann. Hum. Genet. *75*, 694–706.
23. Seldin, M.F., Morii, T., Collins-Schramm, H.E., Chima, B., Kittles, R., Criswell, L.A., and Li, H. (2004). Putative ancestral origins of chromosomal segments in individual african americans: implications for admixture mapping. Genome Res. *14*, 1076–1084.
24. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. *5*, e1000519.
25. Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. Am. J. Hum. Genet. *79*, 1–12.
26. Falush, D., Stephens, M., and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics *164*, 1567–1587.
27. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al.; International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
28. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
29. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.
30. Wright, S. (1931). Evolution in Mendelian populations. Genetics *16*, 97–159.
31. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
32. Tang, H., Peng, J., Wang, P., and Risch, N.J. (2005). Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. *28*, 289–301.
33. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
34. Jin, W., Xu, S., Wang, H., Yu, Y., Shen, Y., Wu, B., and Jin, L. (2012). Genome-wide detection of natural selection in African Americans pre- and post-admixture. Genome Res. *22*, 519–527.
35. Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. *78*, 629–644.
36. Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. Am. J. Hum. Genet. *82*, 290–303.
37. Ihaka, R., and Gentleman, R. (1996). R: A language for data analysis and graphics. J. Comput. Graph. Statist. *5*, 299–314.
38. Lilliefors, H.W. (1967). On Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc. *62*, 399–402.
39. Hitchcock, F.L. (1941). The distribution of a product from several sources to numerous localities. J. Math. Phys. *20*, 224–230.
40. Rubner, Y., Tomasi, C., and Guibas, L.J. (2000). The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. *40*, 99–121.

41. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
42. Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Beckman, K., Thyne, S., Chapela, R., Rodriguez-Santana, J.R., et al. (2009). Ancestry-related assortative mating in Latino populations. Genome Biol. *10*, R132.
43. Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., and Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet. *7*, e1001373.
44. Wang, J. (2003). Maximum-likelihood estimation of admixture proportions from genetic data. Genetics *164*, 747–765.
45. Smith, M.W., Patterson, N., Lautenberger, J.A., Truelove, A.L., McDonald, G.J., Waliszewska, A., Kessing, B.D., Malasky, M.J., Scafe, C., Le, E., et al. (2004). A high-density admixture map for disease gene discovery in african americans. Am. J. Hum. Genet. *74*, 1001–1013.
46. Tian, C., Hinds, D.A., Shigeta, R., Kittles, R., Ballinger, D.G., and Seldin, M.F. (2006). A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am. J. Hum. Genet. *79*, 640–649.
47. Tian, C., Hinds, D.A., Shigeta, R., Adler, S.G., Lee, A., Pahl, M.V., Silva, G., Belmont, J.W., Hanson, R.L., Knowler, W.C., et al. (2007). A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. Am. J. Hum. Genet. *80*, 1014–1023.
48. Wang, S., Ray, N., Rojas, W., Parra, M.V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., Hurtado, A.M., et al. (2008). Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet. *4*, e1000037.
49. Price, A.L., Patterson, N., Yu, F., Cox, D.R., Waliszewska, A., McDonald, G.J., Tandon, A., Schirmer, C., Neubauer, J., Bedoya, G., et al. (2007). A genomewide admixture map for Latino populations. Am. J. Hum. Genet. *80*, 1024–1036.
50. Seybert, T. (2004). Slavery and Native Americans in British North America and the United States: 1600 to 1865. New York Life. http://web.archive.org/web/20040804001522/http://www.slaveryinamerica.org/history/hs_es_indians_slavery.htm.
51. Gallay, A. (2002). The Indian Slave Trade: The Rise of the English Empire in the American South 1670-1717 (New Haven, CT: Yale University Press).
52. Pfaff, C.L., Kittles, R.A., and Shriver, M.D. (2002). Adjusting for population structure in admixed populations. Genet. Epidemiol. *22*, 196–201.
53. Smith, M.W., and O'Brien, S.J. (2005). Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat. Rev. Genet. *6*, 623–632.
54. Xu, S., Jin, W., and Jin, L. (2009). Haplotype-sharing analysis showing Uyghurs are unlikely genetic donors. Mol. Biol. Evol. *26*, 2197–2206.