Published in final edited form as:

Methods Mol Biol. 2012 ; 850: 465–481. doi:10.1007/978-1-61779-555-8_25.

The Analysis of Ethnic Mixtures

Xiaofeng Zhu

Abstract

Populations of ethnic mixtures can be useful in genetic studies. Admixture mapping, or mapping by admixture linkage disequilibrium (MALD), is specially developed for admixed populations and can supplement traditional genome-wide association analyses in the search for genetic variants underlying complex traits. Admixture mapping tests the association between a trait and locus-specific ancestries. The locus-specific ancestries are in linkage disequilibrium (LD) which is generated by the admixture process between genetically distinct ancestral populations. Because of highly correlated locus-specific ancestries, admixture mapping performs many fewer independent tests across the genome than current genome-wide association analysis. Therefore, admixture mapping can be more powerful because of the smaller penalty due to multiple tests. In this chapter, I introduce the theory behind admixture mapping and how we conduct the analysis in practice.

Keywords

Admixture mapping; Population admixture; Ancestry information marker; Hidden Markov model

1. Introduction

In genetic epidemiology, we always want to study the relationship between a phenotype and a genetic marker. A popular design is a retrospective case–control design for a binary trait or a population-based design for a quantitative trait. Association for a genetic marker can be established by performing logistic regression or linear regression analysis. When the study samples are collected from a recent admixed population such as African Americans or Mexican Americans, each subject's chromosome has a mosaic structure of chromosome segments that come from ancestral populations. Intuitively, we are able to test the association between the ancestry at any position of the genome and a disease trait, given such information is available. For example, a 2 by 2 table can be created and standard statistical methods for an association test can be applied (Table 1). The underlying assumption is that the risk allele at a locus occurs at different frequencies among ancestral populations. When this is true, we expect that, in the admixed population, affected individuals share an excess ancestry from the ancestral population with the highest frequency of the risk allele. Figure 1 illustrates such chromosomes sampled at the current generation when admixture occurs in two ancestral populations. Assume that the dark chromosomes are from an ancestral population that has high disease prevalence and the light chromosomes are from the other ancestral population whose disease prevalence is low. As an ideal case, we expect that all the chromosomes at a disease locus in affected cases are inherited from one ancestral population. In comparison, controls will be less likely to carry dark chromosomes. Caution should be taken when performing the association test in Table 1. As in many association studies of population-based samples, confounding is a serious problem. Since the disease prevalence is different in two ancestral populations, an affected

individual is more likely to carry chromosomes from a high-risk ancestral population than from a low risk one, a phenomenon of population structure. In fact, a disease variant, rather than the ancestry itself, contributes to the phenotypic variation. Thus, analyzing the data in Table 1 should take care of the effect of population structure. In general, admixture mapping methods can be simply viewed as testing association between a locus-specific ancestry and a phenotype, meanwhile controlling the effect of population structure. The test statistics can be built by comparing the locus-specific ancestry between cases and controls or by comparing the locus-specific ancestry to the ancestry distribution across the genome among cases only.

1.1. Test Statistics

Mathematical models for admixture mapping can be found in the literature (1–9). Suppose we have an admixed population C resulting from two ancestral populations, X and Y . Let $\Pi_d(\theta)$ and $\Pi_c(\theta)$ be the proportion of alleles that are from ancestral population X among cases and controls in the current admixed population, respectively, where θ represents the genetic distance between the disease location and the candidate marker. The null hypothesis is that the marker is unlinked to the disease risk, or $\theta = 0.5$ between a marker locus and a disease locus. In a case-only design, we test the null hypothesis: $\Pi_d(\theta) = \Pi_d(0.5)$, and in a case-control design, we test the null hypothesis: $\Pi_d(\theta) - \Pi_d(0.5) = \Pi_c(\theta) - \Pi_c(0.5)$.

When we know which ancestral populations an individual's alleles at any marker locus are from, we would be able to estimate Π_d and Π_c at any genomic position, which are estimated by the frequencies of ancestry present in cases and controls, respectively. $\Pi_d(0.5)$ and $\Pi_c(0.5)$ are estimated by the average of ancestry across the genome in cases and controls, respectively. Let $\hat{\Pi}_d(t)$ be the estimated proportion of ancestry from population X at chromosome location t , conditional on the observed marker genotypes. A test statistic for the case-only design is

$$Z_c(t) = \frac{\hat{\Pi}_d(t) - \hat{\Pi}_d(\theta=0.5)}{\sigma(\hat{\Pi}_d(t))}, \quad (1)$$

and a test statistic for the case-control design is

$$Z_{cc}(t) = \frac{(\hat{\Pi}_d(t) - \hat{\Pi}_d(\theta=0.5)) - (\hat{\Pi}_c(t) - \hat{\Pi}_c(\theta=0.5))}{\sigma(\hat{\Pi}_d(t) - \hat{\Pi}_c(t))}, \quad (2)$$

respectively (3, 4, 7). Neither test is affected by population structure because we are testing the excess of ancestry at a marker position.

Consider a study consisting of n_1 unrelated cases and n_2 unrelated controls genotyped at M marker loci. Let x_{ij} be the proportion of alleles from ancestral population X for the i th individual at marker j . At marker j we have

$$\hat{\Pi}_d(j) = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{ij}$$

and

$$\hat{\Pi}_c(j) = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} x_{ij},$$

for cases and controls, respectively. Similarly, we have

$$\hat{\Pi}_d(\theta=0.5) = \frac{1}{Mn_1} \sum_{j=1}^M \sum_{i=1}^{n_1} x_{ij}$$

and

$$\hat{\Pi}_c(\theta=0.5) = \frac{1}{Mn_2} \sum_{j=1}^M \sum_{i=n_1+1}^{n_1+n_2} x_{ij},$$

respectively. Here we assume that only a few loci will contribute to disease disparity among ancestral populations, which is reasonable. We estimate the variance in Eqs. 1 and 2 by

$$\sigma^2(\hat{\Pi}_d) = \frac{1}{M} \sum_{j=1}^M [\hat{\Pi}_d(j) - \hat{\Pi}_d(\theta=0.5)]^2 \quad (3)$$

and

$$\sigma^2(\hat{\Pi}_d - \hat{\Pi}_c) = \frac{1}{M} \sum_{j=1}^M [(\hat{\Pi}_d(j) - \hat{\Pi}_c(j)) - (\hat{\Pi}_d(\theta=0.5) - \hat{\Pi}_c(\theta=0.5))]^2. \quad (4)$$

The rationale for estimating the variance by Eqs. 3 and 4 is that the proportion of X by descent at any locus approximately comes from the same distribution when the locus is not linked with a trait locus. The variance estimated in this way has been theoretically shown to be asymptotically unbiased (10).

A likelihood-based method can be also applied (6). Let $r = \frac{P(\text{disease}|\text{both alleles from } X)}{P(\text{disease}|\text{no both alleles from } X)}$ be the ancestry risk ratio at the locus in a study under the assumption of a multiplicative model. Let λ be the admixture proportion from the high-risk parental population (i.e., population X). Then $\Pi_d(\theta = 0)$ for an intermixture model is (2, 3)

$$\Pi_d(\theta=0) = \frac{\lambda \sqrt{r}}{\lambda \sqrt{r} + 1 - \lambda}.$$

Thus the likelihood of the observed ancestral alleles at marker j is

$$L(r) = \prod_i \frac{(\lambda \sqrt{r})^{x_{ij}} (1 - \lambda)^{1-x_{ij}}}{(\lambda \sqrt{r} + 1 - \lambda)}.$$

A standard likelihood ratio test or score test can be carried out to test the null hypothesis $r = 1$ (6).

For the case-control test, a logistic regression can be applied (6), which is

$$\log \frac{P(y_i|x_{ij})}{1 - P(y_i|x_{ij})} = \beta_0 + \beta_1(x_{ij} - \bar{x}_i),$$

where y_i is the disease status for individual i , $\bar{x}_i$ is the average ancestry for individual i , and β_1 is the log odds ratio of disease for individuals with 2 vs. 0 allele copies from the high-risk parental population. The null hypothesis is $\beta_1 = 0$.

It is straightforward to extend the above method to a quantitative trait (11). A linear regression can be directly applied as

$$y_i = \beta_0 + \beta_1(x_{ij} - \bar{x}_i) + \varepsilon_i,$$

where the null hypothesis is $\beta_1 = 0$.

1.2. Inferring Locus-Specific Ancestry

It is straightforward to perform admixture mapping analysis if we know the locus-specific ancestry. When only marker genotypes are available, statistical methods to infer locus-specific ancestry have been developed (4–7, 12–15). A typical method of inferring locus-specific ancestry is based on a hidden Markov model. Let $\{g_t\}_{t=1}^M$ denote M ordered observed genotypes along a chromosome and $\{v_t\}_{t=1}^M$ the number of alleles being X by descent at the corresponding marker loci. We can model $\{g_t, v_t\}_{t=1}^M$ in an HMM, as illustrated:

$$\begin{array}{ccccc} \text{observed} & \text{genotype} & g_1 & g_2 & \cdots & g_M \\ & & \uparrow & \uparrow & & \uparrow \\ \text{Hidden} & \text{States} & v_1 & \rightarrow v_1 & \cdots \rightarrow & v_M \end{array},$$

by assuming conditional independence given the underlying unobservable states, that is, $P(g_t|g_1, \dots, g_{t-1}, v_1, \dots, v_t) = P(g_t|v_t)$. This assumption may not be true when we have dense markers available, such as the markers used for a genome-wide association analysis, in which pairwise linkage disequilibrium in ancestral populations are often present. In contrast, in the marker hidden Markov model (MHMM) proposed by Tang et al. (12), the observed state g_t depends not only on v_t but also on the past history, as illustrated by

$$\begin{array}{ccccccc} \text{observed} & \text{genotype} & g_1 & \rightarrow & g_2 & \cdots \rightarrow & g_M \\ & & \uparrow & & \uparrow & & \uparrow \\ \text{Hidden} & \text{States} & v_1 & \rightarrow & v_2 & \cdots \rightarrow & v_M \end{array}.$$

For computational tractability, Tang et al. (12) consider only the first-order Markovian dependence, that is,

$$P(g_t|g_1, \dots, g_{t-1}, v_1, \dots, v_t) = \begin{cases} P(g_t|g_{t-1}, v_t) & \text{if } v_t = v_{t-1} \\ P(g_t|v_t) & \text{otherwise} \end{cases},$$

Thus, MHMM is more general than HMM and has the advantage of allowing for background linkage disequilibrium in ancestral populations.

The transition matrix can be obtained based on a continuous gene-flow model as presented by Zhu et al. (3). An alternative flexible transition matrix is implemented in STRUCTURE (15, 16), which assumes an intermixing model, that is, all chromosomes in the sampled admixed subjects descended from a mixed group of ancestral chromosomes n generations ago, who have subsequently mated randomly (27).

2. Methods

In this section, I describe the procedures for conducting admixture mapping analysis in practice when both genotype and phenotype data are available.

2.1. Step 1. Quality Controls

When raw genotype data are obtained, quality controls are necessary before the formal data analysis is performed. The typical genotype data include customer-designed chips, such as iSelect Custom Bead-Chip, or standard arrays used for whole genome association studies such as Affymetrix 5.0 and 6.0 platforms, or Illumina Human669W-Quad or HumanOmni1-Quad. The standard QCs include removing either individuals or SNPs because of low calling rate. For example, an array with calling rate less than 0.9 may be removed and an SNP with calling rate less than 0.95 can also be removed. Illumina platforms use the software GenomeStudio to make the genotype calling. An important parameter is the GenTrain Score ranging from 0 to 1, which is a score calculated from the GenTrain clustering algorithm. SNPs are often sorted by GenTrain score in the SNP table. SNPs with lower scores have poor clustering in the SNP graph and should be excluded in the analysis.

The next level of QC includes examining heterozygosity, which measures the degree of inbreeding. Too low or too high a heterozygosity (defined as < -4 SD or > 4 SD beyond the mean) indicates possible DNA contamination or poor DNA quality. In admixture mapping analysis, the subjects are assumed to be unrelated. To check the relatedness in the samples, a pairwise identity-by-descent (IBD) score is examined for each pair of samples. The outliers of IBD scores can be determined based on the distribution of the IBD scores. One of a pair of subjects within the outliers should be removed. The subject with lower genotyping rate is usually selected to be removed. In addition, samples with IBD 5% or more with other samples are also removed because of possible DNA contamination. Multidimensional scaling (MDS) or principal components (PCs) are used to estimate population substructure, and the identified outliers are excluded in analysis. The results of MDS and PCs are usually robust, as indicated in Fig. 2, where the subjects in the circles indicate these samples are related and may be excluded from the analysis. All the above procedures can be performed by the popular software PLINK (17). However, we do not suggest using HWD to filter SNPs (see Note 1).

2.2. Step 2. Inferring an Individual's Locus-Specific Ancestry (Local Ancestry)

After QCs, we assume all the markers are correctly genotyped. If AIMs are selected, we can use the program ADMIXPROGRAM (4), STRUCTURE (15, 16), or ANCESTRYMAP (5) to estimate the locus-specific ancestry. We usually assume AIMs are in linkage equilibrium in ancestral populations. Thus, we would like to examine whether any AIMs are in linkage disequilibrium in the ancestral populations. We can examine the LD among the AIMs in

¹Hardy-Weinberg Equilibrium (HWE) is often used for QCs to exclude SNPs. However, HWE can be created through the population admixture process. For example, suppose two ancestral populations have genotype frequencies (0.01, 0.18, 0.81) and (0.81, 0.18, 0.01) for genotypes (AA, Aa, aa), respectively. Assuming the admixture rate is 50/50%, then the genotype frequencies in admixed population are (0.41, 0.18, 0.41) with allele frequencies being 0.5 for A and a. HWE is violated in the admixed population. Thus, we do not suggest using HWD to filter SNPs.

HapMap data using the software PLINK. Only one AIM should be selected if several AIMs are in strong LD in ancestral populations.

For the program ADMIXPROGRAM, the input genotype and phenotype files have the same format as STRUCTURE. The first row refers to the genetic distance between two neighboring AIMs. The distance between the last AIM on one chromosome and the first SNP on the next chromosome is coded as -1. The first four columns refer to individual ID, population information, gender, and affected status. The remaining columns are the genotype codes, two columns per AIM, with each column indicating one allele coded by 0 or 1. An example of the input file is presented in Table 2. The parameter file includes the number of individuals, whether the ancestral population allele frequency file is provided, and initial number of generations since admixture events occurred (see Table 3). The parameter file also requires providing which set of AIMs will be used to analyze the data, for example, all the markers, or even or odd markers only. If any AIMs should be excluded in the analysis, it can also be flagged. The program will output either case-only or case-control *Z*-scores for admixture mapping analysis, as requested by the user, as well as the locus-specific ancestries for each individual.

The results from ADMIXPROGRAM can be examined in several ways. The estimated average admixture rate can be examined as the first step. If the AIMs do not provide enough information for estimating the average admixture rate, it is likely to have 50%/50% admixture rate for two ancestral populations. For African Americans, the estimate of admixture rate is usually close to 20% European and 80% African ancestry, that is, a 20/80% admixture. The software also outputs the estimated allele frequencies in ancestral populations. Although these frequencies will not be the same as that observed in HapMap samples, for example, HapMap CEU and YRI, these frequencies should be highly correlated. However, if a region is under strong selection pressure, we may identify substantial differences between the estimated allele frequencies and those observed in the HapMap data. It should be noticed that we have not observed any regions with such strong selection evidence that leads to strong deviation of allele frequencies, except in Puerto Ricans, which are admixed by Europeans, native Indians, and Africans (18). For example, Zhu and Cooper (19) used ADMIXPROGRAM to estimate the allele frequencies of the AIMs in European and African ancestral populations using the 1,743 African Americans enrolled in the Dallas Heart Study (DHS). The estimated allele frequencies in ancestry populations using ADMIXPROGRAM and the corresponding current European and African allele frequencies estimated from European Americans in DHS and Africans from the literature (26) (for the AIMs) were plotted in Fig. 3. High correlation of observed and estimated allele frequencies can be found in Fig. 3 (correlation coefficient >0.97). However, it may happen that the allele frequency estimates in ancestry populations may be flipped (Fig. 4). In this case, different initial allele frequencies should be used in ADMIXPROGRAM. One way is to switch the estimated allele frequencies in the ancestral populations, and these new allele frequencies can be used as the initial allele frequencies in the next run of ADMIXPROGRAM. The software ADMIXPROGRAM usually automatically switches the estimated ancestral population allele frequencies. We also suggest running ADMIXPROGRAM at least three times: (1) without using ancestral population allele frequencies, (2) using one ancestral population frequencies provided, and (3) using both ancestral population frequencies provided. The final results will be compared by examining the negative log-likelihood function values. If the program runs correctly, all the negative log-likelihood function values should be similar.

STRUCTURE (15, 16) and ANCESTRYMAP (5) are popular programs for inferring locus-specific ancestries when only AIMs are available. STRUCTURE allows more than two ancestral populations while both ADMIXPROGRAM and ANCESTRYMAP only allow two

ancestral populations. STRUCTURE provides many features to model admixed populations, including no admixture model, admixture model, and linkage model. For admixed population such as African Americans or Mexican Americans, the linkage model is suggested to give the best results. Both STRUCTURE and ANCESTRYMAP are based on MCMC algorithm. STRUCTURE requires users to provide burn-in length and the number of runs after burn-in. A burn-in of 10,000–100,000, with additional 10,000–100,000 runs, is suggested. This is very time consuming when the number of subjects is over 5,000 and the number of AIMs is over 3,000. When AIMs are well selected (Highly informative AIMs), we have found that a burn-in length of 5,000 followed by an additional 5,000 runs are usually adequate.

When genome-wide scan data are available, such as 500K SNPs or more, the AIMs can be selected from the available data although information may be lost. Statistical methods and software have been developed to analyze such large data sets. SABER (12) applies MHMM to incorporate the dense SNPs with possible background LD, which is typical for dense SNPs such as 500K or more. Other similar softwares are listed in Table 4. Because the dense SNPs are used to reconstruct ancestry blocks, these methods are more accurate than using AIMs but at a cost of more computation time.

2.3. Step 3. Association Analysis of Testing Locus-Specific Ancestry

Once the locus-specific ancestries are inferred, testing the association of locus-specific ancestry and a phenotype is quite straightforward. However, certain data manipulations are necessary. For case-only data, ADMIXPROGRAM, STRUCTURE/MALDSoft, ANCESTRYMAP, and ADMIXMAP can output either *Z*-score or LOD score at AIMs. For case-control data or quantitative traits, the logistic regression or linear regression analysis can be applied using SAS or *R* statistical packages. Let *reg1* be a SAS data set including phenotype, covariates, and ancestry proportion at each marker, which can be obtained from the output of ADMIXPROGRAM or STRUCTURE. We can apply the following SAS macro for association testing.


```

%macro ancestry1(num);

%do i=1 %to &num;

ods listing close;

ods output "Parameter Estimates"=parest_&i;

proc logistic data=reg1;

model y = covariate average_ancestry ancestry&i ;

run; quit;

ods output close;

ods listing;

%end;

%mend;

%macro ancestry2(num);

%do i=2 %to &num;

proc append base=parest_1 data=parest_&i;

%end; %mend;

data prob_1; set parest_1; P = -LOG10(Probt); if Variable
ne 'covariate' and Variable ne 'average_ancestry' and
Variable ne 'Intercept';

run;

```

In the SAS data file prob_1, the following informatin will be outputted for each AIM. The P value is testing the association between a marker specific ancestry and the phenotype.

Estimate	Std. err.	t Value	P	−log10(P)
0.174361	0.657886	0.265032	0.790994	0.101827
0.095116	0.64143	0.148287	0.882121	0.054472
0.019573	0.637086	0.030722	0.975492	0.010776
0.015492	0.629289	0.024618	0.98036	0.008614
−0.02246	0.613737	−0.03659	0.970814	0.012864
0.020716	0.613837	0.033749	0.973078	0.011852
0.245184	0.619785	0.395596	0.692417	0.159633
0.318612	0.620987	0.513075	0.607917	0.216156

The Z -scores from a case-only analysis or P values can be summarized as shown in Fig. 5. When a large Z -score, such as greater than 3, is observed, additional analysis should be performed to determine whether the significant result is due to genotyping error (see Note 2). Since admixture mapping analysis is testing the association between a phenotype and locus-specific ancestries, which can be highly correlated because of the recent admixture,

²When a large Z -score is observed, caution should be exercised before claiming the success of the analysis. It is possible that a large Z -score is driven by some specific SNPs whose genotyping qualities are questionable or if the SNPs are in strong LD. One way to examine this problem is to redo the analysis using half the markers, for example, only even or odd numbered markers. If the result after reducing markers is consistent with that without dropping markers, the observed large Z -score may be robust.

the number of total independent tests in whole genome can be much less than the number of actual tests conducted in admixture mapping analysis. It has been suggested that there are about 1,000 independent tests (7), which is consistent with that in a simulation study for African American samples (4). For case-control analysis or quantitative traits, a permutation test can also be applied to obtain the empirical P values. However, the permutation test should be performed by permuting individual's phenotype and covariates together, keeping the entire marker data unchanged. Thus, we do not need to reestimate the locus-specific ancestries in each permutation. This permutation method can be applied to determine the empirical genome-wide significance level.

2.4. Additional Remarks

Admixture mapping analysis is similar to linkage analysis when an entire admixed population is considered as a large family. Unlike genome-wide association studies, admixture mapping will identify chromosomal regions that may harbor disease variants. Such regions range from 10 to 20 Mb in length, which may include many genes. When genome-wide data are available, several thousands of markers genotyped by an array may also fall into the regions identified by admixture mapping. Presumably, the follow-up direct association tests can be performed for these SNPs using standard association methods for searching which SNPs are responsible for the evidence observed in admixture mapping analysis. This may also be done by accounting for population stratification, using methods such as principal component analysis (20, 21) or the genomic control approach (22). However, only the makers with substantial allele frequency differences in ancestral populations can contribute to the association evidence in admixture mapping analysis. Thus, rather than testing all available SNPs in a region, we only need to test a subset of markers whose allele frequencies differ between ancestral populations by greater than a predefined value (such as 0.2). This procedure can substantially reduce the number of tests. Furthermore, these selected SNPs can be strongly correlated because of the population admixture. The actual number of tests can be estimated by the methods proposed in (23, 24). To define a region in admixture mapping analysis, we can use the 1-unit drop region from a peak of $-\log_{10}(P \text{ value})$. The size of the region is dependent on the peak value, and the number of regions is dependent on how large a peak we want to follow up in further studies. For example, if we choose a peak having negative $\log_{10}(P \text{ value})$ greater than 3, we will expect one region if we assume there are 1,000 independent regions in the genome under the null hypothesis that there is no genetic variant contributing disease disparity among ancestral populations. This number can be increased to 10 if we select regions having negative $\log_{10}(P \text{ value})$ greater than 2. To determine the significance level of a test, we should add the numbers of independent tests in admixture mapping analysis and in all the regions that have been selected for single marker association tests. Even so, the total number of tests is much less than the number of tests in standard genome-wide association studies, which is typically 500K to 1 million. Typically, a genome-wide significant P value is around 10^{-5} to 10^{-6} , depending upon what criteria (the peak of negative $\log_{10}(P \text{ value})$) is used for selecting regions for follow-up association analyses. This is one of the important advantages in admixture mapping, which reduces the penalty due to a larger number of multiple comparisons. Finally, it should be noted that admixture mapping can be used to detect the risk variants that are less frequent in a high risk ancestral population than in a lower risk ancestral population (Note 3).

³It is usually mistakenly believed that only the risk variants more frequent in high-risk ancestral population can be detected in admixture mapping studies. In fact, admixture mapping can also detect the risk variants that are more frequent in a low-risk ancestral population. In fact, the power of admixture mapping is larger when the risk allele is more frequent in the low-risk ancestral population, with a low admixture rate contribution to the admixed population.

Acknowledgments

This work was supported by grant from National Human Genome Research Institute (HG003054).

References

1. Risch NJ. Mapping genes for complex disease using association studies with recently admixed populations. *Am J Hum Genet.* 1992; 51:13.
2. McKeigue PM. 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.* 1998; 63:241–251. [PubMed: 9634509]
3. Zhu X, Cooper RS, Elston RC. Linkage analysis of a complex disease through use of admixed populations. *Am J Hum Genet.* 2004; 74:1136–1153. [PubMed: 15131754]
4. Zhu X, Zhang S, Tang H, Cooper R. A classical likelihood based approach for admixture mapping using EM algorithm. *Hum Genet.* 2006; 120:431–445. [PubMed: 16896924]
5. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, et al. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet.* 2004; 74:979–1000. [PubMed: 15088269]
6. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. *Am J Hum Genet.* 2004; 74:965–978. [PubMed: 15088268]
7. Montana G, Pritchard JK. Statistical tests for admixture mapping with case-control and cases-only data. *Am J Hum Genet.* 2004; 75:771–789. [PubMed: 15386213]
8. Zhang C, Chen K, Seldin MF, Li H. A Hidden Markov modeling approach for admixture mapping based on case-control data. *Genet Epidemiol.* 2004; 27:225–239. [PubMed: 15389926]
9. Zhu X, Tang H, Risch N. Admixture mapping and the role of population structure for localizing disease genes. *Adv Genet.* 2008; 60:547–569. [PubMed: 18358332]
10. Sha Q, Zhang X, Zhu X, Zhang S. Analytical correction for multiple testing in admixture mapping. *Hum Hered.* 2006; 62:55–63. [PubMed: 17047335]
11. Basu A, Tang H, Arnett D, Gu CC, Mosley T, et al. Admixture mapping of quantitative trait loci for BMI in African Americans: evidence for loci on chromosomes 3q, 5q, and 15q. *Obesity (Silver Spring).* 2009; 17:1226–1231. [PubMed: 19584881]
12. Tang H, Coram M, Wang P, Zhu X, Risch N. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet.* 2006; 79:1–12. [PubMed: 16773560]
13. Sundquist A, Fratkin E, Do CB, Batzoglou S. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res.* 2008; 18:676–682. [PubMed: 18353807]
14. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. *Am J Hum Genet.* 2008; 82:290–303. [PubMed: 18252211]
15. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000; 155:945–959. [PubMed: 10835412]
16. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics.* 2003; 164:1567–1587. [PubMed: 12930761]
17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. PLINK: a tool set for whole-genome association and populationbased linkage analyses. *Am J Hum Genet.* 2007; 81:559–575. [PubMed: 17701901]
18. Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, et al. Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am J Hum Genet.* 2007; 81:626–633. [PubMed: 17701908]
19. Zhu X, Cooper RS. Admixture mapping provides evidence of association of the VNN1 gene with hypertension. *PLoS ONE.* 2007; 2:e1244. [PubMed: 18043751]
20. Zhu X, Zhang S, Zhao H, Cooper RS. Association mapping, using a mixture model for complex traits. *Genet Epidemiol.* 2002; 23:181–196. [PubMed: 12214310]
21. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38:904–909. [PubMed: 16862161]

22. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999; 55:997–1004. [PubMed: 11315092]
23. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*. 2005; 95:221–227. [PubMed: 16077740]
24. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet*. 2004; 74:765–769. [PubMed: 14997420]
25. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*. 2009; 5 e1000519.
26. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, et al. A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet*. 2004; 74:1001–1013. [PubMed: 15088270]
27. Long JC. The genetic structure of admixed populations. *Genetics*. 1991; 127:417–428. [PubMed: 2004712]

Fig. 1.

The mosaic structure of chromosomes of cases and controls sampled from an admixed population with two ancestral populations. The *dark* and *light* segments represent chromosome segments inherited from two ancestral populations. The *vertical line* represents the location of a disease susceptibility variant. When the disease variant has high frequency in *dark* population and rare in the other, more *dark* segments are observed at the disease locus in cases than outside of the disease locus as well as among any regions in controls.

Fig. 2. PCA (*left panel*) and MDS (*right panel*) of the genotype data of the 701 African Americans sampled from Maywood, Illinois. The PCA and MDS result in consistent patterns. The subjects within circles may indicate relatedness and special attention should be paid.

Fig. 3.

The comparisons of estimated allele frequencies in ancestral populations and observed allele frequencies in European and African populations (19). A. European ancestral allele frequencies estimated by ADMIXPROGRAM in the African American sample vs. the observed allele frequencies in European Americans. B. African ancestral allele frequencies estimated by ADMIXPROGRAM in the African American sample vs. the observed allele frequencies by the weighted average from Ghana and Cameroon obtained from Smith et al. (26). The points on the off diagonal line suggest the SNPs have switched allele labels.

Fig. 4.
The comparisons of estimated allele frequencies in ancestral populations and observed allele frequencies in European and African populations due to switches of ancestral allele frequency estimated between two ancestral populations.

Fig. 5. The genome-wide Z -scores of admixture mapping in the Dallas Heart Study (19). *Top*: the Z -scores calculated using hypertensive cases only; *Bottom*: the Z -scores calculated based on case-control samples.

Table 1

A 2 by 2 table of testing association between an ancestry and a disease trait in samples from an African American population

Ancestry	Cases	Controls	Odds ratio
AA	$n_{AA}^{(D)}$	$n_{AA}^{(C)}$	Θ_{AA}
EA	$n_{EA}^{(D)}$	$n_{EA}^{(C)}$	
Total	$n_{AA}^{(D)} + n_{EA}^{(D)}$	$n_{AA}^{(C)} + n_{EA}^{(C)}$	$2n$

AA African-ancestral allele, EA European ancestral allele. Superscript D: case; C: control

Table 2

The input file for ADMIXPROGRAM

-1	0.14910	0.00900	0.74240	0.74230	0.00210	0.40690...
1	3	0	1	00	01	00 11 11 11...
2	3	0	0	00	11	00 01 01 11...
3	3	0	1	00	00	00 10 00 11...
...						

The first line represents the genetic distance between two adjacent markers. The distance between the last marker of a chromosome and the first marker of a next chromosome is -1. In the second line, the first column is ID, followed by population ID, gender, affected/not affected, two alleles for each marker separated by a space

Table 3

An example of the parameter file for ADMIXPROGRAM

600	African American sample size
2,806	Number of AIMS
2	Which ancestral population allele frequencies will be used. 0: no information, 1: only European, 2: only African, and 3: both European and African
6	Initial value of the number of generations since population admixture occurred
0	Use all, odd, or even markers. 0: all markers, 1: odd markers, and 2: even markers
0	Is there a file including the information of bad markers? 0: No and 1: yes
2	Estimate case-only Z -score and case-control Z -score. 0: No, 1: case-only Z -score, and 2: both case-only and case-control Z -scores

Table 4

Programs for admixture mapping or inferring locus-specific ancestries

Program	References	Method	AIMs	No of ancestral populations	Background LE
STRUCTURE/MALDISOFT	(7, 15, 16)	HMM MCMC	Microsatellite or SNPs	No limit	Yes
ADMIXMAP	(2, 6)	HMM MCMC	Microsatellite or SNPs	No limit	Yes
ANCESTRYMAP	(5)	HMM MCMC	SNPs	2	Yes
ADMIXPROGRAM	(3, 4)	HMM, ML	SNPs	2	Yes
SABER	(12)	MHMM ML	SNP	No limit	No
HAPMIX	(25)	MHMM ML	SNPs	2	No
LAMP LAMP-ANC	(14)	Moving window	SNPs	No limit	No
HAPAA, uSWITCH	(13)	HMM	SNPs	No limit	No