Hum Genet (2005) 118: 424–433
DOI 10.1007/s00439-005-0067-z

ORIGINAL INVESTIGATION

Hui-Ju Tsai · Shweta Choudhry · Mariam Naqvi
William Rodriguez-Cintron
Esteban González Burchard · Elad Ziv

Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations

Received: 1 June 2005 / Accepted: 10 August 2005 / Published online: 6 October 2005

Abstract Population stratification may confound the results of genetic association studies among unrelated individuals from admixed populations. Several methods have been proposed to estimate the ancestral information in admixed populations and used to adjust the population stratification in genetic association tests. We evaluate the performances of three different methods: maximum likelihood estimation, *ADMIXMAP* and *Structure* through various simulated data sets and real data from Latino subjects participating in a genetic study of asthma. All three methods provide similar information on the accuracy of ancestral estimates and control type I error rate at an approximately similar rate. The most important factor in determining accuracy of the ancestry estimate and in minimizing type I error rate is the number of markers used to estimate ancestry. We demonstrate that approximately 100 ancestry informative markers (AIMs) are required to obtain estimates of ancestry that correlate with correlation coefficients more than 0.9 with the true individual ancestral proportions. In addition, after accounting for the ancestry information in association tests, the excess of type I error rate is controlled at the 5% level when 100 markers are used to estimate ancestry. However, since the effect of admixture on the type I error rate worsens with sample size, the accuracy of ancestry estimates also needs to increase to make the appropriate correction. Using data from the Latino subjects, we also apply these methods to an association study between body mass index and 44 AIMs. These simulations are meant to provide some practical guidelines for investigators conducting association studies in admixed populations.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/10.1007/s00439-005-0067-z and is accessible for authorized users.

Esteban González Burchard and Elad Ziv contributed equally to this manuscript.

H.-J. Tsai (✉) · S. Choudhry · M. Naqvi · E. G. Burchard · E. Ziv
Department of Medicine, University of California, Box 0833,
San Francisco, CA, 94143-0833 USA
E-mail: hut1@itsa.ucsf.edu
Tel.: +1-415-2066541
Fax: +1-415-2063463

H.-J. Tsai (✉) · S. Choudhry · M. Naqvi · E. G. Burchard
Lung Biology Center, San Francisco General Hospital,
San Francisco, CA, USA

E. G. Burchard · E. Ziv
Center for Human Genetics, University of California,
San Francisco, CA, USA

W. Rodriguez-Cintron
San Juan VAMC, University of Puerto Rico School of Medicine,
San Juan, PR, USA

Introduction

Genetic association studies are a powerful approach to identify genetic risk factors associated with complex traits (Risch and Merikangas 1996). However, concern has been raised that population stratification may confound genetic association studies (Lander and Schork 1994; Spielman et al. 1993). This may be especially important and cannot be ignored while conducting association studies in admixed populations such as Latinos or African Americans. If disease risk varies with ancestry proportions, then any marker found at a higher frequency in one ancestral group may be associated with the disease even if it is not a causative allele or at a locus near a causative allele (Burchard et al. 2003; Cardon and Bell 2001; Knowler et al. 1988; Ziv and Burchard 2003).

Population stratification can be identified and corrected by various approaches. In general, the methods of correction for population stratification can be categorized into three classes (1) genomic control (2) "structured association" and (3) a semi-parametric method based on principal components analysis (Bacanu et al. 2000; Chakraborty et al. 1986; Devlin and Roeder 1999; Falush et al. 2003; Hanis et al. 1986; Pritchard et al.

2000; Zhang and Zhao 2001; Zhang et al. 2003). The "structured association" approach estimates individual ancestry by using a set of genetic markers and then tests for association while correcting for individual admixture (IA). This approach is particularly favored by the investigators studying admixed populations and is often used with a set of highly informative markers for estimating ancestral proportions. Several methods are commonly used to estimate ancestry including: (1) maximum likelihood estimation (MLE), (2) *Structure*, and (3) *ADMIXMAP*. The latter two programs use a Markov Chain Monte Carlo (MCMC) approach to estimate ancestry. In addition, *ADMIXMAP* also incorporates a test of association that is performed simultaneously with the estimation of ancestry.

In this study, we examine and compare the performance of these three methods for estimating ancestry and for eliminating the excess type I error rates due to population stratification. We investigate the common case of a three-way population admixture, which is relevant to both Latinos and African Americans, the two largest minority groups in the US. We not only use simulations for most of the investigations but also explore the three methods with real data from an asthma genetic study among the Latinos. The goals are to provide practical recommendations to investigators on the choice of methods, number of markers, and number of ancestral individuals included in the study while performing genetic association studies in admixed populations.

Methods

Maximum likelihood estimation

Consider an admixed population, K_4, resulting from the genetic admixture of subjects from three ancestral populations, K_1, K_2, and K_3. Let s_1, s_2, and $(1-s_1-s_2)$ represent the ancestry proportion from population K_1, K_2, and K_3, separately. Let G_i represent the genotype for an admixed individual at the ith locus. For n loci, likelihood can be defined as:

$$L(s_1, s_2) = \prod_{i=1}^{n} \Pr(G_i). \tag{1}$$

Instead of maximizing likelihood, it is computationally simple to maximize its natural logarithm:

$$\log_e[L(s_1, s_2)] = \sum_{i=1}^{n} \log_e[\Pr(G_i)]. \tag{2}$$

The MLE approach has been implemented in the program IAE3CI, which was kindly provided by Dr. Mark D. Shriver. The program requires the information of allele frequencies from each ancestral population and admixed subjects' genotyping data (Bonilla et al. 2004; Chakraborty et al. 1986; Hanis et al. 1986).

Structure

An admixture model implemented in the program *Structure* assumes each individual inheriting some proportion of its ancestry from each population (Falush et al. 2003). Let K denote the number of populations, p_{kij} denote the frequency of allele i at locus j in population k. Let P denote the multidimensional vector of allele frequencies for all k, i, and j. Let $s_k^{(x)}$ refer to the ancestry proportion of individual x's genome that is derived from population k. Let S refer to the multidimensional vector of ancestry proportions for all subjects of the sample. Let Y be the vector of the populations of origin of every allele copy in each individual with $y_j^{(x,a)}$. Under the admixture model,

$$\Pr[y_j^{(x,a)} = k] = s_k^{(x)}. \tag{3}$$

This admixture model also assumes linkage equilibrium and Hardy-Weinberg Equilibrium (HWE) within populations. Based on a Bayesian approach, it requires priors for P and S. Therefore, Pritchard et al. (2000) assumes that the vector of allele frequencies at locus j in population k is sampled from a Dirichlet distribution with a single hyperparameter α. The vector of ancestry proportions for individual x are sampled from a Dirichlet distribution with a hyperparameter β. We may not have the prior information of the allele frequencies P or the populations that came from origin Y. A MCMC approach is applied to estimate P and Y simultaneously. The admixture model in the later version of *Structure* can handle the situation that is not linkage equilibrium (Falush et al. 2003). For obtaining individual ancestry estimates (IAEs), we input the genotyping data from each ancestral population specified as known populations and admixed subjects specified as an unknown population, assumed admixture model and used default values for other parameters by *Structure* with 50,000 burn-ins and 50,000 further iterations, as suggested by the authors. We also checked that the values of key parameters converged before the end of the burn-in stage.

ADMIXMAP

A combination of Bayesian and classical approaches has been implemented in the program *ADMIXMAP*. For k subpopulations, the ancestry proportions are defined by a vector $\boldsymbol{K}$ with k coordinates. The distribution of $\boldsymbol{K}$ in the population is modeled as a Dirichlet distribution. The stochastic variation of k states of ancestry across all chromosomes in each gamete is modeled by k independent Poisson arrival processes, with intensity parameters summing to τ. Priors are assigned to k parameters for the distribution of admixture in the population, τ and the ancestry-specific allele frequencies at each of j loci in k subpopulations (ancestry proportion vectors: s_{11}, ..., s_{kj}). A MCMC simulation is applied to estimate

the posterior distribution of all unobserved variables, conditional on observed marker data, and phenotype values (Hoggart et al. 2003, 2004; McKeigue et al. 2000). For obtaining IAEs, we input allelic counts of ancestry informative markers (AIMs) calculated from each ancestral population, genetic map distance, and genotyping data and phenotype of admixed subjects to *ADMIXMAP* with 1,000 burn-ins and 20,000 further iterations, as recommended by the authors. Additionally, we evaluated the adequacy of burn-ins by the postscript plots provided in *ADMIXMAP*.

Data sources

Simulation 1

This simulation is meant to replicate a realistic scenario with an admixed population and a trait that varies among the ancestral populations. The differences in the ancestral frequencies of markers being tested for association in this simulation are derived from the differences observed between continental groups (e.g. Africans vs. Europeans and Europeans vs. Asians). A subset of the most informative markers in this simulation is used to estimate ancestry among admixed individuals.

To simulate the allele frequencies in ancestral populations, we first selected 2,000 out of 46,000 random SNPs evenly distributed on chromosome 10 in three different populations, Africans, Caucasians, and Chinese retrieved from the International HapMap Project. We calculated allele frequencies distribution from these observed 2,000 SNPs and identified 101 SNPs with $F_{ST} > 0.3$ between Africans and Caucasians. F_{ST} was calculated as $\delta^2/(\bar{p}(1-\bar{p}))$, where δ^2 denoted variance and $\bar{p}$ was the mean of individual allele frequency (Wright 1969). We then simulated two data sets with 500 and 1,000 admixed subjects, separately, and 30 subjects from each ancestral population. We assumed admixed individuals derived from the admixture of these three ancestral populations. We generated 2,000 markers for subjects from each ancestral population based on the observed allele frequencies and assumed all markers under the HWE. For the admixed subjects, we first simulated their true individual ancestral proportion (TIAP) from each ancestral population based on a uniform distribution. To simulate ancestral proportion, we picked two values (s_1 and s_2) drawn from a uniform distribution with an interval between zero and one. If s_1 was larger than s_2, TIAPs of the subject would be s_2, (s_1-s_2) and $(1-s_1)$ corresponding to each ancestral population. If s_1 was smaller than s_2, TIAPs would be s_1, (s_2-s_1) and $(1-s_2)$. We then generated the marker genotyping data for each admixed subject conditional on the TIAPs and marker allele frequencies in each ancestral population. For each marker, we generated allele 1 independently of allele 2. Next, we simulated the phenotype data for different ancestral groups based on the observed distribution in realistic cases, for instance, bone density (Wagner and Heyward 2000). Simulated phenotype data were followed by a normal distribution with means equal to zero for ancestral population 1 and 3, and one for population 2 with the variance equal to one for all ancestral populations. Simulated phenotype data for admixed subjects were conditional on phenotype distribution in ancestral populations and the corresponding TIAPs.

Simulation 2

This simulation is meant to resemble a "worst case" scenario, in which the markers being tested for association with the phenotype have, on an average, greater ancestral allele frequency differences than the average marker. We simulated the data sets with the combination of different numbers of AIMs ($n = 25$, 50, 100 or 1,000), admixed subjects ($n = 200$ or 500) and subjects from each of three ancestral populations ($n = 15$ or 30). In this simulation scenario, instead of selecting AIMs with $F_{ST} > 0.3$, we generated AIMs with the range of F_{ST} among markers between 0.01 and 0.65 (mean informativeness = 0.15). F_{ST} distribution was provided in Supplement Fig. 1. We defined allele frequencies of markers in three ancestral populations for four marker sets ($n = 25$, 50, 100, and 1,000). The simulation procedures of generating data for admixed subjects and ancestral subjects were the same as the ones in Simulation 1.

Admixture populations from an asthma genetic study

One hundred and eighty-one Mexican and 179 Puerto Rican subjects with asthma, who had participated in the Genetics of Asthma in Latino Americans (GALA) Study were included in this analysis (Burchard et al. 2004). The Mexican and Puerto Rican samples were recruited through primary care clinics in the San Francisco (SF) Bay Area, California, and Puerto Rico (PR), respectively. Subjects were enrolled only if they self-reported that both the biological parents and all biological grandparents were of Puerto Rican or Mexican ethnicity.

We selected 44 AIMs from a panel having large allele frequency differences, δ, between Native American, African, and European ancestral populations. Of the 44 AIMs, 23, 31, and 33 markers had $\delta > 0.3$ for European and Native American ancestry, African and European ancestry, and African and Native American ancestry, respectively (Supplement Table 1). Flanking sequence and other relevant information for all the 44 AIMs can be obtained from the dbSNP website. We genotyped these 44 AIMs for all the Mexican and Puerto Rican subjects.

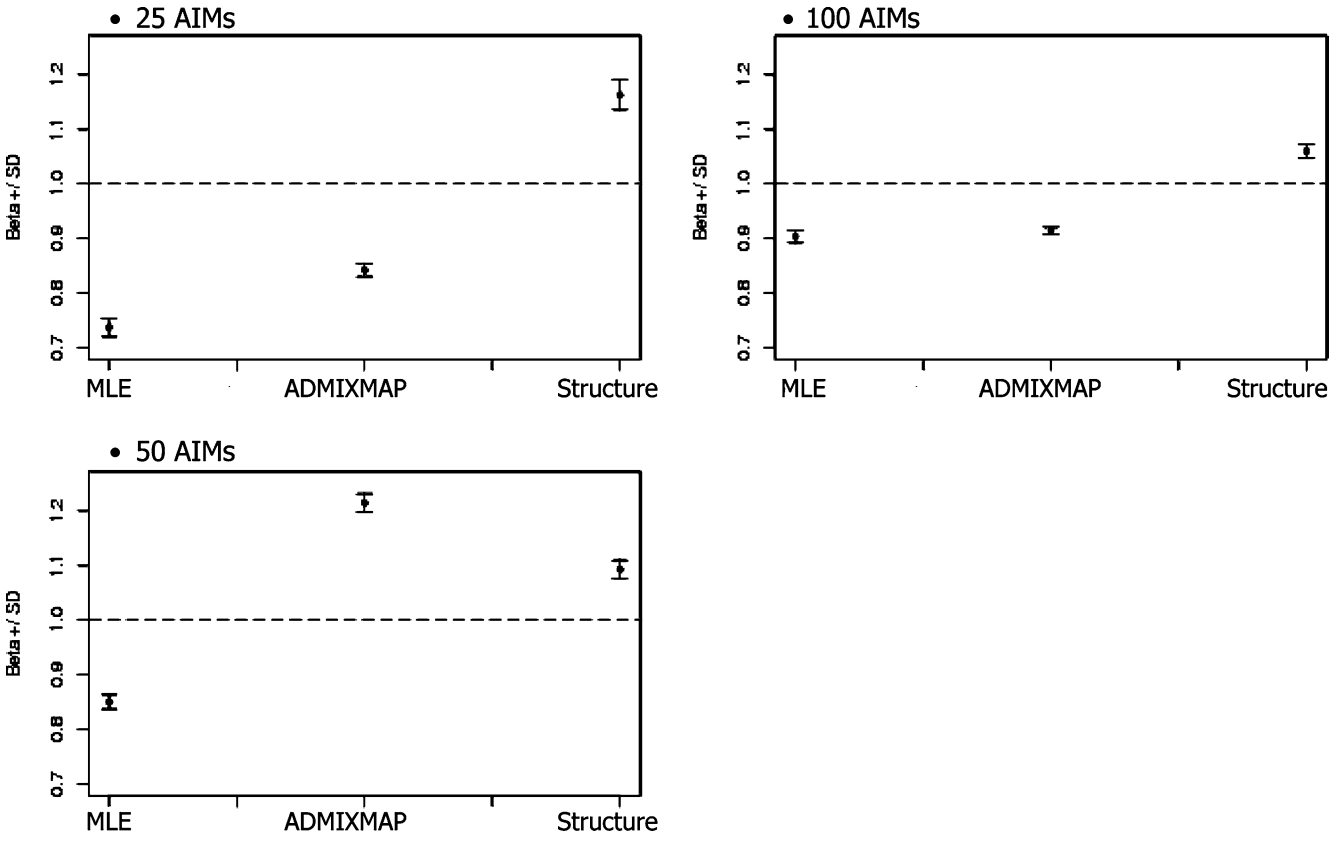


Fig. 1 Precision of individual ancestry estimates (*IAEs*) calculated from each of the three methods. Precision was evaluated by regressing the "TIAPs" on IAEs obtained from each of the three methods in Simulation 1. Note: *Y*-axis represents regression coefficients (β) and corresponding standard deviation (*SD*) obtained from regression models

Data analysis

Evaluating accuracy of IAE

cWe used three programs: IAE3CI, *Structure,* and *ADMIXMAP* to obtain IAEs for simulated data and GALA data. We evaluated the precision and correlation of IAEs obtained from each of the three programs and compared these IAEs to the TIAPs generated by simulation code. We first computed the Pearson's product moment correlation coefficient, *r*, for comparing the "TIAPs" to IAEs generated from each of the three programs. To assess the accuracy of IAEs calculated from the programs, we regressed "TIAPs" on IAEs obtained from the programs, then recorded slopes and the corresponding standard deviations from regression models for each of the simulated sets. We also plotted the errors (TIAP–IAE) for individuals to evaluate the bias in the three methods.

Tests for association and evaluation of the false positive rate

For simulated sets, we performed linear regression models to test for an association between phenotype and markers (all markers in Simulation 1 and AIMs in Simulation 2) under the additive genetic model assumption (genotypes coded as 0, 1, or 2 alleles). For the asthma study, we applied the linear regression models to test for association between body mass index (BMI) and AIMs with covariates: age and gender in the models. For both simulations and the real data, we carried out association tests with and without adjustment for IAEs in the regression models, separately. We only incorporated two out of three IAEs in regression analyses to avoid co-linearity. IAEs included in the models were obtained from the IAE3CI or *Structure*, and the "TIAPs", individually. Score tests for association with traits under the control for population stratification can be obtained in *ADMIXMAP*. We therefore, reported the results of score tests from *ADMIXMAP* for both simulation and real data with respect to evaluating the inflation of type I error rate. We used a *P* value of less than 0.05 as the significance level and recorded positive results from regression analyses according to this threshold.

Data simulation and analyses were carried out using statistical packages R 1.9.0 and STATA 8.0 S/E (College Station, TX, USA). Data simulation R code is available upon request from the author.

Results

Correlation of IAEs among different approaches

We applied three programs by using various simulation sets. We then computed the Pearson's product moment correlation coefficients, *r*, to evaluate the correlation between the "TIAPs" and the IAEs obtained from IAE3CI, *ADMIXMAP,* and *Structure* programs. The results are presented in Table 1 and Supplementary Figs. 2 and 3. All methods increase their correlation with the "TIAPs" at a similar rate with an increasing number of markers (Table 1).

We generated data sets with various numbers of individuals from three ancestral populations ($n = 15$ and 30, separately) in Simulation 2. Increasing the number of ancestral individuals improved IAEs substantially when only using 25 AIMs. However, for the data sets with 50 or more AIMs, adding more ancestral subjects did not significantly increase the accuracy of estimating IA for any of the three methods, even though each method incorporates ancestral allele frequencies of AIMs in a very different manner (data not shown).

Of note, *Structure* provided very poor IAEs when there were 1,000 admixed subjects and 15 ancestral subjects with only 25 AIMs. In this simulation scenario, the correlation coefficients, *r*, were 0.03, 0.04, and 0.05 between *Structure* and the "TIAPs", IAE3CI and *ADMIXMAP*, respectively. However, when we increased the number of ancestral subjects to 30, *r* improved to 0.76 between *Structure* and the "TIAPs", which was comparable to the other methods (data not shown). Thus, *Structure* seems to require a minimal ratio of ancestral to the admixed individuals to appropriately estimate ancestral information in admixed populations.

Accuracy of IAEs from three approaches

To evaluate the accuracy of IAEs, we separately regressed the "TIAPs" on the IAEs obtained from each of the three approaches. When 25 AIMs were used in the regression model, IAEs from *ADMIXMAP* agreed very well with the "TIAPs" (regression coefficient, $\beta = 0.944$, SD = 0.036), compared to the results of IAE3CI or *Structure* ($\beta = 0.642$, SD = 0.025; and $\beta = 1.074$, SD = 0.042, respectively). When 100 or more AIMs were included in the IAE calculation, IAEs from all the three programs were very close to the "TIAPs" and the variance also became smaller (IAE3CI, *ADMIXMAP,* and *Structure*: $\beta = 0.939$, 1.063, 1.097, individually; SD = 0.019, 0.02, 0.021, respectively) (Fig. 1).

In addition, we examined the distribution of differences between "TIAPs" and the estimates from each of the methods. Figure 2 presents the distribution of values for estimated ancestry subtracted from true ancestry ("TIAPs") for each of the three methods. In general, *ADMIXMAP* and *Structure* tended to systematically overestimate ancestry at the low end of the ancestry proportions and to systematically underestimate ancestry at the high end of ancestry proportions. Both the Bayesian methods tended to correlate well in terms of the distribution of deviations. In contrast, maximum likelihood tended to have a greater error throughout the entire distribution of ancestry, but less of a bias at the extremes of ancestry. As expected, when more AIMs were incorporated in admixture estimates, both methods provided less biased estimates.

Control for population stratification

Our simulations produce a very high false positive rate without adjustment for ancestry (Figs. 3a, b). The type I error rate without correction is higher in Simulation 2, since the markers being tested have higher allele frequency differences between the populations. When the sample size of admixed subjects increased the false positive rates also increased (data not shown). In both the simulations, type I error rate was decreased with even as few as 25 AIMs markers and further decreased by 50 AIMs markers. But to achieve a type I error rate as low as the adjustment using the "TIAPs", all methods required 100 AIMs markers for both simulation scenarios (Fig. 3a, b).

Application to the genetics of asthma in Latino Americans (GALA) study

Contemporary Latino populations have been formed by the admixture of three ancestral populations: Africans, Europeans, and Native Americans. We calculated IAEs

Table 1 Correlation of IAEs between two different resources in Simulation 2

	True individual ancestral proportion (m25, m50, m100)[a]	MLE (m25, m50, m100)	*ADMIXMAP* (m25, m50, m100)	*Structure* (m25, m50, m100)
True individual ancestral proportion	–	(0.787,0.870, 0.928)	(0.799, 0.877, 0.931)	(0.787, 0.874, 0.932)
MLE	(0.787, 0.870, 0.928)	–	(0.982, 0.986, 0.991)	(0.991, 0.993, 0.996)
ADMIXMAP	(0.799, 0.877, 0.931)	(0.982, 0.986, 0.991)	–	(0.985, 0.989, 0.994)
Structure	(0.787, 0.874, 0.932)	(0.991, 0.993, 0.996)	(0.985, 0.989, 0.994)	–

[a]IAE estimated by using 25, 50 and 100 AIMs, respectively, in 500 admixed subjects

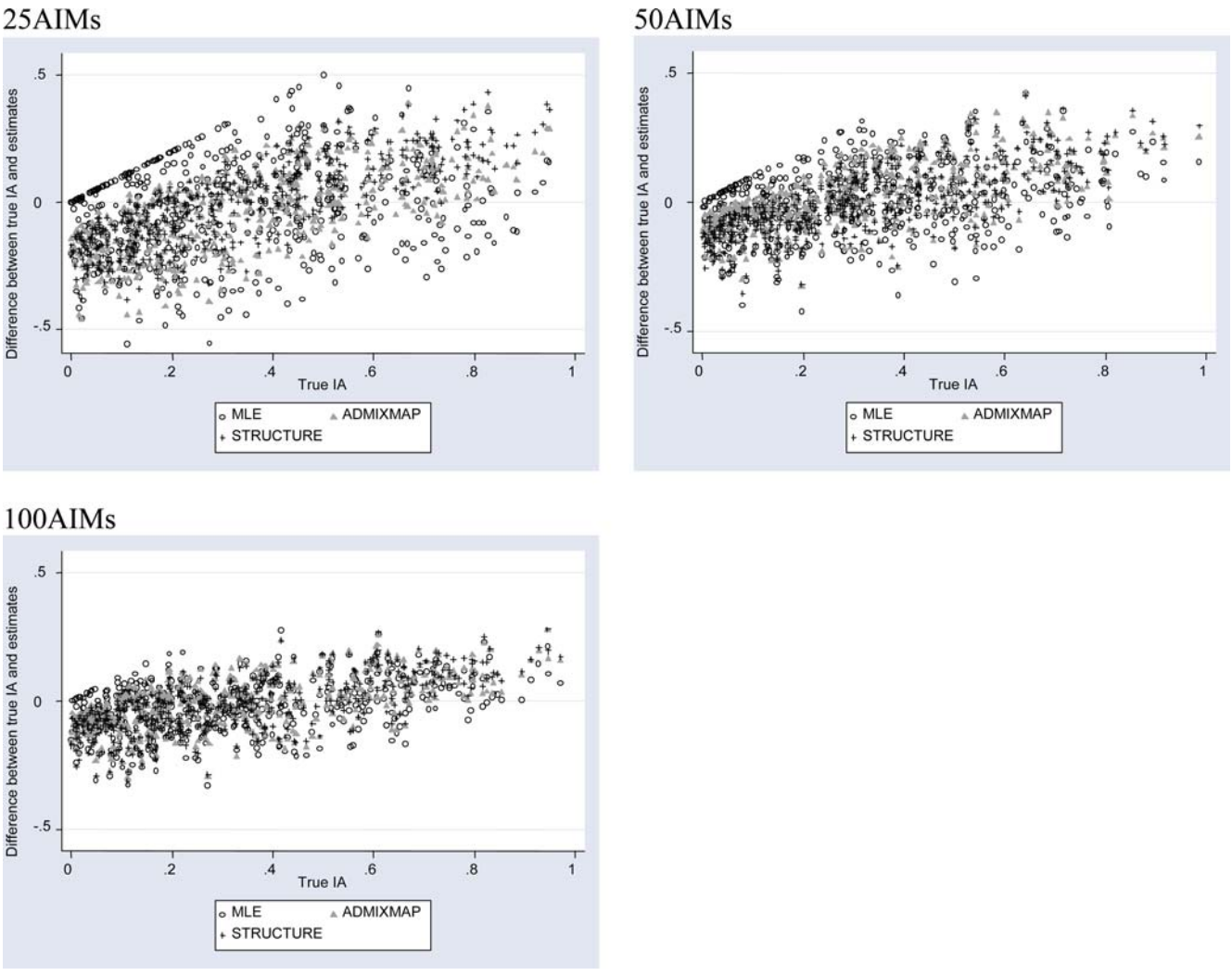


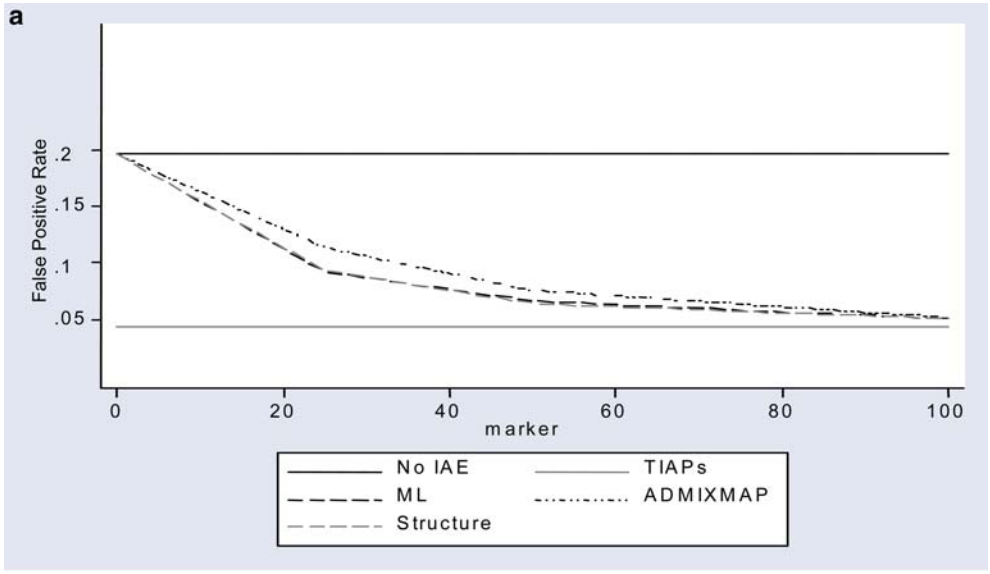
Fig. 2 The distribution of difference between TIAPs and estimates from each of the three methods when 25, 50 and 100 AIMs were used separately to calculate IAEs in Simulation 2. Note: Rate of positive results was calculated from testing association between phenotype and 2,000 markers

of these three ancestral populations in subjects participating in the GALA Study by using IAE3CI, *ADMIXMAP,* and *Structure*. When we compared IAEs obtained from each of the three programs, correlation coefficients between IAE3CI and *Structure* were higher than those observed between *ADMIXMAP* and IAE3CI, and between *ADMIXMAP* and *Structure*. The results were consistent with the results obtained from the simulated data with 50 AIMs (Supplement Fig. 4).

We then applied regression analyses to test the association between BMI and 44 AIMs in the GALA subjects (181 Mexican Americans and 179 Puerto Ricans). We found that 10 out of 44 AIMs were significantly associated with BMI ($P \leq 0.05$) after adjustment for age and gender. After adjusting the IAEs obtained from each of the three programs, only four AIMs remained significantly associated with BMI (Table 2). Furthermore, when we stratified our GALA subjects to two groups: Mexican Americans and Puerto Ricans, there were only four AIMs significantly associated with BMI in the Mexican Americans and one AIM in the Puerto Ricans (Table 3). In addition to stratifying our GALA subjects into two groups based on national origin, we carried out separate models in which we entered both ancestry and nationality by using all the subjects (data not shown). In these models nationality remained a significant predictor, but ancestry did not.

Discussion

In this study, we applied simulations to evaluate the performance of the ''structured association'' approach in admixed populations. We tested three methods: MLE, *ADMIXMAP,* and *Structure* using simulated data with various numbers of AIMs markers, ancestral subjects, and admixed subjects. We also tested these approaches using real data from the GALA Study. These methods



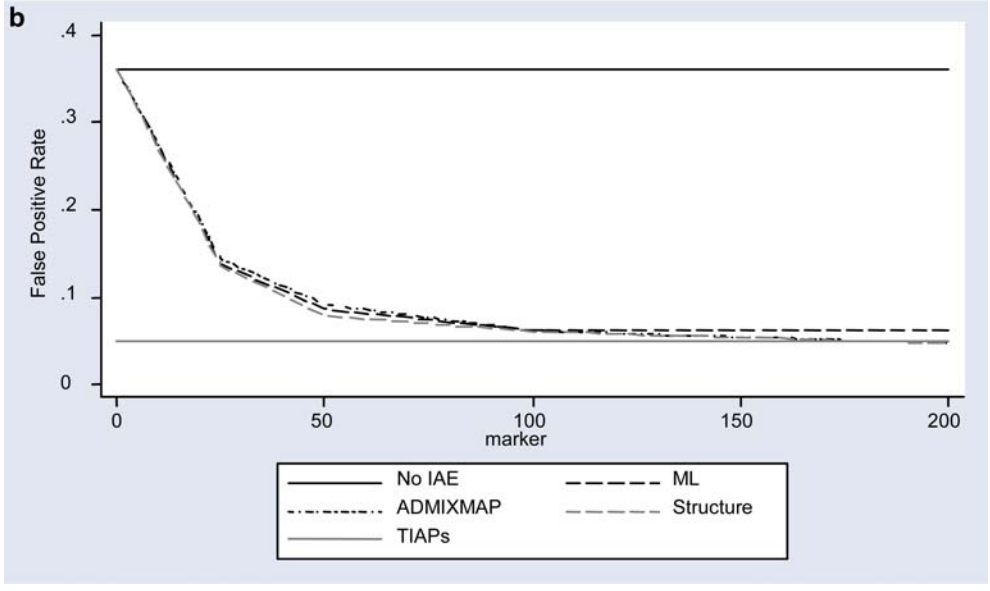


Fig. 3 **a** Rates of positive results in regression models, before and after adjusting IAEs from each of the three methods when 25, 50 and 100 AIMs were used separately to calculate IAEs in Simulation 1 for 500 individuals. **b** Rates of positive results in regression models, before and after adjusting IAEs from each of the three methods when 25, 50 100 and 200 AIMs were used separately to calculate IAEs in Simulation 2 for 500 individuals. Note: Rate of positive results was calculated from testing association between phenotype and 2,000 markers in (**a**) and between phenotype and 1,000 markers in (**b**)

extract the ancestry information from the genotype data of ancestral populations in a different manner. The MLE

Table 2 Associations between AIMs and BMI in Mexicans ($n = 181$) and Puerto Ricans ($n = 179$) combined

Marker	Unadjusted for IAEs		Adjusted for IAEs by MLE		Adjusted for IAEs by *ADMIXMAP*		Adjusted for IAEs by *Structure*	
	t[a]	P[b,c]	t	P	Score	P	t	P
mid93	3.1	0.002	2.27	0.024	252.8	0.028	2.22	0.027
rs223830	−3.07	0.002	−2.54	0.012	−255.7	0.016	−2.54	0.011
wi11153	−2.96	0.003	−2.61	0.009	−289.9	0.014	−2.59	0.01
Ckmm	−2.64	0.009	−1.99	0.048	−213.1	0.051	−2.01	0.045
wi11909	2.61	0.009	1.30	0.196	157	0.16	1.35	0.178
rs6003	2.46	0.014	1.71	0.088	166.6	0.059	1.77	0.078
rs584059	−2.41	0.017	−1.80	0.072	−186.8	0.067	−1.86	0.063
wi9231	2.22	0.027	1.15	0.249	132.2	0.2	1.19	0.234
rs326946	2.15	0.032	1.48	0.139	127.7	0.11	1.51	0.131
Tyr192	2.14	0.033	1.54	0.124	153.4	0.12	1.51	0.131

[a] t = score from Student's t test
[b] P value for each association test, P values less than 0.05 are underlined
[c] Analyses were carried out by regression analyses

Table 3 Associations between AIMs and BMI in Mexicans and Puerto Ricans, separately

Marker	Unadjusted for IAEs		Adjusted for IAEs by MLE		Adjusted for IAEs by *ADMIXMAP*		Adjusted for IAEs by *Structure*	
	t	*P*	*t*	*P*	Score	*P*	*t*	*P*
Mexicans								
mid93	2.27	0.025	2.37	0.019	204.4	0.021	2.36	0.02
Sgc30055	2.72	0.007	2.70	0.008	238.4	0.0084	2.7	0.008
Ckmm	−2.27	0.024	−2.31	0.022	−184.6	0.025	−2.29	0.024
Tyr192	1.98	0.049	2.04	0.043	128.5	0.056	2.04	0.042
Puerto Ricans								
rs223830	−2.31	0.022	−2.16	0.033	−144.9	0.028	−2.16	0.032
Wi11153	−2.10	0.037	−1.12	.265	−141.2	0.069	−1.89	0.06

approach calculates IAEs using allele frequencies of markers from the ancestral populations. *ADMIXMAP* and *Structure* both calculate IAEs through the Bayesian framework by inputting genotyping data from the ancestral populations as a prior information and then estimates a posterior distribution of ancestry information through the MCMC simulation.

The three main factors affecting the accuracy of IAEs are the number of markers, the informativeness of markers, and the number of ancestral subjects. Our results demonstrated that increasing the number of AIMs improved the precision of IAEs in each of these three programs. As the number of markers increased, all three methods improved at a very similar rate. The concordance between all three programs was high regardless of the number of markers. We noted that a high level of agreement between the methods did not imply a high level of accuracy in measuring IAEs. The programs may have a very high level of agreement ($r > 0.95$) despite a low level of agreement with the "TIAPs" ($r < 0.8$) when using very few AIMs markers on IAEs. Since the methods are relatively similar, the similarity in the errors is expected. However, our results demonstrate that the distribution of errors is different among the different methods with a more pronounced bias at the extremes of ancestry with the Bayesian methods and a greater error at the middle ranges of ancestry with the MLE approach. As more markers are added and the error for all methods decreases, these differences become less pronounced.

Our results supported the conclusion from the previous study that these three methods required the prior information from ancestral populations to obtain accurate IAEs (Tang et al. 2005). Moreover, they were generally robust to the number of ancestral subjects, which influenced the accuracy of IAEs only when a limited number of AIMs were used, for instance, 25 AIMs. *Structure* gave poor IAEs when there were more than 1,000 admixed individuals and only 15 ancestral subjects. As the number of ancestral subjects increased to 30, *Structure* performed as well as the other two programs in terms of IAEs. In addition, we observed mild differences of IAEs obtained from *ADMIXMAP* and *Structure*, respectively. Although the same statistical model is implemented in both *ADMIXMAP* and *Structure* programs to model admixed populations, the observed differences are likely due to the prior information from ancestral populations which is treated differently in the programs.

Improving the accuracy of IAEs is important for properly controlling spurious associations in genetic studies. Although IAEs from 50 AIMs did not completely control the inflation of false positive rate, certain level of false positive signals were removed from the association tests. In our simulations, all three programs were able to appropriately correct excess false positive results when using 100 or more AIMs in IAEs. Of interest, with even more informative markers, fewer markers were needed to decrease the type I error rate. For example, when we simulated 25 AIMs with F_{ST} higher than 0.5, the number of AIMs required for controlling the excess of false positives was less than 50 (data not shown). Additionally, previous work has reported when a sample size of admixed subjects increases, more markers are required to adequately control for excess of type I error rate using genomic control (Marchini et al. 2004). Our results demonstrate that the structured association methods require more markers as well.

Our findings demonstrated that the type I error rate was inflated when testing for associations between BMI and AIMs in Latino subjects participating in the GALA Study. False positive associations were corrected after adjusting IAEs. These three approaches provided very similar results for controlling the excess of false positive rates. However, we still observed a slight excess of type I error rate even after correcting for ancestry. Since we only genotyped 44 AIMs, the IAEs for GALA subjects might not be accurate enough for efficiently controlling the excess of false positive rates. It is consistent with our simulation results that 100 AIMs are required for estimating precise individual ancestry and moreover controlling for the type I error rate.

After we stratified the study subjects to two different ethnic groups, the results of association tests did not significantly change with and without IAEs adjustment. It was possible that these four AIMs in the Mexican Americans and one in Puerto Ricans were physically

closed to obesity-related loci (Table 3). However, we examined the location of these five AIMs with the obesity susceptibility loci reported in the human obesity gene map (Snyder et al. 2004). None of them locates closely to the reported obesity candidate genes. Therefore, these positive results were more likely to be an excess of type I error due to chance.

We found that nationality was significantly associated with BMI and seemed to be a stronger predictor of BMI than ancestry, although both eliminated type I error to the same degree. The results of the stratified models suggested that nationality was more proximate to the source of confounding in GALA subjects. In other words, an excess of false positive in our association tests appeared due to nationality rather than ancestry, with the Mexican Americans having a higher BMI compared with the Puerto Ricans. It was possible that environmental factors such as lifestyle and/or diet in the Mexican Americans were different than the lifestyle and/or diet in Puerto Ricans. Although ancestry did not appear to be the most closely correlated factor with BMI, IAEs still effectively controlled the excess false positive rates. By controlling for Native American ancestry, we could eliminate that excess of false positives even if it was not due to any genetic differences between the Native Americans and other populations. This is due to the fact that in order to increase the type I error rate for genetic markers, a particular environmental factor must be associated with ancestry. This example helps to illustrate an important aspect of measuring and adjusting for ancestry: by measuring ancestry, investigators may also eliminate excess false positives whether they are due to genetic or non-genetic confounders.

Previous work has demonstrated the value of using more informative markers (Rosenberg et al. 2003). In the present work, we only simulated single nucleotide polymorphisms (SNPs) for evaluating these three methods. Although the average microsatellite markers are more informative than the average SNPs, the relative abundance of SNPs means that SNPs can be highly selective and we can use the most informative SNPs for ancestry measures (Pritchard and Rosenberg 1999).

We modeled a population with a uniform distribution of ancestry. Although we did this with the explicit purpose of testing each method over the largest possible range of individual ancestry, in real populations the distribution of individual ancestry are more likely to be skewed. In addition, we simulated phenotype data under a normal distribution with means equal to zero or one for different ancestral populations. For the traits with only a subtle difference across different populations, the type I error rate is likely to be lower and thus, the effect of adjustment for stratification less dramatic. Our simulations are based on subjects admixed from distinct ancestral populations for which a large number of markers with large allele frequency differences can be used (i.e., East Asians vs. Caucasians). The number of markers required to distinguish ancestry among more closely related groups (i.e., Japanese vs. Chinese) is likely to be much higher. More work is needed to investigate how these methods perform for populations with subtle substructures.

Our simulations only examined methods that explicitly seek to model individual ancestry in admixed populations. We did not include the approach of adjustment using a latent variable approach (Satten et al. 2001) since this does not allow individual ancestry to vary over a range. Other methods we did not evaluate were the genomic control method (Devlin et al. 2001) and principle component approach (Zhu et al. 2002), which may also be promising approaches for adjusting population stratification but also do not explicitly model individual ancestry.

African American and Latino populations are highly admixed and susceptible to genetic confounding from population stratification. Many clinical and genetic risk factors vary between the racial and ethnic groups (Burchard et al. 2003). African American and Latino populations are underrepresented in biomedical research (King 2002). Association study designs with a proper measurement and adjustment for population stratification are plausible approaches for determining susceptibility genes of complex traits in these two ethnic groups.

Electronic-database information

The URLs for data presented herein are as follows:
dbSNP website, http://www.ncbi.nlm.nih.gov/SNP/
International HapMap Project website, http://www.hapmap.org/
ADMIXMAP program website, http://www.ucd.ie/genepi/software.html
Structure program website, http://pritch.bsd.uchicago.edu/

Acknowledgements Financial support was received from HL07185, GM61390, American Lung Association of California, RWJ Amos Medical Faculty Development Award, NCMHD Health Disparities Scholar, Extramural Clinical Research Loan Repayment Program for Individuals from Disadvantaged Backgrounds, 2001–2003, to Esteban González Burchard, K22CA109351, from the NIH, CRTG 02-0841-CCE from the American Cancer Society, and BCRP030551 from the Department of Defense to Elad Ziv, U19AG23122 from NIH to Steven Cummings, HL51823, HL074204, 3M01RR000083-38S30488, HL56443 and HL51831 to the Asthma Clinical Research Network, U01-CA86117, SFGH General Clinical Research Center M01RR00083-41, U01-HL 65899, UCSF-Children's Hospital of Oakland Pediatric Clinical Research Center (M01 RR01271), Oakland, CA, Sandler Center for Basic Research in Asthma and the Sandler Family Supporting Foundation. The authors would like to acknowledge the families and the patients for their participation. The authors would also like to thank the numerous health care providers and community clinics for their support and participation in the GALA Study. In addition to the primary clinical centers of the investigators, participating community clinics and hospitals include: La Clinica de La Raza, Oakland, CA; UCSF-Children's Hospital of Oakland Pediatric Clinical Research Center, Oakland, CA; General Clinical Research Center, SFGH, San Francisco, CA; Alliance Medical Center, Healdsburg, CA; Santa Clara Valley Medical Center, San José, CA;

Fair Oaks Family Health Center, Redwood City, CA; Clinica de Salud del Valle de Salinas, Salinas, CA; Natividad Medical Center, Salinas, CA; Asthma Education and Management Program, Community Medical Centers, Fresno, CA., Diagnostic Health Centers of: Corozal, Naranjito, Catano, Orocovis, Barranquitas and San Antonio Hospital of Mayaguez. The authors would also like to acknowledge Monica Toscano, MariaElena Alioto, Ivan Gomez, Henry Matallana, Carmen Jimenez, Yannett Marcano, Pedro Yapor, Alma Ortiz, Lisandra Perez and Sheila Gonzalez for their assistance with recruitment and study organization. The authors would like to especially thank Dr. Jeffrey M. Drazen, Dr. Ed Silverman, Dr. Homer A. Boushey, Dr. Dean Shepaprd, Dr. Sylvette Nazario, Dr. Jesus Casal, Dr. Alfonso Torres, Dr. Jose Rodriguez-Santana, Dr. Rocio Chapella, Dr. Scott Weiss, and Dr. Jean G. Ford for all of their effort towards the creation of the GALA Study and to Dr. Mark D. Shriver for assistance in development of the AIMs and for providing ancestral DNA.

References

Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933–1944

Bonilla C, Parra EJ, Pfaff CL, Dios S, Marshall JA, Hamman RF, Ferrell RE, Hoggart CL, McKeigue PM, Shriver MD (2004) Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. Ann Hum Genet 68:139–153

Burchard EG, Avila PC, Nazario S, Casal J, Torres A, Rodriguez-Santana JR, Toscano M, Sylvia JS, Alioto M, Salazar M, Gomez I, Fagan JK, Salas J, Lilly C, Matallana H, Ziv E, Castro R, Selman M, Chapela R, Sheppard D, Weiss ST, Ford JG, Boushey HA, Rodriguez-Cintron W, Drazen JM, Silverman EK (2004) Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. Am J Respir Crit Care Med 169:386–392

Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, Mountain JL, Perez-Stable EJ, Sheppard D, Risch N (2003) The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 348:1170–1175

Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91–99

Chakraborty R, Ferrell RE, Stern MP, Haffner SM, Hazuda HP, Rosenthal M (1986) Relationship of prevalence of non-insulin-dependent diabetes mellitus to Amerindian admixture in the Mexican Americans of San Antonio, Texas. Genet Epidemiol 3:435–454

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155–166

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Hanis CL, Chakraborty R, Ferrell RE, Schull WJ (1986) Individual admixture estimates: disease associations and individual risk of diabetes and gallbladder disease among Mexican–Americans in Starr County, Texas. Am J Phys Anthropol 70:433–441

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492–1504

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. Am J Hum Genet 74:965–978

King TE Jr (2002) Racial disparities in clinical trials. N Engl J Med 346:1400–1402

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) $Gm^{3;5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43:520–526

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36:512–517

McKeigue PM, Carpenter JR, Parra EJ, Shriver MD (2000) Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African–American populations. Ann Hum Genet 64:171–186

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516–1517

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402–1422

Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68:466–477

Snyder EE, Walts B, Perusse L, Chagnon YC, Weisnagel SJ, Rankinen T, Bouchard C (2004) The human obesity gene map: the 2003 update. Obes Res 12:369–439

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 28:289–301

Wagner DR, Heyward VH (2000) Measures of body composition in blacks and whites: a comparative review. Am J Clin Nutr 71:1392–1402

Wright S (1969) Evolution and the genetics of populations, vol 2: the theory of gene frequencies. University of Chicago Press, Chicago

Zhang S, Zhao H (2001) Quantitative similarity-based association tests using population samples. Am J Hum Genet 69:601–614

Zhang S, Zhu X, Zhao H (2003) On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet Epidemiol 24:44–56

Zhu X, Zhang S, Zhao H, Cooper RS (2002) Association mapping, using a mixture model for complex traits. Genet Epidemiol 23:181–196

Ziv E, Burchard EG (2003) Human population structure and genetic association studies. Pharmacogenomics 4:431–441