A Generalized Sequential Bonferroni Procedure for GWAS in Admixed Populations Incorporating Admixture Mapping Information into Association Tests

Wenan Chen^a Chunfeng Ren^a Huaizhen Qin^c Kellie J. Archer^a Weiwei Ouyang^c
Nianjun Liu^d Xiangning Chen^b Xingguang Luo^e Xiaofeng Zhu^f Shumei Sun^a
Guimin Gao^a

^aDepartment of Biostatistics and ^bVirginia Institute for Psychiatric and Behavioral Genetics, School of Medicine, Virginia Commonwealth University, Richmond, Va., ^cDepartment of Biostatistics and Bioinformatics, Tulane University School of Public Health and Tropical Medicine, New Orleans, La., ^dDepartment of Biostatistics, University of Alabama at Birmingham, Birmingham, Ala., USA; ^eBiological Research Center, Beijing Huilongguan Hospital, Beijing, China; ^fDepartment of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA

Key Words

Admixture mapping · GWAS · Sequential Bonferroni procedures · Admixture linkage disequilibrium · Background linkage disequilibrium

Abstract

Objective: To develop effective methods for GWAS in admixed populations such as African Americans. **Methods:** We show that, when testing the null hypothesis that the test SNP is not in background linkage disequilibrium with the causal variants, several existing methods cannot control well the family-wise error rate (FWER) in the strong sense in GWAS. These existing methods include association tests adjusting for global ancestry and joint association tests that combine statistics from admixture mapping tests and association tests that correct for local ancestry. Furthermore, we describe a generalized sequential Bonferroni (smooth-GSB) procedure for GWAS that incorporates smoothed weights calculated from admixture mapping tests into association tests that correct for local ancestry. We have applied the smooth-GSB procedure to analyses of GWAS data on American Africans from the Atherosclerosis Risk in Communities

(ARIC) Study. **Results:** Our simulation studies indicate that the smooth-GSB procedure not only control the FWER, but also improves statistical power compared with association tests correcting for local ancestry. **Conclusion:** The smooth-GSB procedure can result in a better performance than several existing methods for GWAS in admixed populations.

© 2015 S. Karger AG, Basel

Introduction

Admixed populations are populations formed by the recent admixture of two or more ancestral populations. For example, African Americans often have ancestries from West Africans and Europeans. The global ancestry of an admixed individual is defined as the proportion of its genome inherited from a specific ancestral population. The local ancestry of an individual at a specific marker is the proportion of alleles at the marker that are inherited from the ancestral population with a true value of 0, 0.5 or 1. The difference between the local ancestry at a specific marker and the global ancestry of an individual is referred to as the local deviation of ancestry at the marker.

Three Sources of Linkage Disequilibrium in Admixed Populations

For admixed populations, there are three types of linkage disequilibrium (LD) [1]. The first source of LD is variation in the global ancestry among the sampled individuals, which leads to dependence (i.e., LD) among markers across the genome, even though they are from different chromosomes. Individuals with a large global ancestry from a specific ancestral population have an excess of alleles that are common in the ancestral population. This type of LD is called *mixture LD*. In association studies, mixture LD can generate spurious associations (false positive findings) and adjusting for global ancestry is able to control the false positive findings caused by mixture LD.

The second type of LD, *admixture LD*, is formed in local chromosome regions as a result of admixture over the past several hundred years, when large chromosomal segments were inherited from a particular ancestral population, resulting in the temporary generation of long haplotype blocks (usually several megabases, Mb, or longer) [2]. The (proportion of) local ancestry from a specific ancestral population in these blocks may differ from the (proportion of) global ancestry. Admixture LD only exists in local regions and can be used to identify a chromosomal region (usually several Mb) harboring a causal variant.

The third type of LD, *background LD*, is inherited by admixed populations from ancestral populations. Background LD is the traditional LD that exists in much shorter haplotype blocks (usually less than a few hundred kilobases, kbs) in homogeneous ancestral populations, and it is the result of recombination over hundreds to thousands of generations [2]. To illustrate admixture and background LD, we show a special case in figure 1, where a large chromosomal region with admixture LD contains a small region with background LD, and a causal variant is located inside the small background LD region. For association studies, we hope to identify SNPs that are in background LD with causal variants.

Ancestry-Trait Admixture Mapping Tests

Admixture LD has been exploited to locate causal variants that have different allele frequencies among different ancestral populations [3–7]. Mapping by admixture LD is also called admixture mapping [2]. Admixture mapping can only map a causal variant into a wide region of 4–10 cM [2, 8]. Roughly speaking, most admixture mapping tests are based on testing the association between a trait and the local ancestry deviation at a marker. For example, the null hypothesis in admixture mapping tests can be H_{01} : the test SNP is not in admixture LD with the causal vari-

Fig. 1. A large admixture LD (ALD) region (whole white bar) that contains a small background LD (BLD) region (grey bar) with a causal variant inside the small BLD region.

ants. A main advantage of admixture mapping is that only ancestry informative markers (AIMs) are required to be genotyped and tested. An AIM is a marker that has a substantial allele frequency difference between two ancestral populations [9]. If a marker has signal in an admixture mapping test, the marker is called to have an *admixture mapping signal*. A necessary condition for a marker to have an admixture mapping signal is that the marker is in admixture LD with a causal variant that has different allele frequencies between the ancestral populations.

Genotype-Trait Association Tests Correcting for Ancestries

Since admixture mapping tests can only identify wide regions (several Mb) harboring causal variants, (genotype-trait) association tests have been employed to map causal variants into small regions (with a few hundred kbs) by using background LD. For an association test, if no confounding effects exist, the null hypothesis of no association between the test SNP and the trait is approximately equivalent to the null hypothesis H_{02} : the test SNP is not in background LD with causal variants. To control for the confounding effect of global ancestry, some association tests adjusting for global ancestry (or adjusting for principal components of genome-wide genotype scores) have been developed [10–12]. These tests can remove the confounding effect from mixture LD but cannot remove the confounding effect from local admixture LD in a special situation. Here we illustrate the special situation with an example.

Example 1: Suppose that admixture LD extends across a region of 20 Mb. In the region, we consider a causal variant with a strong admixture mapping signal and an AIM SNP that is about 10 Mb away from but not in background LD with the causal variant (fig. 1). By using the association tests adjusting for global ancestry, the AIM may have association signals because the local ancestry at the AIM may be associated with the trait due to admixture LD be-

tween the AIM and the causal variant. This results in an association between the genotype of the AIM and the trait. Therefore, association tests adjusting for global ancestry may identify large regions (with several Mb) harboring causal variants.

To correct for the confounding effect of admixture LD in local regions and therefore map causal variants into small regions with a few hundred kbs, association tests that adjust for local ancestries have been developed [13, 14]. However, these methods can have relatively low power in detecting causal variants with admixture mapping signals.

Joint Association Tests Combining Admixture Mapping Tests and Association Tests with Correction for Local Ancestry

To acquire increased power, Pasaniuc et al. [10] proposed a joint association test, MIXSCORE (or MIX) for binary traits, which combines an admixture mapping test statistic (using admixture LD) with an association test statistic that was conditioned on local ancestry (i.e., correcting for local ancestry). Let $\Omega(R)$ denote an ancestry odds ratio, which is the relative increase in risk per extra allele from an ancestry population such as Europeans. The MIXSCORE is based on an implied assumption that $\Omega(R)$ is a function of the SNP odds ratio R (see the section of Pasaniuc et al. [10]: ‘Mixed SNP and Admixture Association’). Based on this function, if $R = 1$, then $\Omega(R) = 1$. Under this assumption, if a test SNP is not in background LD with a causal variant, then it is neither in admixture LD with the causal variant. However, this assumption may not always be true. Example 1 (see above) shows that we may find an AIM not very far (about 10 Mb) from the causal variant with strong admixture mapping signal, such that the AIM and the causal variant are in admixture LD but not in background LD (fig. 1). Therefore, the AIM has an admixture mapping signal. Using MIXSCORE, the AIM may be called significant (i.e., having an association with the trait). Therefore, the joint association test MIXSCORE may be more suitable for identifying large chromosome regions harboring causal variants (usually several Mb) rather than small chromosome regions (smaller than a few hundred kbs).

Another joint method is the two-stage approach [15, 16], which selects promising regions with admixture mapping signals in the first stage by admixture mapping and which then tests markers in the selected regions in the second stage by association tests that correct for local ancestries. A limitation of this joint method is that it has almost no power to detect causal variants without admixture mapping signals.

It is therefore imperative to develop effective association tests to incorporate information from the admixture mapping test into association tests that correct for local ancestry, so that the new association tests can improve power for mapping causal variants into small regions (smaller than a few hundred kbs in length), while controlling the family-wise error rate (FWER) in GWAS analysis. To achieve this goal, in this study, we propose a novel application of the generalized sequential Bonferroni (GSB) procedure of Holm [17] to GWAS in admixed populations. We propose to calculate smoothed weights by using p values from admixture mapping tests in the GSB procedure; the weights are used to adjust p values of association tests that correct for local ancestry. We have applied the proposed methods to analyze GWAS data on American Africans from the Atherosclerosis Risk in Communities (ARIC) Study.

Methods

First, we describe the concept of a type I error rate and FWER in the strong sense and some existing admixture mapping tests and association tests adjusting for ancestries. Then, we describe a smooth GSB procedure for GWAS that incorporates information from admixture mapping tests into association tests that correct for both local and global ancestries.

Type I Error Rate and FWER in the Strong Sense for Association Tests

Type I Errors under Different Hypotheses

Admixture mapping tests often identify wide chromosomal regions with several Mb harboring causal variants with the corresponding null hypothesis H_{01} : the test SNP is not in admixture LD with the causal variants. If a test SNP is in admixture LD with causal variants, rejecting the null hypothesis (i.e., calling the test SNP significant) is considered as a true positive finding. However, in the context of GWAS analysis, we aim to identify small regions with a few hundred kbs harboring causal variants by testing the null hypothesis H_{02} : the test SNP is not in background LD with the causal variants. Under the null hypothesis H_{02} , if a test SNP is in admixture LD with causal variants but not in background LD with any causal variants, calling the test SNP significant (i.e., rejecting the null hypothesis H_{02} will be regarded as a type I error or a false positive finding. To evaluate this type of error, we describe the concept of the type I error rate and the FWER in the strong sense.

Type I Error Rate and FWER in the Strong Sense

To evaluate the type I error rate of an association test for GWAS, people often assume no causal variants in the genome. In this study, we consider the type I error rate and the FWER *in the strong sense* (see also Westfall and Young [18], p. 10), which are the type I error rate and the FWER for null SNPs, respectively, while some causal variants exist elsewhere in the genome. Under the null hypothesis H_{02} : the test SNP is not in background LD with the causal variants, we define a SNP as a *null SNP* if it is not in back-

ground LD (traditional LD) with any causal variants in both of the ancestral populations, irrespective of whether it is in admixture LD with the causal variants in the admixed population.

Association Tests for Admixed Populations

Below, we describe association tests for admixed populations in generalized linear model (GLM) frameworks. The GLM can be applied to any traits that follow distributions from the exponential family [19], such as binary traits from case-control designs and continuous traits following normal distributions.

Notation

Let Y_i denote the phenotypic value of individual i . For example, in a case-control design, $Y_i = 1$ denotes case status and $Y_i = 0$ denotes control status. For quantitative traits, Y_i has a continuous value. Let G_{ij} denote the coded genotypic score (0, 1, 2) of individual i at the j -th SNP under the assumption of the additive model. Let A_{ij} denote the local ancestry (i.e., proportion of alleles) of individual i at the j -th marker inherited from a given ancestral population. Let Q_i denote the global ancestry proportion of individual i inherited from the given ancestral population. The local deviation of ancestry of individual i at the j -th marker is defined as $D_{ij} = A_{ij} - Q_i$. The ancestries A_{ij} and Q_i can be estimated by existing software such as SABER [20], HAPMIX [21], or LAMP [22, 23].

Ancestry-Trait Admixture Mapping Test T_{admix}

Here, we describe a test T_{admix} for admixture mapping based on a GLM. We assume a link function (see also Zhu et al. [15])

$$h(\mu_i) = \alpha_0 + \alpha_1 Q_i + \alpha_2 D_{ij}, \quad (1)$$

where μ_i is the expected value of Y_i (i.e., $\mu_i = E(Y_i)$), α_0 is the intercept, α_1 is the coefficient for the global ancestry Q_i , and α_2 is the coefficient for local deviation of ancestry D_{ij} at the j -th SNP. For traits following different distributions, we can use different link functions in equation 1. For example, for binary traits from case-control designs, we can use a logit link function, i.e., $h(\mu_i) = \text{logit } \Pr(Y_i = 1) = \log(\mu_i / (1 - \mu_i))$. For quantitative traits following normal distributions, we can use an identity link function, i.e., $h(\mu_i) = \mu_i$.

In the test T_{admix} , a likelihood ratio test statistic (or another related statistic) can be used to test the null hypothesis of no association of the trait with the local deviation of ancestry (i.e. testing the null hypothesis $\alpha_2 = 0$ against the alternative hypothesis: $\alpha_2 \neq 0$). This test T_{admix} can also be used to test the null hypothesis H_{01} : the test SNP is not in admixture LD with the causal variants. However, the T_{admix} test can only map a causal variant with admixture mapping signal into a large admixture LD region with several Mb.

Association Test T_{global} Correcting for Global Ancestry

An association test correcting for global ancestry, T_{global} , can be constructed based on a GLM with a link function (see also McCullagh and Nelder [19])

$$h(\mu_i) = \alpha_0 + \alpha_1 Q_i + \beta G_{ij}. \quad (2)$$

A likelihood ratio test statistic (or another statistic) based on equation 2 can be used to test the null hypothesis of no association between the trait and the genotype (i.e., testing coefficient $\beta = 0$).

The association test T_{global} is approximately equivalent to the following two methods: (1) EIGENSTRAT [11] that adjusts for the first principal component of genome-wide SNP scores and (2) the Armitage trend test with correction for global ancestry (ATT) proposed by Pasaniuc et al. [10]. The ATT method adjusts the phenotypic values and genotypic scores for the effects of global ancestry separately and then constructs a statistic using the adjusted phenotypic values and adjusted genotypic scores. As stated earlier, T_{global} is more suitable for identifying large admixture LD regions with several Mb that harbor causal variants, because it may not be able to control the confounding effect from (local) admixture LD.

Association Test T_{local} Correcting for Both Global and Local Ancestries

We describe a test that corrects for both global and local ancestries (T_{local}), which is based on a GLM with a link function

$$h(\mu_i) = \alpha_0 + \alpha_1 Q_i + \alpha_2 D_{ij} + \beta G_{ij}. \quad (3)$$

The T_{local} test uses a likelihood ratio test statistic (or another related statistic) to test the null hypothesis of no association between the trait and the genotype (i.e., testing coefficient $\beta = 0$). Equation 3 is a variant of the logistic model proposed by Wang et al. [13], who stated that adjusting for local ancestry can control for the confounding effects due to either global or local ancestries. The T_{local} test can also be used to test the null hypothesis H_{02} and to map a causal variant into a small background LD region (with less than a few hundred kbs). However, the T_{local} test may have relatively low power for detecting causal variants with admixture mapping signals. To increase power to identify small background LD regions harboring causal variants, below we propose a novel application of the GSB procedure of Holm [17] to incorporate information from admixture mapping test T_{admix} into the test T_{local} in GWAS for admixed populations.

GSB Procedures for GWAS for Admixed Populations

Suppose that there are m SNPs in GWAS, such that m null hypotheses ($H_1, H_2, \dots, H_m$) are tested, where H_j is the null hypothesis of no association between the j -th SNP and the disease status ($j = 1, 2, \dots, m$). Let α be the nominal level of FWER. The GSB procedure for GWAS can be implemented by the following steps:

- (1) Given a weight w_j for the j -th marker, we adjust the corresponding p value p_j from the association test T_{local} at the j -th marker by weight w_j , i.e., we calculate a B value as $B_j = p_j / w_j$ for the j -th marker.
- (2) Order the B values as $B_{(1)} \leq B_{(2)} \leq \dots \leq B_{(m)}$. Let $w_{(1)}, w_{(2)}, \dots, w_{(m)}$ and $H_{(1)}, H_{(2)}, \dots, H_{(m)}$ denote the corresponding weights and hypotheses of the ordered B values.
- (3) Starting from $j = 1$, given $H_{(1)}, H_{(2)}, \dots, H_{(j-1)}$ have been tested and rejected, if

$$B_{(j)} \leq \alpha / \sum_{k=j}^m w_{(k)},$$

reject $H_{(j)}$ and go to next step if $j < m$; otherwise, accept $H_{(j)}, H_{(j+1)}, \dots, H_{(m)}$, and stop the GSB procedure.

Calculating Weights Using p Values from the Admixture Mapping Test T_{admix}

The original GSB procedure of Holm [17] does not provide a method to calculate weights. Here, we propose to calculate weight

$w_j = 1/q_j$ for the j -th marker, where q_j is the p value from the admixture mapping test T_{admix} . If the weights calculated from the T_{admix} test are independent of the p values calculated from the T_{local} test, then the GSB procedure controls the FWER. For the quantitative traits following normal distributions, we can prove the asymptotic independence between the weights and corresponding p values from the association test that adjusts for local ancestry (see Appendix). However, for other traits, such as binary traits from case-control designs, it is not easy to prove the independence theoretically. From Holm [17], we can see that, to control the FWER of the GSB procedure, the condition ‘weights are independent of the p values of the tests’ is a sufficient but not necessary condition. Our simulation studies [24] showed that in some situations, even when the weights were weakly correlated with the p values, the original GSB procedure still controlled the FWER well. We will show below by simulation studies that the GSB procedure for GWAS can control the FWER well under the null hypothesis H_{02} .

Smooth-GBS Procedure for GWAS Using Smoothed Weights

One concern is that the GSB procedure may give too much weight to SNPs in regions with admixture mapping signals and, therefore, may markedly reduce the power to detect causal variants located outside the regions with admixture mapping signals. To address this concern, we adopt the method of Roeder et al. [25] to smooth weight w_j at the j -th marker. We calculate a smoothed weight at the j -th marker as $(1 - \lambda)w_j + \lambda\bar{w}$ ($0 \leq \lambda \leq 1$), which is a linear combination of the original weight w_j and the average weight of all markers $\bar{w}$. We refer to the GSB procedure using the smoothed weights as smooth-GBS procedure. Parameter λ determines the degree of influence of the original weight w_j on the smoothed weight. If the j -th marker has admixture mapping signal, then w_j provides useful information, and a small λ value will result in increased power of the smooth-GBS procedure; otherwise, if the j -th marker does not have admixture mapping signal, w_j provides only noise, such that a small λ value will result in loss of power. An open question is how to determine the optimal value of λ . We will discuss how to select the value of λ based on our simulation studies.

Simulation Studies

We conducted simulation studies to evaluate the type I error rate and/or FWER in the strong sense and power for several existing methods, the GSB procedure, and the smooth-GBS procedure under the null hypothesis H_{02} : the test SNP was not in background LD with the causal variants. In our simulation studies, we considered binary traits for case-control designs and quantitative traits that follow normal distributions. Below, we focus on describing the simulation studies on binary traits. The simulation results for quantitative traits had similar patterns to those for case-control designs, and they are described in the online supplementary material (see www.karger.com/doi/10.1159/000381474 for all online suppl. material).

Evaluation of Type I Error and FWER in the Strong Sense

It is computationally intensive to evaluate the type I error or the FWER for the whole genome data by simulation studies. Therefore, we estimated the type I error rate and/or FWER for the SNPs in chromosome 1. In addition, we only considered SNPs with no

background LD in the ancestral populations, because it is also computationally intensive to evaluate the type I error rates or FWERs on a large number of simulated dense SNPs with background LD.

Simulating SNP Data Assuming No Background LD for Chromosome 1

We simulated SNP data sets on African Americans assuming that no background LD existed in the ancestral populations. Each data set consisted of 1,000 cases and 1,000 controls. For each individual, we simulated chromosome 1 consisting of 1,030 SNPs selected from the HapMap II data (<http://hapmap.ncbi.nlm.nih.gov/>), which included 100 AIMs. The distance between adjacent SNPs was approximately 200 kb (about 0.2 cM). We assumed that there was no background LD (i.e. no dependence) among the 1,030 SNPs in the two ancestral populations. However, there might be correlations (admixture LD) among the SNPs in the simulated admixed individuals (African Americans) as a result of admixture.

We generated the genotype data in chromosome 1 based on a simulation method of Price et al. [21]. For each admixed individual, we randomly sampled a value from a Beta distribution Beta(3, 12) to represent the global ancestry proportion Q_i that individual i inherited from the given ancestral population (Europeans). The Beta distribution had a mean of 0.2 and a standard deviation of 0.1. To generate each haploid chromosome in an admixed individual, we randomly generated the number of crossover points from a Poisson distribution with mean $\mu = l \times g$, where l is the genetic length (in Morgan) of the chromosome and g is the number of generations since admixture, which we set to 7 in our simulation studies. We then randomly generated an ancestry indicator according to the global ancestry proportion Q_i to represent if the haplotype between two crossover points is of European or West-African ancestry. If the haplotype was from European ancestry, then the haplotype was generated by independently sampling alleles at different SNPs using the allele frequencies in the CEU samples from the HapMap project; otherwise, the YRI samples were used. We used the allele frequencies in the CEU and YRI samples to approximate the frequencies in the two ancestral populations, West Africans and Europeans, respectively.

Simulating Phenotypes Influenced by One Causal Variant with an Admixture Mapping Signal

To evaluate the type I error rate and FWER in the strong sense, when simulating the phenotype for each individual, we chose one AIM (at rs10465723) from the 1,030 simulated SNPs in chromosome 1 as a causal variant. This AIM SNP had a strong admixture mapping signal, with corresponding allele frequencies of 0.858 in the YRI data and 0.283 in the CEU data. The causal AIM SNP was not in background LD with any of the other simulated SNPs in the two ancestral populations, and therefore all simulated SNPs except the causal AIM SNP were null SNPs. Calling any of these null SNPs significant by a test was treated as a type I error.

For case-control designs, the case-control status ($Y_i = 1$ or 0) of individual i was simulated based on a logistic model

$$\text{logit Pr}(Y_i = 1) = \alpha_0 + \beta G_i, \quad (4)$$

where β is the log-odds ratio of the causal allele, and G_i is the genotype score of individual i at the causal SNP. Note that the right side of equation 4 does not include Q_i or A_{ij} because the effects of Q_i and A_{ij} on the phenotypes are included in the effect of genotype

Table 1. Empirical type I error rates of methods for case-control designs at four selected null AIM SNPs based on 10,000 replicated data sets ($\alpha = 0.05$)^a

Distance (Mb) ^b	T_{local}	T_{global}	MIXSCORE	ATT	ES ₁ ^c	ES ₂	ES ₁₀
2.5	0.0508	0.2622	0.4538	0.2613	0.2517	0.2162	0.1378
5	0.0474	0.2221	0.3865	0.2214	0.2102	0.1689	0.0968
9	0.0504	0.1448	0.2375	0.1447	0.1398	0.1171	0.0733
24	0.0483	0.0590	0.0525	0.0509	0.0502	0.0503	0.0491

^a Each data set consisted of 1,000 cases and 1,000 controls. The causal variant was an AIM SNP (with allele frequency $f_E = 0.283$ in the CEU samples and $f_A = 0.858$ in the YRI samples). The odds ratio for the causal allele was 1.5.

^b The distance (Mb) refers to the distance between the null AIM SNPs and the causal AIM SNP.

^c ES₁, ES₂ and ES₁₀ are the EIGENSTRAT tests with correction for the first, the top 2, and the top 10 principal components, respectively. The principal components were computed using 13,056 simulated SNPs across 22 chromosomes (data not shown).

G_i . We set the odds ratio of the risk allele as 1.5, and $\beta = \ln(1.5)$. We set $\Pr(Y_i = 1) = 0.1$, when $G_i = 0$. Equation 4 is different from the models that Wang et al. [13] and Qin et al. [14] used for evaluating the type I error rate in their simulation studies. Their models included a covariate of local ancestry.

Evaluation of Type I Error Rate in the Strong Sense

We simulated 10,000 replicated data sets and estimated empirical type I error rates for the (1,030 – 1) null SNPs in chromosome 1 under the null hypothesis H_{02} for the following five methods: (1) the T_{global} test associated with equation 2; (2) the T_{local} test associated with equation 3; (3) the popular EIGENSTRAT method that corrects for principal components of genome-wide genotypic scores [11]; (4) the MIXSCORE method that combines an admixture mapping test statistic with an association test statistic conditioned on local ancestry [10], and (5) the ATT method with correction for global ancestry [10]. Both the T_{global} and ATT are approximately equivalent to EIGENSTRAT that corrects for the first principal component [11]. For estimating the type I error rate, we set the nominal significance level for single tests at 0.05. To save computing time, we assumed that the global ancestry Q_i and the local ancestry A_{ij} were known. We did not estimate the type I error rates for the GSB procedures because they are multiple testing procedures. Instead, we estimated the FWERs for the GSB procedures (see below).

In table 1, we list the estimated type I error rates in the strong sense at four selected null AIM SNPs with different distances from the causal AIM SNP. From table 1 we can observe that under the null hypothesis H_{02} , T_{local} controlled the type I error rate for all null SNPs. However, T_{global} , MIXSCORE, ATT, and EIGENSTRAT had inflated type I error rates (much larger than the nominal level of 0.05) for null AIM SNPs that were not in background LD but close to the causal AIM SNP. For example, at the null AIM SNP that was 9 Mb away from the causal AIM SNP, the estimated type I error rates for T_{global} , MIXSCORE, and ATT were 0.1448, 0.2357, and 0.1447, respectively. In addition, the estimated type I error rates for ES₁, ES₂, and ES₁₀ were 0.1398, 0.1171, and 0.0733, respectively, where ES _{k} ($k = 1, 2, 10$) denotes the EIGENSTRAT method with correction for the top k principal components of genome-wide ge-

notypic scores. When k increased, the inflated type I error rates of EIGENSTRAT decreased. These methods (T_{global} , MIXSCORE, ATT, and EIGENSTRAT) had inflated type I error rates at the null AIM SNPs, because the null AIM SNPs were not far from the causal AIM SNP and, therefore, were in admixture LD (but not in background LD) with the causal AIM SNP. Since the causal AIM SNP had a strong admixture mapping signal, this could cause the null AIM SNPs to have admixture mapping signals. The results also showed that as the distance from the causal SNP increased, the type I error rates decreased. Therefore, when a causal variant has a strong admixture mapping signal, the four methods (T_{global} , MIXSCORE, ATT, and EIGENSTRAT) are more suitable for identifying a wide chromosome region (with up to about 20 Mb) that harbors the causal variant. On the other hand, our simulation results showed that T_{global} , MIXSCORE, ATT, and EIGENSTRAT all controlled the type I error rate well at all non-AIM null SNPs (data not shown).

Evaluation of FWER in the Strong Sense

We estimated FWERs in the strong sense for the GSB procedures. Because the methods T_{global} , MIXSCORE, ATT, and EIGENSTRAT showed inflated type I error rates, it is straightforward that they will have inflated FWERs. As an example to illustrate the degree of the inflated FWER of these methods, we report the FWER of T_{global} . In addition, we estimated the FWER of T_{local} . We set the nominal FWER equal to 0.05. This is a criterion much stricter than setting the nominal type I error rate at 0.05 for single tests as described above. We simulated 10^5 replicated data sets. The estimated FWERs are listed in table 2. We can see that T_{local} (with the Bonferroni correction), the GSB procedure, and the smooth-GSB procedure controlled the FWERs well. As expected, T_{global} failed to control the FWER, because it cannot control the type I error rate at some null AIM SNPs as shown in table 1.

Evaluation of Power of the GSB Procedures

We evaluated the power of the T_{local} test, the GSB procedure, and the smooth-GSB procedure in the analysis of simulated dense SNP data that have LD patterns similar to the real GWAS data. We did not estimate power for T_{global} , ATT, and the EIGENSTRAT

method ES_k because they cannot control the FWERs well. We simulated replicated GWAS data sets on African Americans with background LD among SNPs in the ancestral populations only for chromosome 1.

Simulating SNP Data with Background LD for Chromosome 1

We simulated SNP data in chromosome 1 with background LD for African Americans in two steps:

Step 1: We used the HapGen software [26, 27] to generate a large pool of 20,000 unrelated haploid chromosomes for each of the two ancestral populations (West Africans and Europeans). We used the haplotype information on the CEU and YRI populations from the phased HapMap2 data, which can be downloaded from the IMPUTE website (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference), for simulating the haploid chromosomes for West Africans and Europeans, respectively. Therefore, these simulated haploid chromosome pools for the two ancestral populations had similar background LD patterns to those in the data sets of CEU and YRI haplotypes. We then thinned the simulated haploid chromosomes and kept only those SNPs that appeared in Affymetrix 6.0. Rare variants (with a minor allele frequency <5%) in each ancestry population were excluded. This resulted in 58,441 SNPs left on chromosome 1. The genetic distance among SNPs was calculated based on the combined genetic map downloaded from the IMPUTE website.

Step 2: We adapted the method of Price et al. [21] for simulating each African American (admixed individual). Each haploid chromosome was generated as a mixture of two ancestral haploid chromosomes, sampled separately from the two pools of ancestral haploid chromosomes (without repeating).

Evaluation of Power when Phenotypes Were Influenced by One Causal Variant with an Admixture Mapping Signal

We estimated power for the T_{local} test, the GSB procedure, and the smooth-GSB procedure based on analyses of the simulated data sets under the assumption that the phenotypes were influenced by one causal variant with an admixture mapping signal.

We simulated 1,000 replicated data sets; each data set was composed of 1,000 cases and 1,000 controls. To simulate the phenotypes, we chose the AIM SNP rs2806404 in chromosome 1 from the Affymetrix 6.0 SNP set as the causal SNP (we did not use the previously selected causal SNP rs10465723, because it was not in the Affymetrix 6.0 SNP set). The risk allele frequencies were 0.292 in the CEU and 0.825 in the YRI samples.

We simulated phenotypes by equation 4 with a set of β values corresponding to different odds ratios at the causal SNP ranging from 1.3 to 1.6 (table 3). We considered two scenarios. In scenario 1, we assumed that the risk allele at the causal SNP had the same odds ratios in the two ancestral populations. In scenario 2, we assumed that the causal SNP was an African-specific causal variant, and we set the odds ratio of the risk allele inherited from Europeans $OR_E = 1$, and the odds ratio of the risk allele inherited from West Africans $OR_A = 1.5$ (see the last row in table 3). We set the genotypic score G_i equal to the number of risk alleles inherited from West Africans, and the coefficient β of G_i equal to $\ln(OR_A)$. Torgerson et al. [28] and Lettre et al. [29] reported African-specific causal variants that are associated with disease in African Americans but not in European Americans.

We estimated power for the T_{local} test, the GSB procedure, and the smooth-GSB procedure based on analyses of the simulated

Table 2. Empirical FWER of methods for case-control designs for (1,030 – 1) null SNPs in chromosome 1 based on 10^5 replicated data sets (nominal FWER = 0.05)

T_{local}	T_{global}	GSB	smooth-GSB ^a		
			$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$
0.0497	0.0819	0.0489	0.0497	0.0498	0.0504

^aFor the smooth-GSB procedure, we calculated power for different λ values.

data sets for chromosome 1. To mimic real GWAS data analysis, we set the significant threshold for the T_{local} test at individual SNPs to be 5×10^{-8} , which is a very strict threshold. Correspondingly, we set the nominal FWER as $5 \times 10^{-8} \times m$ for the GSB procedure and the smooth-GSB procedure, where $m = 58,441$ is the total number of SNPs in chromosome 1.

To illustrate that there was an admixture mapping signal at the causal SNP, we also estimated the power of the admixture mapping test T_{admix} . Following Pasaniuc et al. [10], we used the significance threshold of 10^{-5} for T_{admix} , which was different from the threshold 5×10^{-8} for T_{local} . The less strict threshold for T_{admix} was motivated by the smaller number of independent admixture mapping tests T_{admix} across the genome due to admixture LD.

It usually requires a very long time to estimate an individual's ancestries in the simulated data sets with background LD among SNPs using existing software such as Lamp [23], HAPMIX [21], or SABER [20]. To save computing time, we also assumed that the true global and local ancestries were known.

The estimated power is showed in table 3. Both the GSB procedure and the smooth-GSB procedure had higher power than the T_{local} test. For example, when the odds ratio was 1.6 in both ancestry populations, the power of T_{local} , the GSB procedure, and the smooth-GSB procedure (with $\lambda = 0.7$) were 0.588, 0.732, and 0.687, respectively. The smooth-GSB procedure was robust to different λ values when $0.6 \leq \lambda \leq 0.8$. When λ changed from 0.6 to 0.8, the change of power of the smooth-GSB procedure was always less than 3% in absolute terms (see also 'Influence of Parameter λ on the smooth-GSB Procedure'). We evaluated the power of the admixture mapping test T_{admix} using the threshold 10^{-5} . We can also see that the power was always much greater than the threshold 10^{-5} , ranging from 0.005 to 0.312, which indicated that there was an admixture mapping signal at the causal SNP.

From the last row in table 3, we can observe that at the African-specific causal variant, where the risk allele of European ancestry had an $OR_E = 1$, and the risk allele of West-African ancestry had an $OR_A = 1.5$, the power of the smooth-GSB procedure (with $\lambda = 0.7$) was 0.112, which is much higher than the power of 0.051 of T_{local} .

Evaluation of Power when Phenotypes Were Influenced by Two Causal Variants (One with and the Other without an Admixture Mapping Signal)

As mentioned earlier, one concern is that the GSB procedures may give too much weight to SNPs with admixture mapping signals and, therefore, markedly reduce the power to detect causal

Table 3. Estimated power of methods for case-control designs based on 1,000 replicated data sets for chromosome 1 with one causal SNP that had an admixture mapping signal^a

Odds ratio ^b		Power					
OR _E	OR _A	T_{admix}^c	T_{local}	GSB	smooth-GSB ^d		
					$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$
1.3	1.3	0.005	0.013	0.017	0.012	0.013	0.015
1.4	1.4	0.015	0.102	0.149	0.134	0.127	0.122
1.5	1.5	0.034	0.314	0.429	0.391	0.379	0.376
1.6	1.6	0.059	0.588	0.732	0.696	0.687	0.676
1	1.5	0.312	0.051	0.138	0.116	0.112	0.105

^a Each data set consisted of 1,000 cases and 1,000 controls. The significance threshold for T_{local} at individual SNPs was 5×10^{-8} and the nominal FWER for GSB and smooth-GSB was $5 \times 10^{-8} \times m$, where $m = 58,441$ is the total number of SNPs.

^b OR_E and OR_A denote the odds ratio in Europeans and West Africans at the causal SNP, respectively. The causal SNP rs2806404 had an admixture mapping signal. The risk allele frequency $f_E = 0.292$ in the CEU samples and $f_A = 0.825$ in the YRI samples.

^c The power of T_{admix} was evaluated at a significance level of 10^{-5} , indicating the strength of the admixture mapping signal.

^d For the smooth-GSB procedure, we calculated power for different λ values.

variants without admixture mapping signals. To address this concern, we evaluated the power of T_{local} , T_{admix} , the GSB procedure, and the smooth-GSB procedure based on analyzing simulated case-control data sets under the assumption that the phenotypes were influenced by two causal variants (one with and one without an admixture mapping signal).

Simulating Phenotypes Influenced by Two Causal Variants. To simulate the phenotypes, we chose two specific SNPs (about 150 Mb apart) from chromosome 1 as causal variants. The first causal SNP was SNP rs2806404 with an allele frequency $f_E = 0.292$ in the CEU samples and $f_A = 0.825$ in the YRI samples. The second causal SNP was SNP rs12748791 with an allele frequency $f_E = f_A = 0.133$. There was an admixture mapping signal at the first causal SNP but no admixture mapping signal at the second causal SNP. We simulated 1,000 replicated sets, with each data set composed of 1,000 cases and 1,000 controls. The phenotypic value Y_i of individual i was simulated based on a logistic model

$$\text{logit Pr}(Y_i = 1) = \alpha_0 + \beta_1 G_{i1} + \beta_2 G_{i2}, \quad (5)$$

where G_{i1} and G_{i2} are the genotype scores at the first and second causal SNP, respectively.

Estimated Power when Phenotypes Were Influenced by Two Causal Variants. We estimated the power of T_{local} , T_{admix} , the GSB procedure, and the smooth-GSB procedure based on analyzing the 1,000 simulated case-control data sets described in the previous section. We used the significance threshold of 5×10^{-8} for T_{local} at individual SNPs and the nominal FWER $5 \times 10^{-8} \times m$ for the GSB procedure and the smooth-GSB procedure, where $m = 58,441$ is the total number of SNPs in chromosome 1. We used the significance threshold of 10^{-5} for T_{admix} . The empirical power at the two causal SNPs is shown in table 4. We can see that T_{admix} always had power much higher than the threshold 10^{-5} at the first causal SNP, indicating there was an admixture mapping signal at this causal

variant, and that T_{admix} had power very close to zero at the second causal SNP, indicating no admixture mapping signal at the second causal SNP. Compared with the association test T_{local} , although the GSB procedure had much higher power at the first causal variant with an admixture mapping signal, it had much lower power at the second causal variant without an admixture mapping signal. For example, when the odds ratio was 1.6, the GSB procedure had power of 0.272 at the second causal variant, much lower than the power of 0.414 of the T_{local} test. Therefore, the GSB procedure is not appropriate for GWAS analysis in admixed populations. On the other hand, compared with the association test T_{local} , our smooth-GSB procedure not only had much higher power at the first causal SNP with an admixture mapping signal, but also had comparable (slightly lower) power at the second causal SNP without an admixture mapping signal. For example, when the odds ratio was 1.5, the T_{local} test and our smooth-GSB procedure ($\lambda = 0.7$) had power of 0.264 and 0.351 at the first causal SNP and power of 0.172 and 0.167 at the second causal SNP, respectively.

Influence of Parameter λ on the smooth-GSB Procedure. We also evaluated the influence of different values of parameter λ on the power of the smooth-GSB procedure. We considered nine values of λ : 0.1, 0.2, ..., 0.9, and calculated the corresponding power of the smooth-GSB procedure. We found that when $0.6 \leq \lambda \leq 0.8$, the smooth-GSB procedure had power comparable with that of T_{local} in detecting the causal variant without admixture mapping signals (the second causal SNP), and had significantly higher power than T_{local} at the causal variant with an admixture signal (table 4). In addition, compared with the power of the smooth-GSB procedure with $\lambda = 0.7$, when λ decreased to 0.6 or increased to 0.8, the absolute value of the difference in power at the first causal variant with an admixture mapping signal was always < 0.03 , the absolute value of the difference in power at the second variant without an admixture mapping signal was always less than or around 0.02. In other

Table 4. Estimated power of methods for case-control designs based on 1,000 replicated GWAS data sets for chromosome 1 with two causal SNPs^a

Odds ratio ^b			Power at SNP ₁ ^c						Power at SNP ₂					
OR _{1,E}	OR _{1,A}	OR ₂	T_{admix} ^d	T_{local}	GSB	smooth-GSB ^e			T_{admix}	T_{local}	GSB	smooth-GSB		
						$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$				$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$
1.3	1.3	1.3	0.001	0.016	0.020	0.018	0.018	0.018	0	0.005	0.003	0.004	0.005	0.005
1.4	1.4	1.4	0.012	0.092	0.150	0.135	0.129	0.120	0	0.043	0.029	0.037	0.039	0.040
1.5	1.5	1.5	0.028	0.264	0.392	0.363	0.351	0.332	0	0.172	0.118	0.162	0.167	0.170
1.6	1.6	1.6	0.071	0.524	0.687	0.646	0.634	0.617	0	0.414	0.272	0.389	0.397	0.404
1	1.5	1.5	0.309	0.053	0.141	0.113	0.102	0.096	0	0.167	0.052	0.146	0.148	0.155

^a Each data set consisted of 1,000 cases and 1,000 controls. The significance threshold for T_{local} at individual SNPs was 5×10^{-8} and the nominal FWER for GSB and smooth-GSB was $5 \times 10^{-8} \times m$, where $m = 58,441$ is the total number of SNPs in chromosome 1.

^b OR_{1,E} and OR_{1,A} denote the odds ratio in Europeans and West Africans at the first causal SNP, respectively; OR₂ denotes the odds ratio at the second causal SNP (assuming that the risk allele had the same odds ratio in the two ancestral populations).

^c The first causal SNP had an admixture mapping signal, with an allele frequency $f_E = 0.292$ in the CEU samples and $f_A = 0.825$ in the YRI samples. The second causal SNP had no admixture mapping signal, with an allele frequency $f_E = f_A = 0.133$ in the two ancestral populations.

^d The power of T_{admix} was evaluated at significance level 10^{-5} , indicating the strength of an admixture mapping signal.

^e For the smooth-GSB procedure, we calculated power for different λ values.

words, when the value of λ changed around 0.7, the power of the smooth-GSB procedure only changed slightly. Therefore, we recommend using a value of λ around 0.7 in the smooth-GSB procedure.

Real Data Analysis

We applied the smooth-GSB procedure and T_{local} test to GWAS of data sets on American Africans drawn from the ARIC Study [30]. We did not apply the other existing methods described above due to the possibility of not controlling the FWER. We did a quality control filtering and excluded A/T and G/C SNPs to avoid complementarity issues as described in Pasaniuc et al. [10]. SNPs with a minor allele frequency <0.01 were also removed. After the quality control filtering, there were 3,075 individuals with 584,535 SNPs remaining. We analyzed three phenotypes: type 2 diabetes (T2D), low-density lipoprotein-cholesterol (LDL), and high-density lipoprotein-cholesterol (HDL). The data set for T2D consisted of 531 cases and 1,887 controls. The data sets for quantitative phenotypes LDL and HDL consisted of 2,897 and 2,924 individuals, respectively.

We noted that when the original GSB procedure of Holm [17] is applied to testing association at genome-wide markers, it does not provide a new p value for each mark-

er. In our previous contribution [24], we proposed an adjusted p value, P_{genome} , for each marker for the GSB procedure, which is used to compare sequentially with the nominal FWER level α (see also Westfall and Yong [18], pp. 64, 65). For comparison with the T_{local} test, we calculated a new p value at each SNP for the smooth-GSB procedure that was equal to the adjusted p value P_{genome} divided by the number of test SNPs (m). This new p value at each SNP was used to compare sequentially with the Bonferroni-corrected threshold α/m for single SNPs. In table 5, we listed those SNPs that had p values $<10^{-6}$ either in the GWAS using the T_{local} test or the smooth-GSB procedure. We also listed the p values from the T_{admix} test corresponding to these SNPs to show if there were admixture signals. Because of the small sample size for the T2D phenotype, no SNPs had p values $<10^{-6}$. When we set the genome-wide FWER = 0.05, the corresponding Bonferroni-corrected threshold was 8.55×10^{-8} . The T_{local} test only identified one significant SNP (rs247617) with a p value of 6.02×10^{-22} for the phenotype HDL, and this SNP was also identified by our smooth-GSB procedure with a p value of 8.52×10^{-22} (when $\lambda = 0.7$). SNP rs247617 was also reported to be significantly associated with HDL by Lettre et al. [29] in their meta-analysis on African Americans (see their table 3). In table 5, we can also see that when the value of λ in our smooth-GSB procedure changed from 0.6 to 0.8, the corresponding p values were very close to each other.

Table 5. SNPs with p values $<10^{-6}$ in GWAS for the phenotypes LDL or HDL

Chromosome number	SNP	Allele frequency in HapMap		p value		p value of smooth-GSB ^a		
		CEU samples	YRI samples	T_{local}	T_{admix}	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$
<i>LDL</i>								
2	rs1485059	0.088	0.167	6.15×10^{-7}	0.97	9.75×10^{-7}	8.51×10^{-7}	7.54×10^{-7}
2	rs1485055	0.093	0.197	5.93×10^{-7}	0.97	9.40×10^{-7}	8.20×10^{-7}	7.27×10^{-7}
19	rs1160985	0.566	0.33	2.63×10^{-7}	0.028	1.58×10^{-7}	1.76×10^{-7}	1.97×10^{-7}
<i>HDL</i>								
3	rs4462984	0	0.33	6.20×10^{-7}	0.169	9.66×10^{-7}	8.48×10^{-7}	7.55×10^{-7}
7	rs6950206	0.04	0.686	3.85×10^{-7}	0.073	5.53×10^{-7}	4.98×10^{-7}	4.54×10^{-7}
8	rs2916700	0.531	0.187	4.38×10^{-7}	0.439	7.11×10^{-7}	6.15×10^{-7}	5.42×10^{-7}
16	rs247617	0.341	0.241	6.02×10^{-22}	0.809	9.88×10^{-22}	8.52×10^{-22}	7.48×10^{-22}
16	rs6499863	0.173	0.299	1.57×10^{-7}	0.809	2.58×10^{-7}	2.22×10^{-7}	1.95×10^{-7}

^a p value calculated as the adjusted p value divided by the number of tests (see Chen et al. [24]).

In addition, table 5 shows that at a SNP with an admixture mapping signal (p values from $T_{admix} < 0.05$), the smooth-GSB procedure had a smaller p value than T_{local} . For example, at SNP rs1160985, the p value from T_{local} for LDL was 2.63×10^{-7} . Since this SNP showed a weak admixture signal with a p value of 0.028 (< 0.05) from the T_{admix} test, the p value from our smooth-GSB procedure ($\lambda = 0.7$) was 1.76×10^{-7} , which was closer to the strict Bonferroni-corrected threshold (8.55×10^{-8}) than the p value from the T_{local} test. SNP rs1160985 was also reported to be significantly associated with HDL by Lettre et al. [29] (see their table 3). On the other hand, at a SNP with no admixture signal, the p value from the smooth-GSB procedure was slightly larger than that from the T_{local} test. For example, for the phenotype HDL, at SNP rs2916700, which did not show an admixture signal (p value from T_{admix} was 0.439), the p value from the smooth-GSB procedure was 6.15×10^{-7} , which was slightly larger than the p value from the T_{local} test (4.38×10^{-7}).

Discussion

In this paper, we described a smooth-GSB procedure for GWAS in admixed populations that incorporates information from admixture mapping tests into the association tests that adjust for local ancestry. The smooth-GSB procedure for GWAS can be applied to identify small regions with a few hundred kbs harboring causal variants for binary traits, quantitative traits and any other trait that follow distributions from the exponential family. Our simulation studies indicate that the smooth-GSB

procedure can control the FWER well. Compared with association tests that adjust for local ancestry, the smooth-GSB procedure can attain substantially improved power at the causal variants with admixture mapping signals and has comparable power at the causal variants with no admixture mapping signals.

In the smooth-GSB procedure, we calculated the smoothed weight as $(1 - \lambda)w + \lambda\bar{w}$ ($0 \leq \lambda \leq 1$), which is a linear combination of the original weight w and the average weight of all markers $\bar{w}$. In our future study, we will also investigate whether the smooth-GSB procedure has better performance when calculating the smoothed weight by $(1 - \lambda)w + \lambda E(w)$, where $E(w)$ is the expected value of w under the null hypothesis of no association of the trait with the local deviation of ancestry.

We have evaluated the impact of the values of parameter λ on the power of the smooth-GSB procedure by simulation studies. When $0.6 \leq \lambda \leq 0.8$, the power of the smooth-GSB procedure only changed slightly and had good performance in detecting both causal variants with and without admixture mapping signals. Based on our simulated studies, we recommend using a value of λ around 0.7 in the smooth-GSB procedure. To evaluate the sensitivity of parameter λ in a specific real data analysis, investigators can use different values of λ , such as 0.6, 0.7 and 0.8, and compare the corresponding adjusted p values for each of the most significant SNPs (see also ‘Real Data Analysis’). It is still an open question how to determine the optimal value of parameter λ . We will explore using data-driven methods to determine the optimal value in our future research.

As we mentioned earlier, there are three types of LD in admixed populations: mixture LD, admixture LD and background LD. Mixture LD is caused by variation of global ancestry during the admixture; admixture LD is generated in local regions during the admixture, and background LD is the traditional LD inherited from ancestral populations. For GWAS in admixed populations, we aimed to identify SNPs in background LD with the causal variants and locate each causal variant in a small chromosomal region (often with less than a few hundred kbs). It is well known that the variation of global ancestry or mixture LD can generate confounding effects. In this study, we showed that admixture LD existing in a local region (with several Mb or longer) may also cause spurious association (false positive) findings at null AIMs that are located in the local region, if the local region includes a causal variant with a strong admixture mapping signal. Adjusting for global ancestry can control spurious association findings caused by mixture LD, but may not be able to control the false positive findings caused by admixture LD. Therefore, it is necessary to adjust for local ancestry in the association tests to strictly control the false positive findings caused by admixture LD.

In GWAS analysis, to test for each SNP, the null hypothesis of no association is approximately equivalent to hypothesis H_{02} : the SNP is not in background LD with the causal variants. The association tests that adjust for global ancestry, such as T_{global} , ATT, and the EIGENSTRAT method, are often directly or indirectly used to test H_{02} . However, our simulation studies indicate that when a causal variant is an AIM and has a strong admixture mapping signal, these tests can have inflated type I error rates under the null hypothesis H_{02} at the null AIM SNPs that are in admixture LD but not in background LD with the causal variant. We noticed that similar results have also been reported in Wang et al. [13] and Qin et al. [14]. When evaluating the type I error rate by simulation studies, both groups [13, 14] used models including a covariate of local ancestry to simulate the case-control status. In contrast, in our simulation studies, we used a model including a covariate of genotypic score but not a covariate of local ancestry (see equation 4). In the association tests adjusting for global ancestry, calling a SNP significant cannot guarantee the SNP is in background LD with the causal variants, it can be only in admixture LD with the causal variants. Therefore, it may be more appropriate to use the association tests adjusting for global ancestry to locate a causal variant in a large chromosomal region (several Mb).

As previously described, the joint test MIXSCORE by Pasaniuc et al. [10] is based on an implied assumption that ancestry adds ratio $\Omega(R) = 1$, if the SNP odds ratio is $R = 1$. In other words, the joint association test is approximately based on testing an implied null hypothesis H_{Joint} : H_{01} and H_{02} , i.e., the test SNP is not in admixture LD with the causal variants, and the test SNP is not in background LD with the causal variants. Rejecting the null hypothesis H_{Joint} (i.e., calling the test SNP significant) is equivalent to stating that the test SNP is in admixture LD or in background LD with the causal variants, but one cannot guarantee that the test SNP is in background LD with the causal variants. Therefore, this joint association test is more appropriate for use in identifying large chromosome regions (usually several Mb) harboring the causal variants.

Our smooth-GSB procedures for GWAS are based on using the p values from single SNP-based association tests for genome-wide SNPs and ignoring the correlations among SNPs. However, ignoring the correlations can result in reduced power. In our future research, we will explore using p values from SNP set- (or gene)-based association tests in the smooth-GSB procedures. The SNP set- (or gene)-based association tests can account for the correlations among SNPs.

In addition, in real GWAS data on admixed populations, there may be cryptic relatedness among individuals. In our future research, we will also account for the cryptic relatedness among individuals in the smooth-GSB procedure by using an efficient mixed-model analysis such as the method GEMMA [31].

Appendix

Null Independence between Statistics $T_{admixture}$ and T_{local} for Quantitative Traits

For individual i , we let y_i , Q_i , D_i and G_i denote the quantitative trait value, global ancestry, local ancestry deviation and genotype at a test SNP, respectively. For n unrelated individuals, we write $\mathbf{y} = (y_1, \dots, y_n)'$, $\mathbf{x}_1 = (1, \dots, 1)'$, $\mathbf{x}_2 = (Q_1, \dots, Q_n)'$, $\mathbf{x}_3 = (D_1, \dots, D_n)'$ and $\mathbf{x}_4 = (G_1, \dots, G_n)'$. Moreover, we write $\mathbf{X}_k = [\mathbf{x}_1, \dots, \mathbf{x}_k]$ and $\mathbf{X}_k = \mathbf{X}_k(\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k'$ for $k = 2, 3, 4$. Apparently, all $\mathbf{X}_k$ and $\mathbf{I} - \mathbf{X}_k$ are idempotent matrices, and $\mathbf{X}_j'(\mathbf{I} - \mathbf{X}_k) = \mathbf{0}$ and thus $\mathbf{H}_j(\mathbf{I} - \mathbf{H}_k) = \mathbf{0}$ for each $j \leq k$ ($= 2, 3, 4$). In $T_{admixture}$ test for quantitative traits, we test the null hypothesis $\alpha_3 = 0$ under

$$\text{Model I: } \mathbf{y} = \mathbf{x}_1 \alpha_1 + \mathbf{x}_2 \alpha_2 + \mathbf{x}_3 \alpha_3 + \boldsymbol{\varepsilon},$$

where random vector $\boldsymbol{\varepsilon}$ follows multivariate normal distribution $N(0, \sigma^2 \mathbf{I})$. We adopt the partial F test statistic

$$F_1 = \frac{\mathbf{y}'(\mathbf{H}_3 - \mathbf{H}_2)\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{H}_3)\mathbf{y} / (n - 3)}$$

to measure the admixture mapping information after adjusting for the global ancestry. Model I is a reduced model of

$$\text{Model II: } \mathbf{y} = \mathbf{x}_1\alpha_1 + \mathbf{x}_2\alpha_2 + \mathbf{x}_3\alpha_3 + \mathbf{x}_4\alpha_4 + \boldsymbol{\varepsilon},$$

where random vector $\boldsymbol{\varepsilon}$ follows a multivariate normal distribution $N(0, \sigma^2\mathbf{I})$. For quantitative traits, the T_{local} tests the null hypothesis $\alpha_4 = 0$ under Model II. We use the partial F test statistic

$$F_2 = \frac{\mathbf{y}'(\mathbf{H}_4 - \mathbf{H}_3)\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{H}_4)\mathbf{y}/(n-4)}.$$

We prove that F_1 and F_2 are *asymptotically independent* if the null hypothesis $\alpha_4 = 0$ is true (no matter whether $\alpha_3 = 0$ or $\alpha_3 \neq 0$). For this purpose, we rewrite the F statistics as

$$F_1 = \frac{c_{1n}}{d_{1n}}$$

and

$$F_2 = \frac{c_{2n}}{d_{2n}},$$

where $c_{1n} = \mathbf{y}'(\mathbf{H}_3 - \mathbf{H}_2)\mathbf{y}/\sigma^2$, $d_{1n} = \mathbf{y}'(\mathbf{I} - \mathbf{H}_3)\mathbf{y}/(\sigma^2(n-3))$, $c_{2n} = \mathbf{y}'(\mathbf{H}_4 - \mathbf{H}_3)\mathbf{y}/\sigma^2$ and $d_{2n} = \mathbf{y}'(\mathbf{I} - \mathbf{H}_4)\mathbf{y}/(\sigma^2(n-4))$.

If $\alpha_4 = 0$ is true under Model II, then $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ with $\boldsymbol{\mu} = \mathbf{x}_1\alpha_1 + \mathbf{x}_2\alpha_2 + \mathbf{x}_3\alpha_3$. By Craig's theorem in the general case, c_{1n} and c_{2n} are independent from each other if and only if $(\mathbf{H}_3 - \mathbf{H}_2)(\mathbf{H}_4 - \mathbf{H}_3) = \mathbf{0}$. Since $\mathbf{H}_j(\mathbf{I} - \mathbf{H}_k) = \mathbf{0}$ for each $j \leq k$ ($= 2, 3, 4$), we observe that

$$\begin{aligned} (\mathbf{H}_3 - \mathbf{H}_2)(\mathbf{H}_4 - \mathbf{H}_3) &= [\mathbf{H}_3\mathbf{H}_4 - \mathbf{H}_3^2] - \mathbf{H}_2(\mathbf{H}_4 - \mathbf{H}_3) \\ &= \mathbf{H}_3[\mathbf{H}_4 - \mathbf{I}] - \mathbf{H}_2[(\mathbf{H}_4 - \mathbf{I}) + (\mathbf{I} - \mathbf{H}_3)] = \mathbf{0}. \end{aligned}$$

Since $\mathbf{I} - \mathbf{H}_3$ is an idempotent matrix of rank $n-3$, there is a unique $n \times (n-3)$ matrix $\mathbf{U}_3$ such that $\mathbf{U}_3\mathbf{U}_3' = \mathbf{I} - \mathbf{H}_3$ and $\mathbf{U}_3'\mathbf{U}_3 = \mathbf{I}_{n-3}$. Under Model II, if $\alpha_4 = 0$, then $\mathbf{y} = \mathbf{X}_3(\alpha_1, \alpha_2, \alpha_3)'\mathbf{U}_3' + \boldsymbol{\varepsilon}$, $(\mathbf{I} - \mathbf{H}_3)\mathbf{y} = (\mathbf{I} - \mathbf{H}_3)\boldsymbol{\varepsilon}$, and hence,

$$d_{1n} = \frac{\boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{H}_3)\boldsymbol{\varepsilon}}{\sigma^2(n-3)} = \frac{(\mathbf{U}_3'\boldsymbol{\varepsilon}/\sigma)'(\mathbf{U}_3'\boldsymbol{\varepsilon}/\sigma)}{n-3}.$$

Since $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}_n)$, $\mathbf{z} \stackrel{\text{def}}{=} \mathbf{U}_3'\boldsymbol{\varepsilon}/\sigma \sim N(0, \mathbf{I}_{n-3})$, the $n-3$ elements of $\mathbf{z}$ are independent and identically distributed (i.i.d.) $N(0,1)$ variables. Hence, we prove $d_{1n} \xrightarrow{\text{a.s.}} 1$ according to standard strong large number theory. In words, d_{1n} converges to 1 almost surely. Similarly, we observe that $d_{2n} \xrightarrow{\text{a.s.}} 1$ under Model II (no matter whether $\alpha_4 = 0$ or $\alpha_4 \neq 0$).

Since $\mathbf{H}_3 - \mathbf{H}_2$ and $\mathbf{H}_4 - \mathbf{H}_3$ are orthogonal idempotent matrices of rank 1, there are unit length vectors $\mathbf{u}, \mathbf{v}$ such that $\mathbf{H}_3 - \mathbf{H}_2 = \mathbf{u}\mathbf{u}'$, $\mathbf{H}_4 - \mathbf{H}_3 = \mathbf{v}\mathbf{v}'$ and $\mathbf{u}'\mathbf{v} = \mathbf{0}$. Under the given condition, we obtain $(\mathbf{H}_3 - \mathbf{H}_2)\mathbf{y}/\sigma = (\mathbf{I} - \mathbf{H}_2)\mathbf{x}_3\alpha_3/\sigma + \mathbf{u}\mathbf{u}'(\boldsymbol{\varepsilon}/\sigma)$, and hence

$$\zeta_1 \stackrel{\text{def}}{=} \mathbf{u}'(\mathbf{H}_3 - \mathbf{H}_2)\mathbf{y}/\sigma = \mu_1 + \mathbf{u}'(\boldsymbol{\varepsilon}/\sigma),$$

where $\mu_1 = \mathbf{u}'(\mathbf{I} - \mathbf{H}_2)\mathbf{x}_3\alpha_3/\sigma$. Under the same condition, we obtain the following representation:

$$\zeta_2 \stackrel{\text{def}}{=} \mathbf{v}'(\mathbf{H}_4 - \mathbf{H}_3)\mathbf{y}/\sigma = \mathbf{v}'(\boldsymbol{\varepsilon}/\sigma).$$

From $\mathbf{u}'\mathbf{v} = 0$ and $(\boldsymbol{\varepsilon}/\sigma) \sim N(0, \mathbf{I})$, we obtain that $(\zeta_1 - \mu_1, \zeta_2)' = (\mathbf{u}'(\boldsymbol{\varepsilon}/\sigma), \mathbf{v}'(\boldsymbol{\varepsilon}/\sigma))' \sim N(0, \mathbf{I}_2)$. Since $d_{1n} \xrightarrow{\text{a.s.}} 1$ and $d_{2n} \xrightarrow{\text{a.s.}} 1$, according to Slutsky's theorem, we have

$$\begin{pmatrix} \frac{\zeta_1 - \mu_1}{\sqrt{d_{1n}}} \\ \frac{\zeta_2}{\sqrt{d_{2n}}} \end{pmatrix} = \begin{bmatrix} \frac{1}{\sqrt{d_{1n}}} & 0 \\ 0 & \frac{1}{\sqrt{d_{2n}}} \end{bmatrix} \begin{pmatrix} \zeta_1 - \mu_1 \\ \zeta_2 \end{pmatrix} \xrightarrow{d.} N(0, \mathbf{I}_2).$$

Therefore,

$$(\zeta_1 - \mu_1)/\sqrt{d_{1n}}$$

and

$$\zeta_2/\sqrt{d_{2n}}$$

are asymptotically independent. Since

$$F_1 = \left(\frac{\zeta_1}{\sqrt{d_{1n}}} \right)^2$$

and

$$F_2 = \left(\frac{\zeta_2}{\sqrt{d_{2n}}} \right)^2,$$

the two F statistics are asymptotically independent.

Acknowledgements

This research was supported by the National Institutes of Health grants: R01GM073766 and R01GM081488 from the National Institute of General Medical Sciences, U01HL101064 from the National Heart, Lung, and Blood Institute, and R01 HD060913 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development. The research was also partly supported by grant UL1TR000058 from the National Center for Advancing Translational Sciences.

The ARIC Study was carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC Study for their important contributions.

References

- 1 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.
- 2 Smith MW, O'Brien SJ: Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat Rev Genet* 2005;6:623–632.
- 3 Montana G, Pritchard JK: Statistical tests for admixture mapping with case-control and cases-only data. *Am J Hum Genet* 2004;75:771–789.

- 4 Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D: Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* 2004;74:979–1000.
- 5 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.
- 6 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.
- 7 Zhu X, Tang H, Risch N: Admixture mapping and the role of population structure for localizing disease genes. *Adv Genet* 2008;60:547–569.
- 8 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.
- 9 Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, Seldin MF: A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. *Am J Hum Genet* 2006;79:640–649.
- 10 Pasaniuc B, Zaitlen N, Lettre G, et al: Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARE and a Breast Cancer Consortium. *PLoS Genet* 2011;7:e1001371.
- 11 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904–909.
- 12 Zhu X, Zhang S, Zhao H, Cooper RS: Association mapping, using a mixture model for complex traits. *Genet Epidemiol* 2002;23:181–196.
- 13 Wang X, Zhu X, Qin H, Cooper RS, Ewens WJ, Li C, Li M: Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics* 2011;27:670–677.
- 14 Qin H, Morris N, Kang SJ, Li M, Tayo B, Lyon H, Hirschhorn J, Cooper RS, Zhu X: Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics* 2010;26:2961–2968.
- 15 Zhu X, Young JH, Fox E, et al: Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARE consortium. *Hum Mol Genet* 2011;20:2285–2295.
- 16 Qin H, Zhu X: Power comparison of admixture mapping and direct association analysis in genome-wide association studies. *Genet Epidemiol* 2012;36:235–243.
- 17 Holm S: A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979;6:65–70.
- 18 Westfall PH, Young SS: Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. New York, Wiley, 1993.
- 19 McCullagh P, Nelder JA: Generalized Linear Models. London, Chapman and Hall, 1989.
- 20 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.
- 21 Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S: Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 2009;5:e1000519.
- 22 Sankararaman S, Sridhar S, Kimmel G, Halperin E: Estimating local ancestry in admixed populations. *Am J Hum Genet* 2008;82:290–303.
- 23 Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguez-Cintron W, Chapela R, Ford JG, Avila PC, Rodriguez-Santana J, Burchard EG, Halperin E: Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 2012;28:1359–1367.
- 24 Chen W, Chen X, Archer KJ, Liu N, Li Q, Zhao Z, Sun S, Gao G: A rapid association test procedure robust under different genetic models accounting for population stratification. *Hum Hered* 2013;75:23–33.
- 25 Roeder K, Devlin B, Wasserman L: Improving power in genome-wide association studies: weights tip the scale. *Genet Epidemiol* 2007;31:741–747.
- 26 Spencer CC, Su Z, Donnelly P, Marchini J: Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009;5:e1000477.
- 27 Su Z, Marchini J, Donnelly P: HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* 2011;27:2304–2305.
- 28 Torgerson DG, Ampleford EJ, Chiu GY, et al: Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet* 2011;43:887–892.
- 29 Lettre G, Palmer CD, Young T, et al: Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARE Project. *PLoS Genet* 2011;7:e1001300.
- 30 Musunuru K, Lettre G, Young T, et al: Candidate gene association resource (CARE): design, methods, and proof of concept. *Circ Cardiovasc Genet* 2010;3:267–275.
- 31 Zhou X, Stephens M: Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012;44:821–824.