frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 10 March 2011
doi: 10.3389/fgene.2011.00011

Capitalizing on admixture in genome-wide association studies: a two-stage testing procedure and application to height in African-Americans

Guolian Kang[1†], Guimin Gao[1], Sanjay Shete[2], David T. Redden[1], Bao-Li Chang[3], Timothy R. Rebbeck[3], Jill S. Barnholtz-Sloan[4], Nicholas M. Pajewski[1] and David B. Allison[1]*

[1] *Section on Statistical Genetics, Department of Biostatistics, The University of Alabama at Birmingham, Birmingham, AL, USA*
[2] *Department of Epidemiology, M. D. Anderson Cancer Center, University of Texas, Houston, TX, USA*
[3] *Department of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[4] *Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA*

Edited by:
Dongxiao Zhu, University of New Orleans, USA

Reviewed by:
Dongxiao Zhu, University of New Orleans, USA
Hua Li, Stowers Institute for Medical Research, USA

***Correspondence:**
David B. Allison, Section on Statistical Genetics, Department of Biostatistics, The University of Alabama at Birmingham, Birmingham, AL 35294, USA.
e-mail: dallison@ms.soph.uab.edu

†Present address:
Guolian Kang, Department of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

As genome-wide association studies expand beyond populations of European ancestry, the role of admixture will become increasingly important in the continued discovery and fine-mapping of variation influencing complex traits. Although admixture is commonly viewed as a confounding influence in association studies, approaches such as admixture mapping have demonstrated its ability to highlight disease susceptibility regions of the genome. In this study, we illustrate a powerful two-stage testing strategy designed to uncover trait-associated single nucleotide polymorphisms in the presence of ancestral allele frequency differentiation. In the first stage, we conduct an association scan by using predicted genotypic values based on regional admixture estimates. We then select a subset of promising markers for inclusion in a second-stage analysis, where association is tested between the observed genotype and the phenotype conditional on the predicted genotype. We prove that, under the null hypothesis, the test statistics used in each stage are orthogonal and asymptotically independent. Using simulated data designed to mimic African-American populations in the case of a quantitative trait, we show that our two-stage procedure maintains appropriate control of the family wise error rate and has higher power under realistic effect sizes than the one-stage testing procedure in which all markers are tested for association simultaneously with control of admixture. We apply the proposed procedure to a study of height in 201 African-Americans genotyped at 108 ancestry informative markers. The two-stage procedure identified two statistically significant markers rs1985080 (*PTHB1/BBS9*) and rs952718 (*ABCA12*). *PTHB1/BBS9* is downregulated by parathyroid hormone in osteoblastic cells and is thought to be involved in parathyroid hormone action in bones and may play a role in height. *ABCA12* is a member of the superfamily of ATP-binding cassette transporters and its potential involvement in height is unclear.

Keywords: two-stage, structured association testing, admixture mapping, regional admixture estimate, genome-wide association studies

INTRODUCTION

One of the major focuses of current genomics research is the expansion of association studies beyond populations of European and Asian descent, including African populations and admixed populations such as African-Americans and Hispanics. Although these investigations carry several potential pitfalls such as greater haplotype diversity and lower levels of linkage disequilibrium (LD), one of the most well-known issues is the potential confounding influence of population stratification and admixture (Marchini et al., 2004; Smith et al., 2004; Teo et al., 2010). However, the existence of these phenomena also presents an opportunity, as several recent studies have demonstrated that genetic ancestry need not be viewed as a nuisance quantity. For example, within the context of autoimmune diseases, Richman et al. (2010) illustrated a role for European population substructure across the northwest to southeast cline with endophenotypes of systemic lupus erythematosus. Similarly, Hughes et al. (2008) validated the role of the *HLA-DRB1* shared epitope within African-Americans with rheumatoid arthritis, suggesting an inheritance through admixture with European populations.

Earlier investigators have recognized the value of considering admixture to highlight disease susceptibility regions in the genome, spawning the approach of admixture mapping or mapping by admixture LD (Patterson et al., 2004; Freedman et al., 2006). The basic premise of these approaches is that individuals from admixed populations would have a greater probability of inheriting risk alleles from the ancestral population that carries more of such alleles. The initial appeal of admixture mapping was the potential savings in genotyping costs because the genome could be covered with a few thousand markers with reasonable resolution. However, with the rapid cost decrease for platforms assaying potentially millions of single nucleotide polymorphisms (SNPs), the initial appeal of

admixture mapping has dwindled. Here we offer a new insight that there is benefit to considering the admixture mapping paradigm within genome-wide association (GWA) studies of admixed populations using high-density genotyping arrays.

A major challenge in GWA studies is to balance the control of type I and type II errors. If no adjustment for multiple-testing is used, with hundreds of thousands to millions of tests, the number (and proportion) of false-positives among the results declared significant is likely to be enormous. In contrast, if the Bonferroni correction [or any other method that controls the family wise type I error rate (FWER)] is used, power may be reduced excessively and too many type II errors (false-negatives) may be made (Kang et al., 2009). If there were a way to reduce the number of null hypotheses tested without discarding too many markers that are truly in LD with causative loci, then power could be improved dramatically. With this in mind, several authors have considered various two-stage testing paradigms (Evans et al., 2006; Laird and Lange, 2006; Skol et al., 2006; Wang et al., 2006; Ionita-Laza et al., 2007). Within the context of family based association studies, these approaches entail partitioning the available data into two orthogonal components. The between-family component is used to provide an initial relative ranking of the markers, then using the within-family component to provide a second-stage test of association. Ideally, such two-stage testing paradigms (a) should not require family data, (b) should be robust to confounding by non-random mating (including admixture), (c) should offer strong control of the FWER, and (d) should not arbitrarily split the available data and suffer the attendant loss in power (Allison and Coffey, 2002).

Our intent here is to illustrate that other sources of information, such as admixture, can be used to provide an orthogonal data partition and hence a two-stage testing opportunity satisfying the features listed above. In our method, we divide the association analysis for an admixed population into two parts, one of which tests the association between the phenotype and a predicted genotype based on regional admixture estimates. We then select a subset of promising markers for the second-stage analysis where we test the association between the observed genotype and the phenotype conditional on the predicted genotypes. Because the test statistics used in each stage of the procedure are orthogonal and asymptotically independent under the null hypothesis (see the proof in Appendix), this two-stage procedure maintains appropriate control of the FWER whether or not confounding by admixture exists. As a proof of concept for our new approach, we compare our proposed procedure through simulation to a one-stage procedure in the case of association mapping in an admixed population. We conclude with an illustration of the proposed method within a study of height in African-Americans using ancestry informative markers.

METHODS

We consider the situation of $j = 1, 2, \ldots, J$ ancestry informative SNP markers and $i = 1, 2, \ldots, N$ individuals. We denote $G_{i,j}$ as the observed genotypic value for the ith individual at the jth SNP. For simplicity, we assume that the admixed population sample arises from two ancestral populations (generically labeled as populations A and B). Let $A_{i,j}$ denote a regional admixture estimate, that is, the estimated expected number of alleles inherited from ancestral population A at the jth SNP for the ith individual. Finally, let Y_i denote the observed phenotype for the ith individual.

ONE-STAGE TESTING PROCEDURE

A standard flexible approach for association testing while controlling for admixture is regression within the generalized linear model (GLM) framework, which directly allows for quantitative, binary, ordinal, and time to event (survival) phenotypic distributions through the choice of an appropriate link function (g) (McCullagh and Nelder, 1989; Freedman et al., 2006; Redden et al., 2006; Zhu and Cooper, 2007). This involves a regression model for the jth SNP that assumes that the expected phenotypic value takes the form,

$$E[Y_i] = g^{-1}\left(\gamma_0 + \gamma_1 G_{i,j} + \gamma_2 A_{i,j}\right), \text{for } i = 1,\ldots,N, j = 1,\ldots,J,$$

where $A_{i,j} = 2\omega_{i,2}(j) + \omega_{i,1}(j)$, $\omega_{i,1}(j)$ ($\omega_{i,2}(j)$) are the probabilities that individual i has one (two) allele(s) at the jth marker from ancestral population A, and $g^{-1}(\cdot)$ denotes the inverse link function (Redden et al., 2006; Tiwari et al., 2008). For ease of exposition, we will assume the situation of a quantitative trait, taking $g(\cdot)$ to be the identity link function and introducing residual error terms $\varepsilon_{i,j}$ each independently distributed $N\left(0,\sigma_1^2\right)$. To estimate the admixture proportions $A_{i,j}$, we utilize the Hidden Markov Model approach implemented in *Ancestrymap* (Patterson et al., 2004), although a number of alternative estimation approaches exist and could be readily substituted (Sankararaman et al., 2008; Price et al., 2009). In order to control for multiple-testing, we employ a Bonferroni correction, testing the significance of γ_1 for each SNP at a significance level of α/J. This controls the FWER at the desired upper bound of α. Though we focus on control of the FWER here, one could analogously look at control of other error rates, such as the false discovery rate (Benjamini and Hochberg, 1995, 2000; Storey, 2002).

PROPOSED TWO-STAGE TESTING PROCEDURE

Our proposed two-stage method is predicated on the realization that an association analysis incorporating the ancestry estimate can be divided into two aspects. First, we fit a model using the conditional expectation of genotype, where the expectation is now taken relative to regional admixture estimates for the particular SNP. The second-stage then tests the association of a subset of promising markers based on the first stage screen, where association is now tested conditioned on the conditional genotypic expectation. We prove the orthogonality of the two test statistics used in each stage under the null hypothesis in Appendix, and use simulation to illustrate that the two statistics may be correlated under the alternative hypothesis in the case of admixed populations.

Stage 1: We regress the observed genotypic value ($G_{i,j}$) at each marker on the estimated average number ($A_{i,j}$) of population A-ancestry alleles

$$G_{i,j} = \varphi_0 + \varphi_1 A_{i,j} + e_{i,j}, i = 1,\ldots,N, j = 1,\ldots,J, \quad (1)$$

where $e_{i,j}$ represents residual error terms. This equation is then used to obtain a predicted genotypic value $\widehat{G}_{i,j} = \widehat{E}[G_{i,j}|A_{i,j}] = \widehat{\varphi}_0 + \widehat{\varphi}_1 A_{i,j}$.

We then consider a regression of the quantitative trait on the predicted genotypes as

$$Y_i = \alpha_0 + \alpha_1 \widehat{G}_{i,j} + \tau_{i,j}, i = 1,\ldots,N, j = 1,\ldots,J, \quad (2)$$

where $\tau_{i,j}$ are independently distributed $N\left(0,\sigma_3^2\right)$. We test the significance of α_1 at each marker on the basis of Eq. 2 and select the top q markers for testing in the second-stage. We denote the selected subset of markers here as Φ. Approaches to selecting q will be discussed below.

Stage 2: In the second-stage, we consider a linear regression for the quantitative trait by using the observed genotype as,

$$Y_i = \beta_0 + \beta_1 \hat{G}_{i,j} + \beta_2 (G_{i,j} - \hat{G}_{i,j}) + \vartheta_{i,j}, i = 1,\ldots,n, j \in \Phi, \quad (3)$$

where $\vartheta_{i,j}$ are independently distributed $N(0,\sigma_4^2)$. We test the significance of $\hat{\beta}_2$ at each of the "q" selected markers from stage 1 on the basis of Eq. 3 at a significance level of α/q, where α is the overall significance level. The use of only q in the denominator of the Bonferroni correction is justified by the orthogonality and asymptotic independence under the null hypothesis proved in Appendix (Van Steen et al., 2005; Zheng et al., 2007).

SIMULATION DESIGN

To evaluate the frequency characteristics of our proposed procedure, we simulated an admixed population sample by using *Ancestrymap*. We utilized parameter settings designed to mimic an African-American population (Patterson et al., 2004). The average proportion of alleles inherited from the European ancestral population was set at 1/6, with the number of chromosomal exchanges per Morgan between ancestral segments of the genome since the mixing event set as 10. For the simulations under the alternative hypothesis, we randomly set one marker as a disease marker by setting the "risksim" parameter in *Ancestrymap* (rel8500) (ψ_1) to a value other than 1, where ψ_1 is the increased risk for disease due to carrying one population *A*-ancestry allele at the disease marker (Patterson et al., 2004).

We simulated a quantitative trait by using the equation (Redden et al., 2006)

$$Y_i = sA_{ij} + tG_{ij} + \varepsilon_i, \quad (4)$$

where ε_i is assumed to have a standard normal distribution. s denotes the overall effect of admixture on the trait, while t denotes the mean genotypic effect on the trait. We use simulation to illustrate the correlations of the test statistics in our two-stage procedure under the null and alternative hypotheses. We simulated 200 data sets each with 400 cases and 400 controls genotyped at the 1805 ancestry informative SNPs with one disease-predisposing allele. We then randomly selected one marker and simulated a continuous trait using Eq. 4 above with s set to be 0, 0.1, and 0.3 and t set equal to 0, 0.2, and 0.4 at the selected marker.

FWER EVALUATION

We estimated the FWER as the proportion of replicates in which at least one non-disease-associated SNP was found to be significantly associated with the disease, under two situations: (1) under the null hypothesis that there is no SNP associated with the trait with and without confounding association by admixture and (2) under the non-complete null hypothesis, in which some ancestry SNPs are associated with the trait and the associations are confounded by admixture between these ancestry SNPs and the trait. It is possible to get false-positive results at ancestry SNPs that are not associated with the trait.

To evaluate the FWER under the complete null hypothesis under situation 1, we first simulated 200 cases and 200 controls at 1805 ancestry SNPs under the complete null hypothesis using *Ancestrymap*. Then, we randomly selected one marker and simulated the phenotype by using Eq. 4 with s = 0, 0.1, and 0.3 at the selected marker and t = 0, where the non-zero value of s was chosen to ensure that the phenotype variability explained by admixture was less than 3% (the average value of this value from our simulated data can be found in **Table A1** in the Appendix). The FWER was estimated as the proportion of replicates that identified any one of 1805 ancestry SNPs as significant.

To evaluate the FWER under the non-complete null hypothesis (situation 2), we first simulated 200 (400) cases and 200 (400) controls with one preset disease-associated ancestry SNP by using the software *Ancestrymap*. Then, we simulated the phenotype by using Eq. 4 with s = 0, 0.1, and 0.3 and t = 0.2, 0.4, and 0.6, respectively, where the non-zero value of s was chosen to ensure that the phenotype variability explained by admixture was less than 3% (the average value of this value from our simulated data can be found in **Table A2** in the Appendix). The FWER was estimated by the proportion of replicates where any one of the remaining ancestry SNPs was identified as being significant after the ancestry SNPs located at the same chromosome with the disease-associated ancestry SNP were removed from consideration.

POWER EVALUATION

To estimate the power of the two-stage procedure, we first simulated 200 (400) cases and 200 (400) controls with 1805 ancestry SNPs and randomly chose 1 of the 1805 ancestry SNPs located at chromosome 1 as a specific disease-associated ancestry SNP at which a population *A*-ancestry allele confers 2.4 multiplicative increased risk, where the multiplicative increased risk was chosen to ensure a high power under the scaled sample sizes. Then, we simulated the phenotype by using Eq. 4 with s = 0, 0.1, and 0.3 and t = 0.2, 0.4, and 0.6, respectively, where G in Eq. 4 is the genotype for the specific disease-associated ancestry SNP we chose above. For the estimation of power, we estimated the power level as the proportion of replicates where the specific disease-associated ancestry SNP at chromosome 1 was successfully identified.

SIMULATION RESULTS

CORRELATION EVALUATIONS BETWEEN TWO TEST STATISTICS UNDER THE ALTERNATIVE HYPOTHESIS

Table 1 and **Figures A1 and A2** in the Appendix show that these two test statistics were not correlated under the null and were correlated under the alternative hypothesis based on 200 datasets each with 800 individuals whether confounding by admixture existed or not. The level of correlations seems to increase as the effects of both genotype and the ancestry estimate on the trait increase. The correlations of two test statistics in the two-stage procedure under the alternative hypothesis further support the conclusion that our two-stage procedure has higher power than the one-stage procedure (see below).

FWER EVALUATION

Because the test statistics in each stage of our two-stage procedure are asymptotically independent under the null hypothesis, the FWER of our two-stage procedure should theoretically be

Table 1 | Correlation evaluations of two test statistics in stage 1 and stage 2 for our two-stage procedure.

		s^b		
t^a	Correlation	0	0.1	0.3
NULL HYPOTHESIS				
0	ρ[c]	0.086	0.078	0.072
	p-value[d]	0.228	0.273	0.314
NON-COMPLETE NULL HYPOTHESIS				
0.2	ρ[c]	–0.075	–0.171	–0.156
	p-value[d]	0.290	0.015	0.028
0.4	ρ	–0.176	–0.231	–0.190
	p-value	0.013	0.001	0.007

[a]*The effect of genotype on the trait.*
[b]*The effect of confounding association on the trait.*
[c]*The Spearman's* ρ.
[d]*The p-value for testing correlation between two test statistics in the two-stage procedure based on Spearman's* ρ *statistic under null hypothesis of* ρ = 0.

controlled (Kang et al., 2009). We therefore next evaluated whether our two-stage procedure could effectively control the FWER by the preset limited sample size.

Figures 1 and 2A,B plot the estimated FWERs versus the ratio of the number of ancestry SNPs selected in the first stage (q) to the number of total SNPs (h) under the complete null hypothesis for a quantitative trait with and without association confounding by admixture based on 200 replicates. **Figures 2A,B** are for $s = 0.1$ and 0.3, respectively (confounding by admixture). These figures illustrate that both the one-stage procedure and our two-stage procedure provide adequate control of the FWER.

For the non-complete null hypothesis, refer to the columns labeled FWER in **Tables 2 and 3**. As shown in these two tables, all the estimated FWERs were close to the nominal values of 0.1 and 0.05. Therefore, our two-stage procedure conserved good control of the FWER. On the other hand, we also found that our two-stage procedure still had a conservative FWER when q/h was close to 0 under the alternative hypothesis.

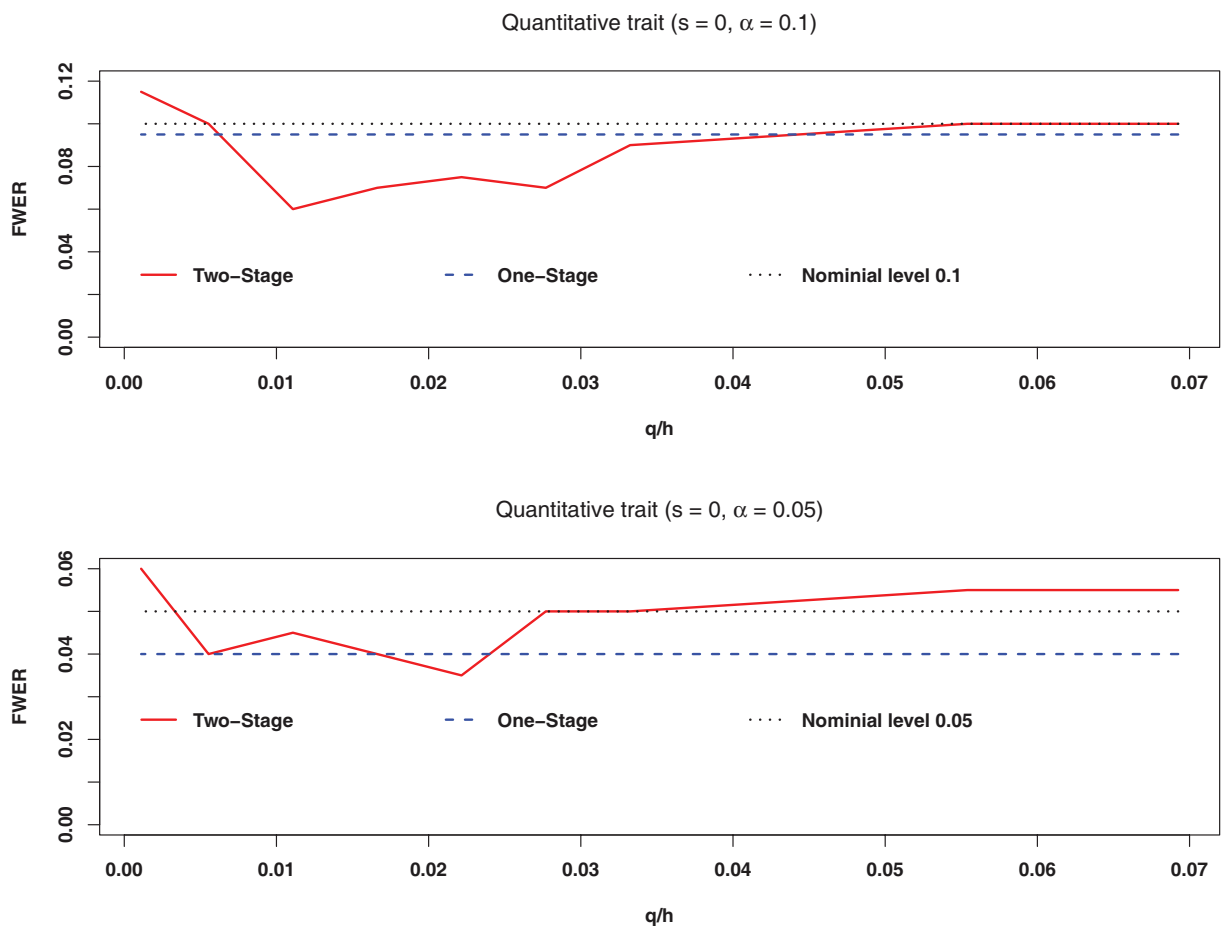

FIGURE 1 | The estimated FWER of the two-stage procedure without confounding association by admixture at significance levels of 0.1 and 0.05 (200 replicates). The red solid line represents the estimated FWER with the two-stage procedure of structured association tests; the blue dashed line represents the estimated FWER with the one-stage procedure of structured association tests; the black dotted line represents the nominal level; q is the number of SNPs selected in the first stage; and h is the number of total SNPs.

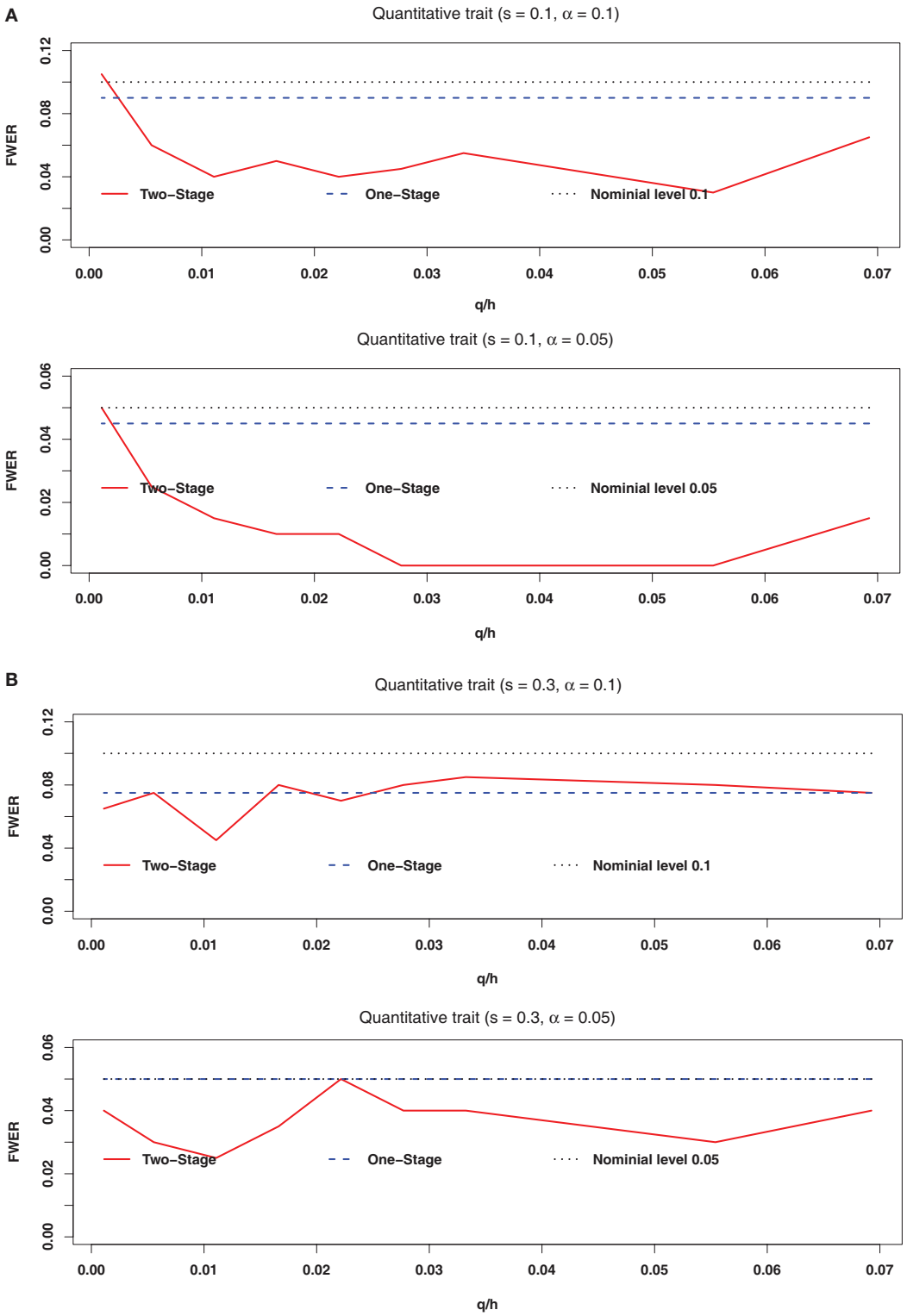

FIGURE 2 | The estimated FWER of the two-stage procedure with confounding association by admixture at significance levels of 0.1 and 0.05 (200 replicates). (A) Is for $s = 0.1$ and **(B)** is for $s = 0.3$. The red solid line represents the estimated FWER with the two-stage procedure of structured association tests; the blue dashed line represents the estimated FWER with the one-stage procedure of structured association tests; the black dotted line represents the nominal level; q is the number of SNPs selected in the first stage; and h is the number of total SNPs.

Table 2 | Empirical power and FWER of the two-stage procedure at a significance level of 0.05 (200 replicates).

t[a]	Method	*q*/*h*[b]	*s*[c] = 0				*s* = 0.1				*s* = 0.3			
			400[d]		800		400		800		400		800	
			Power[e]	FWER[f]	Power	FWER	Power	FWER	Power	FWER	Power	FWER	Power	FWER
0.2	OS[g]		0.005	0.031	0.115	0.02	0.01	0.03	0.09	0.055	0.005	0.045	**0.121**	0.04
	TS[h]	250/1805[i]	0.015	0.061	0.18	0.06	0.015	0.03	0.11	0.04	**0.01**	0.04	0.05	0.03
		125/1805	0.02	0.036	0.2	0.05	0.02	0.05	0.12	0.035	**0.01**	0.045	0.02	0.05
		100/1805	0.031	0.036	0.2	0.05	**0.025**	0.045	**0.125**	0.03	0.005	0.04	0.015	0.06
		25/1805	**0.036**[j]	0.02	**0.27**	0.045	**0.025**	0.04	0.08	0.045	0	0.015	0.005	0.06
		5/1805	0.031	0.036	0.225	0.01	0.005	0.055	0.04	0.04	0	0.01	0.005	0.055
		2/1805	0.026	0.031	0.19	0.025	0	0.065	0.02	0.02	0	0.03	0.005	0.05
0.4	OS[g]		0.35	0.03	0.875	0.05	0.445	0.04	0.865	0.04	0.362	0.05	0.875	0.065
	TS[h]	250/1805	0.505	0.03	0.92	0.03	0.555	0.045	0.92	0.06	**0.447**	0.055	**0.91**	0.06
		125/1805	0.555	0.045	0.935	0.035	0.605	0.05	0.945	0.03	0.412	0.04	0.87	0.04
		50/1805	0.62	0.015	0.95	0.03	**0.68**	0.03	**0.96**	0.04	0.372	0.035	0.795	0.02
		25/1805	0.655	0.02	**0.975**	0.01	0.64	0.02	0.95	0.015	0.322	0.025	0.745	0.02
		5/1805	**0.67**	0.015	0.965	0	0.495	0.02	0.865	0.005	0.196	0.045	0.57	0.015
		2/1805	0.545	0.005	0.945	0	0.395	0.035	0.8	0	0.146	0.045	0.465	0.04
0.6	OS[g]		0.93	0.06	**1**	0.065	0.9	0.06	**1**	0.06	0.915	0.04	**1**	0.06
	TS[h]	250/1805	0.96	0.055	0.985	0.06	0.95	0.06	0.99	0.035	0.96	0.035	0.99	0.065
		125/1805	0.965	0.045	0.985	0.06	0.95	0.04	0.985	0.025	**0.97**	0.01	0.99	0.03
		50/1805	0.985	0.02	0.985	0.045	0.98	0.025	0.985	0.02	**0.97**	0.04	0.99	0.015
		25/1805	**0.995**	0.01	0.985	0.01	**0.985**	0.015	0.985	0.01	0.93	0.015	0.98	0.025
		5/1805	0.97	0.005	0.985	0.005	0.97	0.005	0.985	0	0.785	0.015	0.95	0
		2/1805	0.93	0	0.985	0	0.91	0	0.985	0.005	0.69	0.025	0.93	0

[a] *t is the genotypic effect of the disease marker on the quantitative trait.*
[b] *q is the number of SNPs selected in the first stage; h is the number of total SNPs.*
[c] *s is the effect of confounding association on the trait.*
[d] *400 individuals and 800 individuals.*
[e] *The power is estimated by the proportion of replicates successfully identifying the specific disease SNP.*
[f] *FWER, family wise error rate, which is estimated by the proportion of replicates wrongly identifying any one of the SNPs located at chromosomes 2 to chromosome 22.*
[g] *OS, one-stage procedure.*
[h] *TS, two-stage procedure.*
[i] *250/1805 = 0.139, 125/1805 = 0.069, 50/1805 = 0.028, 25/1805 = 0.014, 5/1805 = 0.003, 2/1805 = 0.001.*
[j] *The maximum power of both the OS and TS is marked in bold.*

POWER COMPARISONS

We compared the power of our two-stage procedure with that of the one-stage procedure described above for a quantitative trait. **Tables 2 and 3** and **Tables A4 and A5** in the Appendix present the empirical power of the two-stage procedure for simulated 200 replicates at significance levels of 0.05 and 0.1, respectively. From these two tables we found that (1) the two-stage procedure generally had higher power than the one-stage procedure; (2) the maximum power of the two-stage procedure was significantly higher than that of the one-stage procedure when there was no or a small or moderate association confounded by admixture and there was a small or moderate true association between disease and marker; (3) as the effect size of association confounded by admixture increased, the power of the two-stage procedure decreased [for example, when $t = 0.4$, $s = 0$, 0.1 and 0.3, $\alpha = 0.05$, and $n = 400$, the difference between the maximum value of the power of the two-stage procedure and that of the one-stage procedure was about 32% (0.67 versus 0.35), 23.5% (0.68 versus 0.445), and 8.5% (0.362 versus 0.447), respectively]; and (4) the selection of q affected the power of our two-stage procedure. As it approaches 1, the power of the two-stage procedure was higher than and close to that of the one-stage procedure. But the selection of q is correlated with the effect sizes of true association and confounding association by admixture. The optimal number of q is approximately 3% (≈50/1805) for all s and t.

In addition, we also noticed that as the effect size of the true association between the trait and the marker increased, the effect of association confounded by admixture on the power first increased and then decreased; but as the sample size increased, the effect of association confounded by admixture on the power decreased. For example, for $\alpha = 0.05$, the difference in the maximum value of the power for $s = 0$ and 0.3 increased first from 2.6% (=0.036 − 0.01) to 22.3.5% (=0.67 − 0.447) and then decreased to 2.5% (=0.995 − 0.97) when $n = 400$. However, when $n = 800$, the above three values were from 15.3% (=0.27 − 0.121) to 6.5% (=0.975 − 0.91) to 0% (=1 − 1).

Table 3 | Empirical power and FWER of the two-stage procedure at a significance level of 0.1 (200 replicates).

t[a]	Method	*q*/*h*[b]	*s*[c] = 0				*s* = 0.1				*s* = 0.3			
			400[d]		800		400		800		400		800	
			Power[e]	FWER[f]	Power	FWER	Power	FWER	Power	FWER	Power	FWER	Power	FWER
0.2	OS[g]		0.025	0.041	0.105	0.05	0.021	0.082	0.09	0.08	**0.015**	0.056	**0.1**	0.07
	TS[h]	250/1805[i]	0.03	0.076	0.2	0.125	0.046	0.103	0.2	0.105	0.005	0.097	0.07	0.08
		125/1805	0.046	0.071	0.225	0.1	0.041	0.056	**0.22**	0.09	0	0.097	0.055	0.085
		50/1805	0.041	0.076	0.255	0.07	**0.056**	0.092	0.17	0.08	0	0.087	0.04	0.1
		25/1805	0.056	0.066	**0.26**	0.08	0.046	0.087	0.14	0.075	0	0.077	0.03	0.115
		5/1805	**0.066**[j]	0.086	0.215	0.07	0.031	0.072	0.11	0.045	0	0.056	0.01	0.1
		2/1805	0.051	0.107	0.195	0.06	0.015	0.062	0.06	0.09	0	0.087	0	0.105
0.4	OS[g]		0.44	0.11	0.925	0.07	0.48	0.09	0.89	0.105	0.431	0.113	0.894	0.086
	TS[h]	250/1805	0.595	0.085	0.95	0.075	0.66	0.11	0.96	0.05	**0.513**	0.103	**0.909**	0.101
		125/1805	0.645	0.035	0.96	0.065	0.67	0.075	0.97	0.06	0.477	0.087	0.828	0.086
		50/1805	0.74	0.04	0.965	0.05	**0.7**	0.06	0.97	0.04	0.374	0.067	0.768	0.061
		25/1805	**0.765**	0.035	**0.97**	0.03	0.69	0.06	**0.975**	0.06	0.333	0.056	0.727	0.071
		5/1805	0.735	0.03	0.96	0.01	0.575	0.045	0.935	0.01	0.185	0.072	0.53	0.051
		2/1805	0.675	0.02	0.925	0.005	0.46	0.075	0.87	0.005	0.138	0.067	0.46	0.051
0.6	OS[g]		0.944	0.101	**1**	0.12	0.949	0.066	**1**	0.135	0.95	0.1	**1**	0.121
	TS[h]	250/1805	0.97	0.076	0.985	0.11	**0.99**	0.076	0.985	0.115	**0.985**	0.065	**1**	0.07
		125/1805	0.99	0.071	0.985	0.075	**0.99**	0.051	0.985	0.075	0.975	0.065	0.99	0.095
		50/1805	**1**	0.066	0.985	0.06	**0.99**	0.04	0.985	0.05	0.945	0.085	0.985	0.065
		25/1805	**1**	0.02	0.985	0	**0.99**	0.02	0.985	0.025	0.925	0.045	0.985	0.01
		5/1805	**1**	0	0.985	0.005	0.965	0.01	0.985	0	0.785	0.025	0.965	0.005
		2/1805	0.97	0	0.975	0	0.919	0.005	0.98	0	0.72	0.01	0.945	0.01

[a] *t is the genotypic effect of the disease marker on the quantitative trait.*
[b] *q is the number of SNPs selected in the first stage; h is the number of total SNPs.*
[c] *s is the effect of confounding association on the trait.*
[d] *400 individuals and 800 individuals.*
[e] *The power is estimated by the proportion of replicates successfully identifying the specific disease SNP.*
[f] *FWER, family wise error rate, which is estimated by the proportion of replicates wrongly identifying any one of the SNPs located at chromosomes 2 to chromosome 22.*
[g] *OS, one-stage procedure.*
[h] *TS, two-stage procedure.*
[i] *250/1805 = 0.139, 125/1805 = 0.069, 50/1805 = 0.028, 25/1805 = 0.014, 5/1805 = 0.003, 2/1805 = 0.001.*
[j] *The maximum power of both the OS and TS is marked in bold.*

Furthermore, it was interesting that under the non-complete null hypothesis our two-stage procedure could have higher power with lower FWER if we chose fewer markers from stage 1 for stage 2 analysis compared with the one-stage procedure, especially when there was moderate or large true association between the trait and the marker. This happens because under the non-complete null hypothesis, if we choose fewer promising markers in stage 1 for stage 2 analysis, there is a smaller chance of the false-positives occurring with nearly no effect on true-positives.

APPLICATION TO HEIGHT IN AFRICAN-AMERICANS

To evaluate the performance of our new two-stage procedure, we applied it to a real data set investigating the association of 108 ancestry informative markers with height in a sample of 201 African-Americans. Detailed information on the 108 ancestry markers can be found in **Table A3** in the Appendix. Participants were part of an ongoing case–control study of genetic risk factors for prostate cancer conducted by investigators at the University of Pennsylvania (Zeigler-Johnson et al., 2004; Stefflova et al., 2009). Height was based on self-report of the subject's tallest height ever reached in inches. Genetic map positions for all markers were evaluated by using a program developed by McKeigue (2006).

For the purpose of comparison, we first conducted a linear regression evaluating the association between height and each SNP. We employed two methods to account for the confounding influence of admixture; Genomic Control (Devlin and Roeder, 1999) and principal components analysis (Price et al., 2006). The genomic control inflation factor was calculated by dividing the median of the test statistics for all SNPs by 0.456. We also conducted one-stage analysis as described before. No SNP was found to be statistically significant at an overall nominal level of 0.05 (0.05/108 for each SNP) by the above three methods. Finally, we conducted our two-stage analysis. On the basis of our simulation results above, we selected the top three SNPs ($\approx 108 \times 0.03$) in stage 1 and tested these three SNPs in stage 2 at an overall nominal level of 0.05 (0.05/3 for each SNP; see Methods). **Table 4** shows the association results at a preset nominal level of 0.05 using our proposed two-stage testing procedure. We

Table 4 | The association results of ancestry informative markers with height at a nominal level of 0.05 by the two-stage procedure.

rs ID	Chromosome	Gene	Physical position	*p*-value in stage 1	Rank in stage 1	*p*-value in stage 2
rs952718	2	*ABCA12*	215714130	0.074	2	0.011
rs1985080	7	*BBS9*	33400099	0.098	3	0.014

found that the two-stage procedure identified two statistically significant ancestry markers (rs952718 and rs1985080) associated with height after controlling for association confounded by admixture.

DISCUSSION

In this study, we have introduced a new two-stage procedure for association mapping in admixed populations. Our simulations indicate that the two-stage procedure had significantly higher power compared with a one-stage procedure and adequately controlled the FWER whether or not the admixture confounded the true association between genotype and trait. Because the performance of our two-stage method depends on the selection of the number of the top markers, we recommend that the top 3% markers be selected in stage 1 for stage 2 analysis in practice. In our real data example, using the one-stage procedure and the other two methods, we found no significant associations; however the two-stage procedure found two ancestry informative SNPs, rs1985080 (*PTHB1/BBS9*) and rs952718 (*ABCA12*), to be significantly associated with height in African-Americans. *PTHB1/BBS9* (parathyroid hormone-responsive B1) is downregulated by parathyroid hormone in osteoblastic cells and is thought to be involved in parathyroid hormone action in bones and may play a role in height (Adams et al., 1999). *ABCA12* [ATP-binding cassette (ABC), sub-family A (ABC1), member 12] is a member of the superfamily of ABC transporters (Annilo et al., 2002). *ABCA12* is a major causative gene for non-bullous congenital ichthyosiform erythroderma (Sakai et al., 2009), but its role in determining height merits further study.

Certain limitations of our proposed method deserve consideration. From empirical data across a range of traits and species, it has been suggested that most genetic variance is additive, which accounts for over half, and often close to 100%, of the total genetic variance (Hill et al., 2008). Thus, in our analysis we focused on the situation of additive genetic effects. If the underlying disease model follows a different mode of inheritance, then the proposed procedure will lose power. However, the proposed models can be straightforwardly adjusted to conduct a 2° of freedom genotypic test, which is robust to the underlying mode of inheritance. In addition, we only carried a subset of promising markers into a second-stage association analysis. Within the context of two-stage family based testing procedures, Ionita-Laza et al. (2007) have suggested that it may be more powerful to test all markers at the second-stage, weighting according to the first stage results. Thus, a point for future research will be to investigate how to optimally conduct two-stage testing procedures based on admixture information.

In addition, our approach is not intended to be used nor is it likely to be useful in all situations. When the correlation between admixture and the observed genotypes is zero, as will happen in regions of the genome that display little to no allele frequency differentiation across populations (or could occur in completely panmictic populations over many generations with no selection, no segregation distortion, and so on), the two-stage approach we propose will have no value. In situations in which the correlation between the adjusted genotypes and the observed genotypes is 1.0, there would also be no value in our two-stage approach because there will be perfect collinearity. Somewhere between zero and one must lie an optimum, and finding that optimum for different circumstances can be a topic for future research.

WEB RESOURCES

R programs implementing the proposed methods can be downloaded from http://www.soph.uab.edu/ssg/

ACKNOWLEDGMENTS

We would like to thank Dr. Nicholas Patterson for his help with modifications of program *Ancestrymap*. This research was supported by grants R01-GM073766 and R01-GM077490 from the National Institute of General Medical Sciences and T32-HL072757 from the National Heart, Lung, and Blood Institute.

REFERENCES

Adams, A. E., Rosenblatt, M., and Suva, L. J. (1999). Identification of a novel parathyroid hormone-responsive gene in human osteoblastic cells. *Bone* 24, 305–313.

Allison, D. B., and Coffey, C. S. (2002). Two-stage testing in microarray analysis: what is gained? *J. Gerontol. A Biol. Sci. Med. Sci.* 57, B189–B192.

Annilo, T., Shulenin, S., Chen, Z. Q., Arnould, I., Prades, C., Lemoine, C., Maintoux-Larois, C., Devaud, C., Dean, M., Denèfle, P., and Rosier, M. (2002). Identification and characterization of a novel ABCA subfamily member, ABCA12, located in the lamellar ichthyosis region on 2q34. *Cytogenet. Genome Res.* 98, 169–176.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 85, 289–300.

Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of the false discovery fate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* 25, 60–83.

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

Evans, D. M., Marchini, J., Morris, A. P., and Cardon, L. R. (2006). Two-stage two-locus models in genome-wide association. *PLoS Genet.* 2, e157. doi: 10.1371/journal.pgen.0020157

Freedman, M. L., Haiman, C. A., Patterson, N., McDonald, G. J., Tandon, A., Waliszewska, A., Penney, K., Steen, R. G., Ardlie, K., John, E. M., Oakley-Girvan, I., Whittemore, A. S., Cooney, K.A., Ingles, S. A., Altshuler, D., Henderson, B. E., and Reich, D. (2006). Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14068–14073.

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* 4, e1000008. doi: 10.1371/journal.pgen.1000008

Hughes, L., Morrison, D., Kelley, J., Padilla, M., Vaughan, L., Westfall, A. O., Dwivedi, H., Mikuls, T. R., Holers, V. M., Parrish, L. A., Alarcón, G. S., Conn, D. L., Jonas, B. L., Callahan, L. F., Smith, E. A., Gilkeson, G. S., Howard, G., Moreland, L. W., Patterson, N., Reich, D. S., and Louis Bridges, Jr. (2008). The HLA-DRB1 shared epitope is associated with susceptibility to rheumatoid arthritis in African Americans through European genetic admixture. *Arthritis Rheum.* 58, 349–358.

Ionita-Laza, I., McQueen, M., Laird, N., and Lange, C. (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am. J. Hum. Genet.* 81, 607–614.

Kang, G. L., Ye, K. Y., Liu, N. J., Allison, D. B., and Gao, G. M. (2009). Weighted multiple hypothesis testing procedures. *Stat. Appl. Genet. Mol. Biol.* 8, 23.

Laird, N. M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* 7, 385–394.

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

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Model*, 2nd Edn. New York: Chapman & Hall/CRC Press.

McKeigue, P. (2006). Smoothing estimates of genetic map distance over short intervals. Available at: http://integrin.ucd.ie/cgi-bin/rs2cm.cgi (accessed July 5, 2010).

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, A., Daly, M. J., and David Reich, D. (2004). Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* 74, 979–1000.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909.

Price, A., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* 5, e1000519. doi: 10.1371/journal.pgen.1000519

Redden, D. T., Divers, J., Vaughan, L. K., Tiwari, H. K., Beasley, T. M., Fernández, J. R., Kimberly, R. P., Feng, R., Padilla, M. A., Liu, N., Miller, M. B., and Allison, D. B. (2006). Regional admixture mapping and structured association testing: conceptual unification and an extensible general linear model. *PLoS Genet.* 2, e137. doi: 10.1371/journal.pgen.0020137

Richman, I. B., Chung, S. A., Taylor, K. E., Kosoy, R., Tian, C., Ortmann, W. A., Nititham, J., Lee, A. T., Rutman, S., Petri, M., Manzi, S., Behrens, T. W., Gregersen, P. K., Seldin, M. F., and Criswell, L. A. (2010). European population substructure correlates with systemic lupus erythematosus endophenotypes in North Americans of European descent. *Genes Immun.* 11, 515–521.

Sakai, K., Akiyama, M., Yanagi, T., McMillan, J. R., Suzuki, T., Tsukamoto, K., Sugiyama, H., Hatano, Y., Hayashitani, M., Takamori, K., Nakashima, K., and Shimizu, H. (2009). ABCA12 is a major causative gene for non-bullous congenital ichthyosiform erythroderma. *J. Invest. Dermatol.* 129, 2306–2309.

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.* 82, 290–303.

Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* 38, 209–213.

Smith, M. W., Patterson, N., Lautenberger, J. A., Truelove, A. L., McDonald, G. J., Waliszewska, A., Kessing, B. D., Malasky, M. J., Scafe, C., Le, E., De Jager, P. L., Mignault, A. A., Yi, Z., De The, G., Essex, M., Sankale, J. L., Moore, J. H., Poku, K., Phair, J. P., Goedert, J. J., Vlahov, D., Williams, S. M., Tishkoff, S. A., Winkler, C. A., and De La Vega, F. M. (2004). A high density admixture map for disease gene discovery in African Americans. *Am. J. Hum. Genet.* 74, 979–1000.

Stefflova, K., Dulik, M. C., Pai, A. A., Walker, A. H., Zeigler-Johnson, C. M., Gueye, S. M, Schurr, T. G., and Rebbeck, T. R. (2009). Evaluation of group genetic ancestry of populations from Philadelphia and Dakar in the context of sex-biased admixture in the Americas. *PLoS ONE* 4, e7842. doi: 10.1371/journal.pone.0007842

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.* 64, 479–498.

Teo, Y., Small, K., and Kwiatkowski, D. (2010). Methodological challenges of genome-wide association analysis in Africa. *Nat. Rev. Genet.* 11. 149–160.

Tiwari, H. K., Barnholtz-Sloan, J., Wineinger, N., Padilla, M. A., Vaughan, L. K., and Allison, D. B. (2008). Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Hum. Hered.* 66, 67–86.

Van Steen, K., McQueen, M., Herbert, A., Raby, B., Lyon, H., DeMeo, D. L., Murphy, L., Su, J., Datta, S., Rosenow, C., Christman, M., Silverman, E. K., Laird, N. M., Weiss, S. T., and Lange, C. (2005). Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.* 37, 683–691.

Wang, H., Thomas, D. C., Pe'er, I., and Stram, D. O. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genet. Epidemiol.* 30, 356–368.

Zeigler-Johnson, C., Friebel, T., Walker, A. H., Wang, Y., Spangler, E., Panossian, S., Patacsil, M., Aplenc, R., Wein, A. J., Malkowicz, S. B., and Rebbeck, T. R. (2004). CYP3A4, CYP3A5, and CYP3A43 genotypes and haplotypes in the etiology and severity of prostate cancer. *Cancer Res.* 64, 8461–8467.

Zheng, G., Song, K., and Elston, R. C. (2007). Adaptive two-stage analysis of genetic association in case-control designs. *Hum. Hered.* 63, 175–186.

Zhu, X., and Cooper, R. S. (2007). Admixture mapping provides evidence of association of the VNN1 gene with hypertension. *PLoS ONE* 2, e1244. doi: 10.1371/journal.pone.0001244

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 18 December 2010; paper pending published: 15 January 2011; accepted: 21 February 2011; published online: 10 March 2011.

Citation: Kang G, Gao G, Shete S, Redden DT, Chang B-L, Rebbeck TR, Barnholtz-Sloan JS, Pajewski NM, Allison DB (2011) Capitalizing on admixture in genome-wide association studies: a two-stage testing procedure and application to height in African-Americans. Front. Gene. ***2****:11. doi: 10.3389/fgene.2011.00011*

This article was submitted to Frontiers in Statistical Genetics and Methodology, a specialty of Frontiers in Genetics.

APPENDIX

PROOF OF ORTHOGONALITY ASYMPTOTIC INDEPENDENCE OF TWO TEST STATISTICS IN STAGE 1 AND STAGE 2 IN THE TWO-STAGE PROCEDURE FOR STRUCTURED ASSOCIATION TESTING

For ease of exposition and to facilitate generalization, we present this proof in the most general terms possible. Let X, Y, and Z be random variables with finite means and variances. Consider the regression with Y as the response variable and $E(X \mid Z)$ as the explanatory variable. Note that $E(X \mid Z)$ is a random variable that is a function of Z. Also, in this equation, all other measured covariates can be included.

$$Y = m + a_b E(X \mid Z)$$

Let $\hat{a}_b$ be an estimator of the regression coefficient in the above equation. Let T_1 be a statistic for testing the significance of this regression coefficient. Note that T_1 is obtained by dividing $\hat{a}_b$ by its estimated standard error. Importantly, note that this estimate of standard error is also a function of Y and Z (because it is obtained from the residuals of the above regression equation, which is function of only Y and Z). Therefore, the distribution of T_1 is a function of Y given Z.

Table A1 | The mean and variance of the percent of variability of a quantitative trait explained by admixture under the complete null hypothesis.

s[a]	Mean(s^2var(A)/var(Y))[b]	Var(s^2var(A)/var(Y))
0.03	6.972548e-05	2.960423e-11
0.1	7.741784e-04	3.644441e-09
0.3	6.924462e-03	2.879201e-07

[a] *a is the confounding effect of admixture on the trait.*
[b] *var(Y) = s^2 var(A) + t^2 var(G) + 1, where var is variance, Y is the quantitative trait, A is the ancestry estimate, and G is the genotype.*

Next, consider a multiple regression equation with Y as the response variable and $E(X \mid Z)$ and $(X - E(X \mid Z))$ as the explanatory variables:

$$Y_i = \beta_0 + \beta_1 E(X \mid Z) + \beta_2 (X - E(X \mid Z))$$

Then, β_2 measures the within-subpopulation correlation between Y and X, and therefore can be estimated by FBAT-type score statistics (Laird et al., 2000; Laird and Lange, 2006)

$$U = \sum (Y - \mu) \times (X - E(X \mid Z)),$$

where μ is the pre-specified user-defined offset parameter (Laird and Lange, 2006). Let us denote the test statistic obtained by dividing U by its estimated standard error (under the null hypothesis) by T_2. Note that the estimated standard error is a function of X conditional on the Y and Z (Lange et al., 2002). Thus, the standard errors are estimated on the basis of X conditional on Y and Z. Therefore, the test statistic T_2 is a random variable whose distribution is a function of X conditional on Y and Z. Our objective is to show that the statistics T_1 and T_2 are independent.

First, let us show that, if the null hypothesis that Y is independent of X conditional on Z is true, the test statistics are uncorrelated. (Independence of two random variables implies uncorrelatedness but the converse is not true.)

$$E(T_1 T_2) = \iint t_1 t_2 f_1(t_1) f_2(t_2) dt_1 dt_2,$$

where the first integral is over t_1 and the second integral is over t_2. Also, f_1 and f_2 are density functions of the random variables T_1 and T_2, respectively. However, one can calculate the above integral in terms of the original density functions of X, Y, and Z.

We know that T_1 is function of Y given Z. Let us denote $T_1 = \psi(Y \mid Z)$. Similarly, we know that T_2 is function of X given Y and Z. Let us denote $T_2 = \varphi(X \mid Y, Z)$.

Define a set

$$A = \{y \mid z : \psi(y \mid z) = t_1\} \text{ and } B = \{x \mid y, z : \varphi(x \mid y, z) = t_2\}.$$

Table A2 | The mean and variance of percent of variability of a quantitative trait explained by admixture under the non-complete null hypothesis.

		400[c]		800	
s[a]	t[b]	Mean(s^2var(A)/var(Y))[d]	Var(s^2var(A)/var(Y))	Mean(s^2var(A)/var(Y))	Var(s^2var(A)/var(Y))
0.03	0.2	8.617742e-05	3.650171e-11	6.879600e-05	2.994778e-11
	0.4	8.119870e-05	3.118215e-11	6.505619e-05	2.611041e-11
	0.6	7.406913e-05	2.465319e-11	5.965373e-05	2.126991e-11
0.1	0.2	9.566892e-04	4.490605e-09	7.637954e-04	3.669371e-09
	0.4	9.014642e-04	3.836954e-09	7.222993e-04	3.199798e-09
	0.6	8.223717e-04	3.034455e-09	6.623497e-04	2.607320e-09
0.3	0.2	8.544490e-03	3.527413e-07	6.832772e-03	2.914053e-07
	0.4	8.054818e-03	3.019340e-07	6.463735e-03	2.544517e-07
	0.6	7.352757e-03	2.393942e-07	5.930145e-03	2.077330e-07

[a] *s is the confounding effect of admixture on the trait.*
[b] *t is the effect of genotype at one disease marker on the trait.*
[c] *Sample size is 400 individuals.*
[d] *var(Y) = s^2 var(A) + t^2 var(G) + 1, where var is variance, Y is the quantitative trait, A is the ancestry estimate, and G is the genotype.*

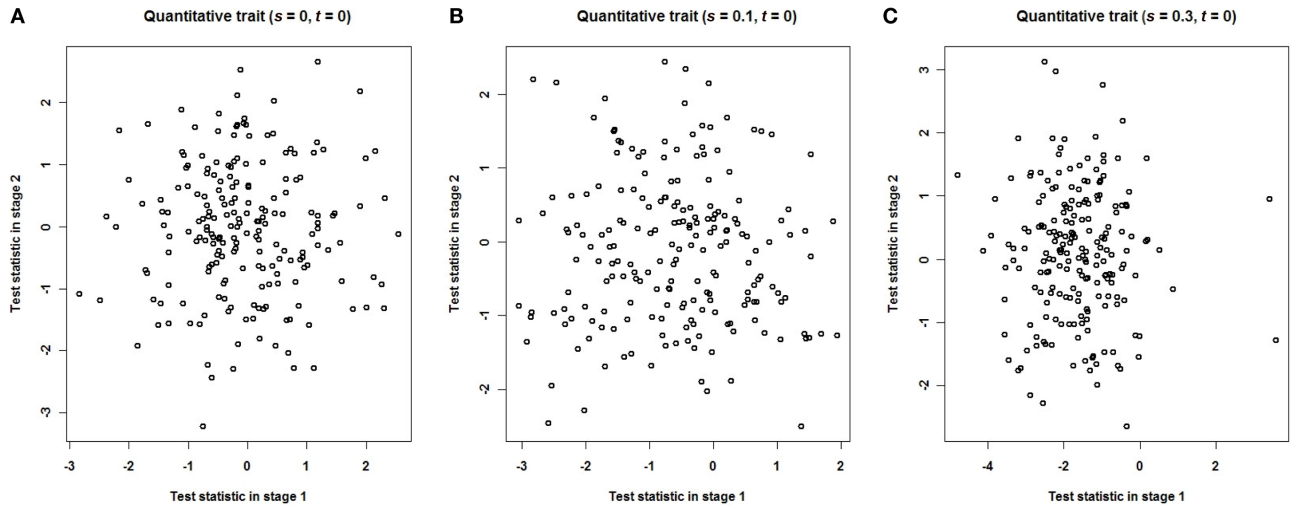

FIGURE A1 | Scatter plot of the test statistics in stage 2 versus the test statistics in stage 1 in the two-stage procedure under the null. (A) is for $s = 0$; **(B)** is for $s = 0.1$; **(C)** is for $s = 0.3$. $t = 0$.

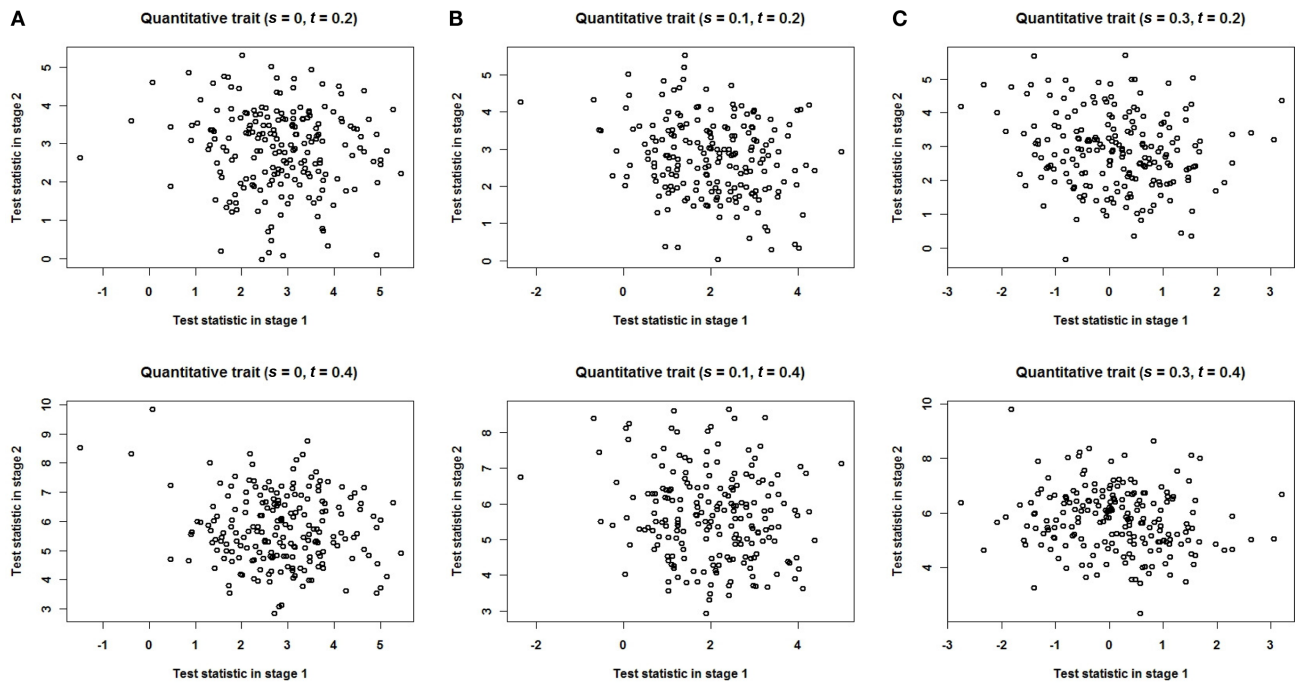

FIGURE A2 | Scatter plot of the test statistics in stage 2 versus the test statistics in stage 1 in the two-stage procedure under the alternative. (A) is for $s = 0$; **(B)** is for $s = 0.1$ and **(C)** is for $s = 0.3$. $t = 0.2$ and 0.4.

Then,

$$E(T_1T_2) = \iint t_1 t_2 f(y \mid z) g(x \mid y, z) dxdy.$$

Here the first integral is over set A and the second integral is over set B. Also f is the conditional density of Y given Z and g is the conditional density of X given Y and Z. We separate the above equation as

$$E(T_1T_2) = \int_A t_1 f(y \mid z) \left[\int_B t_2 g(x \mid y, z)\, dx \right] dy.$$

The second integral in the brackets is essentially $E(T_2)$. It is noteworthy that the $E(X \mid Z)$ in the numerator of T_2 is the expected value of X given Z under the null hypothesis. Also note that under null hypothesis, $E(T_2) = 0$ (the null hypothesis is that X is independent of Y conditional on Z).

The overall numerator is asymptotically normal with a mean of zero and the overall denominator converges to 1. One can then use Slutsky's theorem (Rao, 1973) to show the asymptotic normality of T_2 under the null hypothesis with a mean of zero and variance of 1. Therefore,

Table A3 | Information on 108 ancestry informative markers and the *p*-values of association with height by the two-stage procedure.

rs ID	Chromosome	Gene	Genetic position	Physical position	*p*-values in stage 1	*p*-value in stage 2
rs10202705	2	LOC646324	215.6077	216417394	0.0609	0.0951
rs952718[a]	**2**	**ABCA12**	**214.6798**	**215714130**	**0.0742**	**0.011**
rs1985080	**7**	**BBS9**	**52.12773**	**33400099**	**0.098**	**0.0135**
rs7021690	9	LOC645586	0.193122	534642	0.1055	0.2294
rs9849733	3	C3orf55	180.1571	158876963	0.122	0.4364
rs4350528	15	LOC728292	108.5775	91964704	0.174	0.2227
rs11901793	2	CXCR7	263.0074	237279237	0.1982	0.4396
rs12997060	2	FLJ39660	200.7928	197405233	0.2028	0.744
rs9416026	10	CBARA1	96.30602	74087507	0.2088	0.4445
rs11000419	10	CCDC109A	96.30699	74244696	0.2089	0.8467
rs1462309	3	LOC151760	133.062	112009941	0.2097	0.4675
rs13261248	8	HAS2	139.6438	122583352	0.2133	0.3082
rs12900262	15	LOC723972	29.64596	33272681	0.2177	0.4129
rs6023376	20	DOK5	90.68957	52629121	0.2567	0.8494
rs2426515	20	DOK5	90.49722	52506124	0.2612	0.1514
rs1911999	10	LOC728616	178.5082	132471324	0.2755	0.1609
rs503677	10	HERC4	87.70496	69497018	0.2833	0.2395
rs2816	17	GUCY2D	19.18685	7864289	0.3024	0.8106
rs2246695	14	LOC729637	58.51679	61077818	0.3051	0.0223
rs710052	14	FLJ22447	58.61844	61180428	0.3151	0.2077
rs4896780	6	LOC645749	164.4003	145559100	0.3188	0.3345
rs7187359	16	LOC730183	61.75076	30610656	0.3312	0.9713
rs12926237	16	LOC647086	61.75096	30745097	0.3312	0.4634
rs4811651	20	LOC728922	93.27854	54135335	0.3494	0.8077
rs4529792	10	hCG_2024596	81.85902	65612336	0.3962	0.569
rs7424137	2	COL3A1	194.4305	189709150	0.4053	0.0935
rs6937164	6	MOXD1	148.7214	132737005	0.4066	0.9089
rs4859147	3	DCUN1D1	210.388	184164555	0.4125	0.6253
rs2891	17	C17ORF85	9.398317	3652275	0.4196	0.6011
rs4792105	17	FLJ45455	30.27122	11052086	0.4274	0.6108
rs4489979	15	C15orf53	35.4689	36834731	0.4324	0.2248
rs4659762	1	MT1P2	273.1732	233493259	0.438	0.9298
rs6765491	3	C3orf58	166.8326	145319388	0.4425	0.3359
rs1733731	10	LOC399774	71.0642	53909652	0.4485	0.0848
rs1917028	5	ARHGAP26	157.1685	142106940	0.4486	0.3187
rs33957	5	FGF1	156.1912	141908017	0.4601	0.4253
rs4793237	17	ARL4D	85.4688	38792121	0.4654	0.7775
rs2593595	17	G6PC	84.88509	38309771	0.4719	0.6797
rs228768	17	HDAC5	86.18277	39547419	0.4784	0.348
rs4923940	15	GANC	39.91222	40372666	0.4937	0.8646
rs12594483	15	CDAN1	39.91419	40809278	0.4937	0.4814
rs735480	15	LOC653381	40.30385	42939663	0.4964	0.8786
rs155409	3	CNTN6	2.635718	1330266	0.5017	0.585
rs316598	5	LOC728878	5.460322	2417626	0.5049	0.9373
rs10041728	5	FTMT	136.5294	121164880	0.5065	0.8867
rs1011643	20	MACROD2	39.12778	15492065	0.5089	0.9319
rs645510	12	KSR2	144.2429	116569153	0.5143	0.1442
rs584059	3	LOC646641	160.1478	140313784	0.5177	0.9838
rs798443	2	C2orf46	17.5542	7918873	0.5191	0.8202
rs4885162	13	HCG_1820717	79.18408	73767349	0.549	0.4777
rs13173738	5	FLJ43080	126.6313	110015127	0.5525	0.2062
rs9543532	13	KLF12	79.02257	73599383	0.5551	0.5224
rs13318432	3	GADL1	54.50649	30848144	0.5574	0.2923
rs10056388	5	FLJ43080	126.0682	109533593	0.5586	0.5464
rs3861709	9	BNC2	34.77461	16693100	0.5747	0.8248

(Continued)

Table A3 | Continued

rs ID	Chromosome	Gene	Genetic position	Physical position	*p*-values in stage 1	*p*-value in stage 2
rs10962612	9	LOC648570	34.90021	16794167	0.5755	0.4173
rs1800498	11	DRD2	136.8711	112796798	0.5808	0.1784
rs1885167	9	C9orf39	35.55759	17504515	0.5936	0.6049
rs1982235	2	ATP5G3	186.285	175873413	0.5956	0.376
rs9530646	13	MYCBP2	82.60269	76871502	0.6132	0.9015
rs2814778	1	DARC	164.8221	155987756	0.6205	0.4206
rs9306906	4	LOC727792	55.75533	33788933	0.6277	0.7811
rs4789070	17	CD300A	131.3202	70006271	0.6372	0.9351
rs2184033	10	IPMK	75.94487	59493900	0.6399	0.1966
rs11607932	11	CCND1	88.57647	69059026	0.6467	0.4551
rs2687427	4	LOC133185	55.40398	33048432	0.6625	0.0284
rs1876482	2	LOC442008	38.4539	17284196	0.666	0.8582
rs2777804	9	ABCA1	130.4389	104650796	0.6929	0.4834
rs1412521	9	DBC1	159.8048	118992643	0.699	0.6384
rs1372115	2	ACVR1	169.3729	158503007	0.6991	0.9289
rs7041	4	GC	87.68387	72983369	0.7033	0.6588
rs1508061	2	EXOC6B	97.05812	72867396	0.7045	0.2826
rs17049450	2	LOC402102	145.4099	129901408	0.7169	0.6291
rs12640848	4	ENAM	85.96041	71871447	0.7298	0.0069
rs7134682	12	LOC645253	80.89459	64454418	0.7345	0.6023
rs12692701	2	FIGN	175.3797	164294693	0.7444	0.1651
rs6494466	15	CSNK1G1	61.74714	62295816	0.7454	0.6236
rs857440	6	LOC728961	31.91449	14814156	0.7505	0.456
rs7689609	4	LOC727995	86.35034	72448409	0.7546	0.0392
rs12612040	2	CALM2	73.80653	47363479	0.7645	0.9367
rs870272	9	C9orf18	162.2841	122033114	0.7711	0.4028
rs1858465	17	LOC339209	93.90577	48497919	0.7736	0.3317
rs4823460	22	FAM118A	63.38413	44040171	0.7749	0.801
rs722098	21	LOC388814	5.595659	15607469	0.8444	0.6704
rs4602918	8	CSMD1	7.077355	2610476	0.8447	0.7925
rs1490728	12	CAPZA3	36.7412	18926829	0.8761	0.7462
rs7189172	16	LOC440389	105.4772	78505139	0.8801	0.7853
rs11150219	16	LOC440389	105.4753	78404774	0.8802	0.2759
rs2416791	12	ETV6	25.75001	11592755	0.8869	0.9688
rs1991818	19	KLK7	93.55545	56176825	0.898	0.8252
rs1477921	13	LOC728192	116.9487	105816154	0.8992	0.7171
rs7161	1	DPH2	78.39628	44108067	0.9069	0.9732
rs13169284	5	CMBL	26.64584	10343037	0.9133	0.5776
rs1372894	4	LOC727891	192.5881	171959148	0.9174	0.7485
rs1862819	16	MPHOSPH6	112.0289	80783067	0.9223	0.6052
rs12129648	1	KIF26B	292.7902	241697533	0.9233	0.3829
rs10842753	12	ITPR2	49.58988	26588678	0.9316	0.7366
rs2077863	18	LOC645932	37.92992	11046815	0.9335	0.6782
rs6491743	13	LOC728183	109.1772	102859333	0.9419	0.9647
rs6003	1	F13B	211.9731	193762678	0.9516	0.992
rs10195705	2	CTNNA2	106.2942	80576097	0.9642	0.0683
rs1043809	17	EPN2	47.36621	19180025	0.9716	0.671
rs344454	7	CNTNAP2	167.2209	145839082	0.9799	0.5745
rs10908312	1	CSF3R	68.42725	36774370	0.983	0.71
rs1335826	10	LOC729432	47.59236	24084471	0.9838	0.6698
rs10255169	7	CNTNAP2	167.0514	145456168	0.9875	0.7345
rs2451563	6	LOC643281	97.19765	77170807	0.9911	0.6489
rs1257010	2	LOC643445	112.4947	97055688	0.9994	0.6953

[a]*We chose top three AIMs for stage two association analysis and the significant AIMs are in bold. Here, we chose three because based on our simulation results the optimal proportion of top markers selected in stage one seems equal to 0.03 so that* $0.03 \times 108 \approx 3$.

Table A4 | Empirical power and FWER of the two-stage procedure at a significance level of 0.05 (200 replicates).

t[a]	Method	*q*/*h*[b]	*s*[c] = 0				*s* = 0.1				*s* = 0.3			
			400[d]		800		400		800		400		800	
			Power[e]	FWER[f]	Power	FWER	Power	FWER	Power	FWER	Power	FWER	Power	FWER
0.2	OS[g]		0.005	0.031	0.115	0.02	0.01	0.03	0.09	0.055	0.005	0.045	**0.121**	0.04
	TS[h]	1000/1805[i]	0.01	0.046	0.14	0.035	0.01	0.045	0.095	0.06	**0.01**	0.045	0.07	0.04
		500/1805	0.015	0.051	0.185	0.045	0.015	0.04	0.095	0.06	0.005	0.055	0.055	0.035
		250/1805	0.015	0.061	0.18	0.06	0.015	0.03	0.11	0.04	**0.01**	0.04	0.05	0.03
		125/1805	0.02	0.036	0.2	0.05	0.02	0.05	0.12	0.035	**0.01**	0.045	0.02	0.05
		100/1805	0.031	0.036	0.2	0.05	**0.025**	0.045	**0.125**	0.03	0.005	0.04	0.015	0.06
		75/1805	0.026	0.036	0.215	0.035	0.02	0.045	0.12	0.045	0.005	0.035	0.005	0.045
		50/1805	0.026	0.041	0.25	0.03	**0.025**	0.04	0.11	0.035	**0.01**	0.04	0.005	0.065
		25/1805	**0.036**[j]	0.02	**0.27**	0.045	**0.025**	0.04	0.08	0.045	0	0.015	0.005	0.06
		5/1805	0.031	0.036	0.225	0.01	0.005	0.055	0.04	0.04	0	0.01	0.005	0.055
		2/1805	0.026	0.031	0.19	0.025	0	0.065	0.02	0.02	0	0.03	0.005	0.05
0.4	OS[g]		0.35	0.03	0.875	0.05	0.445	0.04	0.865	0.04	0.362	0.05	0.875	0.065
	TS[h]	1000/1805[i]	0.39	0.035	0.88	0.04	0.465	0.035	0.865	0.065	0.397	0.045	0.91	0.065
		500/1805	0.455	0.03	0.9	0.045	0.52	0.03	0.89	0.06	0.432	0.06	**0.93**	0.06
		250/1805	0.505	0.03	0.92	0.03	0.555	0.045	0.92	0.06	**0.447**	0.055	0.91	0.06
		125/1805	0.555	0.045	0.935	0.035	0.605	0.05	0.945	0.03	0.412	0.04	0.87	0.04
		100/1805	0.575	0.025	0.94	0.03	0.645	0.04	0.945	0.035	0.402	0.045	0.83	0.05
		75/1805	0.605	0.035	0.945	0.03	0.66	0.045	**0.96**	0.04	0.382	0.045	0.83	0.025
		50/1805	0.62	0.015	0.95	0.03	**0.68**	0.03	**0.96**	0.04	0.372	0.035	0.795	0.02
		25/1805	0.655	0.02	**0.975**	0.01	0.64	0.02	0.95	0.015	0.322	0.025	0.745	0.02
		5/1805	**0.67**	0.015	0.965	0	0.495	0.02	0.865	0.005	0.196	0.045	0.57	0.015
		2/1805	0.545	0.005	0.945	0	0.395	0.035	0.8	0	0.146	0.045	0.465	0.04
0.6	OS[g]		0.93	0.06	**1**	0.065	0.9	0.06	**1**	0.06	0.915	0.04	**1**	0.06
	TS[h]	1000/1805[i]	0.945	0.055	**1**	0.065	0.92	0.06	0.995	0.025	0.935	0.03	**1**	0.04
		500/1805	0.96	0.035	0.985	0.055	0.935	0.06	0.99	0.035	0.95	0.03	**1**	0.06
		250/1805	0.96	0.055	0.985	0.06	0.95	0.06	0.99	0.035	0.96	0.035	0.99	0.065
		125/1805	0.965	0.045	0.985	0.06	0.95	0.04	0.985	0.025	**0.97**	0.01	0.99	0.03
		100/1805	0.975	0.035	0.985	0.065	0.96	0.03	0.985	0.035	**0.97**	0.02	0.99	0.025
		75/1805	0.985	0.02	0.985	0.06	0.98	0.035	0.985	0.03	**0.97**	0.03	0.99	0.01
		50/1805	0.985	0.02	0.985	0.045	0.98	0.025	0.985	0.02	**0.97**	0.04	0.99	0.015
		25/1805	**0.995**	0.01	0.985	0.01	**0.985**	0.015	0.985	0.01	0.93	0.015	0.98	0.025
		5/1805	0.97	0.005	0.985	0.005	0.97	0.005	0.985	0	0.785	0.015	0.95	0
		2/1805	0.93	0	0.985	0	0.91	0	0.985	0.005	0.69	0.025	0.93	0

[a] *t is the genotypic effect of the disease marker on the quantitative trait.*
[b] *q is the number of SNPs selected in the first stage; h is the number of total SNPs.*
[c] *s is the effect of confounding association on the trait.*
[d] *400 individuals and 800 individuals.*
[e] *The power is estimated by the proportion of replicates successfully identifying the specific disease SNP.*
[f] *FWER, family wise error rate, which is estimated by the proportion of replicates wrongly identifying any one of the SNPs located at chromosomes 2 to chromosome 22.*
[g] *OS, one-stage procedure.*
[h] *TS, two-stage procedure.*
[i] *250/1805 = 0.139, 125/1805 = 0.069, 50/1805 = 0.028, 25/1805 = 0.014, 5/1805 = 0.003, 2/1805 = 0.001.*
[j] *The maximum power of both the OS and TS is marked in bold.*

$$E(T_1T_2) = \int_A t_1 f(y \mid s)[E(t_2)]dy = 0$$

and

$$Cov(T_1T_2) = E(T_1T_2) - E(T_1)\,E(T_2) = 0.$$

Thus, we have proven that these two statistics are uncorrelated.

We can then prove the asymptotic independence of T_1 and T_2 by noting that the uncorrelatedness (orthogonality) implies independence ***IF*** T_1 and T_2 are normally distributed. T_1 and T_2 are standard linear regression estimators and can be shown to be asymptotically normally distributed by using standard asymptotic arguments. Thus, the joint

Table A5 | Empirical power and FWER of the two-stage procedure at a significance level of 0.1 (200 replicates).

t[a]	Method	*q*/*h*[b]	*s*[c] = 0				*s* = 0.1				*s* = 0.3			
			400[d]		800		400		800		400		800	
			Power[e]	FWER[f]	Power	FWER	Power	FWER	Power	FWER	Power	FWER	Power	FWER
0.2	OS[g]		0.025	0.041	0.105	0.05	0.021	0.082	0.09	0.08	**0.015**	0.056	**0.1**	0.07
	TS[h]	1000/1805[i]	0.02	0.071	0.15	0.05	0.015	0.082	0.125	0.055	**0.015**	0.087	**0.1**	0.08
		500/1805	0.01	0.086	0.185	0.085	0.036	0.097	0.185	0.08	**0.015**	0.097	0.065	0.08
		250/1805	0.03	0.076	0.2	0.125	0.046	0.103	0.2	0.105	0.005	0.097	0.07	0.08
		125/1805	0.046	0.071	0.225	0.1	0.041	0.056	0.22	0.09	0	0.097	0.055	0.085
		100/1805	0.041	0.066	0.215	0.095	0.046	0.062	**0.225**	0.08	0	0.092	0.05	0.09
		75/1805	0.041	0.076	0.23	0.09	0.046	0.056	0.2	0.105	0	0.092	0.045	0.08
		50/1805	0.041	0.076	0.255	0.07	**0.056**	0.092	0.17	0.08	0	0.087	0.04	0.1
		25/1805	0.056	0.066	**0.26**	0.08	0.046	0.087	0.14	0.075	0	0.077	0.03	0.115
		5/1805	**0.066**[j]	0.086	0.215	0.07	0.031	0.072	0.11	0.045	0	0.056	0.01	0.1
		2/1805	0.051	0.107	0.195	0.06	0.015	0.062	0.06	0.09	0	0.087	0	0.105
0.4	OS[g]		0.44	0.11	0.925	0.07	0.48	0.09	0.89	0.105	0.431	0.113	0.894	0.086
	TS[h]	1000/1805[i]	0.47	0.1	0.92	0.06	0.525	0.07	0.9	0.08	0.441	0.123	**0.914**	0.081
		500/1805	0.54	0.105	0.935	0.09	0.575	0.08	0.94	0.075	0.492	0.108	0.909	0.126
		250/1805	0.595	0.085	0.95	0.075	0.66	0.11	0.96	0.05	**0.513**	0.103	0.859	0.101
		125/1805	0.645	0.035	0.96	0.065	0.67	0.075	0.97	0.06	0.477	0.087	0.828	0.086
		100/1805	0.67	0.04	0.96	0.09	0.68	0.075	0.97	0.06	0.462	0.087	0.823	0.086
		75/1805	0.715	0.03	0.965	0.08	**0.71**	0.09	**0.975**	0.04	0.431	0.077	0.808	0.081
		50/1805	0.74	0.04	0.965	0.05	0.7	0.06	0.97	0.04	0.374	0.067	0.768	0.061
		25/1805	**0.765**	0.035	**0.97**	0.03	0.69	0.06	**0.975**	0.06	0.333	0.056	0.727	0.071
		5/1805	0.735	0.03	0.96	0.01	0.575	0.045	0.935	0.01	0.185	0.072	0.53	0.051
		2/1805	0.675	0.02	0.925	0.005	0.46	0.075	0.87	0.005	0.138	0.067	0.46	0.051
0.6	OS[g]		0.944	0.101	**1**	0.12	0.949	0.066	**1**	0.135	0.95	0.1	**1**	0.121
	TS[h]	1000/1805[i]	0.955	0.076	**1**	0.12	0.96	0.061	0.995	0.135	0.95	0.085	**1**	0.121
		500/1805	0.955	0.071	0.99	0.115	0.975	0.066	0.99	0.12	0.955	0.07	**1**	0.095
		250/1805	0.97	0.076	0.985	0.11	**0.99**	0.076	0.985	0.115	**0.985**	0.065	**1**	0.07
		125/1805	0.99	0.071	0.985	0.075	**0.99**	0.051	0.985	0.075	0.975	0.065	0.99	0.095
		100/1805	0.99	0.066	0.985	0.085	**0.99**	0.061	0.985	0.045	0.975	0.065	0.99	0.09
		75/1805	0.995	0.066	0.985	0.095	**0.99**	0.061	0.985	0.05	0.97	0.06	0.99	0.08
		50/1805	**1**	0.066	0.985	0.06	**0.99**	0.04	0.985	0.05	0.945	0.085	0.985	0.065
		25/1805	**1**	0.02	0.985	0	**0.99**	0.02	0.985	0.025	0.925	0.045	0.985	0.01
		5/1805	**1**	0	0.985	0.005	0.965	0.01	0.985	0	0.785	0.025	0.965	0.005
		2/1805	0.97	0	0.975	0	0.919	0.005	0.98	0	0.72	0.01	0.945	0.01

[a] *t is the genotypic effect of the disease marker on the quantitative trait.*
[b] *q is the number of SNPs selected in the first stage; h is the number of total SNPs.*
[c] *s is the effect of confounding association on the trait.*
[d] *400 individuals and 800 individuals.*
[e] *The power is estimated by the proportion of replicates successfully identifying the specific disease SNP.*
[f] *FWER, family wise error rate, which is estimated by the proportion of replicates wrongly identifying any one of the SNPs located at chromosomes 2 to chromosome 22.*
[g] *OS, one-stage procedure.*
[h] *TS, two-stage procedure.*
[i] *250/1805 = 0.139, 125/1805 = 0.069, 50/1805 = 0.028, 25/1805 = 0.014, 5/1805 = 0.003, 2/1805 = 0.001.*
[j] *The maximum power of both the OS and TS is marked in bold.*

distribution of T_1 and T_2 is asymptotically normally distributed. Given this and the fact that we have shown that these two statistics are uncorrelated proves the asymptotic independence of T_1 and T_2.

Having demonstrated the asymptotic independence of T_1 and T_2, we can easily make the specification that Y is a phenotype, X is a genotype, and Z is a variable (e.g., an individual ancestry value or a region-specific admixture value) such that conditional on Z, there can be no confounding of the association between X and Y by admixture. And $E(X \mid Z)$ is the predicted genotype value denoted by $\hat{G}_{i,j}$. If we do so, we now have two tests that can be used in our two-stage procedure that has all of the desirable characteristics [(a) to (d)] that we listed in the introduction.

REFERENCES

Laird, N. M., Horvath, S., and Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genet. Epidemiol.* 19(Suppl. 1), S36–S42.

Lange, C., and DeMeo, D., and Laird, N. M. (2002). Power and design considerations for a general class of family-based association tests: quantitative traits. *Am. J. Hum. Genet.* 71, 1330–1341.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd Edn. New York, NY: John Wiley and Sons.