RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

A Powerful Approach to Test an Optimally Weighted Combination of Rare Variants in Admixed Populations

Xuexia Wang,[1] Shuanglin Zhang,[2] Yun Li,[3,4,5] Mingyao Li,[6] and Qiuying Sha[2*]

[1]Joseph J. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, United States of America; [2]Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America; [3]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, United States of America; [4]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, United States of America; [5]Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina, United States of America; [6]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Received 4 August 2014; Revised 9 January 2015; accepted revised manuscript 26 January 2015.
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21894

ABSTRACT: Population stratification has long been recognized as an issue in genetic association studies because unrecognized population stratification can lead to both false-positive and false-negative findings and can obscure true association signals if not appropriately corrected. This issue can be even worse in rare variant association analyses because rare variants often demonstrate stronger and potentially different patterns of stratification than common variants. To correct for population stratification in genetic association studies, we proposed a novel method to Test the effect of an Optimally Weighted combination of variants in Admixed populations (TOWA) in which the analytically derived optimal weights can be calculated from existing phenotype and genotype data. TOWA up weights rare variants and those variants that have strong associations with the phenotype. Additionally, it can adjust for the direction of the association, and allows for local ancestry difference among study subjects. Extensive simulations show that the type I error rate of TOWA is under control in the presence of population stratification and it is more powerful than existing methods. We have also applied TOWA to a real sequencing data. Our simulation studies as well as real data analysis results indicate that TOWA is a useful tool for rare variant association analyses in admixed populations.

Genet Epidemiol 00:1–12, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: genetic association study; rare variants; admixed population; population stratification

Introduction

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases. However, most identified variants explain only a small proportion of the heritability [Bansal et al., 2010; McCarthy et al., 2008; Schork et al., 2009]. There is growing recognition that many common diseases could be influenced by rare variants [Cohen et al., 2006; Ji et al., 2008; Manolio et al., 2009; Marini et al., 2008; Nejentsev et al., 2009; Zhu et al., 2010]. Advances in next-generation-sequencing technologies allow detecting causal variants by directly examining the effect of rare variants. Several recent studies have achieved promising findings [Dickson et al., 2010; Goldstein et al., 2013; Hershberger et al., 2010; Nejentsev et al., 2009; Zawistowski et al., 2010]. However, due to allelic heterogeneity and the extreme rarity of individual variants, well-established statistical methods for analysis of common variants may not be optimal for detecting rare variants [Li and Leal, 2008]. Moreover, population stratification, which has long been recognized as an issue in genetic association studies, can have even stronger impact in rare variant association studies because rare variants often demonstrate stronger and potentially different patterns of stratification than common variants [Mathieson et al., 2012]. The aforementioned issues prompt an urgent need to develop powerful statistical methods for rare variant association studies, especially, methods that can be applied in admixed populations.

Recently, many statistical methods have been developed for the analyses of rare variants. Most of these methods can be classified into three categories: (1) burden tests [Li and Leal, 2008; Madsen and Browning, 2009; Price et al., 2010; Zawistowski et al., 2010], which collapse rare variants in a genomic region as a single burden variable and then test the cumulative effect of rare variants via regression on the derived burden variable; (2) quadratic tests [Neale et al., 2011; Sha et al., 2012; Wu et al., 2011], which involve the use of a quadratic form of the score vector in test statistics; and (3) combined tests [Derkach et al., 2012; Han and Pan, 2010; Hoffmann et al., 2010; Ionita-Laza et al., 2013; Lee et al., 2012; Li et al., 2010; Lin and Tang, 2011; Sha et al., 2013; Yi and Zhi, 2011] that combine evidence of association from burden tests, quadratic tests, and other possible tests to maximize the use of existing information.

*Correspondence to: Qiuying Sha, Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931, USA. E-mail: qsha@mtu.edu

Burden tests implicitly assume that all rare variants are causal and the directions of the effects are the same. Since such tests cannot differentiate rare variants that associate with the disease in different directions, they may lose power when both risk and protective variants are present. Quadratic tests circumvent this problem by using a quadratic term in the test statistic when modeling complex relationships between rare variants and the phenotype. Burden tests can only outperform quadratic tests when most of the rare variants are causal and the directions of the effects of causal variants are the same. The combined tests attempt to retain the advantages of both burden tests and quadratic tests. They are robust in the presence of opposite effect directions of the causal variants, and are less affected by neutral variants when compared to burden tests.

Although the aforementioned methods have shown promise in detecting rare variants, none of them explicitly model population stratification, which has long been recognized as an issue in genetic association studies. Population stratification emerges when there is a systematic difference in allele frequencies among study subjects due to ancestry difference across individuals. Unrecognized population stratification can lead to both false-positive and false-negative findings and can obscure true association signals if not appropriately corrected [Knowler et al., 1988; Lander et al., 1994; Mao et al., 2013]. For rare variants this problem can be more serious because the spectrum of rare variation can be different in diverse populations. It has been shown that rare variants can demonstrate stronger and potentially different patterns of stratification than common variants, and existing methods cannot effectively correct population stratification in rare variant analyses [Mathieson et al., 2012]. Additionally, empirical findings from the 1000 Genomes Project have shown that the numbers of rare variants can differ significantly among different populations [1000 Genomes Project Consortium, 2010]. These empirical findings have important implications for the analyses of admixed population, such as African Americans and Hispanic Americans, who are recently admixed and have inherited ancestry from more than one continent.

Population stratification could emerge due to two types of mechanisms: (1) stratification due to local ancestry difference driven by natural selection at certain genomic regions; (2) stratification due to global ancestry difference driven by the demographic history of a population or genetic random drift due to finite population size [Wang et al., 2011]. Commonly used methods, such as EIGENSTRAT [Price et al., 2006], genomic control [Devlin and Roeder, 1999; Reich and Goldstein, 2001], may fail to correct population stratification in rare variant association studies [Mathieson et al., 2012] because (1) these methods use a set of genomic markers across the entire genome to capture the global population structure, whereas subtle differences in local ancestry might be diluted due to the inclusion of markers from other genomic regions [Qin et al., 2010]; (2) rare variants can show systematically different patterns of stratification from common variants. Wang et al. [2011] empirically and theoretically demonstrated that regardless of the mechanism of population stratification, whether due to local or global ancestry differences, it is sufficient to adjust for local ancestry at the test region.

Although controlling for population stratification has become a routine in common variant association studies [Epstein et al., 2007; Li et al., 2010], currently, there is only one publication on controlling for population stratification in rare variant association analyses [Mao et al., 2013]. The local ancestry-based weighted dosage score (AWDS) test developed by Mao et al. [2013] takes into account local ancestry of rare alleles, uncertainties in rare variant imputation when imputed data are included, and the direction of the effect of rare variants on phenotypic trait. A shortcoming of AWDS is that it needs to split data into training set and testing set: (1) training set (30% data as training set) is used to decide risk, protective, or neutral variants; (2) testing set (70% data as testing set) is used to test for association. Exclusion of the training data when testing association of rare variants could lead to loss of power. In addition, AWDS tests association in each ancestral population separately. This approach is less powerful than methods that detect causal rare variants in all ancestral populations simultaneously.

To correct for population stratification induced by local ancestry difference, we proposed a novel method to Test the effect of an Optimally Weighted combination of variants in Admixed populations (TOWA). The analytically derived optimal weights can be calculated from existing phenotype and genotype data. TOWA up weights rare variants and those that have strong associations with the phenotype [Sha et al., 2012]. Additionally, it can adjust for the direction of the association, and allows for local ancestry difference among study subjects. Based on the optimal weights and the incorporation of the information from both local ancestry and rare alleles, TOWA is able to test the effect of an optimally weighted combination of rare variants in a region (i.e., gene or pathway) and properly control population stratification. To evaluate the performance of the proposed method, we conducted extensive simulation studies using whole-genome sequence data as well as an analysis in a real sequencing data. We compared the power of the proposed method with AWDS and the weighted dosage score (WDS) test proposed in WHaT [Li et al., 2010]. Our results indicate that the type I error rates of TOWA are under control in the presence of population stratification and it is more powerful than AWDS and WDS for most scenarios we considered.

Methods

Notation

Consider a sample of n individuals from an admixed population. We assume admixture has occurred between two ancestral populations, and each individual is genotyped at M SNPs in a genomic region (a gene or a pathway). Denote y_i (1 for case and 0 for control) as the trait value and D_i as the observed genotype of the ith individual. We define

$$\begin{cases} x_{ijk} = 1, & \text{if the } k\text{th allele of the } i\text{th individual at the } j\text{th} \\ & \text{SNP is the minor allele} \\ x_{ijk} = 0, & \text{otherwise} \end{cases}$$

$$\begin{cases} a_{ijk} = 1, & \text{if the } k\text{th allele of the } i\text{th individual at the } j\text{th} \\ & \text{SNP is from the first ancestral population} \\ a_{ijk} = 0, & \text{if the } k\text{th allele of the } i\text{th individual at the } j\text{th} \\ & \text{SNP is from the second ancestral population,} \end{cases}$$

and $b_{ijk} = 1 - a_{ijk}$, where $k = 1, 2$ in x_{ijk}, a_{ijk}, and b_{ijk}. Let $A_i = (A_{i1}, A_{i2}, \ldots, A_{iM})$ denote the ancestral status of the ith individual, where $A_{ij} = 0, 1, 2, 3$ represents $(a_{ij1}, a_{ij2}) = (0, 0), (0, 1), (1, 0), (1, 1)$ correspondingly. Let p_{ijk}, $k = 1, 2$, denote the probability of the kth allele being from the first ancestral population. This probability can be estimated using HAPMIX [Price et al., 2009], which has an option (DIPLOID for HAPMIX_MODE) that gives 16 probabilities of all 4×4 values of ancestry and genotype for each individual, from which one can infer the corresponding local ancestry for each allele by calculating the corresponding conditional probabilities.

Model for a Homogenous Population

For a homogenous population, we can test the effect of a weighted combination of M variants using logistic regression

$$\text{logit}(\Pr(y_i = 1)) = \alpha + \beta \sum_{j=1}^{M} w_j X_{ij},$$

where X_{ij} is the genotype of the ith individual at the jth SNP, and $w_1, \ldots, w_M$ are the weight functions.

Model for an Admixed Population

To use the logistic regression framework for an admixed population, we need to account for the population ancestry of each allele. When the ancestral status of each allele is known, we can modify the logistic regression model as

$$\begin{aligned} \text{logit}(\Pr(y_i = 1)) = \alpha + \beta_1 \sum_{j=1}^{M} (x_{ij1} a_{ij1} + x_{ij2} a_{ij2}) w_{1j} \\ + \beta_2 \sum_{j=1}^{M} (x_{ij1} b_{ij1} + x_{ij2} b_{ij2}) w_{2j}, \quad (1) \end{aligned}$$

where the weight functions $w_{11}, \ldots, w_{1M}$ and $w_{21}, \ldots, w_{2M}$ will be determined later according to some optimal criteria.

Testing Genetic Association in Both Ancestral Populations Simultaneously

We are interested in testing $H_0 : \beta_1 = \beta_2 = 0$. Significant results indicate that the test region is associated with the disease in at least one of the two ancestral populations.

When the ancestral status is unknown, we need to account for the uncertainty of the ancestral status. In this situation, the likelihood function contributed by the ith individual can be written as

$$\begin{aligned} p_i = \Pr(y_i = 1 | D_i) &= \sum_{A_i} \Pr(y_i = 1 | D_i, A_i) \Pr(A_i | D_i) \\ &= \sum_{A_{i1}=0}^{3} \cdots \sum_{A_{iM}=0}^{3} \frac{\exp(\delta_i)}{1 + \exp(\delta_i)} \Pr(A_i | D_i), \end{aligned}$$

where $\delta_i = \alpha + \beta_1 \sum_{j=1}^{M} (x_{ij1} a_{ij1} + x_{ij2} a_{ij2}) w_{1j} + \beta_2 \sum_{j=1}^{M} (x_{ij1} b_{ij1} + x_{ij2} b_{ij2}) w_{2j}$. The log-likelihood is given by

$$\log L = \sum_{i=1}^{n} (y_i \log p_i + (1 - y_i) \log(1 - p_i)).$$

In Appendix, we have shown that the score statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ is given by

$$\begin{aligned} T_{\text{score}} = (w_1^T u_1, w_2^T u_2) &\begin{pmatrix} w_1^T A w_1 & w_1^T C w_2 \\ w_1^T C w_2 & w_2^T B w_2 \end{pmatrix}^{-1} \\ &\times (w_1^T u_1, w_2^T u_2)^T / \hat{\sigma}^2, \end{aligned}$$

where $u_1 = \sum_{i=1}^{n} (y_i - \bar{y}) x_i$, $u_2 = \sum_{i=1}^{n} (y_i - \bar{y}) z_i$, $A = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$, $B = \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})^T$, $C = \sum_{i=1}^{n} (x_i - \bar{x})(z_i - \bar{z})^T$, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$, $x_i = (x_{i1}, \ldots, x_{iM})^T$, $x_{ij} = x_{ij1} p_{ij1} + x_{ij2} p_{ij2}$, $z_i = (z_{i1}, \ldots, z_{iM})^T$, and $z_{ij} = x_{ij1}(1 - p_{ij1}) + x_{ij2}(1 - p_{ij2})$.

One potential problem with the score statistic T_{score} is that for rare variants, it could be problematic to use A, B, and C to estimate the variance-covariance matrix of the genotypes of the first ancestral population, the variance-covariance matrix of genotypes of the second ancestral population, and the covariance matrix between genotypes of the first and second ancestral populations due to the sparsity of the rare variants. Following Pan [2009] and Sha et al. [2012], we replace A by $A_0 = \text{diag}(A)$, B by $B_0 = \text{diag}(B)$, and C by $C_0 = 0$. Then, the score test statistic becomes

$$T_0(w_1, w_2) = \sigma^{-2} \left(\frac{w_1^T u_1 u_1^T w_1}{w_1^T A_0 w_1} + \frac{w_2^T u_2 u_2^T w_2}{w_2^T B_0 w_2} \right).$$

The statistic $T_0(w_1, w_2)$ reaches its maximum $\sigma^{-2}(u_1^T A_0^{-1} u_1 + u_2^T B_0{}^{-1} u_2)$ when $(w_1, w_2) = (A_0^{-1} u_1, B_0{}^{-1} u_2)$, suggesting the weight is optimal. We define the statistic TOWA as

$$T_{TOWA} = u_1^T A_0^{-1} u_1 + u_2^T B_0^{-1} u_2 = \sum_{j=1}^{M} \left(\frac{u_{1j}^2}{v_{1j}} + \frac{u_{2j}^2}{v_{2j}} \right),$$

where $u_{1j} = \sum_{i=1}^{n} (y_i - \bar{y}) x_{ij}$, $u_{2j} = \sum_{i=1}^{n} (y_i - \bar{y}) z_{ij}$, $v_{1j} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$, and $v_{2j} = \sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2$.

We use a permutation test to evaluate the significance of T_{TOWA}. Let T_{TOWA}^0 denote the observed value of the test statistic T_{TOWA} based on the original data. In each permutation, we randomly shuffle $y_1, \ldots, y_n$ and denote the value of the corresponding test statistic by T_{TOWA}^{per}. We perform the permutation procedure many times. The P-value of the test can be calculated as the proportion of the number of permutations with $T_{TOWA}^{per} \geq T_{TOWA}^0$.

Table 1. Tests compared in this article

TOWA	Testing the effect of the Optimally Weighted combination of rare variants in Admixed populations (proposed in this article) which is able to test the combined effects of two ancestral populations. It is robust to the directions of the effects of causal variants and can control for population stratification.
TOWA$_{AFR}$	Testing the effect of the Optimally Weighted combination of rare variants in AFR ancestral population (proposed in this article) which is able to test the association in the AFR ancestral population. It is robust to the directions of the effects of causal variants and can control for population stratification.
TOWA$_{EUR}$	Testing the effect of the Optimally Weighted combination of rare variants in EUR ancestral population (proposed in this article) which is able to test the association in the EUR ancestral population. It is robust to the directions of the effects of causal variants and can control for population stratification.
WDS	The weighted dosage score test (proposed by Li et al. [2010]).
AWDS$_{AFR}$	The ancestry-based weighted dosage test to test the association in the AFR ancestral population (proposed by Mao et al. [2013]).
AWDS$_{EUR}$	The ancestry-based weighted dosage test to test the association in the EUR ancestral population (proposed by Mao et al. [2013]).

Testing Genetic Association in Each Ancestral Population

In addition to testing the combined effects of the two ancestral populations, we can also test association in each ancestral population. To test the association in the first ancestral population $H_0 : \beta_1 = 0$, we change Equation (1) to

$$\text{logit}(\Pr(y_i = 1)) = \alpha + \beta_1 \sum_{j=1}^{M} (x_{ij1} a_{ij1} + x_{ij2} a_{ij2}) w_{1j},$$

which leads to the test statistic

$$T^{(1)}_{TOWA} = \sum_{j=1}^{M} \left(\frac{{u_{1j}}^2}{v_{1j}} \right).$$

A significant result from this test statistic indicates that the test region is associated with the disease in the first ancestral population. Association analysis in the second ancestral population can be performed in a similar fashion.

The R code of the TOWA method is available at Shuanglin Zhang's homepage http://www.math.mtu.edu/˜shuzhang/software.html.

Comparisons of Methods

AWDS proposed by Mao et al. [2013] is the only method thus far dealing with population stratification for rare variants in admixed populations. AWDS is similar to the WDS test in WHaT [Li et al., 2010]. Both of these methods take into account the direction of association. The key difference is that AWDS is able to adjust for local ancestry in the tested region, and thus can control for population stratification. In order to evaluate the performance of TOWA, we compare the type I error rate and power of TOWA with AWDS and WDS. Table 1 summarizes the tests that are compared in this article.

Simulation

Simulation of Admixed Haplotypes

To evaluate the performance of the proposed method, we conducted extensive simulations following the simulation setup of Mao et al. [2013]. The 1000 Genomes data (http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-2010-08.html) provide a good resource to form sequenced admixed samples with African and Caucasian ancestry. The downloaded dataset includes 348 AFR (78YRI + 67LWK + 24ASW + 5PUR) phased haplotypes and 566 EUR (90 CEU + 92TSI + 43GBR + 36 FIN + 17MXL + 5PUR) phased haplotypes. The number of overlapping SNPs between AFR and EUR is 8,952,982. We simulated 20,000 haplotypes of admixed individuals with African and Caucasian ancestry using the phased haplotypes of AFR and EUR. We employed a two-stage approach proposed by Price at al. [2009]. At the first stage, we determined the ancestry state for each marker from chromosomes 1 to 22. For admixed populations, a person's genome is a mosaic of ancestral chromosomes. Therefore, a person's genome can be partitioned into different ancestry blocks. The breakpoints between the ancestry blocks were determined by the recombination events. We assumed that the probability $1 - e^{-\lambda d}$ of the recombination events follows a Bernoulli distribution, where d denotes the genetic distance (in Morgan) between adjacent SNPs and λ represents the number of generations since admixture. We set $\lambda = 6$ in our simulation. The ancestry of each block was assigned as AFR ancestry with probability p or EUR with probability $1 - p$, where p follows a beta distribution with mean 0.8 and standard deviation 0.1. This allows us to generate an admixed population with an average of 20% Caucasian and 80% African ancestries that is similar to the population in African Americans [Smith et al., 2004]. At the second stage, we generated 20,000 admixed haplotypes based on the ancestry state of each marker across chromosomes 1–22.

Since we have phased haplotype pool from AFR and EUR and we know the ancestry state for each block at the first stage, we can assign a haplotype for each ancestry block following the rule below. For a given ancestry block, if its ancestry state was AFR, we sampled a haplotype from the haplotype pool of AFR and assigned the sequence in that block to the admixed haplotype. We used similar procedure for an ancestry block of EUR ancestry. We repeated this procedure many times and generated a pool of 20,000 admixed haplotypes for each of the 22 autosomal chromosomes. Then, we randomly selected two haplotypes and assigned them to an individual and generated 10,000 individuals' two haplotypes.

Assignment of Disease Status

We considered a sample of n individuals from the simulated admixed population and determined the disease status for each individual following the simulation setup in Mao et al. [2013]. First, we partitioned the genome into 44,620 nonoverlapping segments with 200 SNPs in each segments.

In a given segment among SNPs with minor allele frequencies in the range (0.001, 0.05), we randomly selected c SNPs as causal variants. We followed Mao et al. [2013] by defining the genotype relative risk (GRR) of an individual at variant j of population a descent as

$$GRR_j^a = \left[\frac{PAR_j^a}{(1-PAR_j^a)f_j^a}\right]^{(-1)I\{\xi_j^a=1\}},$$

where PAR represents the population attributable risk; f denotes the adjusted minor allele frequency at variant j of population a ancestry; $\xi_j^a = 1/10$ indicates that the rare allele at SNP j decreases/increases disease risk in population a. We set 1 for GRR in noncausal SNPs. GRR can differentiate the risk of rare variants based on their ancestry state, which means that two individuals carrying the same allele may have different disease risk if they have different ancestral populations. We assigned an individual's disease status according to the equation below

$$\begin{aligned} &P(Z=1|x_{ijk}, a_{ijk}) \\ &= b_0 \times \prod_{j=1}^{200}\prod_{k=1}^{2}\left(GRR_j^{AFR}\right)^{I\{x_{ijk}^{AFR}=1, a_{ijk}^{AFR}=1\}} \\ &\quad \times \left(GRR_j^{EUR}\right)^{I\{x_{ijk}^{EUR}=1, a_{ijk}^{EUR}=1\}}, \end{aligned}$$

where $k = 1, 2$. x_{ijk} denotes if the allele k for individual i at marker j in population AFR (or EUR) is minor allele; a_{ijk} indicates if the allele k for individual i at marker j is from AFR (or EUR). b_0 is the baseline penetrance and is assigned as 10%. We repeated the sampling procedure until the desired number of cases and controls was obtained.

To evaluate the type I error rate of each method, we randomly selected 2,000 individuals from the simulation pool, and designated 1,000 as cases and other 1,000 as controls. This ensures that there is no systematic difference with regard to their global ancestry but there may exist local ancestry difference. For power comparisons, we considered two scenarios: (1) risk variants only, in which all variants increase the disease risk, and (2) mixture of risk and protective variants. In the two scenarios, we assumed: (1) the risk of causal variants in AFR and EUR is different; (2) PARs in AFR and in EUR for causal variants is the same. We further considered two models for the risk variants only case: (1) one-sided disease model in which the causal variants were from only one ancestral population; (2) two-sided disease model in which causal variants were from both ancestral populations. For the one-sided disease model, we selected 20 causal variants from AFR and EUR, respectively. We considered four different values (0.002, 0.003, 0.004, and 0.005) for PAR in AFR and four different values (0.008, 0.012, 0.016, and 0.02) for PAR in EUR. For the two-sided disease model, we selected 40 causal variants in total among which 20 causal variants were from AFR and another 20 causal variants were from EUR. We considered four different scenarios for the two-sided disease model ($PAR_{AFR} = 0.003$ and $PAR_{EUR} = 0.004$; $PAR_{AFR} = 0.003$ and $PAR_{EUR} = 0.008$; $PAR_{AFR} = 0.001$ and $PAR_{EUR} = 0.012$; $PAR_{AFR} = 0.002$ and $PAR_{EUR} = 0.012$). For the mixture of risk and protective variants case, we assumed that causal variants of one ancestral origin contributed to disease risk. We selected 20 causal variants, among which we considered the number of protective variants as 0, 5, 10, and 20, respectively. In each simulation scenario, P-values were estimated by 10,000 permutations and powers were evaluated using 1,000 replicated samples in a prespecified segment at a significance level of 0.05.

Table 2. Type I error rates of TOWA, TOWA$_{AFR}$, and TOWA$_{EUR}$

	Type I error		
Category	TOWA	TOWA$_{AFR}$	TOWA$_{EUR}$
I (0.001–0.003)	0.034	0.046	0.032
II (0.01–0.02)	0.042	0.050	0.042
III (0.03–0.05)	0.056	0.054	0.056

Significance level is assessed at 5%. The three categories indicate the mean local ancestry difference between cases and controls.

Results

Comparison of Type I Error Rates

In order to assess the impact of local ancestry difference on type I errors of the proposed method, we estimated the type I error rates by considering three different scenarios. We randomly selected 2,000 individuals from the simulation pool which was generated as aforementioned, and designated 1,000 as cases and other 1,000 as controls. The genome of a randomly selected individual was partitioned into 44,620 segments of 200 SNPs. The local ancestry of each segment is the average of ancestry proportions across all SNPs within the segment. The local ancestry differences of the 44,620 200-SNP segments based on one simulation run range from 0 to 0.0427. In order to evaluate the performance of the proposed method in the situation when small, medium, and large local ancestry differences exist, we classified the mean local ancestry difference into three categories: I (0.001–0.005), II (0.01–0.02), and III (0.03–0.05). Five hundred segments were retained from each category. We repeated the simulation procedure 20 times and obtained a total of 10,000 segments for each of the three categories. In each category, we analyzed 10,000 segments using TOWA, TOWA$_{AFR}$, and TOWA$_{EUR}$. P-values were estimated by 10,000 permutations and type I error rates in each category were evaluated using 10,000 P-values for the 10,000 segments at a significance level of 0.05 for each category. The type I error rates of each test are shown in Table 2. Our results indicate that the type I error rates for all the tests are under control.

Comparison of Power Under the Risk-Variant Only and One-Sided Disease Model

Figure 1 shows that TOWA$_{AFR}$ is the most powerful test if all causal variants were from AFR and TOWA$_{EUR}$ is the most powerful test if all causal variants were from EUR. AWDS$_{AFR}$ is less powerful than TOWA and is much less powerful than TOWA$_{AFR}$ when all causal variants were from

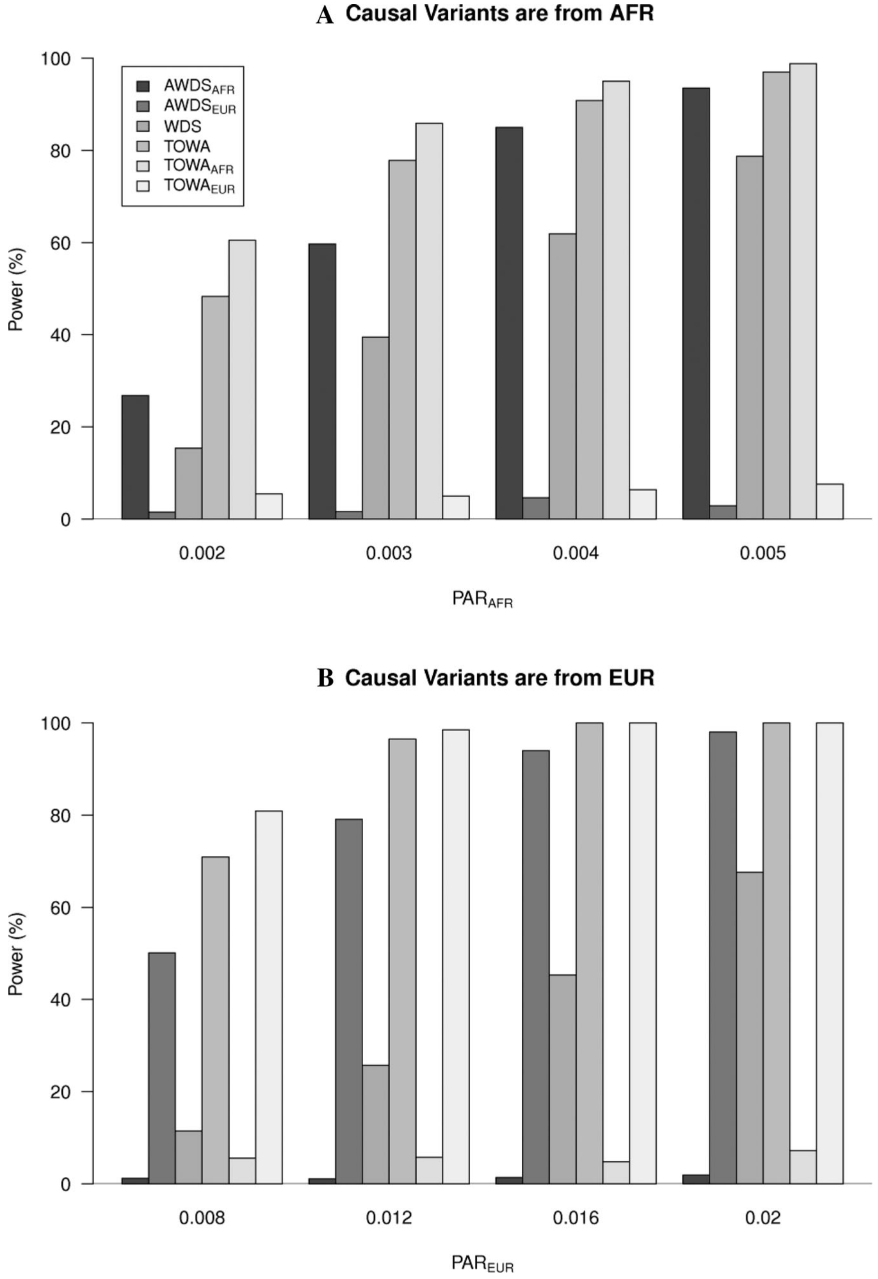

Figure 1. Power comparison under the risk-variant only and one-sided disease models. (A) Power of the six tests under risk-variant only models and all causal variants were from AFR. (B) Power of the six tests under risk-variant only models and all causal variants were from EUR. Significance was assessed at the 5% level.

AFR. AWDS_{EUR} is less powerful than TOWA and is much less powerful than TOWA_{EUR} when all causal variants were from EUR. The power of TOWA_{AFR} is more than twice the power of AWDS_{AFR} when all causal variants were from AFR and PAR was small. The power loss of AWDS is due to AWDS splitting data into two parts: training data and testing data. AWDS used 30% data as training set to classify variants as risk, protective, or neutral variants, and only 70% of the data were used in the association test, which reduces the power of AWDS substantially. Under the risk-variant only and one-sided disease model, TOWA_{AFR} (or TOWA_{EUR}) is based on a more accurate null hypothesis than TOWA. Therefore, TOWA_{AFR} (or TOWA_{EUR}) is more powerful than TOWA when all causal variants were from AFR (or EUR). In the range of PARs, we considered, for each ancestry population (AFR or EUR), the corresponding TOWA_{AFR}- (or TOWA_{EUR}) and AWDS-type tests consistently have higher power than the WDS test though the latter has inflated type I error rate in admixed populations. This pattern is expected because WDS is designed for the collective effect of causal variants ignoring their ancestry while TOWA and AWDS can test rare variant associations in each ancestral population of an admixed population.

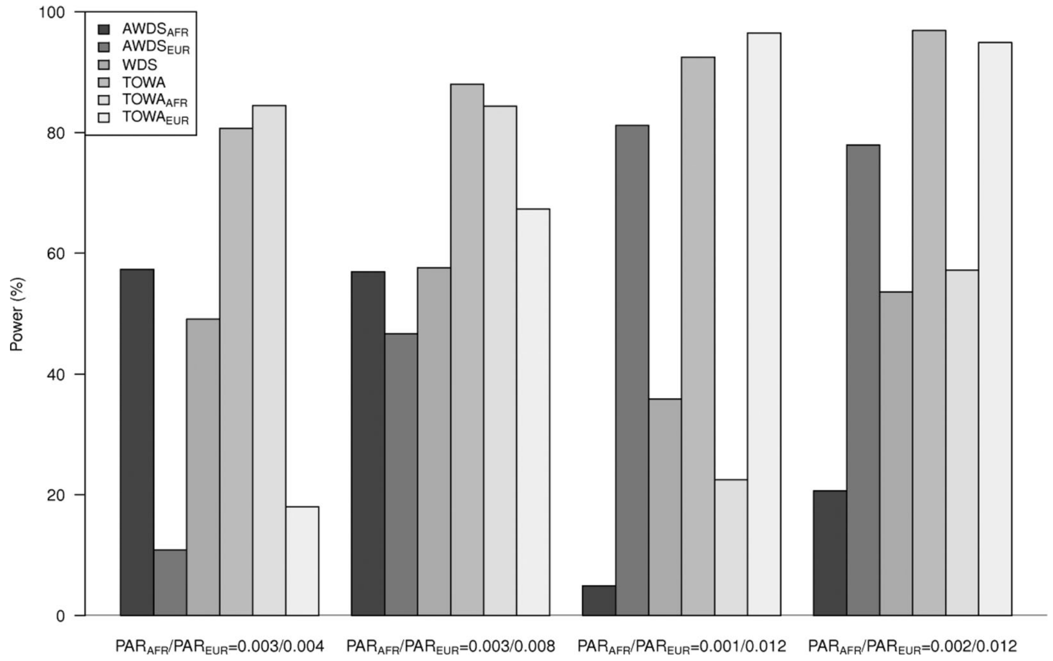

Figure 2. Power comparison under the risk-variant only and two-sided disease models. Power of the six tests under risk-variant only models considering 40 causal variants with 20 causal variants from AFR and 20 causal variants from EUR. Significance was assessed at the 5% level.

Comparison of Power Under the Risk-Variant Only and Two-Sided Disease Model

We considered four different scenarios under the risk-variant only model assuming a total of 40 causal variants with 20 causal variants from AFR and 20 from EUR, respectively. For simplicity, we chose moderate PARs: PAR$_{AFR}$ = 0.003 and PAR$_{EUR}$ = 0.004; PAR$_{AFR}$ = 0.003 and PAR$_{EUR}$ = 0.008; PAR$_{AFR}$ = 0.001 and PAR$_{EUR}$ = 0.012; and PAR$_{AFR}$ = 0.002 and PAR$_{EUR}$ = 0.012. Figure 2 shows that TOWA is about 30% more powerful than AWDS$_{AFR}$ when PAR$_{AFR}$ = 0.003 and PAR$_{EUR}$ = 0.004, PAR$_{AFR}$ = 0.003 and PAR$_{EUR}$ = 0.008. When PAR$_{AFR}$ = 0.002 and PAR$_{EUR}$ = 0.012, both TOWA and TOWA$_{EUR}$ are about 10% more powerful than AWDS$_{EUR}$ and 70% more powerful than AWDS$_{AFR}$. When PAR$_{AFR}$ = 0.001 and PAR$_{EUR}$ = 0.012, TOWA$_{EUR}$ is 10% more powerful than AWDS$_{EUR}$ and 80% more powerful than AWDS$_{AFR}$. AWDS-type tests (AWDS$_{AFR}$/AWDS$_{EUR}$) lost power dramatically in detecting association when there is one population ancestry AFR or EUR with less contribution to disease risk because AWDS-type tests can only test the effect of one single ancestral population. TOWA is able to test the combined effects of two ancestral populations. In addition, TOWA can use all the information in the whole sample in testing for association. However, AWDS-type tests only use 70% of the information from the dataset in testing for association due to the splitting procedure of their method. In the two-sided disease model, WDS is still less powerful than TOWA$_{AFR}$ and AWDS$_{AFR}$ when the side of AFR had larger contribution to disease risk (PAR$_{AFR}$ = 0.003). This pattern is expected because WDS is designed for the collective effect of causal variants ignoring their ancestry while TOWA$_{AFR}$ and AWDS$_{AFR}$ specifically test for associations in the AFR ancestral population.

Comparison of Power When Both Risk and Protective Variants Are Present

In order to clearly demonstrate the difference of the powers among different tests, we chose moderate PARs for AFR as 0.003, 0.004, and 0.005; for EUR as 0.012, 0.016, and 0.02 under one-sided disease model. We fixed the number of causal variants at 20 and chose the number of protective variants as 0, 5, 10, and 20, respectively. We set the same three PAR values for protective variants along with the risk variants. Figure 3 shows that although the power of all the tests is negatively correlated with the number of protective variants, TOWA$_{AFR}$ is the most powerful test when causal variants are from AFR and TOWA$_{EUR}$ is the most powerful test when causal variants are from EUR. In general, all the corresponding TOWA-type tests outperform the corresponding AWDS tests and the WDS test despite that the WDS test has inflated type I error rate in admixed populations.

Impact of Uncertainty in Ancestry Probability Estimation

In the aforementioned analyses, we assumed the ancestry states in the tested region are known. In real studies, we need to estimate the ancestry states in the tested region using ancestry informative markers' genotypes in that region.

Several software packages are available for estimating the ancestry states, including ANCESTRYMAP [Patterson et al., 2004], MALDSOFT [Montana et al., 2004], ADMIXPROGRAM [Zhu et al., 2006], SABER [Tang et al., 2006], LAMP

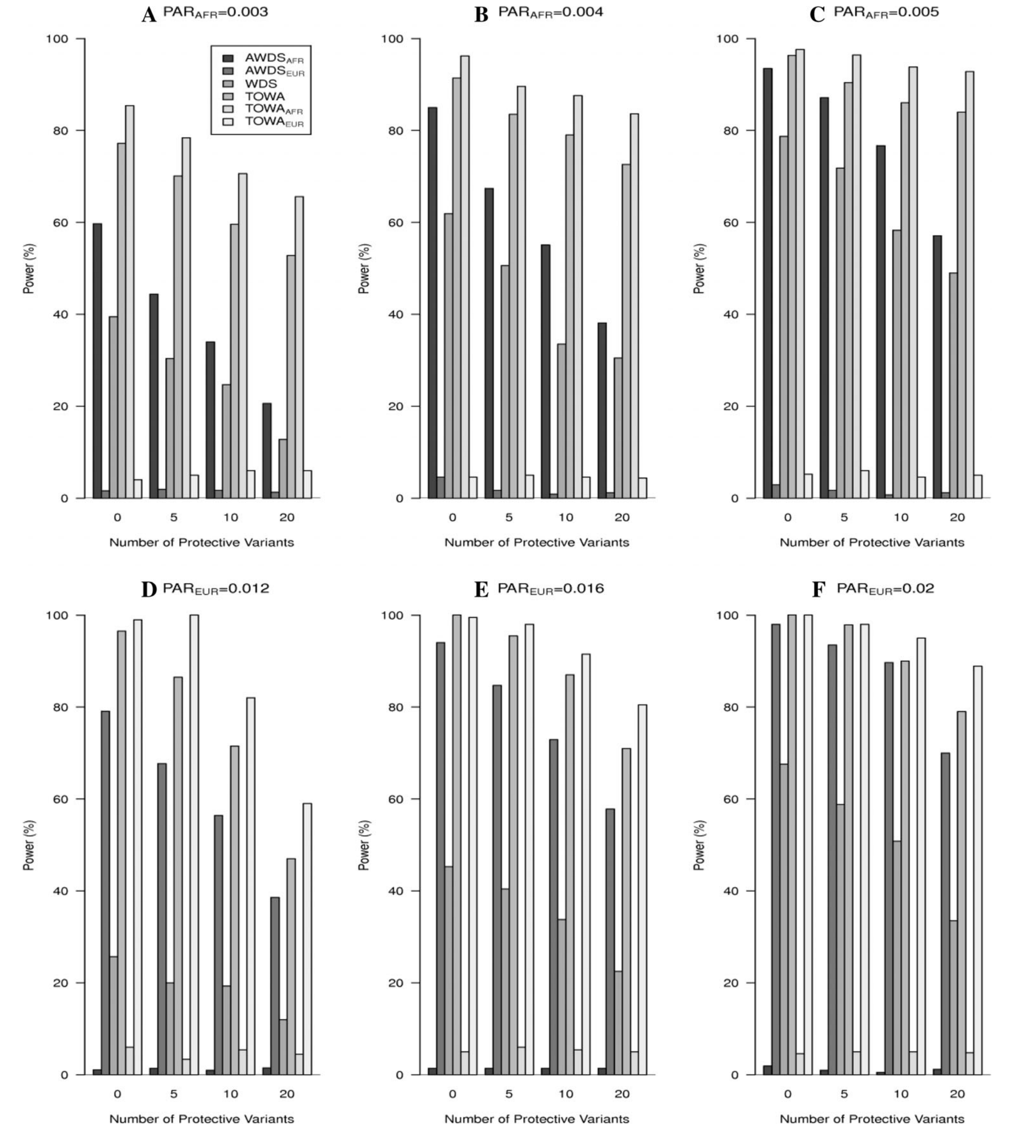

Figure 3. Power comparison under the mixture of risk and protective variants models. We fixed the number of causal variants as 20 for all scenarios. (A–C) Causal variants were all from AFR ancestry background; (D–F) causal variants were all from EUR ancestry background. Significance was assessed at the 5% level.

[Sankararaman et al., 2008], HAPAA [Sundquist et al., 2008], HAPMIX [Price et al., 2009], and SEQMIX (local-ancestry inference for sequenced admixed individuals) [Hu et al., 2013]. The choice of which program to use will depend on the nature of the data. The current state-of-the-art method is HAPMIX, which can yield an estimated ancestry that has as high as 98% correlation with the true ancestry. We chose to use HAPMIX to estimate local ancestries in the tested region given its high quality and its ability to provide allele-specific ancestry estimation. In order to verify the quality of HAPMIX, we have tested the software by using 200 individuals across 5,000 consecutive SNPs on chromosome 22. All the AFR and EUR haplotypes from the 1000 Genome Project are severed as references in our testing. As expected, the inferred ancestry states with HAPMIX are more than 98% identical to the true states. Ideally, we should use programs such as HAPMIX to estimate local ancestry probabilities for all of the 10,000 simulated individuals' genomes. However, these programs are computationally intensive and it is not feasible to run these programs on all simulated datasets. To circumvent this difficulty, we added uncertainties to the ancestry states according to the error model [Mao et al., 2013] in Table 3, which indices similar patterns of uncertainties for local ancestry estimates as HAPMIX.

We re-evaluated the type I error rates and power for TOWA, TOWA$_{\text{AFR}}$, and TOWA$_{\text{EUR}}$ based on two risk-variant only and one-sided disease models by incorporating random errors in local ancestry estimates in the simulated datasets following the error model in Table 3. As presented in Table 4, the type I error rates of TOWA, TOWA$_{\text{AFR}}$, and TOWA$_{\text{EUR}}$ are

Table 3. Error model for ancestry probability estimation

	True ancestry probability	Probability with uncertainty
Without recombination events	$P(A = \text{AFR}) = 0$	e_1
	$P(A = \text{EUR}) = 1$	$1 - e_1$
	$P(A = \text{AFR}) = 1$	$1 - e_1$
	$P(A = \text{EUR}) = 0$	e_1
With recombination events	$P(A = \text{AFR}) = 0$	e_2
	$P(A = \text{EUR}) = 1$	$1 - e_2$
	$P(A = \text{AFR}) = 1$	$1 - e_2$
	$P(A = \text{EUR}) = 0$	e_2

e_1 is generated from uniform (0, 0.01). It is used when there is no recombination between different ancestral populations. e_2 is generated from uniform (0.3, 0.7). It is used when there are recombinations between different ancestral populations. A is the ancestry state of an allele.

acceptable when we used the ancestry probabilities instead of the true ancestry states in association studies. The type I error of TOWA with estimated ancestry and with mean local ancestry difference in category III is slightly excessive as 0.062, possibly due to inaccurate estimation of local ancestry. This indicates how to accurately estimate local ancestry is critical to perform TOWA. The powers of TOWA, TOWA$_{\text{AFR}}$, and TOWA$_{\text{EUR}}$ are reduced slightly (<5%) when using the ancestry probabilities instead of the true ancestry states.

Application to the Hematocrit Dataset

We applied the proposed methods to a combined dataset on hematocrit (HCT) from Candidate-gene Association Resource (CARe) and Women's Health Initiative (WHI) [Reiner et al., 2011]. Exome sequence data were available on a total of 1,692 African American participants of the two cohorts (Duan et al., 2013; Fu et al., 2013). Among them, 1,314 participants have HCT phenotypes (in percentage). HCT, also known as packed cell volume, which is the volume percentage (%) of red blood cells in blood. Previous genome-wide association study identified strong association between HCT and a common SNP rs7312105 on gene *CACNA1C* (Harst et al., 2012; *P*-value = 4×10^{-9}), a gene on chromosome 12 encoding an alpha-1 subunit of a voltage-dependent calcium channel. Calcium channels mediate the influx of calcium ions into the cell upon membrane polarization. An abnormally high HCT, usually called polycythemia, is a life-threatening disorder.

In order to evaluate the performance of the proposed methods in real data, we applied TOWA to the HCT exome-sequencing data and tested association between gene *CACNA1C* and HCT. We assigned individuals with HCT > 41.7 as cases (n = 376) and those with HCT < 41.7 as controls (n = 938). The allele-specific population ancestry in the *CACNA1C* region was inferred using HAPMIX for all cases and controls. We used 1000 Genomes Project phased haplotypes of EUR and AFR as reference. Due to the limitation of the HAPMIX, we can only infer ancestry for 259 SNPs on *CACNA1C* that have both EUR and AFR reference panels in the 1000 Genomes Project. The average African ancestries for cases and controls are 0.6302 (SD = 0.13) and 0.6247 (SD = 0.16), respectively. Our analysis was performed on 22 less common SNPs with MAF < 0.05. As shown in Table 5, TOWA$_{\text{AFR}}$ revealed marginal association for the African side between rare variants in *CACNA1C* and HCT.

Table 5. Real data analysis of *CACNA1C*

	TOWA	TOWA$_{\text{AFR}}$	TOWA$_{\text{EUR}}$	AWDS$_{\text{AFR}}$	AWDS$_{\text{EUR}}$
Variants (MAF < 0.05)	0.065	0.041	0.219	0.076	0.228

P-values of TOWA and AWDS are based on 10,000 permutations and 10,000 bootstrap samples, respectively.

Discussion

Population stratification has long been recognized as an issue in genetic association studies. Unrecognized population stratification can lead to both false-positive and false-negative findings and can obscure true association signals if not appropriately corrected. For rare variants, this problem can be even more serious since the spectrum of rare variation can be very different in diverse populations. It has been shown that there exists noticeable local ancestry difference in samples collected from admixed populations even when their global ancestry patterns are similar [Mao et al., 2013]. Global ancestry adjustment is not sufficient for correcting population stratification induced by local ancestry difference and could lead to inflated type I errors for even a small disparity of local ancestry [Mao et al., 2013; Qin et al., 2010; Wang et al., 2011]. In this article, we proposed a novel method, TOWA, to correct for population stratification. TOWA takes into account the local ancestry of each allele in the tested region, and thus allows the local ancestry difference among study subjects to be appropriately modeled.

TOWA corrects for population stratification by incorporating local ancestry state of the tested variants in the test statistic. How to efficiently and accurately estimate local ancestry

Table 4. Type I error and power of TOWA when (1) true ancestry state was known; (2) ancestry state was estimated with uncertainty

		True ancestry			Estimated ancestry		
		TOWA	TOWA$_{\text{AFR}}$	TOWA$_{\text{EUR}}$	TOWA	TOWA$_{\text{AFR}}$	TOWAF$_{\text{EUR}}$
Type I	Category I	0.034	0.044	0.043	0.040	0.045	0.046
	Category II	0.042	0.046	0.045	0.050	0.048	0.050
	Category III	0.056	0.055	0.054	0.062	0.057	0.060
Power	PAR$_{\text{AFR}}$ = 0.003	0.777	0.859	0.050	0.773	0.855	0.046
	PAR$_{\text{EUR}}$ = 0.012	0.964	0.058	0.984	0.960	0.053	0.941

Uncertainty was introduced according to the error model in Table 3. Significance was assessed at the 5% level.

is critical to perform TOWA. Several software packages are available for estimating the ancestry states, including ANCESTRYMAP [Patterson et al., 2004], MALDSOFT [Montana et al., 2004], ADMIXPROGRAM [Zhu et al., 2006], SABER [Tang et al., 2006], LAMP [Sankararaman et al., 2008], HAPAA [Sundquist et al., 2008], HAPMIX [Price et al., 2009], and SEQMIX [Hu et al., 2013]. The current state-of-the-art method is HAPMIX, which can yield an estimated ancestry that has as high as 98% correlation with the true ancestry [Price et al., 2009]. However, we note that the software is time consuming. If a larger number of haplotypes are present in the reference panels, it may take months or years to complete. In addition, due to the extremely low MAFs of many variants in the sequence data, some observed rare variants in admixed populations may not be present in the reference panel. To circumvent these issues, one may need to preselect a subset of SNPs or ancestral informative markers covering the whole genome and then use HAPMIX to estimate the ancestry states of the subset of SNPs. The ancestry states for those nonselected SNPs can be interpolated by the inferred ancestry states of the flanking SNPs. When there is a switch of ancestry states between two preselected SNPs, one can analyze the original set of SNPs in that region and pinpoint the switch point of ancestry states. If sequencing data are available, one may also consider using SEQMIX [Hu et al., 2013], a recently developed tool, which can accurately infer local ancestry in exome-sequenced or targeted sequenced admixed individuals via the use of off-target sequence reads. Off-target reads generated during exome-sequencing experiments can be combined with on-target reads to accurately estimate the ancestry of each chromosomal segment in an admixed individual. With SEQMIX, accurate ancestry calls (squared correlation between true ancestry and SEQMIX result is ~0.9) can be generated with as little as 0.1-fold coverage of the nontargeted part of the genome.

For a genome-wide rare variants analysis, the computational time of the proposed method is acceptable though the *P*-value is determined by the permutation method. In order to estimate the computational time needed using TOWA in a genome-wide rare variants analysis, we performed a whole-genome scan using TOWA. We partitioned the whole genome into 44,620 nonoverlapping regions with 200 SNPs in each region. We used 1,000 permutations to estimate the *P*-value of TOWA for each region. For a sample with 1,000 cases and 1,000 controls, the computational time of *P*-value estimation of TOWA for all the 44,620 regions was 45 h using a 2.40 GHz single central processing unit (CPU) with 58 MB average memory. If we use parallel computing with 100 CPUs, the computational time for a whole-genome scan with TOWA would be 27 min.

The method proposed in this article is applicable for qualitative traits. We can use the similar idea for qualitative traits to develop a method for quantitative traits. For a quantitative trait, a linear regression can be used to model the relationship between the trait and genetic variants. For an admixed population, we can incorporate the ancestry status of each allele of an SNP into the linear regression model. Then, a score test can be developed to test the association between a combination of rare variants and the trait in admixed populations. However, the performance of the method for quantitative traits needs further investigation.

Appendix

In the following discussion, we assume that, for given genotypes, the ancestral statuses in different variants are independent. That is, $\Pr(A_i|D_i) = \prod_{j=1}^{M} \Pr(A_{ij}|D_i) \overset{\Delta}{=} \prod_{j=1}^{M} \Pr(A_{ij})$.

Let $\delta_i = \alpha + \beta_1 \sum_{j=1}^{M} (x_{ij1}a_{ij1} + x_{ij2}a_{ij2})w_{1j} + \beta_2 \sum_{j=1}^{M} (x_{ij1} b_{ij1} + x_{ij2}b_{ij2})w_{2j}$ and $p_i = \Pr(y_i = 1|\mathrm{D}_i) = \sum_{A_i} \Pr(y_i = 1| D_i, A_i) \Pr(A_i|D_i) = \sum_{A_{i1}=0}^{3} \cdots \sum_{A_{iM}=0}^{3} \frac{\exp(\delta_i)}{1+\exp(\delta_i)} \prod_{j=1}^{M} \Pr(A_{ij})$.

The log-likelihood is given by

$$\begin{aligned}
\log L &= \sum_{i=1}^{n} (y_i \log p_i + (1 - y_i) \log(1 - p_i)) \\
&= \sum_{i=1}^{n} \left(y_i \log \left(\sum_{A_{i1}=0}^{3} \cdots \sum_{A_{iM}=0}^{3} \frac{\exp(\delta_i)}{1 + \exp(\delta_i)} \prod_{j=1}^{M} \Pr(A_{ij}) \right) \right. \\
&\quad \left. + (1 - y_i) \log \left(\sum_{A_{i1}=0}^{3} \cdots \sum_{A_{iM}=0}^{3} \frac{1}{1 + \exp(\delta_i)} \prod_{j=1}^{M} \Pr(A_{ij}) \right) \right).
\end{aligned}$$

Under null hypothesis, the maximum likelihood estimate (MLE) of α is $\hat{\alpha} = \log \bar{y} - \log(1 - \bar{y})$. Then, we can have

$$\begin{aligned}
\frac{d^2 \log L}{d\alpha^2} &= \sum_{i=1}^{n} \left(y_i \frac{\frac{d^2 p_i}{d\alpha^2} p_i - \left(\frac{dp_i}{d\alpha}\right)^2}{p_i^2} \right. \\
&\quad \left. + (1 - y_i) \frac{-\frac{d^2 p_i}{d\alpha^2} (1 - p_i) - \left(\frac{dp_i}{d\alpha}\right)^2}{(1 - p_i)^2} \right)
\end{aligned}$$

$$\frac{dp_i}{d\alpha} = \sum_{A_{i1}=0}^{3} \cdots \sum_{A_{iM}=0}^{3} \frac{\exp(\delta_i)}{(1 + \exp(\delta_i))^2} \prod_{j=1}^{M} \Pr(A_{ij})$$

$$\frac{d^2 p_i}{d\alpha^2} = \sum_{A_{i1}=0}^{3} \cdots \in \sum_{A_{iM}=0}^{3} \frac{\exp(\delta_i) - \exp(\delta_i)^2}{\left(1 + \exp\left(\delta_i\right)\right)^3} \prod_{j=1}^{M} \Pr(A_{ij})$$

$$p_i|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \bar{y}$$

$$\left.\frac{dp_i}{d\alpha}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \bar{y}(1 - \bar{y})$$

$$\left.\frac{d^2 p_i}{d\alpha^2}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \bar{y}(1 - \bar{y})(1 - 2\bar{y})$$

$$\left. - E \frac{d^2 \log L}{d\alpha^2}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = n\bar{y}(1 - \bar{y})$$

$$\frac{d \log L}{d\beta_1} = \sum_{i=1}^{n} \left(y_i \frac{dp_i/d\beta_1}{p_i} + (1 - y_i) \frac{-dp_i/d\beta_1}{1 - p_i} \right)$$

$$\frac{d^2 \log L}{d{\beta_1}^2} = \sum_{i=1}^{n} \left(y_i \frac{\frac{d^2 p_i}{d{\beta_1}^2} p_i - \left(\frac{dp_i}{d\beta_1}\right)^2}{p_i^2} \right.$$

$$\left. + (1-y_i) \frac{-\frac{d^2 p_i}{d{\beta_1}^2}(1-p_i) - \left(\frac{dp_i}{d\beta_1}\right)^2}{(1-p_i)^2} \right)$$

$$\frac{dp_i}{d\beta_1} = \sum_{A_{i1}=0}^{3} \cdots \sum_{A_{iM}=0}^{3} \frac{\exp(\delta_i)}{(1+\exp(\delta_i))^2} \sum_{j=1}^{M} (x_{ij1} a_{ij1}$$

$$+ x_{ij2} a_{ij2}) w_{1j} \prod_{j=1}^{M} \Pr(A_{ij})$$

$$\frac{d^2 p_i}{d\beta_1^2} = \sum_{A_{i1}=0}^{3} \cdots \sum_{A_{iM}=0}^{3} \frac{\exp(\delta_i) - \exp\,(\delta_i)^2}{(1+\exp(\delta_i))^3}$$

$$\times \left(\sum_{j=1}^{M} (x_{ij1} a_{ij1} + x_{ij2} a_{ij2}) w_{1j} \right)^2 \prod_{j=1}^{M} \Pr(A_{ij})$$

$$\left.\frac{dp_i}{d\beta_1}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \bar{y}(1-\bar{y}) x_i^T w_1$$

$$\left.\frac{d^2 p_i}{d{\beta_1}^2}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \bar{y}(1-\bar{y})(1-2\bar{y}) w_1^T x_i x_i^T w_1$$

$$\left.\frac{d \log L}{d\beta_1}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \sum_{i=1}^{n} \left((y_i - \bar{y}) x_i^T w_1 \right)$$

$$-E \left.\frac{d^2 \log L}{d\beta_1^2}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \bar{y}(1-\bar{y}) {w_1}^T \left(\sum_{i=1}^{n} x_i x_i^T \right) w_1$$

Similarly,

$$\frac{d \log L}{d\beta_2}|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \sum_{i=1}^{n} \left((y_i - \bar{y}) z_i^T w_2 \right)$$

$$-E \left.\frac{\partial^2 \log L}{\partial{\beta_2}^2}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \bar{y}(1-\bar{y}) w_2^T \left(\sum_{i=1}^{n} z_i z_i^T \right) w_2$$

$$-E \left.\frac{\partial^2 \log L}{\partial\beta_1 \partial\beta_2}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = \bar{y}(1-\bar{y}) w_1^T \left(\sum_{i=1}^{n} x_i z_i^T \right) w_2$$

$$-E \left.\frac{\partial^2 \log L}{\partial\beta_1 \partial\alpha}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = n\bar{y}(1-\bar{y}) \bar{x}^T w_1$$

$$-E \left.\frac{\partial^2 \log L}{\partial\beta_2 \partial\alpha}\right|_{\beta_1=0,\beta_2=0,\alpha=\hat{\alpha}} = n\bar{y}(1-\bar{y}) \bar{z}^T w_2.$$

The score test statistic is given by

$$T_{\text{score}} = \left(w_1^T u_1, w_2^T u_2 \right) \begin{pmatrix} w_1^T A w_1 & w_1^T C w_2 \\ w_1^T C w_2 & w_2^T B w_2 \end{pmatrix}^{-1}$$

$$\times \left(w_1^T u_1, w_2^T u_2 \right)^T / \hat{\sigma}^2$$

where $A = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$, $B = \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})^T$, $C = \sum_{i=1}^{n} (x_i - \bar{x})(z_i - \bar{z})^T$, and $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \bar{y}(1-\bar{y})$.

Acknowledgment

This research was supported by the National Institutes of Health [R01HG004517 and R01HG005854 to ML; R01HG006292 and R01HG006703 to YL].

The authors have no conflict of interests to declare.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population scale sequencing. *Nature* 467:1061–1073.

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11:773–785.

Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH. 2006. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma lowdensity lipoprotein levels. *Proc Natl Acad Sci USA* 103:1810–1815.

Devlin B. Roeder K. 1999. Genomic control for association studies. *Biometrics* 55(4):997–1004.

Derkach A, Lawless J, Sun L. 2012. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epi* 37(1):110–121.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide sssociations. *PLoS Biol* 8:e1000294.

Duan Q, Liu EY, Auer PL, Zhang G, Lange EM, Jun G, Bizon C, Jiao S, Buyske S, Franceschini N and others. 2013. Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. *Bioinformatics* 29:2744–2749.

Epstein MP, Allen AS, Satten GA. 2007. A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet* 80:921–930.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson DA and others. 2013. Analysis of 6515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–220.

Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petroviski S, Sunyaev S. 2013. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* 14(7):460–470.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70:42–54.

Harst PVD, Zhang W, Leach IM, Rendon A, Verweij N, Sehmi J, Paul DS, Elling U, Allagee H, Li X and others. 2012. Seventy-five genetic loci influencing the human red blood cell. *Nature* 492(7429):369–375.

Hershberger R, Norton N, Morales A, Li D, Siegfried J, Gonzalez-Quintana J. 2010. Coding sequence rare variants identified in MYBPC3, MYH6, TPM1, TNNC1, and TNNI3 from 312 patients with familial or idiopathic dilated cardiomyopathy. *Circ Cardiovasc Genet* 3:155–161.

Hoffmann TJ, Marini NJ, Witte JS. 2010. Comprehensive approach to analyzing rare genetic variants. *PLoS ONE* 5(11):e13584.

Hu Y, Willer C, Zhan X, Kang HM, Abecasis GR. 2013. Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. *Am J Hum Genet* 93(5):891–899.

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. 2013. Family-based association tests for sequence data, and comparisons with population-based association test. *Eur J Hum Genet* 21:1158–1162.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. 2008. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* 40:592–599.

Knowler WC, Williams RC, Petitt DJ, Steinberg AG. 1988. Gm and Type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 43:520–526.

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. *Science* 265(30):2037–2048.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 91:224–237.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83:311–321.

Li M, Reilly MP, Rader DJ, Wang LS. 2010. Correcting population stratification in genetic association studies using a phylogenetic approach. *Bioinformatics* 26:798–806.

Li Y, Byrnes AE, Li M. 2010. To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. *Am J Hum Genet* 87:728–735.

Lin D-Y, Tang Z-Z. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89:354–367.

Madsen BE, Browning SR. 2009. A group-wise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5:e1000384.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–753.

Mao X, Li Y, Liu Y, Lange L, Li M. 2013. Testing genetic association with rare variants in admixed populations. *Genetic Epi* 27(1):38–47.

Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 44:243–246.

Marini NJ, Gin J, Ziegle J, Keho KH, Ginzinger D, Gilbert DA, Rine J. 2008. The prevalence of folate-remedial MTHFR enzyme variants in humans. *Proc Natl Acad Sci USA* 105:8055–8060.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9:356–369.

Montana G, Pritchard JK. 2004. Statistical tests for admixture mapping with case control and cases-only data. *Am J Hum Genet* 75:771–789.

Neale BM, Rivas MA, Kathiresan S, Voight BF, Alshuler D, Devlin B, Orho-Melander M, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet* 7:e1001322.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 324:387–389.

Pan W. 2009. Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genet Epidemiol* 33(6):497–507.

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D and others. 2004. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* 74:979–1000.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 5:e1000519.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86:832–838.

Qin H, Morris N, Kang SJ, Li M, Tayo B, Lyon H, Hirschhorn J, Cooper RS, Zhu X. 2010. Integrating local population structure for fine mapping in genome-wide association studies. *Bioinformatics* 26:2961–2968.

Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 20(1):4–16.

Reiner AP, Lettre G, Nalls MA, Ganesh SK, Mathias R, Austin MA, Dean E, Arepalli S, Britton A, Chen Z and others. 2011. Genome-wide association study of white blood cell count in 16388 African Americans: the continental origins and genetic epidemiology network (COGENT). *PLoS Genet* 7:e1002108.

Sankararaman S1, Sridhar S, Kimmel G, Halperin E. 2008. Estimating local ancestry in admixed populations. *Am J Hum Genet* 82:290–303.

Schork NJ, Murray SS, Frazer KA, Topol EJ. 2009. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19:212–219.

Sha Q, Wang X, Wang X, Zhang S. 2012. Detecting association of both rare and common variants by testing an optimally weighted combination of variants. *Genet Epidemiol* 36:561–571.

Sha Q, Wang S, Zhang S. 2013. Adaptive clustering and adaptive weighting methods to detect disease associated rare variants. *Eur J Hum Genet* 21(3):332–337.

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E and others. 2004. A high-density admixture map for gene discovery in African Americans. *Am J Hum Genet* 74:1001–1013.

Sundquist A, Fratkin E, Do CB, Batzoglou S. 2008. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res* 18:676–682.

Tang H, Coram M, Wang P, Zhu X, Risch N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* 79:1–12.

Wang X, Zhu X, Qing H, Cooper R, Ewens W, Li C, Li M. 2011. Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics* 27:670–677.

Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am J Hum Genet* 89:82–93.

Yi N, Zhi D. 2011. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol* 35:57–69.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S. 2010 Extending rare variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 87:604–617.

Zhu X, Zhang S, Tang H, Cooper R. 2006. A classical likelihood based approach for admixture mapping using EM algorithm. *Hum Genet* 120(3):431–445.

Zhu X, Feng T, Li Y, Lu Q, Elston RC. 2010. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol* 34:171–187.