ARTICLE

# ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure

Timothy Thornton[1] and Mary Sara McPeek[2,3,*]

Genome-wide association studies are routinely conducted to identify genetic variants that influence complex disorders. It is well known that failure to properly account for population or pedigree structure can lead to spurious association as well as reduced power. We propose a method, ROADTRIPS, for case-control association testing in samples with partially or completely unknown population and pedigree structure. ROADTRIPS uses a covariance matrix estimated from genome-screen data to correct for unknown population and pedigree structure while maintaining high power by taking advantage of known pedigree information when it is available. ROADTRIPS can incorporate data on arbitrary combinations of related and unrelated individuals and is computationally feasible for the analysis of genetic studies with millions of markers. In simulations with related individuals and population structure, including admixture, we demonstrate that ROADTRIPS provides a substantial improvement over existing methods in terms of power and type 1 error. The ROADTRIPS method can be used across a variety of study designs, ranging from studies that have a combination of unrelated individuals and small pedigrees to studies of isolated founder populations with partially known or completely unknown pedigrees. We apply the method to analyze two data sets: a study of rheumatoid arthritis in small UK pedigrees, from Genetic Analysis Workshop 15, and data from the Collaborative Study of the Genetics of Alcoholism on alcohol dependence in a sample of moderate-size pedigrees of European descent, from Genetic Analysis Workshop 14. We detect genome-wide significant association, after Bonferroni correction, in both studies.

## Introduction

It is well known that problems can arise in case-control genetic association studies when there is population structure.[1] At its most basic, case-control association testing can be thought of as a comparison of the allele (or genotype) frequency distribution between cases and controls, and markers that are not directly associated with the trait of interest can be spuriously associated with the trait if ancestry differences between cases and controls are not properly accounted for. Similarly, failure to account for population structure can also reduce power. To correct for population structure in case-control studies with samples of unrelated individuals, a number of methods have been proposed, including genomic control (GC),[2] structured association,[3] spectral analysis,[4–8] and other approaches.[9–13] However, many genetic studies include related individuals. Several methods have been proposed for case-control association testing in related samples from a single population with known pedigrees[14–16] or with unknown or partially known pedigrees.[17] However, these methods might not be valid in the presence of population heterogeneity. For certain types of study designs, family-based association tests such as the TDT[18] and FBAT[19] have been used to protect against potential problems of unknown population substructure. Family-based tests, however, are generally less powerful than case-control association methods[20,21] and are more restrictive because they typically require genotype data for family members of an affected individual. In contrast, case-control designs can allow, but do not require, genotype data for relatives of affected individuals.

We address the general problem of case-control association testing in samples with related individuals from structured populations. We do not put constraints on how the individuals might be related, and we allow for the possibility that the pedigree information could be partially or completely missing. We propose a new method, ROADTRIPS, where this name is inspired by the description of the method as a robust association-detection test for related individuals with population substructure. ROADTRIPS uses a covariance matrix estimated from genome-screen data to simultaneously correct for both population and pedigree structure. The method does not require the pedigree structure of the sampled individuals to be known, but when pedigree information is available, the method can improve power by incorporating this information into the analysis. ROADTRIPS is computationally feasible for genetic studies with millions of markers. Other features of ROADTRIPS include (1) appropriate handling of missing data and (2) the ability to incorporate both unaffected controls and controls of unknown phenotype (i.e., general population controls) in the analysis.

In order to compare ROADTRIPS to other methods, on the basis of type 1 error and power, we simulate case-control samples containing both related and unrelated individuals with various types of population structure, including admixture. We also apply ROADTRIPS to identification of SNPs associated with rheumatoid arthritis (RA [MIM 180300]) in small UK pedigrees[22] from Genetic

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [2]Department of Statistics, [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA
*Correspondence: mcpeek@galton.uchicago.edu
DOI 10.1016/j.ajhg.2010.01.001. 

Analysis Workshop (GAW) 15, and we apply it to identification of SNPs associated with alcohol dependence (MIM 103780) in a sample of moderate-size pedigrees of European descent from the Collaborative Study of the Genetics of Alcoholism (COGA) data[23] of GAW 14.

## Material and Methods

We first describe a class of testing procedures suitable for known structure. Then we describe the ROADTRIPS method for extending these tests to the contexts of unknown and partially known structure.

### Overview of Association Testing with Known Structure

Consider the problem of testing for association of a genetic marker with a particular phenotype in a sample of $n$ genotyped individuals. For simplicity, we assume that the marker to be tested is a SNP, with alleles labeled "0" and "1." (The extension to multiallelic markers can be obtained as in previous work.[16]) Let $\mathbf{Y} = (Y_1, \ldots Y_n)^T$ where $Y_i = \frac{1}{2} \times$ (the number of alleles of type 1 in individual $i$), so the value of $Y_i$ is 0, $\frac{1}{2}$, or 1. We treat the genotype data on the $n$ individuals as random and the available phenotype information as fixed in the analysis, an approach that is appropriate, for example, with either random or phenotype-based ascertainment. Under the null hypothesis of no association and no linkage between marker and trait, we assume that the expected value of $\mathbf{Y}$ is $E_0(\mathbf{Y}) = p\mathbf{1}$, where $\mathbf{1}$ is a column vector of 1s of length $n$ and $p$ is a parameter representing the frequency of the type 1 allele. In models incorporating population structure, $p$ would typically be interpreted as an "ancestral" allele frequency or some kind of average allele frequency across subpopulations. We denote by $\mathrm{Var}_0(\mathbf{Y}) = \mathbf{\Sigma}$ the $n \times n$ covariance matrix of $\mathbf{Y}$ under the null hypothesis of no association. It is often convenient to write $\mathbf{\Sigma} = \sigma^2\mathbf{\Psi}$, where $\sigma^2$ is defined to be the variance of $Y$ for an outbred individual in the absence of population structure, and $\mathbf{\Psi}$ accounts for relatedness, inbreeding, and population structure. We use the term "known structure" to refer to the case when the matrix $\mathbf{\Psi}$ is known. We always take $\sigma^2$ to be unknown and estimate it from the sample. Denote by $\widehat{\sigma}^2$ a suitable estimator of $\sigma^2$ (where two examples of suitable estimators are given in the next subsection). Then, in the case of known structure, we consider test statistics for association that have the rather general form

$$\frac{(\mathbf{V}^T\mathbf{Y})^2}{\left(\widehat{\sigma}^2\mathbf{V}^T\mathbf{\Psi}\mathbf{V}\right)} \qquad \text{(Equation 1)},$$

where $\mathbf{V}$ is a fixed, nonzero column vector of length $n$ such that $\mathbf{V}^T\mathbf{1} = 0$. Note that $\mathrm{Var}_0(\mathbf{V}^T\mathbf{Y}) = \sigma^2\mathbf{V}^T\mathbf{\Psi}\mathbf{V}$, so the denominator in Equation 1 can be viewed as an estimator of $\mathrm{Var}_0(\mathbf{V}^T\mathbf{Y})$. In a test for association, $\mathbf{V}$ would naturally include phenotype information and could also include pedigree information. One could include covariate information in $\mathbf{V}$ as well, although we do not treat that situation in the present work. There are a number of case-control association test statistics that have the general form in Equation 1, including the Pearson $\chi^2$ test, the Armitage trend test,[24] the corrected $\chi^2$ test,[15] the $W_{QLS}$ test,[15] and the $M_{QLS}$ test[16] (details on how these tests can be written in the form of Equation 1 are given in subsection Examples of Association Tests with Known Structure). Under standard regularity conditions, the test statistic given in Equation 1 has an asymptotic ${\chi_1}^2$ distribution under the null hypothesis of no association and no linkage.

### Estimation of $\sigma^2$ when Structure Is Known

In the context of Equation 1, when structure is known, we have two general approaches for estimating $\sigma^2$ under the null hypothesis. If we assume that, for an outbred individual in the absence of population structure, HWE holds at the marker, then $\sigma^2 = \frac{1}{2}\,p(1-p)$, where $p$ is the frequency of allele 1 at the SNP being tested, and a reasonable estimator of $\sigma^2$ under this assumption is $\widehat{\sigma}_1^2 = 0.5\widehat{p}(1-\widehat{p})$, where $\widehat{p}$ is a suitable estimator of $p$, the frequency of allele 1 at the SNP being tested. Examples of suitable estimators of $p$ are (1) the sample frequency, $\overline{Y}$; (2) the best linear unbiased estimator (BLUE),[25] given by

$$\widehat{p} = \left(\mathbf{1}^T\mathbf{\Psi}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}^T\mathbf{\Psi}^{-1}\mathbf{Y} \qquad \text{(Equation 2)};$$

and (3) a Bayesian estimator[26] such as $(n\overline{Y} + 0.5)/(n+1)$.

Alternatively, an approach to estimation of $\sigma^2$ that does not assume $\sigma^2 = 0.5p(1-p)$ could be used. When $\mathbf{\Psi}$ is known, a reasonable estimator is

$$\widehat{\sigma}_2^2 = (n-1)^{-1}\left[\mathbf{Y}^T\mathbf{\Psi}^{-1}\mathbf{Y} - \left(\mathbf{1}^T\mathbf{\Psi}^{-1}\mathbf{1}\right)^{-1}\left(\mathbf{1}^T\mathbf{\Psi}^{-1}\mathbf{Y}\right)^2\right] \qquad \text{(Equation 3)},$$

which is RSS/$(n-1)$ for generalized regression of $\mathbf{Y}$ on $\mathbf{1}$, where RSS is the residual sum of squares. Note that when $\mathbf{\Psi} = \mathbf{I}$, the $n \times n$ identity matrix, e.g., with unrelated individuals in the absence of population structure, then $\widehat{\sigma}_2^2$ is just the sample variance of $\mathbf{Y}$.

### Examples of Association Tests with Known Structure

#### *Corrected Pearson $\chi^2$ and Armitage Trend Tests*

In the standard Pearson $\chi^2$ and Armitage trend tests for allelic association, one assumes that the individuals are unrelated with no population structure, so that $\mathbf{\Psi} = \mathbf{I}$. A corrected version of the Pearson $\chi^2$ test has previously been described[15] for the situation when sampled individuals are related with all relationships known, in which case $\mathbf{\Psi} = \mathbf{\Phi}$, where $\mathbf{\Phi}$ is the kinship matrix, which is obtained as a function of the known pedigree information and is given by

$$\mathbf{\Phi} = \begin{pmatrix} 1+h_1 & 2\phi_{12} & \ldots & 2\phi_{1n} \\ 2\phi_{12} & 1+h_2 & \ldots & 2\phi_{2n} \\ \vdots & \ldots & \ldots & \vdots \\ 2\phi_{n1} & 2\phi_{n2} & \ldots & 1+h_n \end{pmatrix} \qquad \text{(Equation 4)},$$

where $h_i$ is the inbreeding coefficient of individual $i$, and $\phi_{ij}$ is the kinship coefficient between individuals $i$ and $j$, $1 \le i, j \le n$. We propose to use this same choice of $\mathbf{\Psi}$ in the corrected Armitage test. (More generally, for either test, one might consider known structure $\mathbf{\Psi}$ that does not necessarily equal $\mathbf{\Phi}$.) In both tests, one further assumes that every individual in the sample can be classified as either case or control. In that context, let $\mathbf{1}_c$ be the case indicator, i.e., the vector of length $n$ whose $i$th entry is 1 if individual $i$ is a case and 0 if individual $i$ is a control. Then both the corrected Pearson $\chi^2$ and corrected Armitage test statistics are obtained as special cases of Equation 1 with the choice

$$\mathbf{V} = \mathbf{1}_c - \frac{n_c}{n}\mathbf{1} \qquad \text{(Equation 5)},$$

where $n_c$ is the number of case individuals among the $n$ total individuals. (In the most general specification of the Armitage test for genetic association, mean-zero, nonlinear functions of $\mathbf{Y}$ are

allowed in place of $\mathbf{V}^T\mathbf{Y}$, but in practice, the test is almost always performed with the $\mathbf{V}$ given in Equation 5.) The difference between the two tests is in the estimation of $\sigma^2$. The corrected Pearson $\chi^2$ test uses the estimator $\hat{\sigma}_1^2$ described in the previous subsection, with $\hat{p}$ taken to be either $\overline{Y}$ or the BLUE given in Equation 2, whereas the corrected Armitage test uses the estimator $(1-n^{-1})\,\hat{\sigma}_2^2$, where $\hat{\sigma}_2^2$ is given in Equation 3. In the special case, $\mathbf{\Phi}=\mathbf{I}$, the corrected Pearson $\chi^2$ and Armitage test statistics equal the standard Pearson $\chi^2$ and Armitage test statistics, respectively. When we calculate the corrected Armitage test statistic, we actually use estimator $\hat{\sigma}_2^2$ instead of $(1-n^{-1})\,\hat{\sigma}_2^2$. For the large values of $n$ typically encountered in human genetic studies, the difference between $\hat{\sigma}_2^2$ and $(1-n^{-1})\,\hat{\sigma}_2^2$ is negligible. In the context of complex trait mapping in samples of related individuals with known pedigrees, the corrected $\chi^2$ test has been demonstrated[15,16] to have correct type 1 error, generally higher power than the $W_{QLS}$ test, and generally somewhat lower power than the $M_{QLS}$ test. However, with additional unknown population structure, we would expect both the corrected Pearson $\chi^2$ and corrected Armitage tests to have inflated type 1 error.

*$W_{QLS}$ Test*

The $W_{QLS}$ test[15] was proposed in the context of related individuals without additional population structure, in which case $\mathbf{\Psi}=\mathbf{\Phi}$ given in Equation 4. The $W_{QLS}$ test statistic is formed from Equation 1 by taking

$$\mathbf{V}=\mathbf{\Phi}^{-1}\mathbf{1}_c-\mathbf{1}_c^T\mathbf{\Phi}^{-1}\mathbf{1}\left(\mathbf{1}^T\mathbf{\Phi}^{-1}\mathbf{1}\right)^{-1}\mathbf{\Phi}^{-1}\mathbf{1} \qquad \text{(Equation 6).}$$

This choice of $\mathbf{V}$ can be motivated by generalized least-squares regression, because $\mathbf{V}^T\mathbf{Y}$ is proportional to the estimated regression coefficient for $\mathbf{1}_c$ in the generalized least-squares regression of $\mathbf{Y}$ on $\mathbf{1}_c$ with intercept. The $W_{QLS}$ test uses the estimator $\hat{\sigma}_1^2$ of $\sigma^2$, where $\hat{p}$ is taken to be the BLUE given by Equation 2. An alternative formulation could be obtained by using the estimator $\hat{\sigma}_2^2$ of Equation 3.

In the context of trait mapping in samples of related individuals with known pedigrees, the $W_{QLS}$ test generally has lower power[16] than the corrected $\chi^2$ and $M_{QLS}$ tests. Nonetheless, we include it in the present work because the ROADTRIPS extension of the $W_{QLS}$ (described in subsection Association Tests when Structure Is Partially or Completely Unknown) is equivalent to the method, recently proposed by Rakovski and Stram,[13] for association testing in the presence of hidden population structure and hidden relatedness. Thus, we include the ROADTRIPS extension of the $W_{QLS}$ in our simulation studies to compare its power and type 1 error to those of our proposed methods.

*$M_{QLS}$ Test*

In contrast to the preceding tests, the $M_{QLS}$ test[16] allows three possible values for an individual's phenotype: "affected," "unaffected," and "unknown," where the label "unknown" is used to represent unphenotyped individuals, e.g., general population controls, or individuals who are deemed too young to have developed an age-related trait such as Alzheimer's, whereas the label "unaffected" is reserved for true unaffecteds. As they have different expected frequencies of predisposing alleles, the two types of controls are treated differently in the analysis. Furthermore, whereas the preceding tests use the phenotype information only for individuals who have genotype data at the marker being tested, the $M_{QLS}$ also uses the phenotype information for individuals with missing genotype data at the marker being tested, provided that those individuals have a sampled relative who is genotyped at the marker.

As a result of these considerations, instead of using the phenotype vector $\mathbf{1}_c$ that is used in the preceding four tests, the $M_{QLS}$ uses the vector $\mathbf{A}=(\mathbf{A}_N{}^T,\mathbf{A}_M{}^T)^T$, which contains more information than $\mathbf{1}_c$. Here, $\mathbf{A}_N$ is the phenotype vector for the $n$ individuals with nonmissing genotype data at the marker being tested, and $\mathbf{A}_M$ is the phenotype vector for the $m$ individuals with missing genotype data at the marker being tested, where individual $i$'s phenotype is coded as $A_i=1$ if $i$ is affected, $-k/(1-k)$ if $i$ is unaffected, and 0 if $i$ is of unknown phenotype, where $0<k<1$ is a constant that represents an external estimate of the population prevalence of the trait for a suitable reference population. (The prevalence estimate is permitted to be very rough; the $M_{QLS}$ test is valid for arbitrary fixed $k$.)

The $M_{QLS}$ test was proposed in the context of related individuals without additional population structure, in which case $\mathbf{\Psi}=\mathbf{\Phi}$, the $n\times n$ matrix given in Equation 4. In order to incorporate the information of $\mathbf{A}_M$ into the $M_{QLS}$ test, one also needs the $n\times m$ matrix, $\mathbf{\Phi}_{N,M}$, whose $(i,j)$th entry is $2\phi_{ij}$, where $\phi_{ij}$ is the kinship coefficient between the $i$th nonmissing and $j$th missing individuals. The $M_{QLS}$ test statistic can be obtained from Equation 1 by choosing

$$\mathbf{V}=\mathbf{A}_N+\mathbf{\Phi}^{-1}\mathbf{\Phi}_{N,M}\mathbf{A}_M-\left(\mathbf{A}_N+\mathbf{\Phi}^{-1}\mathbf{\Phi}_{N,M}\mathbf{A}_M\right)^T\mathbf{1}\left(\mathbf{1}^T\mathbf{\Phi}^{-1}\mathbf{1}\right)^{-1}\mathbf{\Phi}^{-1}\mathbf{1} \qquad \text{(Equation 7)}$$

and using the estimator $\hat{\sigma}_1^2$ of $\sigma^2$, where $\hat{p}$ is taken to be the BLUE given by Equation 2. An alternative formulation could be obtained by using the estimator $\hat{\sigma}_2^2$ of Equation 3. Two different justifications for the choice of $\mathbf{V}$ in Equation 7 have previously been described; one[16] is based on maximizing the noncentrality parameter among all tests of the type in Equation 1 when a two-allele model in outbreds (or an additive model in inbreds) with effect size tending to 0 is used, and the other[27] is based on a relationship with the score test for the retrospective likelihood based on logistic regression with an additive model. In the context of complex trait mapping in samples of related individuals with known pedigrees, the $M_{QLS}$ test has been demonstrated[16] to have generally higher power than both the corrected $\chi^2$ and $W_{QLS}$ tests. However, with additional unknown population structure, we would expect the $M_{QLS}$ test to have inflated type 1 error.

## Outline of ROADTRIPS Approach for Unknown Structure

The idea behind ROADTRIPS is to extend tests of the form given in Equation 1 to the situation when there could be unknown population structure and/or cryptic relatedness in the sample. To do this, we use genome-screen data to form an appropriate estimator $\hat{\mathbf{\Psi}}$ of $\mathbf{\Psi}$ and consider various tests of the form

$$\frac{\left(\mathbf{V}^T\mathbf{Y}\right)^2}{\left(\hat{\sigma}^2\mathbf{V}^T\hat{\mathbf{\Psi}}\mathbf{V}\right)} \qquad \text{(Equation 8),}$$

where we can allow $\mathbf{V}$ to take into account any known pedigree information, in addition to phenotype information, while simultaneously accounting for pedigree errors and additional unknown structure through $\hat{\mathbf{\Psi}}$. This approach allows us to easily adapt to different patterns of missing genotypes at different tested markers, by including only the rows and columns of $\hat{\mathbf{\Psi}}$ (and the entries of $\mathbf{V}$ and $\mathbf{Y}$) that correspond to the individuals genotyped at the particular marker being tested. In what follows, we first describe the population genetic modeling assumptions that underlie our estimation and testing procedures, then we describe the estimators $\hat{\mathbf{\Psi}}$ and $\hat{\sigma}^2$.

## Population Genetic Modeling Assumptions

The modeling assumptions we make are weak and are satisfied by commonly used models of population structure and commonly used models for related individuals. We consider $S$ SNPs in a genome screen, and we generalize the notation of the previous subsections to a set of $S$ SNPs by letting $\mathbf{Y}^s$ be the genotype vector corresponding to SNP $s$, namely, $\mathbf{Y}^s = (Y_1{}^s, \ldots, Y_n{}^s)^T$, $s = 1, \ldots, S$, where $Y_i^s = \frac{1}{2} \times$ (the number of alleles of type 1 at SNP $s$ in individual $i$). Our modeling assumption on the null mean, generalized from the preceding subsections, can be stated as

$$E_0(\mathbf{Y}^s) = p_s\mathbf{1}, \text{ for } 1 \le s \le S \qquad \text{(Equation 9).}$$

We make the following assumption regarding the null covariance matrix:

$$\mathrm{Var}_0(\mathbf{Y}^s) \equiv \mathbf{\Sigma}_s = \sigma_s^2\mathbf{\Psi}, \text{ for } 1 \le s \le S \qquad \text{(Equation 10),}$$

where $\mathbf{\Psi}$ is an arbitrary, positive semidefinite matrix, and $\sigma_s{}^2 > 0$ for all $1 \le s \le S$. Here, the key point is that the correlation structure, captured by $\mathbf{\Psi}$, is assumed to be the same across SNPs, whereas the scalar multiplier $\sigma_s{}^2$ is allowed to vary across SNPs. (Of course, this presumes that the same individuals are genotyped at all SNPs. When some individuals have missing genotypes at SNP $s$, the entries of $\mathbf{Y}^s$ and the rows and columns of $\mathbf{\Psi}$ that correspond to individuals with missing genotypes would be deleted.)

Note that $\mathbf{\Psi}$ and $\sigma_s{}^2$ are defined only up to a constant multiple, in the sense that $c\mathbf{\Psi}$ and $c^{-1}\sigma_s{}^2$ would give the same value of $\mathbf{\Sigma}_s$. By convention, $\sigma_s{}^2$ is usually chosen to be the variance of an outbred individual in the absence of population structure. We now give two examples of population genetic models that satisfy our assumptions.

## Example 1: Related Individuals without Additional Population Structure

An example of a simple model that satisfies the assumptions of the previous subsection is the model for related individuals in an unstructured population. In this model, individuals in the sample can be related by pedigrees, where the pedigree founders are assumed to be independently drawn from a population that is in Hardy-Weinberg equilibrium (HWE). Mendelian inheritance is assumed in the pedigrees. In this case, it has previously been shown[15] that Equations 9 and 10 hold, where $p_s$ is interpreted as the allele frequency of SNP $s$ in the population from which the founders are drawn, $\sigma_s^2 = 0.5p_s(1-p_s)$, and $\mathbf{\Psi} = \mathbf{\Phi}$, where $\mathbf{\Phi}$ is the kinship matrix given in Equation 4.

If the pedigrees are fully known, then the structure matrix $\mathbf{\Psi}$ is known, but if some genealogical information is missing, then $\mathbf{\Psi}$ might be partially or completely unknown.

## Example 2: Balding-Nichols Model with Admixture

In the Balding-Nichols model[6,28,29] with admixture, we let $p_s$ denote the "ancestral" allele frequency at SNP $s$ and let $q_k{}^s$ denote the allele frequency of SNP $s$ in subpopulation $k$, $1 \le k \le K$. We assume that the $q_k{}^s$ are random variables that are independent across both $k$ and $s$, with $q_k{}^s \sim \mathrm{Beta}(p_s(1-f_k)/f_k, (1-p_s)(1-f_k)/f_k)$, where $f_k \ge 0$ can be viewed as Wright's standardized measure of variation[30] for subpopulation $k$. For SNP $s$, let $\mathbf{q}^s = (q_1{}^s, \ldots, q_K{}^s)^T$ denote the vector of subpopulation-specific allele frequencies. Individual $i$ is assumed to have admixture vector $\mathbf{a}_i = (a_{i1}, \ldots, a_{iK})^T$, where $a_{ik} \ge 0$ for all $i$ and $k$, and $\sum_{k=1}^K a_{ik} = 1$ for all $i$. Conditional on the random variable $\mathbf{q}^s$, the two alleles of individual $i$ at SNP $s$ are assumed to be independent, identically distributed (i.i.d.) Bernoulli($\mathbf{a}_i{}^T\mathbf{q}^s$) random variables. In the Balding-Nichols model with admixture, Equations 9 and 10 hold, where $p_s$ is interpreted as the "ancestral" allele frequency, $\sigma_s^2 = 0.5p_s(1-p_s)$, and the entries of $\mathbf{\Psi}$ are given by $\Psi_{ii} = 1 + \sum_{k=1}^K a_{ik}^2 f_k$ and $\Psi_{ij} = 2\sum_{k=1}^K a_{ik}a_{jk}f_k$ if $i \ne j$. In this context, the population structure captured by $\mathbf{\Psi}$ would typically be unknown.

## Estimation of the Matrix $\mathbf{\Psi}$

The matrix $\mathbf{\Psi}$ is a function of the genealogy of the sampled individuals, where genealogy is broadly interpreted as including both population structure and the pedigree relationships of close relatives. The matrix $\mathbf{\Psi}$ will be unknown when there is hidden population structure and/or cryptic relatedness in the sample. We allow a completely general form for $\mathbf{\Psi}$, assuming only that it is positive semidefinite (psd). When genome-screen data are available on the sampled individuals, this information can be used to estimate $\mathbf{\Psi}$. For any pair of individuals $i$ and $j$, let $\mathcal{S}_{ij}$ be the set of markers for which both $i$ and $j$ have nonmissing genotype data. Then if the allele frequencies $p_s$ were known, and if $\sigma_s{}^2 = 0.5p_s(1-p_s)$ as in Examples 1 and 2, an unbiased estimator of $\Psi_{ij}$ would be

$$\tilde{\Psi}_{ij} = \frac{1}{|\mathcal{S}_{ij}|}\sum_{s\in\mathcal{S}_{ij}} \frac{\left(Y_i^s - p_s\right)\left(Y_j^s - p_s\right)}{.5p_s(1-p_s)} \qquad \text{(Equation 11),}$$

where $|\mathcal{S}_{ij}|$ is the number of elements of $\mathcal{S}_{ij}$. If one assumed, for example, that genotypes at different SNPs were independent with $|\mathcal{S}_{ij}| \to \infty$ and $p_s$ known and that $\sigma_s{}^2 = 0.5p_s(1-p_s)$ held at all but a finite number of SNPs, then Equation 11 would provide a consistent estimator of $\Psi_{ij}$. However, $p_s$ will generally not be known, so we propose to further restrict $\mathcal{S}_{ij}$ to those markers that are polymorphic in the sample; let $\hat{p}_s = \overline{Y}^s$, the observed proportion of type 1 alleles in the sample at marker $s$; and use estimator

$$\hat{\Psi}_{ij} = \frac{1}{|\mathcal{S}_{ij}|}\sum_{s\in\mathcal{S}_{ij}} \frac{\left(Y_i^s - \hat{p}_s\right)\left(Y_j^s - \hat{p}_s\right)}{.5\hat{p}_s(1-\hat{p}_s)} \qquad \text{(Equation 12).}$$

The estimator $\hat{\mathbf{\Psi}}$ of Equation 12 is essentially the same as the estimated covariance matrix used in EIGENSTRAT.[6] An alternative estimator could be obtained by using the sample variance of $\mathbf{Y}^s$ in the denominator instead of $0.5\hat{p}_s(1-\hat{p}_s)$.

If every sampled individual were genotyped at the same markers, with no missing genotypes, then $\tilde{\mathbf{\Psi}}$ would be psd and singular, with $\tilde{\mathbf{\Psi}}\mathbf{1} = 0$, i.e., $\mathbf{1}$ would be in the null space of $\tilde{\mathbf{\Psi}}$. With missing genotypes, it is possible for $\tilde{\mathbf{\Psi}}$ to be nonsingular and non-psd. The fact that $\tilde{\mathbf{\Psi}}$ might be non-psd is not, in itself, particularly problematic from a practical point of view, provided $\mathbf{V}^T\tilde{\mathbf{\Psi}}\mathbf{V} > 0$ for the chosen $\mathbf{V}$ in Equation 8 and assuming this provides a sufficiently accurate estimator of $\mathrm{Var}_0(\mathbf{V}^T\mathbf{Y}^s)/\sigma^2{}_s$. The fact that $\tilde{\mathbf{\Psi}}$ might be singular (e.g., in the case of no missing data) or close to singular, with $\tilde{\mathbf{\Psi}}$ orthogonal or approximately orthogonal to the vector $\mathbf{1}$, means that in those cases, one would not be able to directly plug $\tilde{\mathbf{\Psi}}$ into formulae such as Equations 2, 3, 6, and 7. This is discussed further in the next subsection. With substantially different amounts of missing data at different markers, as in the RA and COGA data sets analyzed in the Results section, the matrix $\tilde{\mathbf{\Psi}}$ might be nonsingular and so could be directly used in Equations 2, 3, 6, and 7.

### Estimation of $\sigma^2$ when Structure Is Unknown

In this subsection, we drop the subscript $s$ and use notations $\mathbf{Y}$, $p$, and $\sigma^2$ for the SNP being tested; e.g., we assume $E_0(\mathbf{Y}) = p\mathbf{1}$ and $\mathrm{Var}_0(\mathbf{Y}) = \boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Psi}$. As we did for the case of known structure, we consider two general approaches to estimation of $\sigma^2$ when structure is unknown. The first approach is to take estimators of the form $\hat{\sigma}_1^2 = 0.5\hat{p}(1-\hat{p})$, where $\hat{p}$ is a suitable estimator of $p$. When $\hat{\boldsymbol{\Psi}}$ is orthogonal or approximately orthogonal to the vector $\mathbf{1}$, we cannot plug it into Equation 2 to obtain the BLUE of $p$. (Note that use of the Moore-Penrose generalized inverse $\hat{\boldsymbol{\Psi}}^-$ in place of $\hat{\boldsymbol{\Psi}}^{-1}$ also does not work, because $\hat{\boldsymbol{\Psi}}^-$ is also orthogonal or approximately orthogonal to $\mathbf{1}$, so plugging into Equation 2 would result in both numerator and denominator being exactly or approximately zero.) Instead, we use the more stable estimator $\hat{p} = \overline{Y}$, the sample allele frequency. Thus, our first estimator becomes

$$\hat{\sigma}_1^2 = 0.5\overline{Y}(1-\overline{Y}) \qquad \text{(Equation 13)}.$$

As we did in the case of known structure, we also consider an estimator of $\sigma^2$ that does not assume $\sigma^2 = 0.5p(1-p)$ at the SNP being tested. When $\hat{\boldsymbol{\Psi}}$ is orthogonal or approximately orthogonal to the vector $\mathbf{1}$, we replace Equation 3 by

$$\hat{\sigma}_2^2 = (n-1)^{-1}\mathbf{Y}^T\hat{\boldsymbol{\Psi}}^-\mathbf{Y} \qquad \text{(Equation 14)},$$

where $\hat{\boldsymbol{\Psi}}^-$ is the Moore-Penrose generalized inverse of $\hat{\boldsymbol{\Psi}}$.

### Association Tests when Structure Is Partially or Completely Unknown

We apply Equation 8 to extend association tests developed for the situation of known structure (as described in subsection Examples of Association Tests with Known Structure) to association tests that are appropriate for situations of partially or completely unknown structure. We call the tests based on Equation 8 the ROADTRIPS versions of the corresponding tests given by Equation 1, and we now give several examples. Table 1 gives the weight vector for each statistic defined below.

*$R\chi_1$ and $R\chi_2$, the ROADTRIPS Versions of the Corrected $\chi^2$ and Corrected Armitage Tests*

To extend the corrected $\chi^2$ and corrected Armitage tests to the situation of unknown structure, we apply Equation 8 with $\mathbf{V}$ given in Equation 5 and $\hat{\boldsymbol{\Psi}}$ given in Equation 12. We define $R\chi_1$ to be the ROADTRIPS version of the corrected $\chi^2$ test, where this is obtained by using $\hat{\sigma}_1^2$ given in Equation 13, and we define $R\chi_2$ to be the ROADTRIPS version of the corrected Armitage test, where this is obtained by using $\hat{\sigma}_2^2$ given in Equation 14. When all SNPs have the same pattern of missing genotypes, we expect $R\chi_2$ to perform similarly to GC, because in this case, both $R\chi_2$ and GC are equivalent to correcting all the Armitage $\chi^2$ statistics across the genome by a common factor, though this factor differs between the two methods. However, when different SNPs have different rates of missing genotypes, we expect the $R\chi_2$ statistic to do better than GC, in terms of both type 1 error and power, because the $R\chi_2$ statistic allows different SNPs to have different correction factors appropriate to the level of genotype information available, whereas GC applies the same correction factor to all SNPs.

*RM, the ROADTRIPS Version of $M_{QLS}$ when Structure Is Partially Known*

As the $M_{QLS}$ is generally the most powerful of the statistics for complex trait mapping when structure is known,[16] we expect the ROADTRIPS version of $M_{QLS}$ to be powerful when structure is unknown. We consider separately the cases of partially known structure and completely unknown structure. An example of partially known structure occurs when reliable pedigree information on sampled individuals is available, but one wants to allow for the possibility of additional cryptic relatedness or unknown population structure not captured by the pedigree information. In the context of partially known structure, we compute the matrix $\boldsymbol{\Phi}$ of Equation 4 as a function of the known pedigree information. At the same time, we also calculate the estimator $\hat{\boldsymbol{\Psi}}$ in Equation 12 as before, with the expectation that it will capture the full structure in the data, including structure not explained by $\boldsymbol{\Phi}$. Then we obtain *RM*, the ROADTRIPS version of the $M_{QLS}$ test when structure is partially known, by applying Equation 8 with $\mathbf{V}$ given in Equation 7. The idea is that we create a powerful test by using the known pedigree structure given in $\boldsymbol{\Phi}$ to obtain weights $\mathbf{V}$ that will be optimal[16] when $\boldsymbol{\Psi} = \boldsymbol{\Phi}$. Then we preserve the validity of the test in the presence of additional structure, not captured by $\boldsymbol{\Phi}$, through use of the estimator $\hat{\boldsymbol{\Psi}}$ in the denominator of Equation 8. As we did for the $R\chi$ test, we could add subscripts 1 and 2 to the name of the test to distinguish the use of estimators $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, given in Equations 13 and 14, respectively.

*$RM_{NI}$, the ROADTRIPS Version of $M_{QLS}$ when Structure Is Completely Unknown*

For the case of completely unknown structure, we define the $RM_{NI}$ test, which is a ROADTRIPS version of the $M_{QLS}$, where "NI" stands for "no information." We form $RM_{NI}$ from Equation 8, where we take $\mathbf{V} = \mathbf{A}_N - \overline{A}_N\mathbf{1}$, where $\overline{A}_N$ is the sample average of the elements of $\mathbf{A}_N$. This choice of $\mathbf{V}$ is the natural analog to Equation 7 when $\hat{\boldsymbol{\Psi}}$ is used in place of $\boldsymbol{\Phi}$, for the case when $\hat{\boldsymbol{\Psi}}$ is orthogonal to the $\mathbf{1}$ vector and where we ignore the contribution of $\mathbf{A}_M$. The reason we ignore the contribution of $\mathbf{A}_M$ is that the expected gain by including this term for individuals not known to be related is not high enough to justify the computational cost involved in obtaining the inverse or generalized inverse of $\hat{\boldsymbol{\Psi}}$. (Note that $\hat{\boldsymbol{\Psi}}$ tends to be much more costly to invert than $\boldsymbol{\Phi}$, because in typical applications, $\boldsymbol{\Phi}$ is block-diagonal with small blocks.) We could add subscripts 1 and 2 to the name of the test to distinguish the use of estimators $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ given in Equations 13 and 14, respectively. Note that when the amount of missing data varies across SNPs, the matrix $\hat{\boldsymbol{\Psi}}$ might be nonsingular, and there is the possibility of using Equation 7 with $\hat{\boldsymbol{\Psi}}$ plugged in for $\boldsymbol{\Phi}$. For instance, this occurs in both the data sets we analyze in Results. In this case, one might still choose to ignore the information provided by $\mathbf{A}_M$ for the computational reasons mentioned.

*$RW_{NI}$, the ROADTRIPS Version of $W_{QLS}$ when Structure Is Completely Unknown*

We form $RW_{NI}$ from Equation 8, where we take $\mathbf{V} = \hat{\boldsymbol{\Psi}}^-\mathbf{1}_c$, which is the natural analog to Equation 6 for the case when $\hat{\boldsymbol{\Psi}}$ is orthogonal to the $\mathbf{1}$ vector. If we use estimator $\hat{\sigma}_2^2(n-1)/(n-2)$ of $\sigma^2$, then we obtain the test recently proposed by Rakovski and Stram.[13] (In our simulation study, we actually use estimator $\hat{\sigma}_2^2$ instead of $\hat{\sigma}_2^2(n-1)/(n-2)$, but the difference is completely negligible for the size of $n$ we consider.)

**Table 1. Weight Vectors V for the ROADTRIPS Statistics**

| Statistic | V |
|---|---|
| $R\chi$ | $\mathbf{1}_c - \frac{n_c}{n}\mathbf{1}$ |
| $RM$ | $\mathbf{A}_N + \boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}_{N,M}\mathbf{A}_M - (\mathbf{A}_N + \boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}_{N,M}\mathbf{A}_M)^T\mathbf{1}(\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1})^{-1}\boldsymbol{\Phi}^{-1}\mathbf{1}$ |
| $RM_{NI}$ | $\mathbf{A}_N - \overline{A}_N\mathbf{1}$ |
| $RW_{NI}$ | $\hat{\boldsymbol{\Psi}}^-\mathbf{1}_c$ |

$\hat{\boldsymbol{\Psi}}^-$ is the generalized inverse of $\hat{\boldsymbol{\Psi}}$ and $\overline{A}_N$ is the average of the elements of $\mathbf{A}_N$.

### GAW 15 Rheumatoid Arthritis Data

We apply ROADTRIPS to perform association analysis of RA data provided for GAW 15 by a UK group led by Jane Worthington and Sally John (these data are described in detail elsewhere).[22] Data are available on 157 nuclear families, where 156 of these have at least two affected individuals. Individuals were diagnosed as affected according to the American College of Rheumatology (ACR) criteria. There are 550 individuals with available genotype data. After exclusion of 2 duplicate individuals and 4 outlier individuals who have estimated inbreeding coefficients more than 3 standard deviations (SDs) above the average (where the estimated inbreeding coefficient of individual $i$ is taken to be $\hat{\Psi}_{ii} - 1$), there are 339 affected individuals, 198 unaffected controls, and 7 controls of unknown phenotype in the analysis. The data set includes 10,156 autosomal SNPs that passed quality control filters. We exclude 285 SNPs that are not polymorphic (minor allele frequency less than 0.01). The remaining 9871 SNPs are tested for association.

### GAW 14 COGA Data

We apply ROADTRIPS to identify SNPs associated with alcohol dependence in data provided by the COGA for GAW 14.[23] These data were previously analyzed with association methods that assume known structure.[16] There are a total of 1614 individuals from 143 pedigrees, with each pedigree containing at least three affected individuals. We include in our analysis only those individuals who are coded as "white, non-Hispanic." We designate as cases those individuals who are affected with ALDX1 or who have symptoms of ALDX1, where ALDX1 is defined to be DSM-III-R alcohol dependence with the Feighner Alc Definite phenotype. By these criteria, there are 830 cases with available SNP data. We designate as "unaffected controls" those individuals who are labeled as "pure unaffected," and we designate as "controls of unknown phenotype" those individuals who are labeled as "never drank alcohol." Among individuals with available SNP data, these criteria result in 187 unaffected controls and 13 unknown controls. The data set includes 10,810 autosomal SNPs. We exclude 403 SNPs that are not polymorphic and analyze the remaining 10,407 SNPs.

## Results

### Simulation Studies

We perform simulation studies, in which population structure and related individuals are simultaneously present in the case-control sample, in order to compare the performance of ROADTRIPS to that of previously proposed association methods that correct in some way for either population structure or related individuals or both. The methods to which we compare ROADTRIPS are GC, FBAT, the method of Rakovski and Stram,[13] EIGENSTRAT, $M_{QLS}$, and the corrected Armitage $\chi^2$ test. We also include the standard (uncorrected) Armitage test in the type 1 error study. We simulate four different settings of population structure, including admixture, and two different settings of relationship configuration, where the latter refers to pedigree relationships among sampled individuals.

### Relationship Configurations

Both relationship configurations 1 and 2 include 100 unrelated affected individuals, 400 unrelated unaffected individuals, and individuals sampled from 120 outbred, three-generation pedigrees, where each pedigree has a total of 16 individuals, the phenotypes of all individuals in the pedigrees are observed, and genotypes are observed for only a subset of individuals in each pedigree. We sample six types of these pedigrees; the types are described in Table 2. Relationship configuration 1 has 40 pedigrees of type 1, 40 of type 2, and 40 of type 3, as well as 100 unrelated affected and 400 unrelated unaffected individuals. Relationship configuration 2 contains all six types of pedigrees in Table 2, with 20 of each type, and it also contains 100 unrelated affected and 400 unrelated unaffected individuals.

**Table 2. Pedigree Configuration Types Used in Simulations**

| Type | $N_{af}$ | $N_{un}$ | Genotyped Individuals |
|---|---|---|---|
| 1 | 4 | 12 | Unaffected sib pair and their unaffected first cousin |
| 2 | 5 | 11 | 1 affected parent, 2 affected offspring |
| 3 | 6 | 10 | 1 aff. parent, 2 aff. offspr., unaff. sib pair who are 1st cousins to the latter |
| 4 | 4 | 12 | 1 affected parent with 2 affected and 1 unaffected offspring |
| 5 | 5 | 11 | 1 affected and 2 unaffected sibs, unaffected aunt and her affected spouse |
| 6 | 6 | 10 | 1 aff. and 1 unaff. parent with 2 unaff. offspr., 2 other affecteds |

$N_{af}$ and $N_{un}$ are the total numbers of affected and unaffected individuals in the pedigree, respectively, among whom only the indicated individuals are genotyped.

### Population Structure Settings

Each simulation setting specifies a particular relationship configuration combined with a particular setting of population structure. Each setting of population structure is a special case of the Balding-Nichols model with admixture described in Example 2, in which we take $f_k = 0.01$ for every subpopulation. Population structure 1 has individuals sampled from two subpopulations, with 60% of the pedigrees and affected unrelated individuals sampled from subpopulation 1 and the remaining 40% of the pedigrees and affected unrelated individuals sampled from subpopulation 2. Among the unrelated unaffecteds, 40% are sampled from subpopulation 1 and 60% from subpopulation 2. Population structure 2 is similar to population structure 1, except that the proportions 60% and 40% are replaced by 80% and 20%, respectively. Population structure 3 is similar to population structure 1, except that there are three subpopulations, with all of the unrelated unaffecteds sampled from subpopulation 3. Population structure 4 has individuals sampled from an admixed population, formed from two subpopulations. Individuals in the admixed population are assumed to have i.i.d. admixture vectors of the form $(a, 1 - a)$, where $a$ is a Uniform(0,1) random variable.

**Table 3. Empirical Type 1 Error, at Level 0.0001, in the Presence of Both Related Individuals and Population Structure**

| | Empirical Type 1 Error of Tests[a] | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Setting[b]** | $R\chi$ **or** $RM_{NI}$ | $RM$ | $RW_{NI}$ | **GC** | **EIG** | $M_{QLS}$ | **Corr Arm** | **Arm** |
| (1,1) | 0.00009 | 0.00004 | 0.00011 | 0.00007 | **0.00118** | **0.00027** | **0.00018** | **0.00202** |
| (1,2) | 0.00011 | 0.00012 | 0.00011 | 0.00010 | **0.00059** | **0.00043** | 0.00016 | **0.00073** |
| (2,1) | 0.00008 | 0.00012 | 0.00010 | 0.00004 | **0.00116** | **0.01690** | **0.00485** | **0.01970** |
| (2,2) | 0.00013 | 0.00015 | 0.00012 | 0.00005 | **0.00054** | **0.02281** | **0.00723** | **0.01375** |
| (3,1) | 0.00010 | 0.00010 | 0.00010 | 0.00007 | **0.00155** | **0.06752** | **0.02409** | **0.06094** |
| (3,2) | 0.00015 | 0.00012 | 0.00008 | 0.00008 | **0.00058** | **0.08464** | **0.03189** | **0.05028** |
| (4,1) | 0.00007 | 0.00011 | 0.00012 | 0.00007 | **0.00086** | **0.00056** | **0.00032** | **0.00277** |
| (4,2) | 0.00008 | 0.00007 | 0.00012 | 0.00012 | **0.00036** | **0.00028** | **0.00019** | **0.00069** |

Empirical type 1 error rates are calculated based on 100,000 simulated random SNPs. Rates that are significantly different from the nominal 0.0001 level are in bold. $R\chi$ and $RM_{NI}$ are equivalent statistics for all the settings shown. For the ROADTRIPS statistics, the $\hat{\sigma}_2^2$ of Equation 14 is used.
[a] Abbreviations of test names are genomic control (GC), EIGENSTRAT with outlier removal (EIG), corrected Armitage $\chi^2$ (Corr Arm), and Armitage trend test (Arm).
[b] Setting (*i*, *j*) denotes population structure setting *i* and relationship configuration setting *j*.

Within a given setting of population structure, to simulate the unrelated individuals needed in relationship configurations 1 and 2, we first simulate genotypes according to the chosen setting of population structure, simulate phenotypes conditional on genotypes, and then randomly ascertain 100 affected and 400 unaffected individuals. To sample particular pedigree types within a given setting of population structure, we first simulate genotypes for pedigree founders according to the chosen setting of population structure, drop alleles down the pedigree, simulate phenotypes conditional on genotypes, and then do rejection sampling to obtain 100 replicates of each of the pedigree configuration types. Then, in the simulations, pedigrees of each type are sampled with replacement from the 100 previously obtained replicates of that pedigree type.

### Random and Causal SNPs

We use a trait model that has two unlinked causal SNPs (which we call "SNP 1" and "SNP 2") with epistasis between them.[16] The ancestral frequencies of the type 1 alleles at SNPs 1 and 2 are taken to be 0.1 and 0.5, respectively. Individuals with at least one copy of allele 1 at SNP 1 and at least one copy of allele 1 at SNP 2 have a penetrance of 0.3. All other individuals have a penetrance of 0.05. In the power studies, association is tested with SNP 2. In contrast, "random" SNPs are assumed to be unlinked and unassociated with the trait, and their ancestral allele frequencies are obtained as i.i.d. draws from a uniform (0.1, 0.9) distribution.

### Assessment of Type 1 Error

For each of the eight combinations of population structure and relationship configuration, we generate genotype data for 100,000 random SNPs that are neither linked nor associated with the trait, and we test each of them for association at the 0.0001 level, using various test statistics. In Table 3, for each test statistic, we report the empirical type 1 error, which we calculate as the proportion of simulations in which the test statistic exceeds the ${\chi_1}^2$ quantile corresponding to nominal type 1 error level 0.0001. The statistics compared are the four ROADTRIPS statistics, $R\chi_2$, $RM_2$, $RM_{NI2}$, and $RW_{NI2}$, where the subscript "2" denotes use of the estimator $\hat{\sigma}_2^2$ of Equation 14; GC; EIGENSTRAT; the $M_{QLS}$; the corrected Armitage $\chi^2$; and the uncorrected Armitage trend test. Note that because there are only two types of controls, the $R\chi_2$ and $RM_{NI2}$ tests are identical. The method of Rakovski and Stram[13] corresponds to the ROADTRIPS statistic $RW_{NI2}$, which is included in the comparison. The correct type 1 error of FBAT has been established previously.[31] Using an exact binomial calculation, we determine that empirical type 1 error rates falling in the range of 0.00004–0.00016 are not significantly different from the nominal 0.0001 level.

For GC and all of the ROADTRIPS statistics, empirical type 1 error is not significantly different from the nominal level. In contrast, type 1 error is inflated for EIGENSTRAT, $M_{QLS}$, the corrected Armitage $\chi^2$, and the Armitage trend test. This is to be expected, because these tests either (1) correct for related individuals but not for population structure ($M_{QLS}$, corrected Armitage $\chi^2$), (2) correct for population structure but not for related individuals (EIGENSTRAT), or (3) correct for neither (Armitage trend test). In particular, the top principal components in EIGENSTRAT are not able to capture the complicated covariance structure due to the related individuals in the samples. The results for EIGENSTRAT in Table 3 are obtained with the default setting of ten principal components and with outlier removal. The results without outlier removal and with different numbers of principal components are similar (results not shown). We also performed all the ROADTRIPS tests with variance estimator $\hat{\sigma}_1^2$ of Equation 13 instead of $\hat{\sigma}_2^2$ and obtained nearly identical empirical type 1 error rates (results not shown).

**Table 4. Empirical Type 1 Error at Level 0.0001 when Genotypes are Missing at Random in the Presence of Both Related Individuals and Population Structure**

| | | Empirical Type 1 Error | |
|---|---|---|---|
| **Population Structure** | **Relationship Configuration** | $R\chi_2$ | **Genomic Control** |
| 2 | 1 | 0.00008 | **0.00031** |
| 2 | 2 | 0.00007 | **0.00036** |
| 3 | 1 | 0.00006 | **0.00057** |
| 3 | 2 | 0.00013 | **0.00058** |

Empirical type 1 error rates are calculated based on 100,000 simulated random SNPs. Rates that are significantly different from the nominal 0.0001 level are in bold.

**Table 5. Power to Detect Association in the Presence of Both Related Individuals and Population Structure**

| | Power (Standard Error) | | | | |
|---|---|---|---|---|---|
| **Population Structure** | $R\chi$ **or** $RM_{NI}$ | ***RM*** | $RW_{NI}$ | **GC** | **FBAT** |
| 1 | 0.79 (0.006) | **0.94 (0.003)** | 0.59 (0.007) | 0.78 (0.006) | 0.0012 (0.0005) |
| 2 | 0.42 (0.007) | **0.48 (0.007)** | 0.36 (0.007) | 0.43 (0.007) | 0.0016 (0.0006) |
| 4 | 0.70 (0.007) | **0.80 (0.006)** | 0.48 (0.007) | 0.70 (0.007) | 0.0002 (0.0002) |

Power is assessed at significance level 0.0001 on the basis of 5000 simulated replicates. The highest power for each simulation setting is in bold. Relationship configuration 2 is used in each case. $R\chi$ and $RM_{NI}$ are equivalent statistics for all the settings shown. For the ROADTRIPS statistics, the $\hat{\sigma}_2^2$ of Equation 14 is used.

### Type 1 Error with Missing Genotypes: $R\chi_2$ and GC

When all SNPs have the same pattern of missing genotypes, we expect $R\chi_2$ to perform similarly to GC, because in this case, both $R\chi_2$ and GC are equivalent to correcting all the Armitage $\chi^2$ statistics across the genome by a common factor, though this factor differs between the two methods. However, when different SNPs have different rates of missing genotypes, we expect the $R\chi_2$ statistic to have better control of type 1 error than GC, because the $R\chi_2$ statistic allows different SNPs to have different correction factors appropriate to the level of genotype information available, whereas GC applies the same correction factor to all SNPs.

To assess the magnitude of this effect, we perform a simulation study under four different settings of population structure and relationship configuration, given in columns 1 and 2 of Table 4. For each combination of settings, genotype data are generated for 100,000 random SNPs that are neither linked nor associated with the phenotype. The proportions of individuals with missing genotype data at different SNPs are taken to be i.i.d. random variables drawn from a Beta(3, 12) distribution, which has a mean of 0.2 and a SD of 0.1. Given the proportion of missing genotypes at a marker, the individuals whose genotypes will be set to missing for that marker are chosen uniformly at random from the sample.

The empirical type 1 error rates for $R\chi_2$ and GC are given in Table 4. GC has inflated type 1 error for all of the simulation settings, because of undercorrection of test statistics from SNPs that have relatively low levels of missing genotypes, whereas the empirical type 1 error for $R\chi_2$ is not significantly different from the nominal 0.0001 level. The results illustrate that ROADTRIPS is not only robust to cryptic population and pedigree structure, but also to varying rates of randomly missing genotype data. This results from the fact that the entire empirical covariance matrix is estimated in ROADTRIPS, so when some individuals have missing genotype data at the SNP being tested, the corresponding rows and columns can be deleted from the empirical covariance matrix, allowing one to obtain a variance estimator that accounts for missing genotype data at the marker being tested.

### Power Comparison

We assess power to detect association in the presence of population and pedigree structure only for those tests that maintain correct nominal type 1 error, namely the four ROADTRIPS tests, FBAT, and GC (although, as can be seen in Table 4, the type 1 error of GC might not be correct when the rate of missing genotype data varies across markers). The simulations are performed with relationship configuration 2 under each of the four different settings of population structure. For each setting, 5000 simulated replicates are performed, and SNP 2 of the trait model is tested for association with the trait by each method. Table 5 reports power for each statistic for settings 1, 2, and 4 of population structure. Here, power is calculated as the proportion of simulations for which the statistic exceeds the $\chi_1^2$ quantile corresponding to nominal type 1 error level 0.0001. Power for population structure 3 is close to 0 for all of the statistics (data not shown). The ROADTRIPS statistics are all calculated with estimator $\hat{\sigma}_2^2$ of Equation 14. The FBAT and *RM* methods are given the information of the correct pedigree structure, but not the population structure. The FBAT statistic is calculated with offset value set equal to the prevalence, and the *RM* statistic is calculated with $k$ set equal to the prevalence. As expected, the *RM* test is the most powerful in all settings, because it uses the known pedigree information to incorporate phenotype information about relatives with missing genotype data, and it corrects for additional unknown population structure by means of the empirical covariance matrix. In contrast, the FBAT test, which was given all the same information as the *RM* test, performs very poorly in these simulations, because it is not able to incorporate the data on the 500 unrelated individuals, and it is also not able to incorporate the data from pedigree types 1, 2, and 3, because they do not meet the FBAT criteria for "informative families." If we assume that no pedigree information is available, then the $R\chi$, $RM_{NI}$, and GC tests all give identical or almost identical power, assuming that all markers have comparable amounts of missing genotype data. When different SNPs have different amounts of

**Table 6. Power to Detect Association with Related Individuals in the Absence of Population Structure**

| Power (Standard Error) | | | | | | |
|---|---|---|---|---|---|---|
| $R\chi$ *or* $RM_{NI}$ | *RM* | $RW_{NI}$ | GC | FBAT | $M_{QLS}$ | Corr Arm |
| 0.90 (0.004) | **0.98 (0.002)** | 0.81 (0.006) | 0.91 (0.004) | 0.0002 (0.0002) | **0.98 (0.002)** | 0.90 (0.004) |

Power is assessed at significance level 0.0001 on the basis of 5000 simulated replicates of relationship configuration 2 with no population structure. The highest power is in bold. $R\chi$ and $RM_{NI}$ are equivalent statistics in this simulation setting. For the ROADTRIPS statistics, the $\hat{\sigma}_2^2$ of Equation 14 is used. Abbreviations of test names are genomic control (GC) and corrected Armitage $\chi^2$ (Corr Arm).

missing genotype data, GC might not adequately control type 1 error, so $R\chi$ and $RM_{NI}$ would be preferred. The $RW_{NI}$ test, which corresponds to the test of Rakovski and Stram,[13] has lower power than $R\chi$ and $RM_{NI}$, which is not surprising in light of the fact that the $W_{QLS}$ test, of which the $RW_{NI}$ is an extension, was shown[16] to have generally lower power than the $M_{QLS}$ and corrected $\chi^2$ tests, of which the $RM_{NI}$ and $R\chi$, respectively, are extensions.

We also assess power to detect association when there is pedigree structure but no population structure. The simulation is carried out as in the preceding paragraph except that relationship configuration 2 with no population structure is simulated. In addition to the tests compared in Table 5, we also calculate power for the $M_{QLS}$ and corrected Armitage $\chi^2$ tests, both of which have correct type 1 error in this setting, provided that the pedigree structure is known. Estimated power for this setting is given in Table 6. The $M_{QLS}$ and *RM* tests are the most powerful among the tests considered. The *RM* test is able to match the power of $M_{QLS}$ even though *RM* uses the estimator $\hat{\boldsymbol{\Psi}}$ in the variance calculation, whereas $M_{QLS}$ uses the true $\boldsymbol{\Psi}$. As in the previous power comparison, FBAT has almost no power in this setting, and the $RW_{NI}$ test has lower power than all of the other statistics except FBAT. There is no significant difference in power among $R\chi$, $RM_{NI}$, GC, and the corrected Armitage $\chi^2$ tests. The $M_{QLS}$ and the corrected Armitage $\chi^2$ tests have power identical to their corresponding ROADTRIPS extensions, *RM* and $R\chi$, which illustrates that power is not compromised by use of the empirical matrix $\hat{\boldsymbol{\Psi}}$ in the variance correction for the ROADTRIPS statistics.

### GAW 15 Rheumatoid Arthritis Data

We apply *RM*, $R\chi$, GC, and FBAT to the GAW 15 data to test for association of SNPs with RA. Using a previously reported[22] prevalence of 0.8% for RA in people of European descent, we set both the offset value in FBAT and the prevalence value in *RM* to 0.008. The entries of the empirical covariance matrix $\hat{\boldsymbol{\Psi}}$ and the correction factor for GC are calculated using SNP data across the autosomal chromosomes for the study individuals. Table 7 gives the results of all tests for those SNPs for which at least one of the tests has a nominal p value $<5.0 \times 10^{-5}$. For 8 of these 10 SNPs, the *RM* test has the smallest p value among the four tests used. After Bonferroni correction to adjust for four different tests of association at each of 9,871 SNPs, the *RM* test is significant at the 5% level for 2 SNPs: snp264363 on chromosome 11 ($p = 5.2 \times 10^{-7}$ uncorrected, 0.021 corrected) and snp152076 on chromosome 9 ($p = 5.4 \times 10^{-7}$ uncorrected, 0.021 corrected). The $R\chi$ test is significant at the 5% level for an additional SNP, snp547632 on chromosome 11 ($p = 1.2 \times 10^{-6}$ uncorrected, 0.047 corrected).

**Table 7. Rheumatoid Arthritis Data Results: SNPs with p Value $<0.00005$ for at Least One of the Four Tests**

| | | | | | | p Value | | | |
|---|---|---|---|---|---|---|---|---|---|
| Chr | Marker | Pos. (cM) | $N_{CA}$ | $N_{CO}$ | *p* | $R\chi$ | *RM* | GC | FBAT |
| 11 | snp264363 | 105.57721 | 208 | 133 | 0.86 | 4.7e-3 | 5.2e-7[a] | 1.7e-2 | 1.2e-2 |
| 9 | snp152076 | 77.047858 | 267 | 162 | 0.92 | 3.0e-2 | 5.4e-7[a] | 1.1e-1 | 9.9e-3 |
| 11 | snp547632 | 63.954804 | 230 | 135 | 0.80 | 1.2e-6[a] | 6.2e-2 | 1.8e-4 | 3.0e-1 |
| 3 | snp151721 | 114.80384 | 279 | 174 | 0.84 | 2.8e-4 | 2.2e-6 | 6.8e-4 | 3.0e-3 |
| 18 | snp511091 | 107.60685 | 219 | 130 | 0.57 | 4.2e-2 | 1.1e-5 | 6.6e-2 | 6.6e-2 |
| 14 | snp51741 | 68.2942 | 298 | 188 | 0.95 | 1.4e-3 | 1.2e-5 | 1.3e-3 | 1.5e-2 |
| 15 | snp66639 | 51.864772 | 296 | 187 | 0.97 | 1.5e-5 | 1.0e-2 | 1.4e-4 | NA |
| 3 | snp71651 | 40.817259 | 188 | 118 | 0.74 | 3.5e-2 | 1.8e-5 | 6.3e-2 | 3.6e-2 |
| 4 | snp570108 | 66.453602 | 224 | 142 | 0.71 | 5.1e-2 | 3.9e-5 | 1.2e-1 | 2.4e-1 |
| 8 | snp261673 | 67.427701 | 192 | 131 | 0.61 | 2.9e-1 | 4.6e-5 | 4.0e-1 | 3.7e-1 |

The chromosome (Chr), the name of the marker (Marker), the position of the marker on the chromosome (Pos.), the number of genotypes available in cases ($N_{CA}$) and controls ($N_{CO}$), and the major allele frequency in the case-control sample as a whole (*p*) are displayed. An insufficient number of informative families for the FBAT analysis are indicated by NA.
[a] Genome-wide significance after Bonferroni correction.

A histogram of the estimated self-kinship values (where the estimated self-kinship value of individual *i* is taken to be $.5\hat{\Psi}_{i,i}$) of the individuals included in the analysis can be found in Figure 1. The histogram shows that the values are not centered around 0.5, which is the self-kinship value in the absence of population structure and inbreeding. The self-kinship mean is 0.512, and the majority of the kinship values (77%) are greater than 0.5. There are a few pairs of individuals that should have a kinship coefficient value equal to 0.25 based on the available pedigree information (i.e., parent-offspring pairs and sibling pairs), but have estimated kinship coefficient values close to 0 and thus appear to be unrelated. There are also a few pairs that are not members of the same pedigree but have kinship coefficient estimates that indicate that they are related. Both these phenomena could be caused by sample switches. An attractive feature of the ROADTRIPS methods is that they automatically correct for misspecified relationships in a sample, in addition to hidden population structure, while simultaneously allowing for different SNPs to have different rates of missing genotypes.

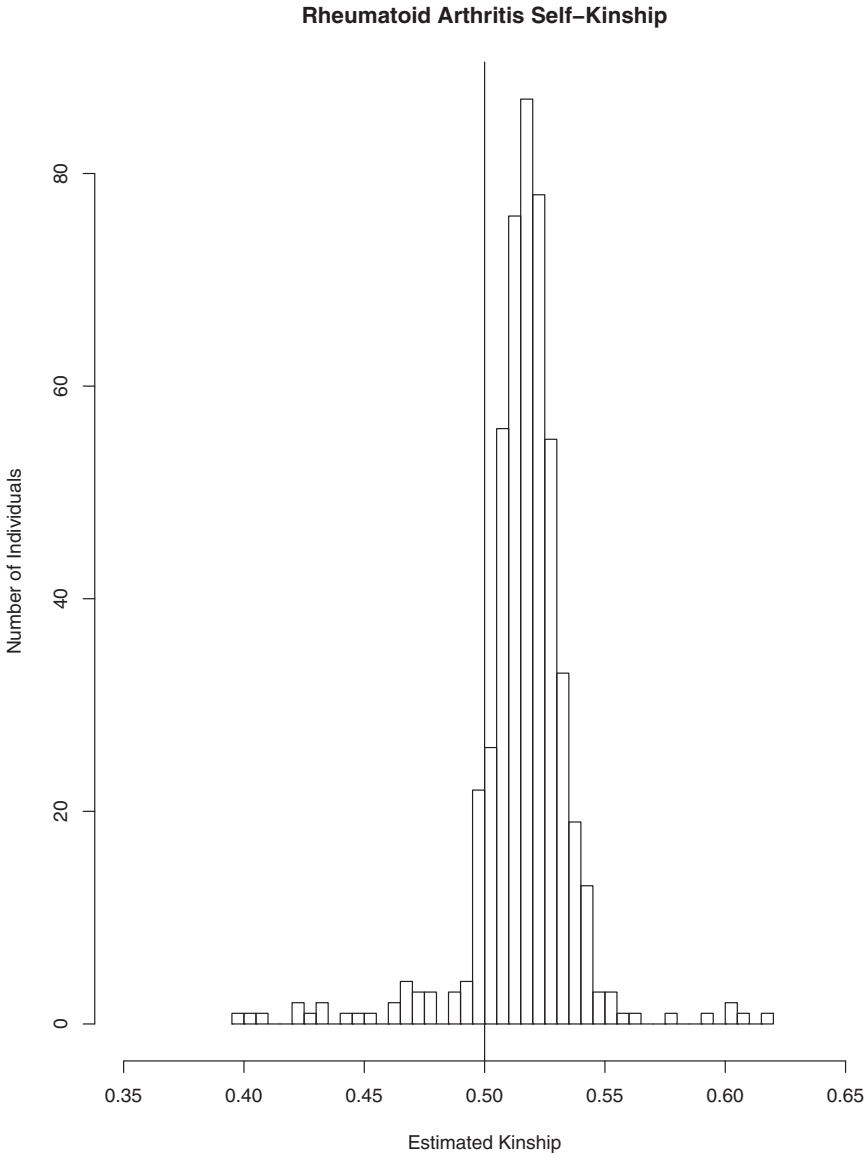


**Figure 1. Histogram of Estimated Self-Kinship Coefficient Values for the GAW 15 UK RA Data**
The vertical line at 0.5 represents the self-kinship value in the absence of population structure and inbreeding.

($p = 1.5 \times 10^{-7}$ uncorrected, 0.005 corrected), and tsc1637642 on chromosome 5 ($p = 6.8 \times 10^{-7}$ uncorrected, 0.02 corrected). The $R\chi$ test is significant at the 5% level for an additional SNP, tsc0571038 on chromosome 11 ($p = 8.0 \times 10^{-7}$ uncorrected, 0.025 corrected). Of the 5 SNPs that were previously[16] identified as genome-wide significant by $M_{QLS}$ or $W_{\chi^2_{corr}}$, 4 were also identified by ROADTRIPS. The exception is tsc0057290 on chromosome 18, which was identified by $M_{QLS}$ and is no longer significant after correcting for cryptic structure. ROADTRIPS identifies an additional SNP, tsc1637642 on chromosome 5, which was not identified in the previous analysis. GC did not identify any significant SNPs. There are some SNPs in Table 8 that are in or near genes of interest; the details have previously been reported.[16] A previous analysis[32] of these data with FBAT, in which a slightly larger sample of individuals was used, detected one SNP with nominal p value $6 \times 10^{-5}$, which is not significant after Bonferroni correction for the number of SNPs tested.

### GAW 14 COGA Data

We apply *RM*, $R\chi$, and GC to test for association with an alcoholism-related phenotype in the GAW 14 COGA data. A previous analysis[16] of these data used the $M_{QLS}$ (with $k$ set to 0.05) and corrected $\chi^2$ tests; these tests correct for known pedigree information, but do not make any correction for unknown structure. In our reanalysis, we compare the results of the tests with and without the correction for unknown structure. To make the results comparable, we set $k = 0.05$ in the *RM* test. Table 8 lists SNPs for which at least one of the *RM*, $R\chi$, and GC tests has a nominal p value $<1.0 \times 10^{-5}$. For 11 of these 12 SNPs, the *RM* test has the smallest p value among the three tests used. After Bonferroni correction to adjust for three different tests of association at each of 10,407 SNPs, the *RM* test is significant at the 5% level for 4 SNPs: tsc1750530 on chromosome 16 ($p = 2.3 \times 10^{-8}$ uncorrected, 0.0007 corrected), tsc0046696 on chromosome 18 ($p = 1.4 \times 10^{-7}$ uncorrected, 0.004 corrected), tsc1177811 on chromosome 1

Figure 2 gives a histogram of the estimated self-kinship coefficient values for the genotyped individuals who were included in the analysis. The center of the histogram is shifted from 0.5 (the self-kinship coefficient in the absence of population structure and inbreeding). Seventy-one percent of the values are greater than 0.5, and the mean self-kinship value is 0.506. Just as there were in the UK RA data, in the COGA data there are a few pairs of individuals who appear to have misspecified relationships or be cryptically related. As previously mentioned, ROADTRIPS adjusts for this in the variance correction.

### Assessment of Computation Time

Using a single processor on a shared machine with eight quad-core AMD Opteron 8384 25 GHz processors with 64 GB RAM, analysis of 10,156 SNPs from the RA data and 10,810 SNPs from the COGA data with four tests ($R\chi$, *RM*, GC, and a ROADTRIPS version of $W_{QLS}$) took approximately

**Table 8. COGA Data Results: SNPs with p Value <0.00001 for at Least One of the Three Tests**

| | | | | | | p Value | | |
|---|---|---|---|---|---|---|---|---|
| **Chr** | **Marker** | **Pos. (cM)** | $N_{CA}$ | $N_{CO}$ | $p$ | $R_\chi$ | *RM* | GC |
| 16 | tsc1750530 | 59.8297 | 644 | 145 | .85 | 3.6e-4 | 2.3e-8[a] | 1.6e-3 |
| 18 | tsc0046696 | 104.665 | 459 | 118 | .60 | 4.0e-1 | 1.4e-7[a] | 4.8e-1 |
| 1 | tsc1177811 | 105.535 | 587 | 149 | .68 | 2.9e-2 | 1.5e-7[a] | 3.4e-2 |
| 5 | tsc1637642 | 95.4901 | 419 | 159 | .84 | 3.6e-1 | 6.8e-7[a] | 4.3e-2 |
| 11 | tsc0571038 | 95.3968 | 581 | 122 | .56 | 8.0e-7[a] | 4.5e-3 | 5.4e-5 |
| 18 | tsc0057290 | 33.9594 | 497 | 126 | .71 | 6.1e-2 | 1.8e-6 | 5.3e-2 |
| 6 | tsc0808295 | 47.1522 | 681 | 162 | .76 | 9.4e-1 | 1.8e-6 | 9.4e-1 |
| 11 | tsc0569292 | 6.78451 | 455 | 127 | .74 | 6.0e-1 | 3.6e-6 | 5.7e-1 |
| 19 | tsc1189131 | 68.94 | 478 | 121 | .55 | 7.4e-1 | 4.2e-6 | 7.3e-1 |
| 3 | tsc0175005 | 158.199 | 594 | 152 | .82 | 1.6e-2 | 4.8e-6 | 2.2e-2 |
| 3 | tsc1519933 | 167.431 | 515 | 134 | .64 | 9.5e-2 | 5.3e-6 | 7.6e-2 |
| 13 | tsc0056748 | 73.9934 | 530 | 133 | .84 | 6.5e-1 | 6.5e-6 | 6.7e-1 |

The chromosome (Chr), the name of the marker (Marker), the position of the marker on the chromosome (Pos.), the number of genotypes available in cases ($N_{CA}$) and controls ($N_{CO}$), and the major allele frequency in the case-control sample as a whole ($p$) are displayed.
[a] Genome-wide significance after Bonferroni correction.

5 and 21 min, respectively. The large difference in computing time is due to the COGA data having extended pedigrees and a sample size that is almost twice that of the RA data. The slowest step is the Cholesky decomposition[33] of $\mathbf{\Phi}$ (for the calculation of *RM* and a ROADTRIPS version of $W_{QLS}$), which we compute at every SNP because the pattern of missing genotype data varies. The computing time scales linearly with the number of SNPs. The speed could presumably be improved, as we have not made serious attempts to optimize the code.

## Discussion

Technological advances in high-density genome scans have made it feasible to perform case-control association studies on a genome-wide basis with hundreds of thousands or millions of markers. The observations in these studies can have several sources of dependence, including population structure and relatedness among the sampled individuals, some of which might be known and some unknown. Failure to properly account for this structure can lead to spurious association or reduced power. We develop ROADTRIPS, a case-control association testing method that simultaneously corrects for both pedigree and population structure, including admixture, where some or all of the structure can be unknown. The method also automatically adjusts for pedigree errors and sample switches. ROADTRIPS is computationally feasible for the analysis of genome-screen data with millions of markers, and it is applicable to association studies with completely general combinations of family and case-control designs. The method does not require the genealogy of the sampled individuals to be known, but when pedigree data are available, ROADTRIPS can incorporate this information to improve power. Our simulation studies indicate that including known pedigree information in ROADTRIPS (by use of the *RM* test) provides an overall and, in some cases, substantial improvement in power over other available methods. In an analysis of GAW 15 RA data from small UK pedigrees, ROADTRIPS detected three SNPs that have significant association with a RA phenotype. In a reanalysis of the GAW 14 COGA data, ROADTRIPS detected five SNPs that have significant association with alcoholism, one of which had not been identified as significant in the previous analysis, and another SNP identified as significant in the previous analysis is no longer identified when cryptic structure is accounted for.

We have shown that when different SNPs have different rates of missing genotype data, ROADTRIPS is still valid, whereas GC is not properly calibrated for this setting. The ROADTRIPS method takes into account both the structure in the data and the particular missing genotype pattern at each SNP to construct a valid test. In contrast, the uniform inflation factor applied to all SNPs by GC can result in both an increase in type 1 error, due to undercorrection of SNPs with lower rates of missing genotype data, as well as a loss of power, due to overcorrection of SNPs with higher rates of missing genotype data. One approach to dealing with this problem in samples of unrelated individuals is to impute missing genotype data. However, there are special difficulties that arise with the use of imputation methods in samples with related individuals and hidden structure. First, Mendelian errors and incompatible genotypes can be introduced with this approach, unless the imputation is performed jointly among related individuals. Second, in samples with hidden population structure, e.g., samples from admixed populations, it is often unclear what the reference population should be, because an individual's ancestry at a particular SNP will generally be unknown. Finally, imputed genotypes are dependent among relatives, where the dependence among imputed genotypes differs from the ordinary dependence among genotypes and is affected by the type and amount of information available for each individual for each SNP. Thus, unlike the pedigree and population structure we consider, the dependence structure among imputed genotypes for different individuals differs across SNPs. This complex dependence among imputed genotypes for related individuals would need to be taken into account in the analysis in order to construct a valid test. However, to our knowledge, the current generation of imputation methods gives information only on the marginal accuracy (e.g., marginal posterior probabilities and not joint posterior probabilities) of imputed genotypes across individuals, so these methods would not allow valid assessment of uncertainty in the general

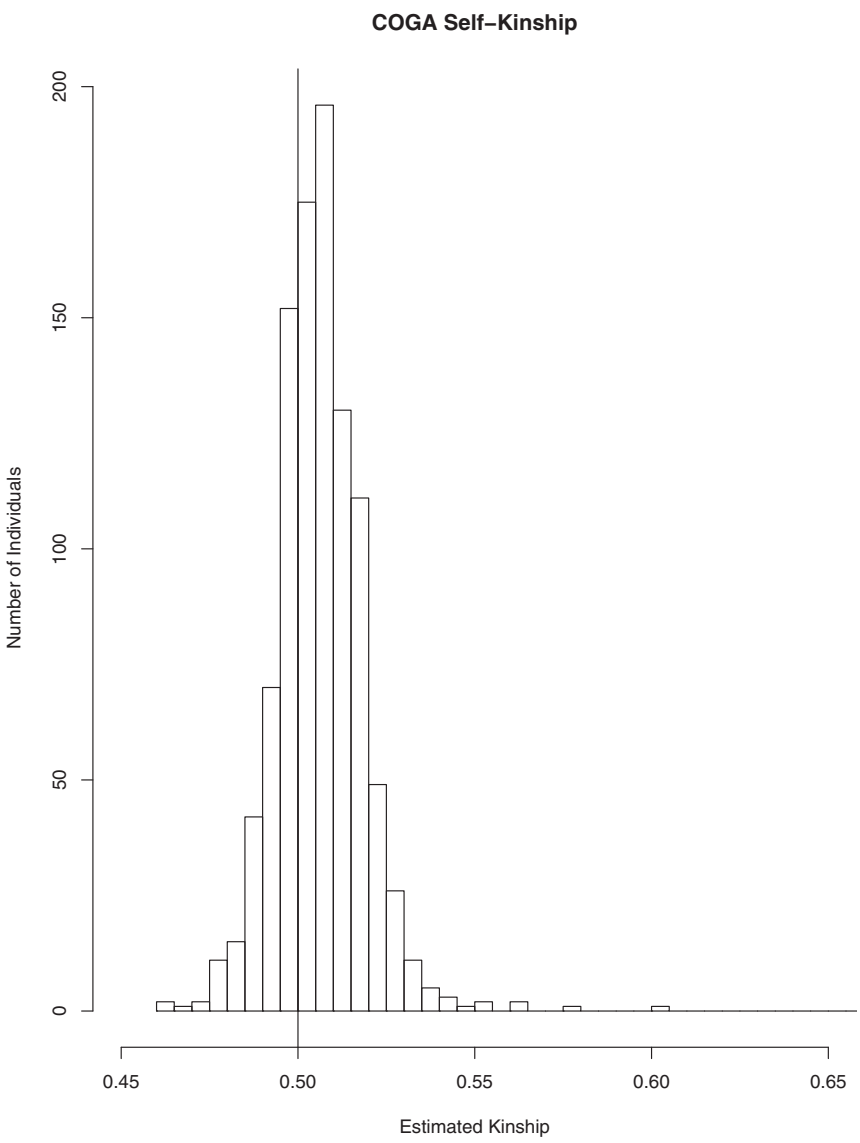


**Figure 2. Histogram of Estimated Self-Kinship Coefficient Values for the GAW 14 COGA Data**
The vertical line at 0.5 represents the self-kinship value in the absence of population structure and inbreeding.

setting of case-control association testing with related individuals.

The ROADTRIPS method uses an estimator of $\boldsymbol{\Psi}$ that is closely related to the estimated covariance matrix used in EIGENSTRAT.[6] Recently, Choi et al.[17] have used a different estimated kinship matrix in the context of association testing when pedigree information is missing. The kinship estimation approach of Choi et al. is suitable for close relatives in the absence of population structure but is not particularly well suited to accounting for population structure. To make this statement more precise, we consider the case of unrelated individuals with population structure based on the Balding-Nichols model with or without admixture. In this context, if one assumed that genotypes at different SNPs were independent with $|S_{ij}| \to \infty$ and $p_s$ known and that ${\sigma_s}^2 = 0.5p_s(1 - p_s)$ held at all but a finite number of SNPs, then the ROADTRIPS estimator of $\Psi_{ij}$ would be consistent, whereas the estimator of Choi et al. would generally be inconsistent. The estimator of Choi et al. is also substantially more computationally intensive to calculate than the $\hat{\boldsymbol{\Psi}}$ we use in ROADTRIPS.

## Acknowledgments

This study was supported in part by the University of California President's Postdoctoral Fellowship and the Lamond Family Foundation Postdoctoral Fellowship (to T.T.) and by National Institutes of Health grant R01 HG001645 (to M.S.M.). Data on alcohol dependence were provided by the Collaborative Study on the Genetics of Alcoholism (U10AA008401) through the Genetics Analysis Workshop (R01GM031575). Data on RA were provided by a UK group led by Jane Worthington and Sally John through the Genetics Analysis Workshop (R01GM031575).

Received: November 15, 2009
Revised: January 6, 2010
Accepted: January 10, 2010
Published online: February 4, 2010

## Web Resources

The URLs for data presented herein are as follows:

ROADTRIPS source code, http://www.stat.uchicago.edu/~mcpeek/software/index.html

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/

## References


1. Lander, E.S., and Schork, N.J. (1994). Genetic dissection of complex traits. Science *265*, 2037–2048.
2. Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.
3. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
4. Zhu, X., Zhang, S., Zhao, H., and Cooper, R.S. (2002). Association mapping, using a mixture model for complex traits. Genet. Epidemiol. *23*, 181–196.
5. Zhang, S., Zhu, X., and Zhao, H. (2003). On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet. Epidemiol. *24*, 44–56.
6. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
7. Lee, A.B., Luca, D., Klei, L., Devlin, B., and Roeder, K. (2010). Discovering genetic ancestry using spectral graph theory. Genet. Epidemiol. *34*, 51–59.
8. Zhang, J., Niyogi, P., and McPeek, M.S. (2009). Laplacian eigenfunctions learn population structure. PLoS ONE *4*, e7928.
9. Satten, G.A., Flanders, W.D., and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am. J. Hum. Genet. *68*, 466–477.
10. Hoggart, C.J., Parra, E.J., Shriver, M.D., Bonilla, C., Kittles, R.A., Clayton, D.G., and McKeigue, P.M. (2003). Control of confounding of genetic associations in stratified populations. Am. J. Hum. Genet. *72*, 1492–1504.
11. Kimmel, G., Jordan, M.I., Halperin, E., Shamir, R., and Karp, R.M. (2007). A randomization test for controlling population stratification in whole-genome association studies. Am. J. Hum. Genet. *81*, 895–905.
12. Luca, D., Ringquist, S., Klei, L., Lee, A.B., Gieger, C., Wichmann, H.-E., Schreiber, S., Krawczak, M., Lu, Y., Styche, A., et al. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am. J. Hum. Genet. *82*, 453–463.
13. Rakovski, C.S., and Stram, D.O. (2009). A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors. PLoS ONE *4*, e5825.
14. Slager, S.L., and Schaid, D.J. (2001). Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. Am. J. Hum. Genet. *68*, 1457–1462.
15. Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., and McPeek, M.S. (2003). Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am. J. Hum. Genet. *73*, 612–626.
16. Thornton, T., and McPeek, M.S. (2007). Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am. J. Hum. Genet. *81*, 321–337.
17. Choi, Y., Wijsman, E.M., and Weir, B.S. (2009). Case-control association testing in the presence of unknown relationships. Genet. Epidemiol. *33*, 668–678.
18. Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. *52*, 506–516.
19. Rabinowitz, D., and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum. Hered. *50*, 211–223.
20. Risch, N., and Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res. *8*, 1273–1288.
21. Bacanu, S.A., Devlin, B., and Roeder, K. (2000). The power of genomic control. Am. J. Hum. Genet. *66*, 1933–1944.
22. Amos, C.I., Chen, W.V., Remmers, E., Siminovitch, K.A., Seldin, M.F., Criswell, L.A., Lee, A.T., John, S., Shephard, N.D., Worthington, J., et al. (2007). Data for Genetic Analysis Workshop (GAW) 15 Problem 2, genetic causes of rheumatoid arthritis and associated traits. BMC Proc *1* (*Suppl 1*), S3.
23. Edenberg, H.J., Bierut, L.J., Boyce, P., Cao, M., Cawley, S., Chiles, R., Doheny, K.F., Hansen, M., Hinrichs, T., Jones, K., et al. (2005). Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. BMC Genet. *6* (*Suppl 1*), S2.
24. Sasieni, P.D. (1997). From genotypes to genes: doubling the sample size. Biometrics *53*, 1253–1261.
25. McPeek, M.S., Wu, X., and Ober, C. (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics *60*, 359–367.
26. Lange, K. (2002). Mathematical and Statistical Methods for Genetic Analysis (New York: Springer).
27. Wang, Z., and McPeek, M.S. (2009). An incomplete-data quasi-likelihood approach to haplotype-based genetic association studies on related individuals. Journal of the American Statistical Association. *104*, 1251–1260.
28. Balding, D.J., and Nichols, R.A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica *96*, 3–12.
29. Pritchard, J.K., and Donnelly, P. (2001). Case-control studies of association in structured or admixed populations. Theor. Popul. Biol. *60*, 227–237.
30. Wright, S. (1951). The genetical structure of populations. Ann. Eugenics *15*, 323–354.
31. Lake, S.L., Blacker, D., and Laird, N.M. (2000). Family-based tests of association in the presence of linkage. Am. J. Hum. Genet. *67*, 1515–1525.
32. Zhu, X., Cooper, R., Kan, D., Cao, G., and Wu, X. (2005). A genome-wide linkage and association study using COGA data. BMC Genet. *6* (*Suppl 1*), S128.
33. Schott, J.R. (1996). Matrix analysis for statistics (New York: John Wiley).