Retrospective Binary-Trait Association Test Elucidates Genetic Architecture of Crohn Disease

Duo Jiang,[1] Sheng Zhong,[2] and Mary Sara McPeek[2,3,*]

In genetic association testing, failure to properly control for population structure can lead to severely inflated type 1 error and power loss. Meanwhile, adjustment for relevant covariates is often desirable and sometimes necessary to protect against spurious association and to improve power. Many recent methods to account for population structure and covariates are based on linear mixed models (LMMs), which are primarily designed for quantitative traits. For binary traits, however, LMM is a misspecified model and can lead to deteriorated performance. We propose CARAT, a binary-trait association testing approach based on a mixed-effects quasi-likelihood framework, which exploits the dichotomous nature of the trait and achieves computational efficiency through estimating equations. We show in simulation studies that CARAT consistently outperforms existing methods and maintains high power in a wide range of population structure settings and trait models. Furthermore, CARAT is based on a retrospective approach, which is robust to misspecification of the phenotype model. We apply our approach to a genome-wide analysis of Crohn disease, in which we replicate association with 17 previously identified regions. Moreover, our analysis on 5p13.1, an extensively reported region of association, shows evidence for the presence of multiple independent association signals in the region. This example shows how CARAT can leverage known disease risk factors to shed light on the genetic architecture of complex traits.

Introduction

Population structure is widespread in genetic association studies. It can arise when the sample is stratified between population subgroups, is admixed with multiple ancestral populations, and/or contains cryptic relatedness. It is well known that failure to properly adjust for population structure in genetic association testing can lead to severely inflated type 1 error rates and loss of power. Moreover, control for covariate information relevant to the phenotype of interest can be useful in genetic association analysis. These covariates can include clinical variables such as sex, age, and smoking habits that can affect disease risk as well as genetic variants known to be causal. Adjustment for such covariates can be desirable in association testing to protect against spurious associations due to confounding factors, and to improve power by reducing unexplained noise in the phenotype.[1–3]

Case-control studies have been a useful design for identification of genetic variants associated with complex traits. We consider the problem of association testing of a binary trait (such as disease status) with simultaneous adjustment for population structure and covariates. ROADTRIPS[4] is a binary-trait association testing method accounting for population structure, but it does not adjust for covariates or additive polygenic effects. Methods[5–9] based on the linear mixed model (LMM) approach have recently gained much popularity. Compared to alternative methods that model population structure as fixed effects,[1,10,11] LMM accommodates simultaneous adjustment for cryptic relatedness as well as sample stratification and admixture, and it offers potential power gain by implicitly conditioning on associated SNPs other than the tested SNP.[12] However, LMM is primarily designed for quantitative traits, and when applied to case-control data, it imposes a misspecified model on the binary phenotype, which can lead to power loss. Two recently proposed methods, LTMLM[13] and LEAP,[14] specifically focus on ascertained case-control studies assuming known disease prevalence. However, LTMLM does not adjust for covariates, and we find (see Results) that LEAP does not succeed in our simulations.

We propose CARAT (CAse-control Retrospective Association Test), a binary-trait association testing approach, which accounts for relevant covariate information and effectively controls for unknown population structure. CARAT exploits the dichotomous nature of the trait by modeling the phenotypic distribution using a mixed effects quasi-likelihood framework. In contrast to LTMLM and LEAP, CARAT does not require the prevalence of the disease to be known. We propose an estimating equation and score test approach, which is computationally efficient for large-scale genome-wide studies. When assessing significance of the test statistic, CARAT takes a retrospective approach in which genotypes are viewed as random under the null, conditional on the phenotype and the covariates. This approach renders CARAT robust to misspecification of the phenotype model. We perform simulation studies to evaluate the type 1 error and power of CARAT, to compare it with existing methods. Finally, we apply CARAT to a genome-wide association analysis of Crohn disease (CD [MIM: 266600]) data from the Wellcome Trust Case Control Consortium (WTCCC).

[1]Department of Statistics, Oregon State University, Corvallis, OR 97331, USA; [2]Department of Statistics, [3]Department of Human Genetics
University of Chicago, Chicago, IL 60637, USA
*Correspondence: mcpeek@uchicago.edu

http://dx.doi.org/10.1016/j.ajhg.2015.12.012.

Material and Methods

The Mean and Variance Model Underlying CARAT

We consider a binary trait measured on a sample of n individuals possibly subject to unobserved population structure, assuming that genome-wide data are available. We focus on the problem of testing for association between the binary trait and a single SNP. Let $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)^T$ be a vector of the binary trait measurements on the n individuals. $\boldsymbol{G} = (G_1, G_2, ..., G_n)^T$ encodes the genotypes of the sampled individuals at the tested SNP, where G_i equals the minor allele count (0, 1, or 2) of individual i. Let $\boldsymbol{X}$ be an n by k covariate matrix, whose i^{th} row $\boldsymbol{X}_i$ contains an intercept term represented as 1 and the values of $k - 1$ non-constant covariates for individual i.

To model the phenotype vector $\boldsymbol{Y}$ conditional on $\boldsymbol{G}$ and $\boldsymbol{X}$, we take a quasi-likelihood approach, in which we specify only the conditional mean and variance structures of $\boldsymbol{Y}$. For the mean structure, we assume that, for $i = 1,..., n$,

$$E(Y_i \mid \boldsymbol{X}, \boldsymbol{G}) = \mu_i,\ g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} + G_i\gamma, \qquad \text{(Equation 1)}$$

where $g(\bullet)$ is a known function, $\boldsymbol{\beta}$ is a k-dimensional column vector of the unknown fixed effects of the covariates, and γ is the unknown scalar effect of the tested SNP. We take $g(\bullet)$ to be the logit link function given by

$$g(\mu_i) = \log\frac{\mu_i}{1-\mu_i}, \qquad \text{(Equation 2)}$$

in which case the linear coefficients can be conveniently interpreted as the size of an additive effect on the log odds scale. The logit link function offers the additional benefit that it is applicable to case-control samples with ascertainment with the intercept considered a nuisance parameter.

To construct a covariance structure for $\boldsymbol{Y}$, we aim to embed two features that we deem essential: (1) the variance should depend on the mean in a way that is consistent with the dichotomous nature of the outcome and (2) assuming that overall genetic similarity leads to phenotypic similarity, the trait measurements should be subject to correlation due to population structure. With regard to the first feature, one approach is to specify

$$\boldsymbol{\Omega} := \text{Var}(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{G}) = \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{1/2}, \qquad \text{(Equation 3)}$$

where $\boldsymbol{\Gamma} = \text{diag}\{\mu_1(1-\mu_1), \cdots, \mu_n(1-\mu_n)\}$ is an n-dimensional diagonal matrix and $\boldsymbol{\Sigma}$ is an n by n correlation matrix (defined as a positive semi-definite matrix with 1s on the diagonal) that does not involve the mean structure. Under this specification, the marginal variance is determined by the mean according to $\text{var}(Y_i) = E(Y_i)[1 - E(Y_i)]$.

An alternative approach is to add a dispersion parameter, $\sigma^2 > 0$, to the model, to obtain

$$\boldsymbol{\Omega} := \text{Var}(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{G}) = \sigma^2\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{1/2}, \qquad \text{(Equation 4)}$$

where the marginal variance is now determined by the mean according to $\text{var}(Y_i) = \sigma^2 E(Y_i)[1 - E(Y_i)]$. In a generalized linear or quasi-likelihood model,[15–17] a dispersion parameter is often included in the variance structure to allow over-dispersion or under-dispersion. However, as has been pointed out,[18] in the case of Bernoulli data, the dispersion parameter is not interpretable in the context of the model, because for the binary random vector $\mathbf{Y}$, regardless of its joint distribution, the marginal distribution of each Y_i is always a Bernoulli distribution, in which case the marginal variance would be $E(Y_i)[1 - E(Y_i)]$ with no dispersion parameter. On the other hand, in the context of genetic association testing, a justification for possibly including a dispersion parameter is the fact that the true model is not known, and the dispersion parameter could potentially be useful for capturing additional error due to model misspecification, in order to improve robustness of the resulting association test. We will demonstrate through simulation studies, in the context of association testing using quasi-likelihood models, that exclusion of the dispersion parameter as in Equation 3 results in improved analysis over the model in Equation 4.

For the structure specified either by Equation 3 or by Equation 4, feature 2 can be achieved by assuming that $\boldsymbol{\Sigma}$ has two additive components, one corresponding to individual-level variance and the other to additive polygenic variance:

$$\boldsymbol{\Sigma} = \xi\boldsymbol{\Phi} + (1-\xi)\boldsymbol{I}, \qquad \text{(Equation 5)}$$

where $\boldsymbol{\Phi}$ is an n by n correlation matrix representing the overall genetic similarity between the individuals due to population structure, $\boldsymbol{I}$ is an n-dimensional identity matrix, and $\xi \in [0,1]$ measures the relative importance of the two variance components.

$\boldsymbol{\Phi}$ can be estimated based on genome-wide data, for example, by using the genetic relationship matrix

$$\widehat{\boldsymbol{\Psi}} = \frac{1}{S}\sum_{s=1}^{S}\frac{\left(\boldsymbol{G}_{(s)} - 2\hat{p}_s\right)\left(\boldsymbol{G}_{(s)} - 2\hat{p}_s\right)^T}{2\hat{p}_s\left(1-\hat{p}_s\right)}, \qquad \text{(Equation 6)}$$

where S is the total number of genotyped markers, $\boldsymbol{G}_{(s)}$ is the genotype column vector at marker s encoding the minor allele counts, and $\hat{p}_s$ is the estimated minor allele frequency (MAF). Alternatively, one could use the sample variance of $\boldsymbol{G}_{(s)}$, instead of $2\hat{p}_s(1-\hat{p}_s)$, in the denominator. A caveat of using $\widehat{\boldsymbol{\Psi}}$ for $\boldsymbol{\Phi}$ is that the diagonal elements of $\widehat{\boldsymbol{\Psi}}$ might not be exactly 1, in which case the resulting $\boldsymbol{\Sigma}$ will not be a valid correlation matrix. To solve this problem, one could obtain an alternative estimator $\hat{\boldsymbol{\Phi}}$ for $\boldsymbol{\Phi}$ by letting

$$\hat{\Phi}_{ij} = \widehat{\Psi}_{ij}\Big/\sqrt{\widehat{\Psi}_{ii}\widehat{\Psi}_{jj}} \text{ for } i,j = 1,\cdots,n. \qquad \text{(Equation 7)}$$

However, $\hat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Psi}}$ tend to be very close numerically when the data do not involve severe systematic deviation from Hardy-Weinberg equilibrium. We consider both estimators in the simulation studies in Results.

The model specified by Equations 1, 2, 3, and 5 will be used for the main method we propose, CARAT. The model specified with Equation 3 replaced by Equation 4 will be referred to as the model with dispersion, and it is used for the alternative method we consider, CARATd. We compare the performance of the methods for association testing based on these two models in the Results.

Parameter Estimation Based on Estimating Equations

For the model either with or without dispersion, we obtain parameter estimates by solving a system of estimating equations, as we now describe. For simplicity, we bind the notation for the covariate matrix and the genotype vector to define $\tilde{\boldsymbol{X}} = (\boldsymbol{X}, \boldsymbol{G})$ and $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}^T, \gamma)^T$. For the model with dispersion, the parameters to be estimated are $\tilde{\beta}$, ξ, and σ^2. For known ξ and σ^2, the quasi-likelihood function for $\tilde{\boldsymbol{\beta}}$ can be differentiated to obtain the quasi-score function[17]

$$\boldsymbol{U}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{D}^T\boldsymbol{\Omega}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}), \qquad \text{(Equation 8)}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}}) = (\mu_1, \cdots, \mu_n)^T$ and $\boldsymbol{D} = \boldsymbol{D}(\tilde{\boldsymbol{\beta}}) = \partial\boldsymbol{\mu}/\partial\tilde{\boldsymbol{\beta}}$ is a Jacobian of the mean vector with respect to $\tilde{\boldsymbol{\beta}}$. With a logit link function, $\boldsymbol{D} = \boldsymbol{\Gamma}\tilde{\boldsymbol{X}}$. Setting $\boldsymbol{U}(\tilde{\boldsymbol{\beta}}) = 0$ yields the generalized estimating equation for $\tilde{\boldsymbol{\beta}}$ given by

$$\tilde{\boldsymbol{X}}^T \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma}^{-1/2} (\boldsymbol{Y} - \boldsymbol{\mu}) = 0, \qquad \text{(Equation 9)}$$

which is a system of $k + 1$ nonlinear equations. To solve for $\tilde{\boldsymbol{\beta}}$, a modified Newton-Raphson algorithm with Fisher scoring[17] involves iteratively updating $\tilde{\boldsymbol{\beta}}$ by

$$\tilde{\boldsymbol{\beta}}_{(j+1)} = \tilde{\boldsymbol{\beta}}_{(j)} + \left\{\boldsymbol{D}^T\left(\tilde{\boldsymbol{\beta}}_{(j)}\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\left(\tilde{\boldsymbol{\beta}}_{(j)}\right)\right\}^{-1}\left\{\boldsymbol{D}^T\left(\tilde{\boldsymbol{\beta}}_{(j)}\right)\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{Y} - \boldsymbol{\mu}\left(\tilde{\boldsymbol{\beta}}_{(j)}\right)\right]\right\}. \qquad \text{(Equation 10)}$$

Starting at an initial value for $\tilde{\boldsymbol{\beta}}$ estimated under a simple logistic model with no correlation, we run this iterative procedure until convergence to obtain $\widehat{\tilde{\boldsymbol{\beta}}}(\xi)$ for fixed ξ (note that Equation 9 does not depend on σ^2).

To estimate the variance parameters, a common strategy is to establish estimating equations by choosing quadratic forms and then equating their observed values to their expected values. Following this strategy, we propose to estimate ξ and σ^2 for known $\tilde{\boldsymbol{\beta}}$ using

$$\sigma^{-2}(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{\Gamma}^{-1/2} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Phi} - I) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma}^{-1/2} (\boldsymbol{Y} - \boldsymbol{\mu}) = \operatorname{trace}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Phi} - \boldsymbol{I})\right) \qquad \text{(Equation 11)}$$

and

$$\sigma^2 = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{\Gamma}^{-1/2} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma}^{-1/2} (\boldsymbol{Y} - \boldsymbol{\mu}) / n. \qquad \text{(Equation 12)}$$

The choice of quadratic forms resulting in these estimating equations is motivated by the maximum likelihood estimators of ξ and σ^2 in the case when $\tilde{\boldsymbol{\beta}}$ is known, assuming a multivariate normal distribution for $\boldsymbol{Y}$ with the same mean and covariance structure as the model with dispersion. Taking into account the fact that $\tilde{\boldsymbol{\beta}}$ is unknown, one could alternatively use $n - k - 1$ for the denominator in the right side of Equation 12.

Combining Equations 9, 11, and 12 provides a scheme for joint estimation of the parameters $\tilde{\boldsymbol{\beta}}$, ξ, and σ^2. As a numerical procedure, for each fixed $\xi \in [0,1]$, we first make use of Equations 9 and 12 to obtain $\widehat{\tilde{\boldsymbol{\beta}}}(\xi)$ and $\widehat{\sigma}^2(\xi)$, which are then plugged into Equation 11 to evaluate the difference between the two sides of the equality. This difference is minimized in absolute value by a one-dimensional search with respect to $\xi \in [0,1]$ to locate the minimum point $\widehat{\xi}$ as an estimate for ξ and the corresponding $\widehat{\tilde{\boldsymbol{\beta}}}(\widehat{\xi})$ and $\widehat{\sigma}^2(\widehat{\xi})$ as estimates for $\tilde{\boldsymbol{\beta}}$ and σ^2. When the model without dispersion is assumed instead, the parameters $\tilde{\boldsymbol{\beta}}$ and ξ can be estimated by solving Equation 9 combined with Equation 11, in which σ^2 is set to be 1. The previously described numerical algorithm can readily be adapted to solve these equations.

Retrospective Association Testing

To detect association between the trait and the SNP of interest, we test $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$. In general, the construction of a quasi-score test statistic would involve evaluating the corresponding coordinate(s) of the quasi-score function at the quasi-likelihood estimates of the nuisance parameters under the null hypothesis.[19] In our case, for the model without dispersion parameter, we let $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\Omega}}_0$, and $\widehat{\boldsymbol{\Gamma}}_0$ denote the values of $\boldsymbol{\mu}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\Gamma}$ evaluated at $(\gamma, \boldsymbol{\beta}, \xi) = (0, \widehat{\boldsymbol{\beta}}_0, \widehat{\xi}_0)$, where $(\boldsymbol{\beta}, \xi) = (\widehat{\boldsymbol{\beta}}_0, \widehat{\xi}_0)$ represents the solution of the system given by Equations 9 and 11 when γ is set to 0. (Similarly, in the model with the dispersion parameter, we let $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\Omega}}_0$, and $\widehat{\boldsymbol{\Gamma}}_0$ denote the values of $\boldsymbol{\mu}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\Gamma}$ evaluated at $(\gamma, \boldsymbol{\beta}, \xi, \sigma^2) = (0, \widehat{\boldsymbol{\beta}}_0, \widehat{\xi}_0, \widehat{\sigma}_0^2)$, where $(\boldsymbol{\beta}, \xi, \sigma^2) = (\widehat{\boldsymbol{\beta}}_0, \widehat{\xi}_0, \widehat{\sigma}_0^2)$ represents the solution of the system given by Equations 9, 11, and 12 when γ is set to 0.) It follows that, evaluated at the null estimates, the coordinate of Equation 8 corresponding to $\boldsymbol{G}$ becomes

$$U_0 := \boldsymbol{G}^T \widehat{\boldsymbol{\Gamma}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0). \qquad \text{(Equation 13)}$$

To perform a quasi-score test, one would divide U_0^2 by its null variance given $\boldsymbol{X}$ and $\boldsymbol{G}$ with the null estimates plugged in, and the resulting test statistic would then be assumed to have a χ_1^2 distribution under H_0. This method will be referred to as "the prospective version of CARAT" (details in Appendix A). The validity of such a test would be contingent on the accuracy of the assumed null distribution, which relies on the consistency of the null estimators and tends to be very sensitive to misspecification of the variance structure in the prospective model for $\boldsymbol{Y}$. Our simulation studies show that substantial inflation or deflation of type 1 error rates frequently occurs in many realistic scenarios (Table S1).

In order to achieve robust control over type 1 error, we assess significance of the test by using a retrospective analysis, in which the conditional distribution of the phenotype given genotype and covariates is considered. Under the null hypothesis of no association, we build a quasi-likelihood model for $\boldsymbol{G}$ conditional on $\boldsymbol{Y}$ and $\boldsymbol{X}$ specified by the following assumptions[4]

$$E_0(\boldsymbol{G} \mid \boldsymbol{X}, \boldsymbol{Y}) = 2p\boldsymbol{1}_n \text{ and } \operatorname{Var}_0(\boldsymbol{G} \mid \boldsymbol{X}, \boldsymbol{Y}) = \sigma_g^2 \boldsymbol{\Psi}, \qquad \text{(Equation 14)}$$

where $\boldsymbol{1}_n$ is an n-dimensional column vector of 1s, $p \in [0,1]$ is the unknown ancestral MAF of the tested SNP, $\boldsymbol{\Psi}$ is a known positive semi-definite matrix capturing overall genetic similarity among individuals due to population structure, and $\sigma_g^2 > 0$ is an unknown variance parameter. $\boldsymbol{\Psi}$ can be obtained based on genome-wide data via the estimator in Equation 6. Let $\widehat{\sigma}_g^2 = 2\widehat{p}(1 - \widehat{p})$ with $\widehat{p} = 0.5 \cdot \overline{G}$ being the sample average estimator for p. Then, the final CARAT test statistic can be defined as

$$T := \frac{(\boldsymbol{V}^T \boldsymbol{G})^2}{\widehat{\sigma}_g^2 \cdot \boldsymbol{V}^T \boldsymbol{\Psi} \boldsymbol{V}} \text{with } \boldsymbol{V} = \widehat{\boldsymbol{\Gamma}}_0^{1/2} \widehat{\boldsymbol{\Sigma}}_0^{-1} \widehat{\boldsymbol{\Gamma}}_0^{-1/2} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0), \qquad \text{(Equation 15)}$$

where the null estimates in $\boldsymbol{V}$ are based on the model without dispersion. Under regularity conditions, T has an asymptotic χ_1^2 distribution under the null hypothesis. A test statistic of the same form, except that the null estimates in $\boldsymbol{V}$ are obtained from the model with dispersion, will give rise to another valid test, which we will refer to as the CARATd method, with the letter d standing for "dispersion."

Connections with LMM

For definiteness, we first present the most commonly used version of the LMM single-SNP association testing statistic. The method assumes a linear mixed effects model given by

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{G}\gamma + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathrm{N}\left(0, \sigma_a^2 \boldsymbol{\Phi} + \sigma_e^2 \boldsymbol{I}\right), \qquad \text{(Equation 16)}$$

where σ_a^2 and σ_e^2 are variance component parameters corresponding to additive polygenic effects and environmental errors, respectively. Denote by $\widehat{\boldsymbol{\mu}}_L$ and $\widehat{\boldsymbol{\Omega}}_L$ the MLEs for the mean vector, $\boldsymbol{\mu}_L = \boldsymbol{X\beta}$, and the covariance matrix, $\boldsymbol{\Omega}_L = \sigma_a^2 \boldsymbol{\Phi} + \sigma_e^2 \boldsymbol{I}$, of $\boldsymbol{Y}$ based on the null model with $\gamma = 0$ (or alternatively the REML estimator could be used for $\boldsymbol{\Omega}_L$). Then the LMM test statistic is given by

$$T_{LMM} := \frac{\left[(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_L)^T \boldsymbol{G}\right]^2}{\boldsymbol{G}^T \widehat{\boldsymbol{\Omega}}_L \boldsymbol{G}},$$

whose null distribution is assumed to be χ_1^2.

Comparing CARAT (and CARATd) with the LMM approach, the difference is 3-fold. First, in terms of the mean structure, CARAT/CARATd assumes that the covariates contribute to the phenotype linearly and additively on the logit scale, whereas LMM assumes linearity and additivity on the original scale. We expect the former to be more reasonable because the trait is dichotomous. The second aspect of the difference relates to the variance structure. LMM imposes a variance structure that is not dependent on the mean, whereas both CARAT and CARATd assume a variance structure with two components, the marginal variance and the correlation matrix, with the first component dependent on the mean and the second one not. CARAT differs from CARATd in that, in the former, the marginal variance is fully determined by the mean, and in the latter it is determined by the mean only up to a scale parameter σ^2. Third, LMM assesses the p values prospectively, assuming phenotype information to be random, whereas CARAT and CARATd can be viewed as directly taking the score test statistic from the prospective model to be the test statistic, but evaluating its p value in a retrospective manner assuming genotype information to be random. We expect the retrospective approach to offer better calibration for our method than the prospective quasi-score test, because the retrospective quasi-likelihood model specified by Equation 14 is much less susceptible to misspecification than the prospective model. This is confirmed by the results in Table S1; see also Appendix A.

Simulation Studies

We conduct simulation studies to evaluate the type 1 error rates and power of CARAT and to compare its performance with that of CARATd and two existing methods, LMM and LEAP, for binary trait association testing. We do not evaluate ROADTRIPS or LTMLM because they do not account for covariates. Genotype, phenotype, and covariate data are simulated on a sample of individuals with various population structure configurations and trait models. In each setting, we simulate 10,000 non-causal SNPs, which are used to correct for population structure. In addition, we generate two causal SNPs, which are assumed to influence the phenotype with epistasis: individuals holding at least one copy of the minor allele at both SNPs have an elevated disease risk compared to those who do not.

In the type 1 error simulations, we test the 10,000 non-causal SNPs for association. For each given setting, the phenotypes are re-simulated 100 times, and 1,000 SNPs are tested for each phenotype simulation. Altogether, 10 tests are performed for each of the 10,000 non-causal SNPs, totaling 100,000 replicates for the type 1 error simulations. In the power simulations, we test the first of the two causal SNPs, and 5,000 replicates are performed with the phenotypes re-simulated for each replicate. For every test, the genotypes at the untested causal SNP(s) are assumed to be unobserved. In the implementations of CARAT and CARATd, we use $\widehat{\boldsymbol{\Psi}}$ given in Equation 6 as an estimator of the genetic relationship matrix $\boldsymbol{\Psi}$ in Equation 14, which is used to assess significance retrospectively. To obtain an estimator of the correlation matrix $\boldsymbol{\Phi}$ of Equation 5 in the trait model, we try both $\widehat{\boldsymbol{\Psi}}$ given in Equation 6 and $\widehat{\boldsymbol{\Phi}}$ given in Equation 7, which is a standardized version of $\widehat{\boldsymbol{\Psi}}$. We note that, theoretically, how $\boldsymbol{\Phi}$ is estimated in the trait model should not affect the validity of a retrospective test.

Sampled Genotypes

For the non-causal markers, the allele frequencies in different ancestral populations or sub-populations are generated according to the Balding-Nichols model.[20] At each SNP, the ancestral allele frequency p is drawn from a uniform distribution on [0,1], independently across SNPs. Conditioning on p, the allele frequency p_k in population k is then drawn, independently for $k = 1,2$, from a Beta distribution with parameters $p(1 - F)/F$ and $(1 - p)(1 - F)/F$, where F is the fixation index measuring genetic differentiation due to population substructure. We take $F = 0.01$ for all simulations. In scenarios with population admixture, for individual i who is either a pedigree founder or is unrelated to everyone else in the sample and whose admixture proportion from population 1 is A_i, the allele frequency at the SNP is given by $A_i \cdot p_1 + (1 - A_i) \cdot p_2$. For descendants in a pedigree, SNP genotypes are obtained by gene dropping in the pedigree. For any given simulation scenario, the non-causal SNPs are simulated only once, with the estimated genetic relationship matrix reused for every simulation replicate. In addition to the non-causal SNPs, two causal SNPs are simulated for each replicate, using the same Balding-Nichols model, with the ancestral allele frequencies set to be 0.1 and 0.5, respectively.

Trait and Covariate Models

Three non-constant covariates are included: sex, age, and another continuous covariate with the standard Gaussian distribution. In scenarios with only unrelated individuals, sex is a Bernoulli(0.5) variable and age is a variable uniformly distributed on the interval from 20 to 60. In scenarios with pedigrees, sex is fixed according to the position in the pedigree (see Figure S1), and ages for the individuals labeled 1–16 in Figure S1 are simulated uniformly and independently within 1.5 years of 75, 73, 46, 46, 43, 43, 40, 40, 18, 21, 15, 17, 13, 15, 12, and 9 years, respectively. In all cases, the values of covariates are assumed to be independent across individuals and are regenerated for each phenotype simulation replicate.

To simulate the phenotype given the genotypes, covariates, and population structure, we consider two types of binary trait models. The first type is the liability threshold model, which assumes that the disease corresponds to an underlying continuous liability distribution influenced by covariates and genetic factors with a threshold that divides the population into affected and unaffected individuals. Specifically, the phenotype Y_i of individual i is given by

$$\begin{aligned} &Y_i = 1 \text{ if and only if } L_i > 0, \\ &\text{with } L_i = \boldsymbol{X}_i\boldsymbol{\beta} + \lambda \cdot 1(G_{1,i} > 0, G_{2,i} > 0) + A_i\rho + a_i + e_i, \end{aligned} \qquad \text{(Equation 17)}$$

where $Y_i = 1$ indicates that individual i is a case subject (affected) and $Y_i = 0$ indicates that i is a control subject (unaffected); $\boldsymbol{X}_i$ is the 1×4 covariate vector, which includes an intercept term; $\boldsymbol{\beta}$ contains the fixed covariate effects; $G_{1,i}$ and $G_{2,i}$ encode the individual's genotypes at the two causal SNPs; λ is a parameter scaling the effect of the causal SNPs; $1(G_{1,i} > 0, G_{2,i} > 0)$ is an indicator function that takes value 1 when both the causal SNPs have at least one copy of the minor allele; $(a_1, \ldots, a_n)^T \sim N(0, \sigma_a^2 \mathbf{K})$ represents additive polygenic effects due to cryptic relatedness, where $\mathbf{K}$ is the kinship matrix resulting from the cryptic relatedness; $e_i \sim_{i.i.d.} N(0, \sigma_e^2)$ represents independent noise; A_i is the proportion of ancestry from population 1 for the i^{th} admixed individual; and $A_i\rho$ is an ancestry effect on the phenotype (see details in the next subsection). The ancestry term, $A_i\rho$, and additive polygenic effect, a_i, are included in the trait model only for some simulation scenarios, as a way to incorporate population structure. In other scenarios, population structure is instead introduced via ascertainment.

A second type of trait model we consider is a logistic model, in which Y_i is given by

$$Y_i \mid \boldsymbol{X_i}, G_{i1}, G_{i2}, A_i, a_i \sim \text{Bernoulli}(\mu_i), \ \log\frac{\mu_i}{1-\mu_i} = \boldsymbol{X}_i\boldsymbol{\beta} + \lambda \cdot 1(G_{1,i} > 0, G_{2,i} > 0) + A_i\rho + a_i. \quad \text{(Equation 18)}$$

Again, only some simulation scenarios (see details in the next subsection) include the ancestry proportion, A_i, and additive polygenic effect, a_i, in the logistic model to incorporate population structure, and in other scenarios, population structure is generated through ascertainment. In the simulations, for either generating model, when association is tested with a given SNP, only $\boldsymbol{X}_i$ is included as a covariate vector, whereas A_i and other SNPs are not included as fixed-effect covariates in the fitted model.

Population Structure Settings

We conduct two sets of simulations corresponding to two different types of population structure. In the "2 Subpopulations" setting, we consider a stratified population with two subpopulations, from which we sample 1,000 case subjects and 1,000 control subjects. To create sample structure, we introduce correlation between the phenotype and subpopulation membership by ascertaining the sample based on varying targeted proportions of case and control subjects from each subpopulation. Overall, the sample contains 1,000 individuals from each subpopulation, with the proportion of case subjects from the first subpopulation varying from 50% to 80%, in increments of 10%, representing settings ranging from no stratification to profound stratification. For example, if this proportion is 60%, then among the 1,000 case subjects, 600 are from subpopulation 1 and 400 are from subpopulation 2, and among the 1,000 control subjects, 400 are from subpopulation 1 and 600 are from subpopulation 2. On the population level (before exercising ascertainment), the phenotype is simulated using the models given in Equations 18 and 17, with the ancestry effect ($A_i\rho$) and the additive polygenic effect (a_i) removed from both models. In the logistic model, the values for $\boldsymbol{\beta}$ and λ are chosen so as to satisfy the following conditions: (1) the three covariates each explain an equal amount of variability in disease probability on the logit scale; (2) the covariates in $\mathbf{X}_i$ altogether explain approximately 50% of the variability of the binary trait, with the rest of the variability explained by the causal SNPs and the Bernoulli variance (or the e values, for the liability threshold model); and (3) individuals with at least one copy of the minor allele for both causal SNPs have a mean disease penetrance of approximately 15%, and the mean penetrance is 10% for other individuals. For the liability threshold model, we choose the parameters $\boldsymbol{\beta}$, λ, and σ_e^2 based on the same conditions, except that condition 1 is enforced on the liability scale, rather than on the logit scale.

In the "Admixture and Cryptic Relatedness" setting, individuals are sampled from an admixed population with two ancestral populations, with or without cryptic relatedness. When cryptic relatedness is not present, 2,000 individuals are simulated, with i.i.d. admixture proportions, A_i, sampled from a uniform distribution on [0,1], $i = 1,..., n$. When cryptic relatedness is present, 400 pedigrees of size 16 (as in Figure S1) are simulated, with an admixture proportion sampled for each pedigree (i.i.d. across pedigrees), and with all founders of a given pedigree given the same admixture proportion. In each simulation scenario, the admixture proportions for the entire sample as well as the resulting genotypes for the non-causal SNPs are generated only once and then retained for all simulation replicates. In the "Admixture and Cryptic Relatedness" setting, population structure is incorporated in the trait-generating model by the ancestry effect, $A_i\rho$, and the additive polygenic effect, a_i, in Equations 17 and 18, where $\rho > 0$ is a fixed parameter common across individuals, implying different disease prevalence values in the two ancestral populations, and consequently, that the two ancestral populations will be represented in an unbalanced way among the case subjects (and among the control subjects). Larger values for ρ and σ_a^2 are associated with more severe population structure. In the logistic model, the parameters $\boldsymbol{\beta}$, λ, ρ, and σ_a^2 are chosen to satisfy the following conditions. (1) The three covariates each explain an equal amount of variability in disease probability on the logit scale. (2) On the logit scale, considering the total variability explained by the covariates, the ancestry fixed effect ($A_i\rho$), and the additive polygenic effect (a_i), the proportion explained by the covariates achieves a specified level ranging from 0% to 100%, in increments of 20%. (3) Of the remaining proportion of variability in (2), which we describe as the proportion explained by ancestry, either it is wholly explained by the ancestry fixed effect (in the case of no cryptic relatedness) or half of it is explained by the ancestry fixed effect and half of it is explained by the additive polygenic effect (in the case when there is cryptic relatedness). (4) In the logistic model, the Bernoulli variance (or in the liability threshold model, the random error e_i) explains, on average, approximately 60% of the total variability in the binary case-control status. (5) On average, the number of case subjects in the sample equals the number of control subjects in the sample. And (6) for individuals with at least one copy of the minor allele for both causal SNPs, there is an increase of 10% in disease penetrance compared with other individuals. For the liability threshold model, the parameters are chosen based on the same criteria with conditions on the logit scale adapted to be on the liability scale.

Crohn Disease Data from WTCCC

To illustrate the use of our method, we analyze a genome-wide association study (GWAS) dataset from the WTCCC.[21] The WTCCC1 study is a case-control GWAS undertaken in the British population jointly by multiple research groups. The sample consists of ostensibly unrelated individuals, about 2,000 case subjects for each of 7 different diseases and 3,000 shared control subjects. Genotyping was conducted with the Affymetrix 500K chip. We analyze the CD data from WTCCC1. After quality control, the data set has 1,748 case subjects and 2,938 control subjects, and a total of 360,230 common SNPs (MAF > 0.05).

Kang et al.[6] have previously argued that sample structure in the WTCCC data is not adequately captured by a linear model with 100 PCs, although it is captured by an LMM. Additional motivation for the use of CARAT on the WTCCC data is based on our simulation studies (see Results), which show (1) improved power of CARAT over LMM because of the logistic mean structure of CARAT and (2) improved type 1 error of CARAT over standard logistic regression, with or without PCs, because of robustness to phenotype model misspecification.

We perform single-SNP association tests using CARAT, CARATd, and LMM. For all three methods, the covariates are sex and three SNPs (rs11209026, rs2201841, and rs10512734), all in regions previously reported and replicated to be associated with CD, based

Table 1. Empirical Type I Error Rates for a Liability Threshold Trait Model with Two Subpopulations

Proportion of Cases from[a]		Empirical Type I Error Rate of					
Population 1	Population 2	CARAT	CARATd	LMM	LOG	LOG-1	LOG-10
50%	50%	.0503	.0502	.0489	.0517*	.0523*	.0528*
60%	40%	.0501	.0501	.0509	.0815*	.0817*	.0529*
70%	30%	.0503	.0497	.0498	.169*	.164*	.0518*
80%	20%	.0495	.0488	.0498	.287*	.281*	.0524*

Empirical type 1 error at level 0.05 is evaluated based on 100,000 replicates, so the standard error is 0.00069 for every entry in the table. Asterisk (*) indicates type 1 error rates that are significantly different from the nominal level, using the z-test at level 0.05.

[a] A sample of 2,000 individuals, with 1,000 from each of the two subpopulations and with equal numbers of case and control subjects, is ascertained based on the specified proportions of case subjects from each subpopulation.

on extensive studies conducted independently of WTCCC1. rs11209026 encodes the amino acid mutation c.1172G>A (p.Arg391Gln) (GenBank: NM_153360.2) in *IL23R* (MIM: 607562) on chromosome 1p31 and has been associated with CD in a number of studies.[22–24] rs2201841 is also located in *IL23R*, has been reported as an independent association signal from rs11209026,[22] and has been associated with CD in other studies.[24–26] rs10512734 is in 5p13.1, and both the SNP and the region have been associated and replicated.[23,27–29]

Results

Simulation Studies

We first evaluate the performance of LEAP based on 500 simulation replicates for each scenario. We observe that LEAP frequently encounters algorithm failure, in which case the program terminates without producing any association testing results. Note that, as a multi-step procedure, the LEAP method first estimates the heritability of the trait, which is then assumed known and plugged into subsequent steps that predict disease liabilities and test for genetic association. However, the first step can yield a heritability estimate that is either negative or more than 100%, which is not permitted in models assumed by the steps that follow and will crash the program. In our simulation settings, LEAP has failure rates (defined to be the empirical proportion of simulation replicates in which the algorithm terminates without generating association testing results) ranging from 21.4% to 92.6% depending on the simulation scenario, with a number of settings having failure rates greater than 50%. As a result, we are unable to obtain reliable type 1 error rates or power comparisons for LEAP. The reason is that failure or not of the algorithm depends on the observed data, so the high failure rates would be expected to introduce bias in results based on only the successful runs. Therefore, we do not include LEAP in any of the comparisons with other methods.

Empirical type 1 error rates of CARAT, CARATd, and LMM are evaluated at nominal levels 0.05 and 0.001 for both the "2 Subpopulation" and "Admixture and Cryptic Relatedness" scenarios across a variety of settings under two types of trait models, based on 100,000 replicates. For comparison, we also evaluate in these settings the empirical type 1 error rates of standard logistic regression (LOG), logistic regression with the leading principal component[1] (PC) of the genetic relationship matrix $\widehat{\Psi}$ included as a covariate (LOG-1), and logistic regression with the leading ten PCs included as covariates (LOG-10). Tables 1 and 2 show the type 1 error results, at nominal level 0.05, for the "2 Subpopulation" and "Admixture and Cryptic Relatedness" settings, respectively, with phenotypes generated under the liability threshold model. The remaining type 1 error results are reported in Tables S2–S4. These results demonstrate that all three of the mixed-model methods (CARAT, CARATd, and LMM) effectively control for population stratification and admixture in these simulation settings. However, use of LOG, LOG-PC1, and LOG-PC10 can lead to inflation of type 1 error in some cases. In particular, we find that even in the absence of population structure and relatedness (row 1 of Table 1 and row 1 of Table 2), logistic regression with or without PCs can have inflated type 1 error when the logistic model does not hold, whereas CARAT is more robust and has correct type 1 error in all the simulation settings. The inflated type 1 error for the logistic regression methods probably reflects, in part, the fact that the logistic model is misspecified when the true model is a liability threshold model. In contrast, CARAT is robust to phenotype model misspecification because of the retrospective assessment of significance. These results suggest that CARAT is preferable to logistic regression, with or without PCs, even when the amount of population structure and/or relatedness is low (or absent).

Power results for CARAT, CARATd, and LMM based on 5,000 replicates at level 0.001 are shown in Figure 1. We do not include LOG, LOG-1, or LOG-10 in these power simulations, because these methods do not maintain robust control over type 1 error. For all of the simulation settings, CARAT consistently has the highest power. In particular, CARAT achieves a power gain over LMM that is statistically significant and often substantial under different trait models and with widely ranging extent of population structure of different types.

One question of interest is whether or not including the dispersion parameter (as in CARATd) provides improvement over the CARAT method. In fact, we find the

Table 2. Empirical Type I Error Rates for a Liability Threshold Trait Model with Population Admixture and Cryptic Relatedness

% Variance due to[a]			Empirical Type I Error Rate of					
Ancestry	**Covariates**	**Cryptic Relatedness**	**CARAT**	**CARATd**	**LMM**	**LOG**	**LOG-1**	**LOG-10**
0%	100%	no	.0491	.0492	.0499	.0512	.0514*	.0522*
20%	80%	no	.0493	.0492	.0497	.188*	.189*	.0526*
20%	80%	yes	.0510	.0511	.0507	.106*	.107*	.0106*
40%	60%	no	.0496	.0492	.0493	.110*	.111*	.0520*
40%	60%	yes	.0500	.0499	.0507	.127*	.126*	.0125*
60%	40%	no	.0497	.0497	.0501	.316*	.316*	.0524*
60%	40%	yes	.0488	.0489	.0489	.163*	.163*	.162*
80%	20%	no	.0496	.0495	.0497	.350*	.350*	.0514*
80%	20%	yes	.0492	.0494	.0496	.182*	.182*	.0182*
100%	0%	no	.0506	.0506	.0512	.369*	.369*	.0524*
100%	0%	yes	.0513	.0514*	.0514*	.195*	.195*	.195*

Empirical type 1 error at level 0.05 is evaluated based on 100,000 replicates, so the standard error is 0.00069 for every entry in the table. Asterisk (*) indicates type 1 error rates that are significantly different from the nominal level, using the z-test at level 0.05.

[a]The percentages are defined to be the variance, on the liability scale, explained by either the ancestry effects (admixture and cryptic relatedness) or the covariate effects, divided by the total variance explained by the two types of effects, indicating the relative impact of ancestry versus covariates on the phenotype.

opposite to be true: CARAT has consistently higher empirical power than CARATd in all simulation scenarios. We also observe that the power difference tends to grow as population structure becomes stronger, and in the special case where there is no population structure, CARAT and CARATd perform almost identically. We propose a likely reason for this in the context of a logistic trait model as follows. In the absence of population structure, the quasi-likelihood trait model is a correctly specified model for the sample. It follows that, in the case of no population structure, CARATd can estimate σ_g^2 in a consistent fashion to be close to the true value 1. Therefore, almost nothing is lost by including this redundant nuisance parameter, provided that the sample size is large. As population structure strengthens, however, the prospective model becomes more subject to misspecification, and the estimated σ_g^2 is no longer close to 1, resulting in the divergence of the performance of CARAT and CARATd.

For the "2 Subpopulation" settings, in which the miscalibration of the LOG-10 statistic is not extreme, we also compare the power of LOG-10 and CARAT for detecting association under either the liability threshold or logistic phenotype model. To make a fair power comparison, we first use 500,000 simulated replicates under the null model to recalibrate both statistics to have correct type 1 error. The results, reported in Table S5, show that CARAT consistently outperforms LOG-10 in terms of power, as well as in type 1 error (Tables 1 and S2) in this scenario.

In the results presented, the correlation matrix $\mathbf{\Phi}$ of Equation 5 in the trait model is estimated by $\widehat{\mathbf{\Psi}}$ given in Equation 6. Use of the alternative estimator, $\widehat{\mathbf{\Phi}}$ of Equation 7, for $\mathbf{\Phi}$ produces very similar results in all scenarios (results not shown for brevity).

Analysis of Crohn Disease Data from WTCCC

Genome-wide association testing for CD is performed on 360,230 SNPs using CARAT, CARATd, and LMM, with adjustment for covariates (sex and three SNPs previously identified as associated), as described in Material and Methods. The resulting genomic control inflation factors for the three methods are $\lambda_{GC} = 1.002$, 1.002, and 1.004, respectively. Table 3 shows the 17 genetic regions with at least one SNP for which the p value is less than 10^{-6} using at least one of the three tests. All 17 regions identified by our analysis have previously been associated with CD or inflammatory bowel disease (IBD [MIM: 266600]). These include many extensively replicated loci and genes.[23,30–34] For example, at 16q12.1, a cluster of associated SNPs are within or in close proximity to *NOD2* (*CARD15* [MIM: 605956]), a well-known CD-susceptibility gene.[30,34–36] The associated SNPs in the 2q37.1 region are close to *ATG16L1* (MIM: 610767), which has been widely associated with CD.[21,32]

In Table 3, a somewhat surprising result is that there is still significant association with a SNP in 5p13.1, even though rs10512734, a known risk SNP from the same region, has been controlled for as a covariate in the analysis. In fact, after rs10512734 is controlled for, four SNPs in close proximity of rs10512734 retain evidence of association (Table 4). In Table 4, we compare the p values of these SNPs based on the analyses with and without rs10512734 included as a covariate (in both analyses, we keep sex and the other two covariate SNPs, rs11209026 and rs2201841, as covariates). We observe that adjustment for rs10512734 results in a boosted association signal for three of the nearby SNPs. For example, rs6883686, which is 8,928 base pairs from rs10512734, shows no evidence of association (p value > 0.9) when rs10512734 is not adjusted for.

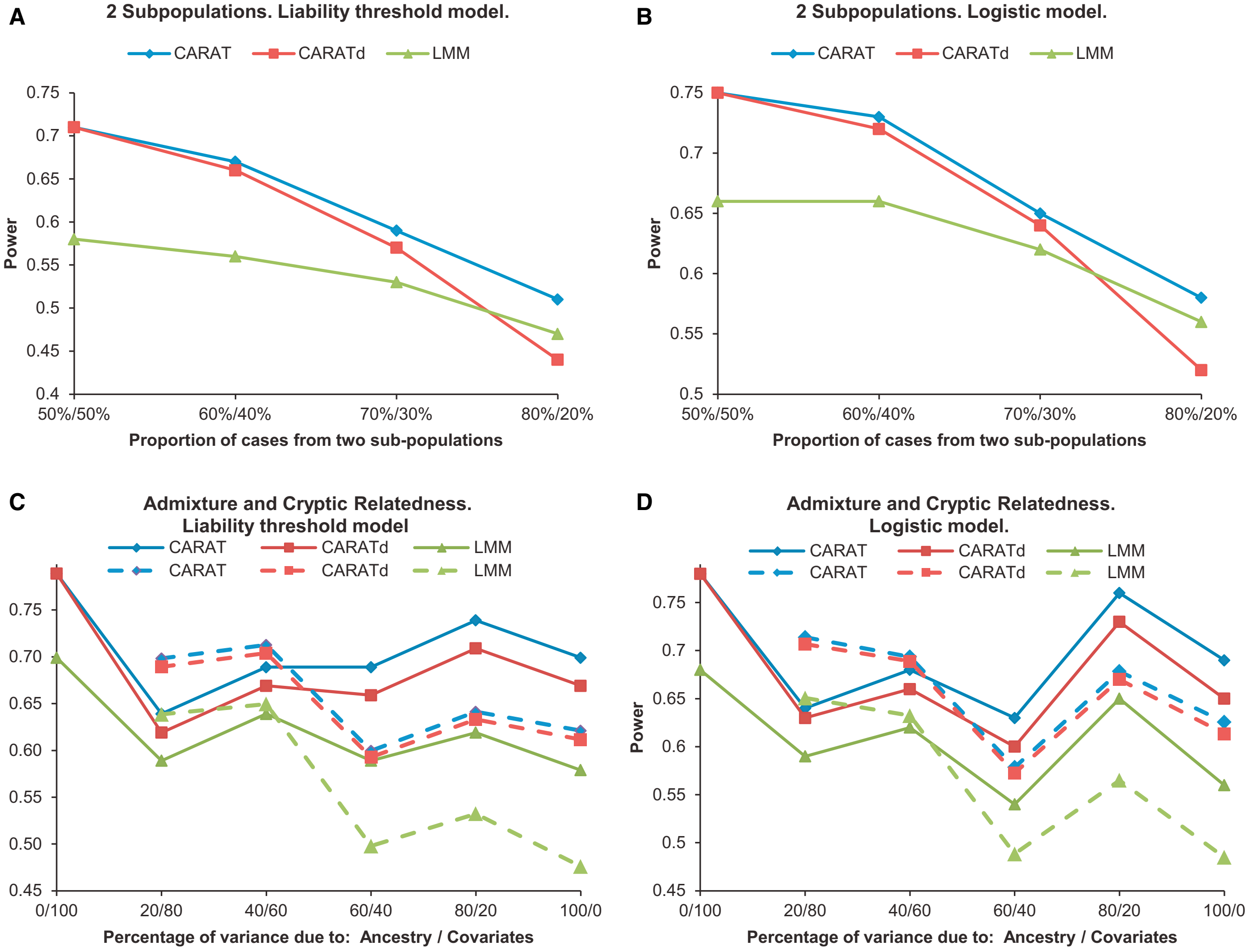

Figure 1. Empirical Power of CARAT, CARATd, and LMM

Empirical power is based on 5,000 replicates. An upper bound for the standard error of the empirical power is 0.007. For a difference between two empirical power values, the upper bound on the standard error is 0.01.

(A and B) Power results in the "2 Subpopulations" setting, in which the leftmost value on the horizontal axis (50%/50%) corresponds to no population stratification and the rightmost value (80%/20%) corresponds to profound stratification. Power results when the phenotype is generated according to a liability threshold model with two subpopulations (A) or to a logistic regression model with two subpopulations (B).

(C and D) Power results in the "Admixture and Cryptic Relatedness" setting, for which the percentages on the horizontal axis are defined so that the numerator is the variance, on the liability or logistic scale, explained by the ancestry and cryptic relatedness effects (the first number) or by the covariate effects (the second number), and the denominator is the total variance explained by the two types of effects. The horizontal scale indicates the relative impact of ancestry versus covariates on the phenotype, with the far left corresponding to no effect of ancestry and strong effects of covariates and the far right corresponding to no effect of covariates and a strong effect of ancestry. The dotted lines denote settings with cryptic relatedness, and the solid lines denote settings without cryptic relatedness. Power results when the phenotype is generated according to a liability threshold model (C) or according to a logistic regression model (D) in the "Admixture and Cryptic Relatedness" setting.

After adjusting for rs10512734, however, suggestive evidence of association is found for rs6883686, with a dramatic change in the p value from 0.99 to 5.1×10^{-4} using CARAT. To investigate the association signals in 5p13.1 more closely, we include each of the individual SNPs in 5p13.1 that are present in the data as a covariate and examine the p values of the other SNPs. We find that no single SNP within 5p13.1 is able to explain all the association signal in this region. These results suggest either that the 5p13.1 region contains more than one independently associated SNP or that there exists at least one untyped CD susceptibility variant in this region that is not well tagged by any single SNP in the dataset.

To further investigate the possibility of untyped causal SNPs in 5p13.1, we obtain dense genotype data through imputation based on the phase 3 data of the 1000 Genomes project.[37] We apply MaCH[38,39] to phase the WTCCC genotype data and use Minimac3[40,41] for genotype imputation. We examine all SNPs present in the 1000 Genomes data that are located in a 20 Mb region around 5p13.1 to determine whether the association signal in that region can be attributed to any single imputed SNP.

Table 3. **Regions of Strongest Association with Crohn Disease in WTCCC Data**

Region	Top SNP	Position[a]	CARAT	CARATd	LMM
1p13.1	rs12078461	117,273,572	4.7×10^{-15}*	4.8×10^{-15}*	5.8×10^{-10}
1q41	rs1933641	216,535,584	1.0×10^{-17}*	1.0×10^{-17}*	1.8×10^{-11}
2p12	rs11887827	81,519,665	9.5×10^{-7}	9.8×10^{-7}	2.3×10^{-7}*
2q37.1	rs10210302	233,823,578	4.4×10^{-12}	4.4×10^{-12}	1.7×10^{-12}*
3p24.3	rs9839841	16,454,562	1.7×10^{-16}	1.6×10^{-16}	4.1×10^{-17}*
4p15	rs1553460	17,804,959	$<10^{-18}$	$<10^{-18}$	$<10^{-18}$
5p13.1	rs17234657	40,437,266	3.7×10^{-8}	3.8×10^{-8}	1.8×10^{-8}*
5q23.1	rs2416472	117,033,845	2.6×10^{-10}	2.6×10^{-10}	3.0×10^{-10}
7p14.1	rs1525791	39,123,083	7.9×10^{-7}	8.1×10^{-7}	5.8×10^{-7}
10q21.2	rs10761659	64,115,570	6.1×10^{-7}	6.1×10^{-7}	6.5×10^{-7}
10q24.2	rs10883371	101,282,445	3.4×10^{-7}	3.4×10^{-7}	2.7×10^{-7}
11q23.2	rs17116117	113,306,801	$<10^{-18}$	$<10^{-18}$	$<10^{-18}$
14q13.2	rs10483456	35,105,918	6.3×10^{-16}	6.4×10^{-16}	6.0×10^{-16}
16p11.2	rs4471699	30,227,808	7.6×10^{-13}	7.6×10^{-13}	1.1×10^{-13}*
16q12.1	rs2076756	49,314,382	5.9×10^{-12}*	6.1×10^{-12}*	2.4×10^{-11}
18p11.21	rs2542151	12,769,947	9.4×10^{-9}*	9.6×10^{-9}*	2.1×10^{-8}
219q13.2	rs41537748	44,449,412	2.2×10^{-8}*	2.2×10^{-8}*	6.3×10^{-8}

Association results are based on an analysis adjusting for sex and three known associated SNPs (rs11209026, rs2201841, and rs10512734) as covariates. Asterisk (*) indicates a p value less than half of the largest of the p values yielded by the three methods.
[a]Base pair position is based on assembly NCBI36.

Specifically, we include each of the imputed SNPs as a covariate (sex is included as an additional covariate) when testing the other SNPs, typed and imputed, within 5p13.1. We find that no single imputed SNP in the region is able to fully explain the association signal at 5p13.1. Among the imputed SNPs, the smallest CARAT p value, 1.2×10^{-12}, is achieved by two SNPs in perfect LD in the imputed data, located at 40,410,043 and 40,410,739 on chromosome 5, respectively. After adjusting for these two SNPs, there remain additional association signals in 5p13.1, with the smallest p value, 4.8×10^{-6}, achieved by a typed SNP, rs11957215. Overall, our results indicate the presence of multiple susceptibility variants within 5p13.1. The 5p13.1 region has previously been reported and replicated as associated with CD,[23,27] and evidence has been found that variants in this region correlate with the expression level of *PTGER4* (MIM: 601586),[23] a gene that has been implicated in IBD in murids. We are the first to report evidence suggesting the existence of multiple independent risk variants within 5p13.1.

The aforementioned observation illustrates that, in association mapping, including known associated SNPs as covariates has the potential to enable discoveries of new association signals and to generate interesting insights into the genetic architecture of the trait. More examples are provided in Table 5, which compares the p values using models with no covariates and those using models with sex and three SNPs as covariates.

Computation Time

The main computational burden of implementing CARAT comes from the eigendecomposition of the genetic relationship matrix. Despite the potential complexity of this decomposition, its impact on the computational feasibility of the method is mitigated by two factors: (1) the decomposition incurs no extra cost beyond LMM and PC-based methods, which require the same decomposition or one of comparable computational complexity;[1,8,9] (2) it needs to be done only once per genome scan. Once the eigendecomposition is available for the genetic relationship matrix, the additional computational cost of CARAT is $O(n^2)$ for fitting the null model and $O(n)$ for calculating the test statistic. As for LMM, the time complexity can be further reduced in the case of a low-rank genetic relationship matrix.[8] Overall, CARAT enjoys the same computational scalability as LMM[8,9] in large genetic association data.

CARAT is implemented in a freely downloadable software package (see Web Resources). We report some example run times of CARAT for analysis of real and simulated data. Using a single processor on a machine with 6 core Intel Xeon 3.50 GHz CPUs and 32 GB RAM, CARAT takes less than 6 s to fit the null model on the WTCCC data with 4,686 individuals, and approximately 131 s to analyze a total of 360,230 genome-wide SNPs, assuming the genetic relationship is available. For a sample of 20,000 individuals, fitting the null models takes 107 s, and the genome-wide analysis takes an additonal 9.4 min

Table 4. Association Signals at 5p13.1 in WTCCC Data

SNP	Position	Controlled for rs10512734?	p Value via CARAT
rs10473185	40410924	yes	1.8×10^{-6}
		no	1.0×10^{-4}
rs10512734	40429362	yes	–
		no	1.7×10^{-7}
rs17234657	40437266	yes	3.7×10^{-8}
		no	4.5×10^{-13}
rs6883686	40438290	yes	5.1×10^{-4}
		no	.99
rs6885315	40465299	yes	5.2×10^{-4}
		no	.93

After rs10512734 is included as a covariate (together with three other covariates), four SNPs in close proximity still show association signal. p values for these SNPs are compared with and without adjusting for rs10512734. For three of these SNPs, controlling for rs10512734 boosts association signal.

(566 s). These results demonstrate that CARAT computes rapidly for large-scale genome-wide association studies.

Discussion

Population structure is a widespread confounding factor in genetic association studies. It is well known that failure to account for population structure can lead to both false positives and false negatives in genetic association mapping. We have developed CARAT, an association testing method for binary traits in samples with population structure. Compared to existing methods such as ROADTRIPS and LTMLM, CARAT features the ability to account for covariate information. Like LMM, CARAT includes an additive polygenic variance component to account for population structure. Unlike LMM, CARAT specifically accommodates binary traits by modeling covariate effects on the logit scale and by accounting for the dependence of the variance on the mean. As a result, CARAT gains power over LMM in case-control association testing. Moreover, we take an estimating equation and score test approach, which ensures that CARAT is computationally rapid for large-scale studies. In addition, CARAT uses a retrospective analysis to assess significance of the test statistic, which equips the method with robustness to misspecification of the phenotype model. We provide a computationally efficient implementation of CARAT in a freely downloadable software package.

Table 5. Including Covariates Can Help Boost Association Signal

Chr	SNP	Covariates[a] Included?	p Value via CARAT
5	rs12657249	yes	1.8×10^{-6}
		no	5.3×10^{-4}
18	rs12966840	yes	9.6×10^{-9}
		no	1.8×10^{-7}
18	rs1942868	yes	1.0×10^{-7}
		no	1.2×10^{-6}
18	rs9954415	yes	6.7×10^{-8}
		no	8.1×10^{-7}
19	rs10421478	yes	2.1×10^{-8}
		no	1.7×10^{-7}

[a] We compare p values based on the analysis with no covariates with those based on the analysis with the following covariates: sex, rs11209026, rs2201841, and rs10512734.

We demonstrate the validity and power of CARAT through simulation studies. In particular, CARAT consistently outperforms LMM under diverse simulation scenarios with different types of trait models, widely ranging levels of covariate effects, and varying degrees of population stratification and admixture. In our simulations, we found the recently developed program LEAP to be unstable, as reflected in frequent failures that prevented the algorithm from producing association testing results. Therefore, we were not able to compare LEAP with the other methods. We compare CARAT to logistic regression with or without PCs included as covariates. We find that CARAT is more robust to misspecification of the phenotype model, while the type 1 error of the logistic regression methods can be inflated in the presence of model misspecification. In addition to demonstrating the improvement of CARAT over previously proposed methods, we also consider the question of whether there is any benefit to including a dispersion parameter in the CARAT model. To test this, we develop CARATd, a version of CARAT that includes the dispersion parameter. We find that the power of CARAT is consistently higher than that of CARATd in our simulations, suggesting that the dispersion parameter should not be included in the model.

We applied CARAT to association mapping of CD in the WTCCC dataset. In the analysis, we included as covariates sex and three known associated SNPs: rs11209026, rs2201841, and rs10512734. We found 17 CD-associated regions, all of which have been previously associated with CD or with a closely related phenotype, IBD. In addition, we observed in the CD analysis that adjusting for relevant covariates can enhance the ability to detect association and can lead to interesting insights into the genetics of a disease. In particular, by including known associated SNPs as covariates, we have found evidence that 5p13.1 might contain multiple independent associated SNPs.

Although the trait model assumed by CARAT uses a logit link function, the method is adaptable to other link functions as deemed appropriate. Ideally, a good choice of the link function should provide a scale on which the covariate effects are additive. As an example, the complementary log-log link might be considered when the case-control status in the data is asymmetric, which might be the case if the binary trait is a truncation of some unobserved count (for instance, "no tumor" versus "some tumors," where the number of tumors is the unobserved count). Moreover, CARAT can be generalized to the analysis of non-binary traits by assuming Equation 1 and

$$\mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{G}) = \sigma^2 \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Sigma} \boldsymbol{\Gamma}^{1/2}, \quad \text{where } \boldsymbol{\Gamma} = \mathrm{diag}\{V(\mu_1), \cdots, V(\mu_n)\},$$

where the link function $g(\bullet)$ in Equation 1 and the marginal variance function $V(\bullet)$in the equation above can be chosen based on a marginal distribution from the exponential family tailored to the data type under consideration. For example, to analyze Poisson count data, one could use $g(\mu) = \log(\mu)$ and $V(\mu) = \mu$. We note that although CARAT sets the dispersion parameter σ^2 to be 1 for binary traits, it could be beneficial to include σ^2 as a free parameter for non-binary data to capture possible over-dispersion.

In our presentation of CARAT, the genetic relationship matrix $\boldsymbol{\Phi}$ in Equation 5 is estimated based on either Equation 6 or 7. However, the method is able to accommodate other choices. For example, Yang et al.[12] show that the power of LMM can be improved by estimating the genetic relationship matrix using genome scan data with the tested SNP excluded. The same strategy can be applied in CARAT. In addition, CARAT can be adapted to incorporate more than two additive correlation components in $\boldsymbol{\Sigma}$, with the extra variance component parameters estimated using estimating equations constructed in a similar way as for Equation 11. It has been argued that including additional variance components could provide improved protection against spurious association with highly differentiated SNPs[42] and could provide simultaneous adjustment for multiple levels of sample structure including population stratification and familial relatedness.[5] Furthermore, CARAT can be combined with methods that model population structure as fixed effects, such as EIGENSTRAT, by including inferred ancestry information as covariates.

Recently, the GCAT method, which models genetic variation in structured populations, has been described.[43] In the context of quantitative traits, GCAT was shown to provide an advantage for association analysis in the presence of population structure, compared to LMM and linear regression with PCs included as covariates. For binary traits, Song et al.[43] do not show their simulation results, but report that "all methods performed similarly well in terms of producing correct P values that were robust to structure," indicating that GCAT does not provide improvement over existing methods for binary traits. Furthermore, the way in which covariates would be incorporated for a binary trait in GCAT is not explicitly described in Song et al. Our results on CARAT indicate that when covariates play an important role for a binary trait, the way that covariates are incorporated can affect power. Specifically, when covariate effects are large, a logistic model for the effect of covariates on the trait is more successful than a linear one in providing high power for association.

For testing association with rare variants, all the methods we compare (LMM, CARAT, CARATd, and logistic regression with PCs) will encounter similar challenges. One potential problem that has been pointed out[44] is that in some cases, rare variants might have different population histories than common variants, so genetic relationship matrices and PCs calculated based on common variants might not be applicable to rare variants. Low MAF of the tested variant does not by itself cause problems for any of these analysis methods when the sample size is sufficiently large. However, if the minor allele count of a tested variant is too low, the asymptotic assumptions underlying the assessment of significance for CARAT, for example, might fail.

Appendix A: Type I Error with the Prospective Version of CARAT

Based on the prospective quasi-likelihood model described in Material and Methods and Equation 6, the prospective quasi score test statistic is given by

$$T_{\mathrm{pro}} = U_0^2 \Big/ \left[\boldsymbol{G}^T \boldsymbol{K} \boldsymbol{G} - \boldsymbol{G}^T \boldsymbol{K} \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{K} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{K} \boldsymbol{G} \right],$$

where $\boldsymbol{K} = \widehat{\boldsymbol{\Gamma}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\boldsymbol{\Gamma}}_0$. To perform the association test with the prospective version of CARAT, we assume that T_{pro} follows a χ_1^2 distribution under the null hypothesis. The accuracy of this null distribution is not robust against misspecification of the prospective model. Our simulation studies show that under many realistic scenarios, the prospective version of CARAT does not offer correct control over type 1 error. Examples of such scenarios are shown in Table S1.

Supplemental Data

Supplemental Data include one figure and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.12.012.

Acknowledgments

This study was supported by the NIH grant R01 HG001645 (to M.S.M.). This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from the WTCCC website (see Web Resources). Funding for the project was provided by the Wellcome Trust under awards 076113 and 085475.

Received: August 17, 2015
Accepted: December 11, 2015
Published: January 28, 2016

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes Project, http://www.1000genomes.org/

CARAT, http://www.stat.uchicago.edu/~mcpeek/software/index.html

MaCH software, http://csg.sph.umich.edu/abecasis/MACH/index.html

Minimac3 software, http://genome.sph.umich.edu/wiki/Minimac3

OMIM, http://www.omim.org/

RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq

WTCCC, http://www.wtccc.org.uk

References

1. 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.
2. Chanock, S.J., and Hunter, D.J. (2008). Genomics: when the smoke clears. Nature *452*, 537–538.
3. Xing, G., and Xing, C. (2010). Adjusting for covariates in logistic regression models. Genet. Epidemiol. *34*, 769–771, author reply 772.
4. Thornton, T., and McPeek, M.S. (2010). ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am. J. Hum. Genet. *86*, 172–184.
5. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. *38*, 203–208.
6. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354.
7. Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., and Buckler, E.S. (2010). Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. *42*, 355–360.
8. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods *8*, 833–835.
9. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
10. Zhang, J., Niyogi, P., and McPeek, M.S. (2009). Laplacian eigenfunctions learn population structure. PLoS ONE *4*, e7928.
11. 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.
12. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. *46*, 100–106.
13. Hayeck, T.J., Zaitlen, N.A., Loh, P.-R., Vilhjalmsson, B., Pollack, S., Gusev, A., Yang, J., Chen, G.-B., Goddard, M.E., Visscher, P.M., et al. (2015). Mixed model with correction for case-control ascertainment increases association power. Am. J. Hum. Genet. *96*, 720–730.
14. Weissbrod, O., Lippert, C., Geiger, D., and Heckerman, D. (2015). Accurate liability estimation improves power in ascertained case-control studies. Nat. Methods *12*, 332–334.
15. Liang, K.-Y., and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. Biometrika *73*, 13–22.
16. Zeger, S.L., and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics *42*, 121–130.
17. McCullagh, P., and Nelder, J. (1989). Generalized Linear Models, Second Edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability (Taylor & Francis).
18. Skrondal, A., and Rabe-Hesketh, S. (2007). Redundant overdispersion parameters in multilevel models for categorical responses. J. Educ. Behav. Stat. *32*, 419–430.
19. Heyde, C. (1997). Quasi-likelihood and Its Application: A General Approach to Optimal Parameter Estimation (Springer).
20. Balding, D., and Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. In Human Identification: The Use of DNA Markers In Human Identification: The Use of DNA Markers, B.S. Weir, ed. (Springer).
21. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., et al.; Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
22. Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science *314*, 1461–1463.
23. Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., Vermeire, S., Dewit, O., de Vos, M., Dixon, A., et al. (2007). Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. *3*, e58.
24. Glas, J., Seiderer, J., Wetzke, M., Konrad, A., Török, H.-P., Schmechel, S., Tonenchi, L., Grassl, C., Dambacher, J., Pfennig, S., et al. (2007). rs1004819 is the main disease-associated IL23R variant in German Crohn's disease patients: combined analysis of IL23R, CARD15, and OCTN1/2 variants. PLoS ONE *2*, e819.
25. Lappalainen, M., Halme, L., Turunen, U., Saavalainen, P., Einarsdottir, E., Färkkilä, M., Kontula, K., and Paavola-Sakki, P. (2008). Association of IL23R, TNFRSF1A, and HLA-DRB1* 0103 allele variants with inflammatory bowel disease phenotypes in the Finnish population. Inflamm. Bowel Dis. *14*, 1118–1124.
26. Wang, M.-H., Okazaki, T., Kugathasan, S., Cho, J.H., Isaacs, K.L., Lewis, J.D., Smoot, D.T., Valentine, J.F., Kader, H.A., Ford, J.G., et al. (2012). Contribution of higher risk genes and European admixture to Crohn's disease in African Americans. Inflamm. Bowel Dis. *18*, 2277–2287.
27. Yamazaki, K., Onouchi, Y., Takazoe, M., Kubo, M., Nakamura, Y., and Hata, A. (2007). Association analysis of genetic variants in IL23R, ATG16L1 and 5p13.1 loci with Crohn's disease in Japanese patients. J. Hum. Genet. *52*, 575–583.
28. Franke, A., Hampe, J., Rosenstiel, P., Becker, C., Wagner, F., Häsler, R., Little, R.D., Huse, K., Ruether, A., Balschun, T., et al. (2007). Systematic association mapping identifies NELL1 as a novel IBD disease gene. PLoS ONE *2*, e691.
29. Kenny, E.E., Pe'er, I., Karban, A., Ozelius, L., Mitchell, A.A., Ng, S.M., Erazo, M., Ostrer, H., Abraham, C., Abreu, M.T., et al. (2012). A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. PLoS Genet. *8*, e1002559.

30. Hugot, J.-P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J.-P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M., et al. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature *411*, 599–603.
31. McGovern, D.P., Hysi, P., Ahmad, T., van Heel, D.A., Moffatt, M.F., Carey, A., Cookson, W.O., and Jewell, D.P. (2005). Association between a complex insertion/deletion polymorphism in NOD1 (CARD4) and susceptibility to inflammatory bowel disease. Hum. Mol. Genet. *14*, 1245–1250.
32. Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F.M., Briggs, J., et al. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat. Genet. *39*, 207–211.
33. Parkes, M., Barrett, J.C., Prescott, N.J., Tremelling, M., Anderson, C.A., Fisher, S.A., Roberts, R.G., Nimmo, E.R., Cummings, F.R., Soars, D., et al.; Wellcome Trust Case Control Consortium (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat. Genet. *39*, 830–832.
34. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al.; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955–962.
35. Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H., et al. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature *411*, 603–606.
36. Elding, H., Lau, W., Swallow, D.M., and Maniatis, N. (2011). Dissecting the genetics of complex inheritance: linkage disequilibrium mapping provides insight into Crohn disease. Am. J. Hum. Genet. *89*, 798–805.
37. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
38. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.
39. Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. Annu. Rev. Genomics Hum. Genet. *10*, 387–406.
40. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.
41. Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2015). minimac2: faster genotype imputation. Bioinformatics *31*, 782–784.
42. Sul, J.H., and Eskin, E. (2013). Mixed models can correct for population structure for genomic regions under selection. Nat. Rev. Genet. *14*, 300.
43. Song, M., Hao, W., and Storey, J.D. (2015). Testing for genetic associations in arbitrarily structured populations. Nat. Genet. *47*, 550–554.
44. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.