Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models

Han Chen,[1,8] Chaolong Wang,[1,2,8] Matthew P. Conomos,[3] Adrienne M. Stilp,[3] Zilin Li,[1,4] Tamar Sofer,[3] Adam A. Szpiro,[3] Wei Chen,[5] John M. Brehm,[5] Juan C. Celedón,[5] Susan Redline,[6] George J. Papanicolaou,[7] Timothy A. Thornton,[3] Cathy C. Laurie,[3] Kenneth Rice,[3] and Xihong Lin[1,*]

Linear mixed models (LMMs) are widely used in genome-wide association studies (GWASs) to account for population structure and relatedness, for both continuous and binary traits. Motivated by the failure of LMMs to control type I errors in a GWAS of asthma, a binary trait, we show that LMMs are generally inappropriate for analyzing binary traits when population stratification leads to violation of the LMM's constant-residual variance assumption. To overcome this problem, we develop a computationally efficient logistic mixed model approach for genome-wide analysis of binary traits, the generalized linear mixed model association test (GMMAT). This approach fits a logistic mixed model once per GWAS and performs score tests under the null hypothesis of no association between a binary trait and individual genetic variants. We show in simulation studies and real data analysis that GMMAT effectively controls for population structure and relatedness when analyzing binary traits in a wide variety of study designs.

Introduction

Population stratification is a major source of confounding in genetic association studies.[1] With the recent development of computationally efficient algorithms, linear mixed models (LMMs) have become popular in genome-wide association studies (GWASs) for controlling population stratification, as well as familial or cryptic relatedness.[2–10] However, in human genetics, GWASs are often conducted using binary traits; for example, case-control studies sample disease subjects (case subjects) and disease-free subjects (control subjects) and collect their genotype and exposure data retrospectively. Cohort studies follow a group of healthy subjects prospectively and collect their subsequent status evaluations with respect to the binary disease outcome. Despite the popularity of the use of LMMs in binary trait GWASs, their reliance on a generally invalid assumption appears to have been overlooked. Specifically, as typically used, LMMs assume that the trait has constant residual variance, which is usually violated by binary traits in the presence of covariates. As a consequence, we will show that in the presence of population stratification, fitting LMMs to binary traits can fail to control for type I error rates and yield incorrect p value estimates. Typical use of LMMs also ignores the biased sampling in case-control studies, which leads to biased effect estimates.

Our study of LMMs for binary traits was motivated by analyses of the binary trait asthma in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), originally performed using LMMs with three random effects to account for genetic relatedness as well as household and block group membership from its multi-stage sampling design. For asthma, ancestry is a known confounder of genetic associations in Hispanic/Latino populations, and in HCHS/SOL the proportion of asthma cases varies widely by ancestry group (e.g., 25.6% in Puerto Ricans versus 4.1% in South Americans, see Table 1).[11] Despite regression adjustment for ancestry by including principal components (PCs)[12,13] and indicators for ancestry group as covariates, the LMM results for this trait appear invalid (Figure 1), showing clear conservatism/anti-conservatism for variants with the minor allele that is respectively less common/more common in Puerto Rican heritage versus all other ancestry groups.

Based on published case-control GWASs,[14–17] this concerning behavior does not appear to be well known. Users of LMMs for binary traits have appealed to Pirinen et al.,[9] who showed that linear models are a sufficiently close approximation to logistic models for p value calculations when the effect size of a genetic variant is small and in the absence of population stratification. We will show that in the presence of population stratification, fitting LMMs to binary traits in both cohort and case-control studies can yield incorrect type I error rates in GWASs due to heteroscedasticity, that is, residual variances in a model that are not constant with respect to different values of covariates.[18] Like linear regression, LMMs assume

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [2]Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672, Singapore; [3]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [4]Department of Mathematics, Tsinghua University, Beijing 100084, P. R. China; [5]Division of Pediatric Pulmonary Medicine, Allergy and Immunology, Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh, Pittsburgh, PA 15224, USA; [6]Division of Sleep and Circadian Disorders, Departments of Medicine and Neurology, Brigham and Women's Hospital, Boston, MA 02115, USA; [7]Prevention and Population Sciences Program, Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA
[8]These authors contributed equally to this work
*Correspondence: xlin@hsph.harvard.edu

http://dx.doi.org/10.1016/j.ajhg.2016.02.012.

©2016 by The American Society of Human Genetics. All rights reserved.

Table 1. Numbers of Asthma Case Subjects and Control Subjects in Six Hispanic/Latino Background Groups in HCHS/SOL after Quality Control of Samples

Group	Case Subjects	Control Subjects	Sample Size	Proportion of Case Subjects	Trait Variance
Central American	55	1,173	1,228	4.5%	0.043
Cuban	182	1,722	1,904	9.6%	0.086
Dominican	99	933	1,032	9.6%	0.087
Mexican	172	4,189	4,361	3.9%	0.038
Puerto Rican	478	1,390	1,868	25.6%	0.190
South American	34	795	829	4.1%	0.039

Trait variance is calculated as the product of proportions of case and control subjects.

homoscedasticity, that is, residual variances are constant regardless of covariate values (Figure 2).[19]

Recently, liability threshold linear mixed models have been proposed for case-control studies.[20,21] Fitting these models require knowing disease prevalence and accurate heritability and liability estimates in the underlying study population, which might not be available or might be difficult to specify in practice, especially when disease prevalence differs between subpopulation groups. For example, in the HCHS/SOL, the asthma prevalence rates were different in different ancestry background groups;[11] in case-control studies in the presence of population stratification, it is difficult to specify the disease prevalence that can be used for analysis when underlying subpopulation groups are unknown and the disease prevalence rates vary between underlying subpopulations. We will show that these methods can fail to control for type I error rates in the presence of moderate or strong population stratification.

To address these issues, we propose using logistic mixed models[22] to account for both population stratification and relatedness in genetic association studies of binary traits, while naturally allowing for the non-constant variance of these traits. Because logistic mixed models are computationally more expensive than LMMs and regular logistic models, they have not been used in large-scale GWASs. Although SAS PROC GLIMMIX allows for fitting logistic mixed models with a genetic relationship matrix, it is not feasible for large-scale GWASs because of its computational burden associated with fitting a large number of logistic mixed models, one per variant, across the genome. We develop the generalized linear mixed model association test (GMMAT), which requires fitting a logistic mixed model under the null hypothesis only once per GWAS, and is hence computationally feasible for large-scale GWASs.

Specifically, GMMAT first fits the null logistic mixed model that includes as fixed effects only covariates, including ancestry PCs, but no individual genetic variants, and uses random effects to account for residual population stratification not captured by fixed effects PCs as well as relatedness. When fitting this null model, GMMAT uses penalized quasi-likelihood (PQL)[22] and the computationally efficient average information restricted maximum likelihood (AI-REML) algorithm.[6,23] This fitted null model is the same for all genetic variants in a GWAS. GMMAT then applies a score test[24] for each individual genetic variant to test for its association with a binary trait. The computational cost of the score test for each genetic variant is negligible compared to the cost of fitting the null logistic mixed model, so the procedure is computationally fast, even for large-scale GWASs.

As a full-modeling approach, GMMAT accounts for the binary nature of the trait, in particular its non-constant variance in the presence of covariates, and so correctly controls type I error rates in the presence of population stratification and relatedness. GMMAT can also allow for complex sampling designs such as hierarchical designs and allow for unobserved shared environmental effects among sampled individuals by incorporating multiple random effects.

Material and Methods

Logistic Mixed Models and Score Tests

For a single-variant test, we consider the following logistic mixed model:

$$\text{logit}(\pi_\text{i}) = \mathbf{X}_i\boldsymbol{\alpha} + G_i\beta + b_i,$$

where $\pi_i = P(y_i = 1 \,|\, \mathbf{X}_i, G_i, b_i)$ is the probability of a binary phenotype (e.g., disease status) for subject i, conditional on their covariates, genotype, and random effects b_i, $\mathbf{X}_i$ is a $1 \times p$ row vector of covariates for subject i, $\boldsymbol{\alpha}$ is a $p \times 1$ column vector of fixed covariate effects including an intercept, G_i is the genotype of a genetic variant for subject i, and β is the genotype effect. We assume that $\mathbf{b} \sim N(0, \sum_{k=1}^{K}\tau_k\mathbf{V}_k)$ is an $n \times 1$ column vector of random effects, where τ_k are the variance component parameters and $\mathbf{V}_k$ are known $n \times n$ relatedness matrices. When the number of variance components $K = 1$, $\mathbf{V}_1$ is usually the genetic relationship matrix estimated from a large number of genetic variants. We allow for multiple random effects to account for complex sampling designs, e.g., hierarchical designs, and environmental covariance structure. The binary phenotypes y_i are assumed to be independent conditional on the random effects $\mathbf{b}$.

To perform the score test for the null hypothesis H_0: $\beta = 0$, we need to fit the null logistic mixed model, which is the same for all genetic variants, as

$$\text{logit}(\pi_{i0}) = \mathbf{X}_i\boldsymbol{\alpha} + b_i, \qquad \text{(Equation 1)}$$

where $\pi_{i0} = P(y_i = 1 \,|\, \mathbf{X}_i, b_i)$. We fit Equation 1 using the penalized quasi-likelihood (PQL) method.[22] Specifically, let $\mathbf{W} = diag\{v_{i0}\}$, where $v_{i0} = \pi_{i0}(1 - \pi_{i0})$, $\mathbf{X} = \begin{pmatrix}\mathbf{X}_1^T & \mathbf{X}_2^T & \cdots & \mathbf{X}_n^T\end{pmatrix}^T$ be an $n \times p$ covariate matrix including an intercept, and $\tilde{\mathbf{Y}}$ be the "working vector" with components $\tilde{Y}_i = \mathbf{X}_i\boldsymbol{\alpha} + b_i + v_{i0}^{-1}(y_i - \pi_{i0})$. Under the null hypothesis H_0: $\beta = 0$, we iteratively fit the working linear mixed model $\tilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{b} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \mathbf{W}^{-1})$. We use the computationally efficient AI-REML algorithm[6,23] to estimate $\hat{\tau}_k$.

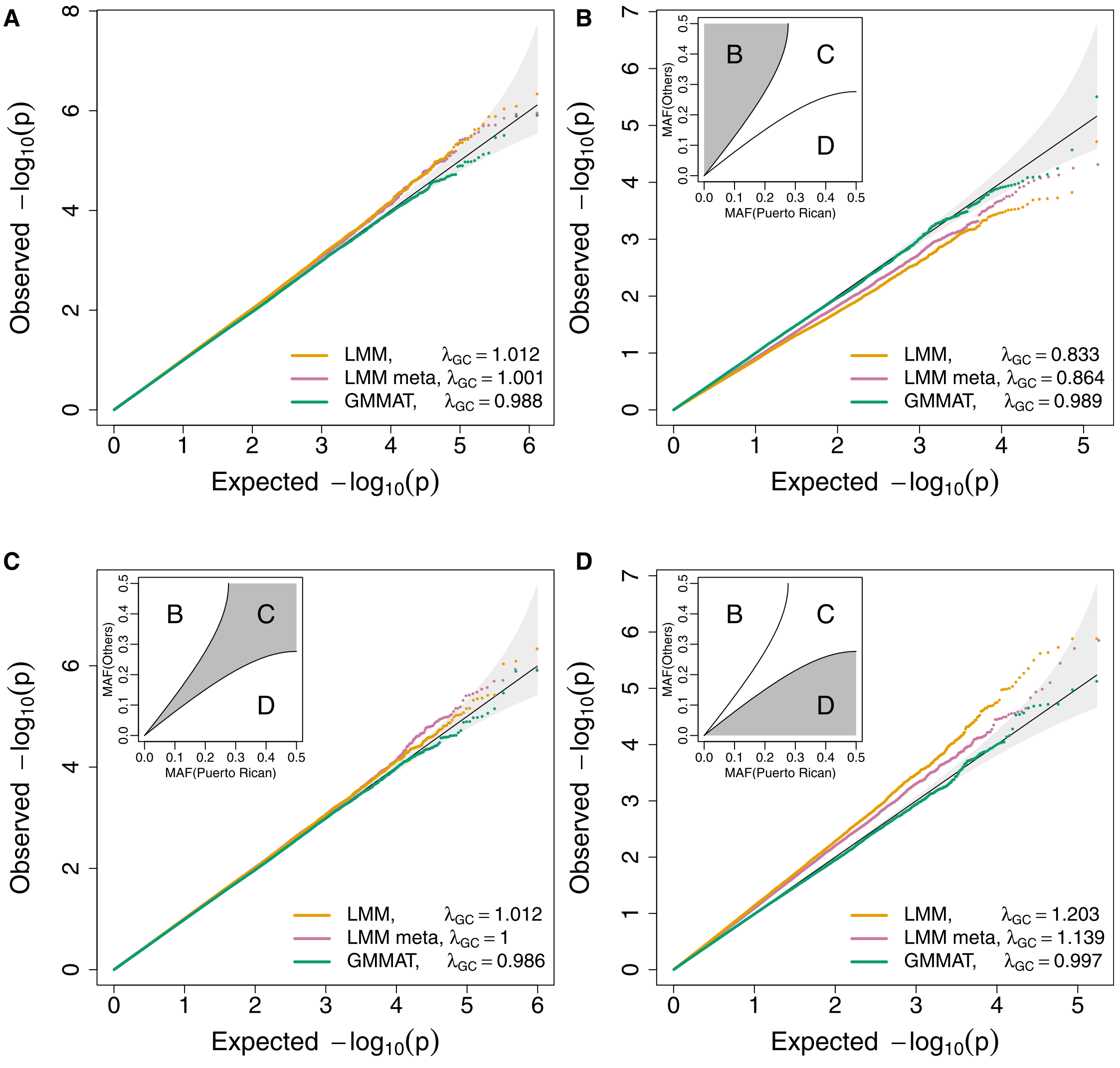

Figure 1. Quantile-Quantile Plot of Association Test p Values from the Asthma GWAS Analysis in HCHS/SOL
(A) All SNPs.
(B) Category 1: SNPs with the ratio of expected variances in Puerto Ricans over non-Puerto Ricans less than 0.8.
(C) Category 2: SNPs with the ratio of expected variances in Puerto Ricans over non-Puerto Ricans between 0.8 and 1.25.
(D) Category 3: SNPs with the ratio of expected variances in Puerto Ricans over non-Puerto Ricans greater than 1.25.
Abbreviations are as follows: LMM, a joint analysis using LMM on the combined samples; LMM meta, an inverse-variance weighted fixed effects meta-analysis approach to combine LMM results from analyzing Puerto Ricans and non-Puerto Ricans separately.

After obtaining the variance component estimates, the estimates of the fixed effects $\widehat{\boldsymbol{\alpha}}$ and random effects $\widehat{\mathbf{b}}$ can be updated, followed by the working vector $\widetilde{\mathbf{Y}}$. The process continues until convergence.

At each iteration, we perform a matrix inversion based on Cholesky decomposition with complexity of $O(n^3)$ and matrix multiplications with complexity of $O(pn^2 + K^2n^2 + p^2n)$, where n is the sample size, p is the number of covariates, and K is the number of variance components. Thus, the total complexity of fitting a logistic mixed model with K known relatedness matrices is $O(in^3 + ipn^2 + iK^2n^2 + ip^2n)$, where i is the number of iterations required to reach convergence.

The score for H_0: $\beta = 0$ is $T = \mathbf{G}^T(\mathbf{y} - \widehat{\boldsymbol{\pi}}_0)$, where $\mathbf{G} = (G_1 \quad G_2 \quad \cdots \quad G_n)^T$ is the n × 1 column vector of genotypes, $\mathbf{y} = (y_1 \quad y_2 \quad \cdots \quad y_n)^T$ is the n × 1 column vector of outcomes, and $\widehat{\boldsymbol{\pi}}_0$ is a vector of fitted values under H_0, which is the same for all SNPs. The estimated variance of the score is $Var(T \mid H_0) = \mathbf{G}^T\widehat{\mathbf{P}}\mathbf{G}$ under the null hypothesis, where $\widehat{\mathbf{P}} = \widehat{\boldsymbol{\Sigma}}^{-1} - \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}(\mathbf{X}^T\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\boldsymbol{\Sigma}}^{-1}$, and $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{W}}^{-1} + \sum_{k=1}^{K}\widehat{\tau}_k\mathbf{V}_k$. The test for each genetic variant involves a vector-matrix-vector multiplication and an inner product calculation for two vectors, and thus the score test step has complexity of $O(qn^2 + qn)$, where q is the total number of genetic variants tested. Also, the score test step can be easily parallelized if necessary. We use the C++ library Armadillo to perform matrix and vector calculations. More details about fitting the logistic mixed model and performing the score test are given in Appendix A.

HCHS/SOL Genotyping and Statistical Analysis

HCHS/SOL subjects who consented to genetic studies had DNA extracted from blood. These DNA samples were genotyped on the SOL HCHS Custom 15041502 B3 array (custom content designed and developed by Papanicolaou, Rotter, and Taylor) at

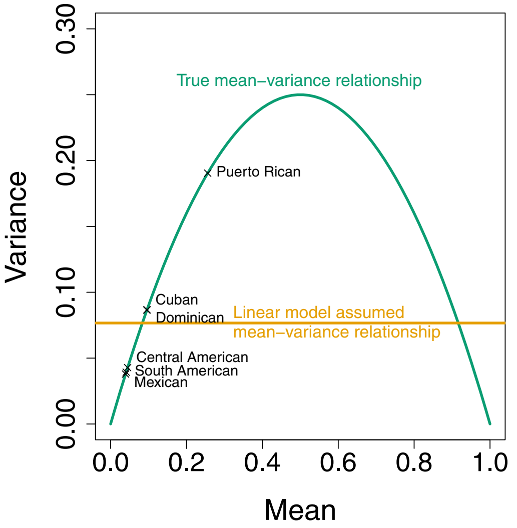

Figure 2. True Mean-Variance Relationship for a Binary Trait and the Constant Mean-Variance Relationship Assumed by Linear Models, Illustrated by the Example from the Asthma Data in HCHS/SOL

For a binary trait with the mean π, its variance is $\pi(1-\pi)$, which varies with the mean. This heteroscedasticity is properly accounted for by logistic regression. Linear models inappropriately assume that the variance of the binary trait does not change with the mean and is a constant (homoscedasticity). For example, the variance of the binary trait (asthma status) in Puerto Ricans is considerably larger than the variances in the other five populations, because Puerto Ricans have a much higher asthma disease proportion than the other populations. This heteroscedasticity caused by population stratification results in the p values calculated from LMMs being likely to be incorrect, but is properly taken into account by logistic mixed models using GMMAT.

Illumina Microarray Services. This array comprised the Illumina Omni 2.5M array (HumanOmni2.5-8v1-1) and additional custom content selected for HCHS/SOL, including ancestry-informative markers, variants characteristic of Amerindians, and known GWAS hits.[25] Illumina Microarray Services, LA Biomed, and the SOL Genetic Analysis Center (GAC) performed quality control to generate recommended SNP- and sample-level quality filters. Samples were checked for annotated versus genetic sex, gross chromosomal anomalies, relatedness and population structure, missing call rates, batch effects, duplicate sample discordance, and Mendelian errors. At the SNP level, checks were performed for Hardy-Weinberg equilibrium, minor allele frequency (MAF), duplicate probe discordance, and missing call rate.

Study participants were recruited via a multi-stage survey sampling procedure, for which individuals were sampled within households that were sampled within block groups.[26] The study includes genetic data from 12,803 individuals of Hispanic/Latino origin, belonging to six different Hispanic/Latino background groups.[25] The HCHS/SOL study was approved by institutional review boards at participating institutions, and written informed consent was obtained from all participants. Standard quality control, similar to that described by Laurie et al.,[27] was performed to filter SNPs and samples of poor quality. Additionally, samples missing information on asthma diagnosis and 56 samples identified as ancestry outliers from principal-component analysis were removed, leaving a sample size of 11,222 for analysis. We also filtered SNPs with a MAF less than 5% in this overall sample, resulting in a set of 1,299,221 autosomal SNPs to be analyzed. The proportions of asthma cases in analyzed participants within each of the genetic background groups of Central American, Cuban, Dominican, Mexican, Puerto Rican, and South American descent are 4.5%, 9.6%, 9.6%, 3.9%, 25.6%, and 4.1%, respectively (Table 1). The Puerto Rican group is a clear outlier from the rest, so we compared it to the collection of all other background groups, which has a combined proportion of 5.8%. To account for potential shared environment effects resulting from the sampling design, relatedness matrices representing household and block group membership were included in addition to the standardized genetic relationship matrix, totaling three random effects for the linear and logistic mixed models. The top five PCs, as well as Hispanic/Latino background group indicators, were used to adjust for ancestry in both models. Additional fixed effects covariates included field center, sex, age, cigarette use, cigarette pack years, and sampling weight (i.e., weights to account for disproportionate selection of the sample according to the sample survey design[26]). Treatment of sampling weight as a fixed effect in this way can effectively adjust for the marginal effect of design variables.[28]

Simulation Studies

We perform coalescent simulations[29] to generate genotypes for a total of 8,000 founders with 1,000,000 independent SNPs from a 20 × 20 grid (Figure S1) to mimic spatially continuous populations (20 individuals per cell) with migration rate between adjacent cells $M = 10$ to represent population structure within Europe.[30,31] To simulate genotypes of an offspring cohort, we first sampled (without replacement) 10 pairs of parents for each cell of the 20 × 20 grid from the original cohort and then simulated two children for each family using the gene dropping algorithm,[32] thus obtaining 8,000 individuals in the offspring cohort. We combined the two cohorts to get 16,000 individuals with both population structure and relatedness. For individual j in family i, the probability of being a case π_{ij} was calculated from

$$\mathrm{logit}(\pi_{\mathrm{ij}}) = \alpha_0 + \alpha_1 Z_i + a_{ij},$$

where $Z_i = 1$ if family i was from a 10 × 10 grid in the top left, and $Z_i = 0$ otherwise. The parameters α_0 and α_1 were chosen such that the disease prevalence was 0.28 in the high-risk population group in the top left given no random effects and 0.05 in the low-risk population group given no random effects. The random effects were simulated as

$$\mathbf{a}_i = \begin{pmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \\ a_{i4} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix} \right),$$

where the variance component parameter σ^2 was set to 2. We randomly sampled 10,000 individuals from the combined original and offspring cohorts to form a simulated cohort study, and their standardized genetic relationship matrix was calculated using 625,583 genetic variants with MAF greater than 5% in the founder population. We removed family indicators and compared linear and logistic mixed models using the genetic relationship matrix. We included the top ten ancestry PCs as covariates in both models. We analyzed common genetic variants with MAF greater than 5%. In the simulated case-control study for rare disease, we chose α_0 and α_1 such that the disease prevalence was 0.045 in the high-risk population group in the top left given no random effects and 0.005 in the low-risk population group given no random effects. We randomly sampled 1,667 case subjects and 8,333 control subjects from the combined original and offspring cohorts to form a case-control study with a total sample size of 10,000 and a case-control ratio of 1:5. We performed the same analysis as described above.

Results

Analysis of HCHS/SOL Asthma GWAS Data

We compared the results from LMMs and logistic mixed models using GMMAT for an analysis of physician-diagnosed asthma in the population-based HCHS/SOL cohort study.

Figure 1A shows the behavior of the overall quantile-quantile (QQ) plots from each method in the asthma analysis. The inflated results from the LMM are caused by violation of its constant residual variance (homoscedasticity) assumption. For binary traits, population stratification affects both population-specific means (disease prevalences in cohort studies) and variances of the trait; population groups with disease prevalence closer to 0.5 have larger variances (Table 1 and Figure 2). The mean-variance relationship assumed in linear models and LMMs is misspecified for binary traits. Although confounding by population structure can be accounted for by adjusting for population groups or ancestry PCs, unequal binary trait variances across different populations are not appropriately modeled in linear models and LMMs. The variance of the standard LMM-based test statistic for no genetic association is underestimated when population groups with larger binary trait variances (i.e., higher disease risk) also have higher MAFs, thus larger genotypic variances. This leads to inflation in the test statistic, and vice versa.

To demonstrate this, we partitioned all common SNPs across the genome into three categories based on their genotypic variances in Puerto Ricans and non-Puerto Ricans: (1) the SNPs with the ratio of expected variances, that is $\mathrm{Var(SNP)} = 2MAF(1-MAF)$, in Puerto Ricans (high-risk group) over non-Puerto Ricans (low-risk group) less than 0.8; (2) the SNPs with the ratio of Var(SNP) in Puerto Ricans over non-Puerto Ricans between 0.8 and 1.25; and (3) the SNPs with the ratio of Var(SNP) in Puerto Ricans over non-Puerto Ricans greater than 1.25. In this classification, the category 1 SNPs ($n = 144{,}815$ [11%]) have appreciably lower MAFs in Puerto Ricans (high risk) than non-Puerto Ricans (low risk). The category 2 SNPs have similar MAFs in both groups ($n = 982{,}805$ [76%]). The category 3 SNPs ($n = 171{,}601$ [13%]) have appreciably higher MAFs in Puerto Ricans (high risk) than non-Puerto Ricans (low risk). Both LMMs and GMMAT perform well for category 2 SNPs (Figure 1C). The problem with LMMs is more apparent for category 1 and 3 SNPs: Figure 1B shows conservative p values for category 1 SNPs, and Figure 1D shows anti-conservative p values for category 3 SNPs, as expected. This indicates the p values calculated from LMMs are likely to be incorrect for at least 24% of SNPs in HCHS/SOL asthma GWASs. Meta-analysis by analyzing Puerto Ricans and non-Puerto Ricans separately after removing related individuals across the two groups improves the LMM performance, but the results are still not well calibrated for category 1 and 3 SNPs (Figures 1B and 1D), probably due to the heteroscedasticity issue caused by within-group population stratification. In contrast, GMMAT performs well for all categories of SNPs considered.

Simulation Studies

We also conducted extensive simulation studies under the null hypothesis to illustrate this issue in various GWAS case-control and cohort study designs, to exclude the possibility that LMM-inflated results in HCHS/SOL were caused by polygenic effects. We restricted analyses to common SNPs with a MAF greater than 5%. First, in a simulated cohort study of 10,000 individuals with cryptic relatedness, we simulated two population groups with disease prevalences of 28% (population 1) and 5% (population 2), respectively, from a map of spatially continuous populations. The disease prevalences were chosen to mimic the asthma disease proportions of Puerto Ricans and non-Puerto Ricans in HCHS/SOL (Table 1). In both LMM and GMMAT, we adjusted for the top ten ancestry PCs as fixed effects covariates. Figure 3A includes results from 3,200 null simulated datasets with 625,583 SNPs and 10,000 related subjects per dataset. LMM has a type I error rate of 1.26×10^{-7} at the nominal 5×10^{-8} level, compared to 5.0×10^{-8} for GMMAT. Note that because more than 2 billion p values are included, type I error rate estimates at this level are accurate with standard error 5×10^{-9}.

Following the HCHS/SOL example, we partitioned all the common SNPs into three categories: (1) the SNPs with the ratio of expected variances in population 1 (high risk) over population 2 (low risk) less than 0.8; (2) the SNPs with the ratio of expected variances in population 1 over population 2 between 0.8 and 1.25; and (3) the SNPs with the ratio of expected variances in population 1 over population 2 greater than 1.25. Category 1 includes the SNPs that have appreciably lower MAFs in population 1 than population 2 (24%). Category 2 includes the SNPs that have comparable MAFs in the two populations (58%). Category 3 includes the SNPs that have appreciably higher MAFs in population 1 than population 2 (18%). For category 2 SNPs, both LMM and GMMAT properly control for type I error rates (Figure 3C). However, despite adjusting for ancestry PCs, LMMs have deflated type I error rates for category 1 SNPs (Figure 3B) and inflated type I error rates for category 3 SNPs (Figure 3D). LMMs often fail to control type I error rates in the presence of moderate or strong population stratification when differences in prevalence by population groups cause large differences in binary trait variances between populations, as demonstrated by our simulation studies that were motivated by the HCHS/SOL study. However, this issue is not always evident in the overall QQ plot using all the SNPs in a GWAS, because the number of SNPs in a

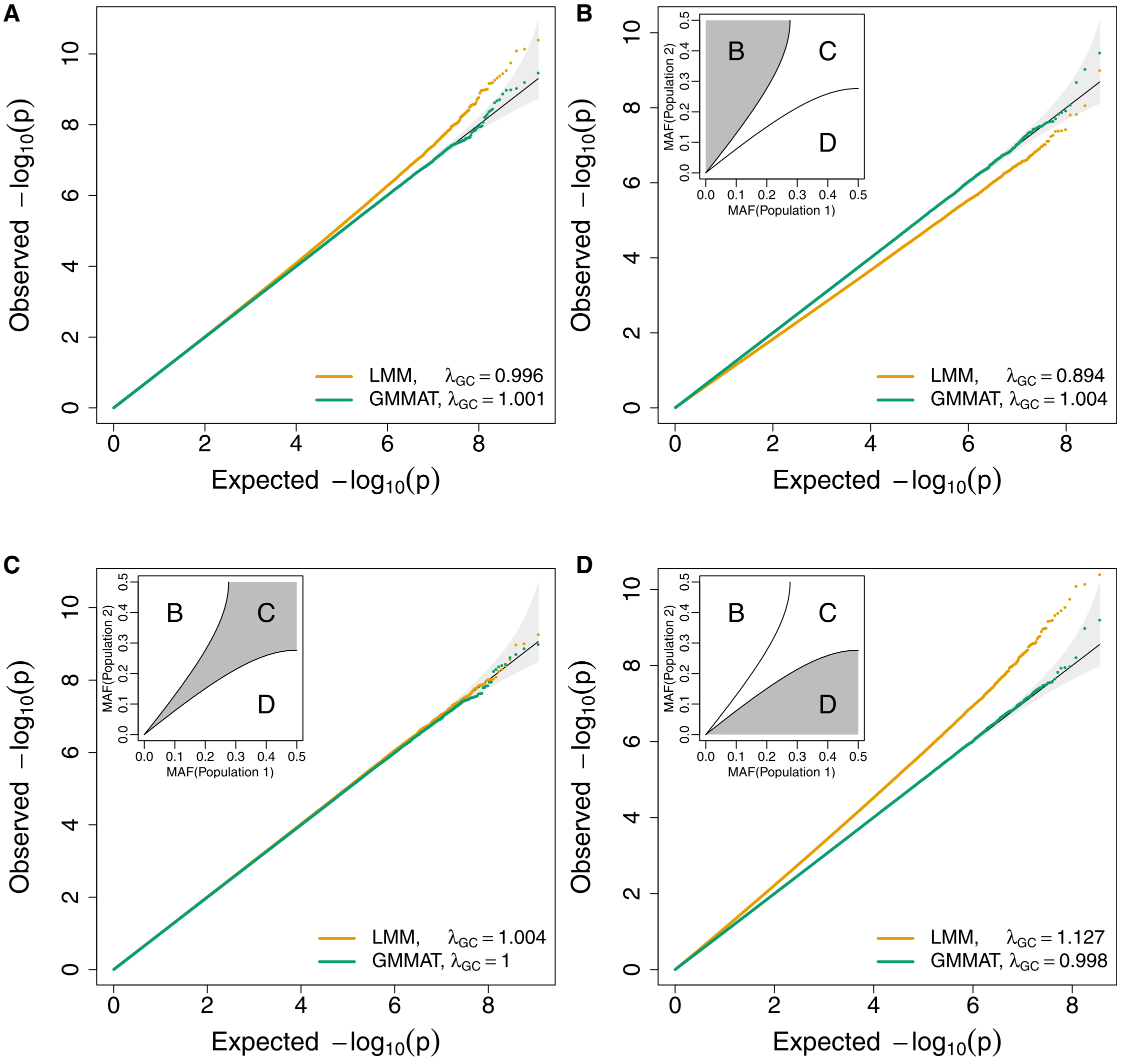

Figure 3. A Simulated Cohort Study with 10,000 Related Individuals

Quantile-quantile plots of association test p values from 3,200 simulation replicates under the null hypothesis of no genetic association, each with 625,583 common SNPs, were combined to get more than 2 billion null p values.

(A) All SNPs.

(B) Category 1: SNPs with the ratio of expected variances in population 1 (high risk) over population 2 (low risk) less than 0.8.

(C) Category 2: SNPs with the ratio of expected variances in population 1 (high risk) over population 2 (low risk) between 0.8 and 1.25.

(D) Category 3: SNPs with the ratio of expected variances in population 1 (high risk) over population 2 (low risk) greater than 1.25.

typical GWAS is usually in the range of 10^5 and 10^6, confidence intervals are relatively wide, and inflation and deflation are likely to mask each other in the overall QQ plot.

To illustrate this, Figure S2 shows results from only one simulated dataset with 625,583 common SNPs and 10,000 subjects in the same setting as Figure 3. There is a slight indication of inflation for p values from LMMs in Figure S2A, but the results are conservative for category 1 SNPs (n = 151,206 [24%]) (Figure S2B) and anti-conservative for category 3 SNPs (Figure S2D) (n = 111,455 [18%]). These results show that the p values from LMMs might be incorrect for at least 42% of SNPs in this analysis, even though inflation and deflation might not be obvious in the overall QQ plot.

We next simulated a case-control study with a total sample size of 10,000 and a case-control ratio of 1:5 from the same two populations as in our simulated cohort study in the presence of cryptic relatedness. Compared to cohort studies, case-control studies oversample cases. The disease prevalence was 4.5% and 0.5% in population 1 (high risk) and population 2 (low risk), respectively. Figure S3A shows that GMMAT works well according to the QQ plot of p values. The genomic control inflation factors show that both LMM and GMMAT have appropriate median p values. However, type I error rates at the nominal level of 5×10^{-8} are 1.75×10^{-7} for LMM and 4.6×10^{-8} for GMMAT, respectively. Moreover, when we divided the SNPs into three categories as we previously described, we observed strong deflation for category 1 SNPs (Figure S3B) and strong inflation for category 3 SNPs (Figure S3D) from LMM analysis, as seen in the simulated cohort study. Additional simulation studies show that the same problem for

LMMs also exists in cohort studies without cryptic relatedness in the presence of population stratification (Figure S4). Although confounding by discrete populations can be appropriately accounted for by analyzing each population separately via LMMs followed by meta-analysis (Figure S5), this strategy does not perform well in the presence of confounding by continuous population structure (Figure S6). Moreover, LMMs do not work well in case-control studies of unrelated individuals with moderate to strong population stratification (Figure S7). These suggest that despite its wide use in both population-based and family-based genetic association studies, LMMs are generally not appropriate for binary traits due to misspecified phenotype variance in the model, probably yielding incorrect p values (see Appendix B for details).

We also performed additional simulations to compare GMMAT with ROADTRIPS, which performs a retrospective test for association in case-control data,[33] and the recently developed liability estimator as a phenotype (LEAP) approach. LEAP fits a liability mixed model that accounts for case-control ascertainment.[20] In the same case-control setting as that described for Figure S7, ROADTRIPS was not well calibrated due to failure to fully account for population stratification. LEAP was found to have well-behaved overall QQ plots but inflated type I error rates for category 1 SNPs, regardless of whether top ten ancestry PCs were adjusted as covariates or not (Figure S8).

We next simulated case-control studies with the disease prevalence of 1% by varying the case-control sampling ratios in two population groups. We first considered balanced cases and controls in the two populations. Specifically, when the case-control ratio is 1:1 in both groups, there is no population stratification, and ROADTRIPS, LEAP, LMMs, and GMMAT all properly control for type I error rates (Figure S9). When the case-control ratios are different but flipped in two populations (4:1 in population 1 and 1:4 in population 2), there is population stratification. However, because the variance of the binary trait is the same in both population groups in this situation, LMMs and GMMAT both perform well, whereas ROADTRIPS has inflated type I error rates in the tail and LEAP has an appropriate overall QQ plots but inflated or deflated type I error rates for the SNPs whose MAFs are different in the two population groups (Figure S10).

We also considered a situation where the case-control ratios are different in two populations (25:2 in population 1 and 25:48 in population 2) in a way that led to a smaller variance of the binary trait in population 1 than in population 2. ROADTRIPS and LEAP adjusting for the top ten PCs show inflation in the overall QQ plot, and LEAP without covariates and LMMs have inflated type I error rates for category 1 SNPs and deflated type I error rates for category 3 SNPs (Figure S11). In terms of required computational resources, LEAP requires more than 20 times the amount of memory compared to GMMAT for an analysis of a case-control study of sample size 10,000.

We also performed simulation studies to evaluate the performance of GMMAT in estimating odds ratios of genetic variants. We found that GMMAT had minor bias in estimating odds ratios when fitting logistic mixed models under the null and alternative hypotheses, compared to the true parameter values in large samples ($n = 10,000$). In addition, the performance in terms of absolute bias appears to be similar to that using MACAU,[34] a recently developed Markov Chain Monte Carlo-based mixed model approach for binomial count data, while the GMMAT odds ratio estimates are less variable (Figure S12).

In the absence of population stratification, LMMs and GMMAT have comparable power, but they are both less powerful than logistic regression (Table S1). In the presence of population stratification, LMMs have less power than GMMAT for causal genetic variants with a lower MAF in the high-risk population than in the low-risk population (Table S1), due to its conservative type I error rate control for such variants, as shown in Figure S4B.

Computational Speed and Memory Usage

We benchmarked our GMMAT package against SAS PROC GLIMMIX regarding computational speed and memory usage. To fit a null model with sample size 2,000, GMMAT takes less than 1.5% of the time required by SAS PROC GLIMMIX when fitting a logistic mixed model with one variance component, and less than 0.6% of the time with three variance components (Table S2), yielding the same numerical results to at least the fourth decimal place. On average, with one variance component of random effects, as is commonly used to account for genetic relatedness in GWASs, SAS PROC GLIMMIX requires about 28 min to fit a null model on a single core of an Intel Xeon E5-2690 CPU (2.90 GHz), compared to about 22 s using GMMAT. With three variance components of random effects that account for complex sampling designs in addition to genetic relatedness, SAS PROC GLIMMIX requires about 1.2 hr, whereas GMMAT needs still about 22 s. It takes about 14 min for GMMAT to perform score tests for 1,000,000 genetic variants on the same core, without parallelization. In practice, score tests for different variants can be easily run in parallel in a computing cluster, and testing each genetic variant takes only about 0.8 ms.

GMMAT requires less than 1 GB memory in analyzing 2,000 individuals. With sample size 10,000, it takes about 18 min to fit the null model with one variance component and 34 min with three variance components using GMMAT, and about 3.6 hr to perform score tests for 1,000,000 genetic variants on a single core (about 13 ms for testing each genetic variant) (Table S2). SAS PROC GLIMMIX reports insufficient memory when 100 GB memory is specified to fit the null model for 10,000 individuals, whereas GMMAT requires less than 14 GB memory for one variance component and 21 GB for three variance components.

Discussion

We have proposed in this paper using logistic mixed models to correct for population stratification and relatedness in analyzing binary traits in GWASs. The proposed GMMAT performs computationally efficient score tests for genetic associations in cohort and case-control GWASs with binary traits. We demonstrate that GMMAT is effective in controlling type I error rates. In contrast, even when ancestry terms are included as covariates, applying LMMs to binary traits can lead to incorrect type I error rates in the presence of population stratification, particularly when population groups have heterogeneous disease risks or case-control ratios that result in different binary trait variances. In such scenarios, LMMs are approximately valid only when the MAF of the genetic variant being tested for association is roughly the same in all population groups, i.e., in the presence of no or weak confounding by population structure. The homoscedasticity assumption underlying standard LMMs is essential but has largely been ignored in previous genetic association studies that used LMMs for binary traits. Despite the widespread belief that LMMs can be used to account for population stratification for both continuous and binary traits, our results show that LMMs can lead to incorrect type I error rates and p values in the presence of population stratification and relatedness. Recently, Conomos et al.[25] showed that violation of the homoscedasticity assumption could also happen for continuous traits, and a model allowing for group-specific residual variances outperformed standard LMMs in such scenarios. For binary traits in discrete population groups, when there is no or weak confounding by population structure within each group, we have shown that meta-analysis is a valid and effective approach for combining results from analyzing each group separately using LMMs (Figure S5). Moreover, we can use logistic regression to analyze homogeneous unrelated samples, which is more powerful than LMMs and GMMAT (Table S1). In reality, however, population groups with no or weak within-group confounding might not always be clearly defined in large-scale genetic association studies, especially for admixed populations. When within-group population stratification remains, we have also shown that a meta-analysis approach using LMMs is still mis-calibrated similarly to the standard LMM approach (Figure S6). Furthermore, it is often difficult to check the validity of LMMs by comparing group-specific binary trait variances and MAFs. In contrast, such checking is not required by GMMAT when fitting logistic mixed models.

In practice, QQ plots have been widely used for model diagnostics in GWASs. Our results show that a well-behaved QQ plot is not sufficient to identify invalid results due to model misspecification. Specifically, an overall QQ plot can appear to show LMMs properly controlling type I error rates for binary traits (Figure S2A), but this might just reflect balancing out of conservative p values and anti-conservative p values of different subsets of SNPs (categories 1 and 3) (Figures S2B and S2D). It is important to note that these errors (incorrect p values) do not “cancel out;” spuriously significant results, i.e., type I errors, are not “corrected” by omission of true signals, i.e., type II errors. More generally, looking just at the overall QQ plot of all SNPs can lead to unwarranted beliefs that analyses (e.g., use of standard LMMs for binary traits) are appropriate, when they can have serious deficiencies for large proportions of the results.

In case-control studies, which oversample cases, LMMs are subject to incorrect type I error rates due to unequal variances of binary traits caused by population stratification introduced by unequal case-control ratios from different sampling schemes across population groups, even if the disease prevalence is the same in all the subpopulations and the overall case-control ratio is 1:1 (Figure S11). Because ROADTRIPS currently does not allow for covariate adjustment, it does not work well in the presence of moderate to strong population stratification.[35] The recently developed liability mixed models[20,21] require accurate estimation of the disease prevalence as well as heritability and liability in the underlying overall population, which can be difficult to obtain in practice in the presence of population stratification and unknown subpopulation groups. These models are generally applicable in the presence of no or weak population stratification, e.g., when the case-control ratios are the same across populations, but can fail to control for type I error rates in the presence of moderate or strong population stratification, e.g., when the case-control ratios differ between populations. They also currently cannot handle multiple random effects in addition to those accounting for genetic relatedness, such as household and block groups in HCHS/SOL. The liability threshold mixed linear model (LTMLM) approach is applicable only to population-based case-control study designs with no confounders and low levels of relatedness, because it cannot adjust for covariates or handle family data.[21] The LEAP method can accommodate covariates and be applicable to family data, but inclusion of covariates presents both technical and statistical challenges.[20] GMMAT provides a flexible method that does not require knowing disease prevalence or heritability and liability estimates, and provides valid p values while properly controlling for type I error rates.

Recently, Song et al.[36] proposed a genotype-conditional association test that accounts for population structure in association tests. However, as pointed out by the authors, their approach does not account for family or cryptic relatedness. In contrast, our logistic mixed model approach is more flexible and can account for both population structure and relatedness in population-based and family-based cohort and case-control studies, as well as complex sampling designs (as illustrated in our HCHS/SOL asthma example). Therefore, our approach can be applied to a much wider range of genetic association studies with family data, cryptic relatedness, unobserved shared environmental effects, and non-random sampling study designs, in addition to population stratification, without the need to model them in different ways.

We provide an open-source R package GMMAT for fitting logistic mixed models and performing score-based tests in GWASs. The package can also be applied to other types of continuous and discrete traits in the general framework of generalized linear models[37] that allow for different link functions and different mean-variance relationships. Furthermore, the score statistics obtained from different studies of the same disease can be easily combined in meta-analysis.[38]

The recently proposed MACAU algorithm[34] implemented Wald tests in mixed models for binomial count data using a Markov Chain Monte Carlo-based approach. However, performing Wald and likelihood ratio tests for a large number of individual variants via logistic mixed models is currently computationally impractical for moderate- to large-sample GWASs and would require future research on developing efficient algorithms. Unlike linear mixed models, logistic mixed models can be directly used to estimate odds ratios by fitting the models under the alternative hypothesis. It is computationally feasible to estimate odds ratios by fitting alternative logistic mixed models for a subset of candidate genetic variants of interest. If computational issues can be resolved in the future, logistic mixed models can also be useful for risk prediction in GWASs.

Appendix A: Derivation of GMMAT

The Generalized Linear Mixed Model

The derivations below are based on generalized linear mixed models (GLMMs), and logistic mixed models are a special case of GLMMs when the link function is logit and the dispersion parameter is fixed at 1. In the context of single-variant test, we consider the following GLMM

$$\eta_i = g(\mu_i) = \mathbf{X}_i\boldsymbol{\alpha} + G_i\beta + b_i,$$

where $\mathbf{X}_i$ is a 1 × p row vector of covariates for subject i, $\boldsymbol{\alpha}$ is a p × 1 column vector of fixed covariate effects including the intercept, G_i is the genotype of the genetic variant of interest for subject i, and β is the fixed genotype effect. We assume that $\mathbf{b} \sim N(0, \sum_{k=1}^{K}\tau_k\mathbf{V}_k)$ is an n × 1 column vector of random effects, τ_k are the variance component parameters, $\boldsymbol{\tau}$ is a K × 1 column vector of τ_k, and $\mathbf{V_k}$ are known n × n matrices. We also assume that given the random effects $\mathbf{b}$, the outcome y_i is conditionally independent with mean $E(y_i \mid \mathbf{b}) = \mu_i$ and variance $Var(y_i \mid \mathbf{b}) = \phi a_i^{-1}v(\mu_i)$, where ϕ is the dispersion parameter (for binary and Poisson data $\phi = 1$), a_i are known weights, and $v(\cdot)$ is the variance function. The linear predictor η_i is a monotonous function of the conditional mean μ_i via the link function $\eta_i = g(\mu_i)$. For binary traits y_i, $\mu_i = \pi_i = P(y_i = 1 \mid \mathbf{X}_i, G_i, b_i)$ is the probability of the binary outcome (e.g., disease status) for subject i.

For subject i, the quasi-likelihood given random effects $\mathbf{b}$ is

$$ql_i(\boldsymbol{\alpha}, \beta; \mathbf{b}) = \int_{y_i}^{\mu_i} \frac{a_i(y_i - \mu)}{\phi v(\mu)} d\mu.$$

The log integrated quasi-likelihood function of $(\boldsymbol{\alpha}, \beta, \phi, \boldsymbol{\tau})$ is

$$\begin{aligned} ql(\boldsymbol{\alpha}, \beta, \phi, \boldsymbol{\tau}) = \log \int & \exp\left\{ \sum_{i=1}^{n} ql_i(\boldsymbol{\alpha}, \beta; \mathbf{b}) \right\} \\ & \times (2\pi)^{-\frac{n}{2}} \left| \sum_{k=1}^{K} \tau_k \mathbf{V}_k \right|^{-\frac{1}{2}} \\ & \times \exp\left\{ -\frac{1}{2}\mathbf{b}^T \left(\sum_{k=1}^{K} \tau_k \mathbf{V}_k \right)^{-1} \mathbf{b} \right\} d\mathbf{b}. \end{aligned}$$

(Equation A1)

Let

$$f(\mathbf{b}) = \sum_{i=1}^{n} ql_i(\boldsymbol{\alpha}, \beta; \mathbf{b}) - \frac{1}{2}\mathbf{b}^T \left(\sum_{k=1}^{K} \tau_k \mathbf{V}_k \right)^{-1} \mathbf{b},$$

we can use Laplace method to approximate the n-dimensional integral

$$\int \exp\{f(\mathbf{b})\} d\mathbf{b} \approx (2\pi)^{\frac{n}{2}} \left| -f''\left(\tilde{\mathbf{b}}\right) \right|^{-\frac{1}{2}} \exp\left\{ f\left(\tilde{\mathbf{b}}\right) \right\},$$

thus Equation A1 becomes

$$\begin{aligned} ql(\boldsymbol{\alpha}, \beta, \phi, \boldsymbol{\tau}) = & -\frac{1}{2}\log\left| \sum_{k=1}^{K} \tau_k \mathbf{V}_k \right| - \frac{1}{2}\log\left| -f''\left(\tilde{\mathbf{b}}\right) \right| \\ & + f\left(\tilde{\mathbf{b}}\right), \end{aligned}$$

(Equation A2)

where

$$\tilde{\mathbf{b}} = \underset{\mathbf{b}}{\operatorname{argmax}} f(\mathbf{b})$$

is the solution of $f'(\mathbf{b}) = 0$.

The first partial derivative of $ql_i(\boldsymbol{\alpha}, \beta; \mathbf{b})$ with respect to $\mathbf{b}$ is

$$\frac{\partial ql_i}{\partial \mathbf{b}} = \frac{\partial ql_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \mathbf{b}} = \frac{a_i(y_i - \mu_i)}{\phi v(\mu_i)} \frac{1}{g'(\mu_i)} \mathbf{Z}_i^T,$$

where $\mathbf{Z}_i$ is a 1 × n vector of indicators such that $b_i = \mathbf{Z}_i\mathbf{b}$, $\mathbf{I}_n = \left(\mathbf{Z}_1^T \quad \mathbf{Z}_2^T \quad \cdots \quad \mathbf{Z}_n^T \right)$, and the second derivative is

$$\begin{aligned} \frac{\partial^2 ql_i}{\partial \mathbf{b} \partial \mathbf{b}^T} = & -\frac{a_i \mathbf{Z}_i^T \mathbf{Z}_i}{\phi v(\mu_i)[g'(\mu_i)]^2} - \frac{a_i(y_i - \mu_i)v'(\mu_i)\mathbf{Z}_i^T \mathbf{Z}_i}{\phi v^2(\mu_i)[g'(\mu_i)]^2} \\ & - \frac{a_i(y_i - \mu_i)g''(\mu_i)\mathbf{Z}_i^T \mathbf{Z}_i}{\phi v(\mu_i)[g'(\mu_i)]^3}. \end{aligned}$$

For canonical link functions, $\nu(\mu_i)g'(\mu_i)=1$, the last two terms become 0. Let

$$\boldsymbol{\Delta} = diag\{g'(\mu_i)\},$$

$$\mathbf{W} = diag\left\{\frac{a_i}{\phi\nu(\mu_i)[g'(\mu_i)]^2}\right\},$$

then Equation A2 becomes

$$\begin{aligned} ql(\boldsymbol{\alpha},\beta,\phi,\boldsymbol{\tau}) &= -\frac{1}{2}\log\left|\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right| -\frac{1}{2}\log\left|\sum_{i=1}^{n}\frac{a_i\mathbf{Z}_i^T\mathbf{Z}_i}{\phi\nu(\mu_i)[g'(\mu_i)]^2} + \left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)^{-1}\right| + \sum_{i=1}^{n}ql_i\left(\boldsymbol{\alpha},\beta;\tilde{\mathbf{b}}\right) - \frac{1}{2}\tilde{\mathbf{b}}^T\left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)^{-1}\tilde{\mathbf{b}} \\ &= -\frac{1}{2}\log\left|\sum_{k=1}^{K}\tau_k\mathbf{V}_k\mathbf{W}+\mathbf{I}\right| + \sum_{i=1}^{n}ql_i\left(\boldsymbol{\alpha},\beta;\tilde{\mathbf{b}}\right) \\ &\quad -\frac{1}{2}\tilde{\mathbf{b}}^T\left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)^{-1}\tilde{\mathbf{b}}. \end{aligned}$$

(Equation A3)

We assume that the weight matrix $\mathbf{W}$ changes slowly with respect to the conditional means (following Breslow and Clayton[22]), that is

$$\frac{\partial\mathbf{W}}{\partial\mu_i}\approx 0,$$

then we take the derivatives of Equation A3:

$$\frac{\partial ql(\boldsymbol{\alpha},\beta,\phi,\boldsymbol{\tau})}{\partial\boldsymbol{\alpha}} = \sum_{i=1}^{n}\frac{a_i(y_i-\mu_i)}{\phi\nu(\mu_i)}\frac{1}{g'(\mu_i)}\mathbf{X}_i^T = \mathbf{X}^T\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}),$$

$$\frac{\partial ql(\boldsymbol{\alpha},\beta,\phi,\boldsymbol{\tau})}{\partial\beta} = \sum_{i=1}^{n}\frac{a_i(y_i-\mu_i)}{\phi\nu(\mu_i)}\frac{1}{g'(\mu_i)}G_i = \mathbf{G}^T\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}),$$

(Equation A4)

$$\begin{aligned} \frac{\partial ql(\boldsymbol{\alpha},\beta,\phi,\boldsymbol{\tau})}{\partial\mathbf{b}} &= \sum_{i=1}^{n}\frac{a_i(y_i-\mu_i)}{\phi\nu(\mu_i)}\frac{1}{g'(\mu_i)}\mathbf{Z}_i^T - \left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)^{-1}\mathbf{b} \\ &= \mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) - \left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)^{-1}\mathbf{b}. \end{aligned}$$

Under the null hypothesis $H_0:\beta=0$, if ϕ and $\boldsymbol{\tau}$ are known, we jointly choose $\widehat{\boldsymbol{\alpha}}(\phi,\boldsymbol{\tau})$ and $\widehat{\mathbf{b}}(\phi,\boldsymbol{\tau})$ to maximize Equation A3, then $\widehat{\mathbf{b}}(\phi,\boldsymbol{\tau}) = \tilde{\mathbf{b}}(\widehat{\boldsymbol{\alpha}}(\phi,\boldsymbol{\tau}),\beta=0)$ because $\tilde{\mathbf{b}}$ maximizes $f(\mathbf{b})$ for given $(\boldsymbol{\alpha},\beta)$. Defining the working vector $\tilde{\mathbf{Y}}$ with elements $\tilde{Y}_i = \eta_i + g'(\mu_i)(y_i-\mu_i)$, the solution of

$$\begin{cases} \mathbf{X}^T\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = 0 \\ \mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = \left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)^{-1}\mathbf{b} \end{cases}$$

can be written as the solution to the system

$$\begin{bmatrix} \mathbf{X}^T\mathbf{W}\mathbf{X} & \mathbf{X}^T\mathbf{W} \\ \mathbf{W}\mathbf{X} & \left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)^{-1}+\mathbf{W} \end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^T\mathbf{W}\tilde{\mathbf{Y}}\\ \mathbf{W}\tilde{\mathbf{Y}}\end{bmatrix}.$$

Let $\boldsymbol{\Sigma} = \mathbf{W}^{-1} + \sum_{k=1}^{K}\tau_k\mathbf{V}_k$, $\mathbf{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}$, then

$$\begin{cases} \widehat{\boldsymbol{\alpha}} = \left(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{Y}} \\ \widehat{\mathbf{b}} = \left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)\boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{Y}}-\mathbf{X}\widehat{\boldsymbol{\alpha}}\right) \end{cases}$$

is the solution that maximizes Equation A3. We note that

$$\begin{aligned} \tilde{\mathbf{Y}}-\widehat{\boldsymbol{\eta}} = \tilde{\mathbf{Y}}-\mathbf{X}\widehat{\boldsymbol{\alpha}}-\widehat{\mathbf{b}} &= \left\{\mathbf{I}-\left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)\boldsymbol{\Sigma}^{-1}\right\}\left(\tilde{\mathbf{Y}}-\mathbf{X}\widehat{\boldsymbol{\alpha}}\right) \\ &= \mathbf{W}^{-1}\boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{Y}}-\mathbf{X}\widehat{\boldsymbol{\alpha}}\right) = \mathbf{W}^{-1}\mathbf{P}\tilde{\mathbf{Y}}. \end{aligned}$$

Estimation of Variance Component Parameters

Following Breslow and Clayton,[22] we ignore the dependence of $\mathbf{W}$ on $\boldsymbol{\tau}$ and use Pearson chi-square statistic to approximate the deviance

$$\begin{aligned} -2\phi\sum_{i=1}^{n}ql_i(\boldsymbol{\alpha},\beta;\mathbf{b}) &= \sum_{i=1}^{n}-2\int_{y_i}^{\mu_i}\frac{a_i(y_i-\mu)}{\nu(\mu)}d\mu \\ &\approx \sum_{i=1}^{n}\frac{a_i(y_i-\mu_i)^2}{\nu(\mu_i)}. \end{aligned}$$

Then Equation A3 at the maximum becomes

$$\begin{aligned} ql(\widehat{\boldsymbol{\alpha}}(\phi,\boldsymbol{\tau}),\beta=0,\phi,\boldsymbol{\tau}) &\approx -\frac{1}{2}\log\left|\sum_{k=1}^{K}\tau_k\mathbf{V}_k\mathbf{W}+\mathbf{I}\right| \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\frac{a_i(y_i-\widehat{\mu}_i)^2}{\phi\nu(\widehat{\mu}_i)} - \frac{1}{2}\widehat{\mathbf{b}}^T\left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)^{-1}\widehat{\mathbf{b}} \\ &= -\frac{1}{2}\log\left|\sum_{k=1}^{K}\tau_k\mathbf{V}_k\mathbf{W}+\mathbf{I}\right| - \frac{1}{2}(\mathbf{y}-\widehat{\boldsymbol{\mu}})^T\boldsymbol{\Delta}\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\widehat{\boldsymbol{\mu}}) \\ &\quad -\frac{1}{2}\left(\tilde{\mathbf{Y}}-\mathbf{X}\widehat{\boldsymbol{\alpha}}\right)^T\boldsymbol{\Sigma}^{-1}\left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)\boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{Y}}-\mathbf{X}\widehat{\boldsymbol{\alpha}}\right) \\ &= -\frac{1}{2}\log|\boldsymbol{\Sigma}\mathbf{W}| - \frac{1}{2}\left(\tilde{\mathbf{Y}}-\widehat{\boldsymbol{\eta}}\right)^T\mathbf{W}\left(\tilde{\mathbf{Y}}-\widehat{\boldsymbol{\eta}}\right) \\ &\quad -\frac{1}{2}\left(\tilde{\mathbf{Y}}-\mathbf{X}\widehat{\boldsymbol{\alpha}}\right)^T\boldsymbol{\Sigma}^{-1}\left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)\boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{Y}}-\mathbf{X}\widehat{\boldsymbol{\alpha}}\right) \\ &= -\frac{1}{2}\log|\mathbf{W}| - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\tilde{\mathbf{Y}}^T\mathbf{P}\mathbf{W}^{-1}\mathbf{P}\tilde{\mathbf{Y}} \\ &\quad -\frac{1}{2}\tilde{\mathbf{Y}}^T\mathbf{P}\left(\sum_{k=1}^{K}\tau_k\mathbf{V}_k\right)\mathbf{P}\tilde{\mathbf{Y}} \\ &= c - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\tilde{\mathbf{Y}}^T\mathbf{P}\tilde{\mathbf{Y}}. \end{aligned}$$

Similarly, the restricted maximum likelihood (REML) version is

$$ql_R(\widehat{\boldsymbol{\alpha}}(\phi,\boldsymbol{\tau}),\beta=0,\phi,\boldsymbol{\tau}) = c_R - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log|\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}| - \frac{1}{2}\tilde{\mathbf{Y}}^T\mathbf{P}\tilde{\mathbf{Y}}.$$

Let $\mathbf{V}_0 = diag\{a_i^{-1}v(\mu_i)[g'(\mu_i)]^2\} = \phi^{-1}\mathbf{W}^{-1}$, then $\boldsymbol{\Sigma} = \phi\mathbf{V}_0 + \sum_{k=1}^{K}\tau_k\mathbf{V}_k$, and the first derivatives of $ql_R(\widehat{\boldsymbol{\alpha}}(\phi,\boldsymbol{\tau})$, $\beta=0,\phi,\boldsymbol{\tau})$ with respect to ϕ and $\boldsymbol{\tau}$ are

$$\frac{\partial ql_R(\widehat{\boldsymbol{\alpha}}(\phi,\boldsymbol{\tau}),\beta=0,\phi,\boldsymbol{\tau})}{\partial\phi} = \frac{1}{2}\left\{\tilde{\mathbf{Y}}\mathbf{P}\mathbf{V}_0\mathbf{P}\tilde{\mathbf{Y}} - tr(\mathbf{P}\mathbf{V}_0)\right\},$$

$$\frac{\partial ql_R(\widehat{\boldsymbol{\alpha}}(\phi,\boldsymbol{\tau}),\beta=0,\phi,\boldsymbol{\tau})}{\partial\tau_k} = \frac{1}{2}\left\{\tilde{\mathbf{Y}}\mathbf{P}\mathbf{V}_k\mathbf{P}\tilde{\mathbf{Y}} - tr(\mathbf{P}\mathbf{V}_k)\right\}.$$

We define the average information[6,23] matrix **AI** with the following entries

$$\mathbf{AI}_{\phi\phi} = \frac{1}{2}\tilde{\mathbf{Y}}\mathbf{P}\mathbf{V}_0\mathbf{P}\mathbf{V}_0\mathbf{P}\tilde{\mathbf{Y}},$$

$$\mathbf{AI}_{\phi\tau_k} = \frac{1}{2}\tilde{\mathbf{Y}}\mathbf{P}\mathbf{V}_0\mathbf{P}\mathbf{V}_k\mathbf{P}\tilde{\mathbf{Y}},$$

$$\mathbf{AI}_{\tau_k\tau_l} = \frac{1}{2}\tilde{\mathbf{Y}}\mathbf{P}\mathbf{V}_k\mathbf{P}\mathbf{V}_l\mathbf{P}\tilde{\mathbf{Y}}.$$

Let $\boldsymbol{\theta}$ be the variance component parameters to estimate, when $\phi\neq 1$, $\boldsymbol{\theta} = (\phi,\boldsymbol{\tau})$, and **AI** is a $(K+1)\times(K+1)$ matrix. For binary and Poisson data, $\phi = 1$, $\boldsymbol{\theta} = \boldsymbol{\tau}$, and **AI** is a $K \times K$ matrix containing only $\mathbf{AI}_{\tau_k\tau_l}$.

We use the following algorithm to fit the null GLMM:

1. Fit a generalized linear model with $\boldsymbol{\tau} = 0$ and get $\widehat{\boldsymbol{\alpha}}^{(0)}$ and working vector $\tilde{\mathbf{Y}}^{(0)}$;
2. Use $\boldsymbol{\theta}^{(0)} = Var(\tilde{\mathbf{Y}}^{(0)})/K$ (if $\phi = 1$) or $\boldsymbol{\theta}^{(0)} = Var(\tilde{\mathbf{Y}}^{(0)})/(K+1)$ (if $\phi\neq 1$) as the initial value of $\boldsymbol{\theta}$;
3. For each $k = 0,1,\ldots,K$, update $\boldsymbol{\theta}$ using $\theta_k^{(1)} = \theta_k^{(0)} + 2n^{-1}\{\theta_k^{(0)}\}^2(\partial ql_R(\boldsymbol{\theta}^{(0)})/\partial\theta_k)$;
4. Use $\tilde{\mathbf{Y}}^{(1)} = \tilde{\mathbf{Y}}^{(0)}$ as $\tilde{\mathbf{Y}}$ and update $\boldsymbol{\theta}^{(2)} = \boldsymbol{\theta}^{(1)} + \{\mathbf{AI}^{(1)}\}^{-1}(\partial ql_R(\boldsymbol{\theta}^{(1)})/\partial\boldsymbol{\theta})$;
5. Calculate $\widehat{\boldsymbol{\alpha}}^{(2)}$ and $\widehat{\mathbf{b}}^{(2)}$ using $\tilde{\mathbf{Y}}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$;
6. Update $\tilde{\mathbf{Y}}^{(2)}$ using $\widehat{\boldsymbol{\alpha}}^{(2)}$ and $\widehat{\mathbf{b}}^{(2)}$;
7. Repeat steps 4–6, until $2\max\{|\widehat{\boldsymbol{\alpha}}^{(i)} - \widehat{\boldsymbol{\alpha}}^{(i-1)}|/(|\widehat{\boldsymbol{\alpha}}^{(i)}| + |\widehat{\boldsymbol{\alpha}}^{(i-1)}|), |\widehat{\boldsymbol{\theta}}^{(i)} - \widehat{\boldsymbol{\theta}}^{(i-1)}|/(|\widehat{\boldsymbol{\theta}}^{(i)}| + |\widehat{\boldsymbol{\theta}}^{(i-1)}|)\} \leq$ tolerance.

The Score Test

Once $(\widehat{\boldsymbol{\alpha}},\widehat{\phi},\widehat{\boldsymbol{\tau}})$ is estimated under the null hypothesis $H_0: \beta = 0$, the score test can be constructed by evaluating Equation A4 at $(\widehat{\boldsymbol{\alpha}},\beta=0,\widehat{\phi},\widehat{\boldsymbol{\tau}})$, that is

$$T = \frac{\partial ql(\widehat{\boldsymbol{\alpha}},\beta=0,\widehat{\phi},\widehat{\boldsymbol{\tau}})}{\partial\beta} = \mathbf{G}^T\widehat{\mathbf{W}}\widehat{\Delta}(\mathbf{y}-\widehat{\boldsymbol{\mu}}) = \mathbf{G}^T\widehat{\mathbf{W}}(\tilde{\mathbf{Y}}-\widehat{\boldsymbol{\eta}}) = \mathbf{G}^T\widehat{\mathbf{P}}\tilde{\mathbf{Y}}.$$

Its variance under the null hypothesis is

$$Var(T \mid H_0) = E\left\{\frac{\partial ql(\widehat{\boldsymbol{\alpha}},\beta=0,\widehat{\phi},\widehat{\boldsymbol{\tau}})}{\partial\beta}\frac{\partial ql(\widehat{\boldsymbol{\alpha}},\beta=0,\widehat{\phi},\widehat{\boldsymbol{\tau}})}{\partial\beta^T}\right\} = E\left(\mathbf{G}^T\widehat{\mathbf{P}}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\widehat{\mathbf{P}}\mathbf{G}\right) = \mathbf{G}^T\widehat{\mathbf{P}}\mathbf{G},$$

the last equality holds because $\widehat{\mathbf{P}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{P}} = \widehat{\mathbf{P}}$.

Appendix B. Additional Simulation Studies

Unrelated Individuals with Population Stratification

We performed additional simulation studies to compare LMM and GMMAT in unrelated individuals in the presence of population stratification. We used the coalescent model[29] to simulate genotypes for a total of 16,000 unrelated individuals with 1,000,000 genetic variants from a 20 × 20 grid (Figure S1) of spatially continuous populations (40 individuals per cell) with migration rate between adjacent cells $M = 10$ to mimic population structure within Europe.[30,31] For individual i, the probability of being a case π_i was calculated from

$$\text{logit}(\pi_i) = \alpha_0 + \alpha_1 Z_i,$$

where $Z_i = 1$ if individual i was from a 10 × 10 grid in the top left (population 1) and $Z_i = 0$ otherwise (population 2). The parameters α_0 and α_1 were chosen such that the disease prevalence was 0.28 in population 1 (high risk) and 0.05 in population 2 (low risk). Note that the mean model is not mis-specified when a linear link function is used, thus our simulation setting is not in favor of logistic models. We randomly sampled 10,000 individuals to form a simulated cohort study for common disease (Figure S4) and calculated their standardized genetic relationship matrix using 625,504 genetic variants with MAF greater than 5%. We adjusted for top ten PCs in both models and combined results from 3,200 null simulation replicates. We compared our method with the strategy in which population 1 and population 2 were analyzed separately using LMMs and the results were subsequently combined in a meta-analysis (which we referred to as LMM meta), using 100 null simulation replicates (Figure S5). Moreover, we considered continuous population stratification where Z_i for each cell is the minimum of its row and column coordinates in Figure S1, which ranges from 0 to 19, and α_0 was chosen such that the disease prevalence in populations with $Z_i = 0$ was 0.02 and $\alpha_1 = 0.2$ (Figure S6).

We also simulated a case-control study for rare disease (Figure S7). We chose α_0 and α_1 such that the disease prevalence was 0.045 in population 1 (high risk) and 0.005 in population 2 (low risk). We randomly sampled 1,667 case subjects and 8,333 control subjects to form a case-control study with a total sample size of 10,000 and a case-control ratio of 1:5. We performed the same analysis as described above and combined results from 3,200 null simulation replicates.

Comparison with the Existing Methods

We also compared GMMAT with ROADTRIPS[33] and the recently developed liability (probit) mixed models.[20,21] ROADTRIPS does not allow for covariates. Because LTMLM[21] does not allow covariate adjustment, we used LEAP[20] and set the disease prevalence to 0.015, which is the pooled prevalence in both high- and low-risk population groups. We compared ROADTRIPS and LEAP without covariates as well as LEAP adjusting for top ten PCs as covariates with LMM and GMMAT. The simulation settings were the same as in Figure S7, but p values from only one simulation replicate were shown (Figure S8).

Moreover, we simulated three case-control settings with the same prevalence but different case-control sampling schemes in two population groups in Figure S1. The disease prevalence was set to 0.01 in both groups. We compared ROADTRIPS and LEAP with and without adjusting for top ten PCs, with LMM and GMMAT.

In the first setting, we randomly sampled 1,250 case subjects and 1,250 control subjects from population 1 and 3,750 case subjects and 3,750 control subjects from population 2. This was a balanced case-control study with balanced designs in both population groups. Because the case-control ratio was the same, there was no population stratification (Figure S9).

In the second setting, we randomly sampled 2,000 case subjects and 500 control subjects from population 1 and 1,500 case subjects and 6,000 control subjects from population 2. This was an unbalanced case-control study with unbalanced designs but equal binary trait variances in two population groups. Population stratification was created by different case-control ratios in two population groups (4:1 in population 1 and 1:4 in population 2), instead of different disease prevalence (Figure S10).

In the third setting, we randomly sampled 2,500 case subjects and 200 control subjects from population 1 and 2,500 case subjects and 4,800 control subjects from population 2. This was a balanced case-control study with unbalanced designs and unequal binary trait variances in two population groups. Because the disease prevalence was 0.01 in both groups, there were no high-risk or low-risk groups, but population 1 was the low binary trait variance group with variance 0.0686 and population 2 was the high binary trait variance group with variance 0.2252 (Figure S11).

Simulations with Genetic Effects

We first conducted simulation studies to evaluate the performance of GMMAT for estimating the odds ratios of genetic variants in the presence of population stratification. We used the same genotype data as in our null simulations of unrelated individuals with population stratification. For individual i, the probability of being a case π_i was calculated from

$$\mathrm{logit}(\pi_i) = \alpha_0 + \alpha_1 Z_i + \beta G_i,$$

where $Z_i = 1$ if individual i was from population 1, $Z_i = 0$ otherwise (population 2). G_i is the additively coded genotype for the causal genetic variant. The parameters α_0 and α_1 were chosen such that the disease prevalence was 0.28 in population 1 and 0.05 in population 2 for individuals with $G_i = 0$. β was chosen such that the odds ratio varied from 1.0, 1.1, 1.25, 1.5, to 2.0. The total sample size was 10,000. We compared the odds ratio estimates and the p values calculated by GMMAT and MACAU (Figure S12). We then compared the power of LMM and GMMAT for identifying causal variants that had lower MAFs in population 1 than population 2 (Table S1).

We also simulated a case-control study with 1,667 case subjects and 8,333 control subjects with no population stratification. We assumed the disease prevalence was 0.01 for individuals with $G_i = 0$ and used the same sampling scheme in both population 1 and population 2. We compared the powers of logistic regression, LMM, and GMMAT for identifying causal variants (Table S1).

Supplemental Data

Supplemental Data include 12 figures and 2 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.02.012.

Acknowledgments

This work was supported by grants P01 CA134294, R35 CA197449, and R01 HL113338 (to H.C., Z.L., and X.L.), K99 HL130593 (to H.C.), and R37 CA076404 (to C.W., Z.L., and X.L.). Z.L. was also supported in part by a scholarship from the China Scholarship Council. We thank the participants and staff of the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) for their contributions to this study. The baseline examination of HCHS/SOL was carried out as a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). The following institutes, centers, and offices contributed to the first phase of HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research (NIDCR), National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, and NIH Office of Dietary Supplements. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03). Genotyping efforts were supported by NHLBI (HSN26220/20054C), National Center for Advancing Translational Science Clinical Translational Science Institute (UL1TR000124), and NIDDK Diabetes Research Center (DK063491). This manuscript has been reviewed by the HCHS/SOL Publications Committee for scientific content and

consistency of data interpretation with previous HCHS/SOL publications.

Received: October 29, 2015
Accepted: February 17, 2016
Published: March 24, 2016

Web Resources

The URLs for data presented herein are as follows:

Armadillo, http://arma.sourceforge.net/
GMMAT, http://www.hsph.harvard.edu/xlin/software.html#gmmat
MACAU, http://www.xzlab.org/software.html
MS, http://home.uchicago.edu/rhudson1/source/mksamples.html
ROADTRIPS, http://faculty.washington.edu/tathornt/software/ROADTRIPS2/

References

1. Lander, E.S., and Schork, N.J. (1994). Genetic dissection of complex traits. Science *265*, 2037–2048.
2. Aulchenko, Y.S., de Koning, D.J., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics *177*, 577–585.
3. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics *178*, 1709–1723.
4. 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.
5. 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.
6. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
7. 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.
8. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
9. Pirinen, M., Donnelly, P., and Spencer, C.C.A. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Ann. Appl. Stat. *7*, 369–390.
10. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
11. Barr, R.G., Avilés-Santa, L., Davis, S.M., Aldrich, T., Gonzalez, F., Henderson, A.G., Kaplan, R.C., LaVange, L., Liu, K., Loredo, J.S., et al. (2016). Pulmonary disease and age at immigration among Hispanics: results from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Am. J. Respir. Crit. Care Med. *193*, 386–395.
12. 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.
13. Tucker, G., Price, A.L., and Berger, B. (2014). Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. Genetics *197*, 1045–1049.
14. Fingerlin, T.E., Murphy, E., Zhang, W., Peljto, A.L., Brown, K.K., Steele, M.P., Loyd, J.E., Cosgrove, G.P., Lynch, D., Groshong, S., et al. (2013). Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat. Genet. *45*, 613–620.
15. Cortes, A., Hadler, J., Pointon, J.P., Robinson, P.C., Karaderi, T., Leo, P., Cremin, K., Pryce, K., Harris, J., Lee, S., et al.; International Genetics of Ankylosing Spondylitis Consortium (IGAS); Australo-Anglo-American Spondyloarthritis Consortium (TASC); Groupe Française d'Etude Génétique des Spondylarthrites (GFEGS); Nord-Trøndelag Health Study (HUNT); Spondyloarthritis Research Consortium of Canada (SPARCC); Wellcome Trust Case Control Consortium 2 (WTCCC2) (2013). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. Nat. Genet. *45*, 730–738.
16. Fakiola, M., Strange, A., Cordell, H.J., Miller, E.N., Pirinen, M., Su, Z., Mishra, A., Mehrotra, S., Monteiro, G.R., Band, G., et al.; LeishGEN Consortium; Wellcome Trust Case Control Consortium 2 (2013). Common variants in the HLA-DRB1-HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis. Nat. Genet. *45*, 208–213.
17. Liu, J.Z., Hov, J.R., Folseraas, T., Ellinghaus, E., Rushbrook, S.M., Doncheva, N.T., Andreassen, O.A., Weersma, R.K., Weismüller, T.J., Eksteen, B., et al.; UK-PSCSC Consortium; International PSC Study Group; International IBD Genetics Consortium (2013). Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. Nat. Genet. *45*, 670–675.
18. Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (Berkeley, CA: University of California Press), 221–233.
19. Jarque, C.M., and Bera, A.K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ. Lett. *6*, 255–259.
20. 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.
21. 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.
22. Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. *88*, 9–25.
23. Gilmour, A.R., Thompson, R., and Cullis, B.R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics *51*, 1440–1450.

24. Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Math. Proc. Camb. Philos. Soc. *44*, 50–57.
25. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernández-Rhodes, L., Justice, A.E., Graff, M., et al. (2016). Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos. Am. J. Hum. Genet. *98*, 165–184.
26. Lavange, L.M., Kalsbeek, W.D., Sorlie, P.D., Avilés-Santa, L.M., Kaplan, R.C., Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. Ann. Epidemiol. *20*, 642–649.
27. Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., et al.; GENEVA Investigators (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. Epidemiol. *34*, 591–602.
28. Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it. Surv. Methodol. *37*, 115–136.
29. Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338.
30. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.
31. Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H.M., Stambolian, D., Chew, E.Y., Branham, K.E., Heckenlively, J., Fulton, R., Wilson, R.K., et al.; FUSION Study (2014). Ancestry estimation and control of population stratification for sequence-based association studies. Nat. Genet. *46*, 409–415.
32. MacCluer, J.W., VandeBerg, J.L., Read, B., and Ryder, O.A. (1986). Pedigree analysis by computer simulation. Zoo Biol. *5*, 147–160.
33. 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.
34. Lea, A.J., Tung, J., and Zhou, X. (2015). A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. PLoS Genet. *11*, e1005650.
35. Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. *11*, 459–463.
36. Song, M., Hao, W., and Storey, J.D. (2015). Testing for genetic associations in arbitrarily structured populations. Nat. Genet. *47*, 550–554.
37. McCullagh, P., and Nelder, J.A. (1989). Generalized Linear Models, Second Edition (Chapman & Hall/CRC).
38. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. *95*, 5–23.