OXFORD

Briefings in Bioinformatics, 2022, **23(2)**, 1–12
https://doi.org/10.1093/bib/bbab572
Problem Solving Protocol

Rare variant association tests for ancestry-matched case-control data based on conditional logistic regression

Shanshan Cheng[†], Jingjing Lyu[†], Xian Shi, Kai Wang, Zengmiao Wang, Minghua Deng, Baoluo Sun and Chaolong Wang

Corresponding authors. Chaolong Wang, Ministry of Education Key Laboratory for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, 13 Hangkong Road, Wuhan 430030, China. E-mail: chaolong@hust.edu.cn; Baoluo Sun, Department of Statistics and Data Science, National University of Singapore, 6 Science Drive 2, Singapore 117546, Republic of Singapore. E-mail: stasb@nus.edu.sg.
[†]These authors are joint first authors.

Abstract

With the increasing volume of human sequencing data available, analysis incorporating external controls becomes a popular and cost-effective approach to boost statistical power in disease association studies. To prevent spurious association due to population stratification, it is important to match the ancestry backgrounds of cases and controls. However, rare variant association tests based on a standard logistic regression model are conservative when all ancestry-matched strata have the same case-control ratio and might become anti-conservative when case-control ratio varies across strata. Under the conditional logistic regression (CLR) model, we propose a weighted burden test (CLR-Burden), a variance component test (CLR-SKAT) and a hybrid test (CLR-MiST). We show that the CLR model coupled with ancestry matching is a general approach to control for population stratification, regardless of the spatial distribution of disease risks. Through extensive simulation studies, we demonstrate that the CLR-based tests robustly control type 1 errors under different matching schemes and are more powerful than the standard Burden, SKAT and MiST tests. Furthermore, because CLR-based tests allow for different case-control ratios across strata, a full-matching scheme can be employed to efficiently utilize all available cases and controls to accelerate the discovery of disease associated genes.

Keywords: rare variant association tests, common controls, population stratification, matched analysis, conditional logistic regression

Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants robustly associated with human genetic diseases or quantitative traits [1, 2]. However, a majority of these variants have small to moderate effects. GWAS of complex disease typically require a large sample size to detect the weak association signals. With sequencing technologies, human genetic studies have shifted the focus to rare variants, especially those in the coding region, because protein-altering mutations are more likely to be causal [3–7]. These studies require an even larger sample size to achieve reasonable statistical power due to the low allele frequencies of rare variants [8]. For example, the International AMD Genomics Consortium identified 16 novel genetic loci, including several rare protein-altering mutations, associated with age-related macular degeneration (AMD) by genotyping 16 144 cases and 17 832 controls on a customized exome chip [9]. Recent large-scale exome sequencing of near half million UK Biobank participants further highlighted the role of rare variants in numerous complex traits and diseases [3, 10]. Although both genotyping and sequencing costs have dropped dramatically, it is still prohibitively expensive for many studies when the required sample size is very large.

A cost-effective approach to increase sample size is to leverage existing data of common controls, so that more experimental efforts can be directed toward cases in the study [11–13]. This approach requires careful selection of controls of genetic ancestry background similar to the cases to avoid false positive signals caused by population stratification. Several matching-based methods have been proposed for GWAS, including algorithms based on principal component analysis (PCA) [13], genetic similarity score [14] and stratification score [15, 16]. To enable ancestry matching for both

Shanshan Cheng is an associate professor at the Ministry of Education Key Laboratory for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.
Jingjing Lyu, **Xian Shi** and **Kai Wang** are graduate students at the School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.
Zengmiao Wang is a lecturer at the State Key Laboratory of Remote Sensing Science, Center for Global Change and Public Health, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China.
Minghua Deng is a professor at the Center for Quantitative Biology, School of Mathematical Sciences and Center for Statistical Sciences, Peking University, Beijing, China.
Baoluo Sun is an assistant professor at the Department of Statistics and Data Science, National University of Singapore, Singapore.
Chaolong Wang is a professor at the Ministry of Education Key Laboratory for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.

Received: September 8, 2021. **Revised:** November 29, 2021. **Accepted:** December 13, 2021

genotyped and sequenced samples, we have developed a method called LASER, which can use small amounts of sequence reads or genotypes to accurately estimate individual ancestry in a predefined ancestry space [17, 18]. Application of LASER in a target sequencing study has led to the discovery of a rare protein-altering variant associated with the risk of AMD [12]. As genetic data from population-scale biobanks accumulate, advanced statistical methods to efficiently analyze matched case-control data are crucial to accelerate genetic discovery.

For matched case-control data, conditional logistic regression (CLR) test and Cochran–Mantel–Haenszel (CMH) test are typically used to account for the matching structure. It is well known that naïve analysis ignoring the matching structure will result in biased and often conservative estimation of the relative risk [19]. Compared with the CMH test, CLR test allows adjustment of covariates in a flexible regression framework, making it appealing to statistical geneticists [13, 14]. Previous matching-based genetic association studies focused on single-variant analysis, which has little power for rare variants. Instead, rare variant association studies usually adopt more powerful methods to test the joint effect of multiple rare variants within a gene or a pathway [3]. Many gene-based rare variant association tests have been developed [20], which can be generally classified into the weighted burden test [21, 22], the variance component test such as the sequence kernel association test (SKAT) [23], the hybrid test of burden and variance component, such as SKAT-O [24], and the mixed effect score test (MiST) [25]. Extensions of these gene-based tests to matched case-control data have not been discussed in the literature. We will fill the gap in this paper by evaluating and extending gene-based association tests to matched analysis of case-control data with population structure. Under CLR, we develop the analogous CLR-Burden test, CLR-SKAT and CLR-MiST within the variance component framework to account for the matching structure. Through extensive simulation studies, we show that CLR-based tests have better control of type 1 errors and are more powerful than the standard Burden, SKAT and MiST tests in the analysis of ancestry-matched case-control data. Our methods provide a general solution to control for population stratification when utilizing common controls.

Methods

Population disease risk model and matched case-control study

In the context of case-control GWAS studies, let $\mathrm{Y} = 1$ if the disease of interest is present and $\mathrm{Y} = 0$ otherwise. Let G denote the genotype under the additive mode of inheritance, where G is a scalar if interest lies on a candidate SNP, and vector-valued for multiple SNPs in gene-level tests. In addition, let X denote covariates to be adjusted for, and let U denote the unknown population structure in the sample. Assume that the population disease risk model is

$$P\left(\mathrm{Y} = 1|U = u, \mathrm{X} = \mathrm{x}, \mathrm{G} = g\right) \\ = \left\{1 + \exp\left[-\left(\xi(u) + \mathrm{x}^{\mathrm{T}}\gamma + g^{T}\beta\right)\right]\right\}^{-1}, \qquad (1)$$

where A^{T} indicates the transpose of A, and the function $\xi(\cdot)$ that encodes population structure on the odds ratio scale is left unspecified. In order for the parameter β to be identifiable, we assume that Z, a possibly low-dimensional representation of the observed genome-wide markers, is sufficiently informative about U so that Z can replace U in Equation (1) [26]. Since the space spanned by PCA on the genotypic matrix of the study subjects represents an optimal low-dimensional embedding, it is natural to represent Z by the top PCs. For low-coverage or target sequencing data, such ancestry-informative PCs can be obtained using LASER [17, 18]. Under these assumptions, we have

$$P\left(\mathrm{Y} = 1|Z = \mathrm{z}, \mathrm{X} = \mathrm{x}, \mathrm{G} = g\right) \\ = \left\{1 + \exp\left[-\left(\xi^{*}(\mathrm{z}) + \mathrm{x}^{\mathrm{T}}\gamma + g^{T}\beta\right)\right]\right\}^{-1}, \qquad (2)$$

where $\xi^{*}(\cdot)$ is again an arbitrary function distinct from $\xi(\cdot)$. We note that under the null hypothesis $\beta = 0$, Model (2) is semiparametric since no assumptions are made as to the form of $\xi^{*}(\cdot)$ which encodes population structure, conditional on a linear function of covariates on the odds ratio scale. On the other hand, $\xi^{*}(\cdot)$ is often modeled as a linear combination of the top PCs, which may be misspecified and lead to improper control of type 1 error, a major issue called population stratification in GWAS. The optimal parametric approach to incorporate PCs in adjusting for population structure remains challenging [27].

If Z is discrete-valued, $\xi^{*}(\cdot)$ can be easily modeled nonparametrically. If Z contains continuous components, such as top PCs, one could in principle attempt to make $\xi^{*}(\cdot)$ more flexible by incorporating interactions and nonlinearities in Z, but this issue may not be adequately resolved without nonparametric smoothing methods [28], which are complicated when many PCs are included in the model due to the curse of dimensionality. Furthermore, although it is well known that the estimation of the parameter of interest β and its covariance remains valid in prospective analysis when the true disease risk Model (2) is fitted to unmatched case-control data [29], this is no longer the case with matched data [30]. For example, even if the linear model $\xi^{*}(\mathrm{z}) = \theta^{\mathrm{T}}\mathrm{z}$ (where z denotes the vector of top PCs) is true in (Equation 2), a nonlinear function is generally required in the prospective analysis model fitted to matched data for valid estimation of β[31]. Misspecified prospective models for matched retrospective designs may lead to bias and improper type 1 error control.

Gene-based tests under conditional logistic model

CLR methods have been proposed for matched case-control designs, which maximize the retrospective likelihood conditional on the unordered set of exposures for the cases and controls in each matched group [32, 33]. Let $Y_{i\cdot} = (y_{i1}, \ldots, y_{ik(i)})$ be the case indicators of the $k(i)$ individuals in the ith matched group, $i = 1, 2, \ldots, n$. For simplicity, we assume each group has only one case matched to multiple controls; the extension to arbitrary numbers of cases and controls per group is described in the Supplementary Note (see Supplementary Data available online at http://bib.oxfordjournals.org/). Let $X_{i\cdot} = (x_{i1}, \ldots, x_{ik(i)})$ be the $p \times k(i)$ covariate matrix and $G_{i\cdot} = (g_{i1}, \ldots, g_{ik(i)})$ be the $q \times k(i)$ genotype matrix for q SNPs within a gene or pathway. Based on Model (2), the retrospective conditional likelihood of n matched groups is

$$l(\gamma, \beta) = \prod_{i=1}^{n} \frac{\prod_{j=1}^{k(i)} \exp\left[y_{ij}\left(x_{ij}^T\gamma + g_{ij}^T W\beta\right)\right]}{\sum_{j=1}^{k(i)} \exp\left(x_{ij}^T\gamma + g_{ij}^T W\beta\right)}, \tag{3}$$

where $W = \mathbf{diag}(w_1, w_2, \ldots, w_q)$ is a $q \times q$ weight matrix for q SNPs, γ is a $p \times 1$ vector of fixed effects for the covariates and β is a $q \times 1$ vector of random effects for the SNPs. The vector β has mean $\mathbf{1}_q\beta_0$ and covariance matrix $\tau[(1-\rho)\mathbf{I}_q + \rho\mathbf{1}_q\mathbf{1}_q^T]$, where τ is the variance component, ρ encodes the correlation of effect sizes between SNPs, $\mathbf{1}_q$ is a $q \times 1$ vector of ones and $\mathbf{I}_q$ is a $q \times q$ identity matrix. We note that the population structure term $\xi^*(z)$ has been eliminated from both numerator and denominator of the conditional likelihood (Equation 3), by assuming $\xi^*(z)$ is constant in each group matched for z. Thereby, deriving association tests based on (Equation 3) can robustly control for population structure without explicitly specifying the effect of the ancestry-informative matching variable Z on the disease status.

Because the conditional likelihood function (Equation 3) is equivalent to the partial likelihood function for the Cox regression [30], existing off-the-shelf algorithms for fitting the Cox Proportional Hazards model may be used to maximize (Equation 3) when computing the association test statistics described below. We note in the Supplementary Note (see Supplementary Data available online at http://bib.oxfordjournals.org/) that fitting the exact conditional likelihood (Equation 3) may be computationally demanding when there are arbitrary numbers of cases and controls in each matched group, in which case the Efron or Breslow's approximation algorithms may be used to increase efficiency [34].

CLR-SKAT

CLR-SKAT is derived as the variance component score test under the null hypothesis of $\tau = 0$, assuming $\beta_0 = 0$ and $\rho = 0$. Up-weighting causal variants can increase the power of CLR-SKAT. The weights $w_j = \mathbf{Beta}(\mathbf{MAF}_j, 1, 25)$, a beta distribution density function evaluated at sample minor allele frequency (MAF) for the j**th** variant, have been proposed, which up-weighs rare variants while maintaining decent nonzero weights for variants in the 1–5% MAF range [23]. It is shown in the Supplementary Note (see Supplementary Data available online at http://bib.oxfordjournals.org/) that the score statistic with respect to τ is given by

$$Q = \left(Y - \hat{\mu}\right)^T GWWG^T \left(Y - \hat{\mu}\right), \tag{4}$$

where $Y = (Y_{1\cdot}, \ldots, Y_{n\cdot})^T$ and $G = (G_{1\cdot}, \ldots, G_{n\cdot})^T$. The term $\hat{\mu} = (\hat{\mu}_{1\cdot}, \ldots, \hat{\mu}_{n\cdot})^T$ is estimated under the null, where $\hat{\mu}_{i\cdot} = (\hat{\mu}_{i1}, \ldots, \hat{\mu}_{ik(i)})^T$ for the ith matched group, and

$$\hat{\mu}_{im} = \frac{\exp\left(x_{im}^T\hat{\gamma}\right)}{\sum_{j=1}^{k(i)} \exp\left(x_{ij}^T\hat{\gamma}\right)}, m = 1, \ldots, k(i), \tag{5}$$

where $\hat{\gamma}$ maximizes the likelihood (Equation 3) under the null. $\hat{\mu}_{im}$ in Equation (5) corresponds to the retrospective probability of the m**th** individual being the case, conditional on the $k(i)$ covariate values in the group. We show in the Supplementary Note (see Supplementary Data available online at http://bib.oxfordjournals.org/) that the SKAT statistic in (Equation 4) asymptotically follows a mixture chi-square distribution $Q \sim \sum_{j=1}^{q} \lambda_j \chi_{1,j}^2$, where λ_j are the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ of the vector $WG^T(Y - \hat{\mu})$ and $\chi_{1,j}^2$ are independent chi-square distributions with 1 degree of freedom.

The SKAT statistic in (Equation 4) has an equivalent representation as $Q = \sum_{j=1}^{q} \sigma_{jj} z_j^2$, where σ_{jj} is the jth diagonal element of $\boldsymbol{\Sigma}$ and z_j is the marginal CLR standardized score statistic for the jth variant (Supplementary Note, see Supplementary Data available online at http://bib.oxfordjournals.org/). It has been previously shown that replacing the squared score statistics by their corresponding marginal likelihood-ratio test (LRT) statistics leads to better performance in finite samples, although they are asymptotically equivalent [35, 36]. Therefore, we propose the LRT-based SKAT statistic

$$Q_{\mathbf{CLR-SKAT}} = \sum_{j=1}^{q} \sigma_{jj} z_{\mathbf{LRT},j}^2, \tag{6}$$

where $z_{\mathbf{LRT},j}$ is the square root of LRT chi-square statistic from the marginal test for the jth variant under the conditional logistic model. In practice, there may be convergence issues when fitting the conditional logistic model for extremely rare variants. In such cases (specifically for variants with minor allele count MAC < 10), we use the squared score statistic in place of the LRT statistic [36].

CLR-Burden

Collapsing rare variants within genomic regions into a single variable is another common strategy for gene-based testing [13, 22, 37, 38]. CLR-Burden is derived as the mean score test under the null hypothesis of $\beta_0 = 0$

assuming the variance component $\tau = 0$

$$Q_{\mathrm{CLR-Burden}} = \left(Y - \hat{\mu}\right)^T GW\mathbf{1}_q\mathbf{1}_q^T WG^T \left(Y - \hat{\mu}\right), \qquad (7)$$

which asymptotically follows a chi-square distribution of $(\mathbf{1}_q^T\boldsymbol{\Sigma}\mathbf{1}_q)\chi_1^2$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $WG^T(Y - \hat{\mu})$. Similar to the derivation of CLR-SKAT, we propose using the asymptotically equivalent LRT chi-square statistic for the genetic burden to obtain better finite-sample performance.

CLR-MiST

The CLR-Burden test is more powerful when most variants in the genomic region are causal and influence the trait in the same direction, while CLR-SKAT is more powerful in the presence of heterogeneous effects [39]. A computationally efficient hybrid mixed effects score test (MiST), which combines burden and SKAT-type tests to account for potential heterogeneous variant effects, has been shown to be powerful under a wide range of scenarios [25]. We derive an analogous hybrid test under conditional logistic regression (CLR-MiST) by combining CLR-Burden statistic with an asymptotically independent CLR-SKAT-type statistic

$$\widetilde{Q}_{\mathrm{CLR-SKAT}} = \left(Y - \widetilde{\mu}\right)^T GWWG^T \left(Y - \widetilde{\mu}\right),$$

where $\widetilde{\mu}$ is estimated under a null model with the genetic burden included as an additional covariate. Let $p_{\mathrm{CLR-Burden}}$ and $\widetilde{p}_{\mathrm{CLR-SKAT}}$ be the *P*-values of the CLR-Burden and CLR-SKAT-type tests, respectively. By Fisher's method, CLR-MiST rejects the null hypothesis of $\tau = 0$ with $\rho = 0$ and β_0 unrestricted at an overall significance level of α if

$$-2\log\left(p_{\mathrm{CLR-Burden}}\right) - 2\log\left(\widetilde{p}_{\mathrm{CLR-SKAT}}\right) \geq \chi^2_{4,\alpha}.$$

This approach has been previously used to derive a hybrid test under the logistic mixed model [40].

Matching algorithm

Given the ancestry coordinates such as PCs of a set of cases and controls, we compute the pair-wise Euclidean distance between any case-control pair. An optimal match to minimize the overall distance of matched cases and controls can be identified using the R package *optmatch* [41]. Specifically, we minimized the quantity $\sum_{i=1}^{n}\sum_{j\in A(i),k\in B(i)} D_{ijk}$, where $A(i)$ and $B(i)$ are the sets of indices for cases and controls in the *i*th matched group, *n* is the total number of matched groups and D_{ijk} is the Euclidean distance in the top PCs between the *j*th case and kth control in the *i*th matched group. To improve matching speed and accuracy, a caliper can be imposed so that no case-control pair will be matched if their Euclidean distance exceeds a prespecified limit.

Three matching schemes are considered: (i) 1-to-1 pair match where each case is matched to a control; (ii) 1-to-*m* pair match where each case is matched to *m* controls and (iii) full match where each group contains arbitrary numbers of cases and controls. Given a matching configuration, the statistical power is equivalent to a sample of

$$n_e = \sum_{i=1}^{n} \frac{2k_i k_i'}{k_i + k_i'}$$

pairs of 1-to-1 match, where *n* is the number of matching strata, k_i and k_i' are the numbers of cases and controls in the *i*th stratum, respectively, and n_e is often called the effective sample size [42].

In practice, when the sample size is large (e.g. 1000s of cases versus 10 000s of controls), the *pairmatch* function in the *optmatch* R package is computationally slow and may not guarantee to return a solution for 1-to-*m* match. Thus, we propose a heuristic algorithm to perform 1-to-*m* match given $D_{n_1 \times n_0}$, a distance matrix between n_1 cases and n_0 controls (assuming $n_0 > n_1$ and *m* is a positive integer) as follows.

(i) Set elements in *D* to missing if they are greater than the prespecified caliper limit;
(ii) Sort nonmissing elements in *D* and set $m_j = 0$ for $j = 1, 2, \ldots, n_1$;
(iii) For the smallest D_{jk} in the case and control pools, match control *k* to case *j* and remove *k* from control pool, set $m_j = m_j + 1$ and remove *j* from the case pool if $m_j = m$;
(iv) Repeat step 3 until either the control pool or the case pool is empty.

When $n_0 < mn_1$ or when the caliper limit is stringent, some cases might have less than *m* matched controls. These cases will be excluded, resulting in a smaller effective sample size. To minimize loss of sample, we attempt to rematch cases and controls among these partially matched pairs whose m_j lie between 0 and *m*. For each case, its potential controls are the controls in these partially matched pairs and within the caliper to the case. We start from the case with the smallest number of potential controls. If there are ≥ *m* potential controls in the control pool for a case, we assign *m* controls closest to the case and remove these *m* controls from the control pool; otherwise, move to the next case. This rematching algorithm is heuristic but works fine to help increase the number of 1-to-*m* matched pairs for downstream analysis.

Simulation of genotypes

We used the coalescent simulator *ms* to generate genotypes for 20 000 diploid individuals evenly distributed on a 20 × 20 lattice (50 individuals on each grid in the lattice; Figure 1A) [43]. The scaled migration rate between neighboring grids was chosen as $M = 10$ to mimic the population structure in Europe [44, 45]. We simulated 1

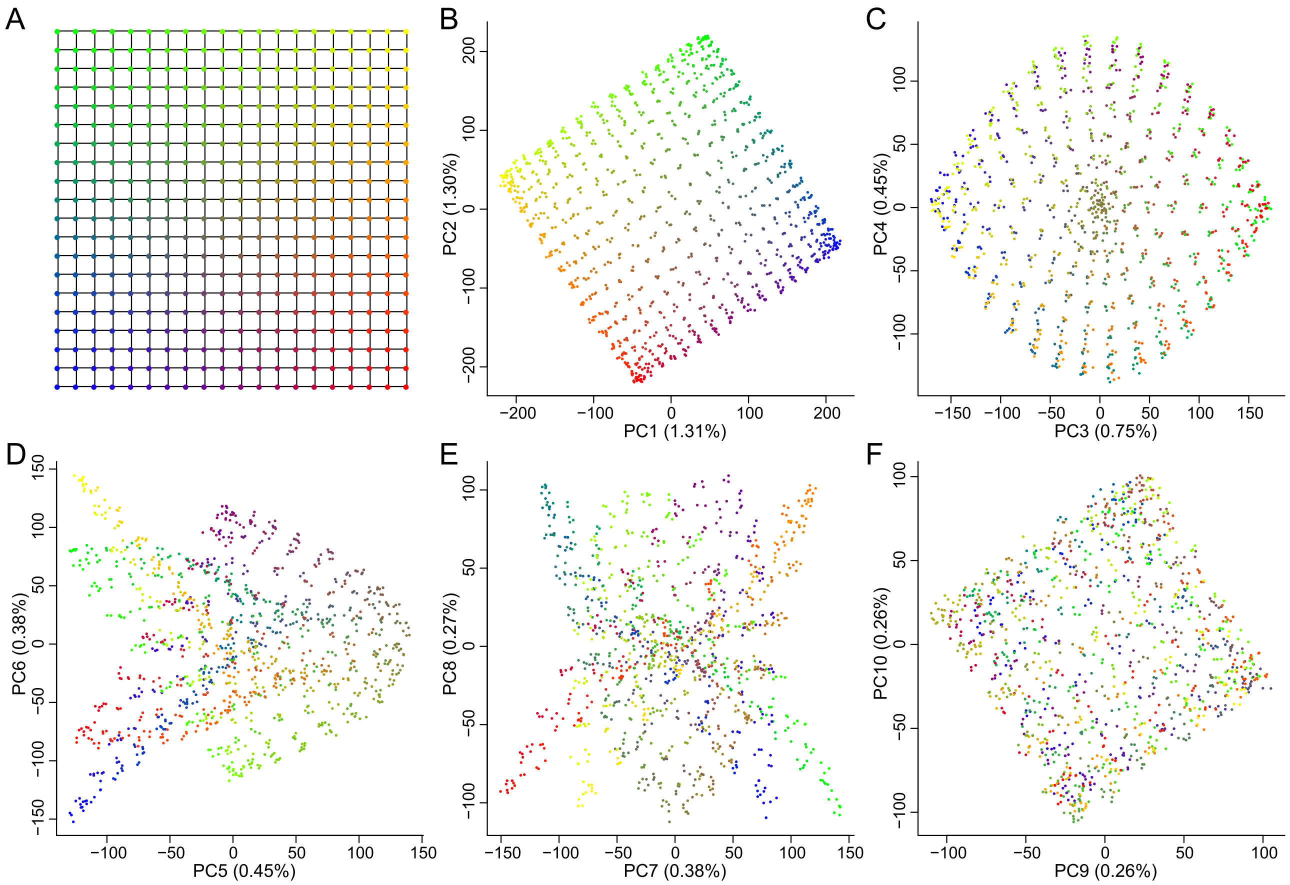

Figure 1. Coalescent simulation of genotype data. (**A**) Geographic map of 400 populations on a 20 × 20 lattice. Each colored dot represents a population. Migrations were allowed between neighboring populations with a scaled migration rate of $M = 10$. (**B**–**F**) Top 10 PCs derived from PCA of the simulated genotypes of 1200 reference individuals (3 individuals per population). Colors follow panel **A** to indicate populations.

million biallelic SNPs with MAF >0.001 on 20 000 independent genealogies. Each genealogy has 50 SNPs and is considered as a single gene for gene-based analysis. Rare variants are defined as SNPs with MAF < 0.05 among all simulated individuals, resulting in an average of 29 rare variants per gene. We randomly selected three individuals from each grid to make up a reference population consisting of 1200 individuals. A reference ancestry map was constructed by PC1 and PC2 of PCA on the reference population (Figure 1B), because the first two PCs mimic the spatial distribution of populations (Figure 1A) while higher order PCs reflected other structure patterns (Figure 1C–F). We projected the remaining individuals onto the ancestry map of top two PCs using LASER [17, 18].

Evaluation of type 1 error

To evaluate type 1 error, we simulated case-control status for each individual under the null disease risk model

$$P\left(Y = 1|U = u, G = g\right) = \left\{1 + \exp\left[-\xi(u)\right]\right\}^{-1}, \qquad (8)$$

where U is a variable denoting the spatial position (i.e. grid coordinates u). We considered two different scenarios of the spatial risk distribution $\xi(u)$: a smooth scenario where the disease risk decays smoothly from a grid in the top right corner of the 20 × 20 lattice (Figure 2A), and a sharp scenario where cases are preferentially drawn from a 7 × 7 grid at the top right corner of the 20 × 20 lattice (Figure 2B). Detailed parameter settings for each scenario are described in the Supplementary Note (see Supplementary Data available online at http://bib.oxfordjournals.org/).

The simulations above generated pools of cases and controls for downstream analyses. In addition to analysis using all simulated individuals, we considered scenarios with controls uniformly distributed across the lattice. Specifically, we randomly kept 20 controls per grid for subsequent matching analysis; the remaining controls were excluded from the control pool. If a grid had less than 20 controls, then all the available controls within that grid were kept. This control down-sampling exercise approximately halved the number of available controls for matching, resulting in fewer available matched controls per case for grids with higher disease risk.

We then randomly sampled 1000 cases from the case pool and either (i) randomly sampled 1000 controls from the control pool to investigate the effects of population stratification without matching or (ii) sampled matched controls for the cases by three matching schemes based on Euclidean distance in the first two PCs: 1-to-1 match, 1-to-3 match and full match, where the caliper width

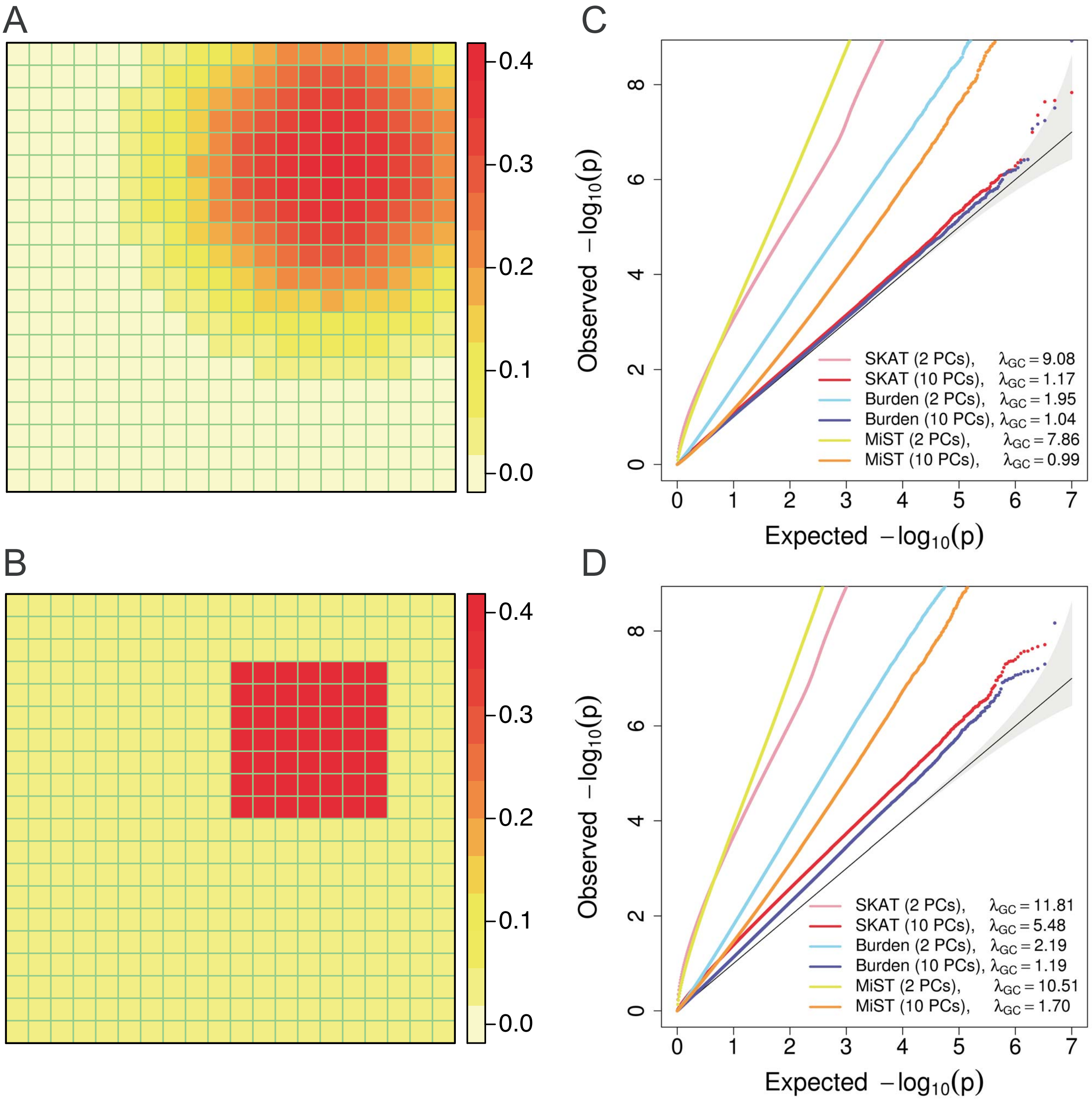

Figure 2. Illustration of population stratification under different simulation settings. (**A**) Map illustrating the smooth spatial distribution of disease risk in a 20 × 20 lattice. Grid color indicates the probability of being a case for individuals in the grid (Equation 8). (**B**) Map illustrating the sharp spatial distribution of disease risk. (**C**) QQ plots of SKAT, Burden and MiST on randomly selected cases and controls under the smooth risk setting, adjusting for the top 2 or 10 PCs. (**D**) QQ plots of SKAT, Burden and MiST on randomly selected cases and controls under the sharp risk setting, adjusting for the top 2 or 10 PCs. The genomic inflation factor λ_{GC} for each test was indicated in the legend.

was set to approximate the 95% prediction interval for individuals from the same grid (Supplementary Note, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Each simulation was repeated to produce 1000 independent case-control phenotype replicates. In each replicate, we performed gene-based association tests for 10 000 genes employing the regular SKAT, Burden and MiST based on the prospective logistic model adjusting for top 2 PCs as linear covariates, and the conditional logistic model-based CLR-SKAT, CLR-Burden and CLR-MiST. Throughout, we used the weights $w_j = \mathbf{Beta}(\mathbf{MAF}_j, 1, 25)$, and we applied the small-sample adjustment procedure to SKAT [46]. We estimated type 1 error based on 10^7 association tests.

Evaluation of power

We followed the same procedure as in type 1 error simulations, except that we simulated case-control status under the alternative model:

$$P(Y = 1|U = u, G = g) = \left\{1 + \exp\left[-\left(\xi(u) + g^T\beta\right)\right]\right\}^{-1}, \quad (9)$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_q)^{\mathrm{T}}$ is the vector of effect sizes for q variants. We varied the proportion of causal variants in a gene by randomly selecting 20 or 50% of the variants to have nonzero effect sizes and setting the remaining variants to have zero effect sizes. Within the causal variants, we varied the directions of their influences on the trait by randomly setting 100/0 or 80/20% of the variants to be deleterious/protective. For causal variants, we set $| \beta_j \Big| = c \Big| \mathbf{log_{10} MAF}_j |$, where c is a scaling constant, so that rarer variants had greater effects [23]. To distinguish the powers, we set $c = 0.877$ and 0.555 when 20 and 50% of the variants are causal, respectively. Detailed parameter settings are described in the Supplementary Note (see Supplementary Data available online at http://bib.oxfordjournals.org/). Based on 1000 simulation replicates, empirical power was estimated as the proportion of tests with P-values smaller than $\alpha = 2.5 \times 10^{-6}$, a threshold considering multiple testing adjustment of ~20 000 genes across the human genome. To investigate the effects of down-sampling controls, we randomly selected half of the available controls in all grids and excluded the remaining ones from the control pool for matching.

Results

Evaluation of type 1 errors

Without case-control matching, the type 1 errors for the SKAT, Burden and MiST are inflated due to population stratification, even after adjustment of 10 PCs, highlighting the challenge to control for population stratification (Figure 2C and D). The inflation is more pronounced for the sharp spatial risk distribution where the nongenetic risk is a highly non-linear function of the PCs, an observation in agreement with previous simulation studies [44].

For case-control matching, because control down-sampling better mimics the situation in practice whereby not all the available controls in a target population are sampled and included in sequencing studies, we present subsequent results under the down-sampling scenario. The results without down-sampling are reported in the Supplementary Data, see Supplementary Data available online at http://bib.oxfordjournals.org/. As shown in Figure 3, SKAT, Burden and MiST are conservative under the 1:1 and 1:3 case to control ratio matching schemes. When cases and controls are proportionally matched, there is no population stratification between cases and controls. The conservativeness is caused by ignoring the matching clusters in the derivation of regular test statistics. In contrast, CLR-SKAT, CLR-Burden and CLR-MiST maintained empirical type 1 error rates near the corresponding nominal levels (Table 1). Under the full matching scheme (on average, 1 case was matched to 4 controls), SKAT, Burden and MiST are highly inflated, while the CLR-based tests appropriately controlled type 1 error rates. SKAT is generally more anticonservative than the Burden test. In fact, the effect of population stratification under the full matching scheme is similar to the analysis without matching because the majority of cases and controls would be included in the analysis. The poor calibration of SKAT, Burden and MiST can be attributed to misspecification of the nonlinear effects of population structure on the disease risk by linear combinations of top PCs. In contrast, CLR-based tests are robust because the effects of population structure have been eliminated from the conditional likelihood (Equation 3). The conclusions are qualitatively similar when controls are not down-sampled (Figure S1, see Supplementary Data available online at http://bib.oxfordjournals.org/, and Table S1**,** see Supplementary Data available online at http://bib.oxfordjournals.org/).

The degree of conservativeness or anticonservativeness for SKAT, Burden and MiST is more pronounced when the spatial distribution of disease risk is sharp (Figure 3D–F) compared with a smooth risk distribution (Figure 3A–C), suggesting that accounting for the matching structure by regressing on linear terms of top PCs may be less effective for sharp distribution of risk. In the smooth risk setting, the top PCs can capture some of the nongenetic risk variation along the axes of the grid and therefore their inclusion as linear covariates in the logistic model serve to partially adjust for the population stratification. On the other hand, parametric modeling of the sharp spatial distribution of nongenetic risk requires highly nonlinear functions of the PCs, and thus, the type 1 error is less effectively controlled under this setting [44]. The empirical type 1 error rates for the CLR-SKAT, CLR-Burden and CLR-MiST are close to the nominal levels across the range of scenarios considered in the simulation (Figures 3 and S1, see Supplementary Data available online at http://bib.oxfordjournals.org/; Tables 1 and S1, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Because SKAT was based on a score statistic Q_{SKAT}[23], while we adopted the LRT statistic for CLR-SKAT, the differences in type 1 error control may be attributed to different finite-sample performances of score and LRT statistics, rather than modeling assumptions. To investigate this possibility, we also included results for the LRT-based SKAT, denoted as $Q_{\mathrm{SKAT-L}}$[36]. The conclusions for $Q_{\mathrm{SKAT-L}}$ and Q_{SKAT} are qualitatively similar, as shown in Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/, confirming that the different performance of SKAT and CLR-SKAT was attributed to different models.

Evaluation of power

Across the range of simulation scenarios, CLR-Burden, CLR-SKAT and CLR-MiST are more powerful than Burden, SKAT and MiST, respectively, especially when cases and controls are matched pairwise (Figures 4 and S3, see Supplementary Data available online at http://bib.oxfordjournals.org/; Tables 2–3 and S2 and S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). This power gain is due to conservativeness of the Burden, SKAT and MiST under

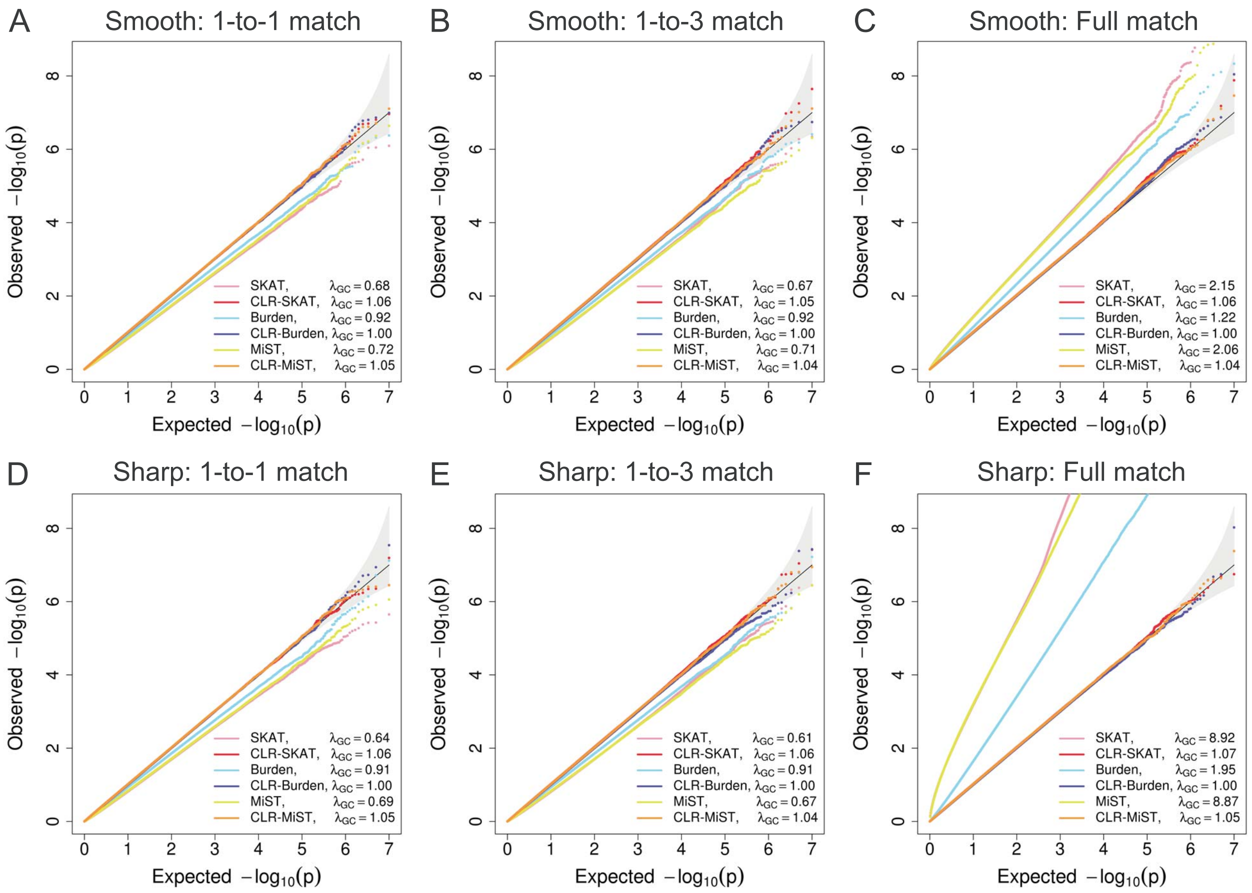

Figure 3. QQ plots for evaluating type 1 errors under the null model with control down-sampling. (**A**) 1-to-1 match under the smooth risk setting. (**B**) 1-to-3 match under the smooth risk setting. (**C**) Full match under the smooth risk setting. (**D**) 1-to-1 match under the sharp risk setting. (**E**) 1-to-3 match under the sharp risk setting. (**F**) Full match under the sharp risk setting. The genomic inflation factor λ_{GC} for each test was indicated in the legend.

Table 1. Empirical relative type 1 errors for simulations with control down-sampling

Spatial risk	Matching scheme	SKAT	CLR-SKAT	Burden	CLR-Burden	MiST	CLR-MiST
Smooth	1-to-1	0.28 (0.36)	1.24 (1.04)	0.24 (0.59)	1.04 (1.04)	0.32 (0.40)	1.28 (1.06)
	1-to-3	0.20 (0.45)	1.36 (1.09)	0.52 (0.61)	1.04 (1.01)	0.20 (0.42)	1.20 (1.08)
	Full	22.3 (5.78)	1.64 (1.08)	7.12 (2.66)	1.76 (1.02)	17.6 (5.64)	1.36 (1.08)
Sharp	1-to-1	0.04 (0.32)	1.04 (1.03)	0.48 (0.55)	1.24 (1.01)	0.20 (0.34)	1.36 (1.04)
	1-to-3	0.24 (0.38)	1.28 (1.08)	0.24 (0.56)	0.56 (1.00)	0.16 (0.35)	1.12 (1.08)
	Full	3590 (119)	1.32 (1.07)	248 (17.0)	0.64 (1.00)	3381 (121)	1.12 (1.06)

Each entry represents the proportion of significant tests out of 10 million simulation replicates at $\alpha = 2.5 \times 10^{-6}$ or 10^{-3} (in the parentheses), divided by α.

the null. The power of all tests increased as more controls are matched to the cases, highlighting the value of aggregating common controls in genetic association tests. In particular, the empirical powers of the CLR-based tests are highest under the full matching scheme. The power simulations for Burden, SKAT and MiST are excluded for full matching because their type 1 errors are not controlled under the null. In general, Burden tests are less powerful than SKAT except when a large proportion of variants are causal and their effects are in the same direction, in agreement with previous simulation results [46]. The hybrid CLR-MiST achieves highest empirical power across the range of scenarios considered in the simulation.

The effective sample size, calculated as the sum of the harmonic mean of the number cases and controls in each group, translates the matched sample into matched pair equivalents [42] and is a good measure of empirical power under each matching scheme. Given 1000 cases, the maximal effective sample sizes are 1000, 1500 and 2000 under 1-to-1, 1-to-3 and full match settings when the number of controls goes to infinity. In practice, due to the limited number of controls, especially under the down-sampling scenario, the effective sample sizes are reduced because some cases are discarded due to insufficient numbers of matched controls (Figures 4 and S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). The gap between the

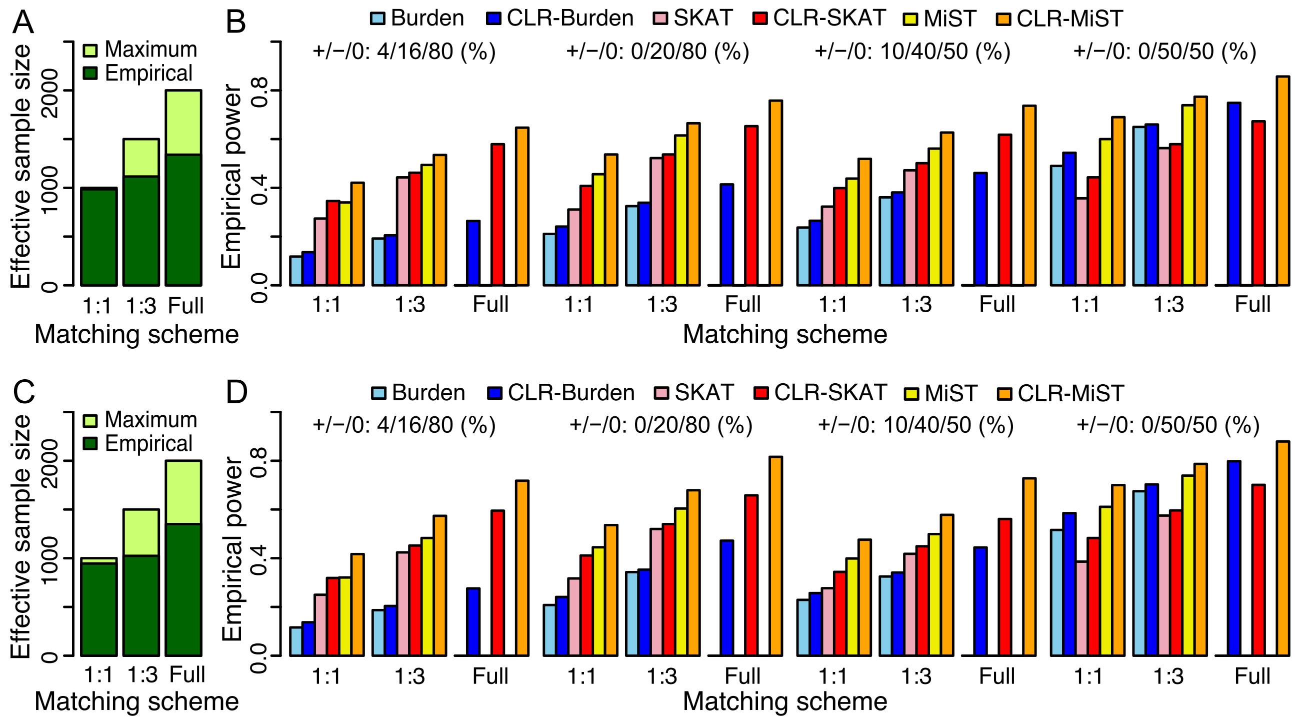

Figure 4. Statistical power evaluated at $\alpha = 2.5 \times 10^{-6}$ based on 1000 simulation replicates with control down-sampling. (**A**) Effective sample size under smooth risk setting. Empirical values are calculated as the average effective sample size across 1000 simulation replicates; maximum values indicate the theoretical upper limit. (**B**) Empirical power under smooth risk for different settings of effect directions for the causal variants. (**C**) Effective sample size under sharp risk setting. (**D**) Empirical power under sharp risk for different settings of effect directions for the causal variants. In (**B**) and (**D**), +/−/0 represent the percentages of protective, deleterious and non-causal variants, respectively.

Table 2. Empirical power for the smooth risk setting with control down-sampling

Effect direction	Matching scheme	SKAT	CLR-SKAT	Burden	CLR-Burden	MiST	CLR-MiST
4/16/80 (+/−/0, %)	1-to-1	0.274	0.346	0.118	0.136	0.340	0.421
	1-to-3	0.443	0.462	0.192	0.205	0.494	0.535
	Full	–	0.579	–	0.264	–	0.647
0/20/80 (+/−/0, %)	1-to-1	0.311	0.408	0.211	0.241	0.456	0.537
	1-to-3	0.522	0.537	0.325	0.339	0.615	0.665
	Full	–	0.653	–	0.414	–	0.758
10/40/50 (+/−/0, %)	1-to-1	0.323	0.399	0.237	0.265	0.438	0.519
	1-to-3	0.472	0.501	0.361	0.381	0.561	0.627
	Full	–	0.618	–	0.461	–	0.737
0/50/50 (+/−/0, %)	1-to-1	0.357	0.443	0.490	0.544	0.600	0.690
	1-to-3	0.563	0.579	0.650	0.660	0.739	0.774
	Full	–	0.673	–	0.749	–	0.857

Each entry represents the proportion of significant tests out of 1000 replicates at $\alpha = 2.5 \times 10^{-6}$. Dashes indicate settings with inflated type 1 error.

empirical effective sample size and the theoretical maximum highlights the potential of power gain as more controls become available.

Discussion

Population stratification is a major challenge in case-control association studies, resulting in poorly calibrated tests and numerous false positive discoveries. Although PCA has been widely used to control for population stratification, especially for tests involving common variants, the effectiveness of PCA depends on the underlying nongenetic risk distribution and population structure [20, 47, 48]. Indeed, the parametric models incorporating top PCs into association tests need to, at least approximately, reflect the risk patterns in each scenario. PCA can effectively capture nongenetic risk factors which vary smoothly in geography because the components with largest genotypic variability often align with geographic maps [49]. But PCA may be less effective when the distributions of the nongenetic risk factors and genetic variants are sharply localized, for instance with rare variants [44]. In the latter case, a highly nonlinear function of the top PCs is required to incorporate the nongenetic risk information into association tests.

We have extended a series of representative rare variant association tests, including Burden, SKAT and MiST,

Table 3. Empirical power for the sharp risk setting with control down-sampling

Effect direction	Matching scheme	SKAT	CLR-SKAT	Burden	CLR-Burden	MiST	CLR-MiST
4/16/80 (+/−/0, %)	1-to-1	0.250	0.319	0.116	0.137	0.321	0.417
	1-to-3	0.424	0.452	0.187	0.204	0.483	0.574
	Full	–	0.595	–	0.276	–	0.718
0/20/80 (+/−/0, %)	1-to-1	0.317	0.411	0.208	0.241	0.445	0.536
	1-to-3	0.520	0.540	0.343	0.353	0.604	0.679
	Full	–	0.658	–	0.472	–	0.816
10/40/50 (+/−/0, %)	1-to-1	0.277	0.344	0.229	0.257	0.399	0.476
	1-to-3	0.418	0.449	0.325	0.341	0.499	0.578
	Full	–	0.561	–	0.444	–	0.728
0/50/50 (+/−/0, %)	1-to-1	0.386	0.483	0.516	0.585	0.611	0.700
	1-to-3	0.575	0.596	0.675	0.703	0.739	0.787
	Full	–	0.701	–	0.798	–	0.879

Each entry represents the proportion of significant tests out of 1000 replicates at $\alpha = 2.5 \times 10^{-6}$. Dashes indicate settings with inflated type 1 error.

under the conditional logistic model, which, by conditioning on the unordered set of exposures for the cases and controls in each matched group, avoids parametric assumptions on the function of PCs. This means that our approach provides a general solution to control for population stratification, regardless of the spatial distribution of disease risks. Across a range of simulation scenarios, our CLR-based tests have been found to be well-calibrated and have greater power over prospective logistic model-based tests, which are conservative under the null. Perhaps more importantly, the logistic model-based tests have inflated type 1 errors under the full matching scheme with arbitrary case to control ratio in each group. In contrast, the CLR-based tests are well-calibrated under the full matching scheme, which enables full utilization of the rich existing data of common controls to boost the statistical power of disease association studies. Besides Burden, SKAT and MiST, there are many other rare variant association tests, each of which has its advantages and disadvantages depending on the study design and genetic architecture of the disease [20]. Thus, it would be useful to extend these rare variant association tests under the CLR framework in the future.

Our implementation of the CLR-based tests is computationally tractable by leveraging existing efficient algorithms for fitting the Cox Proportional Hazards model to maximize the conditional likelihood in CLR, including the Efron or Breslow's approximation algorithms [34] when there are arbitrary numbers of cases and controls in each matched group (see Methods and Supplementary Note, see Supplementary Data available online at http://bib.oxfordjournals.org/). Nonetheless, in our CLR-based tests, we replace the squared score statistic with the LRT statistic for each SNP to achieve better finite-sample performance, which inevitably increases the computational cost, because LRT requires fitting the alternative model for each SNP in the tested genes, while score statistics can be derived by fitting the null model once for all tested genes. We gauged computational time in our simulation studies and found that for the full matching scheme with about 1000 cases and 4000 controls, each test (per gene) costs about 0.2, 1.8 and 2.3 s on a 2.0 GHz CPU for CLR-Burden, CLR-SKAT and CLR-MiST, respectively (Figure S4**,** see Supplementary Data available online at http://bib.oxfordjournals.org/). Considering there are ∼20 000 protein-coding genes in human genome and tests on different genes can be easily run in a massively parallel way on a high-performance computing cluster, our CLR-based tests are computationally tractable in empirical studies with decent sample sizes. Further optimization might be needed when applying to extremely large datasets.

A limitation of the CLR-based tests is that they are developed under the linear kernel. SKAT is able to model epistatic effects of genetic variants under a more flexible kernel machine regression framework [50]. This will entail a more complicated estimation and testing procedure, and we will continue to investigate further into such test statistics. In practice, matching can also be performed based on any covariates, which allows them to be included as a nonparametric component in the logistic model. However, due to the curse of dimensionality, it becomes increasingly difficult to find matched controls as the number of covariates increases, especially if they are continuous or discrete with many categories. We have therefore only considered matching based on ancestry-informative PCs to alleviate confounding by population stratification. Nonetheless, with the rapid expansion of databases for common controls and concomitant lower sequencing costs [3, 51, 52], these obstacles to matching-based studies could eventually be overcome.

Key Points:

- We show that conditional logistic regression (CLR) model coupled with ancestry matching can control for population stratification under arbitrary distribution of disease risks.
- Under CLR model, we develop a weighted burden test (CLR-Burden), a variance component test (CLR-SKAT), and a hybrid test (CLR-MiST), to detect rare variant association signals.
- Our CLR-based tests can facilitate efficient utilization of numerous common controls to

boost the statistical power of disease association studies.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data Availability Statement

No real data were used in this paper. Simulated data are available at https://github.com/chaolongwang/CLoMAT/tree/main/Simulations.

Authors' contributions

C.W. and M.D. conceived and supervised the study. B.S., S.C. and C.W. developed the methods. S.C., J.L., X.S., K.W., Z.W. and B.S. performed simulations and analysis. S.C., B.S. and C.W. drafted the paper. All authors contributed to the revision and approved the final version.

Code availability

R codes of the Conditional Logistic Model Association Tests (CLoMAT) and scripts to reproduce our simulations are available at https://github.com/chaolongwang/CLoMAT.

Funding

Natural Science Foundation of China (81973148).

References

1. Welter D, MacArthur J, Morales J, *et al*. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;**42**:D1001–6.
2. Visscher PM, Wray NR, Zhang Q, *et al.* 10 years of GWAS discovery: biblogy, function, and translation. *Am J Hum Genet* 2017;**101**: 5–22.
3. Wang Q, Dhindsa RS, Carss K, *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* 2021;**597**:527–532.
4. NHLBI Exome Sequencing Project, Do R, Stitziel NO, *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* 2015;**518**:102–6.
5. Lange LA, Hu Y, Zhang H, *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am J Hum Genet* 2014;**94**:233–45.
6. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet* 2012;**13**:135–45.
7. National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC), United Kingdom Inflammatory Bowel Disease Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium, *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 2011;**43**:1066–73.
8. Zuk O, Schaffner SF, Samocha K, *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* 2014;**111**:E455–64.
9. Fritsche LG, Igl W, Bailey JNC, *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet* 2016;**48**: 134–43.
10. Backman JD, Li AH, Marcketta A, *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* 2021;**599**:628–634.
11. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;**447**:661–78.
12. Zhan X, Larson DE, Wang C, *et al.* Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nat Genet* 2013;**45**:1375–9.
13. Luca D, Ringquist S, Klei L, *et al.* On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am J Hum Genet* 2008;**82**:453–63.
14. Guan W, Liang L, Boehnke M, *et al.* Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet Epidemiol* 2009;**33**:508–17.
15. Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet* 2007;**80**:921–30.
16. Epstein MP, Duncan R, Broadaway KA, *et al.* Stratification-score matching improves correction for confounding by population stratification in case-control association studies. *Genet Epidemiol* 2012;**36**:195–205.
17. The FUSION Study, Wang C, Zhan X, *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat Genet* 2014;**46**:409–15.
18. Wang C, Zhan X, Liang L, *et al.* Improved ancestry estimation for both genotyping and sequencing data using projection Procrustes analysis and genotype imputation. *Am J Hum Genet* 2015;**96**:926–37.
19. Breslow NE, Day NE. Statistical methods in cancer research. Volume I - The analysis of case-control studies. *IARC Sci Publ* 1980;5–338.
20. Lee S, Abecasis GR, Boehnke M, *et al.* Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 2014;**95**:5–23.
21. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;**83**:311–21.
22. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009;**5**:e1000384.
23. Wu MC, Lee S, Cai T, *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;**89**:82–93.
24. Lee S, Emond MJ, Bamshad MJ, *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;**91**:224–37.
25. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol* 2013;**37**:334–44.
26. Lin D, Zeng D. Correcting for population stratification in genomewide association studies. *J Am Stat Assoc* 2011;**106**: 997–1008.

27. Peloso GM, Lunetta KL. Choice of population structure informative principal components for adjustment in a case-control study. *BMC Genet* 2011;**12**:64.
28. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, New York. 2009.
29. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979;403–11.
30. Levin B, Paik MC. The unreasonable effectiveness of a biased logistic regression procedure in the analysis of pair-matched case-control studies. *Journal of Statistical Planning and Inference* 2001;**96**:371–85.
31. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Hoboken, New Jersey. 2013.
32. Breslow N, Day N, Halvorsen K, *et al.* Estimation of multiple relative risk functions in matched case-control studies. *Am J Epidemiol* 1978;**108**:299–307.
33. Breslow NE. Statistics in epidemiology: the case-control study. *J Am Stat Assoc* 1996;**91**:14–28.
34. Hertz-Picciotto I, Rockhill B. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics* 1997;**53**:1151–6.
35. Chen H, Lumley T, Brody J, *et al.* Sequence kernel association test for survival traits. *Genet Epidemiol* 2014;**38**: 191–7.
36. Wu B, Pankow JS, Guan W. Sequence kernel association analysis of rare variant set based on the marginal regression model for binary traits. *Genet Epidemiol* 2015;**39**:399–405.
37. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010;**34**:188–93.
38. Price AL, Kryukov GV, de Bakker PIW, *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010;**86**:832–8.
39. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 2011;**35**:606–19.
40. Chen H, Huffman JE, Brody JA, *et al.* Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *Am J Hum Genet* 2019;**104**:260–74.
41. Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. *J Comput Graph Stat* 2006;**15**:609–27.
42. Hansen BB. Propensity score matching to extract latent experiments from nonexperimental data: a case study. In: Dorans NJ, Sinharay S (eds). *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*. New York: Springer, 2011.
43. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 2002;**18**:337–8.
44. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012;**44**:243–6.
45. Chen H, Wang C, Conomos MP, *et al.* Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am J Hum Genet* 2016;**98**: 653–66.
46. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012;**13**:762–75.
47. Babron M-C, de Tayrac M, Rutledge DN, *et al.* Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PLoS One* 2012;**7**:e46519.
48. Liu Q, Nicolae DL, Chen LS. Marbled inflation from population structure in gene-based association studies with rare variants. *Genet Epidemiol* 2013;**37**:286–92.
49. Wang C, Zöllner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet* 2012;**8**:e1002886.
50. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* 2007;**63**:1079–88.
51. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Taliun D, Harris DN, *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;**590**: 290–9.
52. Wu D, Dou J, Chai X, *et al.* Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell* 2019;**179**:736–749 e15.