Submitted to the Annals of Applied Statistics

ADJUSTMENT OF NON-CONFOUNDING COVARIATES IN CASE-CONTROL GENETIC ASSOCIATION STUDIES*

By Hong Zhang[‡,§], Nilanjan Chatterjee[¶], Daniel Rader[§] and Jinbo Chen[§,†]

Fudan University[‡], University of Pennsylvania[§], National Institutes of Health[¶]

It has recently been reported that adjustment of non-confounding covariates in case-control genetic association analyses may lead to decreased power when the phenotype is rare. This observation contrasts a well-known result for clinical trials where adjustment of baseline variables always leads to increased power for testing randomized treatment effects. In this paper, we propose a unified solution that guarantees increased power through covariate adjustment regardless of whether the phenotype is rare or common. Our method exploits external phenotype prevalence data through a profile likelihood function, and can be applied to fit any commonly used penetrance models including the logistic and probit regression models. Through extensive simulation studies, we showed empirically that the power of our method was indeed higher than available analysis strategies with or without covariate adjustment, and can be considerably higher when the phenotype was common and the covariate effect was strong. We applied the proposed method to analyze a case-control genetic association study on human high density lipoprotein cholesterol level.

1. Introduction. Genome-wide association studies in the past decade have led to discovery of hundreds of susceptible genetic variants for a rich collection of human disease phenotypes. To take advantage of these valuable knowledge for further understanding of genetic basis for human phenotypes, a highly pertinent question is to what extent the known risk variants can be exploited for the discovery of new risk variants (Zaitlen et al., 2012a). One common strategy is to adjust for the known risk variants in regression analyses for assessing the significance of new variants. This practice aligns with a widely known result that adjusting for non-confounding covariates, which refer to genetic or non-genetic covariates that are independent of the

*Partially supported by NIH grants R21-ES020811 and R01-ES016626 (HZ and JC) and National Natural Science Foundation of China 11371101 (HZ).

†To whom correspondence should be addressed.

MSC 2010 subject classifications: Primary 62J12; secondary 62P10

Keywords and phrases: case-control studies, Hardy-Weinberg equilibrium, non-confounding covariates, profile likelihood

test genetic variant and known to be associated with the phenotype, may lead to increasing, or at least non-decreasing, power compared with the unadjusted analyses. For example, adjusting for baseline covariates in analyses of clinical trials leads to increased power for testing randomized treatment effects. Interestingly, it has recently been alerted that the contrary may hold true with case-control data (Kuo and Feingold, 2010; Zaitlen et al., 2012b,a; Pirinen, Donnelly and Spencer, 2012). The decrease in power is a function of phenotype prevalence, case versus control sample size ratio, effect size, and distribution of the non-confounding covariates (Pirinen, Donnelly and Spencer, 2012). In particular, covariate adjustment may lead to decreased power for rare phenotypes but increased power for common phenotypes (Pirinen, Donnelly and Spencer, 2012). An ad hoc explanation has been offered for this phenomenon: the non-confounding covariates and the test variant that are independent in the population become correlated in the case-control sample due to the retrospective sampling (Zaitlen et al., 2012b; Pirinen, Donnelly and Spencer, 2012). It is of great interest to understand the theoretical underpinning of the phenomenon, and to develop unified statistical solutions that can guarantee improved power through adjustment of non-confounding covariates.

Since adjustment for non-confounding covariates in prospective logistic regression analyses of case-control data could lead to decreased power for testing genetic association, unadjusted analysis that ignores covariates has often been conducted (Kuo and Feingold, 2010). Recently, Zaitlen et al. (2012a,b) pointed out that the retrospective sampling scheme is responsible for the covariate adjustment dilemma mentioned above. Under a liability threshold penetrance model, they proposed to exploit external information on overall or covariate-specific phenotype prevalence in the adjusted analysis to increase power. Results from simulation studies showed that their method had improved power regardless of phenotype prevalence. But their method was designed specifically for the liability threshold model. In particular, it is not suitable to use under the most widely adopted logistic regression penetrance model. It cannot efficiently accommodate additional non-confounding covariates for which the covariate specific prevalence information is not available. Furthermore, their method cannot provide consistent parameter estimates for genetic effects. In this work, under the frequency-matched case-control study design, we consider testing and estimation of genetic effects based on a general class of penetrance models for binary phenotypes, which includes the liability threshold (probit) and logistic regression models. In addition to integration of phenotype prevalence information, we conjecture that independence between the genetic variables of interest and non-confounding

variables has to be explicitly taken into account in order to realize the power gain through covariate adjustment. Our reasoning is that available information on genetic association in our work is essentially comparable to that provided in a prospective study. Our method maximizes the potential power increase offered by adjusting for non-confounding covariates, in the same sense that adjusting for baseline covariates leads to improved power for testing randomized treatment effects. Our method allows correlation between the non-confounding and matching variables.

In Section 2, we describe a semiparametric maximum likelihood method for fitting flexible regression models to case-control data. Using the Lagrange multiplier method, we derived the profile likelihood of regression parameters that incorporates known phenotype prevalence and gene-covariate independence. We showed that maximization of the profile likelihood for parameter estimation is a saddle point problem that is often difficult to resolve numerically. We therefore proposed a novel modification to the profile likelihood, which led to an estimator that has the same asymptotic efficiency but eradicates the numerical difficulty. In Section 3, we show results from extensive simulation studies that examined the finite-sample performance of our proposed method. In Section 4, we illustrate our method through application to a case-control genetic association study of human high-density lipoprotein cholesterol (HDL-C). Section 5 concludes with final remarks.

2. Method. Consider a frequency-matched case-control study design, where cases and controls are randomly sampled within strata that are defined by variable S. Suppose S takes I values, $S = 1, \ldots, I$, and let n_s denote the total number of cases and controls sampled from stratum $S = s$. Let D denote the case-control status ($D = 1$: case; $D = 0$: control), X the collection of non-confounding covariates, and G the genotype for a di-allelic single nucleotide polymorphism (SNP) coded as 0, 1, and 2 for the number of minor alleles. X can be either univariate or multivariate. Let (X_{si}, G_{si}) denote the genotype and covariate data for the i^{th} subject in stratum $S = s$ ($i = 1, \ldots, n_s$, $s = 1, \ldots, I$). We use a general penetrance model to describe the association between D and (S, X, G):

$$\text{pr}(D = 1|S = s, X = x, G = g) = h(\alpha + \beta_S s + \beta_X^T x + \beta_G g), \tag{2.1}$$

where h is the inverse of a known link function such as the logistic or probit function, and a^T represents the transpose of vector a. Let $\mathcal{B}$ denote the collection of regression coefficients, $\mathcal{B} = (\alpha, \beta_S, \beta_X^T, \beta_G)^T$. Let $\pi_{si} = \text{pr}(X = X_{si}|S = s)$ be the empirical distribution function of X in stratum s, and $q_g(\theta) = \text{pr}(G = g)$ be the distribution function of G indexed by parameter

θ. We consider that neither X nor S confounds the association between D and G, that is, G and (X, S) are independent, although X and S are allowed to be correlated:

$$\text{pr}(X = X_{si}, G = g|S = s) = \pi_{si} q_g(\theta). \tag{2.2}$$

When the Hardy-Weinberg equilibrium (HWE) holds in the underlying population from which cases and controls arise, θ can be the minor allele frequency (MAF) so that q_g's are defined as $q_0(\theta) = (1-\theta)^2$, $q_1(\theta) = 2\theta(1-\theta)$, and $q_2(\theta) = \theta^2$. Alternatively, θ can be a vector of genotype frequencies. We assume HWE in the subsequent developments. The stratum specific phenotype prevalence, $f_s := \text{pr}(D = 1|S = s)$, is known *a priori* for stratum $s = 1, \ldots, I$.

2.1. *The retrospective likelihood function.* Denote $\Theta = (\theta, \mathcal{B}^T)^T$ and $\boldsymbol{\pi} = \{\pi_{si} : i = 1, \ldots, n_s; s = 1, \ldots, I\}$. Note that the number of the elements in $\boldsymbol{\pi}$ is independent of the dimension of X. Under HWE and gene-covariate independence, the retrospective log-likelihood, $\sum_{s=1}^{I} \sum_{i=1}^{n_s} \log \text{pr}(X_{si}, G_{si}|S = s, D_{si})$, as a function of Θ and $\boldsymbol{\pi}$, can be written as

$$\begin{aligned} \ell(\Theta, \boldsymbol{\pi}) \quad &= \quad \sum_{s=1}^{I} \sum_{i=1}^{n_s} \log\{\mathcal{G}(D_{si}, s, X_{si}, G_{si}; \mathcal{B}) q_{G_{si}}(\theta)\} \\ &\quad + \sum_{s=1}^{I} \sum_{i=1}^{n_s} \log(\pi_{si}) - \sum_{s=1}^{I} \sum_{i=1}^{n_s} \log\{f_s^{D_{si}} (1 - f_s)^{1-D_{si}}\}, \end{aligned} \tag{2.3}$$

where $\mathcal{G}(d, s, x, g; \mathcal{B}) := 1 - d + (2d - 1)h(\alpha + \beta_S s + \beta_X^T x + \beta_G g)$ is the probability of $D = d$ given $S = s$, $X = x$, and $G = g$ according to model (2.1). The empirical probability masses $\{\pi_{si} : i = 1, \ldots, n_s; s = 1, \ldots, I\}$ satisfy the following set of constraints:

$$\sum_{i=1}^{n_s} \pi_{si} = 1, \quad s = 1, \ldots, I. \tag{2.4}$$

Another set of constraints that should be satisfied to incorporate the known prevalence data are

$$\text{pr}(D = 1|S = s) = f_s, \quad s = 1, \ldots, I. \tag{2.5}$$

Define $\mathcal{H}_{si}(\Theta)$ as $\sum_{g=0}^{2} \mathcal{G}(1, s, X_{si}, g; \mathcal{B}) q_g(\theta)$. According to the law of total probability, $\text{pr}(D = 1|S = s)$ is equal to $\sum_{i=1}^{n_s} \mathcal{H}_{si}(\Theta)\pi_{si}$, so that (2.5) can be re-written as

$$\sum_{i=1}^{n_s} \{\mathcal{H}_{si}(\Theta) - f_s\}\pi_{si} = 0, \quad s = 1, \ldots, I. \tag{2.6}$$

In Section 2.2, we derive the profile likelihood for parameters Θ based on the Lagrange multiplier formulation, and discuss that it is a saddle point problem to obtain the maximum profile likelihood estimator, "pMLE". To address this numerical challenge, in Section 2.3, we propose to replace the estimated Lagrange multipliers in the profile likelihood by their large sample limits. We show that the resultant estimator "mpMLE" is consistent and has the same asymptotic efficiency as pMLE under some regularity conditions.

2.2. *The profile likelihood method for estimating* Θ. For given Θ, using the method of Lagrange multipliers, we can show that the value of π_{si} satisfying constraints (2.4) and (2.6) takes the form

$$\hat{\pi}_{si}(\Theta, \lambda_s) = \frac{1}{n_s} \frac{1}{1 + \lambda_s\{\mathcal{H}_{si}(\Theta) - f_s\}}, \tag{2.7}$$

where the Lagrange multipliers $\lambda_1, \ldots, \lambda_I$ satisfy the constraints

$$\sum_{i=1}^{n_s} \hat{\pi}_{si}(\Theta, \lambda_s)\{\mathcal{H}_{si}(\Theta) - f_s\} = 0, \quad s = 1, \ldots, I. \tag{2.8}$$

If we write

$$\begin{aligned} \tilde{\ell}(\Theta, \boldsymbol{\lambda}) &= \sum_{s=1}^{I} \sum_{i=1}^{n_s} \log\{\mathcal{G}(D_{si}, S = s, X_{si}, G_{si}; \mathcal{B}) q_{G_{si}}(\theta)\} \\ &\quad + \sum_{s=1}^{I} \sum_{i=1}^{n_s} \log\{\hat{\pi}_{si}(\Theta, \lambda_s)\}, \end{aligned} \tag{2.9}$$

then the profile likelihood function of Θ can be written as

$$\ell_{\mathrm{p}}(\Theta) = \max_{\boldsymbol{\pi}} \ell(\Theta, \boldsymbol{\pi}) = \tilde{\ell}\{\Theta, \boldsymbol{\lambda}(\Theta)\}, \tag{2.10}$$

where $\boldsymbol{\lambda}(\Theta) = \{\lambda_1(\Theta), \ldots, \lambda_I(\Theta)\}$ satisfies constraints (2.8). Consequently, pMLE can be obtained by jointly solving equations (2.8) and

$$\frac{\partial}{\partial \Theta} \tilde{\ell}(\Theta, \boldsymbol{\lambda}) = 0. \tag{2.11}$$

It is noted that equations (2.8) can be written as

$$\frac{\partial}{\partial \boldsymbol{\lambda}} \tilde{\ell}(\Theta, \boldsymbol{\lambda}) = 0. \tag{2.12}$$

The estimating equations (2.11) and (2.12) are the "score" equations derived from the function $\tilde{\ell}(\Theta, \boldsymbol{\lambda})$. Note that $\tilde{\ell}(\Theta, \boldsymbol{\lambda})$ is not a true likelihood

function, since the Lagrange multipliers $\boldsymbol{\lambda}$ are constructed parameters. In fact, the solution $(\tilde{\Theta}, \tilde{\lambda})$ to the "score" equations is not the maximizer of $\tilde{\ell}(\Theta, \boldsymbol{\lambda})$ in general. To illustrate this point, we consider a simple situation where the number of strata is $I = 1$, the covariate X takes only two values 0 and 1, and no genetic effect is considered. In this situation, it turns out that the "profile" log-likelihood function for λ can be obtained explicitly:

$$\ell_\lambda(\lambda) \equiv \max_{\Theta} \tilde{\ell}(\Theta, \lambda) = -n_0 \log(1 - f\lambda) - n_1 \log\{1 + (1 - f)\lambda\} + c, \tag{2.13}$$

where f is the phenotype prevalence, c is a term that is independent of λ, and n_1 and n_0 are the respective numbers of case and control subjects. Furthermore, the value of λ satisfying (2.11) and (2.12) is

$$\tilde{\lambda} = \frac{n_1}{nf} - \frac{n_0}{n(1 - f)}. \tag{2.14}$$

Refer to Appendix S1 (Appendix) for derivation of (2.13) and (2.14). On the other hand, (2.14) happens to be the solution to $\partial \ell_\lambda(\lambda)/\partial\lambda = 0$, and the second derivative function of $\ell_\lambda(\lambda)$ equals $n_0 f^2/(1 - f\lambda)^2 + n_1(1 - f)^2\{1 + (1 - f)\lambda)^2\}$, which is strictly positive. Consequently, the "profile" log-likelihood function $\ell_\lambda(\lambda)$ is convex in λ, so that the solution to the score equations (2.11) and (2.12), $\tilde{\lambda}$, is the minimizer instead of the maximizer of the "profile" likelihood function $\ell_\lambda(\lambda)$ (Figure S1 in Appendix). This is in contradictory to the standard likelihood theory.

The above argument suggests that $(\tilde{\Theta}, \tilde{\boldsymbol{\lambda}})$ be a saddle point of the profile likelihood function $\tilde{\ell}(\Theta, \boldsymbol{\lambda})$. In numerical studies reported in Sections 3 and 4, we found that it was very difficult to obtain estimates by existing numerical optimization algorithms designed for finding global/local maximizer(s), and that the estimates were very sensitive to the initial value used for $(\Theta, \boldsymbol{\lambda})$. The numerical difficulty limits the practical usefulness of the profile likelihood method. In the next subsection, we propose a "limit multiplier" profile likelihood function, based on which we derive a novel estimator of Θ. This new estimator can be numerically obtained in a very reliable manner without compromising statistical efficiency.

2.3. *A limit multiplier profile likelihood method.* We derive the explicit forms of the limit multipliers as follows. Let the true value of Θ be Θ_0 and the limit value of $\boldsymbol{\lambda}$ be $\boldsymbol{\lambda}_0$, which satisfies the following equations:

$$E\left\{\frac{\partial}{\partial\Theta}\tilde{\ell}(\Theta_0, \boldsymbol{\lambda}_0)\right\} = 0 \text{ and } E\left\{\frac{\partial}{\partial\boldsymbol{\lambda}}\tilde{\ell}(\Theta_0, \boldsymbol{\lambda}_0)\right\} = 0. \tag{2.15}$$

The theorem below presents the limit value $\boldsymbol{\lambda}_0$, which depends on both stratum-specific case-control sampling ratios and true phenotype prevalence:

Theorem 1. *Let n_{1s} and n_{0s} denote the numbers of cases and controls, respectively, in the sth stratum. Denote*

$$\lambda_{0s} = \frac{n_{1s}}{n_s f_s} - \frac{n_{0s}}{n_s(1-f_s)}. \tag{2.16}$$

Under certain regularity conditions, $\boldsymbol{\lambda}_0 := (\lambda_{01}, \ldots, \lambda_{0I})$ satisfies equation (2.15).

The proof of Theorem 1 is quite involved and therefore postponed to Appendix S2 (Appendix). Note that the limit multipliers given in (2.16) depend only on the prevalences and sample sizes, and they are free of the unknown parameter vector Θ. Obviously, the limit multipliers λ_{0s}, $s = 1, 2, \ldots, I$, are equal to zero if and only if $f_s/(1-f_s) = n_{1s}/n_{0s}$ (or approximately, a random instead of biased case-control sample is assembled in each matching stratum). When the limit multipliers equal 0, the profile likelihood function $\ell_{\mathrm{p}}(\Theta)$ is exactly the same as the likelihood function of similar data that were collected prospectively, and adding the constraints on the phenotype prevalences does not alter the estimation of Θ. On the other hand, the limit multipliers are non-zero under case-control study design where the sample is not representative of the study population, implying that the constraints on the phenotype prevalences have an impact on the estimation of Θ. The non-zero limit values of multipliers complicate theoretical studies of the asymptotic properties of pMLE $\tilde{\Theta}$, in contrast to many empirical-likelihood based methods, where the limit multipliers were equal to zero (Owen, 1988, 1990; Qin and Lawless, 1994; Qin et al., 2014).

To address the numerical difficulty in calculating pMLE and ease the study of its asymptotic properties, we propose a new estimator "mpMLE", which is defined as the maximizer of the following modified profile likelihood function:

$$\ell_{\mathrm{mp}}(\Theta) = \tilde{\ell}(\Theta, \boldsymbol{\lambda}_0), \tag{2.17}$$

where we simply replace the "estimated" multipliers $\boldsymbol{\lambda}(\Theta)$ with the limit multipliers $\boldsymbol{\lambda}_0$. Obviously, mpMLE is the solution to the following score equation:

$$\frac{\partial}{\partial \Theta} \ell_{\mathrm{mp}}(\Theta) = 0. \tag{2.18}$$

We found that mpMLE was computationally stable in our numerical studies. Theorem 2 below, proved in Appendix S3 (Appendix), shows that mpMLE and pMLE have the same asymptotic efficiency. We first established the

asymptotic properties of mpMLE, which subsequently served as an intermediate step for studying large sample properties of pMLE. The asymptotic variance-covariance matrix can be consistently estimated using a sandwich estimator as described in Appendix S4 (Appendix).

Theorem 2. *Under some regularity conditions, we have the following large sample results. (i) There exists a solution to equations (2.11) and (2.12), denoted as* $(\tilde{\Theta}, \tilde{\boldsymbol{\lambda}})$*, that is consistent for* $(\Theta_0, \boldsymbol{\lambda}_0)$*; there exists a solution to equation (2.18), denoted as* $\hat{\Theta}$*, that is consistent for* Θ_0*. (ii) Both* $\tilde{\Theta}$ *and* $\hat{\Theta}$ *are asymptotically normally distributed, with the same asymptotic expectation* Θ_0 *and the same variance-covariance matrix as given in Appendix S3 (Appendix).*

We recommend mpMLE instead of pMLE in practice. They have the same statistical efficiency, but the compution of mpMLE is stable. We were able to obtain mpMLE reliably from all the datasets that we have so far analyzed. But pMLE appeared to be very sensitive to the choice of initial values in all the numerical algorithms that we attempted. It was difficult to obtain pMLE even if we used estimates from existing methods as initial values. In the HDL-C example reported in Section 4, pMLE failed for all the 64 SNPs analyzed when the results from the standard logistic regression were used for initial values. In the numerical studies described in the next two sections, we used mpMLE as the initial value for pMLE, so that the former was computationally faster. Nevertheless, it remained difficult to compute pMLE had substantially larger variance than mpMLE when the covariate effects were small, the phenotype was rare, or the model was mis-specified; it still failed for one SNP in the HDL-C example.

2.4. *Implementation of the proposed method.* To implement mpMLE, we have developed an R package "CCGA", abbreviation for "Case-Control Genetic Association". CCGA is now freely available at Github, a web-based Git repository hosting service (http://github.com/zhanghfd/CCGA), and users can easily install it with the aid of the R package "devtools". In CCGA, function "SingleSNP" was designed for analyzing a single variant, and function "MultipleSNP" was built upon "SingleSNP" for analyzing GWAs by allowing utilization of multiple CPU cores through parallel processing. The input arguments of the two functions include the link function ("logit" or "probit") and data for the case-control status, stratum membership, SNP genotype(s), and covariate(s). SingleSNP ouputs the estimated log odds ratios (ORs), their standard error estimates, and the p-value for testing the

genetic effect; MultipleSNP outputs the same results for all SNPs. It took SingleSNP only around 0.8 seconds to converge in our numerical studies reported in Sections 3 and 4, and the required memory was quite small. The memory required by MultipleSNP is nearly linear in the number of SNPs.

3. Simulation studies. We conducted extensive simulation studies to evaluate (1) the consistency and efficiency of our proposed method for estimating the OR that measures genetic risk effect and (2) the corresponding type-I error rate and power for testing genetic association. We examined power as a function of phenotype prevalence, case versus control sample size ratio, and ORs that measure genetic and covariate effects. We considered logistic and probit penetrance models and generated data from the frequency-matched case-control design where covariates are available for adjustment. Under both logistic and probit models, we compared our method with relevant existing methods, i.e., the standard logistic regression with ("LOGIT1") or without ("LOGIT0") stratum and covariate adjustment, and a probit-model based method (Zaitlen et al., 2012a,b) with ("LT1") or without "(LT0") adjustment of covariates. We emphasize that no covariate specific prevalence information was available. The stratum specific prevalence information was incorporated in pMLE, mpMLE, LT0, and LT1. The Wald statistic was used for testing statistical significance at the 0.05 level in methods LOGIT0, LOGIT1, pMLE, and mpMLE. In Section 3.1, we described the detailed simulation study design and results under the logistic regression model. We compared two estimators based on either the original profile likelihood ("pMLE.logit") or the limit multiplier profile likelihood ("mpMLE.logit") using the logit link function. In practice, the information on stratum specific prevalences f_s may not be accurate. We therefore evaluated the impact of mis-specifying phenotype prevalence on the relative performance of our method. In Section 3.2, we reported simulation studies under the probit model. We evaluated our method based on the limit multiplier profile likelihood with the probit link function ("mpMLE.probit"), but we did not consider pMLE because of the computation difficulty. We evaluated the impact of mis-specifying the link function on the power of our method under both logistic and probit penetrance models. In Section 3.3, we further evaluated the performance of our method when the non-confounding covariates consisted of 10 SNPs under the logistic penetrance model, aiming to inform the extent of power improvement through adjustment for common susceptible SNPs for testing genetic association.

3.1. *Under a logistic regression model for penetrance.* We generated a stratum variable S from the uniform distribution on $\{1, 2, 3\}$, a SNP genotype

G following HWE with MAF 0.2, and a covariate X from a normal distribution. Here G was independent of S and X, but the latter two were correlated through a hidden variable, i.e., $X \sim N(0.5Z, 1)$ and $S = \sum_{j=1}^{3} I(Z > Z_{(j/3)})$, where $I(\cdot)$ was the indicator function, Z was a standard normal random variable, and $Z_{(r)}$ was the $100r\%$ quantile of the standard normal distribution. The correlation coefficient between X and S was around 0.4. We coded G as the number of minor alleles (0, 1, or 2). The phenotype status D was generated from the logistic regression model:

$$\text{(3.1)} \quad \text{pr}(D = 1|S = s, X = x, G = g) = \frac{\exp(\alpha + \beta_S s + \beta_X x + \beta_G g)}{1 + \exp(\alpha + \beta_S s + \beta_X x + \beta_G g)}.$$

We set the stratum log-odds ratio (log-OR) parameter β_S at $\log(2.0)$, the covariate log-OR β_X at either 0 (zero covariate effect) or $\log(4.0)$ (non-zero covariate effect), and SNP log-OR β_G at either 0 (zero genetic effect) or $\log(1.3)$ (non-zero genetic effect). We chose values of α such that the population phenotype prevalence f was 0.005, 0.05, or 0.2, corresponding to rare phenotype, phenotype of moderate prevalence, and common phenotype. We first generated a population of size 10^7 with each of the 12 parameter combinations. Then in each of the three sampling strata, the stratum specific phenotype prevalence f_s was estimated and assumed known for obtaining estimates pMLE.logit, mpMLE.logit, LT0, and LT1. In each Monte Carlo experiment, 200 cases and 200 controls were randomly drawn from each stratum. To evaluate power improvement with multiple correlated covariates with moderate effects, we generated data similarly but with 10 independent covariates in model (3.1), each following the standard normal distribution and having a log-OR of either 0 (zero covariate effect) or $\log(2.0)$ (non-zero covariate effect). All simulation results were based on 5,000 replicates for each parameter combination.

We first compared the estimators pMLE.logit and mpMLE.logit. Presented in Table 1 are the estimation results of Θ for two parameter combinations, and the other estimation results were similar and not presented. For the first parameter combination ($f = 0.05$, $\beta_X = \log(4.0)$, and $\beta_G = \log(1.3)$), pMLE and mpMLE appeared to perform comparably in terms of bias ("BIAS"), empirical standard error ("SE"), mean estimated asymptotic standard error ("SEE"), and empirical coverage probabilities of the 95% confidence interval ("CP"). For the second parameter combination ($f = 0.005$ and $\beta_X = \beta_G = 0$), the SEs of pMLE.logit were much larger than those of mpMLE, so that the CPs of pMLE were considerably smaller than the nominal level 95%. On the other hand, mpMLE performed reasonably well for both of the two parameter combinations. That is, the average estimates were close to

TABLE 1
Estimation by pMLE.logit and mpMLE.logit under the logistic regression model.

	BIAS	SE	SEE	CP(%)	BIAS	SE	SEE	CP(%)
	$f = 0.05$, $\beta_X = \log(4)$, $\beta_G = \log(1.3)$				$f = 0.005$, $\beta_X = \beta_G = 0$			
	pMLE.logit							
θ	0.000	0.011	0.011	0.947	0.000	0.014	0.011	0.944
α	0.010	0.120	0.122	0.954	0.014	0.102	0.074	0.878
β_S	-0.001	0.062	0.062	0.952	-0.005	0.046	0.031	0.891
β_G	-0.005	0.107	0.106	0.949	0.012	0.124	0.102	0.936
β_X	0.001	0.085	0.084	0.951	-0.006	0.085	0.057	0.899
	mpMLE.logit							
θ	0.000	0.011	0.011	0.947	0.000	0.011	0.011	0.954
α	0.010	0.120	0.122	0.952	0.017	0.075	0.074	0.943
β_S	-0.002	0.062	0.062	0.952	-0.006	0.031	0.031	0.946
β_G	-0.005	0.107	0.106	0.949	0.008	0.102	0.102	0.948
β_X	0.001	0.085	0.084	0.951	-0.004	0.057	0.056	0.947
	mpMLE.logit*							
θ	-0.001	0.011	0.011	0.945	0.000	0.011	0.011	0.954
α	2.813	0.132	0.132	0.000	2.741	0.086	0.085	0.000
β_S	0.693	0.068	0.068	0.000	-1.624	0.037	0.037	0.000
β_G	-0.013	0.107	0.106	0.947	0.008	0.102	0.102	0.948
β_X	0.146	0.094	0.093	0.670	-0.003	0.068	0.068	0.950

*Estimation results with seriously mis-specified f_s. BIAS, mean of the estimate minus the true parameter value; SE, empirical standard error of the estimates; SEE, mean estimated standard error of the estimate; CP, empirical coverage probability of the 95% confidence intervals.

the true values, the mean estimated asymptotic standard errors were close to the empirical standard errors, and the empirical coverage probabilities were close to the nominal level. Furthermore, the estimated multipliers presented in Table S1 (Appendix) were generally close to the limit values when the covariate effect was large. However, they could greatly deviate from the limit values when the covariate effect was small. These large biases could be due to the difficulty in identifying the multipliers in the zero-covariate-effect situation, which consequently resulted in slightly inflated type-I error rates (see Table S2 in Appendix for details). The multiplier estimates also had much larger interquartile ranges in the zero-covariate-effect situation compared with those in the non-zero-covariate-effect situation. To assess the impact of mis-specifying prevalences f_s on estimation, we specified f_1 to be five times its true value, and f_2 and f_3 to be their true values divided by five. As shown in Table 1, the bias in estimating the covariate effect was either small (the first situation) or ignorable (the second situation), and the bias in estimating the intercept and stratum-specific log-ORs can be large. Interestingly, the estimation of the genetic effect and MAF was largely unbiased.

TABLE 2
Power for testing the genetic effect under the logistic regression model (one covariate with non-zero effect, nominal level = 0.05)

f	LOGIT0	LOGIT1	pMLE.logit	mpMLE.logit	LT0	LT1
0.005	0.791	0.628	0.792	0.792	0.784	0.656
0.05	0.639	0.589	0.678	0.670	0.641	0.608
0.2	0.533	0.582	0.651	0.652	0.571	0.603
f				mpMLE.logit*	LT0*	LT1*
0.005				0.789	0.790	0.667
0.05				0.658	0.631	0.606
0.2				0.543	0.463	0.540

*Results with seriously mis-specified f_s.

We then evaluated the type-I error rate and power of all methods for testing the genetic effect ($H_0 : \beta_G = 0$) when a single non-confounding covariate was involved. All the methods maintained the nominal type-I error rate except for pMLE.logit under low or modest phenotype prevalence f and zero covariate effect (Table S3 in Appendix). The type-I error rate inflation might be due to the bias in pMLE.logit as mentioned above. Misspecification in prevalences hardly affected the type-I error rates for testing genetic association (Tables S2 and S4 in Appendix). In general, pMLE.logit and mpMLE.logit had similar or higher power than the other methods in the presence of a strong covariate effect (Table 2), with the power advantage depending on prevalence f. Under rare phenotype ($f = 0.005$) and non-zero covariate effect ($\beta_X = \log(4.0)$), LOGIT0, LT0, and mpMLE.logit had nearly identical power, and the power of mpMLE.logit was higher than LOGIT1 by 16.4% and LT1 by 13.6%. Under moderate f ($f = 0.05$) and non-zero covariate effect ($\beta_X = \log(4.0)$), the power advantage of mpMLE.logit over LOGIT1 and LT1 became smaller (8.9% and 7.4%, respectively), and the power became higher than LOGIT0 and LT0 (the difference was 3.9% and 3.7%, respectively). Under common phenotype ($f = 0.2$) and non-zero covariate effect ($\beta_X = \log(4.0)$), the power gain of mpMLE.logit over LOGIT1 and LT1 further reduced to 7.0% and 4.9%, respectively, and that over LOGIT0 and LT0 further increased to 11.9%, and 8.1%, respectively. As expected, all methods performed quite comparably in the absence of covariate effects (Table S3 in Appendix). With 10 covariates each having a log-OR log(2.0), the type-I error rates were all close to the nominal level (Table S4 in Appendix), and the power advantage of mpMLE.logit became much greater (Table 3). At $f = 0.005$, the power of mpMLE.logit was higher than LOGIT0, LOGIT1, LT0, and LT1 by 5.7%, 29.0%, 5.4%, and 19.1%, respectively. At $f = 0.05$, the power differences were 10.1%, 16.9%, 9.9%, and 12.3%, respectively. At $f = 0.2$, these power differences became 14.9%,

5.0%, 13.8%, and 4.3%, respectively. Here the results for pMLE.logit are not presented because they were almost the same as those for mpMLE.logit. All methods had comparable power in the absence of covariate effects (Table S7 in Appendix). In addition, the mis-specification in prevalences had minimum impact on the power for testing genetic association under low prevalence f ($f = 0.005$), but the power loss was evident as f increased to 0.2 (Tables 2 and 3). Interestingly, such decrease in power seemed to be more serious with stronger covariate effects, and was much smaller with weaker covariate effects (Tables S3 and S5 in Appendix). Similar power decrease was observed for LT0 and LT1 as well.

TABLE 3

Power for testing the genetic effect under the logistic regression model (10 covariates with non-zero effects, nominal level = 0.05)

f	LOGIT0	LOGIT1	mpMLE.logit	LT0	LT1
0.005	0.636	0.403	0.693	0.639	0.502
0.05	0.508	0.440	0.609	0.510	0.486
0.2	0.369	0.468	0.518	0.380	0.475
model			mpMLE.logit*	LT0*	LT1*
0.005			0.678	0.623	0.489
0.05			0.547	0.469	0.408
0.2			0.373	0.259	0.278

*Results with seriously mis-specified f_s.

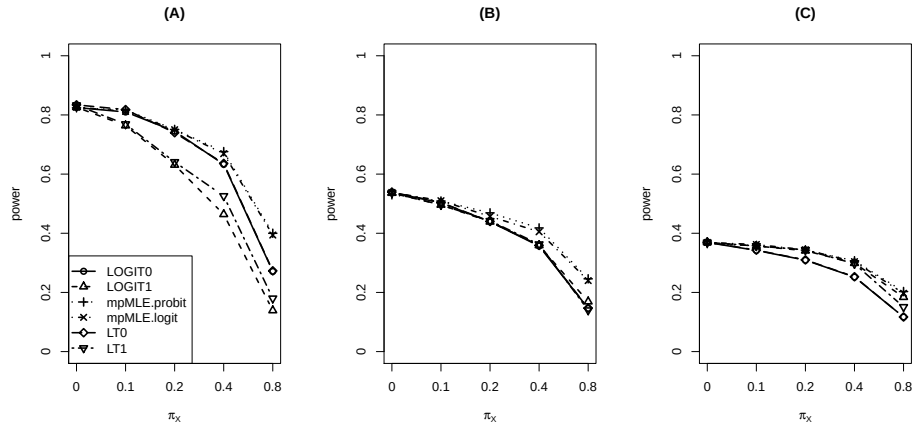

FIG 1. *Power as a function of the percent of variance explained by covariate (π_X) (zero stratum effect, probit model). (A) $f = 0.005$; (B) $f = 0.05$; (C) $f = 0.2$.*

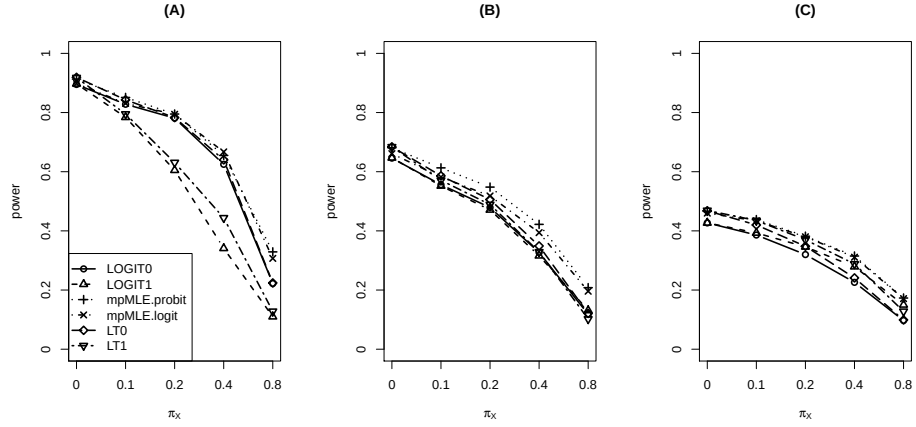

FIG 2. *Power as a function of the percent of variance explained by covariate (π_X) (non-zero stratum effect, probit model). (A) $f = 0.005$; (B) $f = 0.05$; (C) $f = 0.2$.*

3.2. *Under a probit regression model for penetrance.* In this subsection, we evaluated the performance of our method under a probit regression model for penetrance. The phenotype status D was generated from the probit regression model

$$\text{pr}(D = 1|S = s, X = x, G = g) = \Phi(\alpha + \beta_S s + \beta_X x + \beta_G g), \tag{3.2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal variable. The other aspects of the simulation setup were the same as those under the logistic penetrance model. Model (3.2) is equivalent to the liability threshold model. Specifically, the phenotype D can also be defined by dichotomizing a normally distributed variable $Y := \alpha + \beta_S S + \beta_X X + \beta_G G + e$ with e being the standard normal random error term, $D = 1$ if $Y \geq \tau$ and $D = 0$ if $Y < \tau$ for some threshold τ. We note that the matching variable S is the "covariate" in the LT1 and LT0 methods. We fixed β_G at 0.1, so that the genetic effect was small compared with that of the random error. In all simulation scenarios, mpMLE.probit was virtually unbiased, its mean estimated standard errors were close to the empirical standard errors, and its empirical coverage probabilities were close to the nominal level (refer to Table S6 in Appendix for results in one simulation scenario). We focused on the power of mpMLE.probit as a function of the variance of Y explained by the non-confounding covariate X, π_X, which was approximately

$$\pi_X = \frac{\beta_X^2 \text{var}(X)}{\beta_S^2 \text{var}(S) + \beta_X^2 \text{var}(X) + 1}.$$

We considered π_X values of 0, 0.1, 0.2, 0.4, and 0.8. The variance of Y explained by the stratum variable S, denoted as π_S, was approximately

$$\pi_S = \frac{\beta_S^2 \text{var}(S)}{\beta_S^2 \text{var}(S) + \beta_X^2 \text{var}(X) + 1}.$$

We fixed β_S at either 0 or log(2.0) so that $\pi_S = 0$ or $\pi_S > 0$. Setting $\beta_S = 0$ allowed us to assess the impact of including covariate X on the power of the LT methods as no such results were available.

The type-I error rates of all methods were close to 0.05 in the range of 0.04 and 0.06, except that LT1 was slightly conservative at large values of π_X (Figure S2 and Figure S3 in Appendix). For all the methods considered, the power steadily decreased with increasing π_X. The proposed method was generally more powerful, and the relative power depended on both π_X, π_S, and phenotype prevalence f (Figs 1 and 2). In the absence of non-genetic effects ($\pi_S = 0$ and $\pi_X = 0$), all methods had comparable power as expected (Fig 1 (A), (B), and (C): $\pi_X = 0$). With $\pi_S = 0$ and $\pi_X > 0$, mpMLE.probit and mpMLE.logit were uniformly more powerful, and their power advantage generally increased in π_X (Fig 1 (A), (B), and (C)). In the presence of stratum effect but absence of additional covariate effect ($\pi_S > 0$ and $\pi_X = 0$), the three methods that correctly specified the penetrance model (i.e., mpMLE.probit, LT0, and LT1) were most powerful and had comparable power (Fig 2 (A) , (B), and (C)). With $\pi_S > 0$ and $\pi_X > 0$, mpMLE.probit was uniformly more powerful than all the other methods (Fig 2 (A), (B), and (C)). The population phenotype prevalence f appeared to have large influence on the relative power. When f was low (0.005) or moderate (0.05) (Figs 1 and 2, (A) and (B)), two existing methods adjusting for covariates (no covariate specific prevalence information was available), LOGIT1 and LT1, were the least powerful. For example, their power was lower by 12.0% and 11.0%, respectively, compared with mpMLE.logit when $\pi_X = 0.2$, $f = 0.005$, and $\pi_S = 0$. Notably, mis-specifying the penetrance model as the logistic regression model only resulted in minor power loss ($\leq 3.1\%$) for all methods.

3.3. *When non-confounding covariates consist of 10 common SNPs.* To assess the extent of power improvement of mpMLE for testing genetic association through adjustment of common susceptible variants, we used the same setup as Section 3.1 except that 10 independent SNPs were used as covariates. The corresponding 10 MAFs were randomly sampled from the uniform distribution on the interval (0.05, 0.5). Genotypes for each SNP were generated under HWE and coded as the minor allele count in the analysis. The 10 ORs were randomly sampled from the uniform distribution on

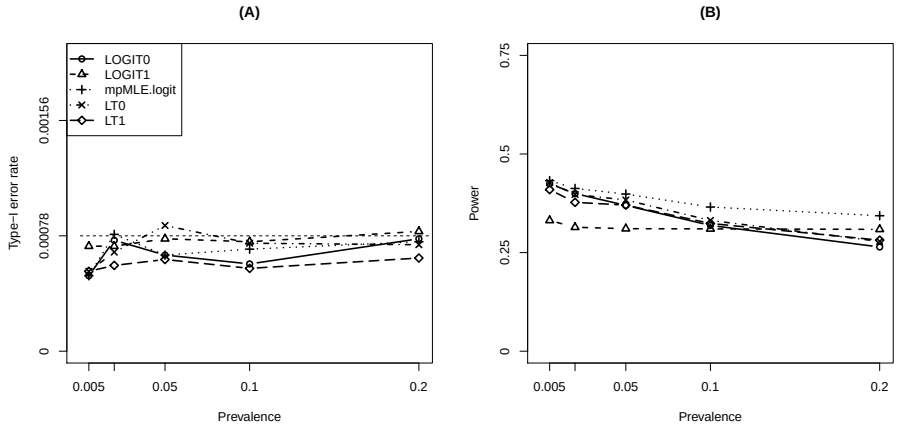

FIG 3. *Type-I error rates and power for testing a common variant (MAF=0.2) at different phenotype prevalence rates when 10 SNPs with moderate effects were adjusted for as covariates (nominal level = 0.05/64). The numbers of cases and controls were equal ($r_{ss} = 1$), and the SNP log-OR β_G was set at* $\log(1.3)$ *for calculating power.*

the interval $(\log(1.3), \log 2)$. We chose to consider larger effect sizes, because SNP covariates with small effects would lead to limited power increase. We expect that the results will inform the analyses when a larger number of SNPs with small effects are adjusted for as covariates, and it is infeasible to analyze for a reasonable sample size even by standard logistic regression method. We evaluated the relative power of mpMLE for testing a common (MAF=0.2) or less common (MAF=0.05) variant with a weak to moderate OR ($\beta_G = 0$, $\log(1.1)$, $\log(1.2)$, $\log(1.3)$, or $\log(1.4)$). We considered a wide range of values for phenotype prevalence ($f = 0.005$, 0.02, 0.05, 0.1, and 0.2) and control versus case sample size ratio common to three strata ($r_{ss} = 0.5$, 1, 2, and 3). To inform the real data example in Section 4 where 64 SNPs were considered, all the tests were performed at significance level $0.05/64 \approx 7.81 \times 10^{-4}$. The total number of cases and controls in each of the three sampling strata was fixed at 600 for testing the common variant and 1200 for testing the less common variant. We generated 100,000 datasets for estimating type-I error rates and 10,000 for estimating power.

Figs 3, 4, and 5 display type-I error rates and power for testing the common variant as a function of prevalence f, control versus case ratio r_{ss}, and genetic effect log-OR β_G, respectively. Results for testing the less common variant were similar (Figures S4, S5, and S6 in Appendix). The type-I error rates were close to the nominal level for all the considered methods. As shown

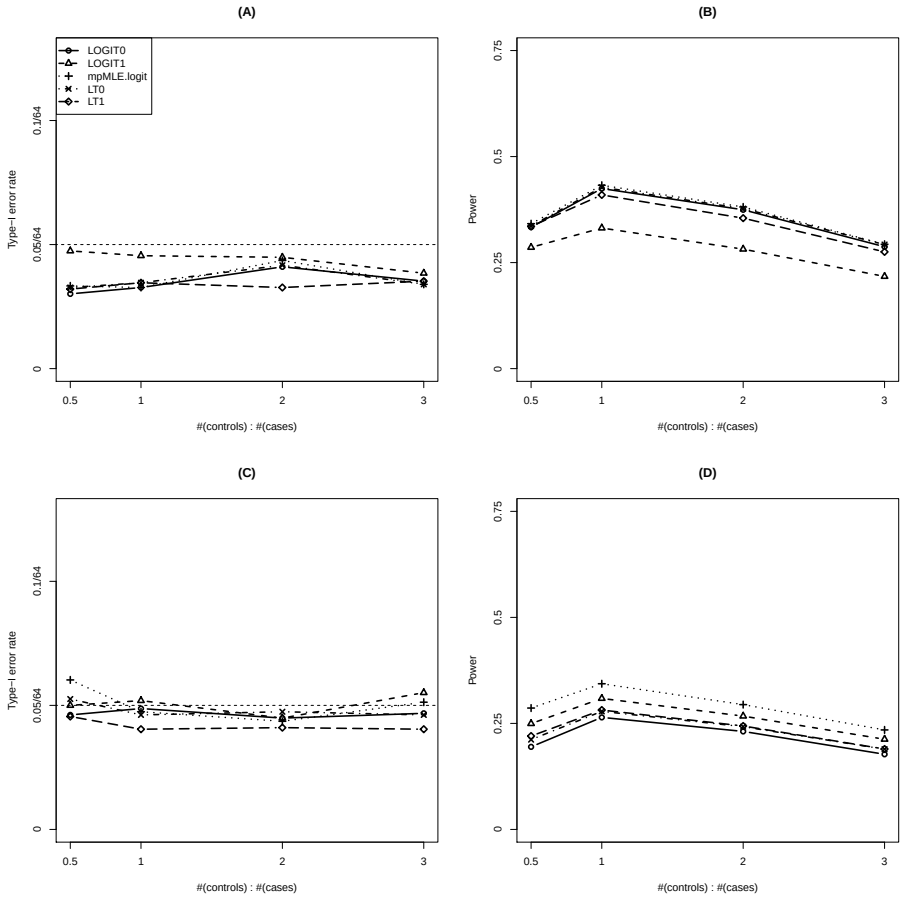

FIG 4. *Type-I error rates and power for testing a common variant (MAF=0.2) at different case versus control ratios when 10 SNPs with moderate effects were adjusted for as covariates. The numbers of cases and controls were equal. The phenotype prevalence f for panels (A) and (B) was 0.005 and for panels (C) and (D)* 0.2.

in Fig 3, at a fixed r_{ss}, the power of all methods decreased slowly with f in a nearly linear fashion, and that of mpMLE was consistently the highest. The two existing methods that do not adjust for covariates (LOGIT0 and LT0) had higher power when $f = 0.005$, but they became less powerful with large f. When f exceeded some threshold, the two existing methods that adjust for covariates (LOGIT1 and LT1) started to be more powerful than LOGIT0 and LT0, and the power difference increased with f. The impact of f on the power difference also appeared to be model dependent. That is, the power difference under the logistic regression model (LOGIT1 and LOGIT0)

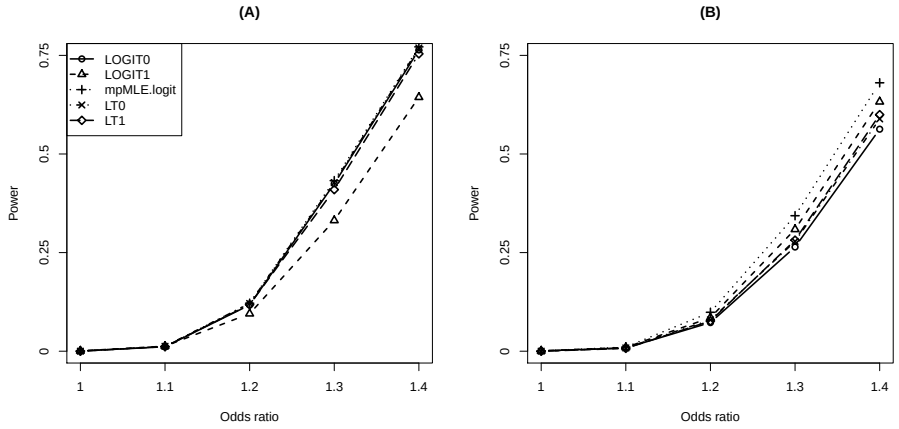

FIG 5. *Power for testing a common variant (MAF=0.2) at different ORs when 10 SNPs with moderate effects were adjusted for as covariates. The numbers of cases and controls were equal. Panels (A) and (B) shows results when the phenotype prevalence $f = 0.005$ and 0.2, respectively.*

can be quite noticeable, while that under the liability threshold model (LT0 versus LT1) was much smaller. This model dependency might be due to the fact that LT0 and LT1 incorporated stratum-specific phenotype prevalence while LOGIT0 and LOGIT1 did not. At given β_G, the maximal power was achieved under equal numbers of cases and controls ($r_{ss} = 1$) regardless of the phenotype prevalence, and the power difference between various methods stayed nearly constant with r_{ss} (Fig 4). When the phenotype was rare ($f = 0.005$), the power of mpMLE was nearly identical to LOGIT0, LT0, and LT1, and the power difference between mpMLE and LOGIT1 increased with β_G (Fig 5). The power difference between mpMLE and all the other methods increased with β_G when the phenotype was common ($f = 0.2$), and that between mpMLE and LOGIT0 was the largest.

4. Analysis of a case-control genetic association study of high-density lipoprotein cholesterol. We apply our proposed method to analyze data from a case-control genetic association study of high-density lipoprotein cholesterol (HDL-C). This study aimed to identify genetic variants contributing to variation in HDL-C levels (Edmondson et al., 2011). Subjects of European ancestry were recruited from the University of Pennsylvania hospital, where cases were defined as those with HDL$> 90^{\text{th}}$ percentile for any fixed age and gender, and controls as those with HDL$< 30^{\text{th}}$ percentile for any fixed age and gender. Data on covariates gender, age,

weight, height, body mass index (BMI), and smoking status were available for 625 cases and 606 controls, who were included in the current analysis. We applied pMLE.logit, mpMLE.logit, LOGIT0, LOGIT1, LT0, and LT1 to analyze 64 di-allelic SNPs in 13 candidate genetic regions (*PCSK5*, *NR1H3*, *FADS1-2-3*, *MVK/MMAB*, *LCAT*, *APOE*, *PLTP*, *GALNT2*, *LPL*, *ABCA1*, *LIPC*, *CETP*, and *LIPG*), which had been previously reported to be associated with the HDL-C level (Edmondson et al., 2011). We do not report the results of pMLE.logit since they are extremely close to those of mpMLE.logit for 63 SNPs and pMLE.logit failed to converged for the other SNP. Standard logistic regression analyses of baseline covariates revealed that gender, age, BMI, and smoking status were significantly associated with HDL-C at significance level 0.05 (Table S7 in Appendix). All the SNPs were on autosomes, and we assumed that the considered SNPs were associated with neither gender nor age. The p-values of the Pearson Chi-squared tests using data from controls for examining associations between each SNP and BMI or smoking were all greater than 0.01 and greater than 0.64 after Bonferroni adjustment. We therefore treated gender, age, BMI, or smoking as non-confounding covariates in our analyses.

TABLE 4
P-values for genetic association analysis in the HDL-C study ($\times 10^{-4}$)

Gene	SNP	LOGIT0	LOGIT1	mpMLE	LT0	LT1
LPL	rs301	12.3	**5.09***	**6.97**	1.18	**7.76**
LPL	rs328	8.71	21.6	8.38	7.89	**4.65**
LPL	rs256	17.3	13.6	**4.77**	16.4	13.8
LPL	rs264	16.4	15.3	**6.19**	15.6	15.9
LPL	rs12679834	13.6	30.8	11.9	12.5	**7.06**
LIPC	rs11635491	12.2	**4.62**	**1.89**	11.7	16.6
LIPC	rs1800588	17.7	**4.41**	**3.92**	17.0	19.6
LIPC	rs2070895	13.0	**3.10**	**3.00**	12.4	16.2
LIPC	rs261332	10.8	**2.46**	**2.41**	10.2	16.8

*Bold numbers indicate significant results (p-value $< 0.05/64 \approx 7.81\times 10^{-4}$).

All the four significantly associated covariates were adjusted for in LOGIT1, LT1, and mpMLE.logit. The "phenotype prevalence" in this study was 0.1/(0.1+0.3)=0.25, and we incorporated it in mpMLE.logit, LT0, and LT1. We compared results from all methods with respect to the number of significant SNPs identified and the standard errors of SNP log-OR estimates. After Bonferroni correction at significance level 0.05, LOGIT0, LT0, LOGIT1, LT1, and mpMLE.logit respectively identified 24, 24, 29, 27, and 31 significantly associated SNPs (p-value $< 0.05/64 \approx 7.81\times 10^{-4}$), of which 24 were identified by all the considered methods (Tables S8-S10 in Appendix). Table 4 displays p-values for SNPs that were identified by at least one method but

not all. In this study, the "phenotype prevalence" 0.25 was high and the covariate "smoking status" had a strong effect on HDL-C (log-OR = 1.873, p-value = 4.5×10^{-25}). According to our simulation results (Fig 1), mpMLE.logit was expected to be the most powerful, while LOGIT0 and LT0 were expected to be the least powerful. Indeed, the relative number of significantly associated SNPs by each method in this study perfectly conformed with the results from our simulation studies. In general, the estimated log-ORs by mpMLE.logit and LOGIT1 were comparable (Figure S4(A) in Appendix), and the standard errors for most SNPs by mpMLE.logit were slightly smaller than those by LOGIT1 (Figure S4(B) in Appendix).

5. Discussion. For analyzing case-control genetic association studies, we proposed a novel profile likelihood method that guarantees power improvement through the adjustment of non-confounding covariates. Our simulation results suggested that the extent of power improvement can be substantial, and that it consistently outperformed the existing methods in the simulation studies. Therefore, our method can lead to increased chance of discovering new genetic variants in future genetic association studies. It relieves data analysts from the burden of having to decide whether to adjust for non-confounding covariates, which is particularly convenient when analyses need to be conducted in multiple ethnic subgroups where phenotype prevalences differ. Furthermore, inconsistent adjustment of covariates across different studies can lead to heterogeneity in estimated effect sizes for a genetic variant. Our method encourages adjustment of non-confounding covariates in all future GWAs to increase power and reduce effect heterogeneity across studies. We have developed an R package CCGA to implement our method. CCGA is freely available at Github (http://github.com/zhanghfd/CCGA).

Our limit multiplier profile likelihood estimator is statistically efficient, computationally stable, and robust to mis-specification in phenotype prevalence. The analysis of the HDL-C study demonstrates the good performance of our method in practical settings. Results from simulation studies supported our conjecture that failure to explicitly incorporate gene-covariate independence into statistical inference resulted in the covariate adjustment dilemma in case-control genetic association studies. We note that Zaitlen et al. (2012a,b) implicitly used gene-covariate independence for estimating covariate effects under the null. It is difficult to make a general statement on the exact circumstances when our method substantially outperforms the existing ones. We are currently deriving analytical results for asymptotic relative efficiency of our method in separate work.

It was interesting that the power of mpMLE.probit decreased with the

amount of variation explained by covariates X, π_X, under the probit penetrance model in our simulation studies. In fact, this same phenomenon has been explained in the literature (Neuhaus and Jewell, 1993; Neuhaus, 1998; Stringer et al., 2011; Pirinen, Donnelly and Spencer, 2012). The power for testing the genetic effect increases with π_X in linear regression analyses as commonly known. On the other hand, in logistic regression, both the estimated log-OR and its standard error estimates become larger with adjustment of X. However, when the phenotype is rare, the increase in the log-OR estimate generally cannot compensate the increase in its standard error estimate, leading to decreased power. The extent of decrease depends on effect sizes of the covariates, as observed in our simulation studies and previous work (Neuhaus and Jewell, 1993; Neuhaus, 1998; Stringer et al., 2011; Pirinen, Donnelly and Spencer, 2012).

Our method was developed for increasing power for testing genetic association. But it can also be used to test any non-genetic exposure of interest. We found that estimation through maximization of the profile likelihood function, which we derived using techniques similar as those in Chatterjee and Carroll (2005), was often infeasible due to computation difficulty. We therefore proposed to replace the estimated Lagrange multipliers in the profile likelihood function with their limit values, which fully resolved the computation issue without compromising statistical efficiency. An important theoretical finding in this work is that the limit multipliers had simple closed forms, which were functions of phenotype prevalence and case-control sampling ratios. We note that mpMLE is a novel method contributed by this work. It involved multipliers with non-zero limits due to the incorporation of prevalence information and distribution constraints. Working with the limits of the Lagrange multipliers instead of the estimated multipliers also allowed much easier assessment of the asymptotic behavior of the estimator.

In this paper, we considered the situation when neither the matching variable S nor covariate X confound the phenotype-gene association. Our method can be extended to allow conditional independence of G and X given S. The HWE may not be satisfied in the presence of population stratification and other confounding factors, but our method is applicable without imposing constraints on the genotype distribution. Our method can be extended to incorporate additional confounding factors that could be correlated with both genetic variants and phenotype status. Let Z be a confounder variable, and assume additionally that X is conditionally independent of (Z, G) given S. We can extend the probability decomposition (2.2) as $\mathrm{pr}(X = X_s, G = g, Z = Z_s|S = s) = \mathrm{pr}(G = g|S = s)\mathrm{pr}(X = X_s|S = s)\mathrm{pr}(Z = Z_s|G = g, S = s)$, where the additional nuisance distri-

bution $\text{pr}(Z|G = g, S = s)$ in the likelihood can be either replaced by the empirical distribution and profiled out or be modelled by some suitable generalized linear model. Consequently, the proposed method can be naturally extended to incorporate Z.

We focused on testing common genetic variants (MAF ≥ 0.05) in the current work. For analyzing rare variants (MAF < 0.05), our method is directly applicable for testing the burden (Li and Leal, 2008) of a set of rare variants. Along the line of variance component based tests (Wu, 2001; Benjamin et al., 2011; Lee et al., 2012; Sun and Hsu, 2013), our method can potentially be extended to accommondate more sophiscated models for rare variants. We will address these additional challenges in future work. It will be pertinent to evaluate the extent of power improvement through adjustment of common susceptible variants.

The interpretations of the genetic effect with or without adjusting for covariates are different, which is not an issue for testing genetic associations. When it is of interest to assess the marginal effect of genetic variants without adjusting for covariates, efficiency gain can be expected through incorporation of covariates into the statistical inference. Approaches along this line have been developed for the estimation and testing of marginal treatment effects by exploiting baseline covariates (Zhang, Tsiatis and Davidian, 2008). We will develop similar methods in the setting of case-control genetic association studies, where the retrospective sampling design poses interesting statistical challenges. Development of such methods can help answer the question whether the discovered genetic variants are of value for improving statistical efficiency in making inference on the marginal effect of new genetic variants.

SUPPLEMENTARY MATERIAL

Appendix: Proof of Theorems 1-2 and equations (2.13) and (2.14), Figures S1-S7, and Tables S1-S10.
(http://homepage.fudan.edu.cn/zhangh/softwares/CCGA/).

References.

Benjamin, M. N., Manuel, A. R., Benjamin, F. V., David, A., Bernie, D., Marju, O., Sekar, K., Shaun, M. P., Kathryn, R. and Mark, J. D. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics* **e1001322**.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92** 399–418.

Edmondson, A. C., Braund, P. S., Stylianou, I. M., Khera, A. V., Nelson, C. P., Wolfe, M. L., DerOhannessian, S. L., Keating, B. J., Qu, L., He, J., Tobin, M. D., Tomaszewski, M., Baumert, J., Klopp, N., Dring, A., Thorand, B.,

Li, M., Reilly, M. P., Koenig, W., Samani, N. J. and Rader, D. J. (2011). Dense genotyping of candidate gene loci identifies variants associated with high-density lipoprotein cholesterol. *Circulation: Cardiovascular Genetics* **4** 145–155.

Kuo, C.-L. and Feingold, E. (2010). What's the best statistic for a simple test of genetic association in a case-control study? *Genetic Epidemiology* **34** 246–253.

Lee, S., Emond, J., M., Bamshad, J., M., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Team, N. G. E. S. P. L. P., Christiani, D. C., Wurfel, M. M. and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics* **91** 224–237.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics* **83** 311–321.

Neuhaus, J. (1998). Estimation efficiency with omitted covariates in generalized linear models. *Journal of the American Statistical Association* **93** 1124–1129.

Neuhaus, J. and Jewell, N. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear-models. *Biometrika* **80** 807–815.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics* **18** 90–120.

Pirinen, M., Donnelly, P. and Spencer, C. C. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics* **44** 848-851.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22** 300–325.

Qin, J., Zhang, H., Li, P., Albanes, D. and Yu, K. (2014). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika* **102** 169–180.

Stringer, S., Wray, N. R., Kahn, R. S. and Derks, E. M. (2011). Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. *PLoS ONE* **6** e27964. doi:10.1371/journal.pone.0027964.

Sun, Z. Y. J. and Hsu, L. (2013). A unified Mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology* **37** 334–344.

Wu, L. S. C. T. L. Y. B. M. e. a. M. C (2001). Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am J Hum Genet* **891** 82–93.

Zaitlen, N., Paşaniuc, B., Patterson, N., Pollack, S., Voight, B., Groop, L., Altshuler, D., Henderson, B. E., Kolonel, L. N., Le Marchand, L., Waters, K., Haiman, C. A., Stranger, B. E., Dermitzakis, E. T., Kraft, P. and Price, A. L. (2012a). Analysis of case-control association studies with known risk variants. *Bioinformatics* **28** 1729–1737.

Zaitlen, N., Lindström, S., Pasaniuc, B., Cornelis, M., Genovese, G., Pollack, S., Barton, A., Bickeböller, H., Bowden, D. W., Eyre, S., Freedman, B. I., Friedman, D. J., Field, J. K., Groop, L., Haugen, A., Heinrich, J., Henderson, B. E., Hicks, P. J., Hocking, L. J., Kolonel, L. N., Landi, M. T., Langefeld, C. D., Le Marchand, L., Meister, M., Morgan, A. W., Raji, O. Y., Risch, A., Rosenberger, A., Scherf, D., Steer, S., Walshaw, M., Waters, K. M., Wilson, A. G., Wordsworth, P., Zienolddiny, S., Tchetgen, E. T., Haiman, C., Hunter, D. J., Plenge, R. M., Worthington, J., Christiani, D. C., Schaumberg, D. A., Chas-

MAN, D. I., ALTSHULER, D., VOIGHT, B., KRAFT, P., PATTERSON, N. and PRICE, A. L. (2012b). Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genetics* **8** e1003032.

ZHANG, M., TSIATIS, A. A. and DAVIDIAN, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64** 707–715.

H. ZHANG
INSTITUTE OF BIOSTATISTICS
FUDAN SCHOOL OF LIFE SCIENCES
SHANGHAI, P.R. CHINA
E-MAIL: zhanghfd@fudan.edu.cn

N. CHATTERJEE
DIVISION OF CANCER EPIDEMIOLOGY AND GENETICS
NATIONAL CANCER INSTITUTE
NATIONAL INSTITUTES OF HEALTH, MA
E-MAIL: chattern@mail.nih.gov

D. RADER
DEPARTMENT OF GENETICS
PERELMAN SCHOOL OF MEDICINE
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PA
E-MAIL: rader@mail.med.upenn.edu

J. CHEN
DEPARTMENT OF BIOSTATISTICS AND EPIDEMIOLOGY
PERELMAN SCHOOL OF MEDICINE
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PA
E-MAIL: jinboche@mail.med.upenn.edu